UNIVERSIDADE FEDERAL DO PARANÁ

ALEXANDRE PEREIRA DE FARIA

AVALIAÇÃO DO DESEMPENHO HUMANO EM AMBIENTES VIRTUAIS PARA TREINAMENTO PROFISSIONAL DE ATIVIDADES CRÍTICAS POR MEIO DE AGRUPAMENTO DE PADRÕES DO ERRO

CURITIBA

2021

ALEXANDRE PEREIRA DE FARIA

AVALIAÇÃO DO DESEMPENHO HUMANO EM AMBIENTES VIRTUAIS PARA TREINAMENTO PROFISSIONAL DE ATIVIDADES CRÍTICAS POR MEIO DE AGRUPAMENTO DE PADRÕES DO ERRO

Trabalho apresentado como requisito parcial para a obtenção do título de Doutor em Ciências pelo Programa de Pós Graduação em Métodos Numéricos em Engenharia do Setor de Tecnologia da Universidade Federal do Paraná.

Orientador: Prof. Sérgio Scheer, Dr.
Coorientador: Prof. Klaus de Geus, Dr

CURITIBA

2021

Catalogação na Fonte: Sistema de Bibliotecas, UFPR
Biblioteca de Ciência e Tecnologia

F224a

Faria, Alexandre Pereira de
Avaliação do desempenho humano em ambientes virtuais para treinamento profissional de atividades críticas por meio de agrupamento de padrões do erro [recurso eletrônico] / Alexandre Pereira de Faria. – Curitiba, 2021.

Tese - Universidade Federal do Paraná, Setor de Tecnologia, Programa de Pós-Graduação em Métodos Numéricos em Engenharia, 2021.

Orientador: Sérgio Scheer – Coorientador: Klaus de Geus.

1. Ambientes virtuais compartilhados. 2. Modelagem – Conhecimento e aprendizagens. 3. Desempenho. 4. Tecnologias de avaliação do desempenho escolar. I. Universidade Federal do Paraná. II. Scheer, Sérgio. III. Geus, Klaus de. IV .Título.

CDD: 006.0785

Bibliotecário: Elias Barbosa da Silva CRB-9/1894

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação MÉTODOS NUMÉRICOS EM ENGENHARIA da Universidade Federal do Paraná foram convocados para realizar a arguição da tese de Doutorado de **ALEXANDRE PEREIRA DE FARIA** intitulada: **AVALIAÇÃO DO DESEMPENHO HUMANO EM AMBIENTES VIRTUAIS PARA TREINAMENTO PROFISSIONAL DE ATIVIDADES CRÍTICAS POR MEIO DE AGRUPAMENTO DE PADRÕES DO ERRO**, sob orientação do Prof. Dr. SÉRGIO SCHEER, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de doutor está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 27 de Outubro de 2021.

Assinatura Eletrônica
28/10/2021 09:47:52.0
SÉRGIO SCHEER
Presidente da Banca Examinadora

Assinatura Eletrônica
28/10/2021 09:13:52.0
PAULO HENRIQUE SIQUEIRA
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
03/11/2021 08:13:50.0
LEONELO DELL ANHOL ALMEIDA
Avaliador Externo (UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ)

Assinatura Eletrônica
28/10/2021 09:20:37.0
CASSIUS TADEU SCARPIN
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
09/11/2021 11:16:02.0
WALMOR CARDOSO GODOI
Avaliador Externo (UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ)

AGRADECIMENTOS

Este trabalho não seria possível sem a colaboração de várias pessoas.

Em primeiro lugar, agradeço aos professores Sérgio Scheer e Klaus de Geus pela oportunidade, confiança, incentivo e paciência ao longo desta jornada. Também agradeço aos professores Leonelo Almeida, Walmor Godoi, Paulo Henrique Siqueira e Cassius Scarpin que participaram na banca de defesa e o professor Awdry Miquelin no exame de qualificação, por suas contribuições.

Agradeço aos colegas e professores do PPGMNE com os quais pude aprender constantemente. Ao Jair e a Jully da secretária do PPGMNE por todo auxílio, sempre prontamente respondido, meu muito obrigado.

Este trabalho foi desenvolvido junto ao OneReal Research Group no âmbito do projeto de P&D PD-6491-0299/2013 proposto pela Copel Geração e Transmissão S.A., sob os auspícios do Programa de P&D da Agência Nacional de Energia Elétrica (ANEEL).

Aos membros do OneReal devo meus sinceros agradecimentos e, em particular, Ricardo Santos e Elton Sato pela ajuda com o banco de dados e imagens do sistema RV2. Um agradecimento especial ao gerente do projeto RV2, Eng. Eletricista Rafael Bee, pela colaboração na análise da tarefa e modelagem do domínio do conhecimento do especialista.

Quem, se eu gritasse dentre as legiões dos Anjos me ouviria?
(Raine Maria Rilke, Elegias de Duíno)

RESUMO

Atividades de manutenção em subestações elétricas em linha viva são consideradas de alto risco, pois implicam a atuação sobre um sistema complexo e sujeito a situações inesperadas e perigosas. Neste cenário crítico, as intervenções humanas demandam atenção aos procedimentos de segurança e à prevenção e controle do erro humano. Se por um lado, a necessidade de treinamento é prontamente reconhecida, por outro lado, a realização de treinamentos em uma subestação real pode ser difícil ou, dependendo do cenário de treinamento, até mesmo impossível. Neste caso, treinamentos em ambientes virtuais representam uma alternativa tanto para a formação profissional quanto para os estudos de confiabilidade. Este trabalho tem como objetivo desenvolver um método de avaliação do desempenho de eletricistas profissionais com base na análise dos erros cometidos durante a execução de atividades críticas em um ambiente de treinamento virtual. O desenvolvimento proposto segue uma estrutura de pesquisa adaptada da *Design Science Research*. A partir da conscientização do problema acerca do erro humano, dois modelos de padrões de dados são propostos, o primeiro, baseado nos padrões de erros extraídos a partir da análise da tarefa e o auxílio de um especialista, o segundo, baseado nos padrões de similaridade, onde o domínio do conhecimento é modelado como um grafo. Os agrupamentos *K-means* definidos sobre espaços de dados modelados tanto como padrões de erros quanto como padrões de similaridade de grafos são analisados e combinados como agrupamentos consensos. O resultado desta análise mostrou que a medida de similaridade entre grafos baseada na distância Resistência Normalizada gera os melhores grupos e apresenta uma boa granularidade representando uma boa alternativa para a avaliação automática do desempenho dos treinandos em sistemas virtuais. Trabalhos futuros devem investigar como a medida de resistência entre grafos captura desvio mínimos associados a certos tipos de erros permitindo a sua identificação automática.

Palavras-chaves: Avaliação do desempenho humano. Mineração de dados educacionais. Similaridade de grafos. Modelagem do conhecimento. Ambientes virtuais de treinamento.

ABSTRACT

Live line maintenance activities in electrical substations are considered high risk since they imply operating in a complex system and are subject to unexpected and dangerous situations. In this critical scenario, human interventions demand attention to safety procedures and the prevention and control of human error. Although the need for training is readily recognized, training in a real substation may be difficult or, depending on the training scenario, even impossible. In this case, training in virtual environments represents an alternative for both professional training and reliability studies. This paper aims to propose a method to assess the performance of professional electricians while performing critical activities by looking for the errors made. The proposed development follows a research framework adapted from *Design Science Research*. From the human error analysis, two data pattern models are proposed, the first, based on the error patterns extracted from the task analysis and the help of an expert, the second, based on the similarity patterns, where the knowledge domain is modeled as a graph. The clusters *K-means* defined over data spaces modeled as error patterns and as graph similarities patterns are analyzed and combined as consensus clusters. The results showed that the graph similarity measure based on the Normalized Resistance distance generates the best clustering and exhibits good granularity representing a good alternative for automatic evaluation of trainee performance in virtual systems. Future work should investigate how the inter-graph similarity measure captures minimum deviations to identify specific error types.

Key-words: Human performance assessment. Educational data mining. Graph similarity. Knowledge modeling. Virtual training environment.

SUMÁRIO

1 INTRODUÇÃO

As aplicações computacionais de jogos digitais em educação e treinamento, conhecidas como jogos sérios, representam um importante ponto de convergência entre o mundo do entretenimento e o mundo do trabalho. A utilização de tecnologias de interação em ambientes virtuais 3D, as lições sobre a experiência do usuário, o design de jogos digitais, o projeto de interfaces adaptadas aos sentidos e a comunicação humana têm proporcionado a construção de experiências simuladas mais próxima do real (BURIOL, 2011; HERNÁNDEZ et al., 2016; GARANT, 1997).

1.1 TEMA

Se em sua gênese, em meados do século XX, as aplicações dessa natureza estavam restritas ao treinamento militar, como os simuladores de voos que reproduziam a cabine do piloto, atualmente, o mesmo conceito é utilizado na formação de condutores de automóveis nas conhecidas “autoescolas”.

A utilização de ambientes virtuais no treinamento profissional tem sido explorado em diversas áreas como plantas industriais (CHENG; HWANG, 2015), saúde (GALLAGHER et al., 2005), militar (MAXWELL, 2015), instalação e manutenção de equipamentos, movimentação de cargas, substâncias perigosas (CAI et al., 2013; MÄÄTTÄ, 2003) e operações de risco (XI et al., 2009). Os ambientes virtuais baseados em técnicas de realidade virtual simulam o mundo real e permitem uma experiência de imersão utilizando métodos de interação diferentes como a manipulação de objetos virtuais e a resposta ao esforço físico (GUPTA et al., 2008b).

Ambientes virtuais 3D que combinam áudio e imagens realísticas e estereoscópicas, estímulos táteis e um sistema de rastreamento podem aumentar a sensação de imersão no ambiente (GUPTA et al., 2008a) e maximizam o seu engajamento às tarefas e objetivos do treinamento. A idealização de cenários críticos em treinamentos reforçam boas práticas e atenção às normas de segurança de forma compartilhada (LOUPOS et al., 2008). A reflexão e o planejamento de soluções específicas de forma coordenada evitam atitudes mecânicas que podem colocar em risco tanto a vida do operador ou da equipe quanto a integridade de equipamentos.

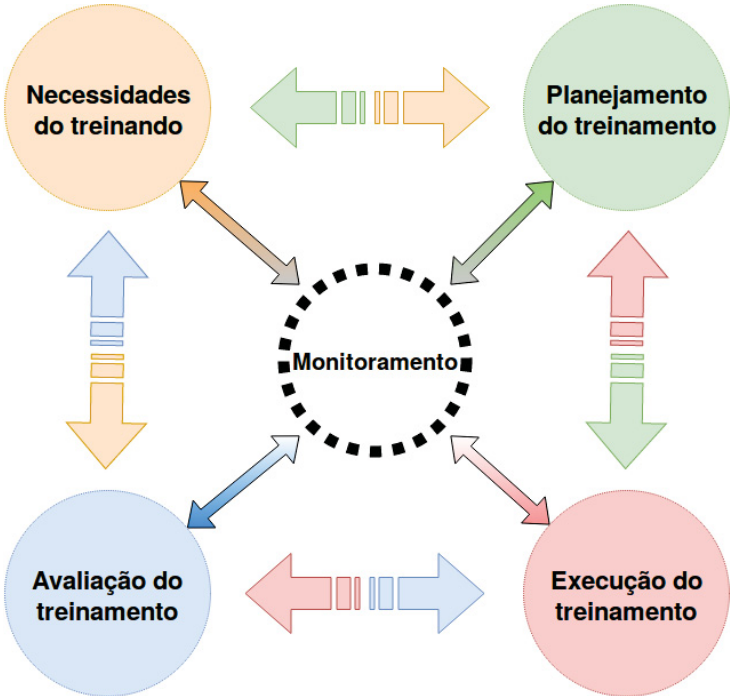
1.2 JUSTIFICATIVA

As atividades de manutenção em subestações elétricas se caracterizam por seu alto grau de periculosidade, trazendo risco à vida dos trabalhadores que atuam no setor elétrico. Os acidentes têm causas diversas, como quedas ou descargas de origem

elétrica, e decorrem de métodos ou procedimentos arriscados. Tal fato torna evidente a importância dos programas de treinamento que abordem práticas reflexivas sobre o trabalho e desenvolvam no treinando a capacidade de avaliação sobre os riscos e meios para o planejamento seguro de sua ação (FUNCOGE, 2013).

A norma regulamentadora 10, NR 10 (BRASIL, 2004), descreve os requisitos e condições para a implementação de medidas e sistemas de prevenção que garantam a segurança e a saúde dos trabalhadores que estão expostos ao risco elétrico. Essa norma exige que os trabalhadores tenham acesso a treinamentos que incluam a introdução à segurança com eletricidade, técnicas de análise e medidas de risco elétrico, utilização de equipamentos de proteção, organização e procedimentos de trabalho. A reciclagem por meio de treinamentos adicionais, que já é obrigatória, por exemplo, no caso de implementação de inovações tecnológicas (BRASIL, 2004), torna-se imprescindível diante de cenários que fujam aos procedimentos padrão ou se configurem numa situação considerada crítica, onde a chance de acidente se torna iminente.

FIGURA 1 – ETAPAS DO TREINAMENTO



FONTE: O Autor, baseado em ABNT (2001)

Conforme a norma NBR ISO 10015, para que o treinamento cumpra seu objetivo de fornecer as competências necessárias ao treinando em condições de excelência, quatro etapas apresentadas na FIGURA 1 devem ser monitoradas pelos gerentes responsáveis (ABNT, 2001):

- Definição das necessidades do treinamento;
- Projeto e planejamento do treinamento;
- Execução do treinamento;
- Avaliação dos resultados do treinamento.

As diretrizes para a elaboração da avaliação do treinamento têm como base princípios instrucionais como a análise e revisão do treinamento executado, o engajamento do treinando e o ajuste das tarefas aos objetivos do treinamento e às necessidades do treinando. A adoção desses princípios envolve processos de decisão sobre o processo de aprendizagem e, portanto, depende de um conjunto de ferramentas que inclua a coleta, processamento, análise e visualização de dados acerca do desempenho do treinando. Na avaliação dos resultados do treinamento devem-se estabelecer critérios e métodos de verificação e classificação do desempenho do treinando com base nos resultados esperados. De modo geral, a avaliação do treinamento fornece o conhecimento necessário para uma análise do treinamento e recomendação de melhorias.

Uma vez que a eficácia do treinamento, incluindo aquele realizado em ambientes virtuais, está relacionada à aprendizagem, a atenção aos aspectos cognitivos e afetivos no processo de aquisição do conhecimento é fundamental para a avaliação do treinando (ALL et al., 2014) (JIA et al., 2012). A natureza digital dos dados provenientes dos ambientes virtuais de treinamento, tem exigido a criação de novos métodos, critérios e instrumentos de avaliação do desempenho do treinando e dos processos de aprendizagem (NETTO et al., 2014a) (HAINEY et al., 2015). A utilização de ambientes virtuais de treinamento em atividades de risco apresentam vantagens em relação ao treinamento em ambientes reais. Conforme Gupta et al. (2008a), algumas destas vantagens, elencadas a seguir, se revelam desejáveis sob o ponto de vista da modelagem do processo de aprendizagem.

- Possibilita o registro, análise e compartilhamento do treinamento, podendo ocorrer e ser repetido a qualquer tempo. O desenho de um sistema de treinamento virtual prevê a comunicação com um banco de dados acerca da interação dos usuários com o sistema. A possibilidade de repetição da atividade permite a padronização de instrumentos de avaliação e a geração de padrões de erros comparáveis entre si.
- Por se tratar de uma simulação virtual, é seguro em relação aos riscos materiais, humanos e ambientais, pois não envolve componentes reais. Dadas as condições de riscos às quais estão expostos os profissionais eletricistas durante a realização

de manobras de manutenção em subestações elétricas, as operações executadas durante estas intervenções exigem o máximo de atenção aos procedimentos previamente planejados, seja numa manutenção de rotina ou emergencial, seja em treinamentos profissionais. O planejamento de um treinamento tradicional desse porte implica uma logística complexa com uma atuação circunscrita por um conjunto limitado de cenários.

- Pode ser personalizado segundo o nível e objetivos de treinamento permitindo a experimentação prévia de novos procedimentos, cenários e tecnologias. Em ambientes virtuais de treinamento, a modularidade embutida em sua arquitetura permite transitar de um foco maior sobre os procedimentos que tenham maior demanda até a elaboração de cenários sujeitos a presença de elementos aleatórios que se apresentem como situações desafiadoras, exercitem o processo de tomada de decisão e, finalmente, postulem uma maior consciência sobre o desempenho.

Entretanto, como apontam De Winter et al. (2012), o uso de simuladores em treinamentos pode ser afetado pela falta de realismo ou simplicidade com que uma tarefa é apresentada. Nesse caso, situações potencialmente perigosas em um ambiente real, podem ser ignoradas pelo treinando. Além disso, as interfaces de interação do treinando com o sistema podem ser desconfortáveis. De um lado, os sistemas de monitoramento das atividades do treinando exigem que sua movimentação esteja limitada pelo alcance na comunicação dos dispositivos de comunicação entre os periféricos do sistema, como controle de manipulação e navegação no ambiente virtual e óculos de imersão. Em relação a este último, e outros sistemas de visualização como monitores 3D, existe um incômodo adicional associado ao senso de orientação provocando perda do equilíbrio e mal-estar. Este fator restringe o tempo de uso dos simuladores, o que em certo sentido, compromete a execução da tarefa em tempo real.

Outras questões importantes são levantadas em termos da eficiência dos simuladores como meio adequado para promover uma aprendizagem significativa. Conforme Pantelidis (2010), a adaptação a tecnologia pode ser um obstáculo exigindo um tempo de aprendizado para o seu uso. Neste sentido, o comportamento dos aprendizes pode denunciar um desvirtuamento dos propósitos instrucionais quando, por exemplo, o aprendiz apenas **joga** a simulação (*gaming the system*) (BAKER et al., 2004). Um dos desafios nesta situação é identificar e classificar tais comportamentos utilizando técnicas de aprendizado de máquina para a sua predição (PAQUETTE; BAKER, 2019).

O estudo do comportamento humano por meio de experimentos simulados é primordial para o desenvolvimento de projetos com foco na segurança. Se inicialmente,

sob um olhar mecanicista, o foco das análises de confiabilidade humana se restringiam ao ambiente induzido ao erro, atualmente agregam também os aspectos relativos à subjetividade da natureza humana. Modelos conceituais e matemáticos que simulam o comportamento humano são referência para o planejamento de ambientes seguros e a criação de barreiras de proteção contra o ato inseguro. Sejam definidos a partir de métodos determinísticos ou lógica fuzzy, relativos a dados históricos ou obtidos por meio de simulações computacionais, a modelagem do erro tem contribuído na formulação de cenários para os quais os dados históricos acerca da experiência operacional são insuficientes para análise de risco. Contudo, conforme Ponte Junior (PONTE JR., 2014), um único modelo não consegue capturar a complexidade de um fenômeno como o erro humano, por isso é necessário manter a criatividade aguçada e ficar sempre atento ao inesperado.

Dados os índices de incidentes causados por falha humana em diversos setores da economia, a recomendação de investimentos em ferramentas de aprimoramento do desempenho do trabalhador do setor elétrico se justifica em, pelo menos, dois estratos diferentes. O primeiro diz respeito à minimização do impacto do erro humano no sistema. Em dissertação sobre a confiabilidade humana em operações em subestações elétricas, (GUEDES, 2017) apresenta alguns casos de falhas ocorridas no Sistema Interligado Nacional (SIN). O impacto dessas falhas no sistema resultaram em *blackouts* em diversos estados do Brasil como apresentado na FIGURA 2. Dentre os erros que contribuíram para estas falhas destaca-se a recorrência dos erros associados a manobras e rotinas de manutenção.

FIGURA 2 – REGISTRO DE FALHAS NO SIN DEVIDO AO ERRO HUMANO

Data	Erro Humano
1/04/1984	Deficiências nas rotinas de manutenção, procedimentos e treinamento das equipes sobre segurança operacional
18/08/1985	Ajustes incorretos do sistema especial de rejeição de carga da usina de Marimbondo
13/12/1994	Falha na operação do sistema durante a realização de testes na subestação Ibiúna
26/03/1996	Erro de manobra na usina de Furnas
11/03/1999	Subestimação dos riscos de blecautes no planejamento de manutenção, arranjo físico das subestações de Extra-Alta-Tensão inadequado
16/05/1999	Configuração inadequada dos dispositivos de proteção e medidas de precauções insuficientes durante as manobras

FONTE: Guedes (2017)

O segundo estrato, um dos principais motivadores deste trabalho, é a prevenção de acidentes. Embora as séries históricas tenham apresentado um decréscimo no número de acidentes no setor até 2012, conforme Silva (2015), os índices não só se mantiveram acima dentre todos os trabalhadores de outras áreas do setor produtivo, como também aumentaram a partir de 2017. Estatísticas compiladas pelo Anuário da Associação Brasileira de Conscientização para os Perigos da Eletricidade (Abracopel) (ABRACOPEL, 2020) ilustram o quadro geral dos acidentes de trabalho no setor apresentado na tabela TABELA 1. A soma de mortes no setor elétrico, correspondente às duas últimas linhas da tabela, coloca a categoria em segundo lugar com 80 vítimas fatais em 2019, ficando apenas atrás somente dos trabalhadores da agricultura.

TABELA 1 – ÓBITOS POR CHOQUE ELÉTRICO POR OCUPAÇÃO

Ocupação	2013	2014	2015	2016	2017	2018	2019
Aposentado	0	0	39	40	33	50	44
Soldador, Serralheiro, Marceneiro, Vidraceiro e Carpinteiro	0	15	9	21	10	7	6
Motorista de Caminhão e Ônibus	0	2	24	20	14	16	25
Curioso/Ladrão	0	33	17	31	30	40	39
Agricultor	0	0	56	90	72	105	82
Pintor/Ajudante	10	23	16	31	18	24	25
Pedreiro/Ajudante	55	60	58	71	67	37	40
Instalador de Telefonia, Placas, Toldos, Calhas e Ar-condicionado	20	20	17	39	10	13	18
Podador de árvore	0	4	1	6	10	7	4
Faxineira, Doméstica, Dona de casa	72	41	33	40	37	56	37
Estudante	117	79	76	93	76	118	74
Eletricista Profissional/Empresa	29	16	22	33	18	5	11
Eletricista, Técnico Eletrotécnico, Ajudante	71	54	61	59	45	57	69

FONTE: Adaptado de (ABRACOPEL, 2020)

Estudo conduzido por Silva e Moreira (2019) sobre dados da Fundação Comitê de Gestão Empresarial (Funcoge) entre 2008 e 2013 concluiu que 67% dos acidentes tiveram como causa falas associadas ao treinamento e supervisão. A utilização de sistemas virtuais de treinamento como ambiente de simulação de atividades críticas pode converter-se numa ferramenta complementar aos métodos tradicionais. O treinamento virtual concebido sobre princípios instrucionais aliado a experiência de imersão e interatividade com o ambiente virtual aproxima o treinando da situação real potencializando o seu engajamento na execução e comprometimento com os objetivos do treinamento. Adicionalmente, como já comentado, um treinamento virtual suporta o desenho de cenários de alto risco com máxima fidelidade permitindo o exercício de um amplo espectro de habilidades em relação à prevenção e análise do erro humano.

No contexto do treinamento profissional em atividades críticas a atenção ao erro é um princípio básico e a sua observação sistemática sobre o erro faz parte dos

objetivos do treinamento de linha viva e antecede qualquer procedimento ou manobra de manutenção em subestações elétricas (COPEL, sd.). A modelagem da aprendizagem a partir da avaliação do desempenho do treinando depende da observação sobre a ocorrência de erros durante a execução de uma tarefa. A identificação e classificação dos erros humanos é fundamental para sua própria descrição, na análise de suas causas e na sua prevenção.

O erro humano tem sido objeto de estudo tanto no campo instrucional quanto da confiabilidade humana. A pesquisa sobre o erro tem aplicações na avaliação do desempenho humano, no desenvolvimento de ambientes de trabalho e estabelecimento de protocolos de ações seguras. No contexto do desenvolvimento de ambientes virtuais de treinamento e, em particular, da incorporação de técnicas de análise de dados na avaliação do desempenho dos treinandos, um modelo do erro humano busca recuperar o sentido original da tarefa realizada em campo. Uma medida do erro na execução da tarefa expande o campo semântico dos dados brutos capturados pelos recursos de telemetria do sistema potencializando a interpretação dos resultados alcançados pelos treinandos.

1. no mapeamento o estado de conhecimento do treinando a partir de dados gerados na interação com o ambiente virtual.
2. como ferramenta auxiliar na análise de risco associado ao erro humano.

Uma lacuna não explorada diz respeito à conjunção destas técnicas para categorizar o desempenho do treinando por meio da visualização e análise dos padrões de erro associados a execução da tarefa.

1.3 DELIMITAÇÃO DO TEMA

O contexto de aplicação deste trabalho circunscreve-se em torno da avaliação do desempenho de eletricistas profissionais de linha viva durante treinamento virtual de atividade de manutenção em subestação elétrica, especificamente, das atividades envolvidas na substituição de isolador de pedestal. Por meio do percurso do treinando durante a execução da tarefa, explora-se com mais ênfase a dimensão relativa ao conhecimento de regras e procedimentos como estratégias de resolução de problemas em situações críticas. Parte-se de um modelo teórico acerca da cognição o qual relaciona o estado de conhecimento do indivíduo por meio de seu desempenho com determinados padrões de comportamento.

Um padrão de comportamento é um reflexo de crenças compartilhadas que influenciam na tomada de decisão e, consequentemente, afetam o desempenho do indivíduo. Dessa forma, a adoção de determinados comportamentos no reconhecimento

e análise de uma situação crítica, ou execução e validação de uma solução proposta, deve corresponder a um certo nível de desempenho. Técnicas de descoberta do conhecimento em conjunto de dados têm sido aplicadas com sucesso em sistemas instrucionais seja para inferir, estática ou dinamicamente, o estado de conhecimento do treinando, ou no mapeamento de padrões de interações do usuário. O objetivo da análise e visualização de tais padrões é revelar uma imagem dinâmica que permita identificar alterações no desempenho individual e classificá-la em termos dos objetivos de aprendizagem. O erro humano tem sido utilizado como parâmetro na avaliação da aprendizagem desde os primórdios do estudo sobre o comportamento humano. Neste sentido, mapeamento e categorização de estratégias cognitivas durante a resolução de problemas e, em particular, a atenção ao erro, são fundamentais para a reflexão e direcionamento do processo de aprendizagem. No contexto da aplicação de treinamento de atividades críticas o erro, ou inversamente, a sua ausência é um dos resultados esperados.

1.4 PROBLEMA

O problema deste trabalho envolve, portanto, a elaboração de um método de avaliação da aprendizagem de eletricistas profissionais em ambiente virtual de treinamento. Este problema será investigado a partir da convergência entre as áreas que têm tratado tanto da questão instrucional quanto da confiabilidade humana, e, em particular, do erro humano. O método desejado deve capturar os erros cometidos pelos treinandos durante a execução da tarefa e categorizar o desempenho de cada treinando segundo classes de desempenho de forma automática. Para alcançar a automação desejada o método deve utilizar ferramentas de análise baseadas na mineração de padrões sobre os dados de interação dos usuários com o sistema de treinamento. Neste sentido, algumas questões orientam o trabalho desta pesquisa como, por exemplo,

- Quais modelos de conhecimento podem representar uma tarefa instrucional?
- Como os padrões de dados devem ser modelados de modo a incorporar os erros cometidos durante a execução da tarefa?
- Quais métodos de mineração podem ser usados para agrupar os treinandos em classes de desempenho?
- Como definir a similaridade entre dois desempenhos diferentes?
- Quais métodos de mineração agrupam os treinandos em classes de desempenho de modo a capturar os padrões de erro dos treinandos?

Estas questões são investigadas a partir dos objetivos enunciado a seguir.

1.5 OBJETIVO GERAL

O objetivo desta tese é a elaboração de um método de avaliação do desempenho humano baseado no agrupamento de padrões de erros no treinamento profissional de atividades críticas em ambiente virtual.

1.5.1 Objetivos específicos

Para alcançar o objetivo geral, três objetivos específicos são definidos:

- Desenvolver um modelo do conhecimento de regras baseado na sua representação padrões de erros,
- Desenvolver um modelo do conhecimento de regras baseado na sua representação como caminhos sobre grafos e do modelo do treinando como passeios sobre o grafo, e
- Identificar os padrões de similaridades entre grafos que se aproximam de padrões de erros por meio da análise de agrupamento de padrões.

1.6 ESTRUTURA DA TESE

A estrutura desta tese está organizada em quatro capítulos, incluindo esta Introdução. No segundo capítulo, a Fundamentação Teórica, são discutidos os subsídios teóricos a partir de quatro grandes núcleos sobre os quais se fundamentam o desenvolvimento da tese: a avaliação do desempenho, a modelagem do erro humano, a representação do conhecimento e a mineração de dados instrucionais. O terceiro capítulo trata dos aspectos metodológicos da análise de dados e o contexto da pesquisa. Neste capítulo são apresentadas as etapas do trabalho referentes a modelagem, as técnicas de análise e visualização dos dados. No último capítulo, são sumarizados os resultados e as limitações da pesquisa as quais, por ora, se apresentam como oportunidades no desenvolvimento de futuras investigações.

2 FUNDAMENTAÇÃO TEÓRICA

Esta tese se fundamenta sobre quatro pilares: a avaliação do desempenho, a modelagem do erro humano, a representação do conhecimento em ambientes virtuais de treinamento e a mineração de dados instrucionais.

FIGURA 3 – PILARES DE SUSTENTAÇÃO DA TESE

FONTE: Guedes (2017)

Neste capítulo da fundamentação teórica eles são apresentados em três seções. A primeira seção, Aprender com os erros, explora a intersecção em torno da reflexão sobre o erro entre os domínios do planejamento da instrução e da análise de confiabilidade humana. Neste sentido estuda-se a relação entre um modelo cognitivo sobre os mecanismos da falha humana e os esquemas de classificação, ou taxonomias do erro. Por meio destas taxonomias, os comportamentos observáveis são mapeados na execução da tarefa instrucional e categorizados conforme os tipos de erro identificados.

As ferramentas de visualização utilizadas na ilustração de conceitos ou técnicas de análise do erro, um aspecto acessório a primeira vista, não deve passar despercebido. O primeiro caso que chama atenção é a utilização de grafos no estudo sobre o erro humano, por exemplo, no monitoramento de estados, como estados de máquinas ou estados do conhecimento, e também na análise da tarefa, como um diagrama de fluxo de atividades ou uma rede de influências. Grafos fazem parte da representação técnica de documentos de padronização para a execução de uma tarefa

e, portanto, são, potencialmente capazes de comunicar as informações a partir de uma campo semântico próximo da atuação do treinando. Além disso, do ponto de vista dos processos instrucionais, os grafos podem ser instanciados como mapas conceituais, os quais são ferramentas muito utilizadas no contexto da aprendizagem significativa (NOVAK; CAÑAS, 2010).

A segunda seção, Aprender com os dados, trata da modelagem do erro e da análise do desempenho humano por meio de técnicas de mineração de dados instrucionais. Nessa seção discutem-se os aspectos quantitativos relacionados a avaliação do desempenho baseada em dados de interação em sistemas de treinamento virtual. Neste contexto de aplicação, o tópico sobre modelagem aborda as questões sobre a representação do conhecimento aprofundando a discussão em torno da utilização de grafos na captura de padrões de desempenho na tarefa instrucional.

A terceira seção, Agrupamento de dados baseados em grafos, é dividida em duas partes. A primeira parte apresenta a fundamentação para a modelagem, tratamento e visualização dos dados, em especial para o agrupamento de dados em classes de similaridade. A segunda parte dedica-se a explorar o conceitos e definições de similaridade de grafos.

2.1 APRENDENDO COM OS ERROS

O desenvolvimento de sistemas computacionais aplicados à educação ou ao treinamento profissional é acompanhado de forma simbiótica pelos estudos sobre a aprendizagem e o desempenho humano. Por um lado, essa intersecção é tecnológica e pode ser representada por uma trajetória que se inicia em meados dos anos de 1960, com as máquinas de ensinar em modo texto e a instrução programada, é marcada pelos sistemas tutores inteligentes, durante as décadas de 1980 e 1990, e chega às atuais plataformas interativas de imersão baseadas em realidade virtual. Por outro lado, segundo Baron et al. (2004), comuns a esta trajetória são a reflexão crítica da psicologia comportamental e a recorrência generalizada às metáforas baseadas no modelo de Processamento de Informação durante a segunda metade do século vinte.

2.1.1 Tecnologias Instrucionais

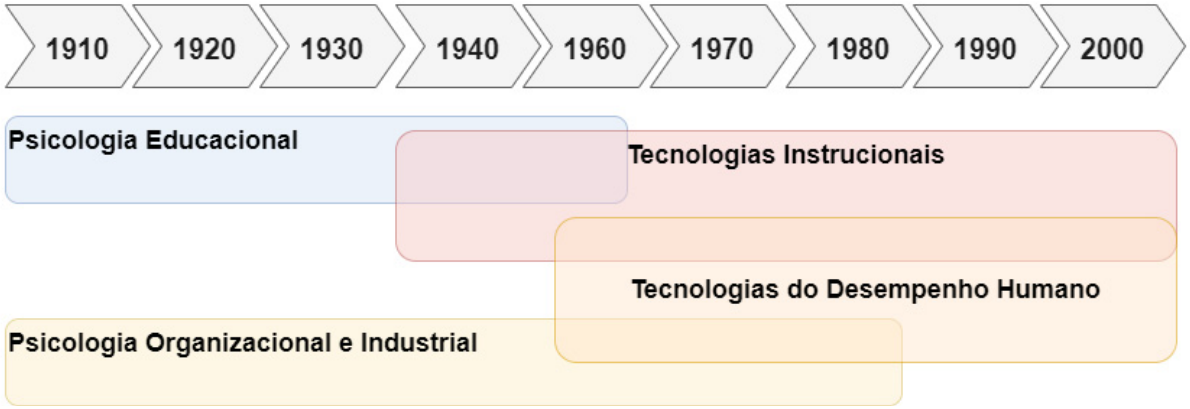
As arquiteturas presentes em diferentes sistemas instrucionais se decompõem em elementos que buscavam mimetizar o processo de ensino tradicional, mas com a atribuição de tarefas automáticas ao sistema computacional, como, por exemplo, a apresentação de uma sequência didática, identificação e correção de erros e ao processamento de visualização do desempenho dos usuários (SOTTILARE, 2018).

Nesta perspectiva, com as máquinas de aprender, os automatismos dos atores

presentes no processo de ensino-aprendizagem cedem espaço para uma participação ativa baseada na interação humano-máquina. Compreende-se daí a importância do desenvolvimento concomitante das técnicas e dispositivos de armazenamento, comunicação e recuperação de dados de interação do usuário com o computador. Conforme Levy (2001), a própria modelagem dos processos de aquisição do conhecimento são mediados pela interação com as tecnologias disponíveis, seja em um livro impresso, seja em uma estrutura de hipertexto na rede mundial de computadores ou no mundo virtual dos jogos digitais.

Do ponto de vista das teorias acerca da aprendizagem e do desempenho humano, é possível reconstruir uma genealogia, que se origina entre as fileiras da psicologia do comportamento, e alcança certa maturidade junto à psicologia da cognição. Diversas áreas se influenciaram mutuamente, muitas vezes trafegando numa zona fronteiriça, fomentando a interdisciplinaridade na pesquisa sobre o tema e permitindo a concepção de esquemas relativamente gerais com largo espectro de aplicação. Ao traçar os fundamentos das chamadas tecnologias instrucionais, um conjunto de aplicações sistemáticas e sistêmicas de teorias, conceitos, princípios e métodos à situação de aprendizagem, Chyung (2008) mapeia as influências sobre as quais diversos autores trataram das questões acerca do desempenho (FIGURA 4).

FIGURA 4 – INFLUÊNCIAS ENTRE AS TECNOLOGIAS INSTRUCIONAIS E DO DESEMPENHO HUMANO



FONTE: Adaptado de Chyung (2008)

Conforme Ponte Jr. (2014), os estudos sobre o desempenho humano e, em particular, sobre o erro humano, sustentam as ações e decisões corporativas, coletivas e individuais, cuja prioridade absoluta é a instalação e manutenção de uma Cultura de Segurança. Boas práticas de segurança implicam a ação certa, na hora certa, para que

um acidente seja evitado. A cultura da segurança vai além do bom senso e se assenta sobre um processo sistemático e sistêmico de investigação das causas e prevenção do erro humano. Esses esforços se consolidaram por meio de técnicas de modelagem, métodos de coleta e análise de dados e estruturas interpretativas acerca dos processos cognitivos e comportamentais que afetam o desempenho humano distanciando-a do desejável nível de excelência.

Ao mesmo tempo em que, ao longo dos anos, os investimentos na pesquisa e na cultura de segurança resultaram em um decréscimo na ocorrência de erro humano, as necessidades e atenção à questão não perderam força. Ao contrário, reconhece-se que o erro humano é inevitável, e, além disso, a engenharia deve trabalhar para que as consequências dos erros não superem níveis de risco aceitáveis. As dimensões e complexidade dos sistemas com os quais o ser humano atualmente interage tornam-no potencialmente capaz de produzir cenários catastróficos. Alguns exemplos clássicos, onde as barreiras contra os erros e a sua propagação falharam, são os casos de acidentes em usinas nucleares, sistemas de transporte, plataformas de petróleo, pesquisa aeroespacial, dentre outros (MOURA et al., 2014). O estudo sobre o erro debruça-se, principalmente, no sentido da identificação, classificação e predição do erro humano com o objetivo de produzir conhecimento para a tomada de decisão em relação às ações corretivas e preventivas (REASON; HOBBS, 2017).

A importância do erro na aprendizagem e sua utilização nos processos instrucionais tem dividido os argumentos a favor e contra. Como lembra Metcalfe (2017) citando David Ausubel, a inclusão intencional de situações que induzem o aprendiz ao erro pode reforçar uma ação indesejada. Porém, o autor adverte que ao evitar ou eliminar o erro do processo de ensino pode limitar a exploração de estratégias alternativas podando os aspectos criativos na tomada de decisão associados a aprendizagem. Diversas abordagens (TULIS et al., 2016; HOGARTH et al., 1991; METCALFE; XU, 2018) têm visto o erro como uma oportunidade inevitável e, como qual, deve ser objeto de uma avaliação propositiva no sentido da correção do erro por meio do aprendiz (OHLSSON, 1996).

2.1.2 Aprendizagem e desempenho

A aprendizagem decorre de processos cognitivos que envolvem mecanismos internos aos indivíduos não podendo, portanto, ser observada diretamente. Apesar disso, aprendizagem pode ser influenciada tanto por fatores externos como o planejamento instrucional e as interações sociais no ambiente de aprendizagem, quanto por fatores internos associados a personalidade e a memória. A ideia de aprendizagem invoca uma dinâmica entre estados, não necessariamente discretos ou bem definidos. Neste sentido, a avaliação da aprendizagem é realizada considerando alguma medida

associada ao desempenho do aprendiz (SODERSTROM; BJORK, 2015). Esta associação entre aprendizagem e desempenho é feita por meio de modelos cognitivos que procuram explicar como determinados fatores afetam a aprendizagem.

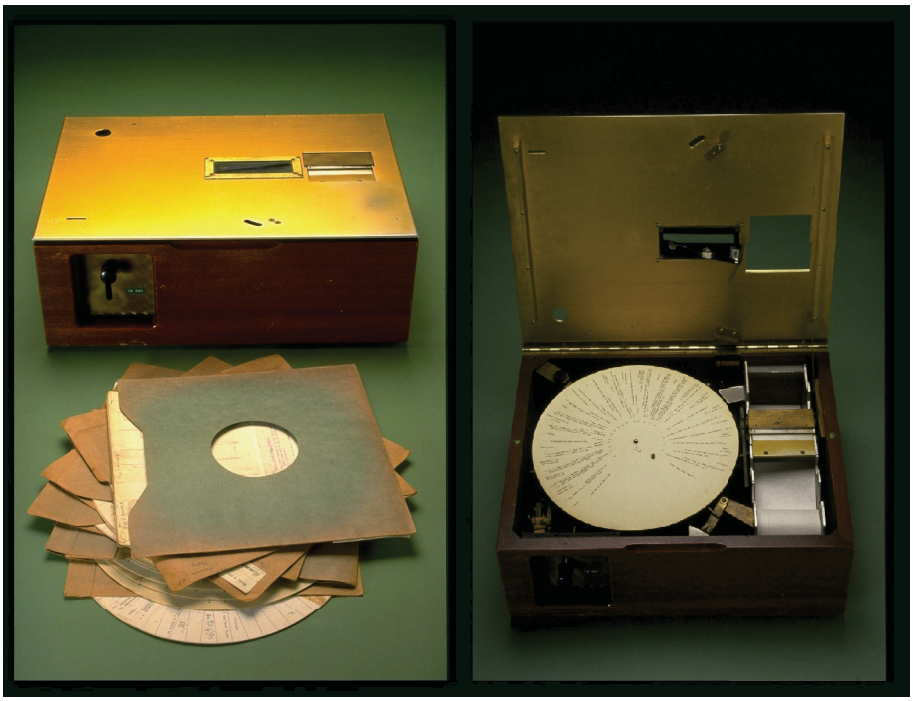
A extensão com a qual estes fatores podem ser incorporados na reflexão sobre a aprendizagem e o impacto desta incorporação no âmbito das tecnologias instrucionais podem ajudar a sintetizar algumas das mais discutidas teorias da aprendizagem ao longo do século vinte: o comportamentalismo (*behaviorism*), o cognitivismo e o construtivismo (DRISCOLL, 2014). Na primeira delas a aprendizagem corresponde a uma mudança no comportamento e os estudos se concentram sobre a construção de situações instrucionais; a segunda, foca nos processos internos à **mente**, onde ocorre o processamento da informação; e a terceira inclui fatores externos (sociais, organizacionais, pessoais) que influenciam a aprendizagem a qual é definida como um fenômeno social e como tal se realiza no seio das relação com outros aprendizes, instrutores e o ambiente de aprendizagem (ERTMER; NEWBY, 1993). Outra teoria, auto-denominada sucessora do construtivismo é o conexionismo, onde a aprendizagem ocorre dentro de uma rede dinâmica de relações. Enquanto o comportamentalismo e o cognitivismo são idealizados a partir da metáfora da máquina, do computador, o modelo conexionista tem inspiração nos processo biológicos (MEDLER, 1998; BELL, 2011).

Entre o final do século 19 e o século 20, a discussão sobre a aprendizagem ganha novos contornos ao se afastar da metafísica e tornar-se um objeto de estudo da psicologia experimental. Ainda que situado historicamente, a partir de então entende-se que o fenômeno da aprendizagem é definido de forma indireta por meio das respostas associadas aos estímulos externos. A aprendizagem é estudada como um problema de **associação** cujas origens são os estudos sobre a aprendizagem em animais sobre a resposta condicionada ou sobre a aprendizagem de associações verbais (GAGNÉ, 1974).

Nas primeiras décadas do século XX, o psicólogo americano Edward Thorndike sistematiza alguns princípios sobre o comportamento humano que podem ser aplicados ao ensino e à aprendizagem. As aspirações de Thorndike contribuíram também para a afirmação de uma crença sobre a qual seria possível o controle, produção ou eliminação de um comportamento específico via uma correlação entre o efeito desejado para um estímulo dado, a prontidão da resposta ao estímulo e a repetição, ou exercício, de condicionamento da resposta. Uma ferramenta importante para caracterizar a evolução da aprendizagem foi introduzida por Thorndike por meio das curvas de aprendizagem, um recurso gráfico para representar a correlação entre o estado do conhecimento e o tempo de aprendizagem. As curvas de aprendizagem evidenciam o seu caráter dinâmico, refletindo a evolução do estado de conhecimento do aprendiz.

A continuidade do trabalho de Thorndike sobre o comportamento condicionado assentado sobre o binômio estímulo-resposta é levada a cabo por outro psicólogo americano, Skinner. Na realidade, Skinner amplia o esquema de Thorndike introduzindo um segundo estímulo de reforço após a obtenção de uma resposta desejada ao estímulo inicial. Em relação ao desenvolvimento de tecnologias instrucionais, Skinner com base nas idéias originais de Sidney Pressey, concebe as primeiras máquinas de ensinar, que além de testar o conhecimento do aluno, auxiliavam no processo de aprendizagem.

FIGURA 5 – MAQUINA DE APRENDER DE SKINNER



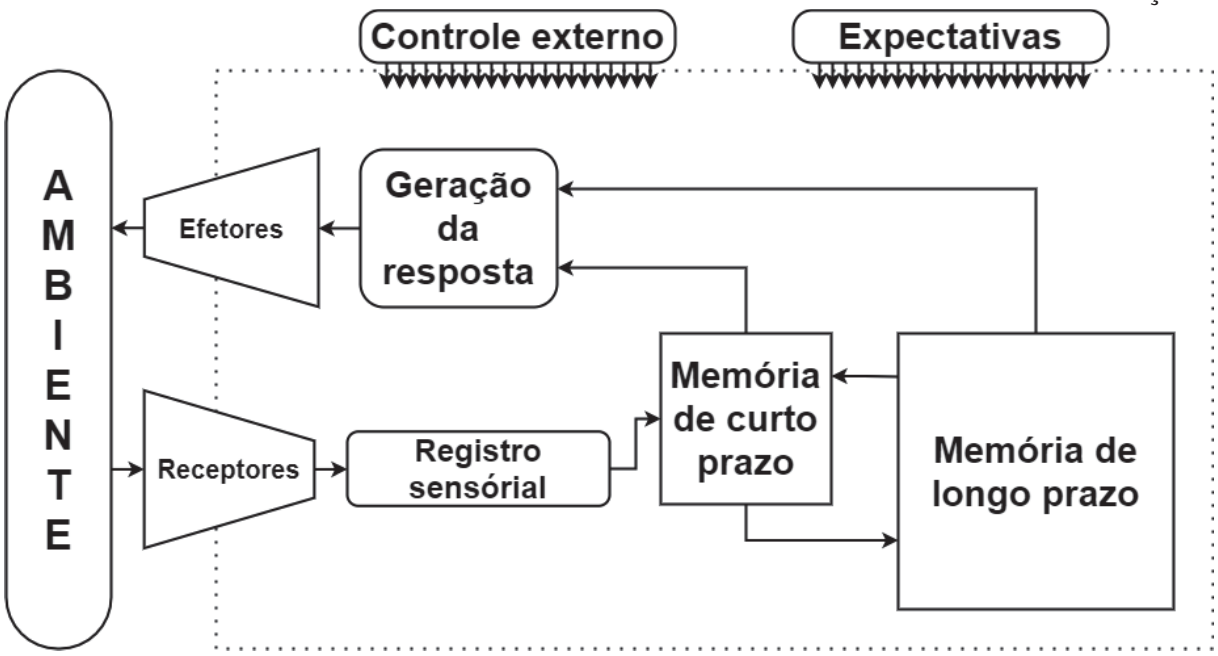
FONTE: https://americanhistory.si.edu/collections/search/object/nmah_690062

As máquinas de ensinar de Skinner, como o exemplar apresentado na FIGURA 5, eram baseadas num programa linear onde o conteúdo, decomposto numa sequência de pequenos passos ou, quadros, era apresentado num visor por meio de respostas a itens de múltiplas escolhas. Apesar da aparente simplicidade, Skinner provê sua máquina com um sistema de revisão das respostas dos usuários acrescentando um passo indispensável para a autoavaliação e reforço da resposta correta. Avanços posteriores permitiram a implementação de outros conceitos como o de **ramificação** de Crowder, de **matética** de Gilbert e de **instrução controlada pelo aluno** de Mager. Os dois primeiros podem ser considerados precursores da adaptabilidade do programa ao estilo de aprendizagem do usuário por meio da identificação e utilização do erro para realimentar o sistema (KAY et al., 1970).

A Instrução Controlada pelo Aluno indica uma participação ativa do aluno no processo de aprendizagem (Aprendizagem centrada no aprendiz). Sinal desta adaptatividade está na retroalimentação do erro do aluno, dado de um indicador de dificuldade no processo de aprendizagem e que, portanto, exigia uma nova etapa intermediária para o esclarecimento das dúvidas antes de prosseguir na sequência programada.

Em meados do século XX, os limites dos estudos sobre o comportamento condicionado revelou-se principalmente em relação à aprendizagem de tarefas complexas e a influência de fatores internos como a memória e a atenção (TENNYSON, 2010). Trata-se de uma virada paradigmática, oportunidade para passar do behaviourismo ao cognitivismo, o que corresponde a uma correção de foco que sai do ambiente controlado em direção ao aprendiz. Com isso a aprendizagem resulta não apenas de uma resposta ao ambiente programando, mas também a complexas formas de comunicação, armazenamento e recuperação do conhecimento a partir do modelo de conhecimento baseado no processamento da informação (GAGNÉ et al., 1992). Neste contexto, Robert Gagné, psicólogo americano, sistematiza um conjunto de fundamentos do design instrucional com forte influência sobre a problemática do treinamento.

FIGURA 6 – MODELO COGNITIVO BASEADO NO PROCESSAMENTO DA INFORMAÇÃO



FONTE: Adaptado de Gagné et al. (1992)

A perspectiva instrucional implica a constituição de meios práticos para auxiliar no processo de aprendizagem do indivíduo, mais do que um carácter explicativo, o

design instrucional é descritivo e prescritivo. Verifica-se a efetividade do processo de aprendizagem comparando os estados subsequentes, antes e depois, do aprendiz passar por uma situação de aprendizagem. Para alcançar um estado de conhecimento desejado (objetivos de aprendizagem/desempenho), que se revele duradouro (retido, mas não imóvel e sujeito a fatores internos ao indivíduo), e se torne parte de um repertório dinâmico, que possa ser mobilizado quando necessário, exige que a instrução seja planejada.

Os princípios instrucionais, conforme Gagné, se organizam segundo três componentes principais:

1. Uma Taxonomia dos Objetivos da Aprendizagem: trata-se de um sistema de classificação hierarquizado sobre os tipos de conhecimentos a serem aprendidos conforme sua natureza como, por exemplo, o conhecimento de regras, a demonstração de habilidades motoras ou o reconhecimento de sinais. Estes tipos de conhecimentos são apresentados segundo uma estratificação do sistema de processamento interno da informação em termos de níveis de complexidade do conhecimento envolvido e resposta as demandas de uma tarefa instrucional.

2. As Condições de Aprendizagem: onde são definidas as condições de aprendizagem para o **planejamento da instrução** aplicada aos tipos de conhecimentos descritos na taxonomia dos objetivos da aprendizagem. Gagné e seus colaboradores integram e ampliam os estudos de outros autores, como é o caso da mútua influência de Benjamim Bloom, na definição da Taxonomia dos Objetivos da Aprendizagem; de David Ausubel, em relação à valorização de conhecimentos prévios; e dos experimentos de Thorndike e Skinner na aprendizagem de sinais. Em sua visão hierarquizada do conhecimento, os tipos de aprendizagem mais complexos, por exemplo, associados a habilidade de resolução de problemas, são precedidos pelo aprendizado prévio de tipos mais simples, como a resposta a um sinal. Smith e Ragan (1996) mapearam as referências de outros pesquisadores na definição dos oito tipos de aprendizagem de Gagné:

    a) Pavlov: aprendizagem de sinais;
    b) Thorndike, Skinner, Kimbal : aprendizagem do tipo estímulo-resposta;
    c) Skinner, Gilbert: aprendizagem de cadeias verbais e motoras;
    d) Underwood: aprendizagem por associação verbal;
    e) Postman: aprendizagem por discriminação múltipla;
    f) Kendler: aprendizagem de conceitos;
    g) Gagné: aprendizagem de princípios;

h) Katona, Maier: aprendizagem para a resolução de Problemas.

Os tipos de aprendizagem identificados por Gagné são estudados segundo as condições “que controlam sua realização” a partir de requisitos prévios. Para cada tipo de aprendizagem, são apresentadas as condições inerentes ao aprendiz, à situação de aprendizagem, sua fenomenologia, e exemplos de aplicação.

3. Os Eventos da Instrução: compilam e ordenam um conjunto de situações instrucionais experimentais, denominadas como **protótipos**, que favorecem a aprendizagem dos tipos de conhecimentos descritos na taxonomia dos objetivos da aprendizagem. O sentido prescritivo no planejamento da instrução como forma de “organização do trabalho pedagógico”, em Gagné, é representado pelos “Nove Eventos de Instrução”:

   a) Ganhar a atenção do aprendiz;
   b) Expor os objetivos de aprendizagem ao aprendiz;
   c) Mobilizar conhecimento prévio do aprendiz;
   d) Apresentar o material de estímulo;
   e) Orientar a aprendizagem;
   f) Produzir/Verificar o desempenho durante a instrução;
   g) Prover a análise e discussão sobre o desempenho;
   h) Avaliar o desempenho ao final da instrução;
   i) Melhorar a retenção e a transferência;

   Estrutura-se por meio desses Eventos um modelo de instrução cientificamente fundamentado em classes de objetivos de aprendizagem que tornem efetivos e incremente os processos de aprendizagem (GAGNÉ; BRIGGS, 1976).

Conforme constatado por Annete e Duncan (ANNETT J.; DUNCAN, 1967), em seu próprio tempo, inúmeros esquemas taxonômicos proliferaram a ponto de se fazer necessário estabelecer critérios para identificar, classificar e selecionar adequadamente alguma dessas taxonomias, ou seja, uma “taxonomia das taxonomias”. Outra crítica, mais aguda, diz respeito à utilidade dessas taxonomias uma vez que os limites difusos das categoriais geram dubiedade ou contradições nas classificações dos objetivos de aprendizagem. É importante ressaltar que, mais do que as peculiaridades dos esquemas propostos, cada qual com suas limitações, valoriza-se o esforço de sistematização dos processos instrucionais por meio de uma análise correlacionada do comportamento humano e da aprendizagem, com forte apelo à aplicação e experimentação, no sentido de promover uma aprendizagem significativa.

É nesse sentido que Merrill (2007), a partir dos estudos que o precedem, compila os “primeiros princípios” do design instrucional. Trata-se de cinco princípios, comuns a diferentes teorias da aprendizagem e aos modelos instrucionais relacionados, os quais diferenciam-se entre si por conta dos enfoques diferentes em relação àqueles princípios. A aplicação de tais princípios, cuja função é a elaboração de ambientes e objetivos de aprendizagem, não depende do tipo de sistema ou modelo instrucional, e tem um caráter prescritivo. Merril elenca seus “primeiros princípios” e as quatro fases do processo instrucional para sua aplicação e, de certa forma, define um procedimento para avaliação e comparação de modelos instrucionais:

Primeiro princípio (fase do engajamento) Engajamento do aprendiz na resolução de problemas do mundo real. A aplicação deste princípio implica

- a apresentação da tarefa (objetivos instrucionais);
- a adequação e ordenação dos níveis de problematização conforme necessidade do aprendiz;
- o sequenciamento da tarefa.

Segundo princípio (fase da ativação) Mobilização de experiências, conhecimentos e estratégias cognitivas prévias para a construção e de novos conhecimentos.

Terceiro princípio (fase da demonstração) Demonstração da tarefa, a orientação ao aprendiz deve manter a coerência relativa aos objetivos da aprendizagem por meio da ilustração de conceitos, a demonstração e visualização da tarefa e seleção dos meios de comunicação.

Quarto princípio (fase da aplicação) Execução da tarefa pelo aprendiz por meio de uma prática consistente aos objetivos de aprendizagem e às avaliações do desempenho. Durante esta fase, espera-se uma gradual diminuição das intervenções do tutor, incluindo a identificação e correção de erros, conforme o aprendiz ganhe autonomia na resolução de outros tipos semelhantes de problemas.

Quinto princípio (fase de integração) Transferência do conhecimento ou habilidades para outros contextos. Nesta fase o aprendiz deve ser estimulado a refletir sobre outras aplicações do conhecimento adquirido.

Seguindo Merrill, ao lado das teorias e modelos delas oriundos sobressaem-se as lições sobre a organização do trabalho de planejamento da instrução baseado não apenas na lógica do conteúdo, mas também em relação às condições de aprendizagem. Numa conjunção de dois mundos, o espaço de instrução passa a ser o laboratório, a situação de aprendizagem constitui um experimento. Especificamente, destaca-se o

desenvolvimento de técnicas e ferramentas de análise das variáveis do comportamento humano em sua interação com outras interfaces (sistemas, dispositivos ou situações). A reflexão sobre a aprendizagem fundamenta-se em modelos cognitivos que explicam os mecanismos internos de aquisição e a retenção do conhecimento.

No campo do design instrucional aplicado ao treinamento, as intersecções com outras áreas em desenvolvimento, como a engenharia e sistemas tutores, ganham contornos relativamente próprios. Noone (NOONE, 1993) alerta sobre as especificidades do design instrucional aplicado ao ambiente de trabalho. Entretanto, sua preocupação maior é que o processo instrucional efetivamente atenda às necessidades do treinamento e da empresa.

As considerações sobre os fatores humanos, como usabilidade e confiabilidade, são indispensáveis em qualquer projeto de engenharia. Dessa forma, a interação humana em sistemas complexos nos setores da indústria, transporte, ou militar, dentre outros, constitui um desafio mais amplo a exigir habilidades específicas como, por exemplo, o reconhecimento de algum tipo de padrão desejável para um elemento dado, a comparação entre padrões, a tomada de decisão sobre adequação ao padrão estabelecido e ação sobre o elemento julgado.

Nos estudos sobre o desempenho humano e, em particular, sobre o erro humano, a relação com o design instrucional é uma via de mão dupla. De um lado, a execução de uma dada tarefa em condições seguras no ambiente profissional depende de um processo de formação de competências especializadas e, portanto, demanda o planejamento do treinamento. De outro lado, a identificação, análise e classificação dos erros dos aprendizes têm um papel fundamental nos modelos instrucionais, como já apresentado em relação aos primeiros princípios de Merrill.

2.1.3 Desempenho e erro humano

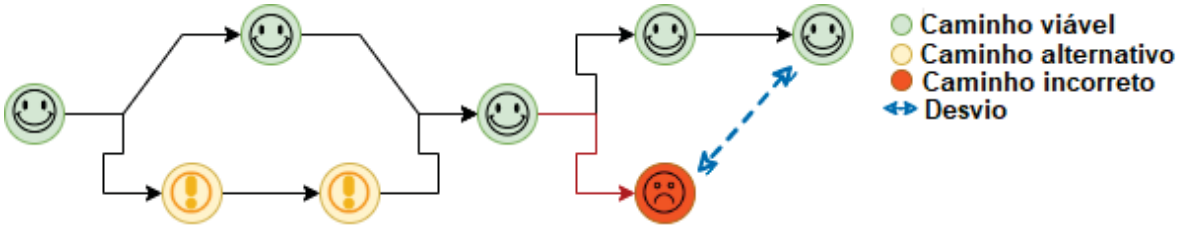
No campo de estudo da psicologia experimental, no início do século vinte, o estudo da aprendizagem atribui ao erro um papel fundamental. Segundo Pereira (1983), "enquanto os psicólogos estudavam os aspectos básicos da aprendizagem, o que eles realmente observavam e mensuravam eram os erros". Uma perspectiva mais abrangente, que incorpora no estudo do erro humano outras área do conhecimento é consolidada a partir dos anos de 1980 com as publicações de John Reason e Donald Norman e da realização dos primeiros encontros internacionais sobre o tema.

Na segunda conferência internacional sobre o erro humano, contando com especialistas das mais diversas áreas, as sessões de trabalho foram organizadas segundo seis eixos principais: Conceitos e definições, Taxonomia, Teoria, Prevenção, Terapêutica e Especulação. Conforme o relato de Pereira (1983), os debates permitiram o esclarecimento em relação à terminologia e ao conceito de erro humano que poderia

ser sintetizado como segue:

O erro humano é um desvio significativo do desempenho esperado o qual pode ser definido a partir de um critério normativo ou estatístico.[1]

FIGURA 7 – ERRO HUMANO COMO UM *MISMATCH* EM RELAÇÃO A UMA AÇÃO ESPERADA.



FONTE: Adaptado de Rasmussen (1980)

Conforme o esquema ilustrado na FIGURA 7, o erro humano se manifesta como um *mismatch* (desvio) da ação esperada (RASMUSSEN, 1980). Uma tarefa poder ser decomposta como uma sequência de atividades (ações) e a sua execução implica escolher uma ordem em que estas atividades serão executadas. Três tipos de sequências estão associadas a execução da tarefa:

- Sequencia esperada: a ordem de execução das atividades é conhecida e utilizada como referência para o medir o desvio entre a ação esperada e a ação executada.

[1] O conceito de erro envolve a ideia de que o desempenho é medido através do comportamento observável utilizando-se na prática número limitado de fatores que caracterizam a sua ocorrência. Conforme apresentado no Apêndice III, 1.1 Definições, do artigo original de Pereira (1983):

> Erro é qualquer desvio significativo da expectativa. O significado do desvio depende quer de um critério normativo quer de um critério estatístico.
> Erro humano é qualquer desvio significativo de um critério específico de expectativa de ação humana (*human performance*).
> Ação (*performance*) é o comportamento observável na interface entre o agente humano e outro sistema aberto, humano ou máquina (interfaces homem-homem ou homem-máquina) que induz mudança no segundo sistema.
> A expectativa da ação tem de ser definida em termos dos contextos da interface em causa e não da própria interface.
> Os contextos da ação humana são hierarquicamente organizados e logicamente analisáveis como níveis de realidade. O número de níveis a considerar é limitado para propósitos práticos (embora ilimitado enquanto possibilidade matemática).

- Sequência alternativa: a sequência alternativa pode ser uma outra estratégia viável para a execução da tarefa. Contudo, uma sequência alternativa deve ser objeto de uma avaliação para verificar se sua execução acarreta em erro.
- Sequência errada: identificada como uma sequência cuja ordem de execução da tarefa viola pelo menos uma das regras da tarefa.

Este esquema fundamenta o trabalho de Jens Rasmussen cujas teorias desenvolvidas nos primeiros anos da década de 1980 e constantemente revistas pelo autor, contribuíram para a consolidação do campo de pesquisa sobre o erro humano (WATERSON et al., 2016). A evolução do modelo é caracterizada principalmente pela incorporação de diferentes categorias de fatores que influenciam a ocorrência do erro como, por exemplo, fatores organizacionais, afetivos, sociais.

Ao conceito geral do erro como um desvio, segue a necessidade de identificar classes de comportamentos observáveis que definem os tipos de erros. O modelo cognitivo baseado no processamento da informação sobre o qual se baseia uma taxonomia do erro humano proposta por Rasmussen foi escolhido a época dos primeiros congressos sendo considerado um dos mais promissores em termos de sua utilização prática.

2.1.4 Taxonomia do Erro Humano

Segundo Pereira (1983), para os participantes da segunda conferência internacional sobre o erro humano, o problema de uma teoria geral do erro pareceu desprovido de interesse. Embora, por fim, tivesse sido descartado enquanto não houvesse uma teoria geral do comportamento, mereceu considerações acerca de elementos para uma teoria futura. Este é o caso da associação do erro como resultado da má adaptação de um comportamento habitual a uma nova situação, segundo John Reason citado por (PEREIRA, 1983). Para Jens Rasmussen, a variabilidade do comportamento humano é um dos elementos fundamentais à sua adaptabilidade e capacidade de aprender em novas e imprevisíveis situações. Neste sentido, qualquer proposta instrucional deve oportunizar a realização de experimentos do tipo tentativa e erro (RASMUSSEN, 1982). Deriva-se daí que o erro é uma consequência necessária de qualquer processo de aprendizagem (PEREIRA, 1983). Assim, da teoria de tipos de erros lógicos e comunicação do psicólogo gestaltista Gregory Bateson, se a aprendizagem é um processo que implica algum tipo de incerteza em graus diferentes, então uma taxonomia da aprendizagem pode ser elaborada sobre uma taxonomia dos tipos de erros. Em suma, o erro é um indicador de mudança na aprendizagem (PEREIRA, 1983).

Aparentemente paradoxal, a posição do erro precisa ser esclarecida. Não se trata de uma relação direta (mais erro implica mais aprendizagem). Pelo contrário,

espera-se que, na medida em que um certo objetivo de aprendizagem tenha sido alcançado, o erro sobre a prática correlata deva ter sido eliminado. Isso, contudo, não permite afirmar que o erro, o mesmo erro que se supunha ter sido eliminado, não volte a acontecer, sob as mesmas circunstâncias e derivado da intervenção do mesmo sujeito. Na verdade, a recorrência do erro, independente do contexto específico, possibilita identificar padrões de erro que fornecem indícios sobre pontos críticos ancorados na interação humano-sistema (REASON; HOBBS, 2017). Adicionalmente, os padrões identificados fundamentam ações de caráter preventivo como as atividades de treinamento e aperfeiçoamento profissional. A mesma percepção é compartilhada por BLOOM et al. (1983) ao realçar a função da análise de erros no planejamento da instrução e, em particular, na elaboração de instrumentos de avaliação e classificação do conhecimento do aprendiz com acurácia e precisão.

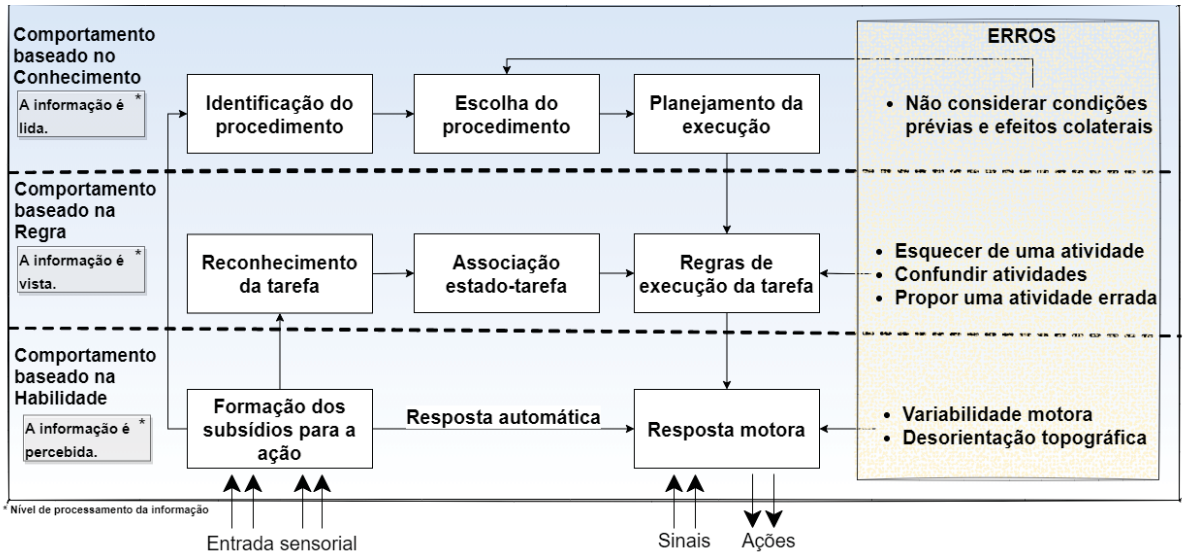
Conforme Reason e Hobbs (2017), a identificação de padrões do erro humano, através da correlação de sua frequência (dados históricos) e de seus tipos (taxonomia), associada à modelagem matemática baseada em métodos estocásticos, têm se constituído como um meio para a análise do erro. Isto têm sido verdade, principalmente, em relação às atividades de manutenção no sentido da identificação dos erros induzidos (inerentes ao sistema) ou provocados (decorrentes da tarefa). Ainda, segundo os autores, na aplicação de uma teoria ou modelo, para a modelagem do erro humano durante a execução de uma tarefa, são necessários a observância e ponderação em relação à complexidade de cadeias causais dos fatores que influenciam o erro. Esses fatores, denominados na literatura como Performance Shaping Factors (PSF), que podem ser endógenos ou exógenos ao sistema, ao humano ou à tarefa, influenciam no desempenho, levam à mudanças no comportamento e podem, segundo sua relevância ou consequências, levar ao erro.

Enfim, embora o erro possa ter como causa uma ação humana considerada insegura, a sua prevenção depende preliminarmente de uma análise diagnóstica sobre o ambiente inseguro, sobre a tarefa, ou atividade, crítica. Efetivada a detecção das fragilidades no processo, a redução dos erros por meio de medidas terapêuticas, prescritivas, são balizadas pela inclusão da retroalimentação (*feedback*) da informação (resposta) da ação humana sobre o sistema (PEREIRA, 1983).

Dentre vários modelos explicativos sobre o fenômeno erro humano, Jens Rasmussen inicia sua discussão baseado no paradigma do processamento da informação (RASMUSSEN, 1982). Os mecanismos internos do erro não podem ser observados diretamente, somente a manifestação do erro na interface da ação humana sobre o ambiente de trabalho. Esse modelo não prescritivo, designado SRK – *Skills-Rules-Knowledge* (FIGURA 8), relaciona os tipos e sequência do processamento interno da informação com as respectivas ativações subjetivas de conhecimentos e funções

mentais, identificadas durante a resolução de um problema e o processo de decisão: observação, identificação, interpretação, definição e planejamento dos procedimentos de execução da tarefa.

FIGURA 8 – MODELO SKR



FONTE: Adapatdo de Rasmussen (1982)

No processamento da informação, as funções mentais assumem diferentes tipos de processamento conforme a respectiva categoria do modelo SRK. Dessa forma, a entrada (*input*) de uma informação é percebida (*look*) no nível da habilidade (*skill*), vista (*see*) no nível das regras (*rules*) e lida (read) no nível do conhecimento (*knowledge*).

Durante a execução de uma tarefa, essa associação permite identificar a função mental correspondente ao efeito externo do erro de forma independente do contexto, ou nível de desempenho. Rasmussen (1982) relaciona às funções mentais (internas) que, porventura, tenham falhado a um comportamento observável associado ao erro. A investigação sobre a cadeia de causas que levam a um acidente provocado pelo erro humano é construída por meio da resposta a três perguntas básicas:

- O que deu errado?, ou seja, qual função interna falhou?;
- Como se deu o erro?, quais mecanismos internos falharam?;
- Por que ocorreu o erro?

Conforme apresentado em Reason e Hobbs (2017), o modelo *Skills-Rules-Knowledge*, SKR, ilustrado na FIGURA 8, define três níveis de desempenho humano facilmente reconhecidas por engenheiros e prontamente aceita pela maioria dos psicólogos cognitivistas. A adoção deste modelo cognitivo, fortemente baseado em controle de sistemas, fornece um esquema de relacionamento do erro a uma falha no processamento de três formas de informações provenientes do sistema – sinais, signos e símbolos –, as quais, independentemente da forma, são percebidas contextualmente a partir de intenções e expectativas humanas (RASMUSSEN, 1983). Os tipos de entrada (sinais, signos e símbolos) e as formas que estas assumem na interface de um sistema de controle onde, por exemplo, todas entradas derivam da Instrumentação, controle e processamento de dados do sistema. Segundo as informações disponíveis na interface do sistema o fluxo da informação deve disparar um procedimento com base numa regra, a qual depende de um plano de trabalho elaborado em um nível superior de processamento baseado na interpretação simbólica.

Em sua formulação original, as ações e tipos de informação baseados na habilidade (*skill-based*), na regra (*rules-based*) e no conhecimento (*knowledge-based*) são descritos da seguinte maneira:

- Ação baseada na habilidade: ação sensório-motora desencadeada na forma de rotinas mais ou menos subconscientes, isto é sem atenção ou controle consciente, que se caraterizam por padrões de comportamento integrados, automatizados e contínuos. Nesse caso, a ocorrência de erros está relacionada à defasagem entre o estado atual e o estado requerido no domínio espaço-tempo de sinais provenientes do sistema durante a interação humana.

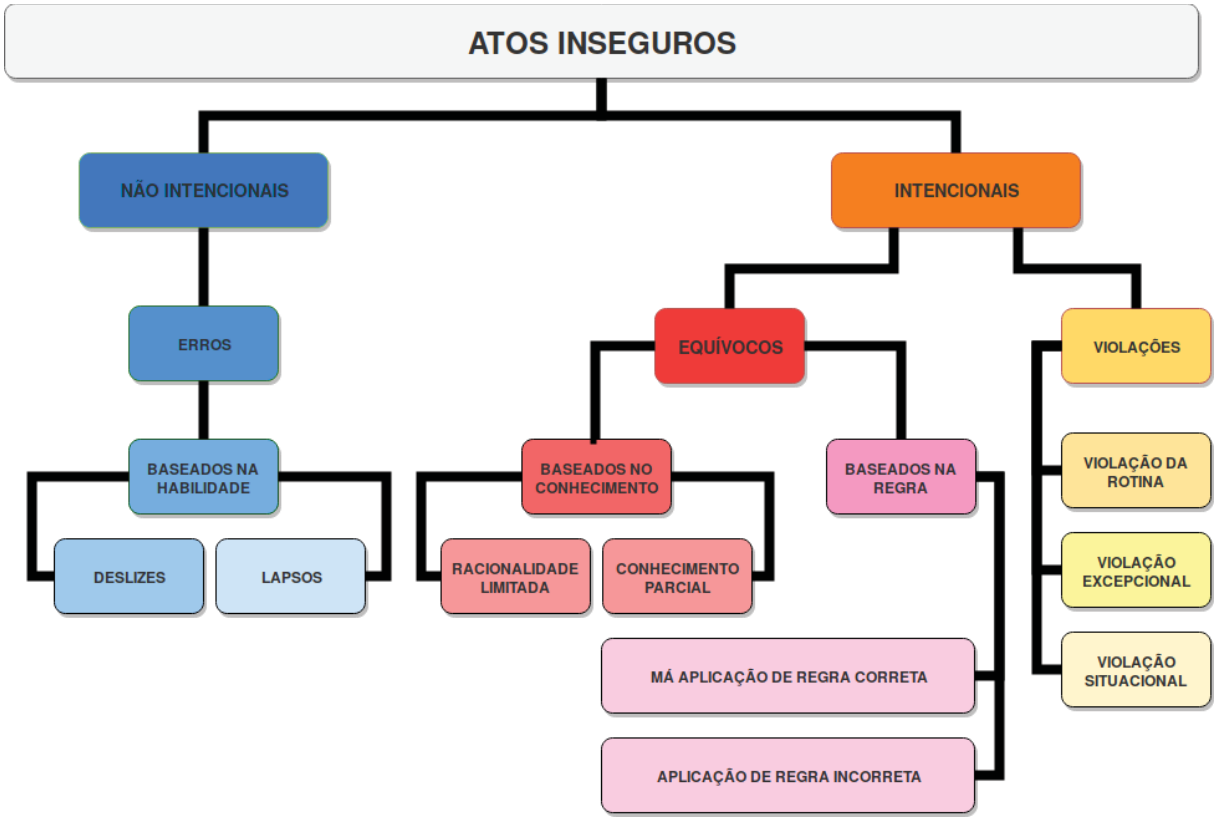
- Ação baseada na regra: caracterizada por uma sequência de rotinas executadas conscientemente em situações conhecidas com base em regras ou procedimentos prontamente armazenados na memória. Esse conhecimento pode ser resultado da experimentação prévia ou adquirida por meio de um processo instrucional. A informação relativa ao sistema é percebida como signos, ou dicas, que podem influenciar o controle da regra que ordena as sequências de rotinas preestabelecidas. Desse modo, o discernimento equivocado de uma situação ou o esquecimento de procedimentos são mecanismos associados à ocorrência de erros.

- Ação baseada no conhecimento: associada ao desempenho em situações inéditas ou desconhecidas que demandam um planejamento de uma ação a partir de análise baseada no conhecimento das propriedades do sistema e dos objetivos gerais do operador. Isso implica a elaboração de hipóteses no processo de decisão e a realização de testes, seja por meio da tentativa e erro, seja concei-

tualmente, por meio da antecipação dos efeitos da execução da ação planejada sobre o sistema. As informações derivadas do sistema são percebidas como símbolos, os quais estão associados a conceitos que possibilitam a adaptação a novas situações e predição do estado futuro do sistema em resposta a uma ação.

Reason e Hobbs (2017) oferecem um quadro compreensivo integrado a esta categorização (FIGURA 9). Uma ação ou ato inseguro pode ser dividido em ação intencional e ação não intencional. Às ações não intencionais estão associados os erros causados pela falta de atenção, ou deslizes (*slips*), as falhas de memória, ou lapsos (*lapses*), e os erros de reconhecimento. Já as ações intencionais derivam erros de dois tipos: equívocos (*mistakes*) e violações (ou transgressões) (*violations*).

FIGURA 9 – CATEGORIZAÇÃO DO ATO INSEGURO



FONTE: (REASON; HOBBS, 2017)

Os erros derivados de ações não intencionais estão relacionados à ação baseada na habilidade e são classificados segundo duas categorias. A primeira categoria inclui os erros de memória, ou lapsos, os quais resultam de uma falha no processamento da informação. Essa falha pode ocorrer tanto na codificação de entrada, por conta da falta de atenção, quanto na saída, devido a problemas na recuperação da

informação armazenada. A segunda categoria de erros não intencionais, os deslizes, ocorrem por conta do conhecimento automatizado da ação diante de uma situação, ou tarefa familiar, o que pode gerar desvio de atenção às mudanças em relação a uma habilidade bem estabelecida.

Os erros associados a ações intencionais são classificados em equívocos, ou *mistakes*, e uma outra categoria, tratada separadamente, as violações ou transgressões. Os equívocos são classificados em relação às ações baseadas em regras. Ainda conforme Reason e Hobbs (2017), atividades de manutenção são altamente estruturadas por meio de procedimentos que podem ser formais, descritos em manuais ou normas, mas também podem fazer parte de um repertório individual ou coletivo, na forma de um conhecimento tácito, ou adquiridos em treinamentos. O desvio em relação às regras se manifestam de duas formas conforme ocorra:

- a aplicação de uma regra corretamente, porém numa situação inadequada, porque não fora identificada corretamente;
- a aplicação de uma regra errada numa situação corretamente identificada.

Os equívocos também se manifestam como erros baseados no conhecimento, ou seja, são falhas de mecanismos de processamento associados a estratégias cognitivas e a falta de informação.

Finalmente, as violações, embora possam se manifestar como uma falha em aplicar regras e procedimentos corretamente, constituem uma categoria diferenciada de ato inseguro. As violações se manifestam sob três formas, a saber:

- Violações às rotinas, no sentido de minimizar esforços, considerados desnecessários;
- Violações excepcionais, de fundo emocional (*thrill-seeking*), caracterizada por superestimar a própria ação ou conhecimento, em função do exibicionismo ou por brincadeira;
- Violações situacionais, quando se verifica a impossibilidade de execução da tarefa seguindo estritamente as regras estabelecidas.

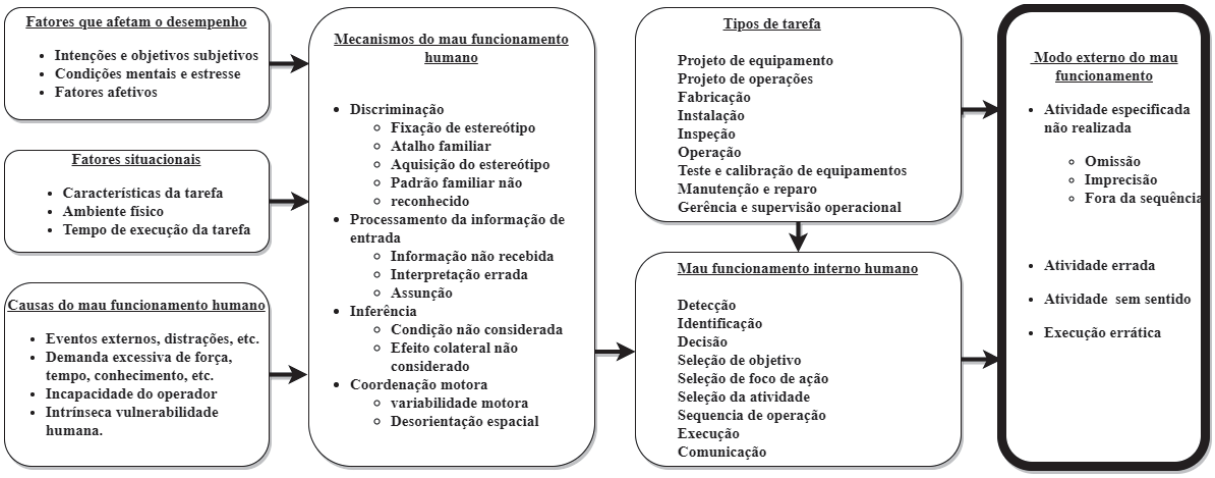
Em termos de uma classificação do mal funcionamento humano, Rasmussen (1982) lista uma série de categorias que podem ser usadas na análise dos eventos que desencadeiam o erro, permitem identificar os itens para os quais serão coletados dados de interação humano-sistema e, finalmente, possibilitam a quantificação do erro humano. As categorias estão organizadas hierarquicamente segundo níveis de

influência. A categoria de mais alto nível, modo externo do mal funcionamento humano (external mode of malfunction, em destaque na FIGURA 10), inclui os tipos de erros:

- Tarefa não executada por conta da omissão da ação;
- Performance não acurada;
- Ação executada na ordem errada;
- Execução fora tempo esperado;
- Ação incorreta;
- Ação errada, executada sobre um componente correto;
- Ação correta executada sobre componente errado;
- Ação executada no tempo incorreto.

Conforme a FIGURA 10, o erro observável, como uma manifestação do comportamento do indivíduo, admite uma rede de causas fundadas nos mais diversos domínios. Esse esquema exige que considerações acerca do erro humano incorporem os fatores presentes na sequência de eventos que levam ao erro.

FIGURA 10 – TAXONOMIA PARA DESCRIÇÃO E ANÁLISE DE FALHAS HUMANAS



FONTE: Adaptado de Rasmussen (1982)

Assim como no caso das taxonomias da aprendizagem, a taxonomia proposta tem mais importância por conta da sua estrutura do que pela apresentação dos itens

que a compõem. Nessa taxonomia, assim como em outras, ressalta-se o caráter pedagógico do estudo do erro humano. Nesse sentido, a classificação de comportamentos observáveis em relação ao desempenho humano tem como finalidade fornecer subsídios para ações que reduzam a ocorrência do erro. Na medida do possível, essa redução depende de ferramentas de coleta e análise do erro por meio de sua predição, ou estimação. Inúmeros modelos e taxonomias têm sido propostos, não existindo um modelo geral que possa ser aplicado de forma irrestrita. Olivares et al. (2018) sistematizaram um sumário dos modelos conceituais e taxonomias para a análise de confiabilidade e as respectivas técnicas de estudo do erro humano e capacidade de detecção de tipos de erros.

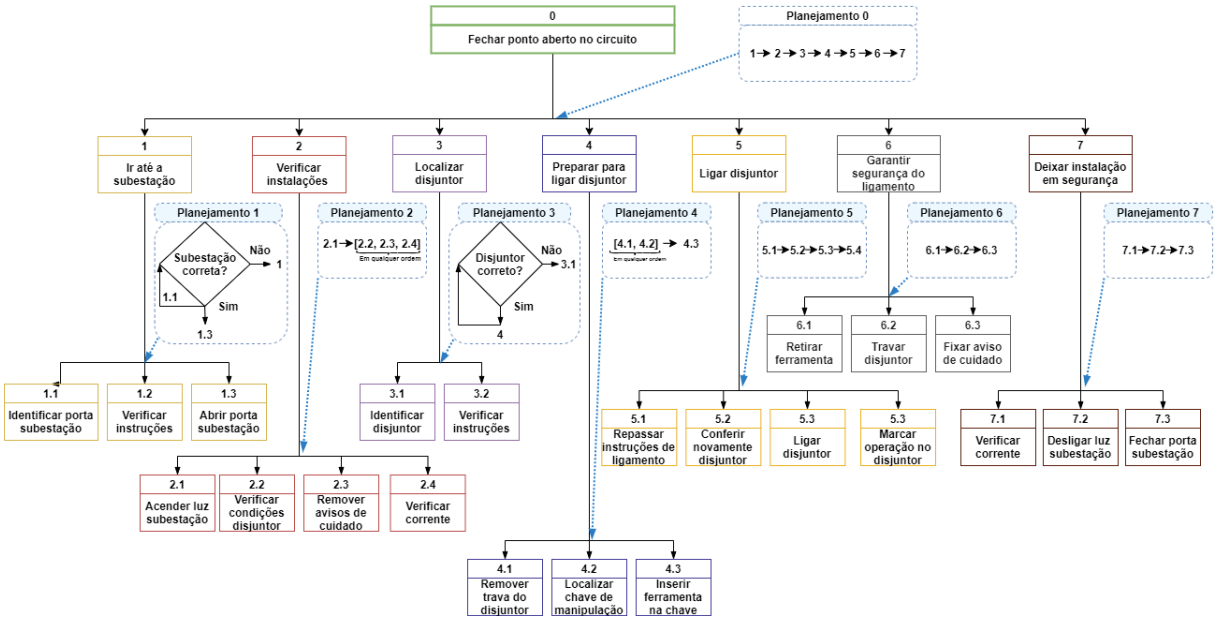
De modo geral, o erro humano é identificado por meio de categorias de comportamentos observáveis. No contexto específico de uma atividade, a identificação do erro, como em Baber e Stanton (2002), é realizada a partir de uma teoria geral dos sistemas. Conforme os autores, os erros acontecem na fronteira, entre os componentes do sistema e o componente humano, no momento da interação, o que corrobora o esforço de procurar na interface homem-sistema a manifestação do erro. Os autores propõem uma técnica de identificação de erro humano onde o planejamento é visto como um processo hierárquico que controla e guia a ordem das sequências de execução da tarefa. A análise formal do domínio do problema converte a tarefa em rotinas de procedimentos que potencialmente podem alterar o estado do sistema.

2.1.5 Análise da tarefa

A técnica de análise, denominada TAFEI, acrônimo para *Task Analysis For Error Identification*, de autoria de Baber e Stanton (2002), é orientada a cenários conforme o mapeamento da atividade humana como estado de máquina durante a execução da tarefa. Seus principais componentes e resultados associados são:

- A Análise Hierárquica da Tarefa (*Hierarchical Task Analysis* – HTA): descrição da atividade humana em termos decomposição da tarefa (FIGURA 11).
- O diagrama de estado-espaço (*State-Space Diagrams* – SSD): descrição dos estados do sistema durante a execução da tarefa.
- Matrizes de Transição (*Transiction Matrices* – TM): matriz de transição de estados permite o mapeamento das transições possíveis entre os estados. Por meio dessa matriz é possível determinar ações (transições entre estados) potencialmente erradas.

FIGURA 11 – ANÁLISE HIERÁRQUICA DA TAREFA



FONTE: (BABER; STANTON, 2002)

A Análise Hierárquica da Tarefa tem como propósito tanto diagnosticar e descrever o erro quanto prescrever medidas para sua prevenção. Trata-se de uma ferramenta de modelagem do comportamento humano, a qual pretende capturar as estratégias de execução da tarefa com o objetivo de identificar um desempenho inadequado que pode ser melhorado. Em termos práticos, a análise leva em consideração que uma tarefa a ser realizada por uma pessoa é definida por objetivos que podem ser decompostos iterativamente numa árvore hierárquica de objetivos e sub-objetivos.

Dada uma tarefa determinada, Shepherd (2003) define uma estrutura de trabalho para a sua análise, na qual, pode-se identificar se um desempenho esperado é aceitável conforme certos objetivos e padrão de execução. A matriz de transição dada na TABELA 2 sumariza as dependências entre as 12 primeiras atividades da tarefa segundo o planejamento ou execução esperada apresentada na FIGURA 11.

A análise da tarefa tem igual importância no design instrucional sendo aplicada nas etapas de planejamento da instrução, incluindo a definição dos resultados esperados de desempenho. Por meio da análise da tarefa o conhecimento é codificado em termos de objetivos a serem alcançados e segundo quais regras. O conjunto de regras que descrevem a tarefa é usado para defini-la formalmente num espaço de possíveis soluções para a sua execução, denominado modelo do conhecimento. Dessa forma, o modelo do conhecimento sintetiza os objetivos da aprendizagem sobre os quais o desempenho é avaliado.

TABELA 2 – MATRIZ DE TRANSIÇÃO

	1.1	1.2	1.3	2.1	2.2	2.3	2.4	3.1	3.2	4.1	4.2	4.3	⋯
1.1	0	L	I	0	0	0	0	I	0	I	0	0	⋯
1.2	0	0	L	I	0	0	0	I	0	I	0	0	⋯
1.3	1	0	0	L	I	0	0	I	0	I	0	0	⋯
2.1	0	0	0	0	L	L	L	I	I	0	0	0	⋯
2.2	0	0	0	0	0	L	L	L	I	I	0	0	⋯
2.3	0	0	0	0	L	0	L	L	I	I	0	0	⋯
2.4	0	0	0	0	L	L	0	L	I	I	0	0	⋯
3.1	0	0	0	0	0	0	0	L	L	I	0	0	⋯
3.2	0	0	0	0	0	0	0	L	0	L	L	I	⋯
4.1	0	0	0	0	0	0	0	0	0	0	L	L	⋯
4.2	0	0	0	0	0	0	0	0	0	0	0	L	⋯
4.3	0	0	0	0	0	0	0	0	0	0	0	0	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	

FONTE: Adapatado de Shepherd (2003)
LEGENDA: **L**: transição legal; **I**: transição ilegal; **0**: transição impossível.

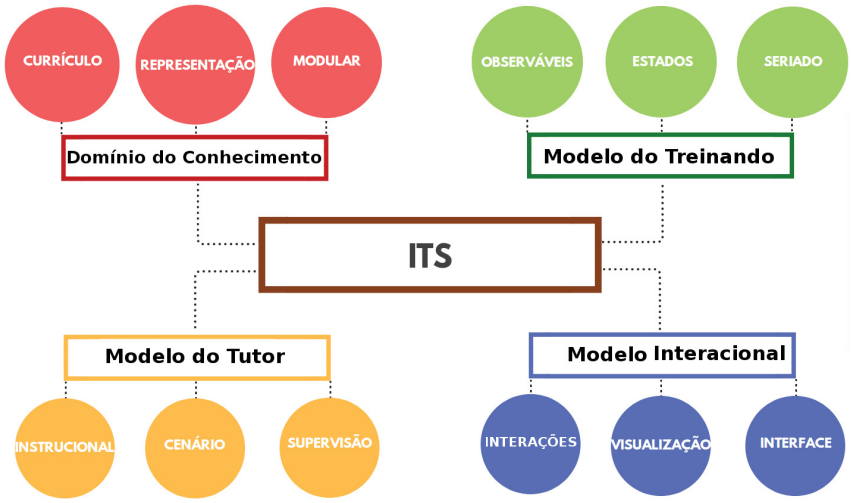
2.1.6 Modelagem de conhecimento e do desempenho

Neste trabalho, a aprendizagem e a avaliação são discutidas a partir de suas aplicações em um sistema computacional cuja arquitetura tem sido bem estudada e definida a partir do desenvolvimento de Sistemas Tutores Inteligentes (*Intelligent Tutor Systems* – ITS). Sistemas Tutores Inteligentes são programas de computador, que incorporam técnicas de inteligência artificial para simular o comportamento de um instrutor em relação ao "o que", "para quem"e "como ensinar"(NTUEN; CHESTNUT, 1995). Entretanto, como lembra Baker (2016), esta visão se mostrou limitada e dela decorrem muitas experiências malsucedidas no uso desta tecnologia. Finalizando, o autor afirma que a despeito dos esforços para o desenvolvimento de sistemas inteligentes, na verdade, o precisa ser feito, é, de forma inteligente, desenvolver sistemas tutores que alavanquem a inteligência humana. A definição, ou mesmo aceitação, do termo **inteligente** é controversa e a adoção de denominações alternativas, como sistemas adaptativos ou baseados no conhecimento ajudam a esclarecer algumas promessas iniciais sobre o que significa a inteligência de tais sistemas (SOTTILARE, 2018). Muitas vezes, a associação é devida ao uso de um método para o tratamento de dados, a automação de uma tarefa instrucional rotineira.

Conforme Pavlik Jr et al. (2013), ao longo do seu desenvolvimento desde a década de 1980, a arquitetura básica de um sistema tutorial inteligente tem sido descrita por meio de quatro componentes ou módulos, conforme apresentado na FIGURA 12 :

- o modelo ou domínio do conhecimento (*knowledge domain* ou *expert model*) refere-se ao conjunto de habilidades, conhecimentos e estratégias que definem os objetivos da aprendizagem de uma tarefa instrucional. O domínio do conhecimento

FIGURA 12 – COMPONENTES DE UM SISTEMA TUTOR INTELIGENTE



FONTE: O Autor com base em Pavlik Jr et al. (2013)

sintetiza o conhecimento do especialista na forma de estados de conhecimento válidos em relação ao qual os aprendizes devem se aproximar. Adicionalmente, as violações às regras e equívocos podem ser incorporados ao modelo (PAVLIK JR et al., 2013).

- o modelo do aprendiz tem sido um dos principais objetos de estudo no desenvolvimento de sistemas tutores inteligentes (SOTTILARE et al., 2013). A modelagem do aprendiz resulta do mapeamento do seu estado de conhecimento e das modificações nesse estado incorporando fatores cognitivos, afetivos, instrucionais e sociais. O modelo do aprendiz pode ser visto como um subconjunto ou como um desvio do domínio do conhecimento (PAVLIK JR et al., 2013). Conforme a aplicação, o modelo do aprendiz recebe outras denominações com o termo aprendiz sendo substituído por estudante, usuário, treinando ou operador (*user model, operator model, learner model, student model, trainee model*). Neste trabalho estes termos serão utilizados de forma equivalente, sendo um ou outro escolhido conforme o contexto em discussão.

- o modelo instrucional (*tutor model*) corresponde as estratégias adotada no planejamento da instrução incluindo a seleção e recomendação de tarefas por meio dos resultados da análise do modelo do apendiz e o domínio conhecimento (PAVLIK JR et al., 2013).

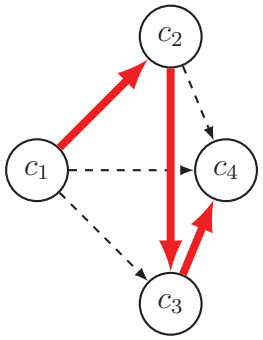
- o modelo de interações (*interaction* ou *communication model*) controla os recursos de interação com o aprendiz como a seleção de alternativas, a visualização do modelo do aprendiz e dos resultados associados ao desempenho (PAVLIK JR et al., 2013).

Dois destes módulos serão discutidos com mais ênfase ao longo do trabalho pois pertencem à classe de problemas relacionados ao objetivo desta tese, o modelo do conhecimento e o modelo do aprendiz. A inter-dependência entre ambos permite a definição do modelo do aprendiz sobre um espaço do conhecimento das soluções, ou estados do conhecimento. O espaço do conhecimento não é equiprovável pois depende não somente da estrutura subjacente à tarefa instrucional, mas também da ordem com a qual a tarefa instrucional foi ensinada. Dessa forma, a rigor, existe uma expectativa de que a tarefa seja executada de forma adequada e defina um estado de conhecimento do aprendiz próximo do conhecimento do especialista.

A relação entre o espaço do conhecimento e o estado do conhecimento é formalizada na chamada teoria do espaço do conhecimento (*Knowledge Spaces Theory*) desenvolvida por Doignon e Falmagne (1999). Uma importante característica dos modelos baseados em espaços de conhecimento é que o domínio do conhecimento é facilmente codificado pois assume uma estrutura de grafo onde os nós representam unidades de aprendizagem e as arestas indicam as suas possíveis combinações na execução da tarefa.

Dada uma tarefa instrucional os possíveis estados do conhecimento incluem cada uma das atividades que compõem a tarefa e as combinações de transição possíveis entre um estado e outro. Assim, sem perda de generalidade para uma tarefa com quatro atividades definimos o espaço de conhecimento como um conjunto $\mathcal{K} = \{\emptyset, \{c_1\}, \{c_2\}, \{c_1, c_2\}, \{c_1, c_3\}, \{c_1, c_4\}, \{c_1, c_2, c_4\}, \{c_1, c_2, c_3\}, \{c_2, c_3, c_4\}, \{c_1, c_2, c_3, c_4\}\}$ cujos elementos c_i, $i = 1, \cdots, 4$,são as unidades curriculares dos estados do conhecimento e $\{\emptyset\}$ corresponde ao estado de conhecimento inicial. (DOIGNON; FALMAGNE, 1999). A estrutura do espaço do conhecimento $\mathcal{K}$ é representada pelo grafo de mesmo nome da FIGURA 13. Neste exemplo, o caminho $\mathcal{P} = \{c_1, c_2, c_3, c_4\}$ sobre o grafo $\mathcal{K}$ representa a única sequência viável para completa execução da tarefa.

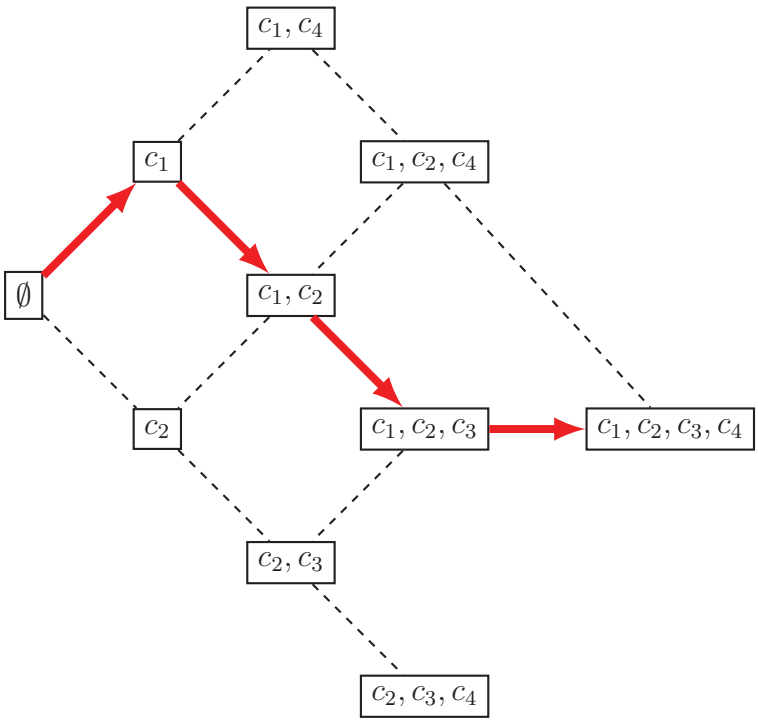
FIGURA 13 – Grafo $\mathcal{K}$ e um caminho $\mathcal{P}$



FONTE: O Autor

Definido dessa forma, o espaço do conhecimento pode ser descrito por seus aspectos combinatórios permitindo a caracterização probabilística dos estados de conhecimento.

FIGURA 14 – Representação do caminho $\mathcal{P}$ sobre o espaço de conhecimento induzido pelo grafo $\mathcal{K}$



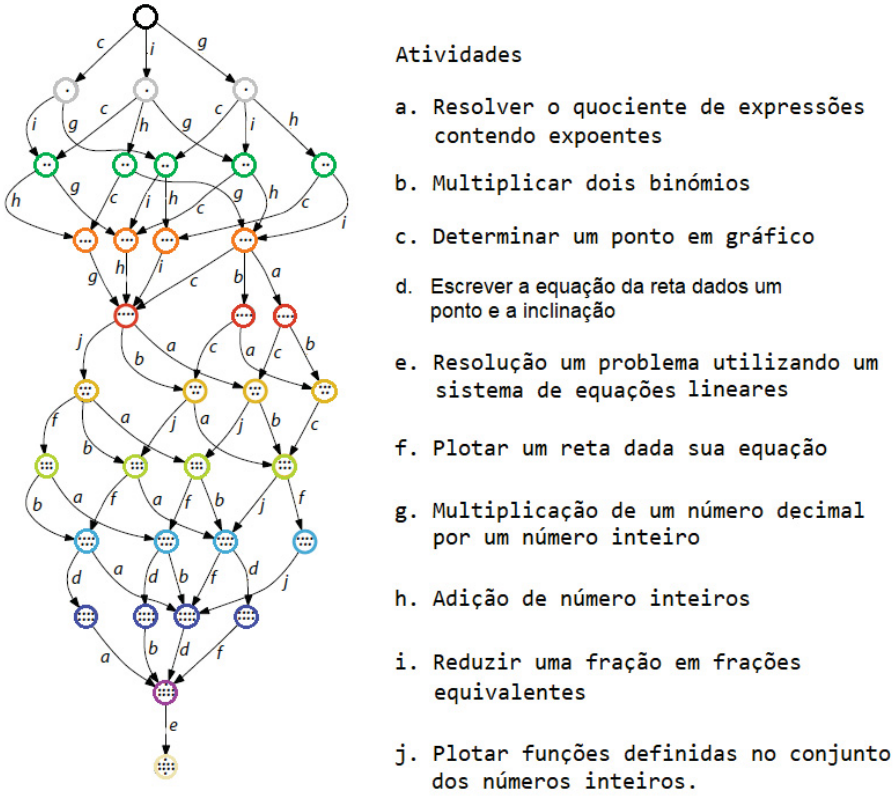
FONTE: O Autor

No grafo da FIGURA 14 a atribuição de probabilidades às conexões entre os estados pode ser **a priori**, quando são conhecidas as probabilidades de que o aprendiz alcance um determinado estado do conhecimento; e **a posteriori** quando essas probabilidades são estimadas a partir da inferência do estado do aprendiz através da modelagem bayesiana do domínio do conhecimento. Nesta concepção, o caminho $\mathcal{P}$ dado pela sequência de flechas no grafo da FIGURA 14 representa a evolução do desempenho do aprendiz.

Um exemplo de codificação da resolução de uma tarefa contendo atividades de matemática básica é apresentado em Falmagne et al. (2013). Como ilustrado, o grafo mapeia as sequências ordenadas das 10 atividades indicadas em ordem alfabética de a. até j.

Como apresentado, o modelo do aprendiz descreve o desempenho na execução da tarefa e pode ser definido a partir do domínio do conhecimento. Conforme

FIGURA 15 – REPRESENTAÇÃO DO ESPAÇO DE CONHECIMENTO COMO UM GRAFO



FONTE: (FALMAGNE et al., 2013)

Chrysafiadi e Virvou (2013), a modelagem do aprendiz pode ser multidimensional associando ao aprendiz não somente os conhecimentos e habilidades, mas também os erros e os equívocos, fatores afetivos, cognitivos e meta-cognitivos como preferências e estilos de aprendizagem. Pavlik Jr et al. (2013) destacam algumas características para a construção de um modelo do aprendiz:

- ser ajustável aos dados, representando qualitativamente e quantitativamente os padrões de desempenho dos aprendizes;
- ser de fácil entendimento para pedagogos, designers instrucionais, instrutores, aprendizes e desenvolvedores dentre outros;
- ser flexível em relação ao contexto, podendo ser estendido e adaptado a novas situações;
- ter baixo custo, não exigindo a obtenção de grande volumes de dados para ser implementado;
- permitir a granularidade por meio do refino de características e incorporação de novas variáveis relativos a sub-processos na execução de uma tarefa;

- ser flexível em relação ao tempo permitindo o seu uso para monitoramento do aprendiz a longo prazo;
- evidenciar os ganhos de aprendizagem na prática por meio da avaliação da eficiência na transferência do aprendizado para a situação real.

Embora estas características funcionem como critérios para o desenvolvimento e validação dos sistemas tutores, os autores reconhecem que a implementação de todas elas é muito difícil, senão impossível. Em especial a última delas para a qual os estudos são escassos e limitados.

As abordagens mais estudadas para a modelagem do aprendiz são

- a modelagem baseada em restrições (MBR), *Constraint Based Modeling*: o modelo do conhecimento é dado por um conjunto de regras que define o padrão de resposta esperado para uma tarefa. O modelo do aprendiz se apresenta como uma pertubação caracterizada por um desvio do modelo do conhecimento. A detecção de um erro o qual se manifesta como uma violação a alguma regra conhecida depende da magnitude e qualidade deste desvio.
- a modelagem baseada no rastreamento do aprendiz (*Tracing Learning Model*): O modelo do aprendiz é codificado por um conjunto de regras que resultam da execução de uma tarefa. Esta abordagem se baseia na Teoria do Controle Adpatativo do Pensamento (*Adaptive Control of Thougth (ACT)*) cujo objetivo é predizer o desempenho dos aprendizes por meio de um modelo cognitivo denominado Tutor Cognitivo (*Cognitve tutor*) com enfase nos controles dos mecanismos internos associados à memória (RITTER et al., 2007).

Ambas abordagens são amplamente usadas e os modelos gerados são satisfatórios no que diz respeito aos critérios de validação elencados anteriormente, mas algumas limitações em relação ao tratamento do erro são apontadas em relação à segunda delas. A modelagem baseada no rastreamento do aprendiz enfatiza mais a falha em relação à resposta esperada provendo subsídios para sua correção do que a explicação do erro. Dessa forma, um erro não permite a continuidade da execução até que a resposta esperada seja alcançada.

Na modelagem do aprendiz baseada em restrições, seguindo Mitrovic e Ohlsson (STELLAN; MITROVIĆ, 2006) a condução do processo de instrução pode ser discutida em termos da organização e representação do conhecimento e a forma de reação do sistema em relação aos erros cometidos pelo usuário. Muitas tarefas não podem ser estritamente codificadas de forma linear, isto é, a transição entre dois estados do sistema não pode ser descrita por meio de uma sequência de ações bem definidas

de forma contínua. Nesse caso, a decomposição da tarefa resulta numa sequência de estados discretos que correlacionam o domínio do conhecimento baseado em regras e o desempenho que não satisfaz essas regras.

No começo da década de 1980, fundado em um modelo cognitivo baseado no processamento da informação, a MBR tinha como objetivo contribuir com o ensino de matemática para crianças e adolescentes. Os estudos realizados desde então relacionam a influência do conhecimento declarativo, como as regras de execução da tarefa, e o respectivo conhecimento procedimental o qual corresponde a efetiva sua execução (OHLSSON, 2015).

A modelagem baseada em restrições baseia-se sobre uma teoria da aprendizagem que evidencia o papel do erro para o processo de aprendizagem por meio de três princípios;

- Aprender é resolver problemas que se movem de um contexto específico para outro geral;
- Aprender é se adaptar gradualmente as novas situações e conhecimentos;
- Aprender é mediar os conflitos cognitivos entre crenças prévias e novas situações.

Dessa forma, contradições, erros, falhas, impasses e respostas erradas são substituídas por meio da detecção e correção dos erros pelo próprio aprendiz de modo que seja alcançado o conhecimento do especialista (OHLSSON, 1993). A aplicação da MBR no desenvolvimento de sistemas tutores inteligentes é realizada em três etapas

1. Das proposições às restrições: o conhecimento do especialista no contexto específico da aplicação da tarefa é codificado em termos das restrições que a descreve;
2. Da detecção do erro a sua correção: os erros dos aprendizes são identificados e rastreados no decurso do desenvolvimento da tarefa;
3. Da aprendizagem à tutoria: os erros assinalados são evidenciados para o aprendiz segundo o seu tipo e impacto na execução da tarefa.

Em sistemas inteligentes, os erros detectados revelam padrões associados ao espaço de busca das soluções corretas. Sua detecção é realizada comparando-se um padrão de execução do treinando com um padrão de execução esperado onde as restrições são todas respeitadas. Dessa forma, o modelo do treinando captura os desvios na execução da tarefa. Outra vantagem associada a a esta concepção é a

simplicidade computacional para mapear e contabilizar os erros cometidos sob a forma de comportamentos não adequados às restrições impostas na execução da tarefa.

A utilização de métodos híbridos, associada à sistemas de treinamento baseados em simulação, tem sido considerada um dos melhores meios complementares para o estudo do erro humano. Conforme Taylor (2004) , a criação de ambientes virtuais fidedignos com a realidade, por meio de tecnologias de realidade virtual para formação e treinamento de profissionais, tem permitido a produção de grandes quantidades de dados, os quais dificilmente poderiam ser obtidos em situações reais (NARANJO et al., 2020). Enquanto a elaboração de cenários de risco em simuladores tem ampliado o repertório sobre os tipos de erros e seus fatores de influência. Essa qualidade soma-se à capacidade dos sistemas de treinamento monitorarem o estado de conhecimento do usuário e capturarem aspectos acerca da aprendizagem do indivíduo.

Para Chrysafiadi e Virvou (2013) a utilização de técnicas de aprendizado de máquina têm possibilitado a inferência de estados de conhecimento desejáveis na forma de uma expectativa sobre o desempenho futuro de forma acurada. Além disso, modelos mais complexos podem avançar sobre outras características do desempenho dos usuários de modo a identificar e mapear atributos subjetivos associados a estratégias cognitivas ou reações afetivas. Na área de operação de sistemas supervisórios de sistema elétricos, diferentes trabalhos têm buscado compreender o fenômeno do erro humano. Destacam-se os trabalhos de Vieira e outros (QUEIROZ VIEIRA et al., 2007; NETO et al., 2009; FOCKING et al., 2012; NETTO et al., 2014a,b) no sentido de modelar o comportamento de operadores de painéis de supervisão de sistemas de potência. Na área de saúde, onde a questão do risco é um fator importante na formação de médicos e enfermeiros, Moraes e outros (FERREIRA et al., 2015; MORAES; SANTOS MACHADO, 2005; MORAES et al., 2009; MACHADO; MORAES, 2012) têm proposto e comparado diferentes técnicas de inteligência artificial para a avaliação da aprendizagem em sistemas de treinamento baseados em realidade virtual, como máquinas de estados finitos, lógica fuzzy, sistemas baseados em regras, redes neurais, redes bayesianas e algoritmos genéticos (COSTA et al., 2001).

2.2 APRENDENDO COM OS DADOS

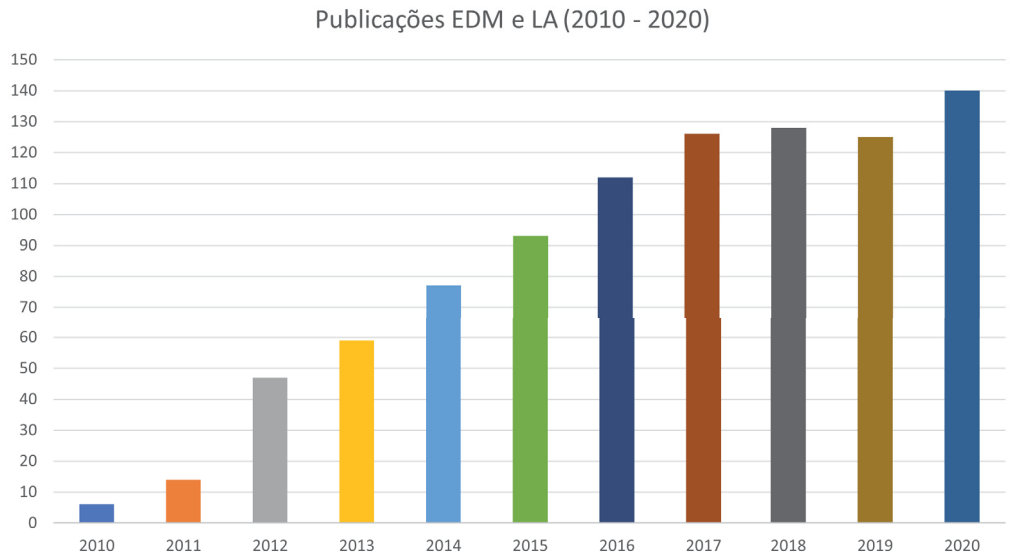
2.2.1 Mineração de dados instrucionais: tendências e aplicações

A mineração de dados aplicada ao contexto da educação e treinamento (*Educational Data Mining – EDM*) é uma área de pesquisa sobre os processos de aprendizagem mediados pelo computador cujo desenvolvimento remonta ao começo dos anos 2000. Conforme Liñán e Pérez (2015), os estudos em EDM são acompanhados de uma vertente denominada Análise da Aprendizagem e Conhecimento (*Learning Analytics and Knowledge – LAK*) cujos objetos e métodos se intersectam. Em ambos a

análise de dados provenientes de ambientes de aprendizagem é realizada por meio da incorporação de técnicas de aprendizado de máquina.

Para ilustrar o aumento no número de pesquisas nestas área foi realizado um levantamento utilizando ("Educational Data Mining"OR "Learning Analytics") como descritores para títulos e palavras-chaves. A pesquisa foi realizada com a ferramenta "Publish or Perish"[2] utilizando o motor de busca do Google Acadêmico. A FIGURA 16 apresenta a distribuição das 980 publicações encontradas para a década de 2010 a 2020.

FIGURA 16 – PUBLICAÇÕES EM EDM E LA NA DÉCADA DE 2010 A 2020



FONTE: O Autor

Ainda segundo Siemens e Baker (2012) as diferenças entre as duas áreas podem ser categorizadas, por exemplo, segundo o tipo de descoberta, a abrangência, as origens e as técnicas e métodos. Na **mineração de dados educacionais** são recorrentes o estudo dos componentes dos sistemas de aprendizagem, como o modelo do aprendiz e o domínio do conhecimento, e a investigação de ferramentas que promovam a adaptação automática destes sistemas às necessidades dos aprendizes utilizando-se de técnicas de agrupamento e classificação, modelagem bayesiana, descoberta de regras de associação e visualização (MANJARRES et al., 2018). A **análise da aprendizagem e conhecimento** tem suas origens nos estudos sobre redes semânticas e

[2] "Publish or Perish"é um software para a análise de citações acadêmicas desenvolvido desde 2007 por Anne-Wills Harzigan, professora de Comércio Exterior (International Management) na universidade de Middlesex na Inglatera. O software pode ser configurado para utilizar diferentes motores de busca como o CrossRef, Scopus, WebOfScience e o Goolge Acadêmico, dentre outros. A versão mais recente é de 2019.

predição de resultados da aprendizagem em um contexto integrado de diferentes fontes de dados. As técnicas que tem se destacado na análise da aprendizagem incluem a análise de redes sociais, a análise de sentimento, a análise do discurso, dentre outras. Embora alguns limites possam ser traçados entre as duas áreas, as suas sobreposições e complementaridades são mais importantes do que qualquer antagonismo (BAKER; INVENTADO, 2014).

Conforme Romero e Ventura (2007) e Romero e Ventura (2020), diversos métodos para análise tem sido utilizados na descoberta de conhecimento em bases de dados instrucionais como, por exemplo, a predição (AGHABABYAN et al., 2018), o agrupamento (SILVERMAN, 1967) e a visualização (DOGAN; CAMURCU, 2009) de estados de conhecimento de aprendizes (SOTTILARE et al., 2018), a detecção de comportamentos anômalos (BAKER; INVENTADO, 2014; LU, 2018) e ferramentas para a análise de redes sociais (SHAFFER; RUIS, 2017). Dentre estes métodos, conforme o objetivo desta tese, identificar classes de desempenho da tarefa instrucional, o agrupamento de dados será discutido mais detalhadamente. O agrupamento de dados tem sido amplamente usado na mineração de padrões de comportamento e estilos de aprendizagem de usuários de ambientes virtuais (JOVANOVIC et al., 2012; HOOSHYAR et al., 2020). Os métodos mais utilizados em dados educacionais são o k-médias (*K-means*) (CHADHA, 2018), os métodos hierárquicos (MCBROOM et al., 2020) e os mapas auto-organizáveis (*Self-organized maps – SOM*) (FOSSEY, 2017; RAMOS et al., 2016; WULANDARI et al., 2020).

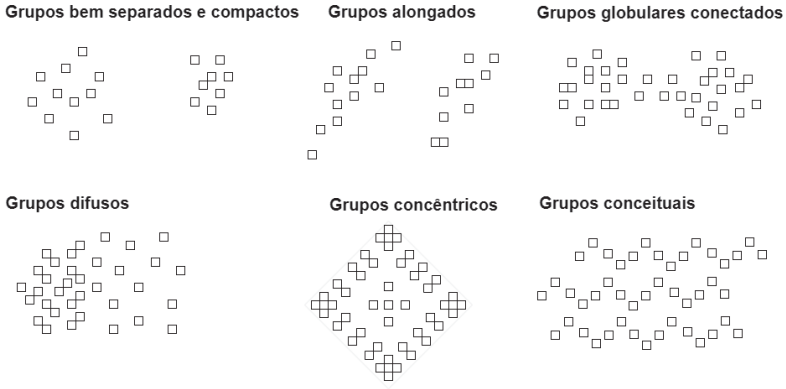
2.2.2 Agrupamento de dados

O agrupamento de dados é utilizado na análise exploratória com o objetivo de categorizar os dados em classes de similaridade por meio do aprendizado não supervisionado, ou seja, sem o conhecimento prévio dos rótulos de classe dos elementos do conjunto de dados. De modo geral, os grupos são definidos a partir da análise da matriz dos padrões de dados, em caso de dados categóricos por exemplo, ou da matriz de similaridade, ou de dissimilaridade, conforme o caso, no caso de dados numéricos (JAIN; DUBES, 1988). Além disso, os rótulos de classe dos grupos em um agrupamento podem ser utilizados no pré-processamento dos dados para a classificação supervisionada.

Conforme Jain e Dubes (1988) os métodos de agrupamentos podem ser classificados segundo os grupos formados definam (1) uma estrutura hierárquica, com grupos aninhados, ou (2) uma estrutura em partições com grupos disjuntos.

A definição de partição como grupo disjunto restringe a ideia de grupos a uma estrutura

FIGURA 17 – ESTRUTURAS GEOMÉTRICAS DE AGRUPAMENTOS



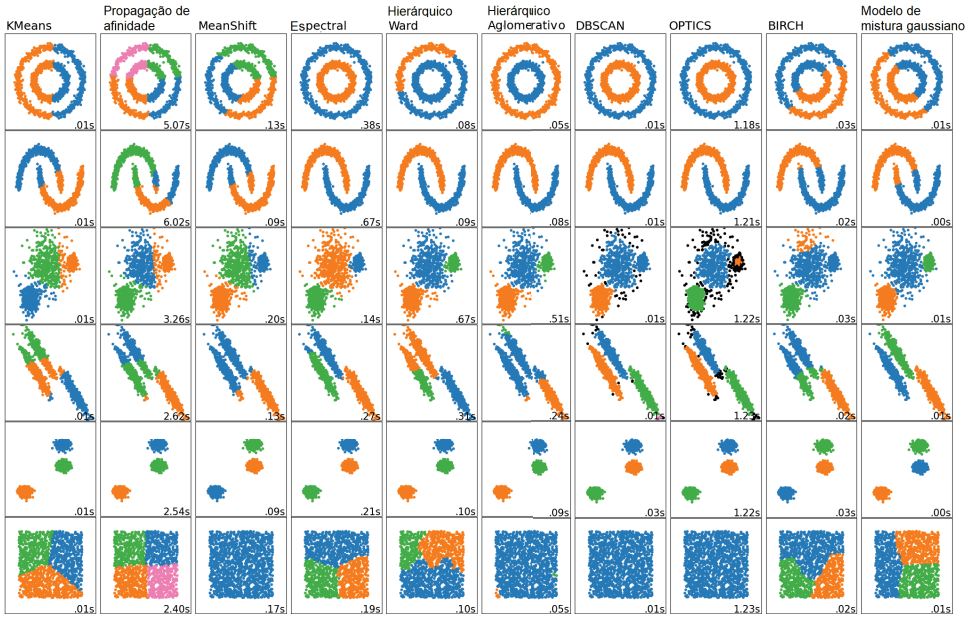
FONTE: Baseado em Murtagh e Heck (2012)

bem separada e compacta não contemplando estruturas como as representadas na FIGURA 17 que ilustram agrupamentos com geometrias globulares conectadas, alongadas, concêntricas, difusas e conceituais ou regradas (MURTAGH; HECK, 2012). Como ilustrado na FIGURA 18, os resultados da aplicação dos métodos de agrupamentos são sensíveis aos tipos de distribuição geométrica e estrutura dos dados (RODRIGUEZ et al., 2019). Na figura, cada linha de gráficos corresponde a um conjunto de dados com estruturas geométricas de grupos conhecidas. Cada coluna de gráficos identifica um método de agrupamento e os grupos formados por estes métodos são diferenciados entre si por meio de cores diferentes. A qualidade dos grupos formados por cada método pode ser inspecionada visualmente por meio da correspondência entre a distribuição do grupos por cores e a distribuição geométrica original de cada conjunto de dados.

O reconhecimentos de partições em distribuições de dados complexas tem sido realizado por outras categorias de métodos como aqueles (3) baseados na densidade do pontos de dados, (4) em modelos de mistura, (5) em grafos, (6) em redes neurais e (7) em conceitos (AGGARWAL, 2013).

Outras características como o número e o tamanho de grupos que o método é capaz de identificar podem servir como critérios para a escolha do método a ser aplicado. O *K-means*, traduzido para o português por **k-médias**, em destaque na TABELA 3, é um dos métodos de agrupamento mais simples e estudado, sendo considerado de propósito geral com melhores resultados quando aplicado sobre conjunto de dados com uma estrutura geométrica simples(JAIN, 2010) .

FIGURA 18 – APLICAÇÃO DE MÉTODOS DE AGRUPAMENTOS SOBRE DIFERENTES DISTRIBUIÇÕES DE DADOS



FONTE: https://scikit-learn.org/stable/modules/clustering.html#overview-of-clustering-methods

TABELA 3 – MÉTODOS DE AGRUPAMENTOS *SCIKIT-LEARN*

Método	**Aplicação**	**Número de grupos**	**Tamanho dos grupos**
K-Means	propósito geral, geometrias simples	poucos	semelhantes
Propagação de afinidade	geometrias complexas	muitos	diferentes
Mean-shift	geometrias complexas	muitos	diferentes
Espectral	geometrias complexas	poucos	semelhantes
Hierárquico (Ward)	grupos hierárquicos, geometrias complexas	muitos	–
Hierárquico (Aglomerativo)	grupos hierárquicos, geometrias complexas	muitos	–
DBSCAN	geometrias complexas	–	diferentes
OPTICS	baseado em densidade	–	diferentes
BIRCH	volume de dados, remoção de valores atípicos	–	–
Modelos de mistura Gaussiana	baseado em densidade	–	–

FONTE: Adaptado de (AGGARWAL, 2013; PEDREGOSA et al., 2011)

2.2.3 Método de agrupamento k-médias (*K-means*)

O método *K-means* pode ser descrito por meio do seguinte problema de otimização: dado um conjunto de dados, identificar K grupos de elementos de modo que cada grupo tenha mínima variância em relação aos seus respectivos representantes, também denominados protótipos ou centroides (REDDY; VINZAMURI, 2018). Nesta formulação, o método foi descrito em 1957 por Stuart P. Lloyd em manuscrito sobre o problema de particionar um conjunto de sinais de modo que a definição das k partições

e respectivos *quantas* minimizem a média quadrada do ruído entre o sinal emitido e o sinal recebido. Por razões computacionais o trabalho foi publicado somente em 1982 (LLOYD, 1982). A formalização do método na forma de aplicação à resolução de problemas de classificação multivariada, entretanto, é sistematizada muito antes por MacQueen et al. (1967) no artigo *Some methods for classification and analysis of multivariate observations*.

A identificação de grupos ótimos é um problema de complexidade computacional não polinomial difícil (*NP-hard*) , ou seja, os cálculos usados em sua resolução cresce exponencialmente em relação ao número de dados[3] (REDDY; VINZAMURI, 2018). Neste contexto, *K-means* é considerada uma heurística gulosa indicada para a identificação de grupos coesos e bem separados e cujo resultado depende de uma estimativa conveniente dos centroides iniciais (JAIN, 2010). Embora a convergência local seja garantida, este efeito exige que, na prática, o método seja aplicado diversas vezes com diferentes conjuntos de centroides iniciais, dentre os quais será selecionado aquele que atende aos critérios de otimização associados à qualidade dos grupos formados no agrupamento.

As propriedades esperadas para agrupamentos ótimos estão associadas à coesão e à separação dos grupos. A coesão ou compacidade, também denominada inércia (*within-cluster sum of squares – wcss*), mede a homogeneidade entre os elementos de um mesmo grupo. A separação ou isolamento (*between-cluster sum of squares – bcss*) estão associados a medida de heterogeneidade entre os grupos (EVERITT, 2011). Além de auxiliar na escolha dos centroides iniciais, estas medidas são aplicadas para determinar o número K de grupos para inicialização do *K-means*, o qual, geralmente, não é conhecido previamente. Da mesma forma que no caso das escolha dos centroides iniciais, o número de grupos que melhor se ajusta aos critérios de otimização do *K-means* pode ser determinado configurando diversos valores para K e avaliando a qualidade do agrupamento em cada caso. Neste sentido, estas propriedades também são utilizadas na definição de medidas de validação da análise de agrupamento. A seguir serão discutidos aspectos sobre a estrutura e planejamento da análise de agrupamento onde aplicação do método *K-means* é apenas uma etapa.

2.2.4 Análise de agrupamentos

A análise de agrupamentos através da aplicação do método *K-means*, assim como em outros métodos de agrupamento baseados na otimização de uma função objetivo, envolve as seguintes etapas (HALKIDI et al., 2001):

- Modelagem dos padrões de dados: os dados brutos nem sempre permitem a

[3] O número de agrupamentos de n elementos em K grupos é aproximadamente igual a $\frac{K^n}{K!}$ para $n \gg K$ (JAIN; DUBES, 1988)

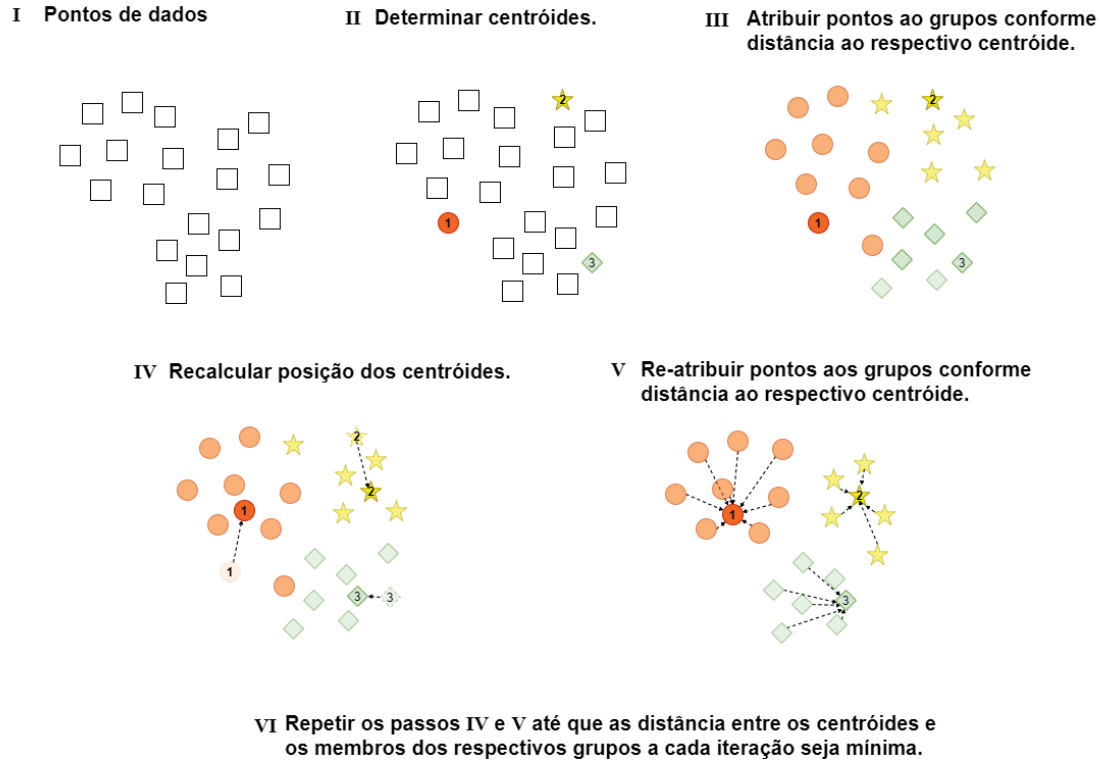
sua comparação direta exigindo o seu pré-processamento como, por exemplo, por meio da seleção, redução ou geração de variáveis, o tratamento de valores faltantes ou atípicos e a transformação do espaço de dados. Ainda nesta etapa pode ser feita a análise de tendência de agrupamentos dos dados. Em especial, espera-se verificar, se existem grupos não aleatórios no conjunto de dados. Para tanto, por exemplo, podem ser usadas abordagens quantitativas baseadas na comparação da média das estatísticas de Hopkins para amostragens sobre o conjunto de dados e uma distribuição aleatória com base no espaço de dados (TAN et al., 2009). Alternativamente, esta verificação pode ser realizada por meio da visualização e identificação de padrões similares no espaço transformado como no caso da utilização da Análise de Componentes Principais para auxiliar a análise de agrupamento.

- Processo de agrupamento
  - Definição da medida de similaridade: a comparação entre os padrões de dados é realizada por meio do cálculo de um índice de proximidade ou distância. Em qualquer caso, dois padrões mais similares correspondem a valores de proximidade maiores ou, de forma equivalente, a valores de distâncias menores;
  - Critério de otimização do agrupamento: uma função objetivo deve atender o critério de otimização da coesão dos grupos garantindo a atribuição de rótulos de classe aos elementos que estejam mais próximos dos centroides dos grupos.
- Validação do agrupamento: em geral, algoritmos de agrupamento definem grupos os quais não são conhecidos previamente. Assim os resultados podem revelar grupos mesmo quando estes não existem ou não podem ser interpretados adequadamente em relação ao contexto de aplicação do método. Na prática, um agrupamento ótimo é selecionado dentre outros resultados da aplicação do método com variações na configuração de seus parâmetros, como o número de grupos ou os centroides iniciais. Além disso, variações sobre o mesmo método podem incluir a transformação do espaço de dados e a definição de medidas de similaridade alternativas.
- Interpretação dos resultados: a última etapa da análise é a interpretação do agrupamento em termos do contexto da aplicação.

O método *K-means*, descrito no **Algoritmo 1**, tem como entrada a matriz de similaridade, ou a matriz de padrões dos dados, e o número K de grupos sobre o qual o método deve tentar agrupar os padrões de dados. O processamento dos dados pode

então ser dividido em duas etapas: a etapa inicial e etapa iterativa. Na etapa inicial, K centroides são selecionados aleatoriamente no espaço de dados como representantes de cada um dos grupos esperados. A seguir, a etapa iterativa do método envolve dois passos: (1) atribuir cada padrão de dados ao grupo cujos centroides sejam mais próximos, e (2) atualizar as posições dos centroides em relação aos membros dos respectivos grupos, como ilustrado na FIGURA 19.

FIGURA 19 – ETAPAS DE CRIAÇÃO DE UM AGRUPAMENTO *K-MEANS*



FONTE: Baseado em Jain e Dubes (1988)

O algorítimo básico do *k-means* pode ser descrito como a seguir

Algoritmo 1 *K-means* básico

Entrada: Matriz de similaridade ou matriz de padrões do conjunto de dados, número K de grupos desejados;
Passo 1: Determinar aleatoriamente K centroides;
Passo 2: Calcular as distâncias dos pontos de dados aos centroides e atribuir o rótulo de classe do grupo cujo respectivo centroide seja o mais próximo;
Passo 3: Atualizar a posição dos centroides em relação aos elementos dos respectivos grupos
Passo 4: Repetir os Passos 2 e 3 até que a posição dos centroides não se altere segundo uma tolerância pré-determinada ou um número de iterações máximo.
Saída: K grupos e centroides dos pontos de dados com os respectivos rótulos de classe.

Jain e Dubes (1988) incluem mais um passo neste algoritmo básico no qual o agrupamento final deve ser ajustado por meio da soma, divisão e exclusão de alguns dos grupos obtidos. Uma prática experimental comum na análise de grupos é a comparação de resultados de diferentes tipos de métodos ou diferentes configurações de parâmetros de um mesmo método utilizando-se medidas de qualidade para os agrupamentos formados. Os resultados da aplicação dos métodos de agrupamentos podem ser manipulados através da análise de meta-agrupamentos com objetivo de identificar um agrupamento consensual (*consensus clustering*) por meio da combinação de agrupamentos (*clustering ensemble*), discutidos com detalhes mais adiante. A seguir é apresentada a formulação matemática e os resultados associados ao *K-means*.

Definição 2.2.1 (Agrupamento de dados) *Dado um conjunto $\mathcal{X}$, cujos elementos $\mathcal{X}_i$, com $i = 1, \cdots, N$ são denominados pontos, instâncias ou padrões. Cada padrão $\mathcal{X}_i$ é determinado por m atributos ou variáveis, assim, $\mathcal{X}_i = \{\mathcal{X}_{i,1}, \cdots, \mathcal{X}_{i,m}\}$. Um agrupamento $\mathcal{C}$ de $\mathcal{X}$ é definido por K partições $\mathcal{C}_k \subset \mathcal{X}$, com $K \geq 2$ tal que $\mathcal{X}$ resulte da união de partições disjuntas não vazias, ou seja,*

$$\mathcal{X} = \bigcup \mathcal{C}_k, \tag{2.1}$$

com $\mathcal{C}_k \neq \emptyset$, e

$$\bigcap_{r \neq s} (\mathcal{C}_r, \mathcal{C}_s) = \emptyset, \tag{2.2}$$

$\forall\ k,\ r,\ s \in \{1, \cdots, K\}$.

Um grupo $\mathcal{C}_k = \{\mathcal{X}_1^{(k)}, \cdots, \mathcal{X}_{n_k}^{(k)}\}$ é definido por um número n_k de elementos $\mathcal{X}_i = \{\mathcal{X}_{i,1}, \cdots, \mathcal{X}_{i,m}\}$ de $\mathcal{X}$. Um grupo $\mathcal{C}_k$ ótimo, representado por seu centroide, é um ponto c^k de $\mathcal{X}^* \supseteq \mathcal{X}$, espaço das soluções que minimizam a função erro (*Sum Squared Error – SSE*).

Um agrupamento $\mathcal{C}$ ótimo fica definido pelo conjunto de centroides $c^{(k)} \in \mathcal{X}^*$, com $k = 1, \cdots, K$ que minimizam a $SSE(\mathcal{C}) = \sum_{k=1}^{K} SSE(\mathcal{C}_k)$. A função erro $SSE(\mathcal{C})$ para o agrupamento $\mathcal{C}$ é dada pela Equação 2.3 (TAN et al., 2009).

$$SSE(\mathcal{C}) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\mathcal{X}_i^{(k)} - c^{(k)})^2 \tag{2.3}$$

Derivando a equação 2.3 em relação à $c^{(k)}$ e igualando a zero, $\frac{\partial}{\partial c^{(k)}} SSE(\mathcal{C}) = 0$, a solução do problema de minimização da função SSE é obtida por meio da resolução da equação diferencial parcial Equação 2.4 em relação à $c^{(k)}$:

$$\frac{\partial}{\partial c^{(k)}} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\mathcal{X}_i^{(k)} - c^{(k)})^2 = 0 \tag{2.4}$$

$$\sum_{k=1}^{K}\sum_{i=1}^{n_k}\frac{\partial}{\partial c^{(k)}}(\mathcal{X}_i^{(k)}-c^{(k)})^2=0$$

$$\sum_{k=1}^{K}\sum_{i=1}^{n_k}\frac{\partial}{\partial c^{(k)}}(\mathcal{X}_i^{(k)})^2-2(\mathcal{X}_i^{(k)})(c^{(k)})+(c^{(k)})^2=0$$

$$\sum_{k=1}^{K}\sum_{i=1}^{n_k}\frac{\partial}{\partial c^{(k)}}(\mathcal{X}_i^{(k)})^2-\frac{\partial}{\partial c^{(k)}}2(\mathcal{X}_i^{(k)})(c^{(k)})+\frac{\partial}{\partial c^{(k)}}(c^{(k)})^2=0$$

$$\sum_{k=1}^{K}\sum_{i=1}^{n_k}-2\mathcal{X}_i^{(k)}+2c^{(k)}=0$$

$$\sum_{k=1}^{K}\sum_{i=1}^{n_k}2(c^{(k)}-\mathcal{X}_i^{(k)})=0$$

$$\sum_{k=1}^{K}\sum_{i=1}^{n_k}(c^{(k)}-\mathcal{X}_i^{(k)})=0$$

$$\sum_{k=1}^{K}\sum_{i=1}^{n_k}c^{(k)}=\sum_{k=1}^{K}\sum_{i=1}^{n_k}\mathcal{X}_i^{(k)}$$

Abrindo o somatório em i no primeiro membro da equação anterior, segue

$$\sum_{k=1}^{K}(\underbrace{c^{(k)}+\cdots+c^{(k)}}_{n_k\text{vezes}})=\sum_{k=1}^{K}\sum_{i=1}^{n_k}\mathcal{X}_i^{(k)},$$

ou seja,

$$\sum_{k=1}^{K}n_kc^{(k)}=\sum_{k=1}^{K}\sum_{i=1}^{n_k}\mathcal{X}_i^{(k)}.$$

Para o k-ésimo grupo, a equação anterior reduz-se a

$$n_kc^{(k)}=\sum_{i=1}^{n_k}\mathcal{X}_i^{(k)}.$$

Logo o valor de $c^{(k)}$ é obtido por meio da Equação 2.5

$$c^{(k)}=\frac{\sum_{i=1}^{n_k}\mathcal{X}_i^{(k)}}{n_k}, \tag{2.5}$$

Conforme pode ser constatado pela Equação 2.5, no método *K-means* os centroides do grupos são os elementos $c^{(k)}$, $k=1,\cdots,K$, cujos atributos correspondem a média dos atributos dos elementos da mesma classe. A adoção de outros representantes dos grupos a partir da mediana e a moda, definem-se as versões *K-medians* e o *K-modes*, respectivamente (JAIN; DUBES, 1988).

2.2.5 Análise de Componentes Principais

Como destacado anteriormente, uma das desvantagens do método *K-means* é a hipótese da existência da estrutura de grupos convexos e isotrópicos com formatos elipsoidais que são melhor ajustados em espaços de dados definidos a partir de poucas varáveis, ou seja, com baixa dimensão. Contrariamente, em espaços de alta dimensão, as medidas de proximidade se degradam ficando próximas de um, enquanto as distâncias tendem a se anular. Assim, os padrões de dados tendem a ser mais similares dificultando a formação dos grupos semanticamente significativos. O comportamento de distâncias baseadas nas normas ℓ foi explorado em artigo de Aggarwal et al. (2001), os quais destacam a análise de diferentes ordens de proximidade definidas por distâncias distintas. O estudo da influência da noção de distância no agrupamento de padrões será um dos pontos centrais desta tese e será discutido em outra seção sob a perspectiva da similaridade entre grafos. A seguir é apresentada a técnica de redução da dimensionalidade e visualização de grupos utilizada em conjunto com o *K-means*.

Uma solução para o problema da dimensionalidade (*curse of dimensinality*) é a utilização da análise de componentes principais (*Principal Components Analysis – PCA*) na fase de pré-processamento dos dados. A *PCA* corresponde a uma transformação que projeta os dados originais em um espaço de dimensão reduzida cujas direções incorporam a maior parte da variabilidade dos dados. Às direções de maior variabilidade correspondem os autovetores associados aos maiores autovalores calculados por meio da decomposição espectral da matriz que guarda informações acerca da variabilidade dos dados. As componentes da matriz podem ser calculadas por meio de medidas de variabilidade como o desvio, a correlação, a variância, a distância ou a similaridade (ELMORE; RICHMAN, 2001; DING; HE, 2004; XU et al., 2014). Conforme Jain e Dubes (1988), a *PCA* deve ser utilizada em conjunto com as técnicas de agrupamentos para análise gráfica de agrupamentos por meio da projeção das duas ou três primeiras componentes principais permitindo a visualização dos grupos em gráficos de duas ou três dimensões.

A proximidade entre o método *K-means* e a *PCA* pode ser explorada sob diversos aspectos. Por exemplo, seguindo Elmore e Richman (2001), a correlação é adotada como medida de similaridade e surge naturalmente da generalização da distância euclidiana utilizada na definição da função SSE. De fato, a Equação 2.3

$$SSE(\mathcal{C}) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\mathcal{X}_i^{(k)} - c^{(k)})^2$$

pode ser escrita em termos da distância de Mahalanobis (JAIN; DUBES, 1988) como na Equação 2.6

$$SSE(\mathcal{C}) = d_{\textit{Mahalanobis}} = \sqrt{(\mathcal{X}_i^{(k)} - c^{(k)})\mathcal{R}^T(\mathcal{X}_i^{(k)} - c^{(k)})} \tag{2.6}$$

onde, para o caso especial da distância euclidiana $R = I$ é a matriz identidade.

Dado o conjunto de n padrões $\mathcal{X}_i = \mathcal{X}_{i,1}, \cdots, \mathcal{X}_{i,m}$, cada qual definido por m atributos, a matriz $[\mathcal{X}]_{n \times m}$ é definida na Equação 2.7

$$[\mathcal{X}] = \begin{bmatrix} \mathcal{X}_{1,1} & \mathcal{X}_{1,2} & \cdots & \mathcal{X}_{1,m} \\ \mathcal{X}_{2,1} & \mathcal{X}_{2,2} & \cdots & \mathcal{X}_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{X}_{n,1} & \mathcal{X}_{n,2} & \cdots & \mathcal{X}_{n,m} \end{bmatrix} \tag{2.7}$$

Uma medida de similaridade (ou dissimilaridade) que indique a proximidade entre dois padrões $\mathcal{X}_i$ e $\mathcal{X}_j$ pode ser definida a partir de sua variabilidade em $\mathcal{X}$. Para tanto, seja a média μ, como definido na Equação 2.5,

$$\mu(\mathcal{X}) = \frac{\sum_{i=1}^{n} \mathcal{X}_i}{n}. \tag{2.8}$$

De modo similar define-se a variância σ que captura o desvio de $\mathcal{X}_i$ em relação à $\mu^{(\mathcal{X})}$

$$\sigma(\mathcal{X}) = \frac{\sum_{i=1}^{n} \mathcal{X}_i - \mu(\mathcal{X})}{n}. \tag{2.9}$$

Seja $\mathcal{X}^*$ o conjunto de padrões normalizados. A matriz $\mathcal{R}_\mathcal{X}$, como na Equação 2.10,

$$\mathcal{R}_\mathcal{X} = \frac{1}{n}[\mathcal{X}^*]^T[\mathcal{X}^*] \tag{2.10}$$

corresponde a matriz de correlação se $\mathcal{X}^* = [\mu_j(\mathcal{X})]$, e a matriz de covariância se $\mathcal{X}^* = \sigma_j(\mathcal{X})$, com $j = 1, \cdots, n$.

As direções de maior variabilidade no espaço de padrões coincidem com as direções dos n autovetores v_i de $\mathcal{R}_\mathcal{X}$ associados aos maiores autovalores λ_i, respectivamente. A matriz $\mathcal{R}_\mathcal{X}$ é simétrica, definida positiva, admitindo, portanto, a sua decomposição em valores singulares (DING; HE, 2004).

$$[\mathcal{R}_\mathcal{X}] = \mathcal{V}^T \Lambda \mathcal{V} \tag{2.11}$$

Na Equação 2.11, as colunas transpostas $v_i = [v_{i,1}, \cdots, v_{i,m}]$ de $\mathcal{V}$ são os autovetores normalizados de $\mathcal{R}_\mathcal{X}$. A matriz Λ é a matriz diagonal cujas entradas $\Lambda_{(i,i)}$ são os autovetores λ_i em ordem decrescente, $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_m \leq 0$. Os autovetores λ_i são calculados a partir da sua definição na Equação 2.12

$$\mathcal{R}_{\mathcal{X}} v_i = \lambda_i v_i. \tag{2.12}$$

A qual pode ser reescrita como

$$\mathcal{R}_{\mathcal{X}} v - \lambda v = [0]. \tag{2.13}$$

Considerando a igualdade $\lambda v = \lambda I v$, com I a matriz identidade, obtemos

$$[\mathcal{R}_{\mathcal{X}} - \lambda I] v = [0]. \tag{2.14}$$

Por definição, na Equação 2.14, o vetor v pertence ao núcleo da transformação linear $\mathcal{R}_{\mathcal{X}} - \lambda I$. Como $v_i \neq 0$, $\forall i$, a transformação $\mathcal{R} - \lambda I$ não é injetiva, portanto não é bijetiva, o que implica que $[\mathcal{R}_{\mathcal{X}} - \lambda I]$ não é inversível. Dessa forma o determinante $|\mathcal{R}_{\mathcal{X}} - \lambda I|$ é nulo, dando origem a expressão da Equação 2.15.

$$|\mathcal{R}_{\mathcal{X}} - \lambda I| = 0. \tag{2.15}$$

A Equação 2.15 é denominada **polinômio característico** de $[\mathcal{R}_{\mathcal{X}}]$ e suas raízes reais são os autovalores que satisfazem a Equação 2.12. Os autovetores v_i são obtidos da resolução dos sistemas lineares da equação matricial 2.13 para cada um dos autovalores já calculados.

$$[\mathcal{R}_{\mathcal{X}} - \lambda I][v] = [0]$$

Os autovetores assim obtidos são normalizados e cumprem a condição de serem ortogonais entre si, dada pela Equação 2.16

$$[v_i]^T [v_j] = \begin{cases} 0 & , se \;\; i \neq j; \\ 1 & , se \;\; i = j. \end{cases} \tag{2.16}$$

Quando aplicada sobre a matriz dos padrões normalizados $\mathfrak{X}$, a matriz dos autovetores unitários $\mathcal{V}$ de $\mathcal{R}_{\mathcal{X}}$ corresponde a uma rotação do espaço dos padrões na direção dos autovetores, cuja variância é medida pelos respectivos autovalores associados. Dessa forma, na Equação 2.17, as colunas y_i de $\mathcal{Y}$ formam uma base de dimensão m para o espaço projetado dos padrões.

$$\mathcal{Y} = \mathfrak{X} \mathcal{V}^T \tag{2.17}$$

onde $\mathcal{Y}^T = [y_1, \cdots, y_n]$, logo y_i pode ser expresso como

$$y_i = \mathcal{V} \mathcal{X}_i = \sum_{i=1}^{n} \sum_{j=1}^{m} \lambda_i \mathcal{X}_{i,j} \tag{2.18}$$

é denominada uma componente principal, trata-se de uma nova variável definida a partir da combinação linear das variáreis originais. As componentes principais guardam a informação sobre a variabilidade dos dados por meio dos valores assumidos por seus autovetores. Dessa forma é possível determinar a porcentagem de variação que cada componente representa em relação à variação total dos dados por meio da Equação 2.19. Dados os autovalores λ_i, proporção de variabilidade $p(\lambda_i)$ é definida como

$$p(\lambda_i) = \frac{\lambda_i}{\sum \lambda_i} \tag{2.19}$$

A variabilidade neste novo espaço Y é definida por meio de sua matriz de covariância $\mathcal{R}_\mathcal{Y}$, como na Equação 2.10, ou seja,

$$\mathcal{R}_\mathcal{Y} = \frac{1}{n}[\mathcal{Y}]^T[\mathcal{Y}]. \tag{2.20}$$

Tomando $\mathcal{Y} = \mathfrak{X}\mathcal{V}^T$, da Equação 2.17, e substituindo na Equação 2.20, obtém-se

$$\mathcal{R}_\mathcal{Y} = \frac{1}{n}[\mathfrak{X}\mathcal{V}^T]^T[\mathfrak{X}\mathcal{V}^T] \tag{2.21}$$

$$\mathcal{R}_\mathcal{Y} = \frac{1}{n}[\mathcal{V}\mathfrak{X}^T][[\mathfrak{X}\mathcal{V}^T], \tag{2.22}$$

Mas, $\frac{1}{n}[\mathfrak{X}^T][\mathfrak{X}] = \mathcal{R}_\mathcal{X}$, logo

$$\mathcal{R}_\mathcal{Y} = \mathcal{V}[\mathcal{R}_\mathcal{X}]\mathcal{V}^T, \tag{2.23}$$

Da Equação 2.11, $[\mathcal{R}_\mathcal{X}] = \mathcal{V}^T\Lambda\mathcal{V}$, segue que a Equação 2.23, torna-se

$$\mathcal{R}_\mathcal{Y} = \mathcal{V}\mathcal{V}^T\Lambda\mathcal{V}\mathcal{V}^T, \tag{2.24}$$

Como o produto $\mathcal{V}\mathcal{V}^T = I$,

$$\mathcal{R}_\mathcal{Y} = \Lambda, \tag{2.25}$$

de onde se conclui que a matriz Λ dos autovetores de $\mathcal{R}_\mathcal{X}$ é a matriz de covariância de $\mathcal{Y}$.

A redução da dimensão do espaço de dados originais para a sua representação, por exemplo, em duas dimensões, é obtida por meio do truncamento da representação de y_i tomando $i = 1, 2$, assim,

$$y_i = \lambda_1\mathcal{X}_{i,1} + \lambda_2\mathcal{X}_{i,2}. \tag{2.26}$$

Esta representação utiliza as duas dimensões de maior variância mantendo a distância relativa entre os padrões de dados. Esta propriedade faz com que a análise

de componentes principais seja recomendada como técnica auxiliar para avaliar a qualidade dos agrupamentos gerados pelo *K-means*. Uma vez que sejam obtidos os rótulos de classe dos padrões como resultado do método *K-means*, estes são atribuídos ao resultado da *PCA* para a identificação visual do grupos e respectivos membros (JAIN; DUBES, 1988).

2.2.6 Validação de agrupamentos

Como lembra Tan et al. (2009), quando aplicados, os métodos de agrupamentos podem encontrar grupos mesmo que eles não existam ou que não correspondam a uma estrutura **natural** de grupos[4]. Geralmente, esta estrutura natural, também denominada estrutura verdadeira, não é conhecida exigindo que os resultados dos métodos de agrupamentos sejam validados. A validação dos resultados da análise de agrupamentos compreende uma etapa em que os grupos formados são avaliados. Esta avaliação é realizada mediante o cálculo e comparação de índices e medidas cuja aplicação deve ser estudada caso a caso podendo variar conforme o tipo do método (hierárquico, por partições, baseado em densidade, etc) ou a estrutura geométrica dos grupos (globulares, difusos, concêntricos, etc) (JAIN; DUBES, 1988). A validação de agrupamentos é frequentemente classificada em dois tipos conforme os índices e medidas utilizados na avaliação do agrupamento dependam de outras informações além daquelas disponíveis nos dados, a **validação externa**, ou caso sejam calculados apenas com as informações contidas nos dados, a **validação interna**.

2.2.7 Validação externa

A validação externa depende do conhecimento prévio sobre os rótulos de classe dos dados na forma de partições de referência. Este conhecimento pode ser baseado na estrutura dos dados ou na opinião de especialistas no contexto de aplicação do método de agrupamento. Os índices e as medidas associados a este tipo de validação devem ser sensíveis a similaridade entre os grupos formados a partir da aplicação de métodos de agrupamentos e as classes em agrupamentos de referência conhecidos *a*

[4] Qualitativos como "naturais"e "verdadeiros"são comumente utilizados na literatura, mas pouco explicados. Embora essa discussão não seja feita aqui, uma visão epistemológica sobre a própria ideia de grupos, grupos naturais ou verdadeiros pode ser acompanhada em Hennig (2015). Um ponto importante neste debate afeta profundamente a atitude do analista em relação aos resultados da análise de agrupamento. Em suma, não se deve imputar ao método a tarefa de encontrar os tais grupos naturais (pois eles não existem). Na verdade, invertendo a perspectiva, são os métodos que podem conformar tipos e estrutura de agrupamentos sobre os dados por meio de transformações do espaço de dados. Um exemplo sobre o impacto da aplicação do método sobre o conjuntos de dados originais é estudada no caso do *K-means* e está associada ao efeito de uniformização da distribuição de dados originais nos grupos *k-means* (XIONG et al., 2008). Se por um lado, este efeito reduz a distância entre os pontos de dados podendo acarretar em grupos mal definidos, principalmente se os centroides de grupos diferentes estiverem relativamente próximos. Por outro lado, é justamente por conta deste efeito que, sob certas condições, *K-means* tem a capacidade formar grupos coesos e bem separados.

priori. Frequentemente, a validação externa é utilizada para comparar resultados de métodos de agrupamentos e outras medidas de validação. Tan et al. (2009) denomina a validação externa como supervisionada e divide as medidas utilizadas neste tipo de validação em dois tipos:

- as medidas orientadas a classificação com origem no campo de estudo de recuperação da informação e que são utilizadas para avaliação de modelos de classificação (TAN et al., 2009). Dois exemplos, a Entropia e a Medida F são apresentados a seguir.

  – A **Entropia** mede o grau de incerteza com o qual os rótulos de classes estão atribuídos aos membros dos grupos. Embora seja utilizada na validação externa, a Entropia também pode ser aplicada na validação interna (GAO; YANG, 2018). A Entropia E_k de um grupo k é calculada pela Equação 2.27

$$E_k = -\sum_{k=1}^{K} p_{kj} log_2 p_{kj}, \tag{2.27}$$

  onde $j = 1, \cdots, K$ e p_{kj}, corresponde a probabilidade de que um membro do grupo k tenha um rótulo de classe j, dada por

$$p_{kj} = \frac{n_{kj}}{n_k}, \tag{2.28}$$

  onde n_{kj} é o número de membros do grupo k com rótulo de classe j e n_k é o número de membros do grupo k. A Entropia E do agrupamento é definida na Equação 2.29

$$E = \sum_{k=1}^{K} \frac{n_k}{n} E_k. \tag{2.29}$$

  Quanto maior for o valor absoluto da Entropia, maior o grau de dissemelhança entre os grupos ou agrupamentos avaliados. A Entropia tem valor igual a 0 quando dois grupos ou agrupamentos, conforme o caso, tem a mesma configuração de classes. A probabilidade p da Equação 2.28 é utilizada na definição da Medida F, apresentada a seguir.

  – A **Medida F** é definida a partir de duas medidas: a **Precisão** e a **Lembrança**. A primeira delas, a **Precisão** P corresponde a probabilidade p definida na Equação 2.28, assim

$$P(k, j) = \frac{n_{kj}}{n_k}. \tag{2.30}$$

  A segunda medida, a **Lembrança** L expressa na Equação 2.31

$$L(k, j) = \frac{n_{kj}}{n_j}, \tag{2.31}$$

corresponde a fração do grupo k que contém membros cujo rótulo de classe é j. A medida F, $F(k, j)$ é calculada por meio da Equação 2.32

$$F = 2\frac{P(i,j) * L(i,j)}{P(i,j) + L(i,j)}. \tag{2.32}$$

A **Medida F** assume valores no intervalo $[0, 1]$, com 1 indicando alta Precisão e Lembrança, e 0, o caso contrário.

- as medidas orientadas a semelhança são baseadas na comparação das quantidades de membros compartilhadas entre o agrupamento resultado da aplicação do método e um agrupamento de referência. De acordo com Tan et al. (2009), os índices, conhecidos como coeficientes de Rand e de Jaccard são os mais utilizados. Silva et al. (2016) adicionam a esta lista o coeficiente de Folkes e Mallows. Estes três coeficientes são calculados tomando como base a comparação dos grupos de um agrupamento de dados com as classes de partições de referência.

  Dado o conjunto de dados $X = \{X_1, \cdots, X_N\}$, sejam um agrupamento **C** com K grupos, **C** $= \{C_i\}$, $i = 1, \cdots, K$, e **P** $= \{P_j\}$, $j = 1, \cdots, L$ as partições de referências. Quatro quantidades estão associadas a relação entre os rótulos dos pontos de dados quando estes são analisados como membros de um grupo C_i e uma classe P_j. Seguindo Silva et al. (2016), seja uma função ϕ tal que $\phi(X_p)$ rotula um ponto de dado X_p com o rótulo de seu grupo C_i para agrupamentos, ou classe P_j no caso de partições. Tomando dois pontos de dados $X_p, X_q \in X$, $p \neq q$, com base nos seus rótulos, podemos definir a seguinte função de contingência definida na Equação 2.33

$$F(X_p, X_q) = \begin{cases} 1, \ se\ X_p\ e\ X_q\ tem\ o\ mesmo\ rótulo \\ 0, \ caso\ contrario. \end{cases} \tag{2.33}$$

  A aplicação desta função simultaneamente a grupos e classes podemos identificar os seguintes resultados possíveis denotados a seguir através de sua codificação binária

  11 : X_p e X_q pertencem a mesmo grupo C_i em **C** e mesma classe P_j em **P**.

  10 : X_p e X_q pertencem a mesmo grupo C_i em **C** e classes P_j diferentes em **P**.

  01 : X_p e X_q pertencem a grupos C_i diferentes em **C** e mesma classe P_j em **P**.

  00 : X_p e X_q pertencem a grupos diferentes em **C** e classes diferentes em **P**.

  Conforme (TAN et al., 2009), denotando por f a frequência dos eventos supra citados, define-se as seguintes quatro quantidades

  f_{11} = número de pares de pontos pertencentes ao evento **11**

f_{10} = número de pares de pontos pertencentes ao evento **10**

f_{01} = número de pares de pontos pertencentes ao evento **01**

f_{00} = número de pares de pontos pertencentes ao evento **00**

a partir das quais são expressos os coeficientes que indicam o grau de similaridade entre o agrupamento **C** e as partições de referência **P**.

- **Rand** (I_R)

$$I_R = \frac{f_{11}}{f_{11} + f_{01} + f_{10} + f_{00}} \tag{2.34}$$

- **Jaccard** (I_J)

$$I_J = \frac{f_{11}}{f_{01} + f_{10} + f_{00}} \tag{2.35}$$

- **Folkes e Mallows** (I_{FM})

$$I_{FM} = \sqrt{\frac{f_{11}^2}{(f_{11} + f_{01})(f_{11} + f_{10})}} \tag{2.36}$$

Assim, como a medida F, os coeficientes de Rand, Jaccard e Folkes -Mallows, são definidos no intervalo $[0, 1]$ com valores próximos de 1 indicando grupos ou agrupamentos similares, e 0 no caso contrário.

2.2.8 Validação interna

A validação interna, também denominada não supervisionada é realizada utilizando-se de critérios que se baseiam apenas nos dados disponíveis. As propriedades de coesão e separação dos grupos, apresentadas na seção 2.2.4 na qual o método *K-means* foi introduzido, são utilizadas na definição de diversas medidas de validação interna.

Dado um conjunto de dados $\mathcal{X} = \{\mathcal{X}_1, \cdots, \mathcal{X}_n\}$ e um agrupamento $C = C_1, \cdots, C_K$ dos pontos de $\mathcal{X}$ com K classes e centroides $c^{(k)}$, tal como definido nas equações 2.1 e 2.2. Seja a medida de proximidade dada pela distância euclidiana como definido na Equação 2.3, a coesão mede a homogeneidade entre os elementos de um mesmo grupo de um grupo C_k é dada pela Equação 2.37

$$Coesão(C_k) = \sum_{i=1}^{n_k} (\mathcal{X}_i^{(k)} - c^{(k)})^2, \tag{2.37}$$

onde $X_i^{(k)}$ são os elementos $X_i \in C_k$ e n_k é número de elementos no grupo C_k.

A separação é uma medida de heterogeneidade sendo definida para dois grupos C_p e C_q por meio da distância entre os respectivos centroides $c^{(p)}$ e $c^{(q)}$ na

Equação 2.38

$$Sep(C_p, C_q) = (c^{(p)} - c^{(q)})^2. \tag{2.38}$$

A separação de um grupo C_k é calculada em relação aos centroides $c^{(k)}$ de cada grupo como na Equação 2.39

$$Sep(C_k) = (c^{(k)} - c)^2 * \frac{n-k}{k-1}, \tag{2.39}$$

onde c é centroide do agrupamento e $\frac{n-k}{k-1}$ é um fator de redução que diminui a influencia do incremento no número K de grupos em relação ao valor de Sep. A separação total do agrupamento e definida diretamente por meio da soma da separação de cada grupo como na Equação 2.40

$$Sep(C) = \sum_{k=1}^{K} Sep(C_k). \tag{2.40}$$

Outras funções de proximidade podem medir a coesão e separação de grupos e agrupamentos como, por exemplo, na definição dos índices de validação **Davies-Bouldin** (Equação 2.41), **Calinski–Harabasz** (Equação 2.43) e **Silhueta** (Equação 2.46).

- O índice **Davies-Bouldin** (DB) mede a similaridade entre dois grupos com base na razão entre a soma da coesão sobre a separação dos grupos (SILVA et al., 2016).

  $$DB(C_p, C_q) = \frac{(Coesão(C_p) + Coesão(C_q))}{Sep(C_p, C_q)} \tag{2.41}$$

  O índice do agrupamento pode ser calculado por meio da Equação 2.42,

  $$DB(\mathbf{C}) = \frac{1}{K} \sum_{k=1}^{K} Max_{p \neq q}(DB(C_p, C_q)), \tag{2.42}$$

  com $p, q = 1, \cdots, K$. O índice DB é estritamente positivo e valores próximos de 0 indicam menor dispersão intra grupos e maior separação entre os grupos o que significa um agrupamento de melhor qualidade (BAARSCH; CELEBI, 2012).

- O índice **Calinski-Harabasz** (CH) é definido como a razão entre a coesão e a separação. Dessa forma, valores maiores de CH indicam grupos com melhor qualidade.

  $$CH(C) = \frac{Sep(C)}{Coesão(C)} \tag{2.43}$$

- O coeficiente de **Silhueta** (SIL) é definido por meio da razão entre a diferença da coesão e a separação. Entretanto a coesão é calculada por meio da diferença $b(X_i) - a(X_i)$ das medidas auxiliares $a(X_i)$ e $b(X_i)$ aplicadas aos pontos de dados

X_i dadas pelas equações 2.44, onde $a(i)$ é distância média entre X_i e os outros membros do grupo C_k,

$$a(i) = \frac{1}{n_k} \sum_{j=1}^{n_k} (X_i^{(k)} - X_j^{(k)})^2, \tag{2.44}$$

e $b(i)$ definido na Equação 2.45 é distância média de X_i aos X_ℓ pertencentes ao grupo C_ℓ mais próximo de do grupo C_k de X_i,

$$b(i) = min_{C_\ell \subset C} \{ \frac{1}{n_\ell} \sum_{j=1}^{n_k} (X_i^{(k)} - X_j^{(k)})^2 \}, \tag{2.45}$$

com $C_\ell \neq C_k$. O coeficiente de silhueta de X_i é calculado por meio da Equação 2.46,

$$SIL(X_i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}. \tag{2.46}$$

Os valores de silhueta para grupos são calculados por meio da média dos valores de silhueta dos membros dos respectivos grupos como na 2.47

$$SIL(C_k) = \frac{1}{n_k} \sum_{i=1}^{n_k} SIL(X_i) \tag{2.47}$$

O coeficiente de silhueta está definido no intervalo $[-1, 1]$ onde valores próximos de 1 indicam melhores grupos ou aderência de um determinado membro ao seu respectivo grupo (AGGARWAL, 2013), e valores próximos de -1, o caso contrário. Grupos e agrupamentos com valores de silhueta próximos de zero indicam que os dados se distribuem uniformemente no grupo ou agrupamento, respectivamente. Quando associado a um membro de o grupo, o valor de silhueta próximo de zero pode ser interpretado geometricamente como um elemento que está localizado na fronteira do grupo. Nestes casos é recomendado rever o resultado de uma classificação potencialmente equivocada (LIM et al., 2016).

O coeficiente de silhueta para agrupamentos é definido de modo similar, com base na média dos coeficientes dos grupos, como dado na Equação 2.48

$$SIL(C) = \frac{1}{K} \sum_{k=1}^{K} SIL(C_k) \tag{2.48}$$

A qualidade do agrupamento pode ser interpretada a partir da TABELA 4 (STRUYF et al., 1997).

O coeficiente de silhueta foi proposto por Rousseeuw (1987) como um método de visualização da qualidade de grupos e agrupamentos. Os gráficos de silhueta, como

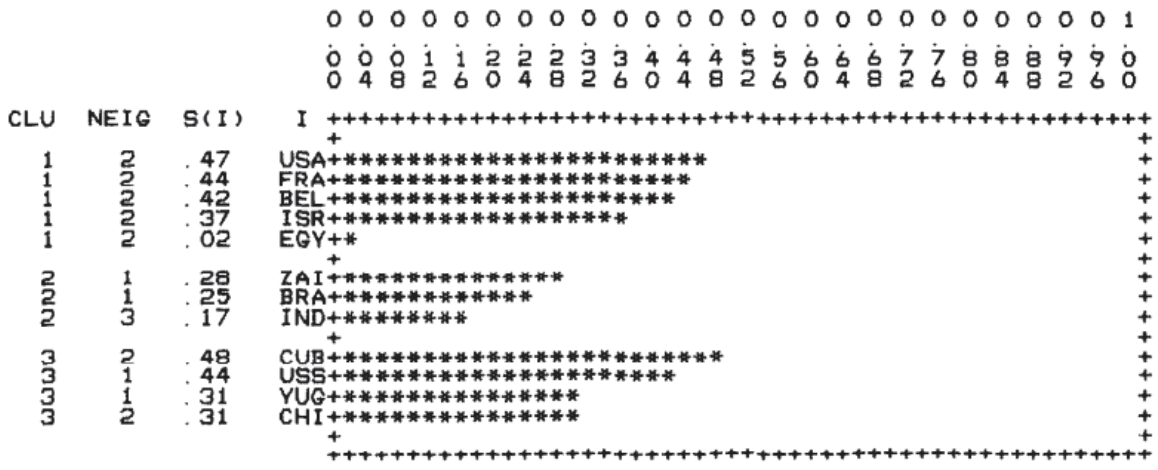
TABELA 4 – INTERPRETAÇÃO DO VALORES DE SILHUETA

Silhueta	Qualidade do agrupamento
(0.7,1.0]	Forte
(0.5,0.7]	Razoável
(0.25,0.5]	Fraca
[-1,0.25]	Nenhuma

FONTE: Adapatado de Struyf et al. (1997)

apresentado originalmente pelo autor e reproduzido na figura 20, são gráficos de barras horizontais agrupadas conforme os respectivos grupos. A largura de cada barra indica o valor de silhueta para cada elemento do conjunto de dados e a altura das faixas de barras agrupadas por grupos representam a quantidade de elementos de cada grupo. Os coeficientes médios de grupos e do agrupamento são mostrados nas duas linhas adicionais externas ao gráfico e emolduradas na figura.

FIGURA 20 – [REPRESENTAÇÃO GRÁFICA DA SILHUETA DE AGRUPAMENTOS



FONTE: (ROUSSEEUW, 1987)

Em seu artigo original, Rousseeuw (1987) desenvolve um exemplo de agrupamento (classificação) de 12 países com base em dados de um experimento com dezoito alunos em uma aula sobre medidas psicológicas na Universidade de Columbia em 1968. Trata-se de um estudo de análise multivariada para 4 dimensões (Alinhamento político e ideológico, Desenvolvimento econômico, Geografia e população e Cultura e raça) associadas a percepção de nação entre os alunos. Seguindo a divisão geo-política da época os países foram agrupados como capitalistas, socialistas e em desenvolvimento.

Na análise do gráfico de silhueta, como pode ser visto na figura 20, o primeiro grupo, países desenvolvidos capitalistas, contém cinco membros representados por

suas siglas: USA, FRA, BEL, ISR, EGY; no segundo grupo, os países em desenvolvimento: ZAI, BRA, IND; e no terceiro grupo, os países socialistas: CUB, USS, YUG, CHI). A classificação do Egito, notado comoEGY no grupo 1 (em destaque), talvez possa ser discutida. De fato, na esquerda da figura, as colunas CLU, $NEIG$ e $S(I)$ indicam o grupo, o grupo vizinho mais próximo e o coeficiente de silhueta de cada ponto de dado, respectivamente. É possível verificar que o valor de silhueta para o elemento Egito é 0.02, um valor muito menor do que a média para o grupo é 0.34. O baixo valor de silhueta do Egito indica que este país tem pouca aderência ao grupo 1. Ademais, a presença do país neste grupo diminui a sua silhueta média. Neste sentido, a sua remoção para o grupo vizinho mais próximo, no caso o grupo 2, além de aumentar a silhueta do grupo 1, parece se ajustar melhor com as características dos países em desenvolvimento do grupo 2. As informações na coluna do grupo vizinho também ajudam a explicar a proximidade de cada país na formação de um bloco ocidental, cujos países membros tem como vizinho mais próximo o grupo 1, ou de um bloco oriental, neste caso, os países membros tem como vizinhos mais próximos os países do grupo 3.

As medidas apresentadas têm sido objeto de investigação em estudos comparativos com outras medidas de validação quando aplicadas a diversos métodos de agrupamentos. Em um estudo abrangente realizado por Arbelaitz et al. (2013), trinta medidas de validação dos dois tipos, interna e externa, são utilizadas para avaliar os agrupamentos obtidos por meio de dois métodos hierárquicos (*Ward* e Ligação completa) e o método *K-means*. Cada método foi aplicado na análise de agrupamento de vinte conjuntos de dados disponíveis no repositório aberto *UCI* da Universidade da Califórnia. Os autores concluem que não foi possível distinguir um pequeno número de medidas melhores do que as outras. Entretanto, dez medidas, dentre as quais encontram-se o índice Davies-Bouldin, o índice Calinski-Harabasz e o coeficiente de silhueta, seriam recomendadas. Uma característica comum a elas é a capacidade de melhor avaliarem grupos coesos e bem separados. Geometricamente, isto significa que estes métodos se adaptam melhor a validação de agrupamentos globulares com tamanhos distribuídos uniformemente. Esta capacidade também pode ser vista como uma limitação, dependendo do ponto de vista.

Em outro estudo, Rendón et al. (2011), compara a utilização de índices internos e externos para avaliação de agrupamentos *k-means*. Os autores comparam cinco medidas internas e quatro medidas externas (dentre as quais, aquelas apresentadas anteriormente neste trabalho) e concluem que as medidas de validação interna apresentam desempenho relativamente superior do que as medidas de validação externa. Um problema em relação às medidas de validação externa, como a Entropia, é a sua dificuldade em capturar o efeito de uniformização causado pelo *K-means* quando este é definido utilizando-se a distância Euclidiana (WU et al., 2009). Neste quesito o uso da

medida F é preferível por ser menos sensível a este efeito do *K-means* como mostra Xiong et al. (2008).

Alguns autores descrevem uma terceira categoria de validação dos grupos, denominada relativa, cujas medidas são utilizadas na comparação de diferentes resultados de agrupamentos com diferentes métodos ou configurações de inicialização de um mesmo método (HALKIDI et al., 2000, 2001, 2002). Como lembra Tan et al. (2009), as medidas de validação relativa não são uma nova categoria de medidas, mas uma forma de utilizar as medidas de validação internas e externas. Na validação externa o agrupamento resultante é comparado com um agrupamento de referência cuja estrutura não resulta necessariamente da aplicação de um método de agrupamento. Na validação relativa, por sua vez, a comparação é realizada entre os agrupamentos resultantes da aplicação de um método de agrupamento sobre o mesmo conjunto de dados. Vendramin et al. (2010) avaliam 40 medidas aplicadas a validação relativa dos resultados de cinco métodos de agrupamentos de dados, incluindo o *K-means*. Dentre as medidas avaliadas o coeficiente de silhueta obteve o melhor desempenho. Novamente estes resultados são mais significativos quando os agrupamentos são globulares e bem separados.

A avaliação da qualidade de agrupamentos com origem em diferentes configurações de um método está relacionada a penúltima etapa da análise de agrupamento na qual os resultados da aplicação de um método de agrupamentos com diferentes configurações são comparados. Desta comparação deve surgir um agrupamento final que, além das propriedades esperadas, contribua para extração de padrões representativos que favoreçam a interpretação dos grupos e os relacionamentos entre seus respectivos membros. Nestes termos, a validação relativa será discutida na etapa da análise de agrupamentos denominada análise de meta-agrupamento.

2.2.9 Combinação de agrupamentos

A análise de meta-agrupamento e combinação de agrupamentos (*clustering ensemble*) são extensões de uma análise de agrupamento sobre padrões de dados na qual os resultados da aplicação de diferentes métodos, ou variações na configuração dos parâmetros de um mesmo método, são pontos no espaço de soluções dos agrupamentos possíveis associados a um determinado conjunto de dados.

Em um meta-agrupamento, os agrupamentos são agrupados segundo sua similaridade no espaço de rótulos. Conforme Caruana et al. (2006), o meta-agrupamento é descrito em três etapas:

1. gerar uma diversidade de bons agrupamentos sobre os mesmos dados;
2. medir a similaridade entre os agrupamentos;

3. utilizar um método de agrupamento para obter os agrupamentos mais similares e, por meio de uma técnica de redução da dimensionalidade, visualizar os metagrupos resultantes.

Em alguns casos, os meta-agrupamentos derivados de agrupamentos com uma diversidade moderada produzem metagrupos de melhor qualidade (HADJITODOROV et al., 2006). Quando esta diversidade de agrupamentos é obtida por meio do mesmo método então a combinação é denominada homogênea ou, caso contrário, heterogênea (NALDI et al., 2010). O objetivo do meta-agrupamento não é necessariamente escolher um agrupamento ótimo, mas reduzir o espaço de busca por este agrupamento circunscrevendo-o a algumas classes que atendam o propósito da análise (CARUANA et al., 2006).

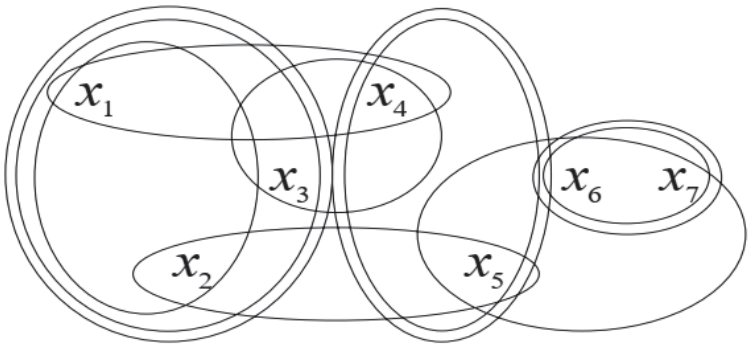
A combinação de agrupamentos (*clustering ensembles* (STREHL; GHOSH, 2002) ou *clustering aggregation* (GIONIS et al., 2007)) pode ser descrita de modo semelhante aquelas do meta-agrupamento, exceto pela etapa três. No caso da combinação, a terceira etapa tem como objetivo definir um agrupamento a partir de agrupamentos ou grupos diversos de modo que a qualidade do agrupamento final seja superior a qualidade dos agrupamentos ou grupos tomados individualmente (ZHANG; LI, 2011). Para alcançar este resultado de um agrupamento superior, os agrupamentos originais, ou agrupamentos base, são avaliados segundo uma função **consenso** responsável pela rotulação dos membros do agrupamento final, denominado **agrupamento consenso** (NALDI et al., 2010) .

O meta-agrupamento e a combinação não são excludentes. O primeiro pode ser utilizado como uma etapa anterior a geração do consenso permitindo a seleção de agrupamentos com qualidades desejadas para serem combinados no agrupamento consenso, como proposto por Zhang e Li (2011):

1. geração de agrupamentos base por meio da aplicação de diferentes métodos e parâmetros de agrupamento;
2. cálculo da similaridade dos agrupamentos base;
3. definição do meta-agrupamento;
4. geração do consenso.

Em outra abordagem (STREHL; GHOSH, 2002), o meta-agrupamento é considerado um tipo de combinação cujo objetivo é a obtenção de uma estrutura de hiper-grafos a partir da similaridade entre os agrupamentos base. A representação dos agrupamentos como hiper-grafos onde os pontos de dados são os vértices e os grupos em cada um dos agrupamentos são as hiper arestas, como ilustrado na FIGURA 21.

FIGURA 21 – AGRUPAMENTOS REPRESENTADOS COMO UM HIPER-GRAFO



FONTE: (STREHL; GHOSH, 2002)

Formalmente, conforme Aggarwal (2013), dado um conjunto $\mathcal{X}$ dos elementos $\mathcal{X}_i$, com $i = 1, \cdots, n$, e os agrupamentos base $\mathcal{C} = \{\mathcal{C}^b\}$ de $\mathcal{X}$ com $b = 1, \cdots, n_b$. Cada agrupamento $\mathcal{C}^b$ é definido, respectivamente, por seus K grupos[5], assim $\mathcal{C}^b = \{\mathcal{C}^b_k\}$, com $k = 1, \cdots, K$. Dada uma função $\mathfrak{L}^b : \mathcal{X} \subset \mathcal{C}^b \mapsto \{1, \cdots, K\}$ que associa cada elemento $\mathcal{X}_i$ ao seu rótulo de grupo k, como na Equação 2.49,

$$\mathfrak{L}^b(\mathcal{X}_i) = \mathfrak{l}^b_i = k. \tag{2.49}$$

Um agrupamento $\mathcal{C}^b$ pode ser representado por um vetor dos rótulos de seus respectivos membros $\mathfrak{l}^b = [\mathfrak{l}^b_i]$, com $i = 1, \cdots, n$. A aplicação de $\mathfrak{L}$ em todos os agrupamentos base $\mathcal{C}^b$ resulta no conjunto de agrupamentos dos rótulos de $\mathcal{X}$ em cada agrupamento $\mathcal{C}^b$ dado por $\mathfrak{l} = \{\mathfrak{l}^b\}$, com $b = 1, \cdots, n_b$. Ainda segundo (AGGARWAL, 2013), outra representação de um agrupamento é dada por sua matriz de filiação, ou co-associação,$\mathbf{H}^b$ de cada elemento $\mathcal{X}_i$ ao grupo k no agrupamento $\mathcal{C}^b$, cujas componentes são definidas pela Equação 2.50

$$h^b_{i,k} = \begin{cases} 1, \ se \ \mathcal{X}_i \in \mathcal{C}^b_k \\ 0, \ caso \ contrário \end{cases}, \tag{2.50}$$

onde $i = 1, \cdots, n$ e $k = 1, \cdots, K$.

O agrupamento consenso $\mathcal{C}^\Phi$ resulta da aplicação de uma função consenso Φ sobre o conjunto dos agrupamentos dos rótulos de $\mathcal{X}$. Uma função consenso $\Phi(\mathcal{C}^p, \mathcal{C}^q)$ entre dois agrupamentos pode ser definida como uma função de similaridade (AGGARWAL, 2013). Dessa forma, o agrupamento consenso $\mathcal{C}^\Phi$ é resultado da maximização da função consenso expressa por meio da Equação 2.51.

[5] De modo geral, um número diferente K_b de grupos poderia ser definido para cada agrupamento $\mathcal{C}^b$.

$$\Phi(\mathcal{C}) = \frac{1}{n_b} \sum_{b=1}^{n_b} \Phi(\mathcal{C}^b, \mathcal{C}^\Phi). \tag{2.51}$$

A função consenso definida na Equação 2.51 pode ser aplicada a partir de várias perspectivas dando origem a métodos diferentes dentre os quais, conforme destaca Zhou (2019), estão

- funções baseadas na matriz de co-associação ou na matriz de filiação $\mathbf{H}^b$ tomadas como uma matriz de similaridade dos agrupamentos. A matriz de similaridade pode ser utilizada com métodos de agrupamentos como *K-means* ou hierárquicos. Os métodos baseados nesta matriz não são indicados para a combinação de um número grande de agrupamentos.

- funções baseadas na matriz de adjacência do hiper-grafo dada pela matriz em bloco $[\mathbf{H}] = [\mathbf{H}^1, \cdots, \mathbf{H}^{n_b}]$ cuja matriz de similaridade S é dada por $S = [\mathbf{H}^b][\mathbf{H}^b]^\dagger$, onde $[]^\dagger$ é a inversa generalizada. O agrupamento consenso é obtido por meio de métodos de agrupamentos aplicados a grafos e não são indicados quando os agrupamentos apresentam alta diversidade.

- funções baseadas na re-rotulação dos elementos dos grupos, caso seja necessária. Este procedimento, aplicado a cada agrupamento, visa identificar similaridades entre os agrupamentos por meio da permutação dos rótulos de grupos no mesmo agrupamento e sua posterior comparação com outros grupos nos agrupamentos restantes.

- funções baseadas na transformação do espaço de dados em um espaço dos rótulos dos dados. O agrupamento consenso é obtido por meio de métodos de agrupamentos com base na similaridade dos rótulos entre os agrupamentos base.

Assim, como os métodos de agrupamentos e suas medidas de validação, a combinação de agrupamentos tem sido objeto de estudos comparativos. Os métodos de combinação baseados na matriz de filiação e na transformação para o espaço de rótulos são considerados intuitivos, simples e de fácil aplicação, mas estão limitados a um número de agrupamentos de pequeno a médio com diversidade mediana (ZHOU, 2019). Pelo mesmo motivo muitos estudos utilizam métodos de agrupamentos mais simples como *K-means* e método de agrupamento hierárquicos aglomerativos (HE et al., 2005; KUNCHEVA et al., 2006; HADJITODOROV et al., 2006; VEGA-PONS; RUIZ-SHULCLOPER, 2011).

Kuncheva et al. (2006) examinou 24 arranjos experimentais sobre vinte e quatro conjuntos de dados com diferentes estruturas geométricas e tamanhos. Os métodos

de agrupamento *K-means* e hierárquicos foram utilizados para a geração dos agrupamentos base e na geração do consenso nas combinações de agrupamentos baseadas na matriz de coassociação. Para a geração do agrupamento consenso foram utilizados métodos de combinação baseados nas funções consenso citadas. Os rótulos dos agrupamentos consenso foram comparados com os rótulos conhecidos e os melhores arranjos experimentais foram obtidos por meio da aplicação do método *K-means* na geração dos agrupamentos base e a aplicação dos métodos hierárquicos aglomerativos. Entretanto, como lembram Kuncheva e Hadjitodorov (2004) estes métodos são simples e têm sido utilizados sobre conjunto de dados com estruturas geométricas conhecidas e os resultados encontrados talvez não sejam repetidos com o uso de dados com estruturas mais complexas.

De modo geral, nesta seção foi discutido como a análise de agrupamento, incluindo a combinação de agrupamentos, exigem a sua aplicação repetidas vezes de modo que a diversidade desejada para a geração de um agrupamento consenso com boa qualidade seja alcançado. Na sequência deste trabalho são apresentadas as características específicas da análise de agrupamentos sobre dados que são representados como grafos. Neste sentido, serão expostos os conceitos e definições sobre grafos, medidas de similaridade entre grafos e as peculiaridades do agrupamento de grafos.

2.3 AGRUPAMENTO DE DADOS BASEADOS EM GRAFOS

O tipo de um dado depende dos atributos que o define, alguns dados podem ser definidos em espaços n-dimensionais na forma de vetores e matrizes. As informações associadas a estes atributos podem se referir a uma sumarização do dado por meio dos valores das variáveis que representam estes atributos ou medidas estatísticas que caraterizam este dado em relação à sua posição na distribuição do conjunto de dados ou sua amostra. Um dos atributos de um dado pode ser a sua relação com outro dado.

Dados que são descritos por meio de suas relações com outros dados podem ser representados por meio de um grafo. Um grafo é uma abstração matemática que representa entidades, denominados vértices ou nós, e de suas relações conhecidas como arestas ou arcos do grafo. Assim, vários fenômenos podem ser modelados na forma de um grafo como

- uma rede social pode ser representada como um grafo, onde as pessoas são os vértices e seus relacionamentos definem as arestas do grafo;
- na rede mundial de computadores, as páginas da web são vértices e seus hiperlinks são as arestas;

- numa rede rodoviária, as cidades definem os vértices de um grafo e as estradas as arestas que conectam as cidades;
- os átomos são vértices de um grafo que representa uma molécula e as arestas estão associadas as forças de ligação entre os átomos;
- os comandos em um programa de computador pode ser vistos como vértices de um grafo cujas arestas indicam as possíveis ordens em que eles podem ser implementados
- em um circuito eletrônico os componentes são vértices em um grafo que o representa, e as conexões entre os componentes são as arestas do grafo.

Neste contexto, uma possível abordagem na aplicação da análise de agrupamento de dados baseados em grafos tem como objetivo, por exemplo, a identificação de comunidades. As comunidades são representadas por grupos de vértices que se agrupam em torno de alguns vértices que são tomados como os elementos representativos das comunidades (AGGARWAL; WANG, 2010).

Em uma segunda abordagem, que será objeto de discussão nesta tese, a análise de agrupamento que pode ser aplicada a conjunto de dados cujos elementos são grafos. Neste caso, dado um conjunto de grafos, o objetivo da análise de agrupamento é agrupar os grafos similares a partir da definição de uma medida de similaridade ou dissimilaridade (distância) entre dois grafos (KOUTRA; FALOUTSOS, 2018). Em certa medida, a solução para o problema de similaridade entre dois grafos pode ser vista como uma aproximação para o problema mais geral e difícil no qual se procura saber se dois grafos quaisquer são isomorfos (AGGARWAL; WANG, 2010). Na seção a seguir são apresentados os conceitos fundamentais e definições de similaridade entre grafos.

2.3.1 Grafos

Um grafo $\mathcal{G}(\mathcal{V}, \mathcal{E})$ é definido por um conjunto de N vértices $\{\mathcal{V}\} = \mathcal{V}_1, ..., \mathcal{V}_N$ e um conjunto de M arestas $\{\mathcal{E}\} = \mathcal{E}_1, ..., \mathcal{E}_M$, onde $\mathcal{E}_k = \mathcal{E}(\mathcal{V}_i, \mathcal{V}_j) \subseteq \mathcal{V} \times \mathcal{V}$. Uma aresta é uma conexão, ou adjacência, $i \sim j$ entre $\mathcal{V}_i$ e $\mathcal{V}_j$, com $i, j = 1, \cdots, N$.

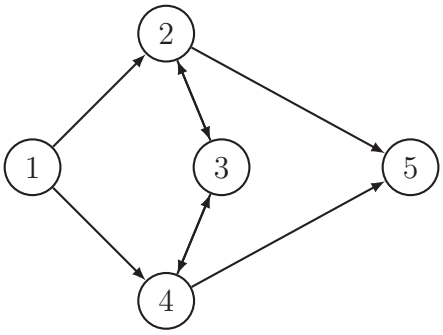
Um grafo é denominado ponderado quando aos seus vértices ou arestas estão associados a atributos quantitativos e, não ponderado, em caso contrário. Em relação à suas arestas um grafo pode ser dirigido, quando $\mathcal{E}(\mathcal{V}_i, \mathcal{V}_j) \neq \mathcal{E}(\mathcal{V}_j, \mathcal{V}_i)$, e não dirigido quando $\mathcal{E}(\mathcal{V}_i, \mathcal{V}_j) = \mathcal{E}(\mathcal{V}_j, \mathcal{V}_i)$, com $i \neq j$. Um grafo dirigido também é conhecido como **dígrafo**. Os vértices e as arestas ainda podem ter atributos qualitativos como, por exemplo, rótulos que os nomeiem.

Um grafo dirigido pode ser representado graficamente como um diagrama cujos vértices são pontos e as arestas são segmentos orientados entre os vértices

conectados. Os segmentos orientados introduzem as categorias de vértices sucessores e vértices antecessores de um vértice dado.

Exemplo 2.3.1 *Dado o grafo dirigido e não ponderado $\mathcal{G}$ da FIGURA 22, onde $\mathcal{V}_{\mathcal{G}} = \{1, 2, 3, 4, 5\}$, os vértices 2 e 4 são sucessores do vértice 1, e este é um antecessor daqueles.*

FIGURA 22 – Grafo $\mathcal{G}$



FONTE: O Autor

Algebricamente, o grafo $\mathcal{G}$ fica completamente definido por meio de sua matriz de adjacência $A_{\mathcal{G}}$ onde

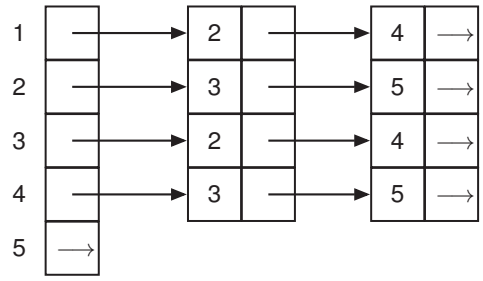
$$A_{ij} = \begin{cases} 1, \ se \ i \sim j \in \mathcal{E} \\ 0, \ caso \ contrário \end{cases} \tag{2.52}$$

Assim, a matriz de adjacência do grafo G da FIGURA 22 é

$$A_{\mathcal{G}} = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{2.53}$$

Os relacionamentos entre os vértices do grafo $\mathcal{G}$ podem ser representados por uma lista de adjacência. A FIGURA 23 representa uma lista de adjacência onde cada linha é reservada para um vértice do grafo e as setas indicam os respectivos sucessores.

FIGURA 23 – Lista de adjacência de $\mathcal{G}$



FONTE: O Autor

O grau δ_i do vértice i corresponde ao número de arestas incidentes ao vértice i. Portanto, a matriz $D_{\mathcal{G}}$ do grau dos vértices é uma matriz diagonal onde $D_{i,i} = \delta_i$. Para grafos dirigidos pode-se considerar $\delta_i = \mathring{\delta}_i + \hat{\delta}_i$, onde a primeira parcela, denominada grau de entrada, é o número de arestas com destino no vértice i, e a segunda parcela é o número de arestas com origem em i, denominada grau de saída do vértice. Aqui adota-se $\delta_i = \hat{\delta}_i$ o que resulta na seguinte matriz para o grafo do exemplo 2.3.1.

$$D_{\mathcal{G}} = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{2.54}$$

A partir da matriz de adjacência $A_{\mathcal{G}}$ e a matriz grau do vértice $D_{\mathcal{G}}$, a matriz Laplaciana de $\mathcal{G}$ é definida por meio da equação matricial 2.56:

$$L_{\mathcal{G}} = D_{\mathcal{G}} - A_{\mathcal{G}} \tag{2.55}$$

A matriz Laplaciana do grafo $\mathcal{G}$ é

$$L_{\mathcal{G}} = \begin{bmatrix} 2 & -1 & 0 & -1 & 0 \\ 0 & 2 & -1 & 0 & -1 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{2.56}$$

Dada a matriz Laplaciana L_G calcula-se a matriz Laplaciana normalizada por meio da Equação 2.57:

$$\mathcal{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}. \tag{2.57}$$

Na Equação 2.57, o fator $D^{-\frac{1}{2}}$ corresponde a matriz inversa da raiz quadrada de D. Como D é diagonal então

$$D_{i,i}^{-\frac{1}{2}} = \frac{1}{\sqrt{\delta_i}}, \tag{2.58}$$

de onde, segue que

$$D_{\mathcal{G}}^{-\frac{1}{2}} = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{2.59}$$

Dessa forma, a matriz Laplaciana normalizada de G pode ser determinada por meio do produto

$$\begin{bmatrix} -\frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 \\ 0 & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & 0 & -\frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2 & -1 & 0 & -1 & 0 \\ 0 & 2 & -1 & 0 & -1 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} -\frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 \\ 0 & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & 0 & -\frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{2.60}$$

, ou seja,

$$\mathcal{L}_{\mathcal{G}} = \begin{bmatrix} 1 & \frac{-1}{2} & 0 & \frac{-1}{2} & 0 \\ 0 & 1 & \frac{-1}{2} & 0 & 0 \\ 0 & \frac{-1}{2} & 1 & \frac{-1}{2} & 0 \\ 0 & 0 & \frac{-1}{2} & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{2.61}$$

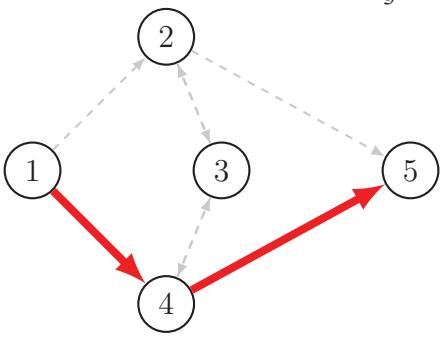
Definição 2.3.2 *Um passeio $\mathcal{W}_{\mathcal{G}}$ em $\mathcal{G}$*

$$\mathcal{W}_{\mathcal{G}} = \{\mathcal{V}_1^{\mathcal{G}}, \mathcal{V}_2^{\mathcal{G}}, ..., \mathcal{V}_r^{\mathcal{G}}\} \tag{2.62}$$

é uma sequência de r vértices conectados.

A rigor, no grafo dirigido $\mathcal{G}$ o conjunto de arestas de $\mathcal{W}_{\mathcal{G}} \subseteq \mathcal{E}_{\mathcal{G}}$. Assim, para o exemplo 2.3.1, um caminho pode ser definido pela sequência de vértices $\mathcal{W}_{\mathcal{G}} = \{1, 4, 5\}$ como apresentado na FIGURA 24.

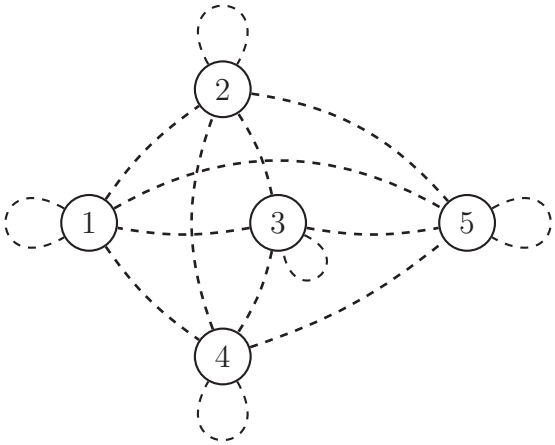
FIGURA 24 – Passeio $\mathcal{W}_{\mathcal{G}}$



FONTE: O Autor

Neste trabalho define-se um grafo $\mathcal{G}^*$ induzido pelos vértices de $\mathcal{G}$, tal que $\mathcal{W}_{\mathcal{G}^*} \subseteq \mathcal{E}_{\mathcal{G}^*}$. O grafo $\mathcal{G}^*$ induzido pelos vértices de $\mathcal{G}$ possui as arestas conectando todos os vértices entre si e arestas que têm origem e destino sobre o mesmo vértice, conectando um vértice a si mesmo por meio de um laço como apresentado na FIGURA 25.

FIGURA 25 – Grafo $\mathcal{G}^*$ induzido pelos vértices de $\mathcal{G}$



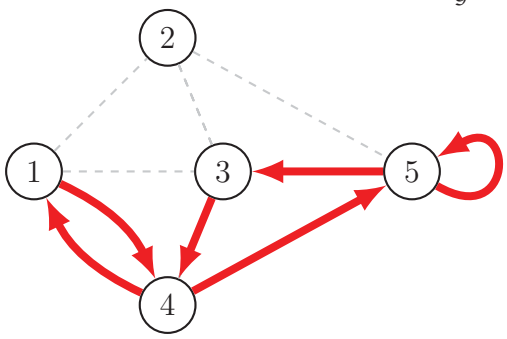
FONTE: O Autor

Como pode ser observado, o número de arestas possíveis entre os vértices de $\mathcal{G}^*$ é maior do que o número de arestas de $\mathcal{G}$. A matriz de adjacência deste grafo é a matriz $[1]_{\mathcal{G}^*}$ (Eq.: 2.63), onde todas as entradas são iguais a um.

$$[1]_{\mathcal{G}^*} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{2.63}$$

Para o exemplo 2.3.1, um passeio $\mathcal{W}_{\mathcal{G}^*} = \{1, 4, 5, 5, 3, 4, 1\}$ é apresentado na FIGURA 26.

FIGURA 26 – Passeio sobre $\mathcal{W}_{\mathcal{G}^*}$



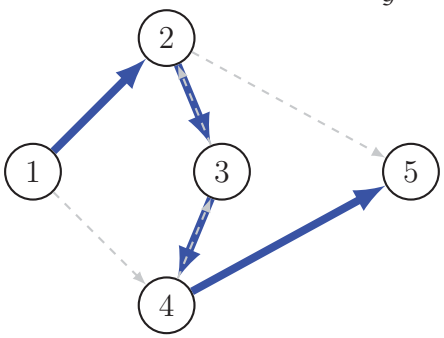
FONTE: O Autor

Definição 2.3.3 *O caminho simples de comprimento K sobre $\mathcal{G}$*

$$\mathcal{P} = \{\mathcal{V}_1^{\mathcal{G}}, \mathcal{V}_2^{\mathcal{G}}, ..., \mathcal{V}_k^{\mathcal{G}}\} \tag{2.64}$$

é dado por uma sequência de vértices e as arestas que os conectam tal que não exista repetição de vértices.

FIGURA 27 – Caminho $\mathcal{P}_{\mathcal{G}}$



FONTE: O Autor

2.3.2 Similaridade entre grafos

Em determinadas aplicações como problemas de roteamento em logística, sequenciamento genético na bioinformática, detecção de erros ou códigos maliciosos em programas de computador, comportamento de usuários em redes sociais, dentre outras, o reconhecimento de padrões pode ser tratado como um problema de se medir o grau de similaridade entre grafos.

Um exemplo de similaridade entre grafos pode ser definido a partir de Silverman (1967). Seguindo o autor, uma tarefa com N atividades pode ser modelada como um grafo $\mathcal{G}$ com N vértices e M arestas que conectam estes vértices indicando possíveis sequências com as quais as atividades que compõem a tarefa podem ser executadas. Um passeio $\mathcal{W}$ no grafo $\mathcal{G}^*$ induzido a partir dos vértices de $\mathcal{G}$ é uma sequência ordenada de vértices que corresponde a uma certa ordem de execução de R atividades. Neste contexto passeios similares indicam um padrão comum na execução da tarefa.

Seja $\{\mathcal{W}_i\}$, com $i = 1, \cdots, L$, um conjunto de L passeios sobre o grafo $\mathcal{G}^*$, onde cada passeio i corresponde a uma execução da tarefa. Dados dois passeios $\mathcal{W}_i$ e $\mathcal{W}_j$, sejam $|\mathcal{V}_{\mathcal{W}_)}|$ e $|\mathcal{V}_{\mathcal{W}_|}|$ os números de vértices dos caminhos i e j, respectivamente. Uma medida de similaridade pode ser definida por meio da função similaridade aplicada sobre $\mathcal{W} \times \mathcal{W}$:

$$Sim(\mathcal{W}_i, \mathcal{W}_j) = \frac{|\mathcal{W}_i \cap \mathcal{W}_j|}{|\mathcal{W}_i| + |\mathcal{W}_j| - |\mathcal{W}_i \cap \mathcal{W}_j|}. \tag{2.65}$$

A aplicação da Equação 2.65) aos passeios i, j de $\{\mathcal{W}\}$ gera uma matriz de similaridade $[S]_{L\times L}$, a partir da qual é possível agrupar os caminhos similares segundo o grau de similaridade. Esta definição de similaridade expressa relação entre os vértices em comum e a união disjunta dos vértices nos dois passeios avaliados, mas não leva em consideração a ordem dos vértices ou, no contexto da aplicação, a ordem das atividades executadas. A ordem com a qual uma tarefa é executada durante um treinamento pode revelar padrões que caracterizam estratégias de ação pelo treinando ou ainda a comparação com um padrão de ordem esperado. Padrões esperados ou permitidos podem igualmente serem representados como passeios sobre estes grafos, permitindo medir a similaridade entre a execução de um usuário e uma execução ótima no sentido de serem completas e as únicas possíveis segundo a ordem correta de execução das atividades.

Na próxima seção são discutidas algumas medidas de similaridade entre grafos dentre as quais algumas delas dependem em sua definição de uma distância entre grafos. Uma medida de similaridade pode então ser dada pela Equação 2.66

Definição 2.3.4 *Dados dois grafos $\mathcal{G}$ e $\mathcal{G}'$, a similaridade $Sim(\mathcal{G}, \mathcal{G}')$ entre dois grafos é definida como*

$$Sim(\mathcal{G}, \mathcal{G}') = \frac{1}{1 + d(\mathcal{G}, \mathcal{G}')}, \tag{2.66}$$

onde $d(\mathcal{G}, \mathcal{G}')$ é a distância entre $\mathcal{G}$ e $\mathcal{G}'$.

Nota-se que quanto menor for a distância $d(\mathcal{G}, \mathcal{G}')$, maior será o grau de similaridade entre $\mathcal{G}$ e $\mathcal{G}'$, dessa forma, $Sim(\mathcal{G}, \mathcal{G}') \in [0, 1]$. Diferentes definições de distâncias podem ser utilizadas conforme apresentado na seção 2.3.3.

2.3.3 Medidas de similaridade de grafos

Seguindo Wills e Meyer (2020), as medidas de similaridade apresentadas a seguir podem ser divididas em três tipos:

- distâncias espectrais, ou **distâncias** λ, são baseadas na decomposição espectral e avaliação dos autovalores das matrizes de adjacência, Laplaciana ou Laplaciana normalizada dos grafos.
- as distâncias matriciais são baseadas em operações sobre as matrizes de adjacência dos grafos, como nos casos da **distância de edição** e da distância **sobreposição vértice-aresta**; ou sobre a matriz Laplaciana do grafo como na definição da distância de **resistência normalizada** ou a matriz de adjacência que entra no cálculo da medida **Deltacon**.
- as distâncias vetoriais capturam as diferenças entre dois grafos por meio da comparação da distribuição estatística dos vértices e arestas dos grafos. Este é o caso da medida de similaridade **Netsimile**.

Distâncias espectrais

O espetro de uma matriz M é a sequência ordenada de seus autovalores λ_i^M. No caso da matriz de adjacência a sequência de autovalores é ordenada em ordem decrescente e para a matriz Laplaciana a ordem dos autovetores é crescente (STANIĆ; JOVANOVIĆ, 2014). Seja λ_i^A e $\lambda_i^{A'}$, com $i = 1, 2, ..., n$ a sequência dos autovetores da matriz M e M' do grafos $\mathcal{G}$ e $\mathcal{G}'$, respectivamente. A distância $d_M(\mathcal{G}, \mathcal{G}')$ é definida como

$$d_\lambda(\mathcal{G}, \mathcal{G}') = \left(\sum_{i=1}^{k} (\lambda_i^M - \lambda_i^{M'})^p \right)^{\frac{1}{p}}, \tag{2.67}$$

onde para $k \leq n$, p define a distância de Minkovski ℓ_p (WILLS; MEYER, 2020). As distâncias espectrais são invariantes a permutação dos rótulos dos vértices e a escolha de valores menores de K reflete a estrutura global do grafo.

Distância de edição de grafos

A distância de edição de grafos, *Gaph Edit Distance* (GED) (GAO et al., 2009), é calculada a partir da quantidade de operações necessárias para transformar $\mathcal{G}$ em $\mathcal{G}'$ com custo mínimo. As operações de edição sobre os grafos incluem apagar um vértice ou aresta ou incluir um vértice ou aresta. Para capturar as edições realizadas na transformação de $\mathcal{G}$ em $\mathcal{G}'$ pode-se usar a distância definida como a diferença entre as matrizes de adjacência A e A' dos grafos $\mathcal{G}$ e $\mathcal{G}'$, respectivamente.

$$d_{GED}(\mathcal{G}, \mathcal{G}') = \|A - A'\| = \sum_{i,j}^{n} |A_{i,j} - A'_{i,j}| \tag{2.68}$$

Sobreposição vértice/aresta

No exemplo apresentado pela Equação 2.65, a similaridade foi definida sobre a quantidade de vértices compartilhados pelos grafos $\mathcal{G}$ e $\mathcal{G}'$. Em sobreposição vértice/aresta, *Vertex/Edge Overlap* (VEO) (FELLOWS et al., 2009) é calculada a partir da razão entre quantidade de vértices e arestas compartilhadas e a soma de vértices e arestas por $\mathcal{G}$ e $\mathcal{G}'$.

$$d_{VEO}(\mathcal{G}, \mathcal{G}') = 2\frac{|\mathcal{V}_{\mathcal{G}} \cap \mathcal{V}_{\mathcal{G}'}| + |\mathcal{E}_{\mathcal{G}} \cap \mathcal{V}_{\mathcal{G}'}|}{|\mathcal{V}_{\mathcal{G}}| + |\mathcal{V}_{\mathcal{G}'}| + |\mathcal{E}_{\mathcal{G}}| + |\mathcal{E}_{\mathcal{G}'}|} \tag{2.69}$$

Deltacon

Deltacon ($\delta - con$) (KOUTRA et al., 2013) é um algoritmo que compara as afinidades entre nós. A similaridade entre dois grafos $\mathcal{G}$ e $\mathcal{G}'$ é definida como na Equação 2.66, onde a distância $d_{Deltacon}(\mathcal{G}, \mathcal{G}')$ é calculada pela expressão

$$d_{Deltacon}(\mathcal{G}, \mathcal{G}') = \left(\sum\sum(\sqrt{S_{ij}} + \sqrt{S'_{ij}})\right)^{\frac{1}{2}}, \tag{2.70}$$

com $i, j = 1, 2, 3, ..., n$ e onde, S e S' são as matrizes de afinidades entre os nós de $\mathcal{G}$ e $\mathcal{G}'$, respectivamente, definidas por

$$S = [I + \epsilon D_G - \epsilon A_G]^{-1}, \tag{2.71}$$

onde D é a matriz diagonal grau dos vértices e A a matriz de adjacência de $\mathcal{G}$. O valor de ϵ na Equação 2.71 é uma constante pequena associada à vizinhança dos nós. As colunas da matriz S contém o vetor de afinidades s_i do nó i. s_i é a solução da equação

$$[I + \epsilon D - \epsilon A]s_i = \boldsymbol{e}_i, \tag{2.72}$$

onde $\boldsymbol{e}_i$ o vetor cuja componente $i = 1$ e 0 em caso contrário. A Equação 2.72 provém do método *Fast Belief Propagation* que modela a difusão da informação em um grafo e sobre o qual Deltacon foi desenvolvido.

Resistência normalizada

A resistência (Res) de um grafo é calculada com base na analogia com um circuito elétrico onde as arestas entre dois vértices $\mathcal{V}_i$ e $\mathcal{V}_j$ representam resistores com resistência $\frac{1}{w_{ij}}$ (KLEIN; RANDIC, 1993). A distância resistência normalizada (CHEN; ZHANG, 2007) entre dois grafos $\mathcal{G}$ e $\mathcal{G}'$ pode ser expressa como

$$d_{res}(\mathcal{G}, \mathcal{G}') = \left(\sum_{i=1}^{k} (R - R')^p \right)^{\frac{1}{p}}, \tag{2.73}$$

onde R é a matriz resistência definida em termos da inversa generalizada da matriz Laplaciana $L^\dagger$ como segue

$$R = diag(L^\dagger)1^T + 1diag(L^\dagger)^T + 2L^\dagger, \tag{2.74}$$

onde $diag(L^\dagger)$ é a matriz diagonal de $L^\dagger$ e 1 é a matriz com todas componentes iguais a 1.

Netsimile

Netsimile (BERLINGERIO et al., 2012) é uma medida de similaridade baseada em características dos vértices dos grafos em relação à sua vizinhança (*egonet features*) como a quantidade e média do grau dos vértices vizinhos. Um vetor s das medidas estatísticas da distribuição destas características contendo a média, a mediana, o desvio-padrão, assimetria e curtose é a assinatura (*signature*) do grafo. A similaridade entre $\mathcal{G}$ e $\mathcal{G}'$ é definida pela distância de Camberra

$$d_{Netsim}(\mathcal{G}, \mathcal{G}') = \sum \frac{|s - s'|}{s + s'} \tag{2.75}$$

Variações nos resultados decorrentes da utilização destas definições de similaridade estão relacionadas a sua capacidade de capturarem diferenças locais, globais, ou ambas.

Conforme sumarizado na TABELA 5, estas medidas diferem quanto a sua eficiência em determinar se dois grafos são similares conforme estas diferenças se revelem em relação à propriedades

1. locais, baseada na forma como cada um dos vértices se conecta com seus vizinhos mais próximos; ou
2. globais, baseada na forma como todos os vértices se conectam entre si.

TABELA 5 – DISTÂNCIAS ENTRE GRAFOS

Distância	Formulação	Entrada	Sensibilidade
$d_{\lambda_M}(G, G')$	$\left(\sum_{i=1}^{k}(\lambda_{Mi} - \lambda_{M'i})^p\right)^{\frac{1}{p}}$	M_G e $M_{G'}$, $M = A$ ou $M = L$	Global/Local
$d_{GED}(G, G')$	$\|A - A'\| = \sum_{i,j}^{n} \|A_{i,j} - A'_{i,j}\|$	A_G e $A_{G'}$	Local
$d_{VEO}(G, G')$	$2\frac{\|V_G \cap V_{G'}\| + \|E_G \cap E_{G'}\|}{\|V_G\| + \|V_{G'}\| + \|E_G\| + \|E_{G'}\|}$	G, G' e $G \cap G'$	Local
$d_{\delta-con}(G, G')$	$\left(\sum(\sqrt{S_{ij}} + \sqrt{S'_{ij}})\right)^{\frac{1}{2}}$	$S = [I + \epsilon D - \epsilon A]^{-1}$	Global/Local
$d_{Res}(G, G')$	$\left(\sum_{i=1}^{k}(R - R')^p\right)^{\frac{1}{p}}$	$R = diag(L^{\dagger})[1]^T + [1]diag(L^{\dagger})^T + 2L^{\dagger}$	Global
$d_{Netsim}(G, G')$	$\sum \frac{\|s - s'\|}{s + s'}$	s = vetor assinatura de G	Local

FONTE: Baseado em Wills e Meyer (2020)

Na análise de grafos, estas diferenças, respectivamente, respondem a perguntas distintas como sobre a detecção de anomalias ou a identificação de comunidades.

2.3.4 Resumo

Neste capítulo, os fundamentos teóricos sobre os quais se apoiam esta tese foram apresentados em três seções. Na primeira seção, Aprendendo com os erros, foram discutidos como o estudo do comportamento humano e dos mecanismos cognitivos presentes no processo de aprendizagem são permeados pelo erro na interface humano-sistema. Dessa forma, destacou-se o papel da modelagem e mapeamento do erro humano durante a execução de uma tarefa na avaliação do desempenho. Na segunda seção, Aprendendo com os dados, foram discutidos os conceitos e as ferramentas de análise e visualização de dados utilizados para a identificação de padrões no desempenho dos treinandos. Na última seção, Agrupamento de dados baseados em grafos, foram expostos os fundamentos da linguagem de grafos aplicados à modelagem de padrões de dados e as definições de similaridade que permitem agrupar estes padrões. No próximo capítulo são discutidos os aspectos metodológicos relativos ao desenvolvimento de soluções para alcançar os objetivos desta tese.

3 ASPECTOS METODOLÓGICOS

Neste capítulo são discutidos os aspectos metodológicos para o desenvolvimento do método de avaliação do desempenho dos treinandos a partir do agrupamento dos padrões de dados. A estrutura de pesquisa que suporta este desenvolvimento é abordado por meio da perspectiva da *Design Science Research* (DSR). As etapas de modelagem e análise dos dados de interação dos treinandos aqui descritas, são discutidas nos dois capítulos seguintes.

3.1 ASPECTOS METODOLÓGICOS DA *DESIGN SCIENCE RESEARCH*

Esta pesquisa segue uma estrutura de investigação adaptada da *Design Science Research* (DSR), composta por cinco atividades, conforme ilustrado na FIGURA 28: identificação e conscientização do problema, sugestão de solução, desenvolvimento da solução, validação da solução e comunicação dos resultados da pesquisa.

FIGURA 28 – ATIVIDADES DO DESIGN SCIENCE RESEARCH

FONTE: Adaptado de Dresch et al. (2015)

O objetivo deste trabalho é desenvolver um método. Na abordagem epistemológica da Design Science Research – DSR[1], um **método** é definido como um "artefato",

[1] Conforme Bax (BAX, 2013), a DSR é uma metateoria, "que fundamenta e operacionaliza a condução da pesquisa quando o objetivo a ser alcançado é um artefato ou uma prescrição" (DRESCH et al., 2015). Trata-se de uma busca por alternativas metodológicas para a pesquisa aplicada às áreas de

uma construção humana. A DSR é orientada à resolução de um problema específico e, segundo March e Smith (1995) citados por (DRESCH et al., 2015), quatro são os resultados da pesquisa científica: os conceitos, os modelos, os métodos e as instanciações. Um dos principais produtos das pesquisas científicas baseadas em DSR, os métodos, são “uma coleção de etapas (algoritmo ou guia) para a execução de uma tarefa” (MARCH; SMITH, 1995). Os métodos são construídos a partir de

1. conceitos, também denominados **constructos**, subjacentes ao campo semântico do contexto de aplicação;
2. modelos, ou representações, do espaço de soluções de um problema.

Um método deve ser capaz de atuar sobre a representação de um conjunto de dados e produzir uma outra representação mantendo invariantes certas características (MARCH; SMITH, 1995). Em particular, um método deve ser capaz de fornecer uma projeção dos dados que torne visível essas características. Neste sentido, os métodos se comportam como morfismos.

Baseado em Lacerda et al. (2013), a estrutura metodológica deste trabalho é definida em cinco atividades.

Identificação e conscientização sobre o problema Atividade na qual o problema e os requisitos para sua solução são definidos. Aqui também são expostos a justificativa do trabalho e o contexto específico de aplicação. Por meio da consulta às bases bibliográficas e documentais sobre o problema são identificados os elementos que comporão a solução do problema. Neste trabalho, o resultado desta atividade é apresentado na Introdução e no capítulo da Fundamentação Teórica.

Sugestão de uma solução Uma solução para o problema é elaborada a partir dos conceitos, modelos, métodos e instanciações estudados. A solução proposta é baseada na visualização de padrões do erro humano. Nesta etapa é apresentado um modelo conceitual do método de avaliação a partir do agrupamento dos padrões de erro dos treinandos identificados na execução da tarefa.

Desenvolvimento da solução O desenvolvimento da solução é realizado por meio de um experimento, uma instanciação do método, que pode ser aplicado ao contexto específico do problema real ou, alternativamente, a contextos artificiais baseados em dados gerados por simuladores. Neste trabalho, o desenvolvimento da solução tem três fases: a modelagem, a análise e a validação de agrupamentos

engenharia, gestão e tecnologia da informação, dentre outras, em relação à tradição estabelecida nas ciências naturais e sociais.

dos padrões de dados. Neste sentido uma primeira solução é apresentada com base na extração manual de padrões de erro dos treinandos.

Avaliação e validação da solução Nesta atividade a solução é avaliada e os resultados obtidos são realimentados nas atividades anteriores com o objetivo de aperfeiçoamento do artefato. Para a primeira solução, os agrupamentos baseados em padrões de erro são avaliados segundo os critérios de validação de agrupamentos. Novas etapas de sugestão, desenvolvimento e avaliação de novas soluções fundamentadas na modelagem de dados baseada em grafos são propostas com o objetivo de automatizar o agrupamento de padrões.

Conclusão e comunicação dos resultados da pesquisa A última atividade da pesquisa consiste na síntese dos resultados encontrados e na proposição de novas aplicações do artefato produzido. A conclusão da pesquisa inclui um relato dos avanços alcançados e das limitações dos resultados encontrados. A solução aplicada ao contexto específico deve ser generalizada para outros casos na mesma classe de problemas. Finalmente, a conclusão da pesquisa implica sua comunicação à comunidade científica por meio de sua defesa e publicização por meio de artigos e participação em congressos.

Dentre estas cinco atividades, as três intermediárias são apresentadas neste capítulo. As atividades de sugestão, desenvolvimento e avaliação da solução do problema formam um bloco de atividades recorrentes com mútua influência. Na verdade, a busca por alternativas nesta fase de realimentação pode exigir um retorno e aprofundamento no conhecimento do problema e a concepção de outras soluções possíveis. A seguir são apresentados os modelos sobre os quais o método proposto está fundamentado.

3.1.1 Sugestão de solução

O principal desafio desta tese é o desenvolvimento de um método de avaliação do desempenho de treinandos durante a execução de uma tarefa de manutenção em ambiente de treinamento virtual. As operações para a manutenção de subestações elétricas em linha viva são consideradas de alto risco, pois implicam a atuação sobre um sistema complexo e sujeito a situações inesperadas e perigosas. Neste cenário crítico, as intervenções humanas demandam atenção aos procedimentos de segurança e à prevenção e controle do erro humano. Neste sentido, o método proposto deve ser capaz de modelar o fenômeno do erro humano e o estado de conhecimento do treinando a partir dos dados de interação registrados durante a execução da tarefa instrucional.

Conforme a estrutura da *DSR*, uma sugestão para a solução do problema de avaliação do desempenho dos treinandos deve ser apresentada segundo a integração

entre conceitos, modelos, métodos e instanciações identificados na Fundamentação Teórica. A atividade de sugestão é definida como uma fase de **abdução** da solução, em contraste paradigmático, com as atividades de desenvolvimento e validação, as quais fazem parte da fase de **dedução** da solução. Sendo assim, a abdução é tida como um momento de convergência criativa de outros artefatos para a construção de uma solução do problema (PIMENTEL et al., 2020). O método de avaliação sugerido, no sentido da *DSR* é criado com base nos seguintes artefatos:

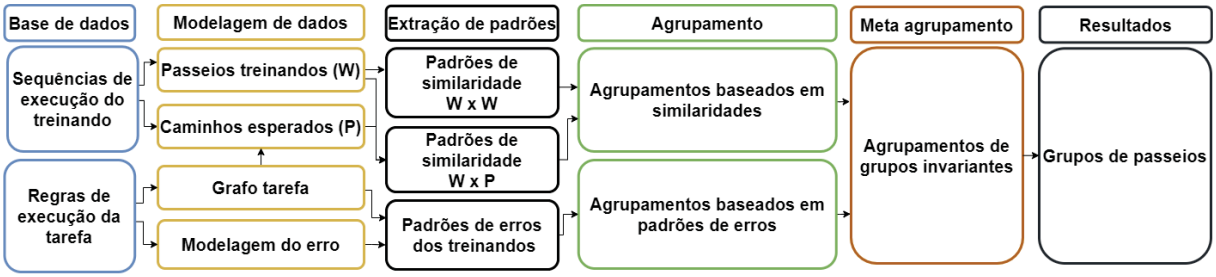
- Um modelo que descreva o fenômeno do erro humano e sua relação com os processos cognitivos. Uma característica importante dos dados disponíveis para análise está relacionada a modalidade de treinamento onde o treinado executa a tarefa. A avaliação de fatores associados as interações coletivas ou com operadores virtuais autônomos, ou *non-player character* não são exploradas. Como apresentado na fundamentação teórica o modelo *SRK* fornece uma descrição do erro humano em termos de três tipos (Habilidades, Regras e Conhecimentos) segundo níveis de processamento da informação. De acordo com o modelo SRK a tarefa instrucional do RV2 está relacionada ao **conhecimento de regras**. Uma taxonomia do erro humano aplicada ao treinamento profissional em linha viva para a identificação e caracterização os tipos de erros cometidos no contexto de aplicação;
- Um modelo que incorpore o domínio do conhecimento do especialista e do conhecimento do treinando. Dada a natureza procedimental da tarefa analisada, a modelagem do conhecimento adota uma abordagem baseada em grafos seguindo as referências sobre **espaços de conhecimento** como **espaços de estado**. A representação de um modelo do domínio do conhecimento na forma de um grafo de estados associa as regras de execução da tarefa instrucional a um **grafo tarefa**. Em uma representação baseada em grafos, as sequências esperadas para ordenação das tarefas são **caminhos viáveis** sobre o grafo tarefa. Segue, deste último, que um modelo do treinando corresponde a **passeios livres** sobre o grafo tarefa. Os passeios livres se diferenciam dos caminhos por seus desvios que podem estar associados a tipos de erros cometidos em relação às regras de execução da tarefa.
- Um modelo de análise e visualização dos dados baseados em grafos. As ferramentas de análise dos dados de interação do treinando devem fornecer informações sobre o desempenho na execução das atividades da tarefa instrucional. A análise e visualização dessas informações têm como finalidade facilitar a comparação do desempenho do treinando em relação a um desempenho esperado, e entre os treinandos entre si. O resultado desta comparação é a formação de classes que agrupem níveis de desempenho semelhantes na mesma classe.

3.1.2 Desenvolvimento, Avaliação e Validação

As atividades de **Desenvolvimento**, da **Avaliação** e da **Validação** da solução, nos termos da *DSR*, toma como variável de controle diferentes definições de similaridade de passeios colocando à prova diversas instanciações do método. Os resultados obtidos são analisados segundo critérios de validação de grupos. Os testes realizados são submetidos a análise de meta agrupamento e a combinação de agrupamentos para a identificação de uma medida de similaridade entre grafos que capture os padrões de erros dos treinandos no agrupamento final.

O processamento e análise dos dados é estruturado segundo um fluxo de trabalho para o agrupamento de dados adaptado de (JAIN, 2010), como apresentado a seguir.

FIGURA 29 – MÉTODO DE AGRUPAMENTO DE PASSEIOS



FONTE: Baseado em Jain (2010)

Conforme a FIGURA 29, a análise de agrupamentos proposta tem 6 etapas:

1. a extração de informações do banco de dados do sistema
2. a modelagem de dados em um grafo da tarefa, passeios do treinando e caminhos viáveis
3. a definição de padrões de similaridades e padrões de erro para os passeios do treinando
4. o agrupamento dos passeios do treinando de acordo com as similaridades dos passeios e dos padrões de erro
5. a análise de meta agrupamento para identificar os agrupamentos de similaridades que estão mais próximos do agrupamento de padrões de erro; e
6. a combinação de agrupamento para formação de um agrupamento consenso.

3.1.3 Recursos computacionais

Neste trabalho, a geração, a manipulação e a visualização de grafos utilizam a biblioteca *Networkx* disponíbeis para a linguagem de programação *Python* (HAGBERG et al., 2020) e o aplicativo *Gephi* (BASTIAN et al., 2009). Os cálculos de similaridade são realizados a partir da biblioteca *Netcomp* (WILLS, 2017) para *Python* que funciona conjuntamente com *Networkx*. O método de agrupamento *k-means*, os cálculos e a visualização da qualidade do agrupamento por meio do coeficiente de silhueta e a visualização de grupos utilizando a projeção das componentes principais, *PCA*, é realizada a partir de sua implementação na biblioteca *Scikit-learn* (PEDREGOSA et al., 2011). Por fim, a biblioteca *DiceR* ((CHIU; TALHOUK, 2018)) para linguagem de programação ***R*** é utilizada na combinação dos agrupamentos consenso.

Os programas criados foram implementados por meio dos ambientes de desenvolvimento *Spyder* para *Python* e *RStudio* para ***R*** a partir de suas instalações na plataforma *Anaconda* 3, versão 2021.11. O computador utilizado foi um Acer Nitro 5 com processador Intel i7-7700HQ, 2.8GHz. 8GB RAM, HD SSD 128 GB, GPU Nvidia GeForce GTX 1050, com sistema operacional *Windows 10 Pro*.

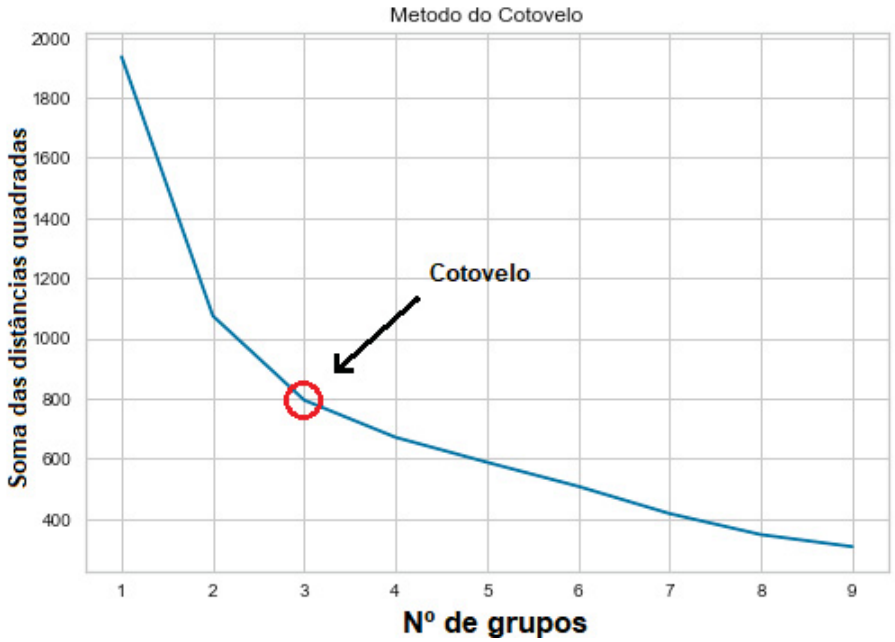
3.1.4 Agrupamentos de padrões: aplicação do método *k-means*

Entre vários métodos de agrupamento, o k-médias (*k-means*) é um método de agrupamento por partições simples com uma ampla gama de aplicações (AGGARWAL, 2013). A implementação computacional da análise de agrupamentos utilizando a linguagem de programação *Python* pode ser feita por meio das funções *fit*, *predict* e *transform* disponíveis na classe *sklearn.cluster.KMeans* da biblioteca *Scikit-learn* (PEDREGOSA et al., 2011). O método *predict* atribui o rótulos de classe aos dados e o método *transform* expressa o espaço de dados em termos de suas distâncias aos centroides de cada grupo. O método *fit* é utilizado para ajustar os dados aos grupos, segundo um número k de grupos pré-definido pelo algoritmo denominado *k-means++* (ARTHUR; VASSILVITSKII, 2007).

A função *k-means* define aleatoriamente os centroides de cada grupo a cada iteração, incrementando o número de grupos até um número arbitrário de n grupos que se deseja testar. A cada iteração *k-means++* roda até um número máximo de vezes ou até que a soma da distância quadrada dos centroides aos outros membros dos grupos iniciais sejam as menores possíveis. Por meio deste algoritmo, para cada número de grupos variando de 2 até n são estimadas as variâncias nos respectivos agrupamentos. Este método, conhecido como método do cotovelo ou do joelho, *knee* ou *elbow method* (JAIN; DUBES, 1988), permite que se escolha um número de grupos adequado para o agrupamento final desejado.

Na análise gráfica, conforme mostra a FIGURA 30, a linha que expressa a relação entre a soma da distância quadrada e o número de grupos não apresenta diferenças significativas segundo se incrementa o número de grupos além do número k identificado. Além dos padrões de dados, ou sua matriz de similaridade, o número de grupos k, é o único parâmetro de entrada para o algoritmo em sua implementação no *Scikit-learn*. Seguindo este método, neste trabalho, foram avaliados valores de k até dez. O número de grupos k = 3 apresentou os melhores resultados para todos os casos analisados. Os gráficos com os resultados destas análises estão disponibilizadas no APÊNDICE 6.

FIGURA 30 – MÉTODO DO COTOVELO



FONTE: O Autor

Na sequência, são pontuadas algumas pesquisas sobre a utilização de sistemas de treinamentos baseados em realidade virtual aplicados ao setor elétrico, aspectos do ambiente virtual de treinamento RV2 e da atividade de substituição do isolador de pedestal.

3.2 CONTEXTO DE APLICAÇÃO

O contexto de aplicação deste trabalho é um sistema de treinamento virtual que reproduz uma subestação elétrica na qual os usuários podem interagir com objetos em cena através de um óculos de imersão 3D e manipuladores interativos. O sistema foi desenvolvido no âmbito do projeto de Pesquisa e Desenvolvimento (P&D) da Agência

Nacional de Energia Elétrica (ANEEL), denominado RV2, cujo objetivo é o desenvolvimento de um ambiente virtual para treinamento de eletricistas em atividades críticas de manutenção em subestações elétricas executado pela Companhia Paranaense de Energia (COPEL), Institutos para o Desenvolvimento Tecnológico (LACTEC) e a Universidade Federal do Paraná (UFPR).

O ambiente virtual oferece uma modalidade de treinamento guiada de uma tarefa de vinte atividades, denominada "substituição de um isolador de pedestal"(GEUS et al., 2020). Os dados de interação utilizados neste trabalho se referem às 22 sessões de treinamento disponíveis no banco de dados do sistema. Os registros contêm informações sobre o tipo, tempo e ordem das atividades realizadas em cada sessão. Uma cena do ambiente RV2 pode vista na FIGURA 31.

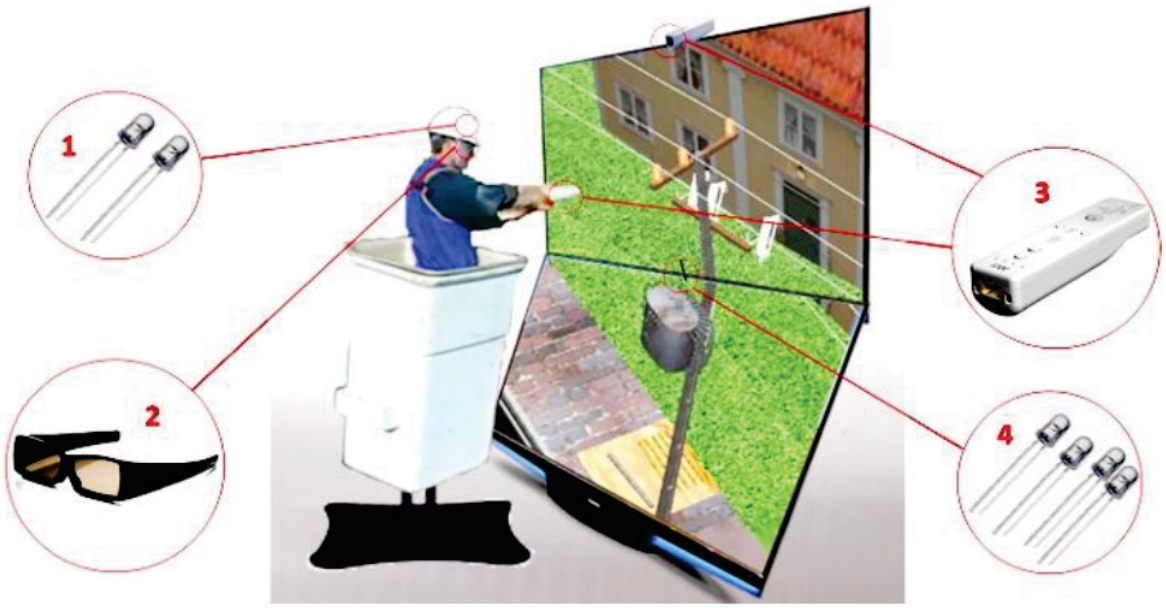
FIGURA 31 – CENA DO AMBIENTE RV2

FONTE: OneReal Research Group

Entre 2008 e 2010, em um projeto similar, o RV1, foi desenvolvida uma plataforma para treinamento de troca de cruzetas em postes em linha viva (BURIOL et al., 2009). Conforme descrito por Buriol (2011), a FIGURA 32 ilustra os componentes tecnológicos do ambiente de treinamento RV1 como (1) LEDs para rastreamento dos movimento da cabeça, (2) óculos para monitores 3D, (3) dispositivo *wiimote* para navegação no ambiente virtual, e (4) LED's para marcadores da câmera do *wiimote*.

Na proposta atual, além do desenvolvimento atualizado de um novo sistema, adaptado à demanda de treinamento em subestações, uma das questões abordadas trata da avaliação do desempenho dos eletricistas em treinamento.

FIGURA 32 – AMBIENTE DE TREINAMENTO RV1



FONTE: (BURIOL et al., 2009)

3.2.1 Trabalhos semelhantes

Assim como em outras áreas, a pesquisa e o desenvolvimento de sistemas virtuais de treinamento no setor de transmissão e distribuição de energia elétrica tem encontrado um amplo campo de aplicações (GUPTA et al., 2008a). O levantamento bibliográfico preliminar revela uma ênfase no desenvolvimento de ambientes para o treinamento em salas de controle em subestações (FOCKING et al., 2012) ou a manutenção em linhas de transmissão (GALVAN et al., 2010; PARK et al., 2006) De modo geral os trabalhos consultados fazem uma descrição do processo e das tecnologias, utilizadas na modelagem do ambiente, e das interações dos usuários (ROMERO et al., 2008; MENDES DE LIMA et al., 2013; TANAKA et al., 2015).

Aguiar et al. (2019) apresentam o resultado da análise do erro humano a partir de relatórios de falha e propõem um modelo do treinando para apoio do estudo do erro humano. Os autores investigaram as relações entre fatores que influenciam o desempenho durante a execução de tarefa como, por exemplo, condições físicas e mentais do usuário, o tipo de manobra a ser executada, as condições de execução. O modelo, alimentado com estas informações, é dinamicamente modificado conforme novo treinamento seja realizado. Neto et al. (2009) propõem estratégias de prevenção do erro humano para atividades de supervisão de painéis de controle. Scherer et al. (2010) desenvolvem um estudo de ampliação dos tipos de erro identificáveis no contexto do setor elétrico e uma taxonomia estendida é proposta. O desenvolvimento de uma ferramenta completa, integrada ao ambiente de simulação, é apresentado em Netto

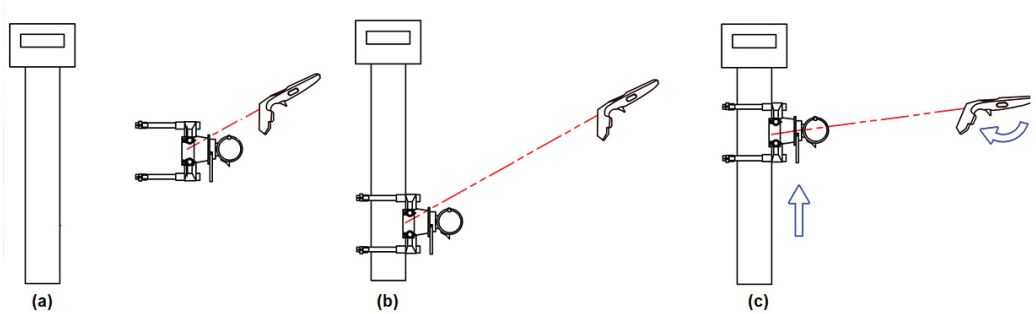
et al. (2014a), onde o sistema de avaliação mapeia o registro de interações do usuário durante a execução da tarefa e confronta os parâmetros correspondentes com um registro considerado desejável.

3.2.2 Ambiente virtual de treinamento RV2

O projeto RV2 foi desenvolvido segundo três linhas principais. As duas primeiras dizem respeito à modelagem geométrica do sistema e aos processos de interação dos treinandos. A terceira linha de desenvolvimento do projeto, na qual se inclui esta pesquisa, representa o esforço de inovação para integrar um sistema de avaliação dos treinandos ao ambiente de simulação. Além de permitir a criação de programas de treinamentos que garantam a retenção e transferência do conhecimento, as ferramentas de registro e análise de dados possibilitam a prospecção do conhecimento tácito de profissionais experientes. Nesse sentido, o mapeamento das interações do treinando com o ambiente pode revelar padrões, ou singularidades, em relação às estratégias cognitivas na resolução de problemas e tomada de decisão em atividades críticas.

O ambiente virtual de treinamento foi modelado no motor de jogos *Unreal Engine*® como recursos de interação em primeira pessoa baseada em realidade virtual. Dentro do ambiente a tarefa de troca do isolador de pedestal é decomposta em atividades e, cada atividade, é definida por meio da montagem (*attach*) de objetos (*assets*) de interatuáveis como ferramentas de manutenção e equipamentos presentes na subestação elétrica virtual (ROSENDO et al., 2019). A FIGURA 33 ilustra o conceito de manipulação e montagem de dois objetos (na figura, o colar sela e o poste do isolador) por meio de controladores externos.

FIGURA 33 – CONCEITO DA MONTAGEM DE DOIS OBJETOS POR MEIO DE MANIPULADORES DE INTERAÇÃO COM AMBIENTE



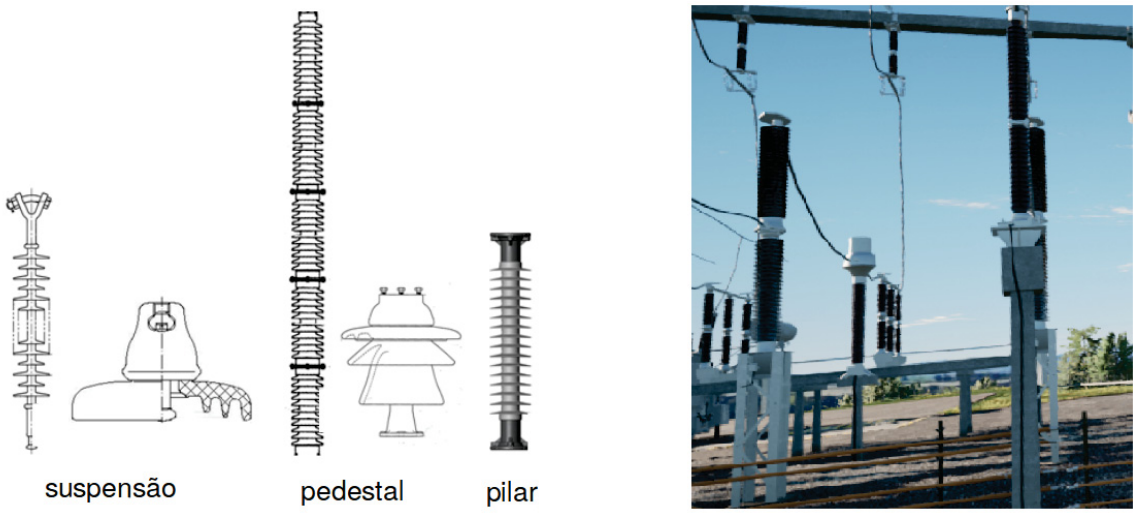
FONTE: Adaptado de Rosendo et al. (2019)
LEGENDA: Figura (a): Seleção do objeto 1; Figura (b): Seleção do objeto 2; Figura (c) Montagem dos objetos 1 e 2.

Dois objetos são definidos de forma complementar como passivos ou ativos conforme o seu papel na relação de montagem (ROSENDO et al., 2019). Um dos objetos pode assumir mais de um papel, desde que o seu correspondente seja complementar. Montagens de objetos fora desta relação de complementaridade não são possíveis, por exemplo, o treinando não pode usar uma chave de fenda para rosquear uma porca. A seguir, a atividade de substituição de isolador de pedestal é apresentada.

3.2.3 Substituição de isolador de pedestal

Isoladores são componentes cuja função é manter a distância segura entre duas estruturas com potenciais elétricos diferentes (ERBETTA, 2015). Conforme Texeira e Bazzo (COSTA TEXEIRA; BAZZO, 2013), os isoladores devem possuir resistência mecânica e elétrica para suportar e isolar os equipamentos numa subestação. Devido ao tempo de uso, manuseio, ou obsolescência tecnológica, qualquer alteração das propriedades isolantes pode comprometer a integridade de equipamentos, colocar em risco a vida de pessoas e, portanto, faz-se necessária sua substituição. Conforme ilustrado na FIGURA 34, existem diferentes tipos de isoladores presentes numa subestação.

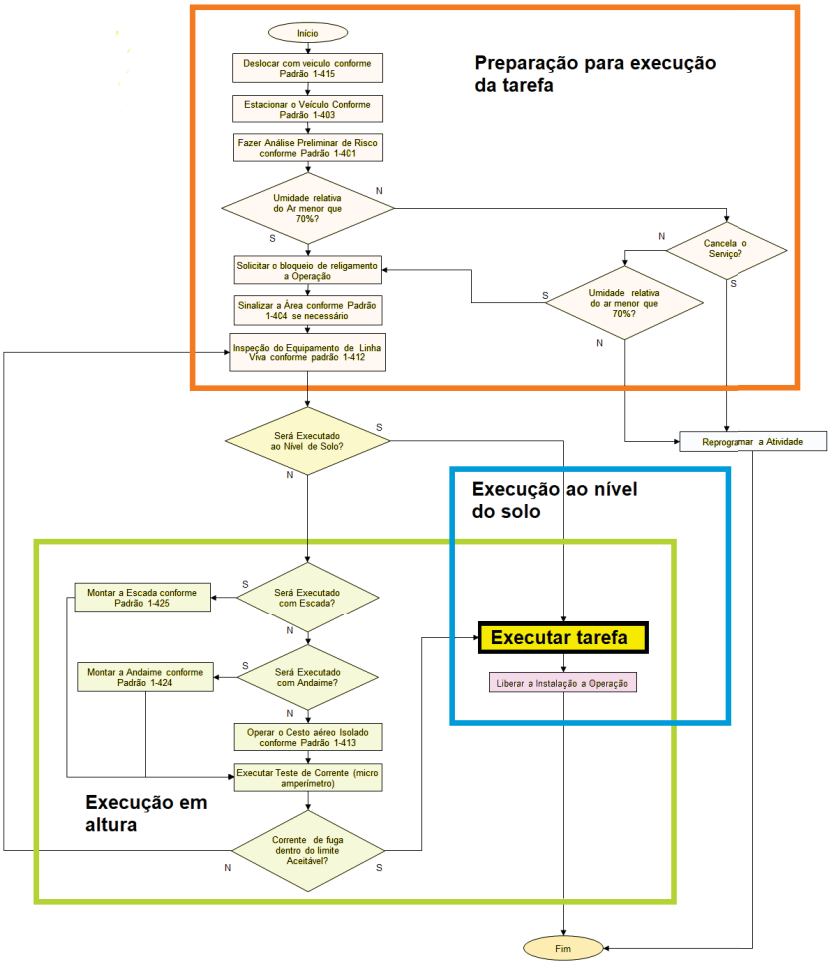
FIGURA 34 – ISOLADORES PRESENTES EM SUBESTAÇÕES



FONTE: (Esq.) (ERBETTA, 2015); (Dir.) OneReal Research Group
LEGENDA: (Esq.) Tipos de Isoladores; (Dir.) Isoladores de pedestal no RV2.

A substituição de isolador de pedestal é considerada crítica e deve ser realizada por uma equipe de, pelo menos, seis pessoas, incluindo um supervisor que não atua diretamente na execução das manobras. Os eletricistas que realizam esta tarefa devem ter completado mais de 500 horas de treinamentos em trabalhos com eletricidade (NR 10) e em altura (NR 35), tecnologia de equipamentos, operação e manobra em subestações, dentre outros (COPEL-GSST, 2020).

FIGURA 35 – TAREFA COMPLETA DA SUBSTITUIÇÃO DO ISOLADOR DE PEDESTAL



FONTE: Modificado a partir de COPEL-GSST (2020)

Conforme ilustrado no fluxograma da FIGURA 35 a substituição do isolador pode ser realizada em altitude por meio de escada, andaime ou cesto aéreo, ou ainda ao nível do solo, modalidade escolhida para ser desenvolvida no RV2. Em campo, esta tarefa envolve uma série de ações como preparação e isolamento do local, verificação e controle de variáveis climáticas e limpeza e teste de equipamentos que podem ocorrer antes, durante e depois da execução da tarefa.

No manual de “Procedimentos Básicos para Manutenção de Subestações com Técnicas de Linha Viva” (COPEL, sd.), a substituição de isolador de pedestal é descrita por meio de cinco conjuntos de atividades, doravante definida como manobras :

1. **Montagem do suporte**

   a) Instalar um bastão garra Ø 64 x 3600 mm na barra junto a estrutura com o pedestal a ser substituído;

b) Amarrar o bastão na parte superior da estrutura com uma corda;
c) Amarrar o olhal do bastão ao cavalo da sela através de uma corda.

2. **Montagem do mastro**

a) instalar um segundo conjunto, bastão e sela, no lado oposto da estrutura, este com o bastão invertido (com a porca olhal para cima) e com a carretilha e corda de serviço na mesma;
b) “Enforcar” um estropo de náilon no corpo do isolador através de ganchos espirais instalados em bastões universais;
c) Colocar a extremidade do estropo no gancho da corda de serviço.

3. **Desconexão da barra**

a) Com auxílio de uma chave com catraca em um bastão universal e soquete adequado, afrouxar os parafusos do conector que prende a barra ao isolador no lado vivo;
b) Terminar de retirar os parafusos com o bastão com soquete multiangular;
c) Elevar levemente a barra através da corda que une a sela ao bastão;
d) Apertar o colar através da porca borboleta;
e) Segurar firmemente a corda de serviço.

4. **Substituição da coluna**

a) Cuidadosamente, e sem invadir o isolamento da coluna, sacar os parafusos da base do isolador;
b) Baixar o isolador ao solo; Içar o isolador com o estropo na mesma posição que se encontrava no isolador anterior;
c) Cuidadosamente, e sem invadir o isolamento da coluna, colocar os parafusos na base do isolador;
d) Baixar a barra para que ela se apoie no novo isolador;
e) Com auxílio de um bastão universal e locador de pino colocar a parte superior do conector;
f) Com um sacador e colocador de pino, em um bastão universal ou com o próprio bastão com soquete multiangular, colocar os parafusos no lado vivo;
g) Com auxílio de uma chave com catraca em um bastão universal e soquete adequado, apertar os parafusos do conector que prende a barra ao isolador no lado vivo.

5. **Conclusão**

a) Retirar o estropo;
b) Retirar os bastões garra;
c) Retirar as selas;

d) Recolher o material, ferramentas e equipamentos.

De acordo com COPEL (sd.), na execução dessa atividade, quando possível, deve-se empregar o “método à distância”. Esse método é considerado mais seguro, pois o eletricista não entra no potencial elétrico de um condutor energizado. Em qualquer caso, a disponibilidade e utilização de equipamento de proteção é obrigatória para todos. Os riscos e ações preventivas descritas na documentação disponível são apresentadas na TABELA 6.

TABELA 6 – SUBSTITUIÇÃO DE ISOLADOR DE PEDESTAL: RISCOS E AÇÕES PREVENTIVAS

Riscos	Ações Preventivas
Choque por contato	Manter distância segura do potencial de terra e das fases adjacentes conforme IAP-040412.
Lesão por queda do isolador	Executar tarefa em equipe, com supervisão e manter atenção,
	Utilizar bastões compatíveis com a atividade.
	Alertar colaboradores do risco de que do material/ferramental, evitar o trânsito de pessoas sob a área de trabalho.
Queda do eletricista	Usar conjunto de segurança para trabalho em altura.
	Utilizar técnicas de escala segura.
Lesão por quebra de materiais	Fazer o aperto de conexões considerando a suportabilidade do componente.
	Inspecionar previamente o componente
Lesão por corte no manuseio do isolador	Inspecionar previamente as condições do isolador, manuseando com luvas.
	Atenção nas partes danificadas do isolador.

FONTE: Adaptado de COPEL (sd.) e COPEL-GSST (2020)

Além desses riscos, a retirada e instalação do isolador ou sustentação do condutor representa um risco ergonômico podendo causar entorses. O manuseio incorreto ou a queda do condutor pode causar choque e curto-circuitos nos equipamentos. Por último, um condutor que apresente falhas pode estar energizado implicando risco de arcos elétricos ou explosão do isolador.

4 MODELAGEM DE PADRÕES

Dentre as manobras enumeradas anteriormente, os dados de treinamento disponíveis no banco de dados do RV2 referem-se a execução das manobras 1, 2, e 3(a). Assim, a tarefa de substituição do isolador de pedestal fica definida pelas atividades descritas na TABELA 7.

TABELA 7 – ATIVIDADES DA TAREFA

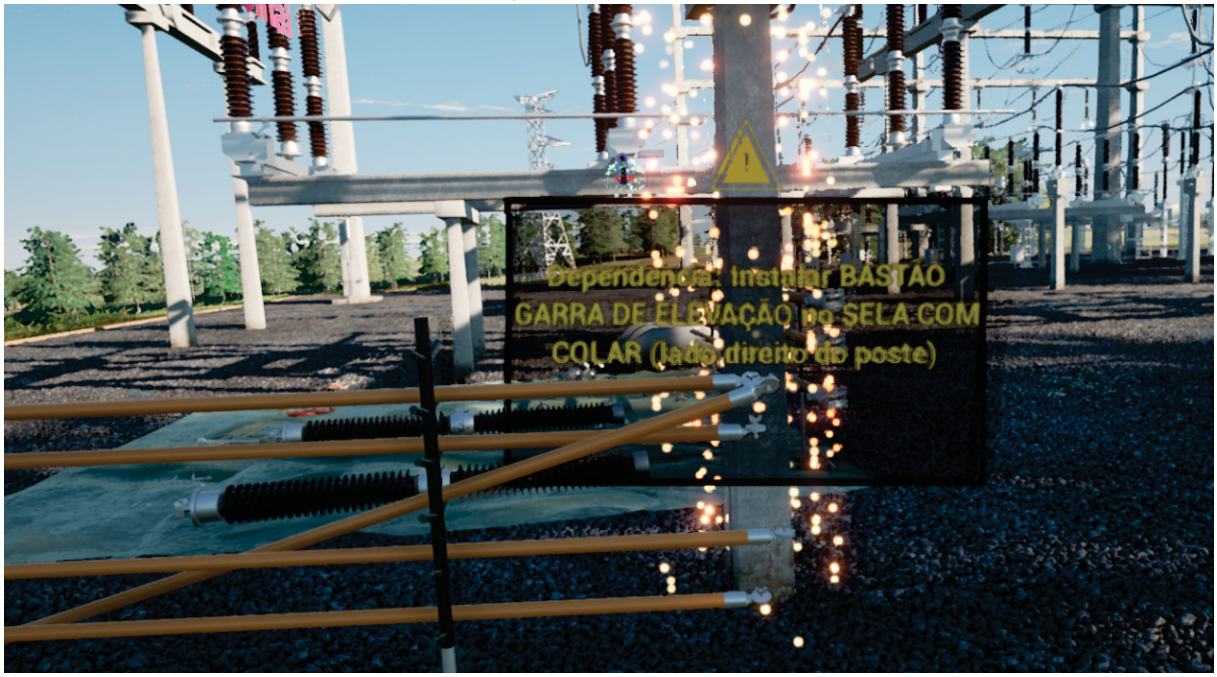
Tarefa	Manobra	ID	Descrição da atividade
Substituição de Isolador de Pedestal	Montagem do suporte e do mastro	1	Instalar SELA COM COLAR no PEDESTAL DO ISOLADOR PEDESTAL (lado direito do pedestal)
		2	Instalar BASTÃO GARRA DE ELEVAÇÃO no SELA COM COLAR (lado direito do pedestal)
		4	Instalar SELA COM COLAR no PEDESTAL DO ISOLADOR PEDESTAL (lado esquerdo do pedestal)
		5	Instalar BASTÃO GARRA DE ELEVAÇÃO em posição vertical no SELA COM COLAR (lado esquerdo do pedestal)
		7	Fixar CORDA em uma CARRETILHA na ponta do BASTÃO GARRA DE ELEVAÇÃO
	Amarração do corpo do isolador	111	Prender ESTROPO 1 em uma ESPINA
		12	Instalar ESTROPO 1 na parte de cima do ISOLADOR DANIFICADO usando o BASTÃO UNIVERSAL com ESPINA
		112	Prender ESTROPO 2 em uma ESPINA
		13	Instalar ESTROPO na parte de baixo do ISOLADOR DANIFICADO usando o BASTÃO UNIVERSAL com ESPINA
	Desconexão do isolador	14	Instalar CHAVE CATRACA em um BASTÃO UNIVERSAL
		15	Instalar ESPINA em um BASTÃO UNIVERSAL
		161	Usar BASTÃO UNIVERSAL com CHAVE CATRACA para retirar PORCA 1 DO CONECTOR
		162	Usar BASTÃO UNIVERSAL com CHAVE CATRACA para retirar PORCA 2 DO CONECTOR
		163	Usar BASTÃO UNIVERSAL com CHAVE CATRACA para retirar PORCA 3 DO CONECTOR
		164	Usar BASTÃO UNIVERSAL com CHAVE CATRACA para retirar PORCA 4 DO CONECTOR
		171	Usar BASTÃO UNIVERSAL com ESPINA para soltar PARAFUSO 1 DO CONECTOR
		172	Usar BASTÃO UNIVERSAL com ESPINA para soltar PARAFUSO 2 DO CONECTOR
		173	Usar BASTÃO UNIVERSAL com ESPINA para soltar PARAFUSO 3 DO CONECTOR
		174	Usar BASTÃO UNIVERSAL com ESPINA para soltar PARAFUSO 4 DO CONECTOR
		18	Usar BASTÃO UNIVERSAL com ESPINA para retirar CONECTOR do topo do ISOLADOR DANIFICADO

FONTE: OneReal Research Group

O campo ID refere-se ao identificador da atividade no banco de dados. No banco de dados original as atividades 111 e 112 são indicadas apenas como 11

repetido duas vezes. O mesmo acontece com as atividades 161 a 164 e 171 a 174, que aparecem como 16 e 17, respectivamente, repetidos 4 vezes cada. Neste trabalho adotou-se a perspectiva de que estas atividades são aplicadas sobre objetos diferentes e, embora sejam descritas da mesma maneira, cada ocorrência deve ser considerada separadamente. Além disso, as atividades foram simplificadas de modo que pudessem ser executadas por um único treinando.

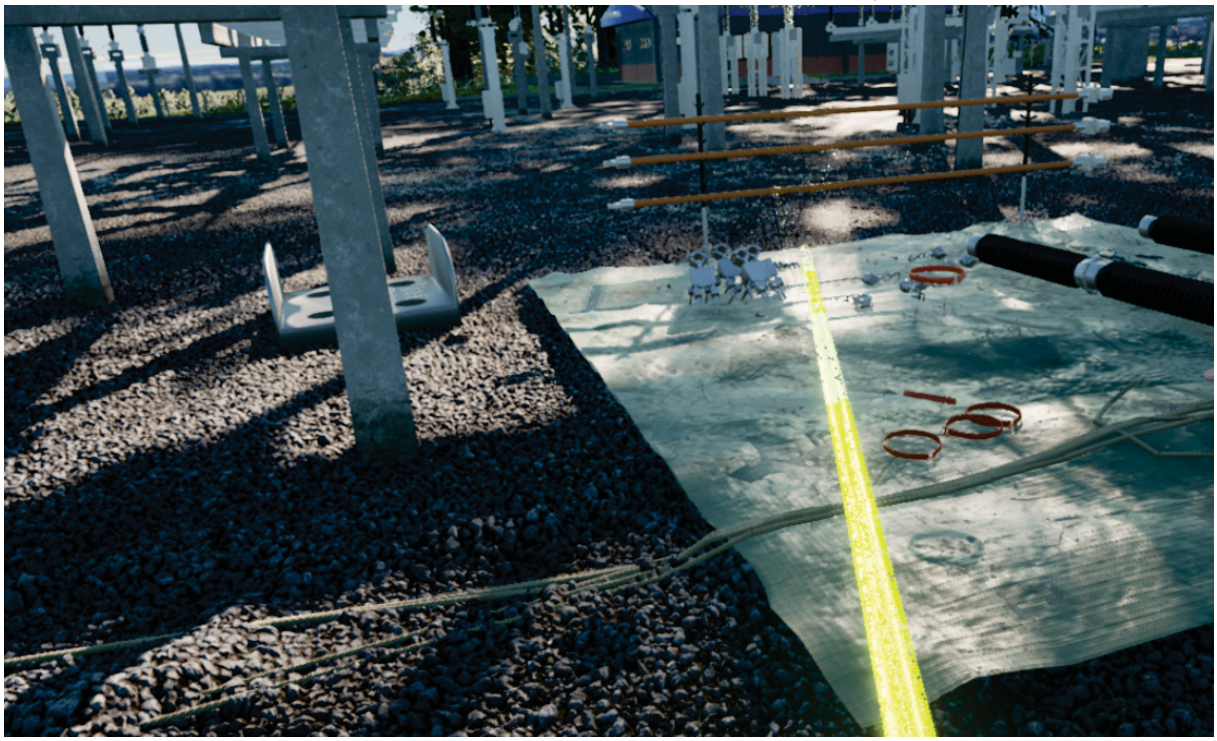
FIGURA 36 – SINALIZAÇÃO DE OBJETO A SER MANIPULADO



FONTE: OneReal Research Group

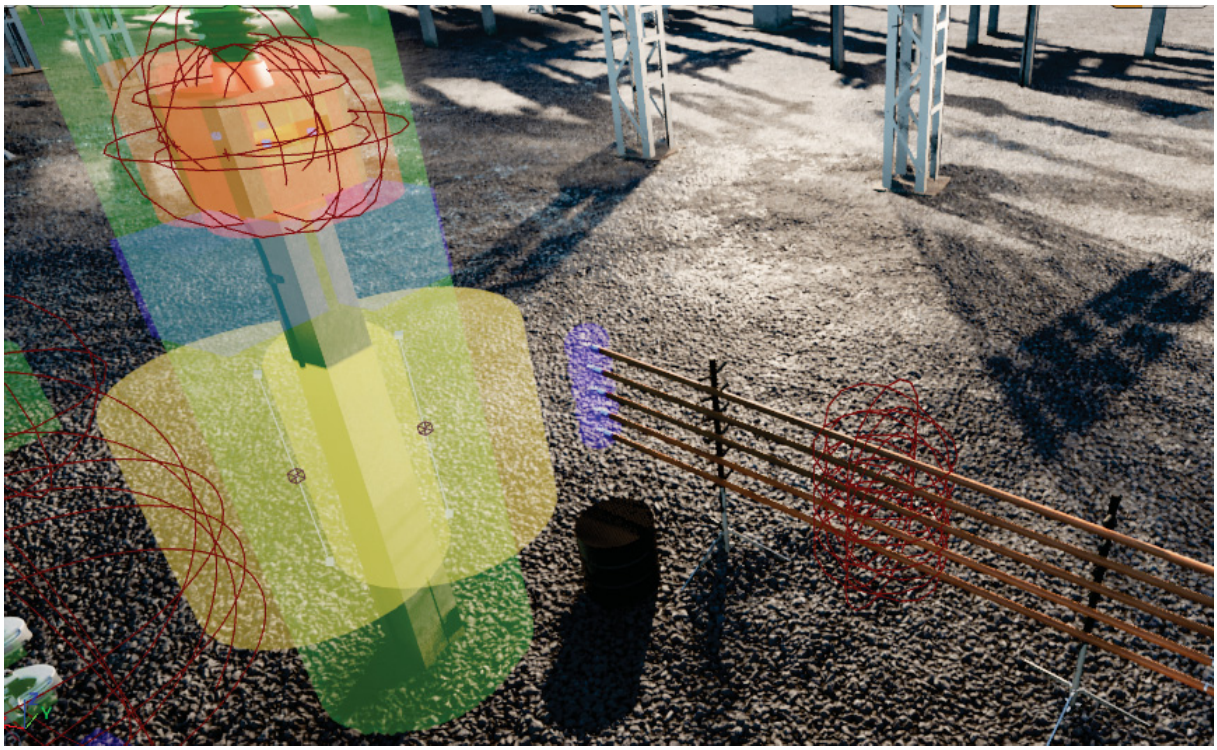
De modo geral, cada atividade compreende a identificação, manipulação e combinação de objetos, ou *assets* na linguagem da *Unreal Engine* ®. Estes objetos estão identificados com letras maiúsculas no campo Descrição. A identificação dos objetos é auxiliada por uma sinalização sobre o objetos como indicado na FIGURA 36. A manipulação dos objetos é feita por meio dos controles manuais (*joystick*) e auxiliada por um "apontador"que ajuda na seleção, movimentação e combinação com outros objetos. A combinação de objetos é realizada por meio da associação de geometrias primitivas auxiliares como cilindros, prismas e esferas. Os objetos são configurados como passivos ou ativos e são combinados no ambiente por meio de conectores, ou *snaps*, na forma de pontos, retas ou planos das geometrias auxiliares .

FIGURA 37 – RECURSO VIRTUAL DE AUXILIO A SELEÇÃO E OBJETOS



FONTE: OneReal Research Group

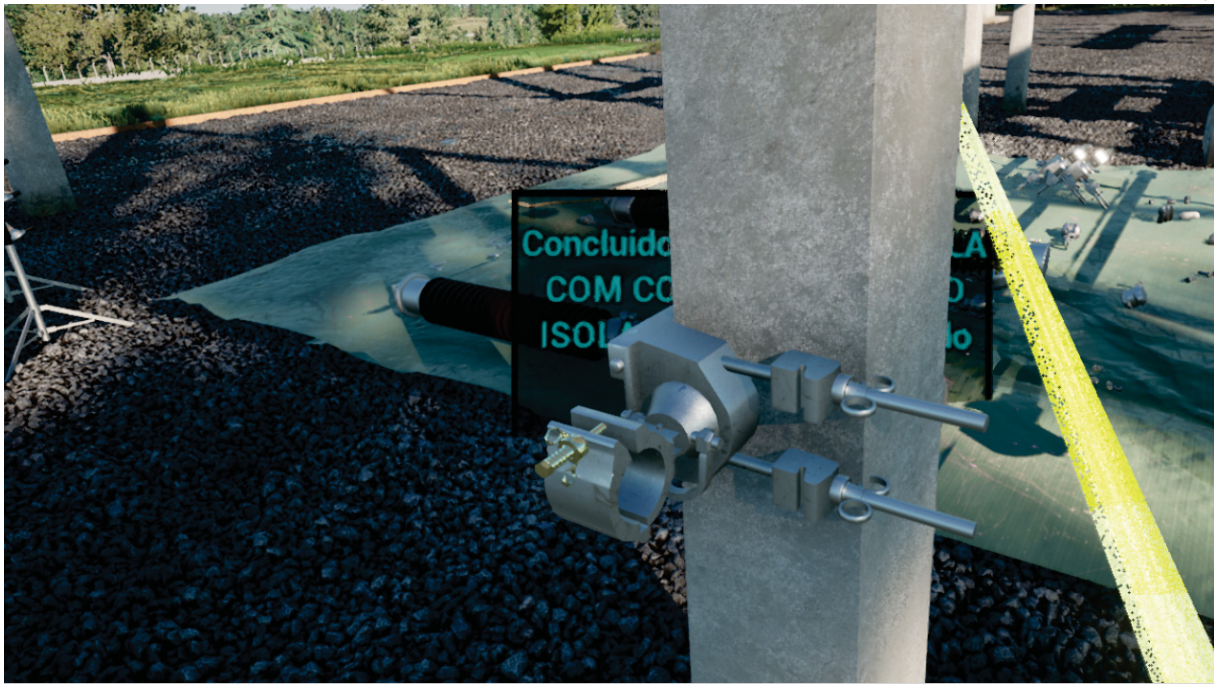
FIGURA 38 – GEOMETRIAS AUXILIARES UTILIZADAS NAS COMBINAÇÕES DOS OBJETOS



FONTE: OneReal Research Group

Ao final de uma atividade completa, dois objetos devem estar unidos. Na FIGURA 39 a primeira atividade da TABELA 7, montagem completa da sela com colar no pedestal do isolador é ilustrada.

FIGURA 39 – ATIVIDADE 1 (MONTAGEM SELA COM COLAR X PEDESTAL DO ISOLADOR)



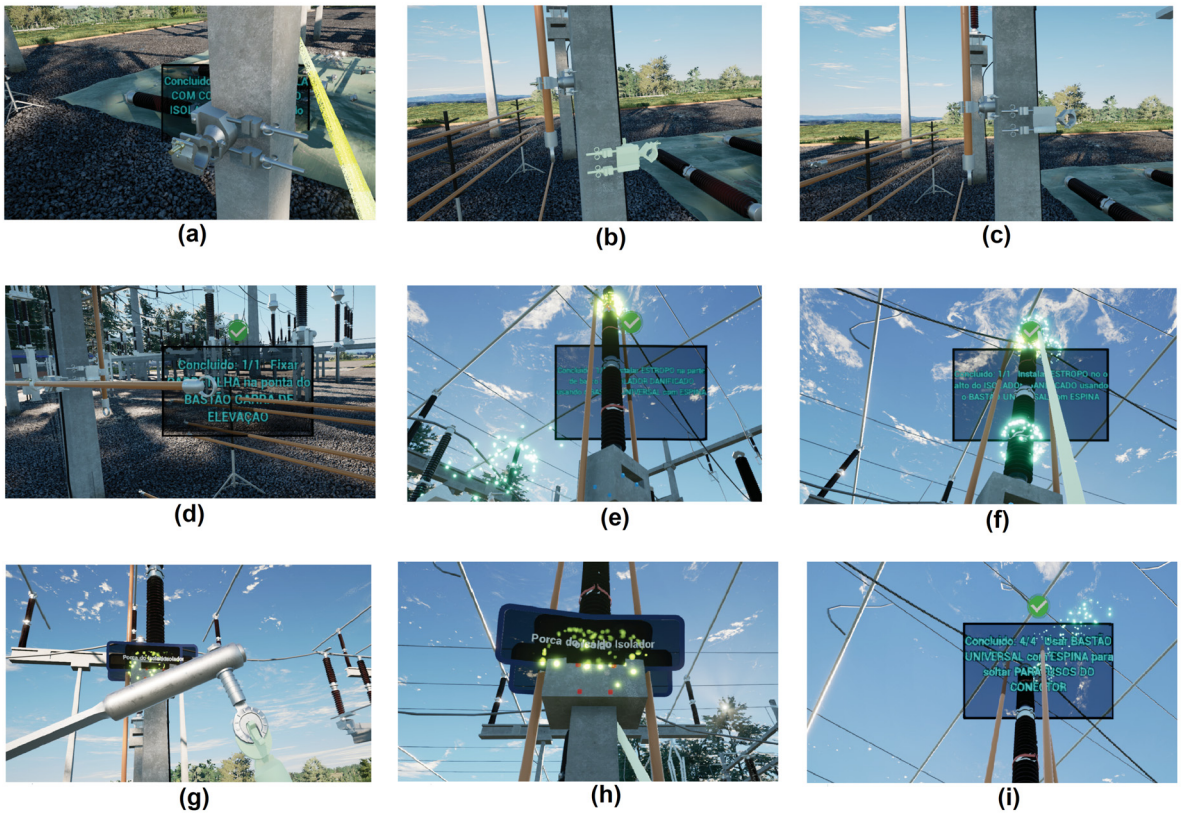
FONTE: OneReal Research Group

Outras atividades da TABELA 7 são apresentadas na FIGURA 40. Na seção seguinte são discutidos os aspectos da modelagem do conhecimento e do pré-processamento dos dados de interação dos treinandos na forma de padrões de erros e padrões de similaridades.

4.1 MODELAGEM DO CONHECIMENTO BASEADO EM REGRAS

Os dados de interação dos treinandos durante as sessões de treinamento foram registrados num banco de dados em linguagem SQLite. O banco de dados é organizado em sete tabelas (FIGURA 41). Os dados sobre a execução da tarefa pelos treinandos está registrado na tabela LogAtividade, emoldurada na figura.

FIGURA 40 – SEQUÊNCIA DE EXECUÇÃO DA TAREFA



FONTE: OneReal Research Group

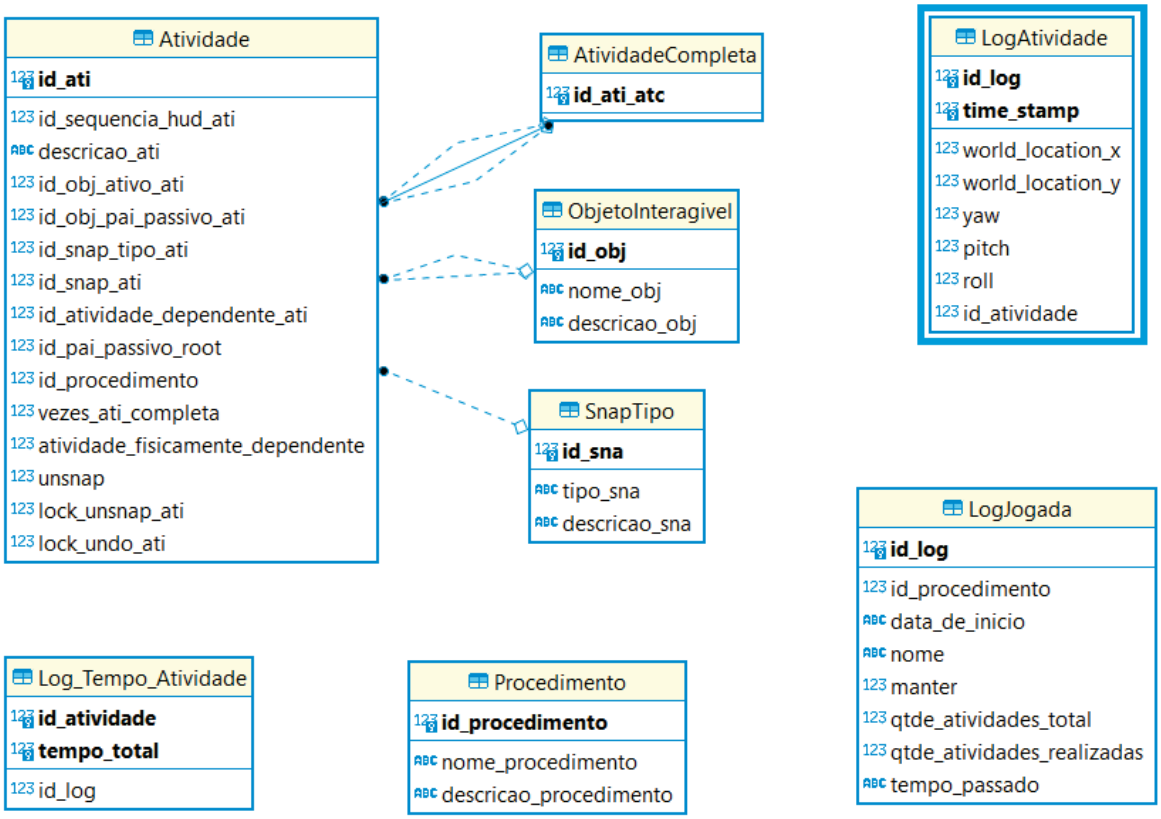
LEGENDA: (a): Instalação de SELA COM COLAR no PEDESTAL DO ISOLADOR PEDESTAL, no lado direito do pedestal, em execução.(b): Instalação de SELA COM COLAR no PEDESTAL DO ISOLADOR PEDESTAL no lado esquerdo do pedestal finalizada. (c): Instalar SELA COM COLAR no PEDESTAL DO ISOLADOR PEDESTAL no lado esquerdo do pedestal. (d): Instalar BASTÃO GARRA DE ELEVAÇÃO em posição vertical no SELA COM COLAR no lado esquerdo do pedestal. (e): Instalar ESTROPO no corpo do ISOLADOR DANIFICADO usando o BASTÃO UNIVERSAL com ESPINA, em execução. (f): Instalar ESTROPO no corpo do ISOLADOR DANIFICADO usando o BASTÃO UNIVERSAL com ESPINA, finalizado. (g): Instalar CHAVE CATRACA em um BASTÃO UNIVERSAL. (h): Usar BASTÃO UNIVERSAL com CHAVE CATRACA para retirar PORCA DO CONECTOR. (i): Usar BASTÃO UNIVERSAL com ESPINA para soltar PARAFUSO DO CONECTOR.

A tabela LogAtividade contém os registros de 8 variáveis para cada treinando (TABELA 8).

Cada uma das variáveis são descritas como:

- id_log : identificação do treinando;

FIGURA 41 – DIAGRAMA UML DO BANCO DE DADOS RV2



FONTE: O Autor

TABELA 8 – LINHAS DE DADOS DE INTERAÇÃO DO TREINANDO ID 02

id_log	*world_location_x*	*world_location_y*	*yaw*	*pitch*	*roll*	*time_stamp*	*id_atividade*
02	-2544.736084	-3093.217529	78.149704	-14.958127	-0.46759	11.754851	1
02	-2563.237793	-3147.081055	74.63694	-5.07397	-0.781067	17.765175	4

FONTE: O Autor

- $time\_stamp$: registro do tempo de execução da atividade;
- $world\_location\_x$, : coordenada **x** do treinando no momento da conclusão da atividade em relação ao plano cartesiano que define a posição de um objeto na cena do ambiente;
- $world\_location\_y$: coordenada **y** do treinando no momento da conclusão da atividade em relação ao plano cartesiano que define a posição de um objeto na cena do ambiente;
- yaw: posição de rotação em torno do eixo **z** da cabeça modelo virtual da câmera em primeira pessoa;
- $pitch$: posição de rotação em torno do eixo **x** da cabeça do modelo virtual da câmera em primeira pessoa;

- $roll$: posição de rotação em torno do eixo **y** da cabeça do modelo virtual da câmera em primeira pessoa;
- $id\_atividade$: identificação da atividade realizada.

As variáveis associadas a identificação do treinando, qual atividade, em que ordem e tempo foi realizada são: id_log, $time\_stamp$ e $id\_atividade$. Os dados relativos a estas variáveis para cada treinando foram sumarizados por meio da TABELA 9.

TABELA 9 – SEQUÊNCIAS DE ATIVIDADES EXECUTADAS PELOS TREINANDOS

ID.tr.	Total At. / Temp. Total	Atex_1	Atex_2	Atex_3	Atex_4	Atex_5	Atex_6	Atex_7	Atex_8	Atex_9	Atex_10	Atex_11	Atex_12	Atex_13	Atex_14	Atex_15	Atex_16	Atex_17	Atex_18	Atex_19	Atex_20	Atex_21	Atex_22	Atex_23
1	14	1	2	4	5	7	4	5	2	111	112	15	15	13	14	14	14	161	162	171	12	172	5	
	330	31,10	8,35	8,05	6,25	7,41	70,25	4,83	5,90	11,62	5,93	30,97	16,39	11,74	25,63	2,49	3,03	12,62	0,17	24,01	0,54	0,04	10,66	
2	3	4	1	5																				
	686	496,05	56,46	116,17																				
3	4	1	4	5	2	2	2	2																
	697	29,42	88,87	31,30	36,76	227,68	187,45	36,11																
4	5	1	4	2	2	4	5	161	162	163	164													
	640	23,54	74,64	63,00	21,74	198,72	4,04	242,04	1,01	0,13	0,11													
5	10	1	4	5	5	5	2	14	161	162	163	164	111	112	15	12								
	596	35,38	11,83	14,54	70,70	79,42	13,63	34,05	87,91	0,02	1,30	0,22	81,38	11,56	24,16	21,76								
6	11	1	4	2	5	2	111	112	15	12	14	161	162	163	164	7	15	13						
	953	44,85	18,70	20,94	35,99	122,74	35,25	7,14	41,91	46,86	50,38	32,61	0,10	0,23	0,07	70,12	138,15	29,96						
7	9	1	4	5	2	14	161	111	15	112	12	15	13	162	163	164								
	675	20,79	19,54	21,51	17,33	70,85	135,84	33,09	23,40	19,45	14,17	105,79	20,29	82,52	0,49	0,09								
8	4	4	4	1	4	5	2	2	2	2														
	677	79,81	98,21	22,48	178,85	38,26	39,23	27,24	42,26	41,42														
9	14	1	14	161	162	163	15	171	172	18	111	13	112	12										
	208	28,48	37,49	11,03	0,08	0,03	32,47	16,81	0,11	0,56	13,40	5,02	9,22	7,93										
10	5	1	4	7	5	2	14	161	162	163	164	15	15	111	112	13	15	12	171	18	172			
	637	16,52	5,46	5,53	21,06	8,18	13,19	21,25	0,03	0,06	0,01	9,77	13,27	7,51	1,99	6,15	4,94	11,44	14,79	0,10	2,38			
11	14	1	4	5	7	2	2	7	15	111	112	12	13	14	161	162	163	164	15					
	320	27,14	7,24	6,57	20,17	31,02	9,48	57,37	32,60	8,34	29,30	25,42	35,85	18,10	49,93	0,06	0,05	3,22	89,59					
12	14	4	1	5	5	1	2	7	15	111	112	12	13	14	161	162	163	171	172	173	174	18		
	644	28,41	5,56	7,61	27,84	22,99	2,68	15,96	14,84	15,11	4,69	11,89	19,14	10,42	15,55	0,10	0,03	37,08	1,34	0,41	0,51	24,75		
13	12	1	4	1	2	5	7	15	15	111	112	12	13	161	162	171	172	173	174	163	164	18	14	
	618	20,93	18,43	5,13	24,80	34,31	11,65	28,49	37,70	31,14	17,63	34,82	23,13	73,86	9,86	25,80	8,00	0,09	2,70	45,66	56,13	26,51	32,74	
14	14	1	4	2	5	7	15	14	111	112	13	12	161	162	163	164	171	172	18	173	174	5	2	1
	286	17,56	8,58	7,00	5,08	72,99	11,16	13,04	5,14	22,29	10,51	18,27	66,62	20,61	0,03	0,04	60,72	0,09	0,16	0,38	0,09	14,96	6,60	41,08
15	1	4	1	2	5	15	111	15	112	7	12	15	13	14	161	162	163	171	172	173				
	610	13,90	4,52	3,55	3,92	13,37	4,83	5,39	5,86	5,93	21,11	24,53	33,81	20,08	39,49	5,96	0,03	40,36	0,02	0,03				
16	13	4	1	2	5	7	15	15	14	111	112	12	13	161	162	171	172	18	2					
	305	15,58	9,20	9,66	20,64	10,67	6,88	4,39	5,19	3,44	10,93	13,70	9,02	22,12	0,62	101,24	0,74	0,02	1,31					
17	14	1	4	1	2	5	7	15	15	14	111	112	14	12	13	161	162	171	172	173	174	18		
	515	27,56	12,79	26,13	6,54	11,85	12,54	9,24	17,33	7,87	11,18	15,64	50,49	42,71	96,76	45,48	0,02	21,70	0,02	1,81	0,07	7,46		
18	14	1	4	5	2	15	15	111	112	12	15	13	7	161	18	171	162	15	14	15				
	571	17,06	6,36	20,61	13,48	21,55	6,28	14,81	8,14	57,38	58,20	50,05	18,64	58,25	1,51	1,38	3,62	40,21	74,89	39,47				
19	13	4	1	5	2	15	15	111	112	12	13	7	14	161	161	171	172	173	18	15	5			
	559	12,80	5,27	7,40	4,52	8,90	35,00	8,56	3,06	19,77	55,77	14,15	34,98	32,03	0,07	20,00	0,62	4,04	0,02	41,06	11,35			
20	14	1	4	1	5	2	7	2	14	15	111	112	15	12	13	161	162	163	164	171	18	5		
	453	27,43	11,28	2,70	12,32	9,78	139,67	15,39	12,30	10,47	9,58	10,98	8,42	44,47	61,39	22,06	0,17	0,13	0,67	16,56	0,36	20,29		
21	14	1	4	5	2	5	7	15	111	112	15	12	13	14	14	15	161	162	163	164	171	172	173	18
	914	100,24	84,06	38,37	51,19	15,19	32,29	35,47	17,62	4,89	27,94	133,70	20,67	27,50	11,34	22,81	81,64	0,43	0,27	16,75	78,19	1,68	0,43	0,13
22	13	1	4	5	2	15	15	111	112	12	13	14	161	162	163	164	171	18	172	7				
	509	12,75	8,85	9,02	14,37	12,94	28,95	37,30	9,43	41,30	67,67	16,92	93,63	0,02	0,11	45,33	0,13	0,02	0,24	16,08				

FONTE: O Autor

Na primeira coluna, está o campo de identificação do treinando. Para cada treinando são reservadas duas linhas: na linha superior estão descritas as atividades executadas e na linha inferior o tempo de execução da tarefa. Dessa forma, a segunda coluna resume o número e o tempo total de atividades, e as colunas seguintes identificam a ordem e o tempo em que cada atividade foi executada.

De acordo com a tabela $Atividade$ do banco de dados apresentada na FIGURA 41, as atividades são definidas, dentre outros, pelo seu identificador, pela sua descrição e pelos objetos (ativo e passivo) que devem ser montados segundo um determinado tipo de conexão (*snap*). Além disso um vetor extraído do campo $id\_atividade\_dependente\_ati$ fornece uma ordenação explícita em que as atividades devem ser executadas. Ainda assim, os treinandos podem executar as tarefas em qualquer ordem, aumentando dessa forma a possibilidade de execução de uma atividade foram da ordem sugerida.

Conforme análise das tarefas descritas na TABELA 7, outras restrições foram adicionadas, incluindo uma atividade virtual, ID 0, representando o início da sessão de treinamento. Assim, como apresentado na TABELA 10, para cada atividade, existem uma atividades predecessoras e sucessoras. A precedência (subsequência) aqui indica que a atividade atual deve ser realizada depois (antes) de alguma das atividades predecessoras (sucessoras). Como cada atividade deve ser executada apenas uma vez, uma atividade executada sai da lista de antecessoras e sucessoras das outras.

TABELA 10 – DEPENDÊNCIA DAS ATIVIDADES

Atividade ID	Dependências	
	Atividades Predecessoras	Atividades Sucessoras
1	0	2, 4, 5
2	1, 4, 5, 7	1, 4, 5, 7, 111, 112
4	0	1, 2, 5
5	1, 2, 5, 7	1, 2, 4, 7, 111, 112
7	2, 5	1, 2, 4, 5, 111, 112
111	2, 5, 7	12, 13
12	111, 112	111, 112, 14
112	2, 5, 7	12, 13
13	111, 112	111, 112, 14
14	13	15, 161
15	14, 161, 162, 162, 163, 164	161, 162, 163, 164, 171
161	14, 15	15, 162, 171
162	15, 161	15, 163, 171, 172
163	15, 162	15, 164, 171, 172, 173
164	15, 163	15, 171, 172, 173, 174
171	15, 161	15, 162, 163, 164, 172
172	15, 162, 171	15, 163, 164, 171
173	15, 163, 172	15, 164, 172
174	164, 173	18
18	174	

FONTE: O Autor

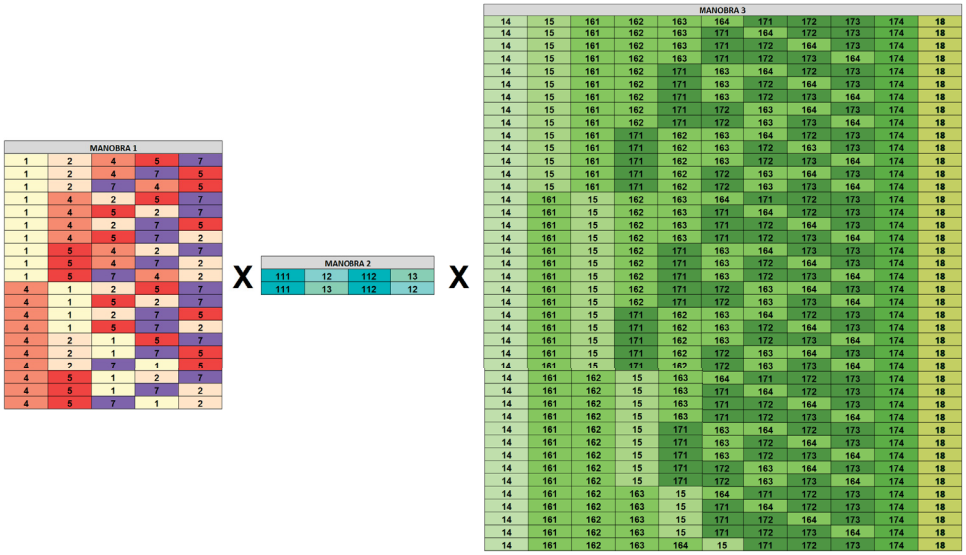
A partir da TABELA 10 foram geradas as sequências de atividades que satisfazem as dependências identificadas por suas atividades sucessoras ou antecessoras. Estas sequências correspondem a um padrão desejado para a execução da tarefa e são consideradas como uma classe padrão de sequências a qual corresponde um desempenho ótimo.

Para obter estas sequências, inicialmente, foram geradas todas as sequências que correspondem a permutação das vinte atividades num vetor de vinte posições. Do resultado destas permutações foram excluídas todas as sequências que não satisfaziam as condições assinaladas pelos sucessores e antecessores de uma atividade. Todos os procedimentos para geração das sequências foram implementados em linguagem

Python.

As sequências foram particionadas em conjunto de atividades divididas por manobras conforme a divisão já indicada na TABELA 7 e ilustrado na FIGURA 42. O conjunto de todas as sequências são obtidas a partir do produto cartesiano das manobras 1, 2 e 3, totalizando 1680 sequências viáveis.

FIGURA 42 – SEQUÊNCIAS DE CAMINHOS ESPERADOS



FONTE: O Autor

Este conjunto de sequências tem as seguintes restrições:

- Todas as sequências contêm as vinte atividades diferentes;
- Cada atividade é executada uma única vez;
- Uma atividade executada sai da lista de predecessores e sucessores das outras atividades.

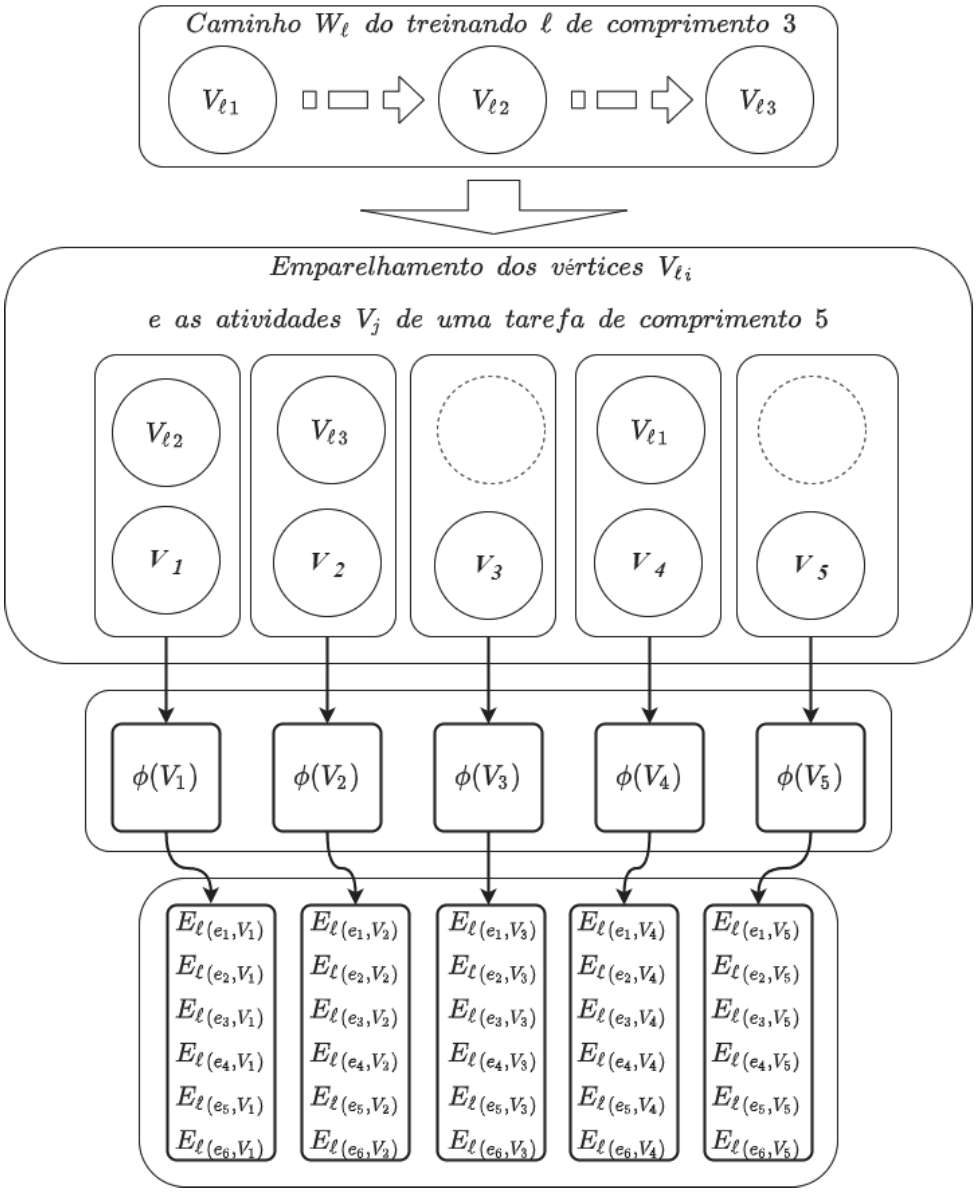
Na sessão seguinte o domínio do conhecimento é modelado como um espaço de erros associados às violações das regras de execução da tarefa. Cada treinando deve ser descrito por um vetor cujas entradas representam os erros cometidos durante a execução da tarefa na sua respectiva sessão.

4.1.1 Modelagem dos padrões de erro

A partir da revisão apresentada no capítulo da Fundamentação Teórica e, em particular, com base em (SCHERER et al., 2010), os erros cometidos por trabalhadores

profissionais na manutenção de linhas vivas foram identificados e mapeados como um vetor erro para cada passeio W_ℓ executado pelo treinando ℓ. A FIGURA 43 ilustra este processo para uma tarefa de cinco atividades e um passeio de um treinando que tenha executado três atividades.

FIGURA 43 – EXTRAÇÃO DOS ERROS DOS CAMINHOS DOS TREINANDOS



FONTE: O Autor

As entradas do vetor erro são componentes de sinal da ocorrência de seis tipos de erros na execução das atividades da tarefa:

- (e_1) Execução em Tempo Inadequado: a execução de uma atividade em tempo inadequado é definida relativamente aos tempos dos treinandos para a mesma atividade.[1]
- (e_2) Repetição: a atividade é executada mais de uma vez na mesma manobra;
- (e_3) Inclusão: a atividade é executada fora da sua respectiva manobra;
- (e_4) Inversão: a ordem das atividades está invertida na manobra;
- (e_5) Incompletude: atividades sucessoras esperadas não executadas;
- (e_6) Omissão: a atividade não é executada.

A partir destes tipos de erros define-se um vetor do erro

$$\{e_i\} = \{i\}, \tag{4.1}$$

com $i = 1, 2, \cdots, 6$ e uma função de extração do erro dada por

$$\phi(V_{\ell j}) = \vec{e}', \tag{4.2}$$

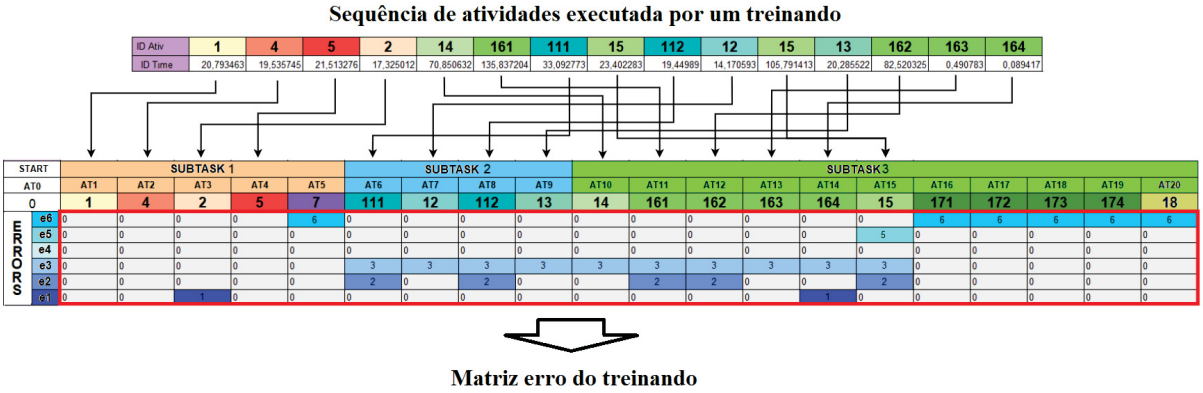
onde a j-ésima posição de $\vec{e}$,

$$\vec{e_j} = \begin{cases} i, \ se \ e_j \ afeta \ V_j \ de \ \ell \\ 0, \ caso \ contrário \end{cases} \tag{4.3}$$

A FIGURA 44 ilustra o processo de extração do padrão de erro a partir da sequência que representa a execução da tarefa por um treinando. Na primeira e segunda linhas da tabela correspondem a sequência de atividades do caminho do treinando e os tempos de execução de cada atividade, respectivamente. As três linhas seguintes referem-se a identificação das manobras e atividades a serem executadas. A quinta linha não representa um caminho, embora a sequência de atividades apresentada seja uma instância de um caminho esperado. Nas últimas seis linhas, para cada atividade indicadas nas colunas, estão assinalados os erros mapeados segundo as definições dadas em 4.1.1.

[1] Conforme opinião de especialistas, por tratar-se de uma atividade crítica, a favor de uma execução segura, o tempo é um fator menos importante dentre os outros apresentados. De qualquer forma, como a média de tempo total para execução da tarefa é dez minutos, a distribuição de tempo por tarefa permitiu estimar um tempo aceitável que pode ser definido como sendo inferior ao terceiro quartil dos tempos de todas as execuções da mesma tarefa.

FIGURA 44 – MAPEAMENTO DE ERROS EM UM CAMINHO



FONTE: O Autor

Para cada treinando ℓ a atribuição do erro leva a uma matriz de erros do treinando (Equação 4.4)

$$E_\ell = \begin{bmatrix} \mathbf{E}_{\ell(e_1,V_1)} & \mathbf{E}_{\ell(e_1,V_2)} & \cdots & \mathbf{E}_{\ell(e_1,V_N)} \\ \mathbf{E}_{\ell(e_2,V_1)} & \mathbf{E}_{\ell(e_2,V_2)} & \cdots & \mathbf{E}_{\ell(e_2,V_N)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{E}_{\ell(e_6,V_1)} & \mathbf{E}_{\ell(e_6,V_2)} & \cdots & \mathbf{E}_{\ell(e_6,V_N)} \end{bmatrix}, \tag{4.4}$$

onde $\mathbf{E}_{\ell(\vec{e},V_j)} = \phi(V_{\ell_j})$.

A matriz $\mathbf{E}_\ell$ pode ser redimensionada como um vetor linha cujas componentes são as colunas transpostas de $[\mathbf{E}_{\ell(e,V_j)}]$ (Equação 4.5):

$$\mathbf{E}_{\ell 1\times|e|*|V|} = \mathbf{E}_{\ell completo} = [\mathbf{E}_{\ell(e,V_1)}]^T, [\mathbf{E}_{\ell(e,V_2)}]^T, \cdots [\mathbf{E}_{\ell(e,V_N)}]^T] \tag{4.5}$$

A matriz dos erros dos treinandos 4.6 é matriz onde cada linha corresponde ao vetor de erro de um treinando dado pela Equação 4.5:

$$\mathbf{E}_{L\times|V|} = \{\mathbf{E}_{\ell completo}\}, \tag{4.6}$$

com $\ell = 1, \cdots, L$.

4.1.2 Matrizes de padrões de erros

Da etapa de modelagem do domínio do conhecimento baseada em grafo temos o seguinte: os caminhos viáveis representam as sequências esperadas para

percorrer o grafo da tarefa, e as sequências de atividades realizadas pelos treinandos foram definidas como passeios induzidos pelos vértices do grafo da tarefa. Um erro na execução da tarefa é um desvio de algum caminho.

O modelo matricial dos erros dos treinandos resulta em um vetor de erros de aprendizagem de alta dimensão (Equação 4.5), assim, adicionalmente, para análise de agrupamento, duas formas reduzidas foram aplicadas às matrizes de erros do treinando. Em ambos os casos, para cada treinando, é atribuído um único vetor de erro, cujos componentes são

- a) o vetor linha do máximo erro dado pela Equação 4.7:

$$\mathbf{E}_{\ell\max} = \left[\begin{array}{cccc} max(\mathbf{E}_{(e_i,V_1)}) & max\mathbf{E}_{(e_i,V_2)}) & \cdots & max(E_{(e_i,V_N)}) \end{array} \right] \tag{4.7}$$

- b) o vetor linha da soma dos erros dado pela Equação 4.8:

$$\mathbf{E}_{\ell soma} = \left[\begin{array}{cccc} \sum_{i=1}^{6} E_{(e_i,V_1)} & \sum_{i=1}^{6} E_{(e_i,V_2)} & \cdots & \sum_{i=1}^{6} E_{(e_i,V_N)} \end{array} \right] \tag{4.8}$$

4.2 MODELAGEM DOS DADOS BASEADA EM GRAFOS

Nesta seção o domínio do conhecimento da execução da tarefa é modelado como um grafo, as sequências de atividades executadas pelos treinandos são modeladas como passeios sobre o grafo os quais apresentam desvios em relação ao caminhos viáveis para a execução da tarefa. A comparação entre caminhos e passeios e passeios entre si são é realizada a partir das definições de similaridade entre grafos apresentada na Fundamentação Teórica.

4.2.1 Modelagem do domínio do conhecimento como um grafo tarefa

Dado um vetor de tarefa V com N entradas de atividade V_j, $j = 1, \cdots, N$, as regras para executar as atividades são restrições de ordem e definem X caminhos viáveis

$$P = \{P_x\} = P_x(V_{P_x}, \mathcal{E}_{P_x}) = \{V_{P_x 1}, \cdots, V_{P_x N}\}, \tag{4.9}$$

com

$$V_{P_x i} \neq V_{P_x j}, \tag{4.10}$$

$\forall x \in X$.

A união dos caminhos resulta no grafo de tarefas, denominado grafo tarefa,

$$G_T = (V_G, \mathcal{E}_G) = \bigcup_{i=1}^{X} P_x, \tag{4.11}$$

onde $V_G = V_P$ e

$$\mathcal{E}_G = \bigcup_{i=1}^{X} \mathcal{E}_{P_x}. \tag{4.12}$$

Com base na lista de atividades predecessores e sucessores de uma atividade na 10, define-se a matriz de adjacência do grafo tarefa:

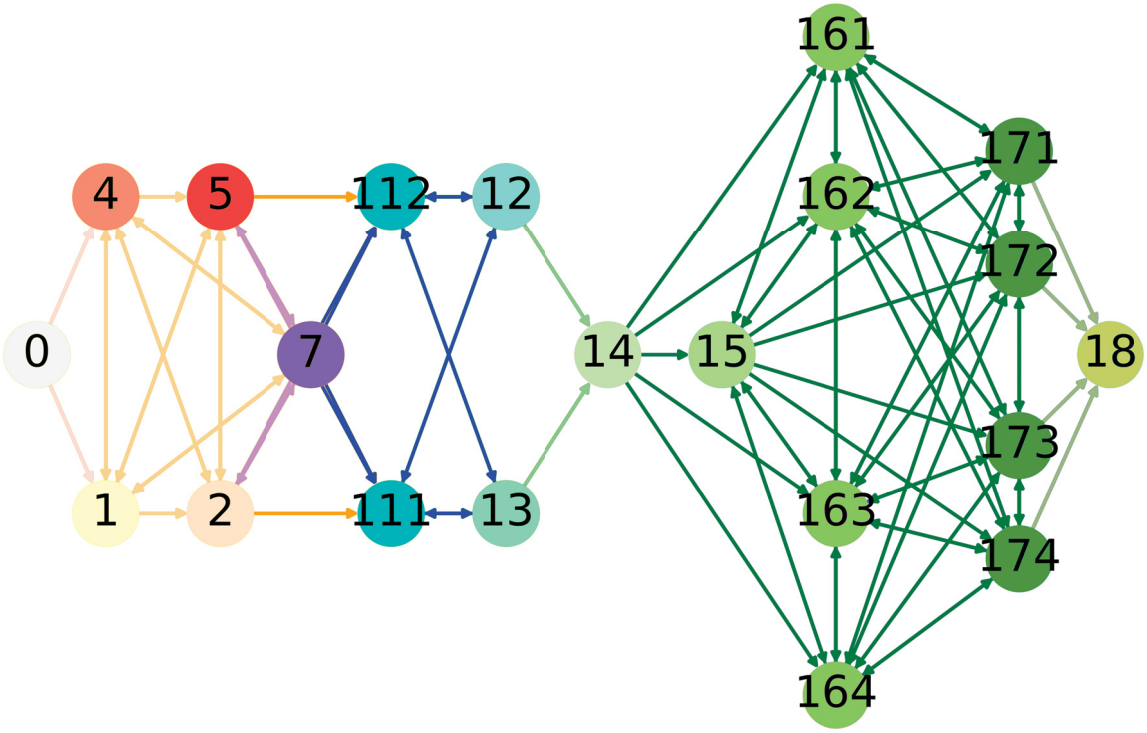
FIGURA 45 – MATRIZ DE ADJACÊNCIA DO GRAFO DA TAREFA

	1	2	7	4	5	111	12	112	13	14	161	162	163	164	15	171	172	173	174	18
1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
7	1	1	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	1	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
111	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0
112	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
161	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0
162	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	1	0
163	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	1	1	0
164	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1	1	1	1	0
15	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	1	1	1	1	0
171	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	1	1	1	1
172	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	1	0	1	1	1
173	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	1	1	0	1	1
174	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	1	1	1	0	1
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

FONTE: O Autor

O grafo tarefa é um multidígrafo, pois conforme ilustrado na FIGURA 46, seus vértices suportam duas arestas direcionadas em sentidos contrários.

FIGURA 46 – GRAFO ATIVIDADE



FONTE: O Autor

4.2.2 Modelagem do treinando como passeios sobre o grafo da tarefa

As sequências realizadas pelos L treinandos são os passeios $\{W_\ell\}$ induzidos pelos vértices do grafo da tarefa, $\ell = 1, \cdots, L$. O passeio do treinando pode ter arestas que não tenha um correspondente no grafo de tarefas. A ordem de atividade realizada pelo ℓ-ésimo treinando correspondente à ordem dos vértices em

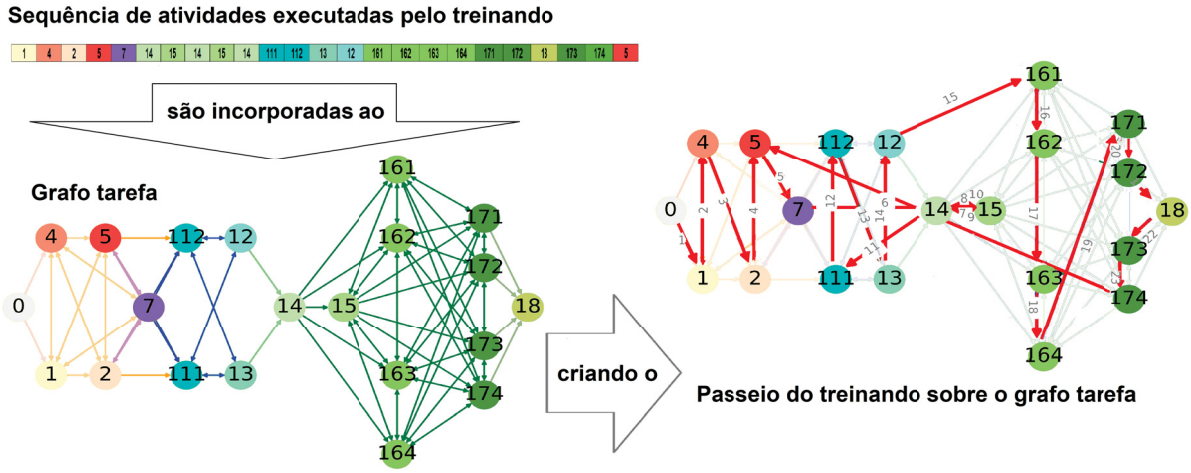
$$W_\ell = \{V_{W_{\ell,1}}, \cdots, V_{W_{\ell,R}}\}, \tag{4.13}$$

onde R é o total das atividades concluídas. Além disso, supõe-se que W_ℓ é definido sobre um grafo super conectado construído a partir de G_T, de tal forma que

$$|V_{P_x}| = |V_{W_\ell}|, \forall \ell, x. \tag{4.14}$$

Desta forma, uma comparação $W \times P$ diz respeito apenas à diferença entre suas arestas direcionadas. A FIGURA 47 ilustra a transformação do espaço dos dados de interação no espaço dos caminhos sobre o grafo tarefa. Os rótulos sobre as arestas indicam a ordem na qual as atividades foram executadas.

FIGURA 47 – CAMINHO DE UM TREINANDO SOBRE O GRAFO DA TAREFA



FONTE: O Autor

4.2.3 Modelagem dos padrões de similaridade

Duas abordagens foram adotadas para os cálculos de similaridade dos passeios, uma relativa e outra absoluta. Na primeira abordagem, tomam-se os passeios dois a dois, por exemplo, $\mathcal{W}_i$ e $\mathcal{W}_j$, como apresentado na FIGURA 48 (a), e verifica-se a similaridade entre eles definida no espaço $(W \times W)$. Na segunda abordagem, são mensuradas as similaridades entre os passeios dos treinandos e os caminhos viáveis (FIGURA 48(b)) definidas no espaço $(W \times P)$. No segundo caso, cada passeio $\mathcal{W}_q$ é comparado com cada um dos K caminhos viáveis $\mathcal{P}_k$.

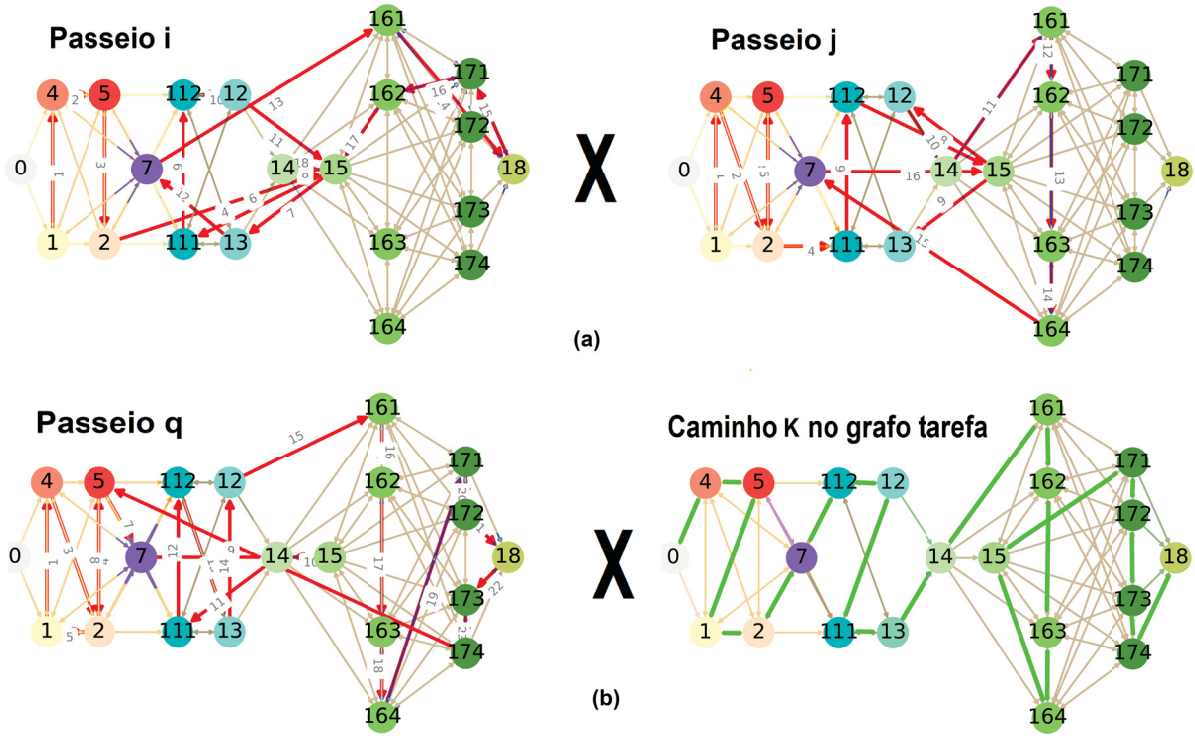
4.2.4 Padrão de similaridade entre passeios e caminhos

Para cada medida de similaridade Sim_d, baseada em uma distância d, ficam definidas as matrizes de similaridade $[S_d]$ definidas na primeira abordagem pela Equação 4.15,

$$[S_{d(W\times P)}] = [Sim_d(W_q, P_k)], \tag{4.15}$$

com $q = 1, \cdots, L$, e $k = 1, \cdots, X$. Cada linha i destas matrizes corresponde a um vetor, ou padrão de similaridade do passeio $\mathcal{W}_q$ com os 1680 caminhos viáveis. Uma visualização parcial da matriz de similaridades baseada na distância de sobreposição vértice-aresta pode ser vista na FIGURA 49

FIGURA 48 – COMPARAÇÃO DE PASSEIOS ENTRE SI E PASSEIOS E CAMINHOS



FONTE: O Autor
LEGENDA: Comparação de passeios entre si (acima) e passeios e caminhos (embaixo)

FIGURA 49 – Matriz do padrão de similaridades $\mathcal{W} \times \mathcal{P}$

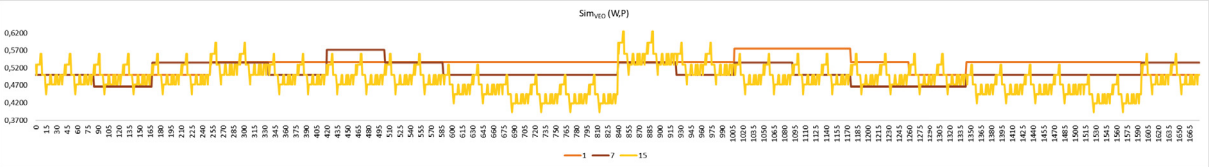
$$\begin{bmatrix}
0.5094 & 0.5094 & 0.5094 & 0.5094 & 0.5385 & 0.5385 & 0.5385 & 0.5385 & 0.5385 & 0.5094 & \cdots & 0.4815 & 0.4815 & 0.4815 & 0.4815 & 0.4815 & 0.4815 & 0.4815 & 0.4815 & 0.4815 \\
0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 & \cdots & 0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\
0.5116 & 0.5116 & 0.5116 & 0.5116 & 0.5116 & 0.5116 & 0.5116 & 0.5116 & 0.5116 & 0.5116 & \cdots & 0.5476 & 0.5476 & 0.5476 & 0.5476 & 0.5476 & 0.5476 & 0.5476 & 0.5476 & 0.5476 \\
0.6047 & 0.5682 & 0.5682 & 0.5682 & 0.5682 & 0.5333 & 0.5333 & 0.5682 & 0.5333 & 0.5682 & \cdots & 0.5333 & 0.5000 & 0.5000 & 0.5333 & 0.5000 & 0.5333 & 0.5333 & 0.5333 & 0.5682 \\
0.5102 & 0.4800 & 0.4800 & 0.4800 & 0.4800 & 0.4510 & 0.4510 & 0.4800 & 0.4510 & 0.4800 & \cdots & 0.5417 & 0.5102 & 0.5102 & 0.5417 & 0.5102 & 0.5417 & 0.5417 & 0.5417 & 0.5745 \\
0.4902 & 0.4615 & 0.4615 & 0.4615 & 0.4615 & 0.4340 & 0.4340 & 0.4615 & 0.4340 & 0.4615 & \cdots & 0.5200 & 0.4902 & 0.4902 & 0.5200 & 0.4902 & 0.5200 & 0.5200 & 0.5200 & 0.5510 \\
0.5000 & 0.4706 & 0.4706 & 0.4706 & 0.4706 & 0.4423 & 0.4423 & 0.4706 & 0.4423 & 0.5000 & \cdots & 0.5306 & 0.5000 & 0.5000 & 0.5306 & 0.5000 & 0.5306 & 0.5306 & 0.5306 & 0.5625 \\
0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 & \cdots & 0.5349 & 0.5349 & 0.5349 & 0.5349 & 0.5349 & 0.5349 & 0.5349 & 0.5349 & 0.5349 \\
0.4815 & 0.4545 & 0.4545 & 0.4545 & 0.4545 & 0.4286 & 0.4286 & 0.4545 & 0.4286 & 0.4545 & \cdots & 0.4815 & 0.4545 & 0.4545 & 0.4815 & 0.4545 & 0.5094 & 0.4815 & 0.4815 & 0.5385 \\
0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 & \cdots & 0.5349 & 0.5349 & 0.5349 & 0.5349 & 0.5349 & 0.5349 & 0.5349 & 0.5349 & 0.5349 \\
0.5714 & 0.5400 & 0.5400 & 0.5400 & 0.5400 & 0.5098 & 0.5098 & 0.5400 & 0.5098 & 0.5400 & \cdots & 0.5098 & 0.4808 & 0.4808 & 0.5098 & 0.4808 & 0.5400 & 0.5098 & 0.5098 & 0.5714 \\
0.6000 & 0.6000 & 0.6000 & 0.6000 & 0.5686 & 0.5385 & 0.5385 & 0.5385 & 0.5094 & 0.5385 & \cdots & 0.5686 & 0.5385 & 0.5385 & 0.5385 & 0.5094 & 0.6000 & 0.5686 & 0.5686 & 0.6000 \\
0.5577 & 0.5000 & 0.5000 & 0.5000 & 0.5577 & 0.5000 & 0.5000 & 0.5577 & 0.5000 & 0.5283 & \cdots & 0.5283 & 0.4727 & 0.4727 & 0.5283 & 0.4727 & 0.5283 & 0.5000 & 0.5000 & 0.5000 & 0.5577 \\
0.6275 & 0.5660 & 0.5660 & 0.5370 & 0.5370 & 0.5091 & 0.4821 & 0.5660 & 0.5091 & 0.5370 & \cdots & 0.4821 & 0.4561 & 0.4310 & 0.5091 & 0.4561 & 0.5370 & 0.5091 & 0.5091 & 0.4821 & 0.5370 \\
0.5294 & 0.5294 & 0.5294 & 0.5600 & 0.5000 & 0.4717 & 0.5000 & 0.4717 & 0.4717 & 0.4717 & \cdots & 0.5294 & 0.5000 & 0.5294 & 0.5000 & 0.5000 & 0.5600 & 0.5294 & 0.5294 & 0.5600 & 0.5600 \\
0.5294 & 0.5000 & 0.5294 & 0.5294 & 0.5294 & 0.5294 & 0.5294 & 0.5600 & 0.5600 & 0.5000 & \cdots & 0.4717 & 0.4717 & 0.4717 & 0.5000 & 0.5000 & 0.5000 & 0.4717 & 0.5000 & 0.5000 & 0.5000 \\
0.5686 & 0.5385 & 0.5385 & 0.5385 & 0.5686 & 0.5385 & 0.5385 & 0.5686 & 0.5385 & 0.5385 & \cdots & 0.5385 & 0.5094 & 0.5094 & 0.5385 & 0.5094 & 0.5686 & 0.5385 & 0.5385 & 0.5385 & 0.5686 \\
0.4444 & 0.4444 & 0.4444 & 0.4444 & 0.4717 & 0.4717 & 0.4717 & 0.4717 & 0.4717 & 0.4717 & \cdots & 0.4717 & 0.4717 & 0.4717 & 0.4717 & 0.4717 & 0.4444 & 0.4444 & 0.4444 & 0.4444 & 0.4444 \\
0.4545 & 0.4286 & 0.4286 & 0.4545 & 0.4545 & 0.4286 & 0.4545 & 0.4545 & 0.4545 & 0.4286 & \cdots & 0.4815 & 0.4545 & 0.4815 & 0.4815 & 0.4815 & 0.5094 & 0.4815 & 0.4815 & 0.5094 & 0.5094 \\
0.4906 & 0.4630 & 0.4364 & 0.4364 & 0.4364 & 0.4107 & 0.4107 & 0.4364 & 0.4107 & 0.4364 & \cdots & 0.4364 & 0.4107 & 0.4107 & 0.4364 & 0.4107 & 0.4630 & 0.4630 & 0.4364 & 0.4364 & 0.4630 \\
0.6275 & 0.5660 & 0.5370 & 0.5660 & 0.5370 & 0.4821 & 0.5091 & 0.5370 & 0.5091 & 0.5370 & \cdots & 0.4561 & 0.4068 & 0.4310 & 0.4561 & 0.4310 & 0.5091 & 0.4821 & 0.4561 & 0.4821 & 0.5091 \\
0.5192 & 0.5192 & 0.5490 & 0.5490 & 0.4906 & 0.4906 & 0.4906 & 0.4906 & 0.4906 & 0.4630 & \cdots & 0.5192 & 0.5192 & 0.5192 & 0.5192 & 0.5192 & 0.5490 & 0.5192 & 0.5490 & 0.5490 & 0.5490
\end{bmatrix}$$

FONTE: O Autor

Cada componente da linha desta matriz é resultado da aplicação da função similaridade Sim_d sobre um passeio $\mathcal{W}_\ell$ e um vetor de caminhos viáveis $\mathcal{P}$. O gráfico de um passeio $\mathcal{W}_\ell$ é um conjunto de pontos do plano cujas coordenadas têm abscissas no vetor de caminhos $\mathcal{P}$ e ordenadas calculadas a partir de suas distâncias a cada caminho em $\mathcal{P}$. Fixando uma definição de similaridade, por exemplo, sobreposição vértice-aresta, VEO, os gráficos de três passeios em relação aos 1680 caminhos são

representados na figura 50.

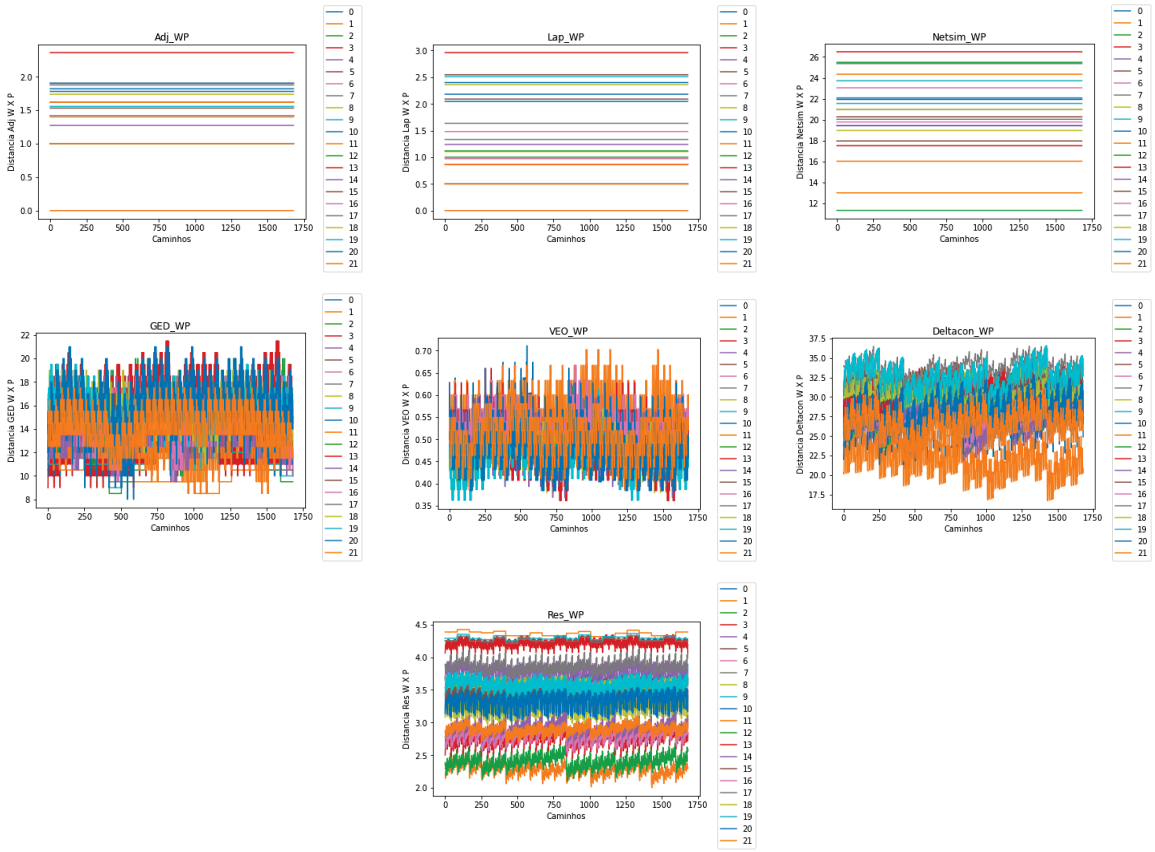
FIGURA 50 – GRÁFICO DE SIMILARIDADE $VEO\ \mathcal{W} \times \mathcal{P}$ DE TRÊS PASSEIOS



FONTE: O Autor

Os gráficos detalhados de todos os padrões por medida de similaridade analisada estão no APÊNDICE 6. Os padrões das sete definições de similaridade entre passeios e caminhos são apresentados na FIGURA 51.

FIGURA 51 – Padrões de similaridade $(\mathcal{W} \times \mathcal{P})$



FONTE: O Autor

Conforme análise gráfica, os padrões obtidos podem ser divididos em dois grupos. Os padrões que resultam das similaridades espectrais – adjacência e laplaciana – e Netsimile apresentam um comportamento similar onde a distância entre um passeio e

qualquer um dos caminhos é constante. Este resultado indica que todos os caminhos são iguais nos espaços de padrões definidos a partir destas distâncias. De fato, o cálculo da distância entre dois caminhos utilizando estas definições resulta sempre nulo.

Os caminhos dos treinandos se diferenciam apenas em relação às suas arestas, mas não quanto ao número de arestas, o número de vértices ou grau dos vértices. Neste sentido, a medida de similaridade Netsimile, definida sobre medidas estatísticas associada à estrutura *egonet* centrada em cada vértice, é incapaz de distinguir entre dois caminhos diferentes tais como definidos aqui. Pela mesma razão, no caso das distâncias espectrais existe uma grande chance de que os caminhos sejam coespectrais, ou seja, compartilham os mesmos autovalores que são utilizados no cálculo das similaridades baseada na distância λ.

As similaridades restantes definem o segundo grupo onde as distâncias entre passeios e caminhos parecem se comportar de forma errática. No gráficos de padrões de similaridades baseadas na distância de edição (GED), sobreposição vértice-aresta (VEO) e Deltacon ($\delta - con$) esta característica é mais acentuada com os padrões se sobrepondo uns aos outros. Já os padrões derivados da similaridade baseada na Resistência Normalizada, embora estejam neste grupo, estão distribuídos em intervalos relativamente bem definidos com pouca sobreposição de padrões.

Na TABELA 11 é apresentado o sumário estatístico dos valores assumidos pelas similaridades entre passeios e caminhos. A variabilidade destas similaridades pode ser analisada segundo o número de valores e a distribuição de seus valores. No primeiro caso comparam-se o número de valores distintos entre as diferentes distâncias. Os números de valores de similaridades distintos para $\delta - con$ e resistência, 23559 e 23697, respectivamente, são relativamente altos em comparação com as outras distâncias, para as quais estes valores estão entre 14 e 98.

TABELA 11 – SUMÁRIO ESTATÍSTICO DOS VALORES DE SIMILARIDADES ENTRE PASSEIOS E CAMINHOS

	Lap_WP		Adj_WP		GED_WP		VEO_WP		δcon_WP		ResN_WP		Netsim_WP	
Valores distintos	19		14		28		98		23697		23559		22	
	Vals	**Freq.**	**Vals**	**Freq.**	**Vals**	**Freq.**	**Vals**	**Freq.**	**Vals**	**Freq.**	**Vals**	**Freq.**	**Vals**	**Freq.**
Média	1.60	1945.26	1.52	2640.00	14.75	1320.00	0.51	377.14	28.90	1.56	3.24	1.57	20.45	1680.00
Desvio padrão	0.79	842.45	0.55	1945.22	4.11	1072.11	0.08	602.09	3.23	5.92	0.49	8.70	3.80	0.00
Mínimo	0.00	1680.00	0.00	1680.00	8.00	2.00	0.36	1.00	16.69	1.00	2.00	1.00	11.34	1680.00
1º Quartil	1.05	1680.00	1.40	1680.00	11.38	230.25	0.44	28.25	26.98	1.00	2.86	1.00	19.09	1680.00
Mediana	1.48	1680.00	1.59	1680.00	14.75	1366.00	0.51	197.00	29.29	1.00	3.37	1.00	20.63	1680.00
3º Quartil	2.27	1680.00	1.81	2940.00	18.13	2248.50	0.57	504.00	31.21	1.00	3.59	1.00	22.80	1680.00
Máximo	2.96	5040.00	2.36	8400.00	21.50	3020.00	0.71	4853.00	36.59	336.00	4.42	672.00	26.49	1680.00
Coef. Var.	0.50	0.43	0.36	0.74	0.28	0.81	0.17	1.60	0.11	3.79	0.15	5.55	0.19	0.00

FONTE: O Autor

No segundo caso, a variabilidade é analisada em relação aos padrões definidos

a partir da mesma similaridade e expressa pelo coeficiente de variação (C.V.)[2] cujos valores são apresentados na última linha da TABELA 11 e em cada coluna **Vals**. Em relação à heterogeneidade dos dados, novamente as distâncias $\delta - con$ e Resistência Normalizada se destacam com os menores valores do coeficiente de variação, 0.11 e 0.15.

4.2.5 Padrões de similaridade entre passeios

As matrizes de similaridade $[S_d]$ são definidas a partir de cada medida de similaridade Sim_d baseada em uma distância d. Entretanto, neste caso, as similaridades são calculadas entre os passeios entre si por meio da equação 4.17.

$$[S_{d(W\times W)}] = [Sim_d(W_i, W_j)], \tag{4.17}$$

com $i, j = 1, \cdots, L$. Cada linha i destas matrizes corresponde a um vetor, ou padrão de similaridade do passeio $\mathcal{W}_i$ com os L passeios.

A FIGURA 4.2.5 representa a matriz de similaridade para a distância de edição entre passeios. A matriz da FIGURA 4.2.5 é uma matriz de similaridade típica cujas características evidentes são a simetria e todas as componentes da diagonal principal iguais a um.

FIGURA 52 – Matriz dos padrões de similaridade $\mathcal{W} \times \mathcal{W}$

1	0.400	0.457	0.371	0.429	0.400	0.286	0.400	0.200	0.429	0.314	0.171	0.200	0.057	0.229	0.314	0.229	0.229	0.257	0.114	0.257	0.257
0.400	1	0.829	0.686	0.571	0.486	0.543	0.829	0.400	0.800	0.457	0.486	0.400	0.257	0.486	0.514	0.429	0.429	0.514	0.486	0.286	0.286
0.457	0.829	1	0.800	0.686	0.543	0.657	0.943	0.457	0.971	0.571	0.314	0.343	0.371	0.371	0.400	0.371	0.543	0.400	0.429	0.400	0.457
0.371	0.686	0.800	1	0.657	0.571	0.571	0.743	0.429	0.771	0.600	0.286	0.314	0.400	0.343	0.314	0.286	0.343	0.257	0.400	0.371	0.600
0.429	0.571	0.686	0.657	1	0.686	0.629	0.629	0.600	0.714	0.543	0.343	0.257	0.286	0.286	0.257	0.229	0.343	0.314	0.571	0.486	0.429
0.400	0.486	0.543	0.571	0.686	1	0.486	0.486	0.457	0.514	0.457	0.257	0.286	0.371	0.314	0.286	0.257	0.257	0.229	0.429	0.457	0.343
0.286	0.543	0.657	0.571	0.629	0.486	1	0.600	0.400	0.686	0.457	0.200	0.171	0.143	0.429	0.171	0.086	0.429	0.229	0.314	0.229	0.343
0.400	0.829	0.943	0.743	0.629	0.486	0.600	1	0.400	0.914	0.514	0.314	0.343	0.314	0.371	0.400	0.371	0.486	0.400	0.429	0.343	0.400
0.200	0.400	0.457	0.429	0.600	0.457	0.400	0.400	1	0.486	0.429	0.171	0.200	0.171	0.171	0.143	0.114	0.229	0.257	0.457	0.257	0.314
0.429	0.800	0.971	0.771	0.714	0.514	0.686	0.914	0.486	1	0.543	0.286	0.314	0.343	0.343	0.371	0.343	0.514	0.371	0.457	0.371	0.429
0.314	0.457	0.571	0.600	0.543	0.457	0.457	0.514	0.429	0.543	1	0.514	0.429	0.286	0.286	0.371	0.286	0.286	0.314	0.457	0.486	0.543
0.171	0.486	0.314	0.286	0.343	0.257	0.200	0.314	0.171	0.286	0.514	1	0.457	0.143	0.486	0.400	0.429	0.086	0.457	0.314	0.286	0.343
0.200	0.400	0.343	0.314	0.257	0.286	0.171	0.343	0.200	0.314	0.429	0.457	1	0.286	0.286	0.657	0.686	0.171	0.429	0.286	0.371	0.257
0.057	0.257	0.371	0.400	0.286	0.371	0.143	0.314	0.171	0.343	0.286	0.143	0.286	1	0.086	0.400	0.371	0.029	0.000	0.143	0.286	0.286
0.229	0.486	0.371	0.343	0.286	0.314	0.429	0.371	0.171	0.343	0.286	0.486	0.286	0.086	1	0.286	0.257	0.143	0.286	0.143	0.229	0.229
0.314	0.514	0.400	0.314	0.257	0.286	0.171	0.400	0.143	0.371	0.371	0.400	0.657	0.400	0.286	1	0.743	0.229	0.429	0.229	0.257	0.257
0.229	0.429	0.371	0.286	0.229	0.257	0.086	0.371	0.114	0.343	0.286	0.429	0.686	0.371	0.257	0.743	1	0.143	0.343	0.200	0.286	0.114
0.229	0.429	0.543	0.343	0.343	0.257	0.429	0.486	0.229	0.514	0.286	0.086	0.171	0.029	0.143	0.229	0.143	1	0.343	0.200	0.171	0.286
0.257	0.514	0.400	0.257	0.314	0.229	0.229	0.400	0.257	0.371	0.314	0.457	0.429	0.000	0.286	0.429	0.343	0.343	1	0.286	0.257	0.371
0.114	0.486	0.429	0.400	0.571	0.429	0.314	0.429	0.457	0.457	0.457	0.314	0.286	0.143	0.143	0.229	0.200	0.200	0.286	1	0.514	0.229
0.257	0.286	0.400	0.371	0.486	0.457	0.229	0.343	0.257	0.371	0.486	0.286	0.371	0.286	0.229	0.257	0.286	0.171	0.257	0.514	1	0.314
0.257	0.286	0.457	0.600	0.429	0.343	0.343	0.400	0.314	0.429	0.543	0.343	0.257	0.286	0.229	0.257	0.114	0.286	0.371	0.229	0.314	1

[2] Coeficiente de variação (C.V.) é uma medida de variabilidade adimensional na qual variância da amostra de dados é escalonada segundo a média dos dados. O coeficiente de variação pode ser utilizado para a comparação da variabilidade entre duas amostras e o seu cálculo pode ser realizado pela Equação 4.16

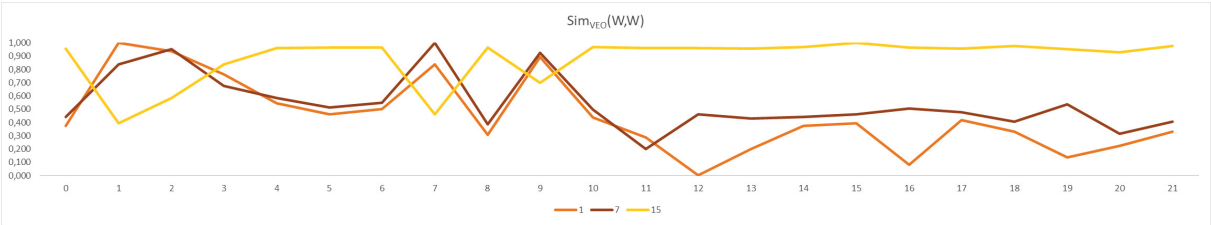
$$CV = \frac{\sigma}{\mu}, \tag{4.16}$$

onde σ e μ são a variância da amostra e a média da amostra. Por um lado, valores de C.V. próximos de zero indicam que os dados da amostra são mais homogêneos. Por outro lado, valores de C.V. acima de $0,3$ indicam grande variabilidade ou heterogeneidade dos dados da amostra (GARCIA, 1989).

FONTE: O Autor

Como na definição das matrizes de similaridade entre passeios e caminhos, cada componente da linha desta matriz corresponde a similaridade Sim_d de um passeio $\mathcal{W}_{\mathcal{W}_\ell}$ e todos os passeios $\mathcal{W}_\}$. Nesse caso, o gráfico de um passeio $\mathcal{W}_\ell$ é um conjunto de pontos do plano cujas coordenadas tem abscissas no vetor de passeios $\mathcal{W}$ e ordenadas dadas por sua distância a cada passeio em $\mathcal{W}$. Para uma definição de similaridade, por exemplo, sobreposição vértice-aresta, VEO, os gráficos de três passeios em relação aos 22 passeios são representados na figura 53.

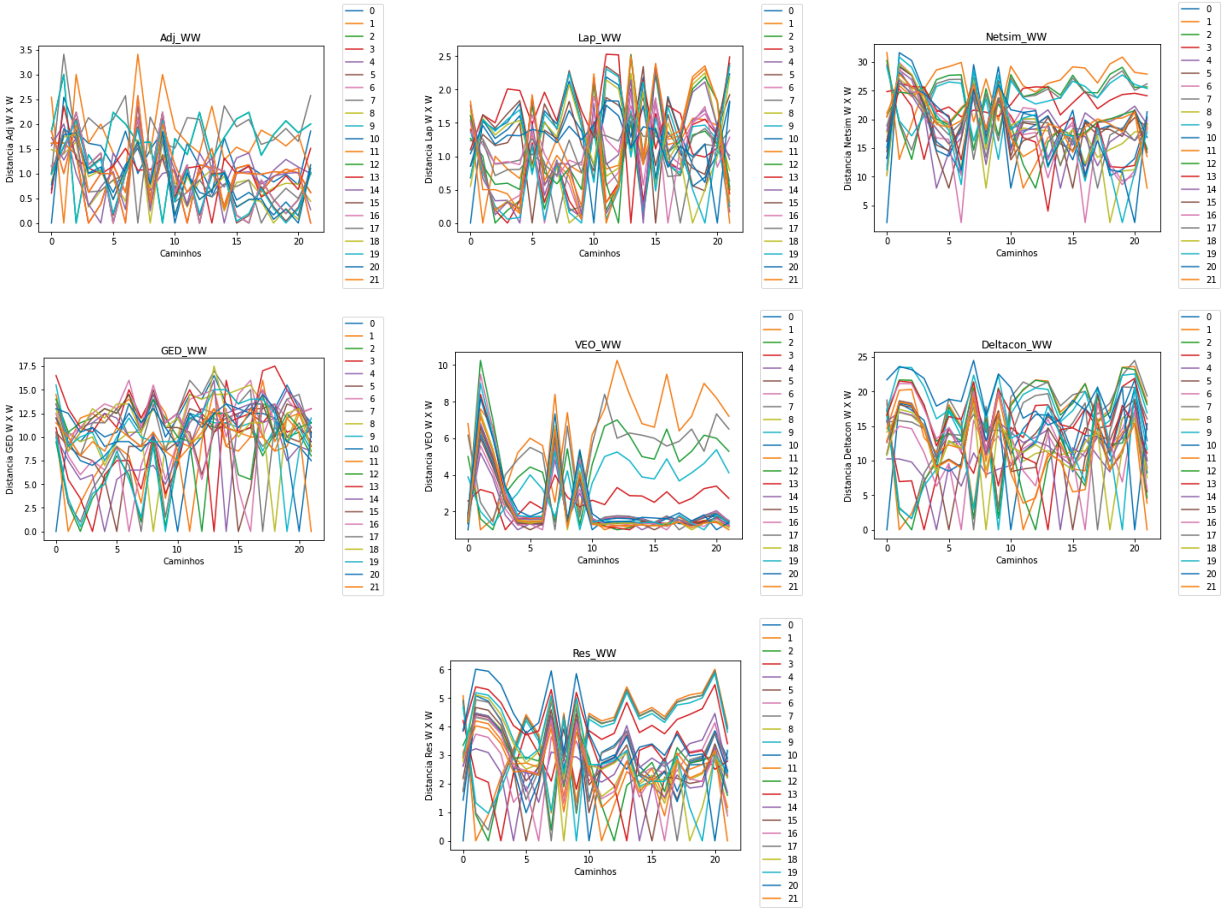
FIGURA 53 – GRÁFICO DE SIMILARIDADE VEO $\mathcal{W} \times \mathcal{W}$ DE TRÊS PASSEIOS



FONTE: O Autor

Os padrões de similaridade entre passeios para as sete distâncias analisadas são apresentados na figura 54. Diferente dos padrões entre passeios e caminhos, a análise gráfica dos padrões entre passeios não permite uma distinção intuitiva das definições de similaridade. Um caso que merece menção, entretanto, é a medida de similaridade Netsimile. Conforme pode ser observado no seu gráfico, Netsimile é a única medida cuja distância entre dois padrões iguais não é nula, embora seja a menor dentre os outros padrões. Dessa forma, a distância nula entre dois padrões iguais pode ser obtida somente depois que as linhas da matriz de padrões são separadamente padronizadas utilizando a norma min-max.

FIGURA 54 – Padrões de similaridade $(\mathcal{W} \times \mathcal{W})$



FONTE: O Autor

A distribuição dos valores distintos para descrever os padrões de similaridade entre passeios é mais uniforme do que a distribuição dos padrões de similaridades entre passeios e caminhos. O número de valores distintos para todas as distâncias não ultrapassa 200. A variabilidade dos valores entretanto segue aquela dos padrões entre passeios e caminhos. Exceto pela menor variabilidade de Netsimile, as distâncias $\delta - con$ e Resistência Normalizada apresentam o menor coeficiente de variação, $0,34$ e $0,37$, respectivamente.

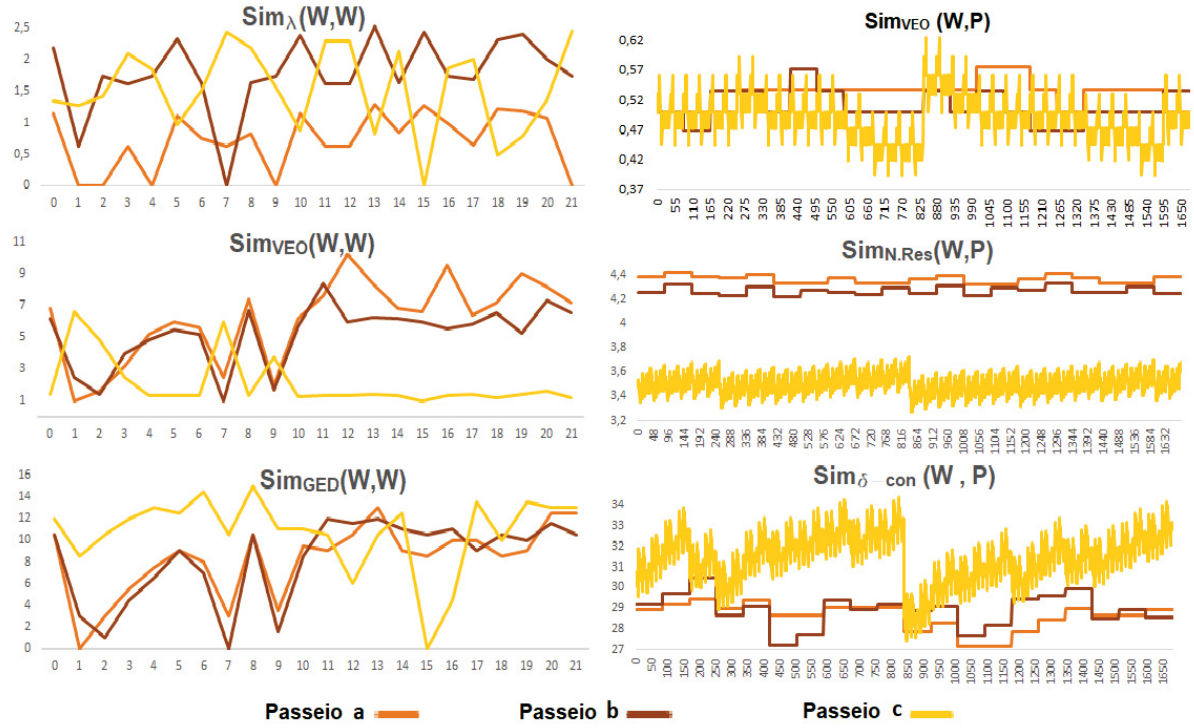
TABELA 12 – FREQUÊNCIA DOS VALORES DE SIMILARIDADES ENTRE DOIS CAMINHOS

	Lap_WW		Adj_WW		GED_WW		VEO_WW		δcon_WW		ResN_WW		Netsim_WW	
Valores distintos	224		127		34		98		161		232		235	
	Vals	**Freq.**	**Vals**	**Freq.**	**Vals**	**Freq.**	**Vals**	**Freq.**	**Vals**	**Freq.**	**Vals**	**Freq.**	**Vals**	**Freq.**
Média	1.22	2.16	1.13	3.81	9.13	14.24	2.95	3.01	14.13	2.09	3.16	2.10	20.45	2.06
Desvio padrão	0.59	1.38	0.65	4.23	5.17	12.70	2.12	2.24	4.84	1.31	1.18	1.32	3.80	0.61
Mínimo	0.00	2.00	0.00	2.00	0.00	2.00	1.00	2.00	0.00	2.00	0.00	2.00	11.34	1.00
1º Quartil	0.81	2.00	0.67	2.00	5.13	4.00	1.42	2.00	10.95	2.00	2.36	2.00	19.09	2.00
Mediana	1.24	2.00	1.02	2.00	9.25	11.00	1.74	2.00	14.57	2.00	2.98	2.00	20.63	2.00
3º Quartil	1.61	2.00	1.58	4.00	13.38	25.00	4.22	4.00	17.73	2.00	4.14	2.00	22.80	2.00
Máximo	2.53	22.00	3.41	34.00	17.50	44.00	10.25	22.00	24.47	22.00	6.01	22.00	26.49	10.00
Coef. Var.	0.49	0.64	0.58	1.11	0.57	0.89	0.72	0.74	0.34	0.63	0.37	0.63	0.19	0.30

FONTE: O Autor

A FIGURA 55 apresenta uma amostra das similaridades entre três passeios a, b e c. Na coluna de gráficos à esquerda na FIGURA 55 estão representados os padrões de similaridades para a Laplaciana normalizada ($Sim_{\lambda_{Lap}}(W,W)$), sobreposição vértice-aresta ($Sim_{VEO}(W,W)$) e distância de edição ($Sim_{GED}(W,W)$) dos três passeios quando medidas entre os passeios entre si.

FIGURA 55 – PADRÕES INVARIANTES ATRAVÉS DE DIFERENTES MEDIDAS DE SIMILARIDADES



FONTE: O Autor

Na coluna de gráficos à direita na figura, são apresentados os padrões de similaridade entre os três passeios a, b e c e os 1680 caminhos. Neste caso as similaridades apresentadas são sobreposição vértice-aresta ($Sim_{VEO}(W,W)$), resistência normalizada ($Sim_{Res}(W,W)$) e Deltacon ($Sim_{\delta-con}(W,W)$). Por exemplo, os padrões de similaridades dos passeios a e b são mais semelhantes do que os vetores dos passeios a e c ou b e c. Em qualquer caso, como pode ser observado na FIGURA 55, dois aspectos devem ser destacados:

1. os padrões de similaridade de um passeio são diferentes para cada medida de similaridade, ou seja, as medidas de similaridade aqui analisadas avaliam de maneira diferente o grau de semelhança ou proximidade entre dois padrões.

2. as distâncias entre dois padrões são invariantes em relação às diferentes medidas de similaridades, ou seja, independente da medida de similaridade utilizada, se dois padrões estão mais próximos ou mais distantes segundo uma medida de similaridade, então esta proximidade ou distância será mantida observada quando outra medida for utilizada.

A análise dos gráficos dos passeios por meio de suas matrizes de similaridade fornecem indícios que os padrões de similaridades de passeios permitem distinguir padrões semelhantes tanto em relação aos passeios entre si quanto em relação aos caminhos viáveis. Além disso, essa distinção é invariante em relação às diferentes medidas de similaridade. Resta saber se estes dois aspectos destacados serão suficientes para agrupar estes padrões da mesma maneira ou se os agrupamentos formados serão diferentes conforme o padrão de similaridade. A seguir será apresentada a análise de agrupamentos para os padrões de similaridade dos passeios dos treinandos.

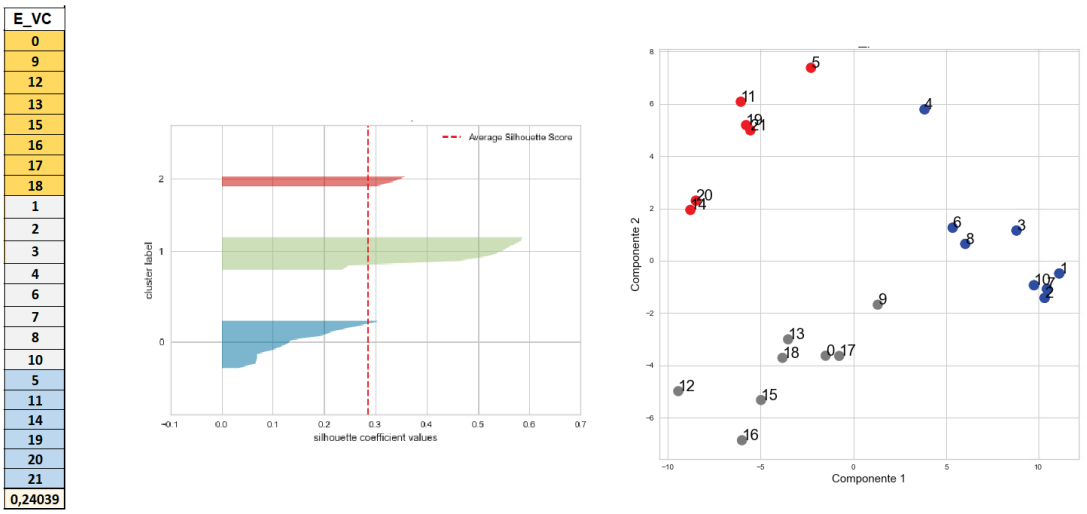
5 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Neste capítulo, são apresentados os resultados da análise de agrupamentos dos padrões de erros e dos padrões de similaridade. Embora o método de agrupamento utilizado seja o mesmo, os espaços de padrões são diferentes resultando em agrupamentos relativamente diversos em relação à sua estrutura geométrica e qualidade. Conforme a *DSR*, a sugestão de solução do problema é representada pelos agrupamentos baseados me padrões de erros, e as alternativas a esta solução são os agrupamentos de padrões similaridade.

5.1 ANÁLISE DE AGRUPAMENTO DE PADRÕES DE ERRO

A partir da definição da matriz de erro dos treinandos e suas formas reduzidas, os padrões de erro são extraídos segundo as três matrizes de erro $\mathbf{E}_{soma}$, $\mathbf{E}_{max}$ e $\mathbf{E}_{completo}$. Os agrupamentos de padrões gerados são apresentados a seguir.

FIGURA 56 – AGRUPAMENTOS DE PADRÕES DE ERRO: ERRO COMPLETO



FONTE: O Autor
LEGENDA: (a) padrões com rótulos de grupos; (b) Gráfico de silhueta; (c) Projeção das componentes principais

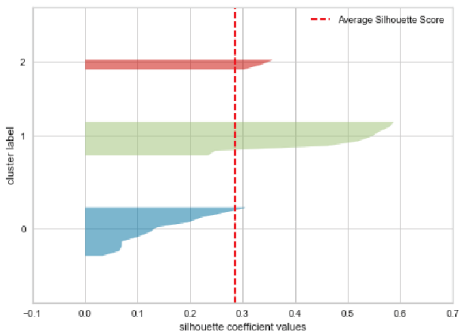
Nos três casos, as figuras (a) identificam as afiliações de padrões aos seus respectivos grupos por meio de três cores, uma para cada grupo. Além disso, por meio desta representação do agrupamento é possível fazer uma avaliação sobre a distribuição

de membros por grupos. Neste caso, é possível perceber que os agrupamentos de padrões do erro máximo e do vetor completo de erros são bem balanceados com um número similar de padrões afiliados por grupo.

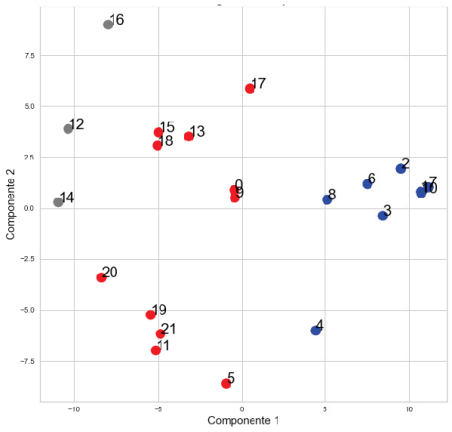
FIGURA 57 – AGRUPAMENTOS DE PADRÕES DE ERRO: SOMA DOS ERROS

E_soma
0
5
9
11
13
15
17
18
19
20
21
1
2
3
4
6
7
8
10
12
14
16
0,2861

(a)



(b)



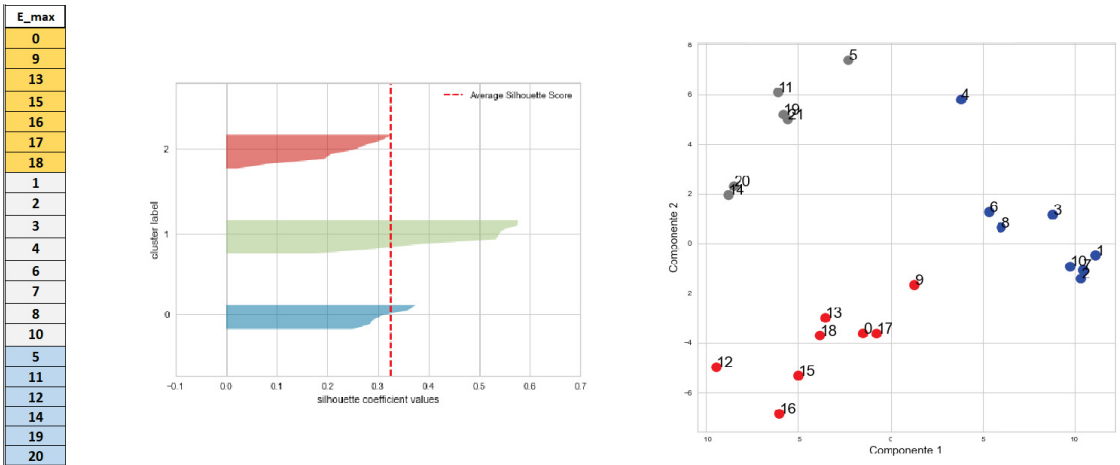
(c)

FONTE: O Autor
LEGENDA: (a) padrões com rótulos de grupos; (b) Gráfico de silhueta; (c) Projeção das componentes principais

A análise da qualidade dos agrupamentos por meio do valor do coeficiente de silhueta entre $0,24$ e $0,2$ revela que a estrutura de grupos gerada não tem as qualidades esperadas para um agrupamento ótimo. Assim, os grupos formados são difusos como pôde ser confirmado por meio dos gráficos da projeção das componentes principais nas figuras (c) para cada caso. Finalmente, a análise dos gráficos de silhueta (Figuras (b)) mostram que apenas um dos grupos em cada um dos agrupamentos ficou acima da média com valores acima de $0,5$ indicando alguma estrutura de grupo com qualidade.

De modo geral, a solução proposta para avaliação dos passeios dos treinandos por meio da modelagem dos erros cometidos e agrupamento dos padrões de erros não é satisfatória. Dessa forma, uma outra abordagem baseada no agrupamento de padrões de similaridades é proposta. Esta abordagem é baseada na modelagem do domínio do conhecimento por meio da sua representação na linguagem dos grafos.

FIGURA 58 – AGRUPAMENTOS DE PADRÕES DE ERRO: MÁXIMO ERRO



(a) (b) (c)

FONTE: O Autor
LEGENDA: (a) padrões com rótulos de grupos; (b) Gráfico de silhueta; (c) Projeção das componentes principais

5.2 ANÁLISE DE AGRUPAMENTOS DE PASSEIOS

A definição formal de agrupamentos de passeios segue Silva et al. (2016).

Definição 5.2.1 *Dados os passeios* $W = \{W_\ell\}$*, com* $\ell = 1, \cdots, L$*, um grupo de passeios* $W^{(i)} = \{W_\ell\}$ *é uma partição de* W*, com* $\ell \in [1, \cdots, L]$ *tal que*

1.

$$W = \bigcup W^{(i)} \tag{5.1}$$

2.

$$W^{(i)} \neq \varnothing, \ \forall \ i; \tag{5.2}$$

3.

$$W^{(i)} \cap W^{(j)} = \varnothing, \ com \ i, j = 1, \ 2, \ 3 \ e \ i \neq j. \tag{5.3}$$

A primeira propriedade da definição 5.2.1 diz que o agrupamento de passeios é formado por todos grupos de passeios, a segunda e a terceira propriedades exigem que os grupos não podem ser vazios e são disjuntos, respectivamente.

5.2.1 Agrupamentos dos padrões de similaridade

Sejam os passeios dos treinandos

$$W = \{W_\ell\}, \tag{5.4}$$

com $\ell = 1, \cdots, L$, e os caminhos esperados

$$P = \{P_x\}, \tag{5.5}$$

com $x = 1, \cdots, X$, como já definido.

As matrizes de similaridade dos erros dos treinandos $[S_d]$, onde

$$d \in \{d_{\lambda Adj}, d_{\lambda Lap}, d_{GED}, d_{VEO}, d_{ResN}, d_{\delta-con}, d_{Netsim}\} \tag{5.6}$$

são calculadas de dois modos:

1. Primeiro, conforme a similaridade entre o passeio executado pelo treinando e os caminhos esperados:
$$[S_{d(W \times P)}], \tag{5.7}$$
2. Segundo, de acordo com a similaridade relativa entre os passeios dos treinandos entre si:
$$[S_{d(W \times W)}]. \tag{5.8}$$

Dessa forma, um total de catorze agrupamentos derivados das medidas de similaridade são avaliados. A medida de avaliação utilizada será o coeficiente de silhueta, pois além de expressar a qualidade do agrupamento e seus respectivos grupos, esta medida avalia a aderência de cada padrão ao seu grupo.

5.2.2 Agrupamento de padrões de similaridades entre passeios e caminhos $(\mathcal{W} \times \mathcal{P})$

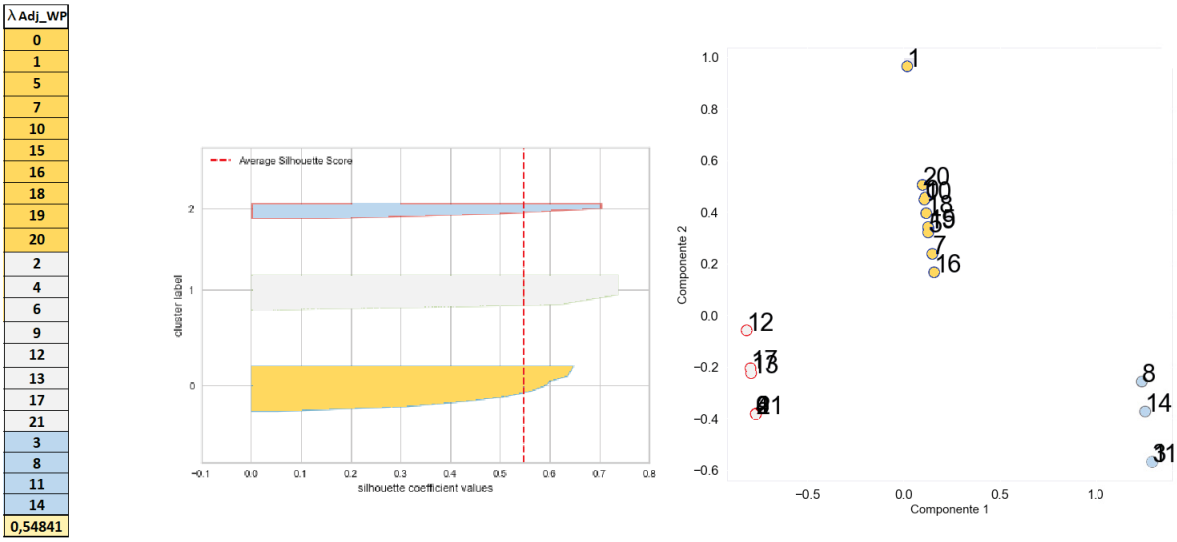
A seguir são apresentados os resultados para os agrupamentos de padrões entre passeios e caminhos. Os resultados dos agrupamentos são apresentados por meio da identificação dos rótulos de classe de cada padrão (figuras (a)), do gráfico de silhueta (figuras (b)) e o gráfico da projeção das componentes principais dos padrões em duas dimensões (figuras (c)).

Agrupamento de similaridade $Sim_{\lambda_{Adj}}(\mathcal{W} \times \mathcal{P})$

O agrupamento $Sim_{\lambda_{Adj}}(\mathcal{W} \times \mathcal{P})$ tem silhueta igual a $0,54841$ o que indica que existe uma estrutura de grupos razoável. Entretanto este valor é somente um pouco maior do limiar de $0,5$ que o definiria como um grupo com uma estrutura artificial. Conforme Rousseeuw (1987), nestes casos, os grupos formados como resultado da aplicação do método não fazem sentido (*make sense*). Um aspecto positivo do

agrupamento é que todos os grupos têm silhueta acima da média, representada por meio da linha tracejada no gráfico na FIGURA 59 (b). Em particular, os grupos 1 e 2, identificados pelas cores cinza e azul, têm silhueta acima de $0,7$ classificando-os como grupos de ótima qualidade neste agrupamento. Em relação à projeção dos grupos segundo suas componentes principais revela uma estrutura de grupos alongados. Esta estrutura se alonga sobre eixos paralelos ao eixo da Componente 2. Cada um dos eixos se apoia sobre um ponto do eixo da Componente 1, um para cada grupo. Assim, cada grupo fica bem caracterizado por membros cujas projeções são próximas do eixo do grupo. Cada eixo poderia ser definido, por exemplo, por meio do valor da Componente 1 do centroide do grupo. A dispersão das projeções dos padrões na direção destes eixos fornece uma estimativa da sua aderência ao grupo.

FIGURA 59 – AGRUPAMENTOS DE SIMILARIDADE $SIM_{\lambda_{ADJ}}(\mathcal{W} \times \mathcal{P})$

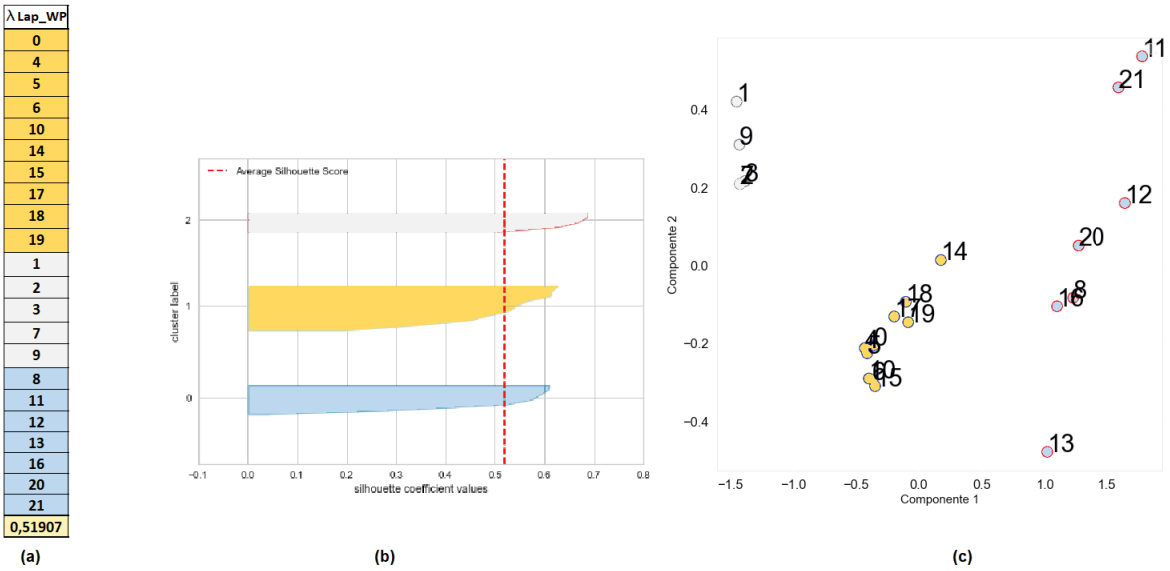


FONTE: O Autor
LEGENDA: (a) padrões com rótulos de grupos; (b) Gráfico de silhueta; (c) Projeção das componentes principais

Agrupamento de similaridade $Sim_{\lambda_{Lap}}(\mathcal{W} \times \mathcal{P})$

O agrupamento $Sim_{\lambda_{Lap}}(\mathcal{W} \times \mathcal{P})$ tem silhueta igual a $0,51907$ indicando uma estrutura razoável de grupos cujos valores de silhueta estão acima da média do grupo. Assim, a qualidade dos grupos é superior a qualidade do agrupamento. A projeção dos grupos mostra estruturas alongadas um pouco mais dispersas do que o no caso do agrupamento baseado na distância espectral adjacência.

FIGURA 60 – AGRUPAMENTOS DE SIMILARIDADE $SIM_{\lambda_{LAP}}(\mathcal{W} \times \mathcal{P})$



FONTE: O Autor
LEGENDA: (a) padrões com rótulos de grupos; (b) Gráfico de silhueta; (c) Projeção das componentes principais

Agrupamento de similaridade $Sim_{Netsimile}(\mathcal{W} \times \mathcal{P})$

O agrupamento $Sim_{Netsimile}(\mathcal{W} \times \mathcal{P})$ tem silhueta igual a $0,56943$ indicando uma estrutura razoável de grupos. Entretanto, os valores de silhueta de dois grupos estão acima da média próximos de $0,75$ o que significa que estes grupos têm uma forte estrutura de grupo. A projeção dos grupos mostra estruturas alongadas segundo sua primeira componente.

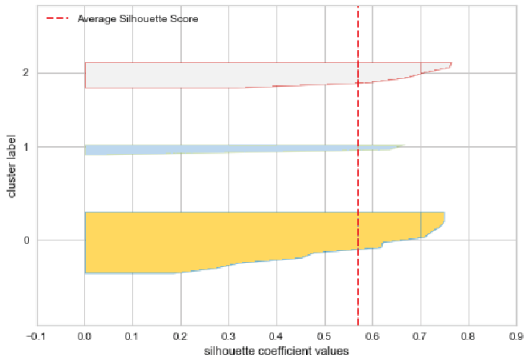
Agrupamento de similaridade $Sim_{GED}(\mathcal{W} \times \mathcal{P})$

O agrupamento $Sim_{GED}(\mathcal{W} \times \mathcal{P})$ tem silhueta igual a $0,63204$ indicando uma estrutura razoável de grupos. Os valores de silhueta dos grupos estão acima de $0,7$ com um dos grupos com silhueta acima de $0,8$, o que significa que estes grupos têm uma forte estrutura de grupo. O gráfico das projeções das componentes principais mostram estruturas alongadas de acordo com as primeiras componentes.

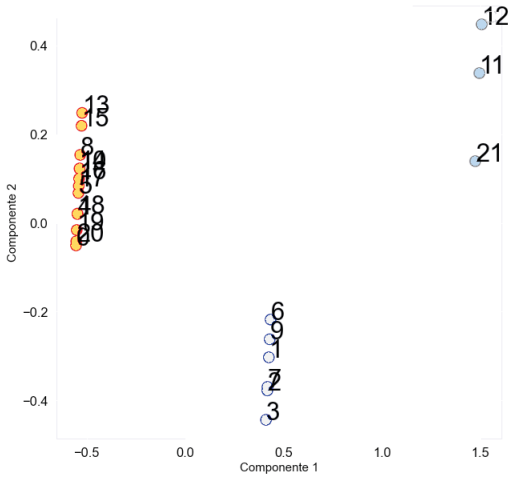
FIGURA 61 – AGRUPAMENTOS DE SIMILARIDADE $SIM_{NETSIMILE}(\mathcal{W} \times \mathcal{P})$

Ntsim_WP
0
4
5
8
10
13
14
15
16
17
18
19
20
1
2
3
6
7
9
11
12
21
0,56943

(a)



(b)



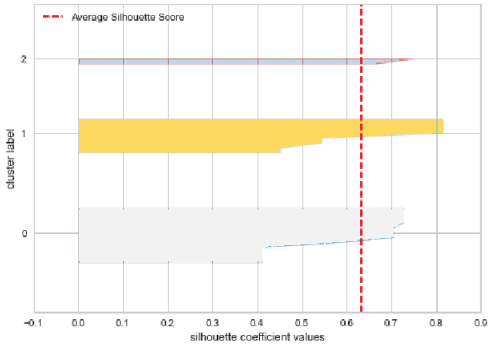
(c)

FONTE: O Autor
LEGENDA: (a) padrões com rótulos de grupos; (b) Gráfico de silhueta; (c) Projeção das componentes principais

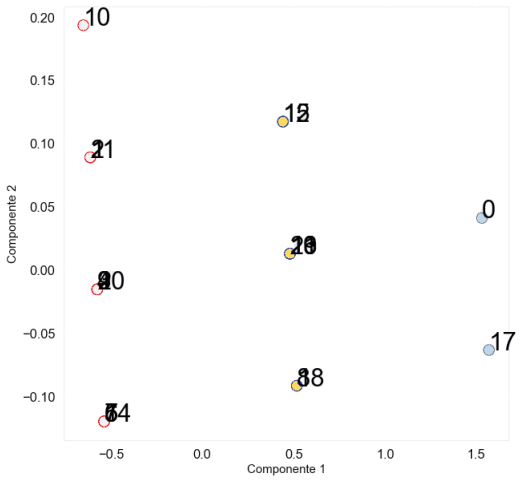
FIGURA 62 – AGRUPAMENTOS DE SIMILARIDADE $SIM_{GED}(\mathcal{W} \times \mathcal{P})$

GED_WP
8
12
13
15
16
18
19
21
1
2
3
4
5
6
7
9
10
11
14
20
0
17
0,63204

(a)



(b)



(c)

FONTE: O Autor
LEGENDA: (a) padrões com rótulos de grupos; (b) Gráfico de silhueta; (c) Projeção das componentes principais

Agrupamento de similaridade $Sim_{VEO}(\mathcal{W} \times \mathcal{P})$

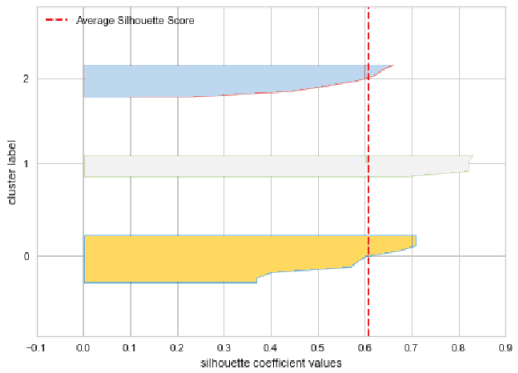
O agrupamento $Sim_{VEO}(\mathcal{W} \times \mathcal{P})$ tem uma estrutura razoável de grupos com silhueta igual a $0,60737$. Os valores de silhueta dos grupos estão acima da média com dois grupos têm forte estrutura de grupo com silhueta maiores do $0,7$ e $0,8$ cada um. O

gráfico das projeções mostra estruturas alongadas segundo a primeira componente.

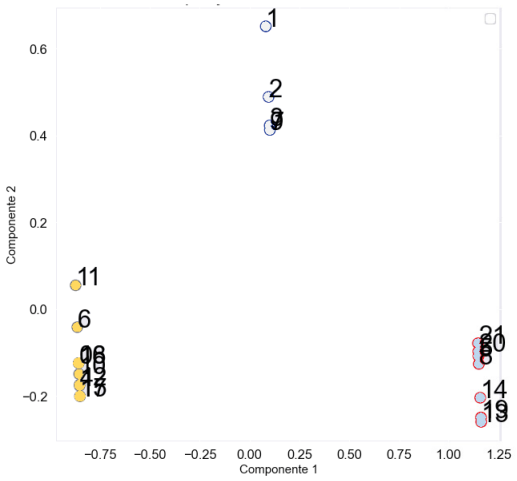
FIGURA 63 – AGRUPAMENTOS DE SIMILARIDADE $SIM_{VEO}(\mathcal{W} \times \mathcal{P})$

VEO_WP
0
4
6
10
11
12
15
16
17
18
1
2
3
7
9
5
8
13
14
19
20
21
0,60737

(a)



(b)



(c)

FONTE: O Autor
LEGENDA: (a) padrões com rótulos de grupos; (b) Gráfico de silhueta; (c) Projeção das componentes principais

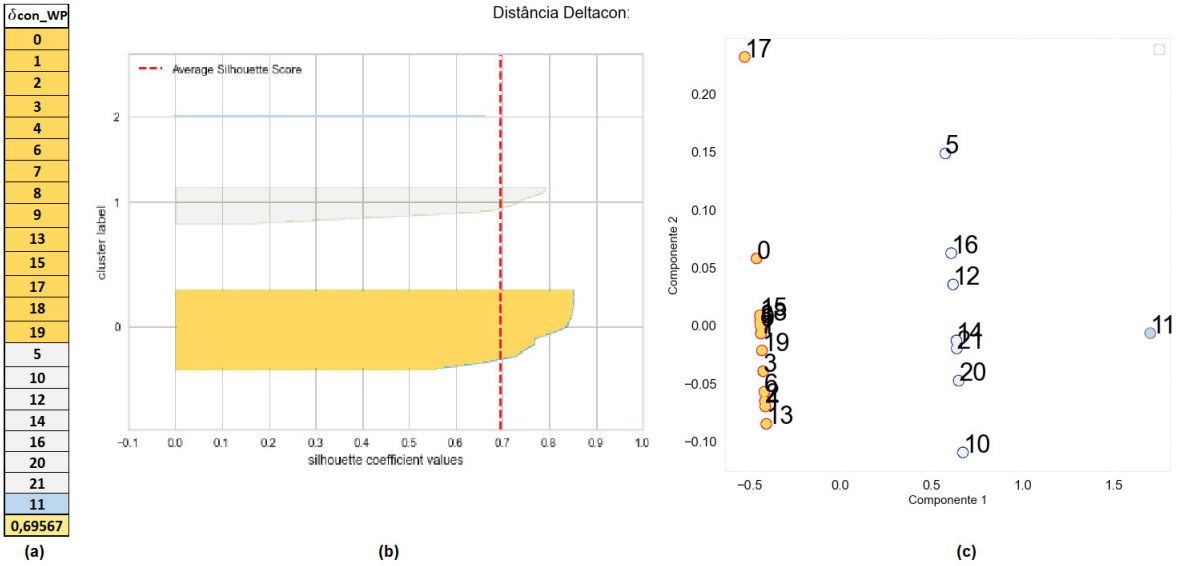
Agrupamento de similaridade $Sim_{\delta-con}(\mathcal{W} \times \mathcal{P})$

O agrupamento $Sim_{\delta-con}(\mathcal{W} \times \mathcal{P})$ tem uma estrutura razoável de grupos com silhueta igual a $0,69567$. Este valor está próximo do limiar de $0,71$ que o caracterizaria como um agrupamento com forte estrutura. Esta categoria entretanto é alcançada por dois dos três grupos. Um problema em relação à este agrupamento é a presença de um grupo com apenas um padrão, identificado na FIGURA 64 pelo número 11. A presença de um grupo com um único elemento pode indicar um padrão atípico. Entretanto, dado o conhecimento dos dados originais, não parece ser este o caso, o que poderia indicar um erro de agrupamento. Para os dois agrupamentos não unitários a projeção das duas primeiras componentes principais mostram um estrutura de grupos alongada.

Agrupamento de similaridade $Sim_{Res}(\mathcal{W} \times \mathcal{P})$

O agrupamento $Sim_{Res}(\mathcal{W} \times \mathcal{P})$ tem uma forte estrutura de grupos com silhueta igual a $0,666886$. Um dos grupos se destaca com silhueta acima de $0,9$. Novamente, neste caso os grupos apresentam uma estrutura alongada como pode ser observado por meio do gráfico da projeção de suas componentes principais em duas dimensões.

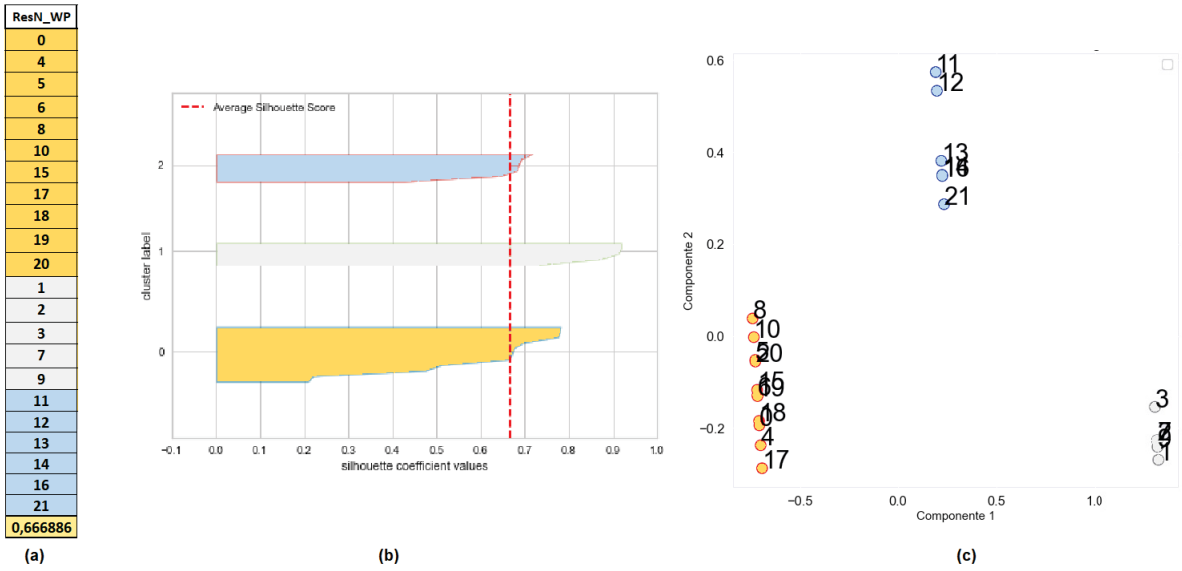
FIGURA 64 – AGRUPAMENTOS DE SIMILARIDADE $SIM_{\delta-CON}(\mathcal{W} \times \mathcal{P})$



FONTE: O Autor

LEGENDA: (a) padrões com rótulos de grupos; (b) Gráfico de silhueta; (c) Projeção das componentes principais

FIGURA 65 – AGRUPAMENTOS DE SIMILARIDADE $SIM_{RES}(\mathcal{W} \times \mathcal{P})$



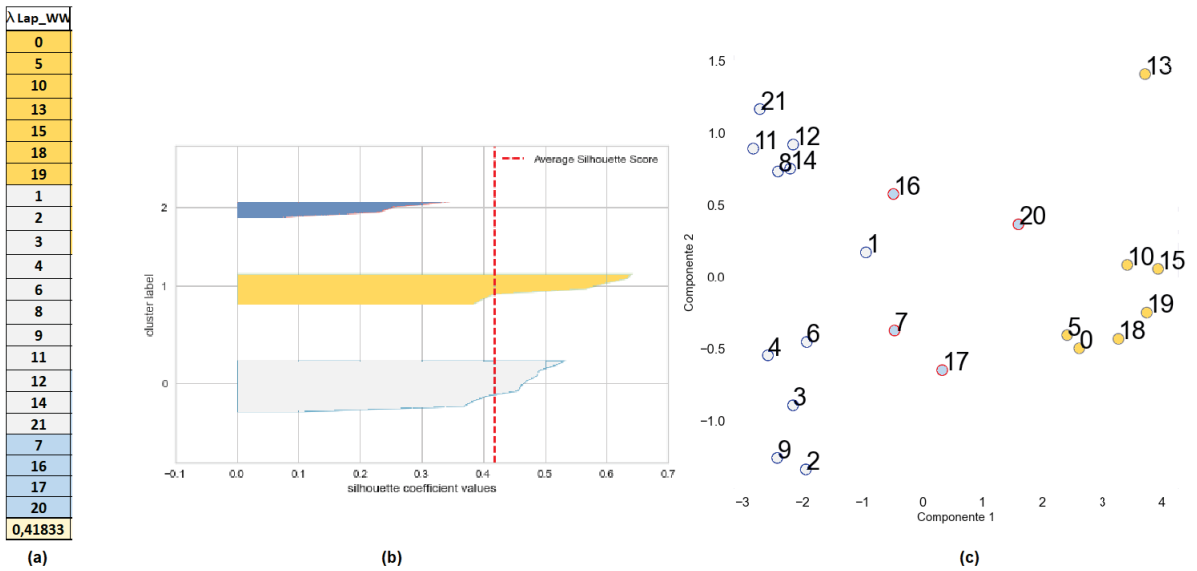
FONTE: O Autor

LEGENDA: (a) padrões com rótulos de grupos; (b) Gráfico de silhueta; (c) Projeção das componentes principais

5.2.3 Agrupamento de padrões de similaridade entre passeios $(\mathcal{W} \times \mathcal{W})$

Até aqui foram discutidos os agrupamentos de padrões de similaridade entre passeios e caminhos. A seguir são apresentados os agrupamentos de padrões de similaridade entre passeios. Ao contrário dos agrupamentos de padrões entre passeios e caminhos, aqueles agrupamentos de padrões de similaridade entre passeios apresentaram resultados inferiores para o coeficiente de silhueta. Os seis agrupamentos seguintes têm silhueta abaixo de $0,5$ o que resulta na interpretação de que não há evidência de que exista uma estrutura de grupos no agrupamento. Os quatro piores agrupamentos nestas condições são os seguintes apresentado em ordem decrescente do valor de silhueta: $Sim_{\lambda_{Lap}}(\mathcal{W} \times \mathcal{W})$ com $0,41$ (FIGURA 66), $Sim_{Res}(\mathcal{W} \times \mathcal{W})$ com $0,37$ (FIGURA 67), $Sim_{\delta-con}(\mathcal{W} \times \mathcal{W})$ com $0,31$ (FIGURA 68) e $Sim_{GED}(\mathcal{W} \times \mathcal{W})$ com $0,24$ (FIGURA 69).

FIGURA 66 – AGRUPAMENTOS DE SIMILARIDADE $SIM_{\lambda_{LAP}}(\mathcal{W} \times \mathcal{W})$



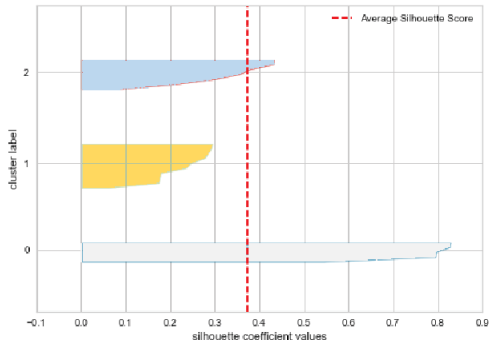
FONTE: O Autor
LEGENDA: (a) padrões com rótulos de grupos; (b) Gráfico de silhueta; (c) Projeção das componentes principais

Os gráficos das projeções das componentes principais destes agrupamentos refletem os baixos valores de silhueta representado os padrões por meio de grupos

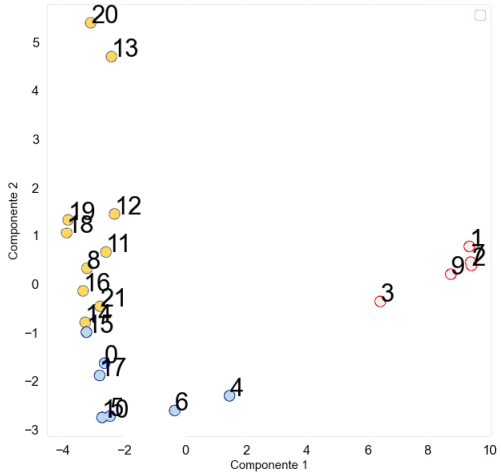
FIGURA 67 – AGRUPAMENTOS DE SIMILARIDADE $SIM_{RES}(\mathcal{W} \times \mathcal{W})$

ResN._WW
8
11
12
13
14
16
18
19
20
21
1
2
3
7
9
0
4
5
6
10
15
17
0,373935

(a)



(b)



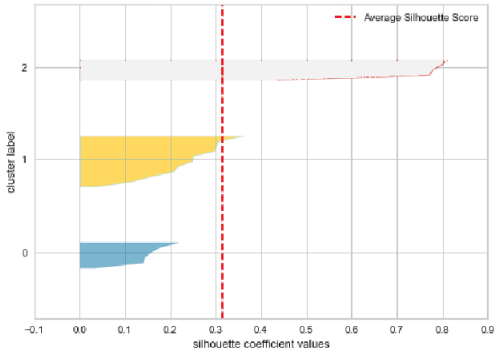
(c)

FONTE: O Autor
LEGENDA: (a) padrões com rótulos de grupos; (b) Gráfico de silhueta; (c) Projeção das componentes principais

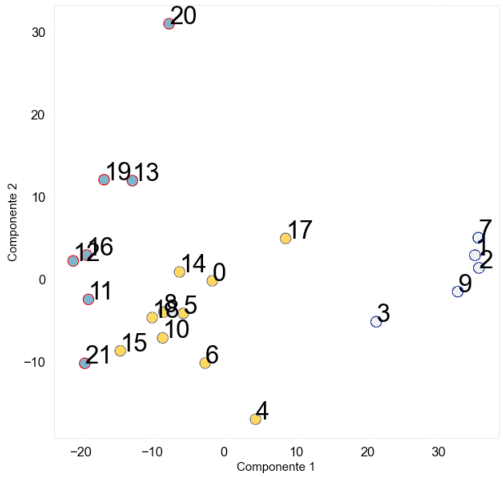
FIGURA 68 – AGRUPAMENTOS DE SIMILARIDADE $SIM_{\delta-CON}(\mathcal{W} \times \mathcal{W})$

δcon_WW
0
4
5
6
8
10
14
15
17
18
21
1
2
3
7
9
11
12
13
16
19
20
0,31584

(a)



(b)

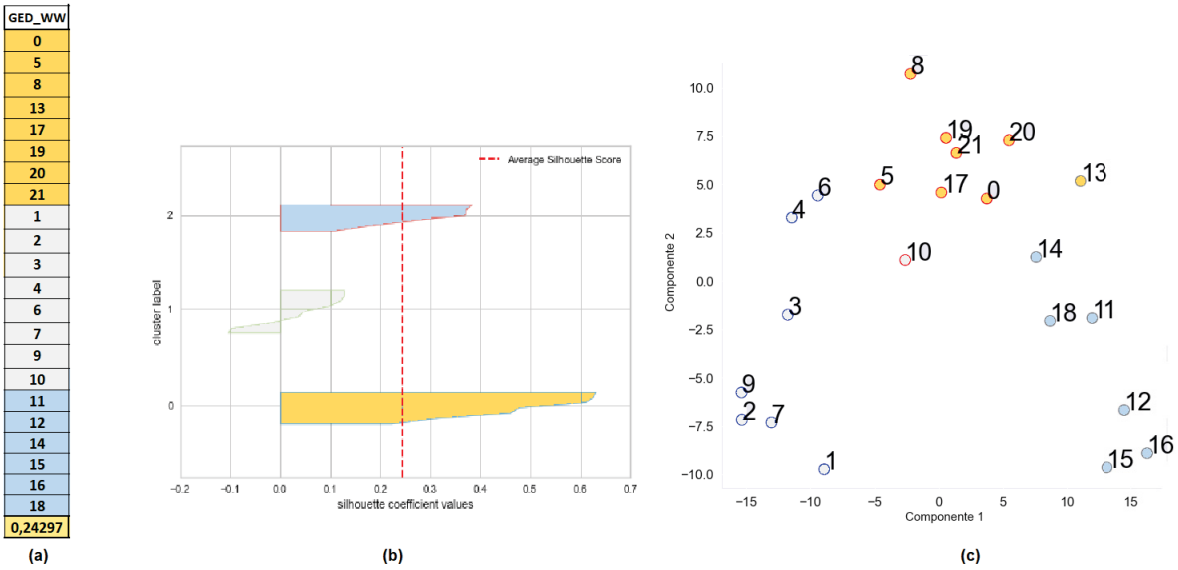


(c)

FONTE: O Autor
LEGENDA: (a) padrões com rótulos de grupos; (b) Gráfico de silhueta; (c) Projeção das componentes principais

difusos com fronteiras mal definidas. De modo geral, a medida que os valores de silhueta diminuem para os agrupamentos apresentados é possível perceber que alguns grupos em seus respectivos agrupamentos apresentam silhueta abaixo da média. As exceções a esta regra são os dois primeiros agrupamentos $Sim_{Netsimile}(\mathcal{W} \times \mathcal{W})$

FIGURA 69 – AGRUPAMENTOS DE SIMILARIDADE $SIM_{GED}(\mathcal{W} \times \mathcal{W})$



FONTE: O Autor
LEGENDA: (a) padrões com rótulos de grupos; (b) Gráfico de silhueta; (c) Projeção das componentes principais

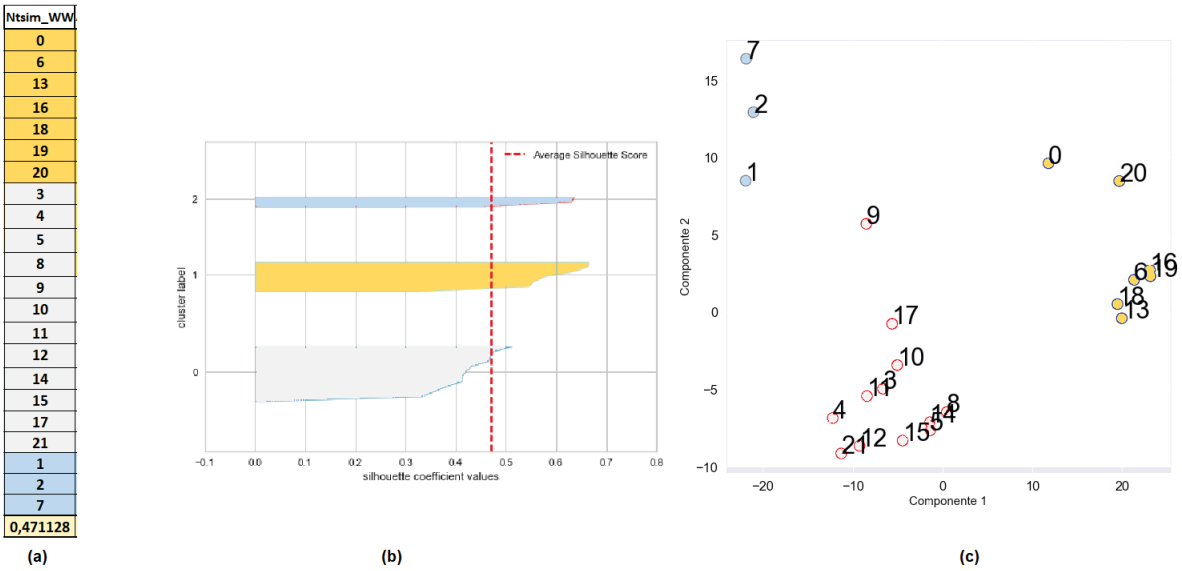
e $Sim_{\lambda_{Adj}}(\mathcal{W} \times \mathcal{W})$. Como pode ser observado nas figuras 70 e 71, nestes casos, embora não sejam tão coesos os grupos formados são relativamente bem separados com uma estrutura de grupos razoável cujo valor de silhueta está acima da média do agrupamento.

O único agrupamento de padrões de similaridade entre passeios que apresentou silhueta acima de $0,5$ foi o agrupamento $Sim_{VEO}(\mathcal{W} \times \mathcal{W})$ com $0,648$. Entretanto, dois grupos com número reduzido de membros deste agrupamento têm silhueta abaixo da média e um grupo com 75% de padrões afiliados tem membros com silhueta acima de $0,8$. A projeção dos padrões segundo suas componentes na FIGURA 72 mostra um grupo com grande quantidade de membros e com uma estrutura alongada. Os dois grupos menores em números de membros do agrupamento são difusos e seus membros estão dispersos, distantes um dos outros.

5.2.4 Tipos de agrupamentos de passeios

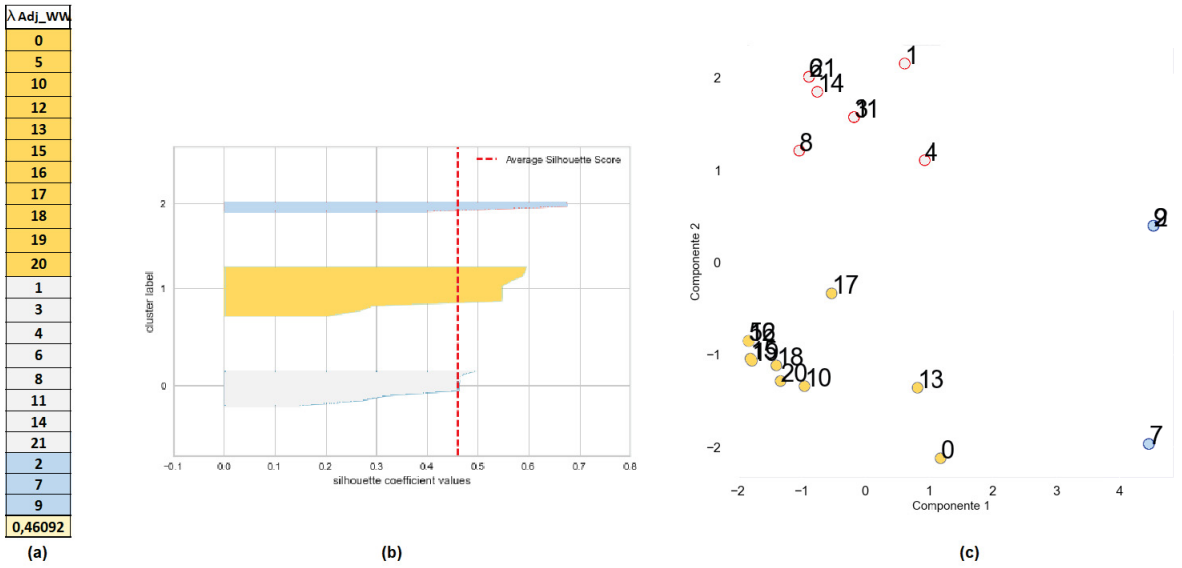
A comparação entre os coeficientes de silhueta e a dispersão das projeções dos dados nos gráficos da *PCA* permite dividir o desempenho dos agrupamentos

FIGURA 70 – AGRUPAMENTOS DE SIMILARIDADE $SIM_{NETSIMILE}(\mathcal{W} \times \mathcal{W})$



FONTE: O Autor
LEGENDA: (a) padrões com rótulos de grupos; (b) Gráfico de silhueta; (c) Projeção das componentes principais

FIGURA 71 – AGRUPAMENTO DE SIMILARIDADE $SIM_{\lambda_{ADJ}}(\mathcal{W} \times \mathcal{W})$

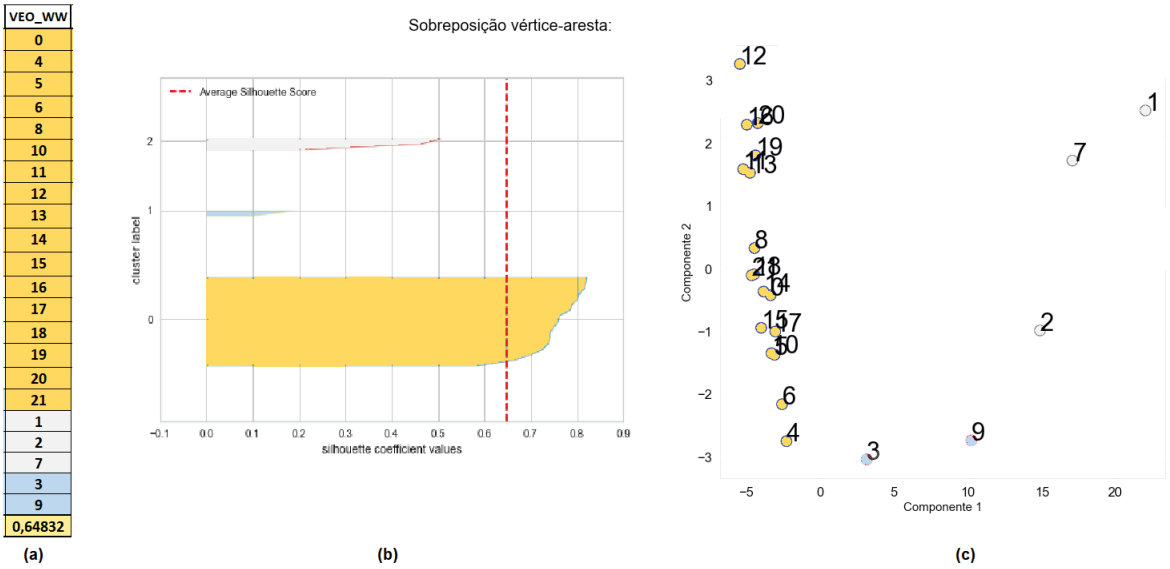


FONTE: O Autor
LEGENDA: (a) padrões com rótulos de grupos; (b) Gráfico de silhueta; (c) Projeção das componentes principais

baseados em similaridades em três tipos. Conforme os gráficos da FIGURA 73

Tipo 1 no primeiro tipo, formado a partir de agrupamentos de padrões de similaridade entre passeios e caminhos ($(\mathcal{W} \times \mathcal{P})$), cujos coeficientes de silhueta média estão acima de $0,6$ e os grupos formados em seus respectivos agrupamentos são

FIGURA 72 – AGRUPAMENTOS DE SIMILARIDADE $SIM_{VEO}(\mathcal{W} \times \mathcal{W})$



FONTE: O Autor
LEGENDA: (a) padrões com rótulos de grupos; (b) Gráfico de silhueta; (c) Projeção das componentes principais

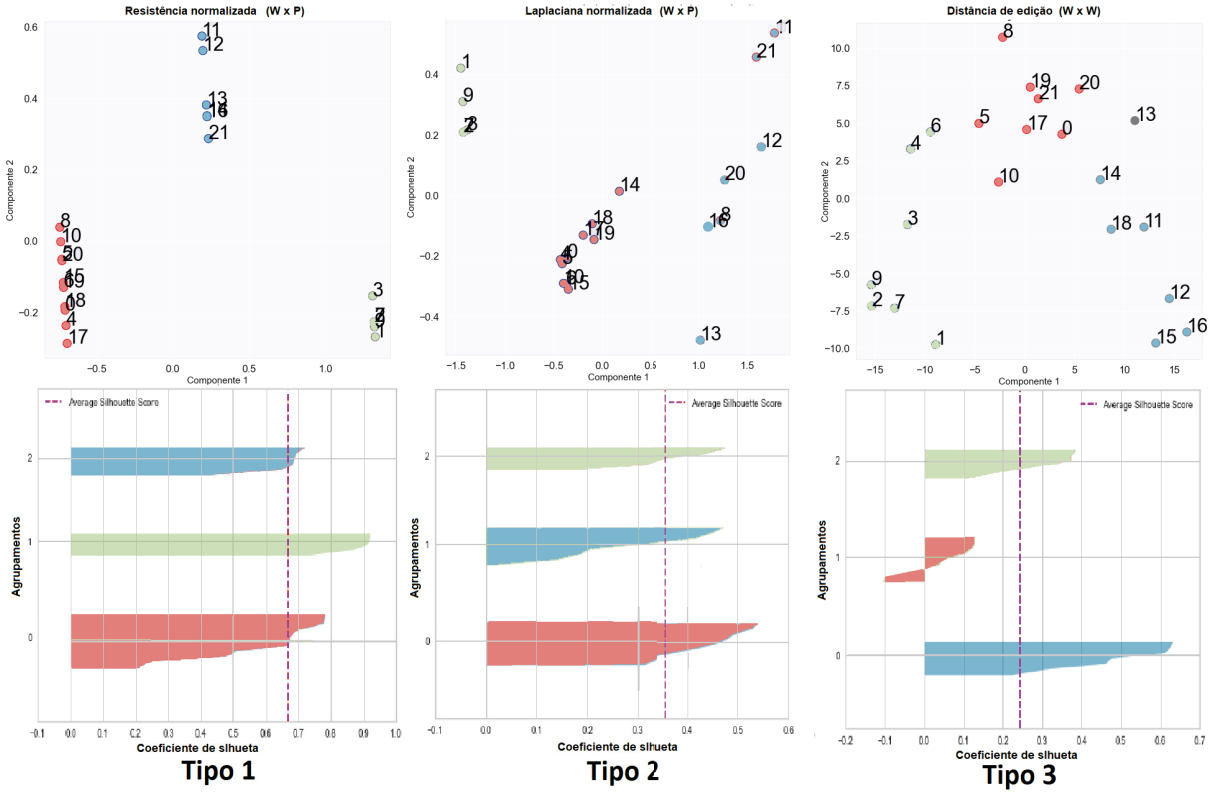
relativamente densos e bem separados com grupos alongados na direção de eixos apoiados sobre os valores da primeira componente no eixo das abscissas;

Tipo 2 no segundo tipo de agrupamento valores de silhueta médio variam de $0,4$ a $0,64$. Nestes agrupamentos, alguns grupos são coesos com qualidade superior, mas os outros são difusos.

Tipo 3 no terceiro tipo de agrupamento, valores médios de silhueta variando entre $0,4$ e $0,2$, e a presença de grupos com silhueta abaixo da média, indicando que os passeios não foram agrupados corretamente.

Na figura 74 os agrupamentos são mostrados de modo contíguo uns aos outros. As células de cores –amarelo, cinza e azul –, identificam os três grupos em cada coluna de agrupamento. As colunas estão ordenadas da esquerda para direita de acordo com o valor do coeficiente de silhueta médio do agrupamento. Os valores de silhueta mais altos foram obtidos por meio do agrupamento de padrões entre passeios e caminhos do **tipo 1**. Nesta categoria estão os agrupamentos $Sim_{Res}(\mathcal{W} \times \mathcal{P})$, $Sim_{\delta-con}(\mathcal{W} \times \mathcal{P})$, $Sim_{GED}(\mathcal{W} \times \mathcal{P})$ e $Sim_{VEO}(\mathcal{W} \times \mathcal{P})$.

FIGURA 73 – AGRUPAMENTOS POR SIMILARIDADES



FONTE: O Autor

FIGURA 74 – AGRUPAMENTOS DE PASSEIOS POR SIMILARIDADE

	Medidas de similaridade														Padrões de erros		
	δcon_WP	ResN_WP	VEO_WW	GED_WP	VEO_WP	Ntsim_WP	λ Adj_WP	λ Lap_WP	Ntsim_WW	λ Adj_WW	λ Lap_WW	ResN._WW	δcon_WW	GED_WW	E_max	E_soma	$E_{comp.}$
Agrupamentos de passeios	0	0	0	8	0	0	0	0	0	0	0	8	0	0	0	0	0
	1	4	4	12	4	4	1	4	6	5	5	11	4	5	9	5	9
	2	5	5	13	6	5	5	5	13	10	10	12	5	8	13	9	12
	3	6	6	15	10	8	7	6	16	12	13	13	6	13	15	11	13
	4	8	8	16	11	10	10	10	18	13	15	14	8	17	16	13	15
	6	10	10	18	12	13	15	14	19	15	18	16	10	19	17	15	16
	7	15	11	19	15	14	16	15	20	16	19	18	14	20	18	17	17
	8	17	12	21	16	15	18	17	3	17	1	19	15	21	1	18	18
	9	18	13	1	17	16	19	18	4	18	2	20	17	1	2	19	1
	13	19	14	2	18	17	20	19	5	19	3	21	18	2	3	20	2
	15	20	15	3	1	18	2	1	8	20	4	1	21	3	4	21	3
	17	1	16	4	2	19	4	2	9	1	6	2	1	4	6	1	4
	18	2	17	5	3	20	6	3	10	3	8	3	2	6	7	2	6
	19	3	18	6	7	1	9	7	11	4	9	7	3	7	8	3	7
	5	7	19	7	9	2	12	9	12	6	11	9	7	9	10	4	8
	10	9	20	9	5	3	13	8	14	8	12	0	9	10	5	6	10
	12	11	21	10	8	6	17	11	15	11	14	4	11	11	11	7	5
	14	12	1	11	13	7	21	12	17	14	21	5	12	12	12	8	11
	16	13	2	14	14	9	3	13	21	21	7	6	13	14	14	10	14
	20	14	7	20	19	11	8	16	1	2	16	10	16	15	19	12	19
	21	16	3	0	20	12	11	20	2	7	17	15	19	16	20	14	20
	11	21	9	17	21	21	14	21	7	9	20	17	20	18	21	16	21
Coef. de Silhueta	0,69567	0,666886	0,64832	0,63204	0,60737	0,56943	0,54841	0,51907	0,471128	0,46092	0,41833	0,373935	0,31584	0,24297	0,32466	0,2861	0,24039

FONTE: O Autor

Por meio da FIGURA 74 é possível avaliar como os padrões estão distribuídos entre os grupos dos agrupamentos. Assim, embora os grupos em cada agrupamento apresentem um número diverso de membros, existem grupos cujos membros compartilham o mesmo rótulo ainda que em agrupamentos diferentes. Esta característica

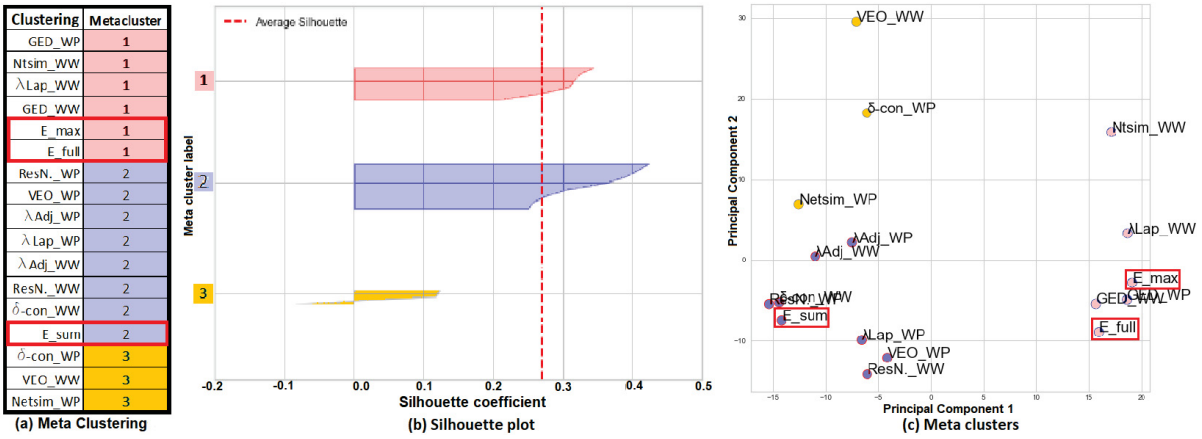
permite afirmar que, embora diferentes, os agrupamentos não são altamente diversos.

O agrupamentos com uma diversidade de grupos variando de média a baixa e boa qualidade são utilizados para a análise subsequente de meta-agrupamento. Nesta análise os agrupamentos de padrões de similaridades são comparados com agrupamentos de padrões de erros. Conforme os objetivos desta tese, a análise de meta-agrupamento deve identificar aqueles agrupamentos de padrões de similaridade que se aproximam dos agrupamentos de padrões de erro.

5.3 ANÁLISE DE META-AGRUPAMENTO: PADRÕES DE SIMILARIDADE $\times$ PADRÕES DE ERRO

A análise de meta-agrupamentos identifica quais agrupamentos de padrões de similaridade $\overline{W}_{S_d}$ podem ser uma aproximação para os agrupamentos $\overline{W}_E$ de padrões de erros cometidos durante a execução da tarefa pelos treinandos. Considerando todos os agrupamentos obtidos definimos o conjunto de agrupamentos base por meio da união $\overline{W} = \overline{W}_{S_d} \cup \overline{W}_E$. Uma vez mais usando *K-means*, o meta-agrupamento é aplicado a $\overline{W}^T$. As Figuras 75 (a) e (c) apresentam a distribuição dos rótulos dos três metagrupos e suas projeções por meio das componentes principais. Em ambos os casos, os agrupamentos baseados em padrões de erro são emoldurados.

FIGURA 75 – META-AGRUPAMENTO



FONTE: O Autor
LEGENDA: (a) metagrupos identificados por seus rótulos; (b) gráfico de silhueta; (c) gráfico da projeção das componentes principais

O coeficiente médio de silhueta dos metagrupos é a linha vertical tracejada na FIGURA 75 (b). Os resultados da análise de meta-agrupamento revelam um valor baixo para o coeficiente de silhueta médio, abaixo de $0,3$. Isso significa que os metagrupos

não são densos, como mostrado na FIGURA 75 (c). Apesar disso, para o primeiro e o segundo metagrupos, assim como para a maioria de seus membros, os coeficientes de silhueta estão acima da média como é mostrado na FIGURA 75 (b). Agrupamentos baseados em padrões de erro, emoldurados na FIGURA 75 (a) e na FIGURA 75 (c),estão neste metagrupo.

Os agrupamentos de padrões de erro $\overline{W}_{E_{max}}$ e $\overline{W}_{E_{completo}}$ estão no metagrupo 1 e $\overline{W}_{E_{sum}}$ no metagrupo 2. Os agrupamentos baseados em similaridades $\overline{W}_S$ mais próximos de $\overline{W}_{E_{max}}$ e $\overline{W}_{E_{max}}$ no metagrupo 1 são os agrupamentos base das similaridades distância de edição entre caminhos e entre caminhos e passeios. No metagrupo 2, o agrupamento baseado na distância de resistência e a similaridade Deltacon entre passeios e caminhos são os agrupamentos mais próximos de $\overline{W}_{E_{sum}}$, o vetor de erro de soma de linhas das matrizes de passeio dos treinandos de erro.

Finalmente, entre os 14 agrupamentos de padrões de similaridade, apenas as similaridades citadas alcançaram a aproximação visada com os padrões de erro dos treinandos. Além disso, os coeficientes de silhueta e a estrutura de alguns dos grupos destes agrupamentos obtiveram melhores resultados do que os outros grupos nos agrupamentos restantes.

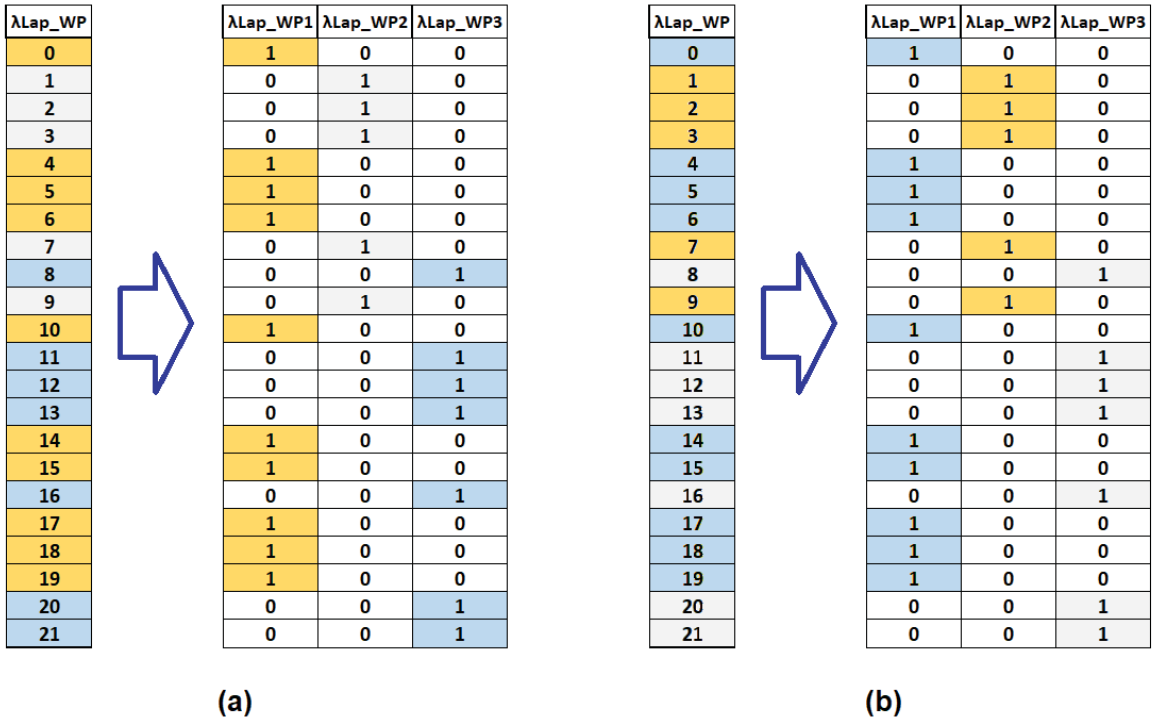
Em relação às distâncias em que se baseiam as definições de similaridades associadas a estes agrupamentos, enquanto a distância de edição, GED, calcula similaridades topológicas através de operações sobre o grafos, tais como adição e exclusão de vértices ou arestas, a resistência normalizada e Deltacon capturam o fluxo de informações em sua propagação sobre o grafo. Quanto à sensibilidade, enquanto o GED pode detectar mudanças locais, a Resistência Normalizada é aplicada às divergências globais na comparação entre os grafos. A similaridade calculada por Deltacon é sensível a ambos.

Os resultados sugerem que os padrões de erro podem ser analisados por meio das propriedades topológicas dos grafos. Como já discutido, outros fatores como o espaço dos padrões de dados, o método de agrupamento e a definição de similaridade entre dois pontos no espaço de dados influenciam os resultados dos agrupamentos. Neste trabalho, como o método e os espaço de dados são os mesmos para cada agrupamento de padrão de similaridade, considera-se que as diferenças entre estes agrupamentos sejam derivadas da definição de similaridade em cada caso. Dessa forma, uma vez que cada definição de similaridade captura um aspecto diferente na comparação entre grafos, a sua combinação pode gerar um agrupamento consenso com qualidade superior aos agrupamentos base utilizados na sua definição. Esta análise é apresentada a seguir.

5.4 COMBINAÇÃO DE AGRUPAMENTOS

Para a geração do agrupamento consenso, os agrupamentos são decompostos segundo os rótulos dos membros dos seus grupos. Assim, neste caso, como o número de grupos em todos os agrupamentos é três, cada agrupamento resulta em três vetores. Assim, como já apresentado na Fundamentação Teórica, este procedimento permite comparar grupos que se diferenciam pelos seus rótulos, mas não pelos seus membros. Um caso hipotético é ilustrado na FIGURA 76 para o agrupamento de padrões de similaridade $Sim_{\lambda_{Lap}}(\mathcal{W} \times \mathcal{P})$. Neste exemplo, na FIGURA 76 (a) o método de agrupamento rotulou os grupos como 1, 2 e 3. Nesta ordem, os membros dos grupos são identificados por suas cores: amarelo, cinza e azul, respectivamente. na FIGURA 76 (b) os mesmos grupos foram rotulados como 3, 1 e 2. Assim, exceto por seus rótulos, os agrupamentos do exemplo são iguais.

FIGURA 76 – DECOMPOSIÇÃO DO AGRUPAMENTO NO ESPAÇO DE RÓTULOS COMO VETORES DE FILIAÇÃO AOS GRUPOS



λLap_WP
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21

λLap_WP1	λLap_WP2	λLap_WP3
1	0	0
0	1	0
0	1	0
0	1	0
1	0	0
1	0	0
1	0	0
0	1	0
0	0	1
0	1	0
1	0	0
0	0	1
0	0	1
0	0	1
1	0	0
1	0	0
0	0	1
1	0	0
1	0	0
1	0	0
0	0	1
0	0	1

(a)

λLap_WP
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21

λLap_WP1	λLap_WP2	λLap_WP3
1	0	0
0	1	0
0	1	0
0	1	0
1	0	0
1	0	0
1	0	0
0	1	0
0	0	1
0	1	0
1	0	0
0	0	1
0	0	1
0	0	1
1	0	0
1	0	0
0	0	1
1	0	0
1	0	0
1	0	0
0	0	1
0	0	1

(b)

FONTE: O Autor
LEGENDA: Permutação dos nomes dos rótulos de classe entre (a) e (b))

Por meio desta representação a combinação dos agrupamentos é realizada sobre uma matriz cujas linhas representam os padrões de dados e as colunas indicam a filiação do padrão a um determinado grupo no agrupamento correspondente. Na

TABELA 13, os vinte e dois caminhos dos treinandos são representados por seus atributos de filiação a um determinado grupo em cada um dos agrupamentos selecionados para a combinação, como no exemplo apresentado os agrupamentos de padrões de similaridade Resistência Normalizada, sobreposição vértice-aresta e distância de edição. Como já discutido, o agrupamento consenso é gerado com base na maximização de uma função consenso que avalia a similaridade entre os padrões de rótulos associados aos caminhos dos treinandos.

TABELA 13 – REPRESENTAÇÃO DE PASSEIOS NO ESPAÇO DE RÓTULOS DE GRUPOS

$\mathcal{W}$	Res			Veo			Ged		
	RNWP1	**RNWP2**	**RNWP3**	**SVAWW1**	**SVAWW2**	**SVAWW3**	**DEWP1**	**DEWP2**	**DEWP3**
$\mathcal{W}_1$	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**1**
$\mathcal{W}_2$	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**1**	**0**
$\mathcal{W}_3$	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**1**	**0**
$\mathcal{W}_4$	**0**	**1**	**0**	**0**	**0**	**1**	**0**	**1**	**0**
$\mathcal{W}_5$	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**1**	**0**
$\mathcal{W}_6$	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**1**	**0**
$\mathcal{W}_7$	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**1**	**0**
$\mathcal{W}_8$	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**1**	**0**
$\mathcal{W}_9$	**1**	**0**	**0**	**1**	**0**	**0**	**1**	**0**	**0**
$\mathcal{W}_{10}$	**0**	**1**	**0**	**0**	**0**	**1**	**0**	**1**	**0**
$\mathcal{W}_{11}$	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**1**	**0**
$\mathcal{W}_{12}$	**0**	**0**	**1**	**1**	**0**	**0**	**0**	**1**	**0**
$\mathcal{W}_{13}$	**0**	**0**	**1**	**1**	**0**	**0**	**1**	**0**	**0**
$\mathcal{W}_{14}$	**0**	**0**	**1**	**1**	**0**	**0**	**1**	**0**	**0**
$\mathcal{W}_{15}$	**0**	**0**	**1**	**1**	**0**	**0**	**0**	**1**	**0**
$\mathcal{W}_{16}$	**1**	**0**	**0**	**1**	**0**	**0**	**1**	**0**	**0**
$\mathcal{W}_{17}$	**0**	**0**	**1**	**1**	**0**	**0**	**1**	**0**	**0**
$\mathcal{W}_{18}$	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**1**
$\mathcal{W}_{19}$	**1**	**0**	**0**	**1**	**0**	**0**	**1**	**0**	**0**
$\mathcal{W}_{20}$	**1**	**0**	**0**	**1**	**0**	**0**	**1**	**0**	**0**
$\mathcal{W}_{21}$	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**1**	**0**
$\mathcal{W}_{22}$	**0**	**0**	**1**	**1**	**0**	**0**	**1**	**0**	**0**

FONTE: O Autor

A combinação dos agrupamentos base e geração do agrupamento consenso foi implementada por meio da biblioteca *DiceR* desenvolvida para a linguagem R (CHIU; TALHOUK, 2018). A biblioteca disponibiliza funções de agrupamento dos dados para criação dos agrupamentos base e funções para o cálculo e avaliação do agrupamento consenso. O consenso pode ser calculado para doze métodos de agrupamento diferentes, incluindo *K-means*, métodos hierárquicos e baseados em modelos de mistura. Para a função consenso da combinação são disponibilizados os métodos *K-modes*[1], votação majoritária e agrupamento hierárquico aglomerativo, dentre outros. A avaliação do agrupamento consenso pode ser realizada por meio de quinze medidas internas para a validação relativa. Dentre estas medidas estão aquelas já apresentadas como

[1] O método *K-modes* é uma versão do *K-means* adaptado para o agrupamento de dados categóricos onde os centroides dos grupos são determinados por meio da moda dos membros do grupos (CHATURVEDI et al., 2001).

os índices de Calinski-Harabasz, Davies-Bouldin e o coeficiente de silhueta (CHIU; TALHOUK, 2018).

Assim, como no caso dos métodos de agrupamentos de padrões de similaridade ou padrões de erros, a estratégia de geração do agrupamento consenso foi baseada em diferentes combinações de agrupamentos de boa qualidade e com diversidade de média a baixa. De acordo com as análises dos agrupamentos e meta-agrupamento, os seguintes agrupamentos foram selecionados para compor as combinações

- $Sim_{\delta-con}(\mathcal{W} \times \mathcal{P})$
- $Sim_{Res}(\mathcal{W} \times \mathcal{P})$
- $Sim_{VEO}(\mathcal{W} \times \mathcal{P})$
- $Sim_{GED}(\mathcal{W} \times \mathcal{P})$

Como já discutido, estes agrupamentos são do **Tipo 1**, com coeficiente de silhueta acima de $0,6$ e uma estrutura de grupos alongados e bem separados. Além disto, conforme a análise de meta-agrupamentos, estes agrupamentos são aqueles que mais se aproximam dos agrupamentos de padrões de erros. A partir destes agrupamentos foram validadas as seguintes combinações:

- $Sim_{Res}(\mathcal{W} \times \mathcal{P}) \times Sim_{\delta-con}(\mathcal{W} \times \mathcal{P})$
- $Sim_{Res}(\mathcal{W} \times \mathcal{P}) \times Sim_{VEO}(\mathcal{W} \times \mathcal{P})$
- $Sim_{\delta-con}(\mathcal{W} \times \mathcal{P}) \times Sim_{VEO}(\mathcal{W} \times \mathcal{P})$
- $Sim_{\delta-con}(\mathcal{W} \times \mathcal{P}) \times Sim_{GED}(\mathcal{W} \times \mathcal{P})$
- $Sim_{GED}(\mathcal{W} \times \mathcal{P}) \times Sim_{VEO}(\mathcal{W} \times \mathcal{P})$
- $Sim_{Res}(\mathcal{W} \times \mathcal{P}) \times Sim_{GED}(\mathcal{W} \times \mathcal{P})$
- $Sim_{Res}(\mathcal{W} \times \mathcal{P}) \times Sim_{GED}(\mathcal{W} \times \mathcal{P}) \times Sim_{VEO}(\mathcal{W} \times \mathcal{P})$
- $Sim_{\delta-con}(\mathcal{W} \times \mathcal{P}) \times Sim_{GED}(\mathcal{W} \times \mathcal{P}) \times Sim_{VEO}(\mathcal{W} \times \mathcal{P})$
- $Sim_{Res}(\mathcal{W} \times \mathcal{P}) \times Sim_{\delta-con}(\mathcal{W} \times \mathcal{P}) \times Sim_{VEO}(\mathcal{W} \times \mathcal{P})$
- $Sim_{Res}(\mathcal{W} \times \mathcal{P}) \times Sim_{GED}(\mathcal{W} \times \mathcal{P}) \times Sim_{\delta-con}(\mathcal{W} \times \mathcal{P}) \times Sim_{VEO}(\mathcal{W} \times \mathcal{P})$

Combinações com um número maior de agrupamentos foram testadas adicionando agrupamentos do **Tipo 2** à análise. Entretanto, os resultados foram inferiores aqueles apresentados nas tabelas 16, 15 e 14. Como pode ser percebido, neste caso,

a inclusão de mais agrupamentos à combinação resultou em uma qualidade inferior em relação à qualidade dos agrupamentos base. Para cada uma das combinações, dois métodos foram utilizados para a função consenso: o *K-modes* (KM) e o agrupamento hierárquico baseado na ligação média (*Average linkage*). A avaliação dos agrupamentos consenso foi realizada utilizando os índices Calinski-Harabasz (CH), Davies-Bouldin (DB) e o coeficiente de silhueta (Sil). De modo geral, as medidas de validação não apresentaram divergências em relação à avaliação dos dois consensos gerados para cada combinação proposta.

TABELA 14 – CONSENSO DAS COMBINAÇÕES DE TRÊS AGRUPAMENTOS

	RESGEDVEO				GEDVEODELTA				GEDRESDELTA				RESGEDVEODELTA			
	HC	**Passeios**	**KM**	**Passeios**	**HC**	**Passeios**	**KM**	**Passeios**	**HC**	**Passeios**	**KM**	**Passeios**	**HC**	**Passeios**	**KM**	**Passeios**
	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**
	0	**4**	0	**4**	0	**8**	0	**1**	0	**4**	0	**4**	0	**1**	0	**8**
	0	**5**	0	**6**	0	**12**	0	**2**	0	**5**	0	**6**	0	**2**	0	**15**
	0	**6**	0	**10**	0	**13**	0	**3**	0	**6**	0	**10**	0	**3**	0	**17**
	0	**8**	0	**11**	0	**15**	0	**4**	0	**8**	0	**11**	0	**4**	0	**18**
	0	**10**	0	**12**	0	**16**	0	**6**	0	**10**	0	**12**	0	**6**	0	**19**
	0	**11**	0	**15**	0	**17**	0	**7**	0	**11**	0	**15**	0	**7**	1	**1**
	0	**14**	0	**16**	0	**18**	0	**9**	0	**14**	0	**16**	0	**8**	1	**2**
	0	**15**	0	**17**	0	**19**	0	**11**	0	**15**	0	**17**	0	**9**	1	**3**
	0	**17**	0	**18**	0	**21**	0	**17**	0	**17**	0	**18**	0	**15**	1	**4**
	0	**18**	1	**1**	1	**1**	1	**5**	0	**18**	1	**1**	0	**17**	1	**5**
	0	**19**	1	**2**	1	**2**	1	**8**	0	**19**	1	**2**	0	**18**	1	**6**
	0	**20**	1	**3**	1	**3**	1	**13**	0	**20**	1	**3**	0	**19**	1	**7**
	1	**1**	1	**7**	1	**4**	1	**14**	1	**1**	1	**7**	1	**5**	1	**9**
	1	**2**	1	**9**	1	**6**	1	**19**	1	**2**	1	**9**	1	**10**	1	**10**
	1	**3**	2	**5**	1	**7**	1	**20**	1	**3**	2	**5**	1	**11**	1	**20**
	1	**7**	2	**8**	1	**9**	1	**21**	1	**7**	2	**8**	1	**14**	2	**11**
	1	**9**	2	**13**	2	**5**	2	**10**	1	**9**	2	**13**	1	**20**	2	**12**
	2	**12**	2	**14**	2	**10**	2	**12**	2	**12**	2	**14**	2	**12**	2	**13**
	2	**13**	2	**19**	2	**11**	2	**15**	2	**13**	2	**19**	2	**13**	2	**14**
	2	**16**	2	**20**	2	**14**	2	**16**	2	**16**	2	**20**	2	**16**	2	**16**
	2	**21**	2	**21**	2	**20**	2	**18**	2	**21**	2	**21**	2	**21**	2	**21**
CH		10.35		12.37		9.08		7.477		10.35		12.37		8.624		10.5
DB		1.014		1.086		1.147		1.264		1.014		1.086		1.097		1.087
Sil		0.56		0.5331		0.4082		0.3285		0.56		0.5331		0.4331		0.4227

FONTE: O Autor

Conforme os resultados da avaliação dos consensos, as combinações de três agrupamentos (TABELA 14) resultaram em consensos com silhueta abaixo de $0,6$ para os dois métodos, *K-modes* e ligação média. Dentre as combinações com dois agrupamentos, dois casos obtiveram silhueta abaixo deste valor: $Sim_{\delta-con}(\mathcal{W} \times \mathcal{P}) \times Sim_{VEO}(\mathcal{W} \times \mathcal{P})$ (Tabelas 16) e $Sim_{Res}(\mathcal{W} \times \mathcal{P}) \times Sim_{GED}(\mathcal{W} \times \mathcal{P})$ (Tabela 15). Os melhores consensos foram obtidos a partir das combinações dos agrupamentos $Sim_{Res}(\mathcal{W}\times\mathcal{P}) \times Sim_{\delta-con}(\mathcal{W}\times\mathcal{P})$ com silhueta $0,7$, e $Sim_{Res}(\mathcal{W}\times\mathcal{P}) \times Sim_{VEO}(\mathcal{W}\times\mathcal{P})$ com silhueta $0,7125$ (TABELA 16).

De acordo com a TABELA 16, deve-se destacar a presença do agrupamento de padrões que resultou da medida de similaridade baseada na distância de Resistência Normalizada entre passeios e caminhos nos consensos com melhores resultados. Importante ressaltar que o consenso foi formado a partir dos padrões de rótulos. Assim,

TABELA 15 – CONSENSO DAS COMBINAÇÕES DE DOIS AGRUPAMENTOS I

$\delta - con \times GED$				$GED \times VEO$				$Res \times GED$			
HC	**Passeios**	**KM**	**Passeios**	**HC**	**Passeios**	**KM**	**Passeios**	**HC**	**Passeios**	**KM**	**Passeios**
0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**
0	**17**	0	**17**	0	**17**	0	**17**	0	**17**	0	**4**
1	**1**	1	**1**	1	**1**	1	**1**	1	**1**	0	**5**
1	**2**	1	**2**	1	**2**	1	**2**	1	**2**	0	**6**
1	**3**	1	**3**	1	**3**	1	**3**	1	**3**	0	**10**
1	**4**	1	**4**	1	**4**	1	**4**	1	**4**	0	**17**
1	**5**	1	**5**	1	**5**	1	**5**	1	**5**	0	**20**
1	**6**	1	**6**	1	**6**	1	**6**	1	**6**	1	**1**
1	**7**	1	**7**	1	**7**	1	**7**	1	**7**	1	**2**
1	**9**	1	**9**	1	**9**	1	**9**	1	**9**	1	**3**
1	**10**	1	**10**	1	**10**	1	**10**	1	**10**	1	**7**
1	**11**	1	**11**	1	**11**	1	**11**	1	**11**	1	**9**
1	**14**	1	**14**	1	**14**	1	**14**	1	**14**	1	**11**
1	**20**	1	**20**	1	**20**	1	**20**	1	**20**	1	**14**
2	**8**	2	**8**	2	**8**	2	**8**	2	**8**	2	**8**
2	**12**	2	**12**	2	**12**	2	**12**	2	**12**	2	**12**
2	**13**	2	**13**	2	**13**	2	**13**	2	**13**	2	**13**
2	**15**	2	**15**	2	**15**	2	**15**	2	**15**	2	**15**
2	**16**	2	**16**	2	**16**	2	**16**	2	**16**	2	**16**
2	**18**	2	**18**	2	**18**	2	**18**	2	**18**	2	**18**
2	**19**	2	**19**	2	**19**	2	**19**	2	**19**	2	**19**
2	**21**	2	**21**	2	**21**	2	**21**	2	**21**	2	**21**
CH	11.99		11.99		11.74		11.74		12.05		16.02
DB	0.7993		0.7993		0.8331		0.8331		0.8202		0.8825
Sil	0.6685		0.6685		0.6323		0.6323		0.6412		0.5442

FONTE: O Autor

TABELA 16 – CONSENSO DAS COMBINAÇÕES DE DOIS AGRUPAMENTOS II

$Res \times \delta - con$				$Res \times VEO$				$\delta - con \times VEO$			
HC	**Passeios**	**KM**	**Passeios**	**HC**	**Passeios**	**KM**	**Passeios**	**HC**	**Passeios**	**KM**	**Passeios**
0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**
0	**4**	0	**4**	0	**4**	0	**4**	0	**1**	0	**1**
0	**5**	0	**5**	0	**6**	0	**5**	0	**2**	0	**2**
0	**6**	0	**6**	0	**10**	0	**6**	0	**3**	0	**3**
0	**8**	0	**8**	0	**11**	0	**8**	0	**4**	0	**4**
0	**10**	0	**10**	0	**12**	0	**10**	0	**6**	0	**6**
0	**15**	0	**15**	0	**15**	0	**15**	0	**7**	0	**7**
0	**17**	0	**17**	0	**16**	0	**17**	0	**8**	0	**9**
0	**18**	0	**18**	0	**17**	0	**18**	0	**9**	0	**15**
0	**19**	0	**19**	0	**18**	0	**19**	0	**13**	0	**17**
0	**20**	0	**20**	1	**1**	0	**20**	0	**15**	0	**18**
1	**1**	1	**1**	1	**2**	1	**1**	0	**17**	1	**5**
1	**2**	1	**2**	1	**3**	1	**2**	0	**18**	1	**8**
1	**3**	1	**3**	1	**7**	1	**3**	0	**19**	1	**13**
1	**7**	1	**7**	1	**9**	1	**7**	1	**5**	1	**14**
1	**9**	1	**9**	2	**5**	1	**9**	1	**10**	1	**19**
1	**13**	2	**11**	2	**8**	2	**11**	1	**12**	1	**20**
2	**11**	2	**12**	2	**13**	2	**12**	1	**14**	1	**21**
2	**12**	2	**13**	2	**14**	2	**13**	1	**16**	2	**10**
2	**14**	2	**14**	2	**19**	2	**14**	1	**20**	2	**11**
2	**16**	2	**16**	2	**20**	2	**16**	1	**21**	2	**12**
2	**21**	2	**21**	2	**21**	2	**21**	2	**11**	2	**16**
CH	21.06		22.17		25.14		23.16		9.54		13.29
DB	0.6651		0.631		0.7248		0.762		0.8286		0.9203
Sil	0.6538		0.7002		0.7125		0.6934		NA		0.5131

FONTE: O Autor

temos um consenso no espaço de rótulos.

Uma vez definidos os agrupamentos consenso, resta validar os agrupamentos a partir do espaço de padrões de similaridade. Dessa forma, os passeios serão

reagrupados conforme os rótulos do agrupamentos consenso e os respectivos valores de silhueta são calculados. Os resultados desta última análise para os melhores consensos são dados na TABELA 17.

TABELA 17 – COEFICIENTE DE SILHUETA DOS AGRUPAMENTOS DE PADRÕES DE SIMILARIDADE RESISTÊNCIA NORMALIZADA COM BASE NOS CONSENSOS

Combinação		$Res \times \delta - con$		$Res \times VEO$		$\delta - con \times VEO$		$\delta - con \times GED$		$GED \times VEO$		$Res \times GED$		$Res \times GED \times VEO$	
Método		HC	KM	HC	KM	HC	KM	HC	KM	HC	KM	HC	KM	HC	KM
Silhueta por caminhos	0	0,63339	0,60671	0,003059	0,60671	0,354224	0,245724	0,671092	0,671092	0,671092	0,671092	0,671092	0,56843	0,385743	0,003059
	1	0,513715	0,872401	0,902911	0,872401	0,527998	0,609452	-0,17608	-0,17608	-0,17608	-0,17608	-0,17608	0,162121	0,891489	0,902911
	2	0,537304	0,930816	0,948904	0,930816	0,55789	0,647003	-0,21688	-0,21688	-0,21688	-0,21688	-0,21688	0,147253	0,942258	0,948904
	3	0,46941	0,860933	0,899065	0,860933	0,554134	0,637432	-0,25382	-0,25382	-0,25382	-0,25382	-0,25382	0,066451	0,885112	0,899065
	4	0,438058	0,334073	-0,21503	0,334073	0,459128	0,430913	-0,68582	-0,68582	-0,68582	-0,68582	-0,68582	0,558508	0,028358	-0,21503
	5	0,733162	0,727113	0,153302	0,727113	-0,06362	0,254025	-0,4181	-0,4181	-0,4181	-0,4181	-0,4181	0,515493	0,56403	0,153302
	6	0,746489	0,745178	-0,2	0,745178	0,130816	-0,1786	-0,56111	-0,56111	-0,56111	-0,56111	-0,56111	0,585984	0,573422	-0,2
	7	0,537304	0,930816	0,948904	0,930816	0,55789	0,647003	-0,21688	-0,21688	-0,21688	-0,21688	-0,21688	0,147253	0,942258	0,948904
	8	0,325607	0,29774	0,383344	0,29774	-0,47833	0,455959	0,210599	0,210599	0,210599	0,210599	0,210599	-0,14093	0,152087	0,383344
	9	0,541589	0,925181	0,944121	0,925181	0,54994	0,637604	-0,19631	-0,19631	-0,19631	-0,19631	-0,19631	0,164404	0,937113	0,944121
	10	0,641235	0,630041	-0,4013	0,630041	0,148898	-0,65055	-0,32787	-0,32787	-0,32787	-0,32787	-0,32787	0,364862	0,472943	-0,4013
	11	0,61719	0,613445	-0,20004	0,613445	0	0,339886	-0,46036	-0,46036	-0,46036	-0,46036	-0,46036	-0,53545	-0,62451	-0,20004
	12	0,659826	0,65924	-0,2305	0,65924	-0,7478	0,345049	0,393096	0,393096	0,393096	0,393096	0,393096	0,328252	0,586411	-0,2305
	13	-0,82064	0,648197	0,307656	0,648197	-0,64071	-0,17252	0,495471	0,495471	0,495471	0,495471	0,495471	0,384234	0,684726	0,307656
	14	0,486091	0,533327	0,383071	0,533327	0,385724	0,11038	-0,58081	-0,58081	-0,58081	-0,58081	-0,58081	-0,67643	-0,65266	0,383071
	15	0,721827	0,72391	-0,1738	0,72391	0,082598	-0,19222	-0,54082	-0,54082	-0,54082	-0,54082	-0,54082	-0,59075	0,572689	-0,1738
	16	0,622584	0,673099	-0,43664	0,673099	0,121384	-0,12741	0,497111	0,497111	0,497111	0,497111	0,497111	0,394194	0,706615	-0,43664
	17	0,365756	0,225095	-0,28554	0,225095	0,482337	0,483128	0,670389	0,670389	0,670389	0,670389	0,670389	0,503776	-0,09057	-0,28554
	18	0,70621	0,695315	-0,06277	0,695315	0,275842	0,085967	-0,71808	-0,71808	-0,71808	-0,71808	-0,71808	-0,66003	0,49876	-0,06277
	19	0,728419	0,723264	-0,05218	0,723264	0,197459	-0,12011	-0,59628	-0,59628	-0,59628	-0,59628	-0,59628	-0,67419	0,539717	-0,05218
	20	0,670818	0,66123	0,242476	0,66123	0,11579	0,395595	-0,38368	-0,38368	-0,38368	-0,38368	-0,38368	0,38846	0,501326	0,242476
	21	0,458785	0,510592	0,397722	0,510592	0,39859	0,146033	0,468518	0,468518	0,468518	0,468518	0,468518	0,335949	0,519946	0,397722
Silhueta média		0,515188	0,660351	0,193488	0,660351	0,180463	0,228625	-0,13303	-0,13303	-0,13303	-0,13303	-0,13303	0,106266	0,45533	0,193488

FONTE: O Autor

Para a obtenção dos valores de silhueta na tabela são calculados com base nos padrões de similaridades baseadas na Resistência Normalizada. Como pode ser observado, os dois consensos formados utilizando o *K-modes* (KM) que envolve o padrão de rótulos desta similaridade alcançaram os valores de silhueta de $0,660351$ tanto em sua combinação com os padrões de rótulos Deltacon e Sobreposição vértice-aresta. Os valores de silhueta também foram calculados com base nos padrões de similaridade Deltacon, Sobreposição vértice-aresta e distância de edição. Entretanto, nestes três casos os valores não ultrapassam $0,3$ (TABELA 18). Segue deste último que os padrões de similaridade Resistência Normalizada podem ser agrupados de forma aproximada em relação aos agrupamentos de padrões de erros. Além disso, os agrupamentos formados obtiveram índices de qualidade que permitem afirmar que os grupos formados tem uma estrutura forte associada a sua coesão e separabilidade.

TABELA 18 – COEFICIENTE DE SILHUETA MÉDIO DOS AGRUPAMENTOS DE PADRÕES DE SIMILARIDADE DELTACON, SOBREPOSIÇÃO VÉRTICE-ARESTA E DISTÂNCIA DE EDIÇÃO COM BASE NOS CONSENSOS.

Combinação		$Res \times \delta - con$		$Res \times VEO$		$\delta - con \times VEO$		$\delta - con \times GED$		$GED \times VEO$		$Res \times GED$		$Res \times GED \times VEO$	
Método		HC	KM	HC	KM	HC	KM	HC	KM	HC	KM	HC	KM	HC	KM
$Sim()_d$	Delta	0,077976	0,031665	-0,10775	0,031665	0,332	0,071058	-0,00046	-0,00046	-0,00046	-0,00046	-0,00046	-0,08076	-0,07221	-0,10775
	VEO	0,080787	0,111668	0,056987	0,111668	0,020039	0,054644	0,06775	0,06775	0,06775	0,06775	0,06775	0,040373	0,050893	0,056987
	GED	0,081655	0,161165	0,16361	0,161165	-0,13174	-0,03093	0,121771	0,121771	0,121771	0,121771	0,121771	0,154118	0,118655	0,16361

FONTE: O Autor

Finalmente, conclui-se que um método para avaliação do desempenho dos treinandos, tal como enunciado nos objetivos desta tese, ou seja, um método que

capture os padrões de erro dos treinandos durante a execução de uma atividade instrucional, pode ser formalizado como segue

Modelagem do conhecimento Análise e modelagem da tarefa como um **grafo tarefa** cujos vértices são as atividades da tarefa. As arestas, por sua vez, são definidas pela união de **caminhos viáveis**, os quais representam possíveis ordenações de atividades que resultem em uma sequência desejável para a execução da tarefa pelo treinando.

Modelagem do treinando Análise e modelagem da execução da tarefa pelo treinando como um passeio sobre o grafo, ou **passeio do treinando**.

Modelagem dos padrões de similaridade Modelagem dos padrões de similaridade por meio de medidas que avaliem a proximidade entre os passeios dos treinandos e os caminhos viáveis. O modelo resultante é uma matriz de padrões de similaridade onde cada treinando é representado por um vetor linha cujas entradas correspondem à distância do passeio do treinando a cada um dos caminhos viáveis. Conforme o resultado desta tese, a medida de similaridade mais promissora para ser utilizada nesta etapa é a distância Resistência Normalizada.

Agrupamento do padrões de similaridade Aplicação do método *K-means* para agrupar os padrões de similaridade dos treinandos.

Validação do agrupamento Avaliação da qualidade do agrupamento por meio do cálculo do coeficiente de silhueta do agrupamento e respectivos grupos e padrões afiliados.

5.4.1 Resumo

Neste capítulo foi apresentada a proposta de um método de avaliação do desempenho de treinandos com base no agrupamento de padrões de similaridade que se aproximam dos padrões dos erros cometidos durante a execução de uma tarefa instrucional em um sistema de treinamento virtual. O método proposto foi desenvolvido utilizando uma estrutura de pesquisa baseada nas atividades do *Design Science Research* (*DSR*). Neste sentido, uma solução é proposta com base na similaridade de padrões de erro. Esta abordagem apresentou dois problemas. O primeiro está relacionado ao tempo gasto na modelagem do erro. O segundo, a qualidade dos agrupamentos obtidos. Conforme a previsto na *DSR*, a partir desta solução inicial foram propostas alternativas para o problema de agrupamento de padrões de erro. Estas alternativas foram construídas com base na modelagem do conhecimento como um grafo associado às regras para a execução da tarefa e um modelo do treinando que pode ser descrito como um passeio sobre este grafo.

A mineração de dados baseados em grafos é bem conhecida por suas aplicações em análise de redes. Nestes casos, é comum pensar no agrupamento sobre grafos como uma busca por comunidades que são definidas pela similaridade dos vértices da rede. Nesta perspectiva os objetos de análise são os vértices e arestas do grafo. Uma outra forma de definir o agrupamento de dados baseado em grafos é tomar os grafos como objeto de um agrupamento. Assim, dois grafos que pertençam ao mesmo grupo são mais similares do que grafos que pertençam a grupos diferentes. Uma vez que passeios e caminhos são tipos de grafos, uma medida de similaridade deve ser capaz de agrupar os passeios que representam a execução da tarefa pelos treinandos segundo classes de desempenho. Um método de agrupamento de padrões baseado na similaridade de grafos que se aproxime dos padrões de erros pode apresentar duas vantagens. A primeira é a automação do processo de avaliação que não dependa da modelagem manual do erro, tal como foi realizada neste trabalho. A segunda vantagem é a possibilidade de obtenção de grupos com uma estrutura bem definida com índices de qualidade superiores àquela obtida por meio dos padrões baseados no erro.

Um problema fundamental sobre o qual esta tese se apoia é o da similaridade entre dois grafos. Por este motivo, diversas concepções de similaridade são utilizadas para definir os padrões de similaridade entre os passeios dos treinandos entre si e entre os passeios e os caminhos viáveis. Os agrupamentos obtidos foram submetidos a avaliação por meio de medidas de validação e utilizados em uma análise de meta-agrupamento e combinação dos agrupamentos para a geração de uma agrupamento consenso. Os resultados desta análise final revelaram que a medida de similaridade baseada na distância Resistência Normalizada gera padrões de similaridade cujo agrupamento se aproxima dos padrões de erro com uma estrutura de grupos bem definida.

6 CONSIDERAÇÕES FINAIS

O último capítulo desta tese trata de sumarizar os principais resultados encontrados pela pesquisa e as limitações do trabalho que impactam na sua generalização. Finalmente, algumas oportunidades que derivam destas limitações são destacadas para desenvolvimentos futuros.

A motivação deste trabalho foi contribuir para o desenvolvimento de um método de avaliação do desempenho de treinandos em um ambiente de treinamento virtual. O desenvolvimento do método foi realizado por meio de uma estrutura de pesquisa baseada na *DSR* cujas atividades envolvem a conscientização sobre o problema, a formulação de uma solução, a geração de soluções alternativas, a validação da solução e a comunicação dos resultados da pesquisa. Neste sentido a conscientização do problema foi realizada por meio da revisão bibliográfica e sistematizada no capítulo da Fundamentação Teórica.

Os tópicos discutidos na Fundamentação Teórica foram agrupados em quatro eixos, ou pilares: a avaliação do desempenho, a modelagem do erro humano, a representação do conhecimento e a mineração de dados instrucionais. Por meio deste tópicos procurou-se mostrar que o erro tem um papel importante tanto nas tecnologias instrucionais quanto nas tecnologias do desempenho humano. Em qualquer uma das áreas o erro humano pode ser estudado segundo um modelo que explique o erro como um fenômeno associado aos processos cognitivos. Dessa forma, a compreensão do erro por meio de sua identificação, descrição e classificação possibilita a elaboração de estratégias instrucionais que promovam a reflexão sobre a ação humana e a sua interação com sistemas em situação críticas.

No contexto das tecnologias instrucionais mediadas pelo computador, as técnicas de mineração de dados, e os estudos sobre a representação do conhecimento têm fornecido um conjunto de ferramentas tanto para modelagem do erro humano quanto para avaliação do desempenho. A análise de agrupamento, por exemplo, por meio da classificação não supervisionada de padrões de desempenho permite extrair estratégias na resolução de problemas e identificar comportamentos atípicos de aprendizes em situação de aprendizagem.

Em relação à representação do conhecimento, os grafos têm sido amplamente utilizados para capturar as relações entre entidades. Na área de estudos conhecida como Mineração de Dados Educacionais, os grafos têm merecido uma atenção especial com a realização de eventos acadêmicos e publicações dedicadas ao tópico e suas aplicações. Assim como em outras áreas, como na análise genética ou de redes, a

mineração de dados é comumente definida sobre um grafo. Ou seja, as entidades a serem agrupadas são identificadas com os vértices do grafo e as arestas guardam as informações sobre suas relações. Em uma outra abordagem, adotada neste trabalho, a mineração de dados é realizada sobre um conjunto de grafos. Dessa forma, um dos problemas no agrupamento de dados baseado em grafos é definir uma medida de similaridade entre grafos.

Neste trabalho, a modelagem dos dados de interação dos treinandos como passeios sobre um grafo se destaca como um meio para gerar soluções alternativas para o método de avaliação que está sendo proposto. De um lado, conforme a atividade de sugestão da solução da *DSR* para o problema de avaliação do desempenho dos treinandos, uma primeira solução é proposta com base na modelagem e agrupamento de padrões de erro. Entretanto, como visto, esta solução apresentou resultados insatisfatórios e de difícil interpretação com grupos estruturalmente difusos. Por outro lado, a modelagem de padrões de similaridades de dados baseados em grafos possibilitou a formação de alguns agrupamentos que se aproximaram do padrão de erros com uma qualidade superior.

Dentre as sete definições de similaridade em quatorze configurações diferentes a medida de similaridade entre grafos que obteve melhores resultados foi a Resistência Normalizada. Esta medida alcançou índices de qualidade superiores tanto na análise de agrupamento quanto na sua extensão por meio da combinação de agrupamentos estando presente também nos agrupamentos consenso.

Dessa forma, conforme o objetivo da pesquisa, mostrou-se que a modelagem e agrupamento de dados baseados em grafos pode ser uma alternativa para a modelagem dos padrões de erro. Em certo sentido, com base nos experimentos realizados, pode-se afirmar que a solução baseada neste método de modelagem e agrupamentos de dados dos treinandos são sensíveis aos erros cometidos durante a tarefa.

Os resultados encontrados devem ser contextualizados a partir do conjunto de restrições sobre o qual eles foram obtidos como:

- a utilização de uma amostra pequena composta por vinte e dois padrões, o que impede afirmar se o método proposto pode ser generalizado ou mesmo escalonado;
- a modelagem dos padrões de erro ficou restrita a três modelos que não funcionaram;
- a utilização das medidas de similaridade e sua implementação na biblioteca *Netcomp*;

- a utilização do método de agrupamento *K-means*, o qual, como discutido, apresenta melhores resultados com grupos coesos e bem separados;
- a utilização do coeficiente de silhueta para validação dos agrupamentos;
- a utilização dos métodos de geração e validação do consenso implementados na biblioteca *DiceR* para a linguagem R.

Cada um destes itens representa uma decisão dentre outras possibilidades e todos, conjuntamente, delineiam o caminho desta pesquisa. Dentro destes limites, entretanto, além de alcançar o objetivo proposto, o trabalho desenvolvido produziu as seguintes contribuições:

- o desenvolvimento do modelo do conhecimento de regras baseado na sua representação como caminhos sobre grafos,
- o desenvolvimento do modelo do treinando como passeios sobre o grafo, e
- a identificação de padrões de similaridades que se aproximam de padrões de erros.

A modelagem do conhecimento como um grafo é amplamente difundida e sua aplicação no contexto das tecnologias instrucionais, dentre outros, tem encontrado um campo em franca expansão. Aliada às técnicas de mineração de dados, como os métodos de agrupamento, os dados baseados em grafos têm sido estudados como uma coleção de objetos relacionados. Os agrupamentos de grafos, entretanto, são menos explorados.

Os estudos e resultados preliminares resultantes da pesquisa sobre a qual esta tese foi construída foram publicizados em eventos nacionais e internacionais na forma dos artigos listados a seguir

1. FARIA, A. P. ; SANTOS, R. C. R. ; GEUS, K. . Treinamento de atividades críticas em Ambientes Virtuais 3D. XLIV CONGRESSO BRASILEIRO DE EDUCAÇÃO EM ENGENHARIA, 2016, Natal. Anais do XLIV Congresso Brasileiro de Educação em Engenharia - XLIV COBENGE, 2016.
2. FARIA, A. P. ; SIQUEIRA, P. H. ; GODOI, W. C. ; GEUS, K. . Redes Neurais Aplicadas à Classificação de Performance em Jogos Digitais. II Simpósio de Métodos Numéricos em Engenharia, 2017, Curitiba. Anais do II Simpósio de Métodos Numéricos em Engenharia. Curitiba: UFPR / Setor de Tecnologia, 2017.

3. GEUS, K.; BEÊ, R.; CORRÊA, V.; SANTOS, R.; FARIA, A.; SATO, E.; SWINKA-FILHO, V.; MIQUELIN, A.; SCHEER, S.; SIQUEIRA, P.; GODOI, W.; ROSENDO, M.; GRUBER, Y. Immersive Serious Game-style Virtual Environment for Training in Electrical Live Line Maintenance Activities. PROCEEDINGS of the 12th International Conference on Computer Supported Education. Science and Technology Publications, 2020.

4. FARIA, A. P. ; GEUS, K.; SCHEER, S.. Agrupamento de padrões de caminhos em treinamento virtual: uma análise de similaridades. XL Congresso Nacional de Matemática Aplicada e Computacional, 2021, Campo Grande.

5. FARIA, A. P. ; GEUS, K.; SCHEER, S. .Error patterns applied to task performance in a virtual training environment: a meta clustering approach. XLII Ibero-Latin-American Congress on Computational Methods in Engineering, CILAMCE, 2021, Rio de Janeiro.

A partir dos resultados encontrados, uma direção para desenvolvimentos futuros é o estudo de um modelo de erro que possa ser extraído diretamente dos dados baseados em grafos e, consequentemente, o desenvolvimento de uma definição de similaridade que seja sensível aos padrões de erros codificados por meio do grafo. Uma metodologia para este desenvolvimento pode ser adaptada de Koutra e Faloutsos (2018) com a caracterização dos erros como alterações topológicas em grafos sintéticos e os seus efeitos nos valores de similaridade. A distância entre grafos dada pela Resistência Normalizada apresentou algumas qualidades notáveis como a granularidade para capturar a diferença entre dois grafos e agrupamentos com estruturas coesas e bem separadas. Neste sentido, o estudo do seu comportamento em relação aos desvios característicos de determinados erros pode fornecer uma base para uma nova medida de similaridade entre grafos.

REFERÊNCIAS

ABRACOPEL. **Anuário Estatítico de Acidentes de Origem Elétrica 2020 – Ano base 2019**. [S.l.], 2020.

AGGARWAL, C. **Data Clustering : Algorithms and Applications**. Hoboken: CRC Press, 2013. ISBN 978-1-4665-5822-9.

AGGARWAL, C. C.; WANG, H. (Ed.). **Managing and Mining Graph Data**. [S.l.]: Springer US, 2010. DOI: 10.1007/978-1-4419-6045-0. Disponível em: https://doi.org/10.1007/978-1-4419-6045-0.

AGGARWAL, C. C.; HINNEBURG, A.; KEIM, D. A. On the surprising behavior of distance metrics in high dimensional space. In: SPRINGER. INTERNATIONAL conference on database theory. [S.l.: s.n.], 2001. P. 420–434.

AGHABABYAN, A.; LEWKOW, N.; BAKER, R. S. Enhancing the Clustering of Student Performance Using the Variation in Confidence. In: SPRINGER. INTERNATIONAL Conference on Intelligent Tutoring Systems. [S.l.: s.n.], 2018. P. 274–279.

AGUIAR, Y. P. C.; FÁTIMA Q. VIEIRA, M. de; GALY, E.; SANTONI, C. Accounting for Individual and Situation Characteristics to Understand the User Behaviour When Interacting With Systems During Critical Situations. In: HUMAN Performance Technology. [S.l.]: IGI Global, 2019. P. 298–325. DOI: 10.4018/978-1-5225-8356-1.ch016. Disponível em: https://doi.org/10.4018/978-1-5225-8356-1.ch016.

ALL, A.; NUNEZ CASTELLAR, E.; CASTELLAR, N.; LOOY, J. Measuring Effectiveness in Digital Game-Based Learning: A Methodological Review. **International Journal of Serious Games**, v. 1, mai. 2014. DOI: 10.17083/ijsg.v1i2.18.

ANNETT J.; DUNCAN, K. Task analysis and training design. **Occupational Psychology**, v. 14, p. 211–21, 1967.

ARBELAITZ, O.; GURRUTXAGA, I.; MUGUERZA, J.; PÉREZ, J. M.; PERONA, I. An extensive comparative study of cluster validity indices. **Pattern Recognition**, v. 46, n. 1, p. 243–256, 2013. ISSN 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2012.07.021.

ARTHUR, D.; VASSILVITSKII, S. K-Means++: The Advantages of Careful Seeding. **Proc. of the Annu. ACM-SIAM Symp. on Discrete Algorithms**, v. 8, p. 1027–1035, jan. 2007. DOI: 10.1145/1283383.1283494.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR ISO 10015**: Gestão da qualidade - Diretrizes para treinamento. Rio de Janeiro, 2001. P. 12.

BAARSCH, J.; CELEBI, M. E. Investigation of internal validity measures for K-means clustering. In: SN. PROCEEDINGS of the international multiconference of engineers and computer scientists. [S.l.: s.n.], 2012. v. 1, p. 14–16.

BABER, C.; STANTON, N. A. Task analysis for error identification: theory, method and validation. **Theoretical Issues in Ergonomics Science**, Taylor & Francis, v. 3, n. 2, p. 212–227, 2002.

BAKER, R. S. Stupid tutoring systems, intelligent humans. **International Journal of Artificial Intelligence in Education**, Springer, v. 26, n. 2, p. 600–614, 2016.

BAKER, R. S.; CORBETT, A. T.; KOEDINGER, K. R.; WAGNER, A. Z. Off-task behavior in the cognitive tutor classroom: when students"game the system". In: PROCEEDINGS of the SIGCHI conference on Human factors in computing systems. [S.l.: s.n.], 2004. P. 383–390.

BAKER, R. S.; INVENTADO, P. S. Educational data mining and learning analytics. In: LEARNING analytics. [S.l.]: Springer, 2014. P. 61–75.

BARON, S.; HUEY, D. K. B.; RING, J. **Qualitative Modeling of Human Performance in Complex, Dynamic Systems**. [S.l.]: Insight, 2004. v. 6, p. 42–43.

BASTIAN, M.; HEYMANN, S.; JACOMY, M. Gephi: an open source software for exploring and manipulating networks. In: THIRD international AAAI conference on weblogs and social media. [S.l.: s.n.], 2009.

BAX, M. P. Design science: filosofia da pesquisa em ciência da informação e tecnologia Design science: philosophy of research in information science and technology. **Ci. Inf.**, v. 42, n. 2, p. 298–312, 2013.

BELL, F. Connectivism: Its place in theory-informed research and innovation in technology-enabled learning. **International Review of Research in Open and**

Distributed Learning, Athabasca University Press (AU Press), v. 12, n. 3, p. 98–118, 2011.

BERLINGERIO, M. .; KOUTRA, D.; ELIASSI-RAD, T.; FALOUTSOS, C. Netsimile: A scalable approach to size-independent network similarity. **arXiv preprint arXiv:1209.2684**, 2012.

BLOOM, B.; HASTINGS, J.; MADAUS, G. **Manual de Avaliação Formativa e Somativa do Aprendizado Escolar**. [S.l.]: Livraria Pioneira Editora, 1983.

BRASIL. Ministério do Trabalho e Emprego. **NR 10**: Segurança em instalações e serviços em eletricidade. Brasilia, DF, 2004.

BURIOL, T. M.; ROSENDO, M.; SCHEER, S.; GEUS, K. de. Proposta de plataforma baseada em realidade virtual para treinamento de atividades em linha viva. **XXX CILAMCE**, p. 1–10, 2009.

BURIOL, T. M. **Convergencia de Games e Realidade Virtual para Treinamento de Manutenção em Redes de Energia em Linha Viva**. 2011. Tese (Doutorado) – Universidade Federal do Parana.

CAI, L.; YANG, Z.; YANG, S. X.; QU, H. Modelling and simulating of risk behaviours in virtual environments Based on multi-agent and fuzzy logic. **International Journal of Advanced Robotic Systems**, v. 10, 2013. ISSN 17298806. DOI: 10.5772/56832.

CARUANA, R.; ELHAWARY, M.; NGUYEN, N.; SMITH, C. Meta clustering. In: IEEE. SIXTH International Conference on Data Mining (ICDM'06). [S.l.: s.n.], 2006. P. 107–118.

CHADHA, A. Efficient clustering algorithms in educational data mining. In: HANDBOOK of Research on Knowledge Management for Contemporary Business Environments. [S.l.]: IGI Global, 2018. P. 279–312.

CHATURVEDI, A.; GREEN, P. E.; CAROLL, J. D. K-modes clustering. **Journal of classification**, Springer, v. 18, n. 1, p. 35–55, 2001.

CHEN, H.; ZHANG, F. Resistance distance and the normalized Laplacian spectrum. **Discrete Applied Mathematics**, Elsevier BV, v. 155, n. 5, p. 654–661, mar. 2007. DOI:

10.1016/j.dam.2006.09.008. Disponível em: https://doi.org/10.1016/j.dam.2006.09.008.

CHENG, C.-M.; HWANG, S.-L. Applications of integrated human error identification techniques on the chemical cylinder change task. **Applied ergonomics**, v. 47C, p. 274–284, mar. 2015. DOI: 10.1016/j.apergo.2014.10.008.

CHIU, D. S.; TALHOUK, A. diceR: an R package for class discovery using an ensemble driven approach. **BMC Bioinformatics**, Springer Science e Business Media LLC, v. 19, n. 1, jan. 2018. DOI: 10.1186/s12859-017-1996-y. Disponível em: https://doi.org/10.1186/s12859-017-1996-y.

CHRYSAFIADI, K.; VIRVOU, M. Student modeling approaches : A literature review for the last decade. **Expert Systems With Applications**, Elsevier, v. 40, n. 11, p. 4715–4729, 2013. ISSN 0957-4174. DOI: 10.1016/j.eswa.2013.02.007. Disponível em: http://dx.doi.org/10.1016/j.eswa.2013.02.007.

CHYUNG, S. Y. **Foundations of Instructional and Performance Technology**. Amherst, Massachusetts: HRD Press, 2008. ISBN 9781599961361.

COPEL. **Procedimentos básicos para trabalhos de manutenção de subestações coma técnica de linha viva**. Curitiba PR, sd.

COPEL-GSST. **MANUTENÇÃO EM LINHAS DE DISTRIBUIÇÃO EM ALTA TENSÃO COM LINHA VIVA GRUPO 5-900**. Curitiba PR, 2020.

COSTA, T. K. d. L.; MACHADO, L. d. S.; MORAES, R. M. d. Inteligência artificial e sua aplicação em serious games para saúde. **RECIIS - Revista Eletrônica de Comunicação, Informação e Inovação em Saúde**, v. 8, n. 4, p. 525–539, 2001.

COSTA TEXEIRA, C. da; BAZZO, F. **Aumento da confiabilidade e segurança do sistema elétrico de potência brasileiro através da substituição de isoladores sujeitos à expansão do cimento**. Curitiba - PR, 2013.

DE WINTER, J.; LEEUWEN, P. M. van; HAPPEE, R. et al. Advantages and disadvantages of driving simulators: A discussion. In: PROCEEDINGS of 8th Measuring Behavior. [S.l.: s.n.], 2012. v. 2012.

DING, C.; HE, X. K-means clustering via principal component analysis. In: PROCEEDINGS of the twenty-first international conference on Machine learning. [S.l.: s.n.], 2004. P. 29.

DOGAN, B.; CAMURCU, A. Y. Visual clustering of multidimensional educational data from an intelligent tutoring system. **Computer Applications in Engineering Education**, Wiley, v. 18, n. 2, p. 375–382, fev. 2009. DOI: 10.1002/cae.20272. Disponível em: https://doi.org/10.1002/cae.20272.

DOIGNON, J.-P.; FALMAGNE, J.-C. **Knowledge Spaces**. [S.l.]: Springer, ago. 1999. ISBN 9783540645016.

DRESCH, A.; LACERDA, D. P.; JÚNIOR, J. A. V. A. **Design science research: método de pesquisa para avanço da ciência e tecnologia**. [S.l.]: Bookman Editora, 2015.

DRISCOLL, M. P. **Psychology of learning for instruction**. Londres: Pearson, 2014.

ELMORE, K. L.; RICHMAN, M. B. Euclidean distance as a similarity metric for principal component analysis. **Monthly weather review**, v. 129, n. 3, p. 540–549, 2001.

ERBETTA, C. D. C. **Caracterização e estudo de envelhecimento de isolador tipo pino em PEAD utilizado no setor elétrico**. 2015. Tese (Doutorado) – Universidade Federal de Minas Gerais.

ERTMER, P. A.; NEWBY, T. J. Behaviorism, cognitivism, constructivism: Comparing critical features from an instructional design perspective. **Performance improvement quarterly**, Wiley Online Library, v. 6, n. 4, p. 50–72, 1993.

EVERITT, B. **Cluster Analysis**. Hoboken: Wiley, 2011. ISBN 978-0-470-74991-3.

FALMAGNE, J.-C.; ALBERT, D.; DOBLE, C.; EPPSTEIN, D.; HU, X. (Ed.). **Knowledge Spaces: Applications in Education**. [S.l.]: Springer Berlin Heidelberg, 2013. DOI: 10.1007/978-3-642-35329-1. Disponível em: https://doi.org/10.1007/978-3-642-35329-1.

FELLOWS, M.; GUO, J.; KOMUSIEWICZ, C.; NIEDERMEIER, R.; UHLMANN, J. Graph-based data clustering with overlaps. In: SPRINGER. INTERNATIONAL Computing and Combinatorics Conference. [S.l.: s.n.], 2009. P. 516–526.

FERREIRA, J. A.; SOARES, E.; MACHADO, L. S.; MORAES, R. M. Assessment of fuzzy gaussian naive bayes for classification tasks. **PATTERNS 2015**, p. 73, 2015.

FOCKING, G. P.; VIEIRA, M. D. E. F. Q.; DA, J. S.; NETO, R. Treinamento teórico-prático de operadores de sistemas elétricos apoiado por ambiente virtual, simuladores e laboratório virtual. **Anais do XIX Congresso Brasileiro de Automática**, p. 2765–2772, 2012.

FOSSEY, W. A. **An Evaluation of Clustering Algorithms for Modeling Game-Based Assessment Work Processes**. 2017. Tese (Doutorado) – University of Maryland.

FUNCOGE. **Relatório de estatísticas de acidentes no setor elétrico brasileiro.** Rio de Janeiro, 2013.

GAGNÉ, R. M. **Como se realiza a aprendizagem**. Rio de Janeiro: Livros Técnicos e Científicos, 1974.

GAGNÉ, R. M.; BRIGGS, L. J. **La planification de la ensenanza: sus principios**. DF - Mexixo: Editorial Trillas, 1976.

GAGNÉ, R. M.; BRIGGS, L. J.; WAGNER, W. W. **Principles of Instructional Design**. Florida - US: Hancourt Brace, 1992.

GALLAGHER, A. G.; RITTER, E. M.; CHAMPION, H.; HIGGINS, G.; FRIED, M. P.; MOSES, G.; SMITH, C. D.; SATAVA, R. M. Virtual reality simulation for the operating room: proficiency-based training as a paradigm shift in surgical skills training. **Annals of surgery**, Lippincott, Williams, e Wilkins, v. 241, n. 2, p. 364, 2005.

GALVAN, I.; AYALA, A.; MUÑOZ, J.; SALGADO, M.; RODRÍGUEZ, E.; PÉREZ, M. Virtual Reality System For Training Of Operators Of Power Live Lines. **October**, v. I, p. 20–23, 2010.

GAO, X.; XIAO, B.; TAO, D.; LI, X. A survey of graph edit distance. **Pattern Analysis and Applications**, Springer Science e Business Media LLC, v. 13, n. 1, p. 113–129, jan. 2009. DOI: 10.1007/s10044-008-0141-y. Disponível em: https://doi.org/10.1007/s10044-008-0141-y.

GAO, X.; YANG, M. Understanding and Enhancement of Internal Clustering Validation Indexes for Categorical Data. **Algorithms**, MDPI AG, v. 11, n. 11, p. 177, nov. 2018. DOI: 10.3390/a11110177. Disponível em: https://doi.org/10.3390/a11110177.

GARANT, E. **Virtual reality training system for substation operator**. 1997. PhD Dissertation – Department ofElectrical Engineering McGill University.

GARCIA, C. H. **Tabelas para classificação do coeficiente de variação**. v. 171. Piracicaba SP, 1989.

GEUS, K.; BEÊ, R.; CORRÊA, V.; SANTOS, R.; FARIA, A.; SATO, E.; SWINKA-FILHO, V.; MIQUELIN, A.; SCHEER, S.; SIQUEIRA, P.; GODOI, W.; ROSENDO, M.; GRUBER, Y. Immersive Serious Game-style Virtual Environment for Training in Electrical Live Line Maintenance Activities. In: PROCEEDINGS of the 12th International Conference on Computer Supported Education. [S.l.]: Science e Technology Publications, 2020. DOI: 10.5220/0009343200420053. Disponível em: https://doi.org/10.5220/0009343200420053.

GIONIS, A.; MANNILA, H.; TSAPARAS, P. Clustering aggregation. **ACM Transactions on Knowledge Discovery from Data (TKDD)**, ACM New York, NY, USA, v. 1, n. 1, 4–es, 2007.

GUEDES, J. P. **Análise da confiabilidade humana na operação de uma subestação do sistema elétrico de potência**. 2017. Tese (Doutorado) – Pós-Graduação em Engenharia de Produção - Universidade Federal de Rio Grande do Sul, Porto Alegre - RS.

GUPTA, S.; ANAND, D.; BROUGH, J.; SCHWARTZ, M.; KAVETSKY, R. **Training in virtual environments: A safe, cost effective, and engaging approach to training**. [S.l.]: CALCE EPSC Press, University of Maryland, College Park, MD, 2008, 2008.

GUPTA, S. K.; ANAND, D. K.; BROUGH, J. E.; KAVETSKY, R. A.; SCHWARTZ, M.; THAKUR, A. A survey of the virtual environments-based assembly training applications. In: VIRTUAL Manufacturing Workshop, Turin, Italy. [S.l.: s.n.], 2008. P. 1–10.

HADJITODOROV, S.; KUNCHEVA, L.; TODOROVA, L. Moderate diversity for better cluster ensembles. **Information Fusion**, v. 7, p. 264–275, set. 2006. DOI: 10.1016/j.inffus.2005.01.008.

HAGBERG, A.; SCHULT, D.; SWART, P. **NetworkX 2.5**. [S.l.: s.n.], 2020. Disponível em: https://networkx.org/.

HAINEY, T.; CONNOLLY, T.; CHAUDY, Y.; BOYLE, E.; BEEBY, R.; SOFLANO, M. Assessment Integration in Serious Games. In: GAMIFICATION: Concepts, Methodologies, Tools, and Applications. [S.l.]: IGI Global, jan. 2015. P. 515–540. ISBN 9781466682016. DOI: 10.4018/978-1-4666-8200-9.ch025.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. Clustering validity checking methods: Part II. **ACM Sigmod Record**, ACM New York, NY, USA, v. 31, n. 3, p. 19–27, 2002.

_____. On Clustering Validation Techniques. **Journal of Intelligent Information Systems**, v. 17, out. 2001. DOI: 10.1023/A:1012801612483.

HALKIDI, M.; VAZIRGIANNIS, M.; BATISTAKIS, Y. Quality scheme assessment in the clustering process. In: SPRINGER. EUROPEAN Conference on Principles of Data Mining and Knowledge Discovery. [S.l.: s.n.], 2000. P. 265–276.

HE, Z.; XU, X.; DENG, S. A cluster ensemble method for clustering categorical data. **Information Fusion**, Elsevier, v. 6, n. 2, p. 143–151, 2005.

HENNIG, C. What are the true clusters? **Pattern Recognition Letters**, Elsevier, v. 64, p. 53–62, 2015.

HERNÁNDEZ, Y.; PÉREZ RAMÍREZ, M.; INGRAM RAMÍREZ, W.; NAVA AYALA, E.; ONTIVEROS HERNÁNDEZ, N. J. Architecture of an Intelligent Training System based on Virtual Environments for Electricity Distribution Substations. **Research in Computing Science**, v. 129, n. 1, p. 63–70, 2016. ISSN 1870-4069. DOI: 10.13053/rcs-129-1-7.

HOGARTH, R. M.; GIBBS, B. J.; MCKENZIE, C. R.; MARQUIS, M. A. Learning from feedback: exactingness and incentives. **Journal of Experimental Psychology: Learning, Memory, and Cognition**, American Psychological Association, v. 17, n. 4, p. 734, 1991.

HOOSHYAR, D.; YANG, Y.; PEDASTE, M.; HUANG, Y.-M. Clustering Algorithms in an Educational Context: An Automatic Comparative Approach. **IEEE Access**, Institute of Electrical e Electronics Engineers (IEEE), v. 8, p. 146994–147014, 2020. DOI:

10.1109/access.2020.3014948. Disponível em: https://doi.org/10.1109/access.2020.3014948.

JAIN, A. K. Data clustering: 50 years beyond K-means. **Pattern Recognition Letters**, Elsevier BV, v. 31, n. 8, p. 651–666, jun. 2010. DOI: 10.1016/j.patrec.2009.09.011. Disponível em: https://doi.org/10.1016/j.patrec.2009.09.011.

JAIN, A. K.; DUBES, R. C. **Algorithms for clustering data**. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988.

JIA, B.; YANG, Y.; ZHANG, J. Study on Learner Modeling in Adaptive Learning System. **J. Comput.**, v. 7, n. 10, p. 2585–2592, 2012.

JOVANOVIC, M.; VUKICEVIC, M.; MILOVANOVIĆ, M.; MINOVIĆ, M. Using data mining on student behavior and cognitive style data for improving e-learning systems: A case study. **International Journal of Computational Intelligence Systems**, v. 5, p. 597–610, mai. 2012. DOI: 10.1080/18756891.2012.696923.

KAY, H.; DODD, B.; SIME, M. **Iniciação à instrução programada e às máquinas de Ensinar.** São Paulo: IBRASA, 1970.

KLEIN, D. J.; RANDIC, M. Resistance distance. **Journal of Mathematical Chemistry**, Springer, v. 12, n. 1, p. 81–95, 1993.

KOUTRA, D.; FALOUTSOS, C. **Individual and collective graph mining : principles, algorithms, and applications**. San Rafael, California: Morgan & Claypool Publishers, 2018. ISBN 1681730391.

KOUTRA, D.; VOGELSTEIN, J. T.; FALOUTSOS, C. DeltaCon: A Principled Massive-Graph Similarity Function. In: PROCEEDINGS of the 2013 SIAM International Conference on Data Mining. [S.l.]: Society for Industrial e Applied Mathematics, mai. 2013. DOI: 10.1137/1.9781611972832.18. Disponível em: https://doi.org/10.1137/1.9781611972832.18.

KUNCHEVA, L. I.; HADJITODOROV, S. T. Using diversity in cluster ensembles. In: IEEE. 2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583). [S.l.: s.n.], 2004. v. 2, p. 1214–1219.

KUNCHEVA, L. I.; HADJITODOROV, S. T.; TODOROVA, L. P. Experimental comparison of cluster ensemble methods. In: IEEE. 2006 9th International Conference on Information Fusion. [S.l.: s.n.], 2006. P. 1–7.

LACERDA, D. P.; DRESCH, A.; PROENÇA, A.; ANTUNES JÚNIOR, J. A. V. Design Science Research: método de pesquisa para a engenharia de produção. **Gestão & produção**, v. 20, p. 741–761, 2013.

LEVY, P. **As tecnologias da Inteligência: o futuro do pensamento na era da informática.** Rio de Janeiro: Editora 34, 2001.

LIM, J.; LINSANGAN, N.; CRUZ, F. R.; CHUNG, W.-Y. Temperature Compensated Electronic Nose for Fruit Ripeness Determination Using Component Correction Principal Component Analysis. **International Journal of Computer and Communication Engineering**, v. 5, p. 331–340, jan. 2016. DOI: 10.17706/IJCCE.2016.5.5.331-340.

LIÑÁN, L. C.; PÉREZ, Á. A. J. Educational Data Mining and Learning Analytics: differences, similarities, and time evolution. **International Journal of Educational Technology in Higher Education**, Springer, v. 12, n. 3, p. 98–112, 2015.

LLOYD, S. Least squares quantization in PCM. **IEEE transactions on information theory**, IEEE, v. 28, n. 2, p. 129–137, 1982.

LOUPOS, K.; PSONIS, P.; AMDITIS, A. Virtual reality: The way ahead in industrial safety. **Proceedings - 22nd European Conference on Modelling and Simulation, ECMS 2008**, p. 548–554, 2008. DOI: 10.7148/2008-0548.

LU, X. **Using behavioral context in process mining : exploration, preprocessing and analysis of event data**. Mai. 2018. Tese (Doutorado) – Mathematics e Computer Science. Proefschrift. ISBN 978-90-386-4484-4.

MÄÄTTÄ, T. **Virtual environments in machinery safety analysis**. [S.l.]: VTT Technical Research Centre of Finland, 2003.

MACHADO, L. d. S.; MORAES, R. M. d. Assessment systems for training based on virtual reality: A comparison study. **Journal on Interactive Systems**, v. 3, n. 1, 2012.

MACQUEEN, J. et al. Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA, 14. PROCEEDINGS of the fifth Berkeley symposium on mathematical statistics and probability. [S.l.: s.n.], 1967. P. 281–297.

MANJARRES, A. V.; SANDOVAL, L. G. M.; SUÁREZ, M. S. Data mining techniques applied in educational environments: Literature review. **Digital Education Review**, v. 06, p. 235–266, 2018.

MARCH, S. T.; SMITH, G. F. Design and natural science research on information technology. Decision Support Systems. **Decision Support Systems**, v. 15, p. 251–266, 1995. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.826.5567&rep=rep1&type=pdf.

MAXWELL, D. B. **Gauging Training Effectiveness of Virtual Environment Simulation Based Applications for an Infantry Soldier Training Task**. 2015. Tese (Doutorado) – University of Central Florida Orlando, Florida.

MCBROOM, J.; YACEF, K.; KOPRINSKA, I. DETECT: A Hierarchical Clustering Algorithm for Behavioural Trends in Temporal Educational Data. In: LECTURE Notes in Computer Science. [S.l.]: Springer International Publishing, 2020. P. 374–385. DOI: 10.1007/978-3-030-52237-7_30. Disponível em: https://doi.org/10.1007/978-3-030-52237-7_30.

MEDLER, D. A. A brief history of connectionism. **Neural Computing Surveys**, Citeseer, v. 1, p. 18–72, 1998.

MENDES DE LIMA, G.; CARDOSO, A.; LAMOUNIER JR, E.; MACEDO JR, J. R. Development of a Virtual Reality environment of electric power substations for quality of service improvement. In: X Conferência Brasileira sobre Qualidade de Energia Elétrica. [S.l.: s.n.], jan. 2013.

MERRILL, M. D. First principles of instruction: A synthesis. **Trends and issues in instructional design and technology**, v. 2, p. 62–71, 2007.

METCALFE, J. Learning from errors. **Annual review of psychology**, Annual Reviews, v. 68, p. 465–489, 2017.

METCALFE, J.; XU, J. Learning from one's own errors and those of others. **Psychonomic Bulletin & Review**, Springer, v. 25, n. 1, p. 402–408, 2018.

MORAES, R.; MACHADO, L.; SOUZA, L. Online assessment of training in virtual reality simulators based on general Bayesian Networks. In: VI International Conference on Engineering and Computer Education (ICECE'2009). [S.l.: s.n.], 2009. P. 8–11.

MORAES, R. M. d.; SANTOS MACHADO, L. dos. Evaluation of Training Executed by Web Using Fuzzy Rule Based Expert Systems. In: GLOBAL Congress on Engineering and Technology Education. [S.l.: s.n.], 2005.

MOURA, R.; BEER, M.; PATELLI, E.; LEWIS, J.; KNOLL, F. Human error analysis: Review of past accidents and implications for improving robustness of system design. In. ISBN 9780429226823. DOI: 10.1201/b17399-147.

MURTAGH, F.; HECK, A. **Multivariate data analysis**. [S.l.]: Springer Science & Business Media, 2012. v. 131.

NALDI, M.; FACELI, K.; CARVALHO, A. de. Uma Revisão Sobre Combinação de Agrupamentos. **Revista de Informática Teórica e Aplicada**, v. 16, mar. 2010. DOI: 10.22456/2175-2745.8522.

NARANJO, J. E.; SANCHEZ, D. G.; ROBALINO-LOPEZ, A.; ROBALINO-LOPEZ, P.; ALARCON-ORTIZ, A.; GARCIA, M. V. A Scoping Review on Virtual Reality-Based Industrial Training. **Applied Sciences**, MDPI AG, v. 10, n. 22, p. 8224, nov. 2020. DOI: 10.3390/app10228224. Disponível em: https://doi.org/10.3390/app10228224.

NETO, J. A. N.; QUEIROZ VIEIRA, M. d. F.; SANTONI, C. Estratégias para Prevenção do Erro em Sistemas Elétricos: Um Estudo de Caso. In: ANAIS do IX SBAI (Simpósio Brasileiro de Automação Inteligente). [S.l.: s.n.], 2009.

NETTO, A. d. S.; TORRES, F.; QUEIROZ VIEIRA, M. d. F.; FERREIRA, F.; SORAGHAN, J. Ferramenta para avaliação do treinamento de operadores de subestação elétrica apoiada por simulador. In: V Simpósio Brasileiro de Sistemas Elétricos. [S.l.: s.n.], abr. 2014.

NETTO, A. d. S.; TORRES, F.; QUEIROZ VIEIRA, M. d. F.; FERREIRA, F. F. V. M. Electrical system operator training evaluation: a sistematic approach. **International Journal of Enhanced Research in Science Technology and Engineering**, v. 3, p. 34–40, ago. 2014.

NOONE, L. Instructional design and workplace performance. **Australasian Journal of Educational Technology**, v. 9, n. 1, 1993.

NOVAK, J. D.; CAÑAS, A. J. A teoria subjacente aos mapas conceituais e como elaborá-los e usá-los. **Práxis Educativa (Brasil)**, Universidade Estadual de Ponta Grossa, v. 5, n. 1, p. 9–29, 2010.

NTUEN, C. A.; CHESTNUT, J. A. An Expert System for Selecting Manufacturing Workers for Training. **Expert Systems With Applications**, v. 9, n. 3, p. 309–332, 1995.

OHLSSON, S. Learning from error and the design of task environments. **International Journal of Educational Research**, Elsevier BV, v. 25, n. 5, p. 419–448, jan. 1996. DOI: 10.1016/s0883-0355(97)81236-0. Disponível em: https://doi.org/10.1016/s0883-0355(97)81236-0.

OHLSSON, S. Constraint-Based Modeling: From Cognitive Theory to Computer Tutoring – and Back Again. **International Journal of Artificial Intelligence in Education**, Springer Science e Business Media LLC, v. 26, n. 1, p. 457–473, out. 2015. DOI: 10.1007/s40593-015-0075-7. Disponível em: https://doi.org/10.1007/s40593-015-0075-7.

_____. The Interaction Between Knowledge and Practice in the Acquisition of Cognitive Skills. In: FOUNDATIONS of Knowledge Acquisition: Cognitive Models of Complex Learning. [S.l.]: Springer US, 1993. P. 147–208. DOI: 10.1007/978-1-4615-3172-2_5. Disponível em: https://doi.org/10.1007/978-1-4615-3172-2_5.

OLIVARES, R. C.; RIVERA, S.; MCLEOD, J. N. A novel qualitative prospective methodology to assess human error during accident sequences. **Safety Science**, v. 103, p. 137–152, mar. 2018. DOI: 10.1016/j.ssci.2017.10.023.

PANTELIDIS, V. S. Reasons to use virtual reality in education and training courses and a model to determine when to use virtual reality. **Themes in Science and Technology Education**, v. 2, n. 1-2, p. 59–70, 2010.

PAQUETTE, L.; BAKER, R. S. Comparing machine learning to knowledge engineering for student behavior modeling: a case study in gaming the system. **Interactive Learning Environments**, Taylor & Francis, v. 27, n. 5-6, p. 585–597, 2019.

PARK, C.-H.; JANG, G.; CHAI, Y.-H. Development of a Virtual Reality Training System for Live-Line Workers. **International Journal of Human–computer Interaction**, v. 20, n. 3, p. 285–303, 2006.

PAVLIK JR, P.; BRAWNER, K.; OLNEY, A.; MITROVIC, A. A review of learner models used in intelligent tutoring systems. In: DESIGN Recommendations for Intelligent Tutoring Systems: learner modeling. [S.l.]: U.S. Army Research Laboratory, abr. 2013. v. 1. P. 39–68.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PEREIRA, O. G. Erro humano : uma conferência internacional. **Análise Psicológica**, v. 3, 1983.

PIMENTEL, M.; FILIPPO, D.; SANTOS, T. M. Design Science Research: pesquisa científica atrelada ao design de artefatos. **RE@ D-Revista de Educação a Distância e eLearning**, v. 3, n. 1, p. 37–61, 2020.

PONTE JR., G. P. **Gerenciamento de riscos baseado em fatores humanos e cultura de segurança: estudo de caso de simulação computacional do comportamento humano durante a operação de escape e abandono em instalações offshore.** Rio de Janeiro: Elsevier, 2014.

QUEIROZ VIEIRA, M.; SCAICO, A.; DO, J.; NETO, N.; SANTONI, C.; MERCANTINI, J.-M. A model based Operator Training Simulator for Electric Systems. In: 3RD Conference on Conceptual Modeling and Simulation (CMS). [S.l.: s.n.], fev. 2007.

RAMOS, J. L. C.; SILVA, R. E. D. e; SILVA, J. C. S.; RODRIGUES, R. L.; GOMES, A. S. A comparative study between clustering methods in educational data mining. **IEEE Latin America Transactions**, IEEE, v. 14, n. 8, p. 3755–3761, 2016.

RASMUSSEN, J. Notes on human error analysis and prediction. In: APOSTOLAKIS, G.; GARRIBBA, S.; VOLTA, G. (Ed.). **Synthesis and analysis methods for safety and reliability studies**. [S.l.]: Plenum Publishing Corporation, 1980. P. 357–389.

RASMUSSEN, J. A Model of Human Decision Making in Complex Systems and its Use for Design of System Control Strategies. **Proceedings of the American Control Conference**, June, p. 270–276, 1982. ISSN 07431619. DOI: 10.23919/acc.1982.4787855.

RASMUSSEN, J. Skills , Rules , and Knowledge Signals , Signs , and Symbols , and Other Distinctions in Human Performance Models. **IEEE Transactions on Systems, Man, and Cybernetics**, v. 13, n. 3, p. 257–266, 1983.

REASON, J.; HOBBS, A. **Managing Maintenance Error**. [S.l.]: CRC Press, mar. 2017. DOI: 10.1201/9781315249926. Disponível em: https://doi.org/10.1201/9781315249926.

REDDY, C. K.; VINZAMURI, B. A survey of partitional and hierarchical clustering algorithms. In: DATA clustering. [S.l.]: Chapman e Hall/CRC, 2018. P. 87–110.

RENDÓN, E.; ABUNDEZ, I. M.; GUTIERREZ, C.; ZAGAL, S. D.; ARIZMENDI, A.; QUIROZ, E. M.; ARZATE, H. E. A comparison of internal and external cluster validation indexes. In: PROCEEDINGS of the 2011 American Conference, San Francisco, CA, USA. [S.l.: s.n.], 2011. v. 29, p. 1–10.

RITTER, S.; ANDERSON, J. R.; KOEDINGER, K. R.; CORBETT, A. Cognitive Tutor: Applied research in mathematics education. **Psychonomic bulletin & review**, Springer, v. 14, n. 2, p. 249–255, 2007.

RODRIGUEZ, M. Z.; COMIN, C. H.; CASANOVA, D.; BRUNO, O. M.; AMANCIO, D. R.; F. COSTA, L. da; RODRIGUES, F. A. Clustering algorithms: A comparative approach. Edição: Hans A Kestler. **PLOS ONE**, Public Library of Science (PLoS), v. 14, n. 1, e0210236, jan. 2019. DOI: 10.1371/journal.pone.0210236. Disponível em: https://doi.org/10.1371/journal.pone.0210236.

ROMERO, C.; VENTURA, S. Educational data mining: A survey from 1995 to 2005. **Expert Systems with Applications**, Elsevier BV, v. 33, n. 1, p. 135–146, jul. 2007. DOI: 10.1016/j.eswa.2006.04.005. Disponível em: https://doi.org/10.1016/j.eswa.2006.04.005.

ROMERO, C.; VENTURA, S. Educational data mining and learning analytics: An updated survey. **WIREs Data Mining and Knowledge Discovery**, Wiley, v. 10, n. 3, jan. 2020. DOI: 10.1002/widm.1355. Disponível em: https://doi.org/10.1002/widm.1355.

ROMERO, G.; MAROTO, J.; FELEZ, J.; CABANELLAS, J. M.; MARTINEZ, M. L.; CARRETERO, A. Virtual reality applied to a full simulator of electrical substations. **Electric Power Systems Research**, 78, Issue, 3 march, p. 409–417, 2008.

ROSENDO, M.; GRUBER, Y. A.; SANTOS, J. C. C.; MANCIA, L. B.; GEUS, K. d.; BEE, R. T.; GODOI, W. C. Smapping: A high level framework for modelling non linear maintenance procedures on power substations using relational database. Aceito no ICVR 2019 Singapore. Não publicado. [S.l.], 2019.

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

SCHERER, D.; VIEIRA, M. F. Q.; NETO, J. A. N. Human Error Categorization: An Extension to Classical Proposals Applied to Electrical Systems Operations. In: IFIP Advances in Information and Communication Technology. [S.l.]: Springer Berlin Heidelberg, 2010. P. 234–245. DOI: 10.1007/978-3-642-15231-3_23. Disponível em: https://doi.org/10.1007%2F978-3-642-15231-3_23.

SHAFFER, D.; RUIS, A. Epistemic network analysis: A worked example of theory-based learning analytics. **Handbook of learning analytics**, 2017.

SHEPHERD, A. **Hierarchial Task Analysis**. [S.l.]: CRC Press, set. 2003. DOI: 10.1201/9781482289206. Disponível em: https://doi.org/10.1201/9781482289206.

SIEMENS, G.; BAKER, R. S. J. d. Learning analytics and educational data mining. In: PROCEEDINGS of the 2nd International Conference on Learning Analytics and Knowledge - LAK '12. [S.l.]: ACM Press, 2012. DOI: 10.1145/2330601.2330661. Disponível em: https://doi.org/10.1145/2330601.2330661.

SILVA, A. Análise organizacional de acidentes de trabalho no setor de distribuição de energia elétrica. **Universidade Estadual Paulista**, p. 209, 2015. Disponível em: https://repositorio.unesp.br/bitstream/handle/11449/139369/000860423.pdf?sequence=1&isAllowe

SILVA, L.; PERES, S.; BOSCARIOLI, C. **Introdução a Mineração de Dados com aplicações em R**. Rio de Janeiro: Elsevier, 2016.

SILVA, L. G. G. da; MOREIRA, J. M. L. Causas possíveis de acidentes de trabalho fatais de origem elétrica. In: ANAIS do ENEGEP 2019 - Encontro Nacional de

Engenharia de Produção. [S.l.]: ABREPO, nov. 2019. DOI: 10.14488/enegep2019_tn_sto_297_1679_37636. Disponível em: https://doi.org/10.14488/enegep2019_tn_sto_297_1679_37636.

SILVERMAN, J. **New Techniques in Task Analysis**. San Diego, California, 1967.

SMITH, P. L.; RAGAN, T. Impact of R. M. Gagne's Work on Instructional Theory. In: PROCEEDINGS of Selected Research and Development Presentations at the 1996 National Convention of the Association for Educational Communications and Technology. [S.l.: s.n.], 1996.

SODERSTROM, N.; BJORK, R. Learning Versus Performance: An Integrative Review. **Perspectives on psychological science : a journal of the Association for Psychological Science**, v. 10, p. 176–199, mar. 2015. DOI: 10.1177/1745691615569000.

SOTTILARE, R. A.; GRAESSER, A.; HU, X.; (EDS.), H. H. **Design Recommendations for Intelligent Tutoring Systems**. [S.l.]: U.S. Army Research Laboratory, ago. 2013. 4 Vols. ISBN 9780989392303.

SOTTILARE, R.; PEREZ, R.; ALEXANDER, T.; FLETCHER, D.; SKINNER, A.; ZOTOV, V.; DURLACH, P. **Assessment of Intelligent Tutoring Systems Technologies and Opportunities**. Neuilly-sur-Seine: North Atlantic Treaty Organization, Research e Technology Organization, 2018. ISBN 978-92-837-2091-1.

SOTTILARE, R. A. A Comprehensive Review of Design Goals and Emerging Solutions for Adaptive Instructional Systems. **Technology, Instruction, Cognition & Learning**, v. 11, n. 1, 2018.

STANIĆ, Z.; JOVANOVIĆ, I. Spectral Distances of Graphs Based on their Different Matrix Representations. **Filomat**, v. 28, p. 723–734, jan. 2014. DOI: 10.2298/FIL1404723J.

STELLAN, O.; MITROVIĆ, A. Constraint-based knowledge representation for individualized instruction. **Computer Science and Information Systems**, v. 3, n. 1, p. 1–22, 2006.

STREHL, A.; GHOSH, J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. **Journal of machine learning research**, v. 3, Dec, p. 583–617, 2002.

STRUYF, A.; HUBERT, M.; ROUSSEEUW, P. et al. Clustering in an object-oriented environment. **Journal of Statistical Software**, v. 1, n. 4, p. 1–30, 1997.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introducao ao datamining: mineracao de dados**. Rio de Janeiro: Ciencia Moderna, 2009.

TANAKA, E. H.; PALUDO, J. A.; CORDEIRO, C. S.; DOMINGUES, L. R.; GADBEM, E. V.; EUFLAUSINO, A. Using immersive virtual reality for electrical substation training. **International Conference e-Learning**, p. 136–140, 2015.

TAYLOR, L. Educational theories and instructional design models. Their place in simulation. **Nursing Education and Research, Southern Health**, 2004.

TENNYSON, R. D. Historical Reflection on Learning Theories and Instructional Design. **Contemporary Educational Technology**, v. 1, n. 1, p. 1–16, 2010.

TULIS, M.; STEUER, G.; DRESEL, M. Learning from Errors: A Model of Individual Processes. **Frontline Learning Research**, ERIC, v. 4, n. 2, p. 12–26, 2016.

VEGA-PONS, S.; RUIZ-SHULCLOPER, J. A survey of clustering ensemble algorithms. **International Journal of Pattern Recognition and Artificial Intelligence**, World Scientific, v. 25, n. 03, p. 337–372, 2011.

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

WATERSON, P.; COZE, J.-C. le; ANDERSEN, H. Recurring themes in the legacy of Jens Rasmussen. **Applied Ergonomics**, v. 59, out. 2016. DOI: 10.1016/j.apergo.2016.10.002.

WILLS, P. **NetComp v 0.2**. [S.l.: s.n.], 2017. Disponível em: https://github.com/peterewills/NetComp.

WILLS, P.; MEYER, F. G. Metrics for graph comparison: A practitioners guide. Edição: Pin-Yu Chen. **PLOS ONE**, Public Library of Science (PLoS), v. 15, n. 2, fev. 2020. DOI: 10.1371/journal.pone.0228728. Disponível em: https://doi.org/10.1371/journal.pone.0228728.

WU, J.; CHEN, J.; XIONG, H.; XIE, M. External validation measures for K-means clustering: A data distribution perspective. **Expert Systems with Applications**, Elsevier, v. 36, n. 3, p. 6050–6061, 2009.

WULANDARI, D. A. N.; ANNISA, R.; YUSUF, L.; PRIHATIN, T. EDUCATIONAL DATA MINING FOR STUDENT ACADEMIC PREDICTION USING K-MEANS CLUSTERING AND NAÏVE BAYES CLASSIFIER. **Jurnal Pilar Nusa Mandiri**, v. 16, n. 2, p. 155–160, 2020.

XI, C.; WU, H.; JOHER, A.; KIRSCH, L.; LUO, C.; KHASAWNEH, M. 3-D virtual reality for education, training and improved human performance in nuclear applications. In: 6TH American Nuclear Society International Topical Meeting on Nuclear Plant Instrumentation, Control, and Human-Machine Interface Technologies 2009. [S.l.: s.n.], 2009. P. 2347–2356.

XIONG, H.; WU, J.; CHEN, J. K-means clustering versus validation measures: a data-distribution perspective. **IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)**, IEEE, v. 39, n. 2, p. 318–331, 2008.

XU, Q.; DING, C.; LIU, J.; BIN, L. PCA-guided search for K-means. **Pattern Recognition Letters**, v. 54, dez. 2014. DOI: 10.1016/j.patrec.2014.11.017.

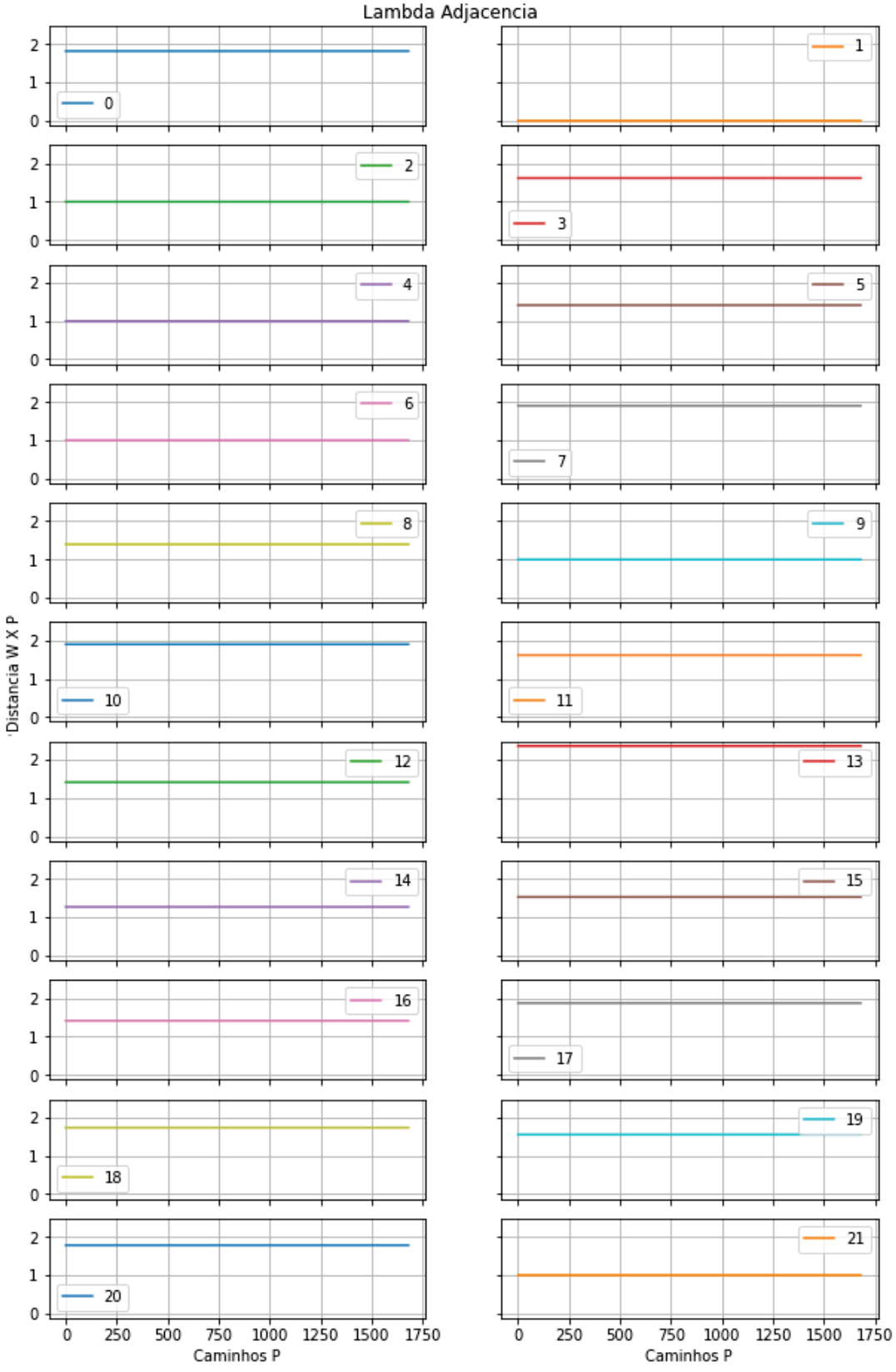
ZHANG, Y.; LI, T. Consensus Clustering + Meta Clustering = Multiple Consensus Clustering. **Proceedings of the 24th International Florida Artificial Intelligence Research Society, FLAIRS - 24**, jan. 2011.

ZHOU, Z.-H. **Ensemble methods: foundations and algorithms**. [S.l.]: Chapman e Hall/CRC, 2019.

APÊNDICE 1 - PADRÕES DE SIMILARIDADE

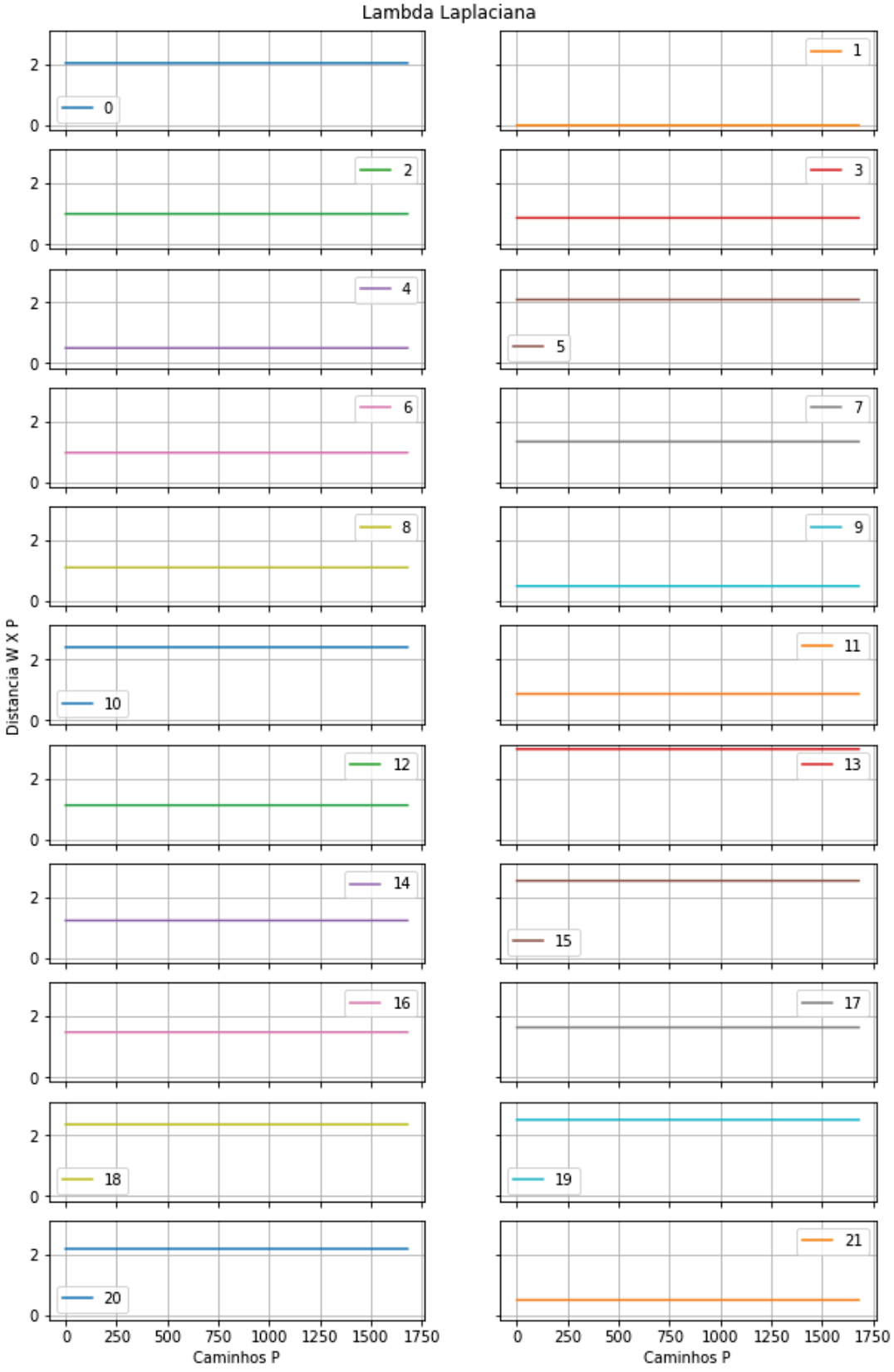
PADRÕES DE SIMILARIDADE $(\mathcal{W} \times \mathcal{P})$

FIGURA 77 – Padrão de similaridade $(\mathcal{W} \times \mathcal{P})_{\lambda_{Adj}}$



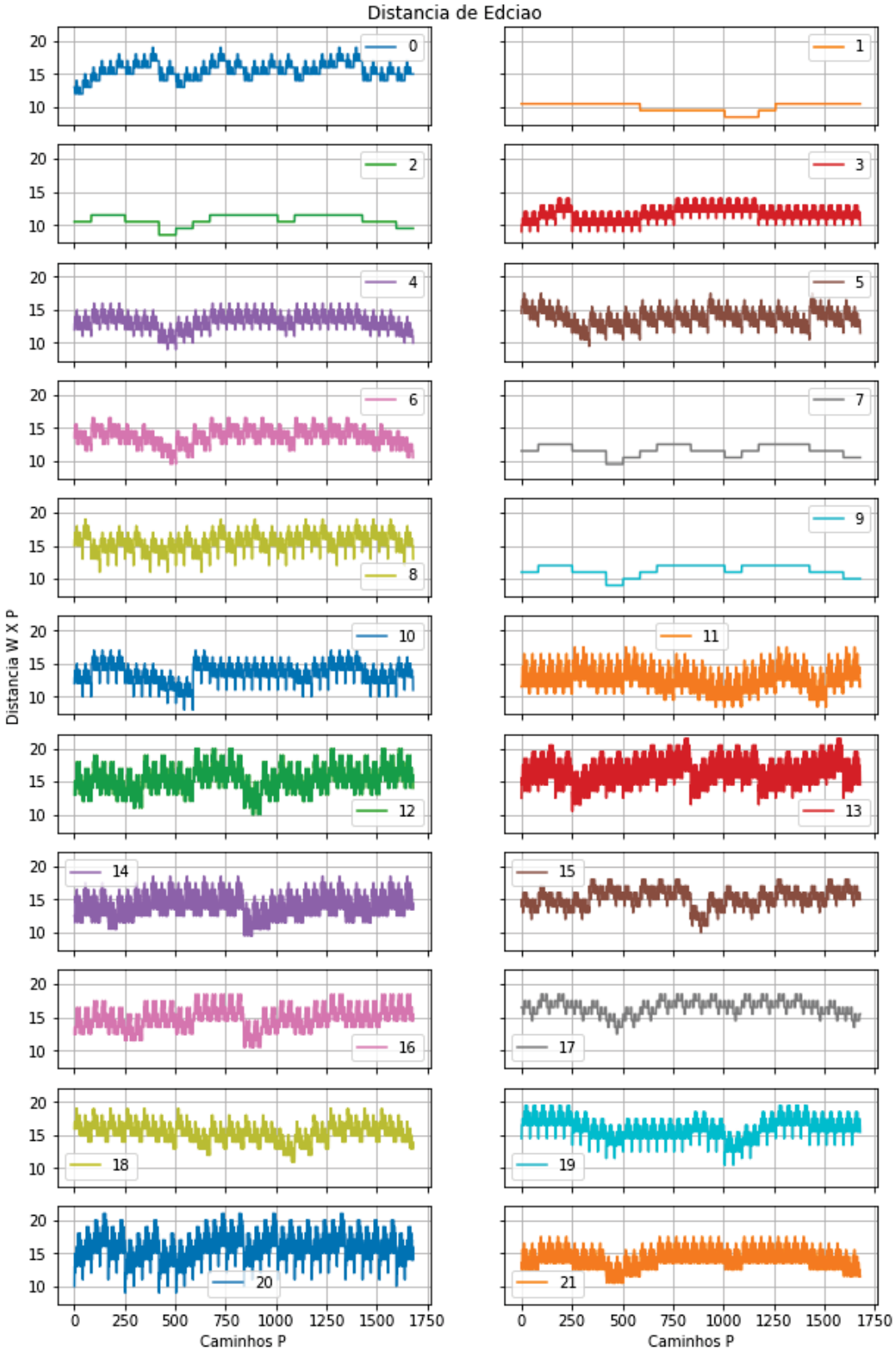
FONTE: O Autor

FIGURA 78 – Padrão de similaridade $(\mathcal{W} \times \mathcal{P})_{\lambda_{Lap}}$



FONTE: O Autor

FIGURA 79 – Padrão de similaridade $(\mathcal{W} \times \mathcal{P})_{GED}$



FONTE: O Autor

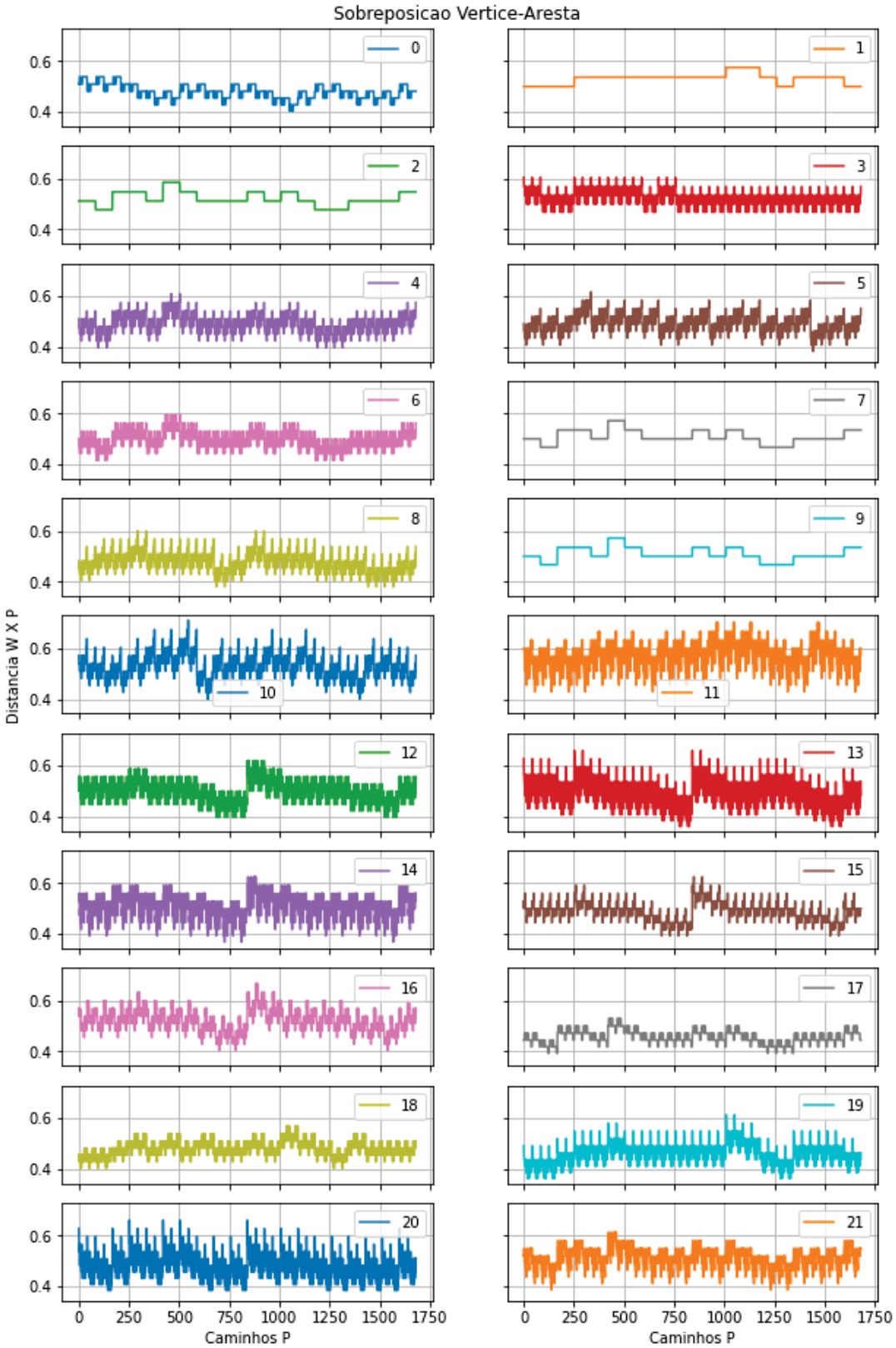
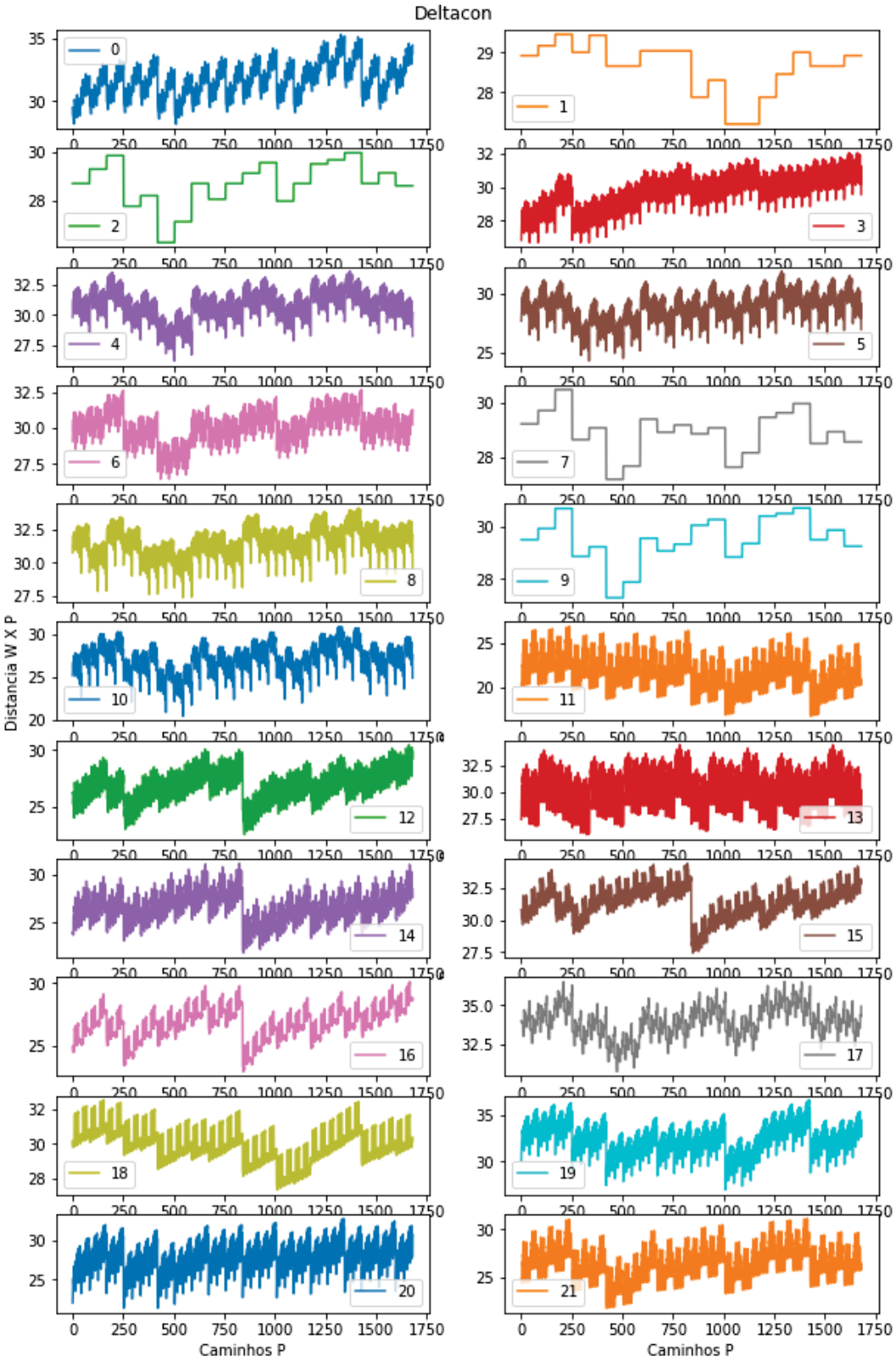


FIGURA 80 – Padrão de similaridade $(\mathcal{W} \times \mathcal{P})_{VEO}$

FONTE: O Autor

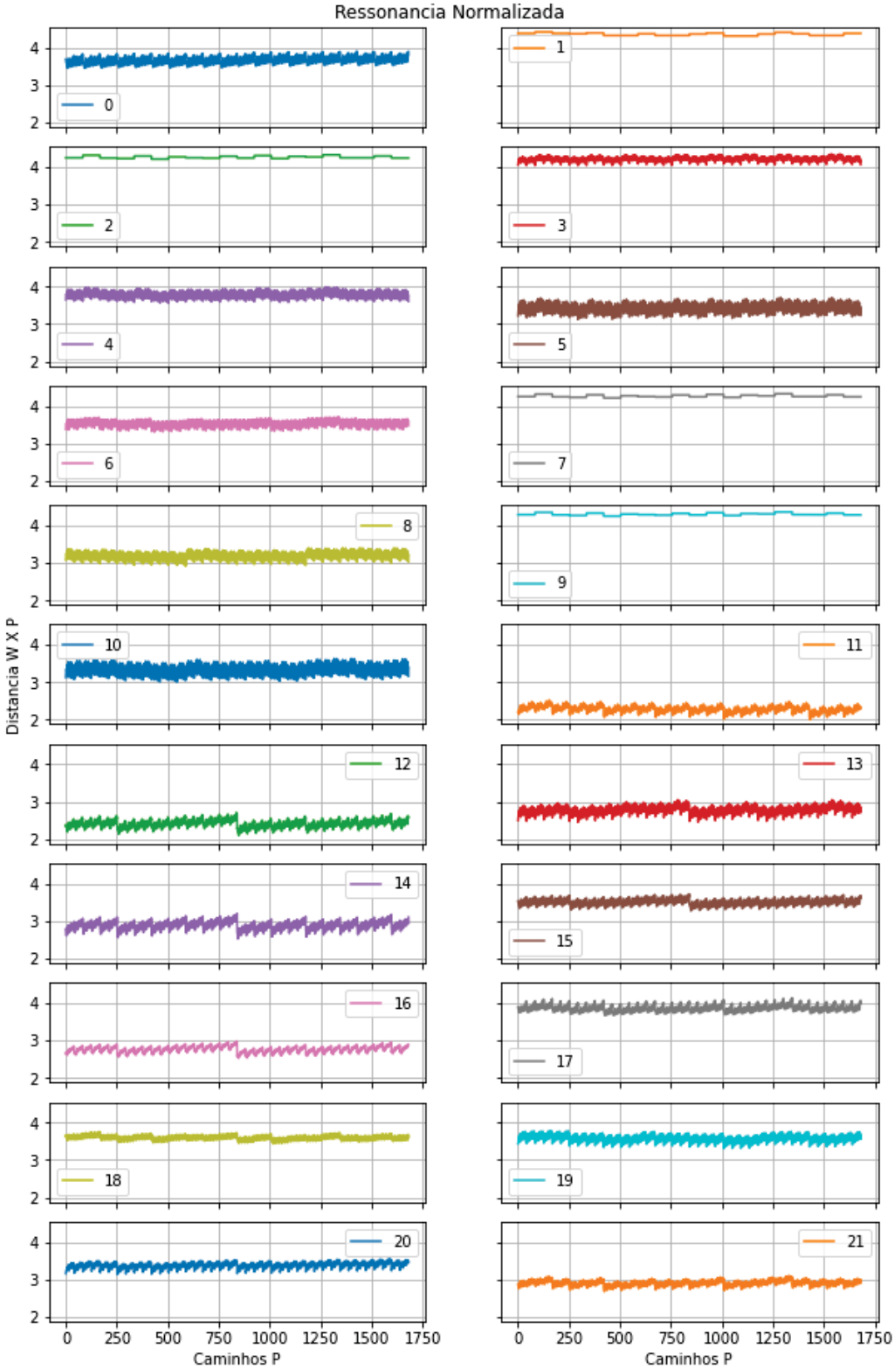
PADRÃO DE SIMILARIDADE $(\mathcal{W} \times \mathcal{P})$

FIGURA 81 – Padrão de similaridade $(\mathcal{W} \times \mathcal{P})_{\delta-con}$



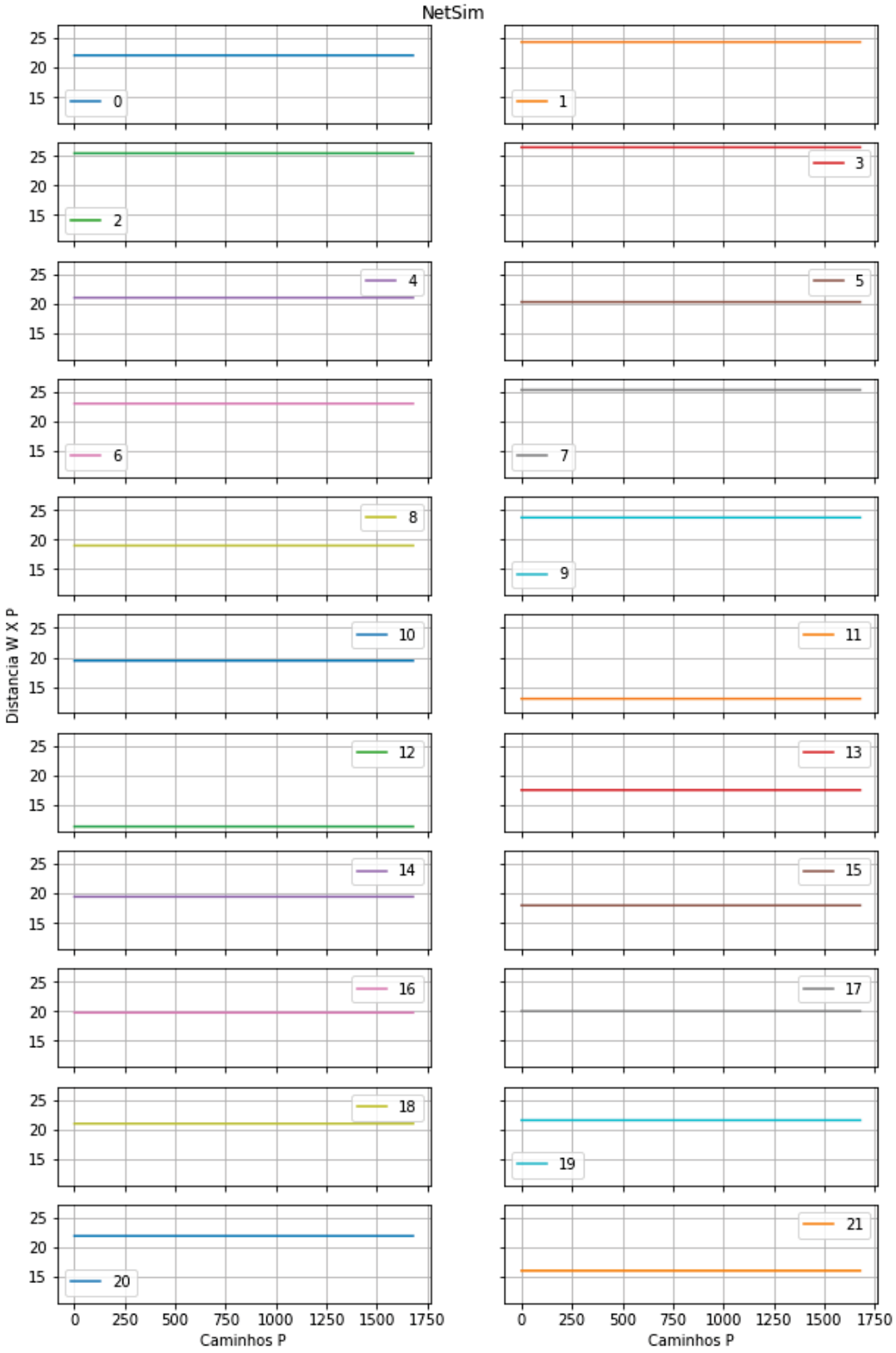
FONTE: O Autor

FIGURA 82 – Padrão de similaridade $(\mathcal{W} \times \mathcal{P})_{Res}$



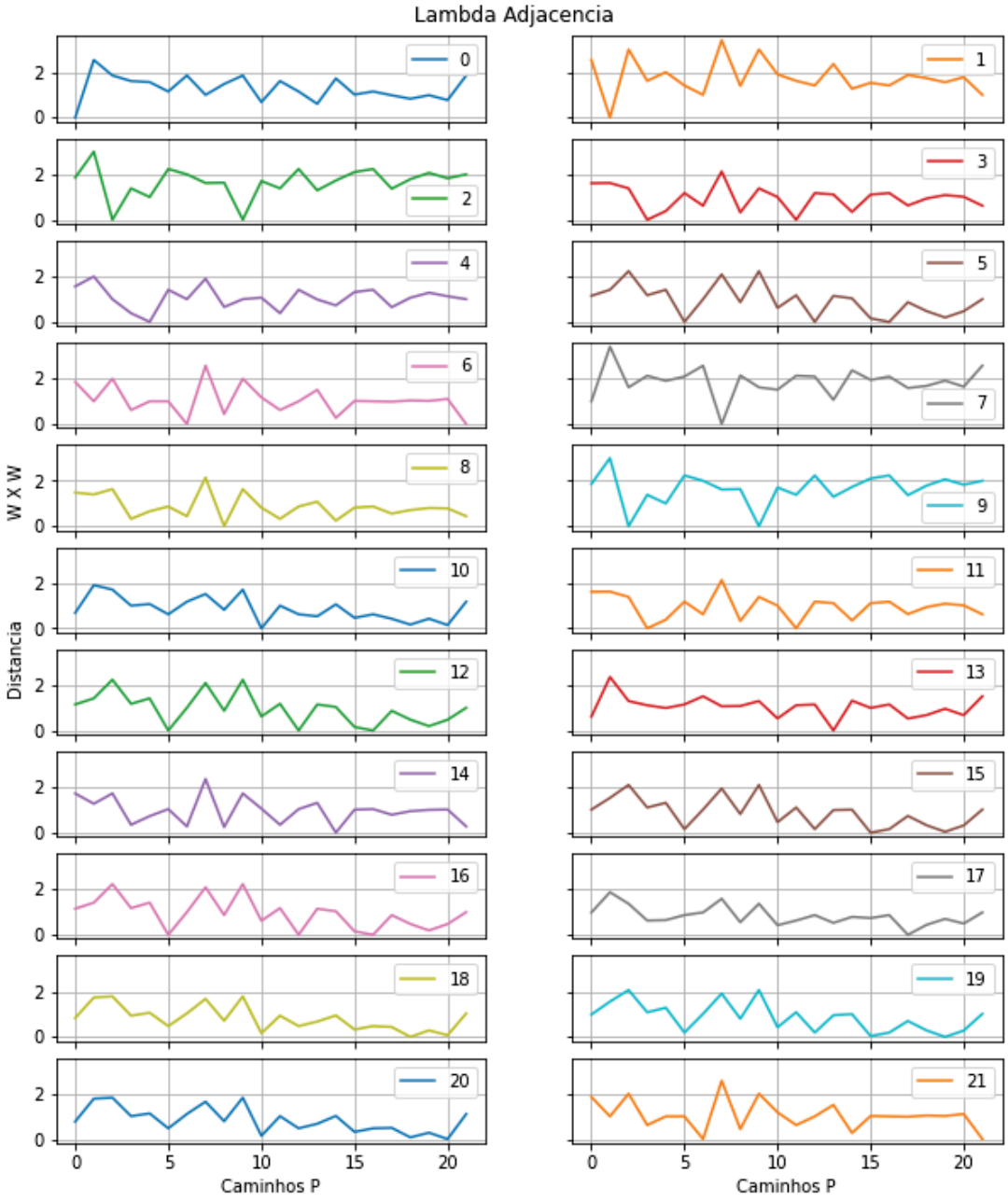
FONTE: O Autor

FIGURA 83 – Padrão de similaridade $(\mathcal{W} \times \mathcal{P})_{Netsim}$



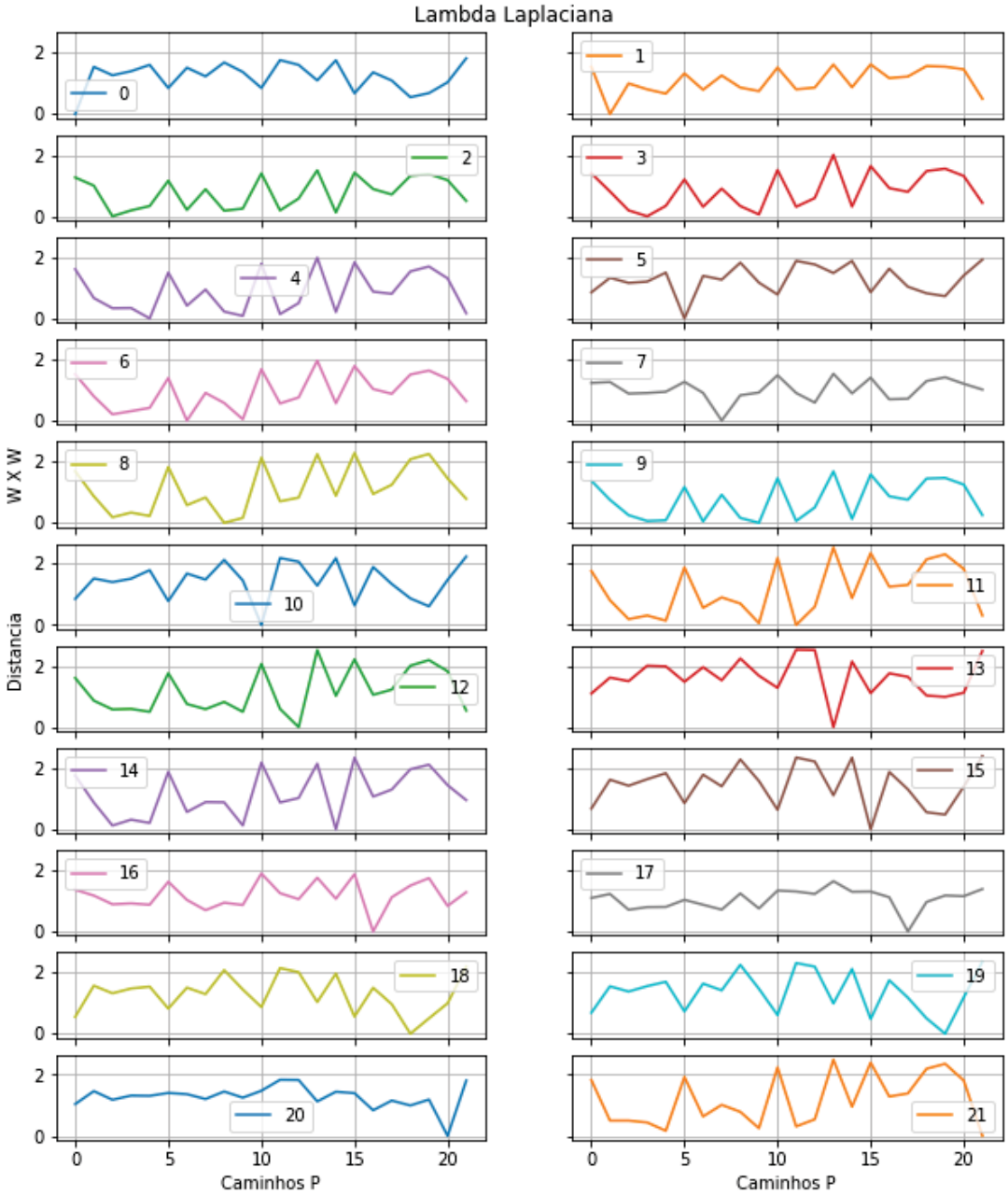
FONTE: O Autor

FIGURA 84 – Padrão de similaridade $(\mathcal{W} \times \mathcal{W})_{\lambda_{Adj}}$



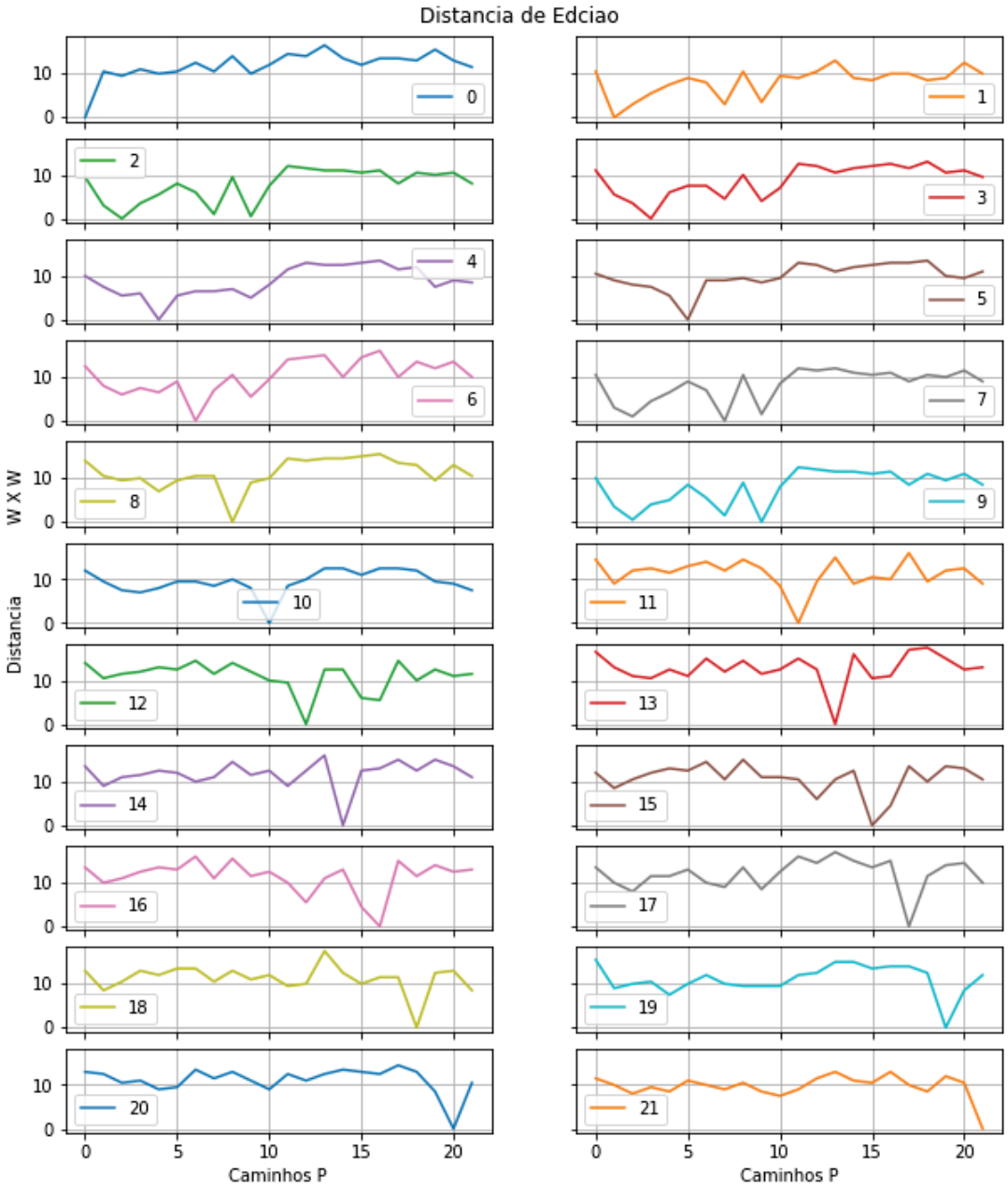
FONTE: O Autor

FIGURA 85 – Padrão de similaridade $(\mathcal{W} \times \mathcal{W})_{\lambda_{Lap}}$



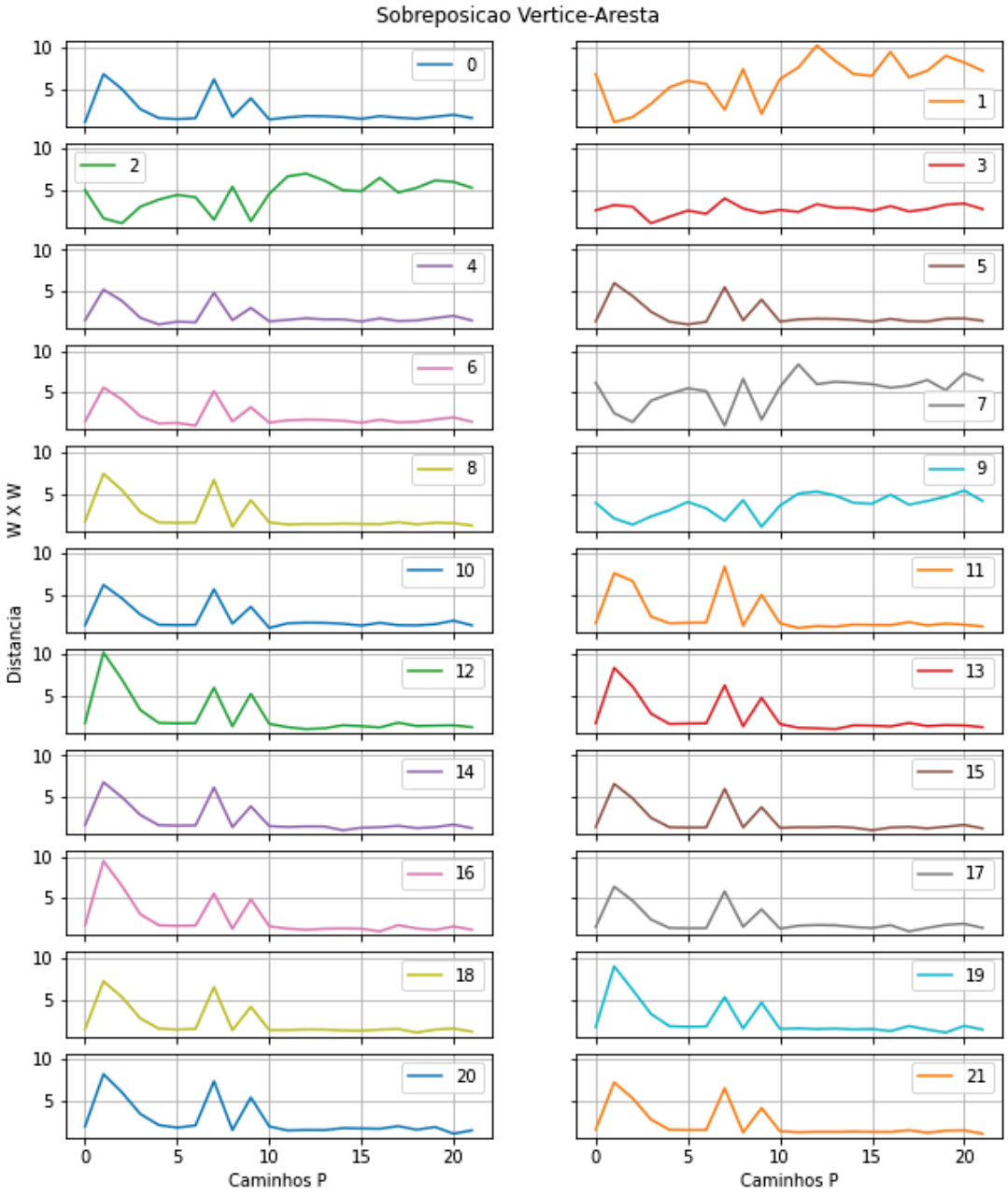
FONTE: O Autor

FIGURA 86 – Padrão de similaridade $(\mathcal{W} \times \mathcal{W})_{GED}$



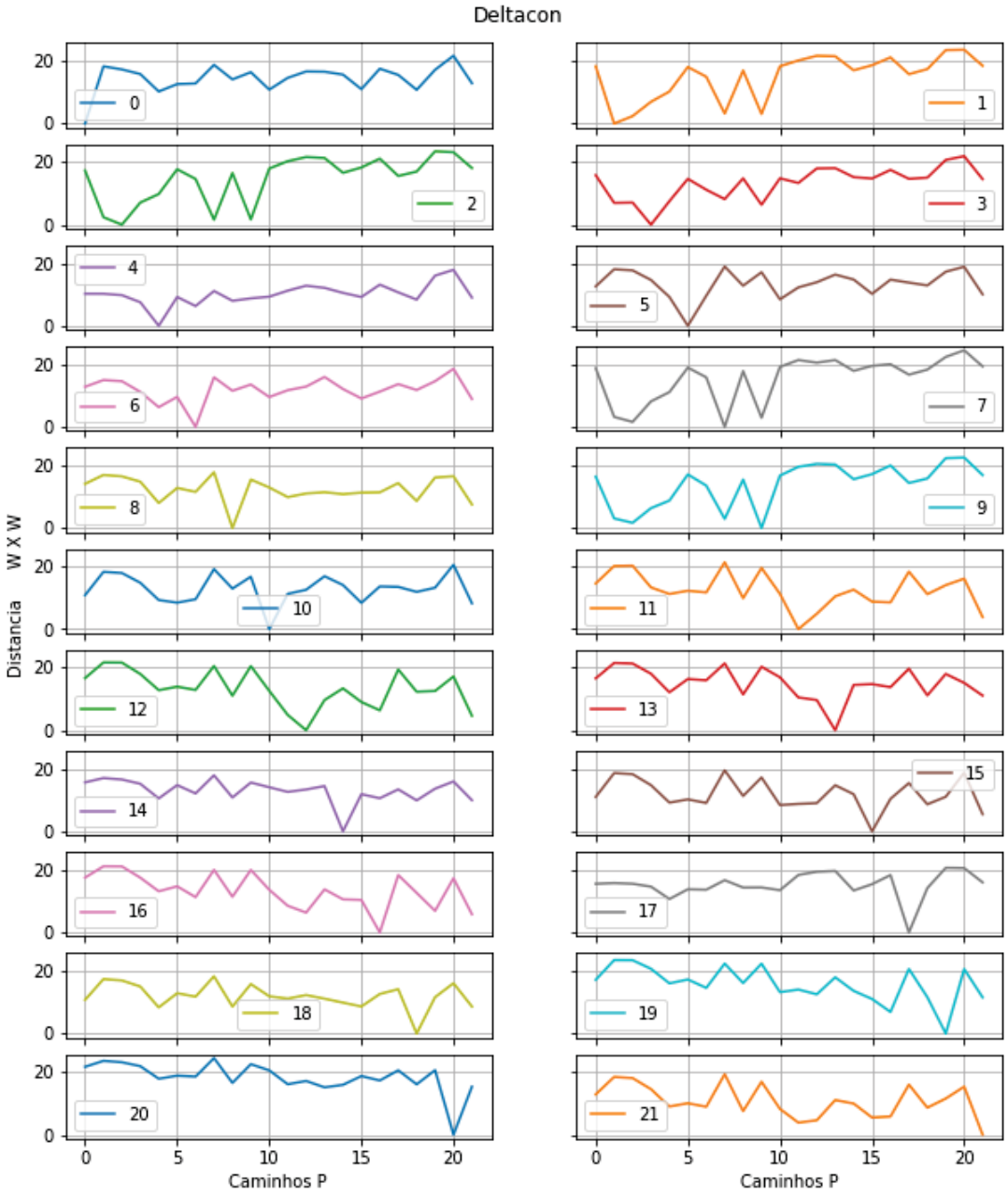
FONTE: O Autor

FIGURA 87 – Padrão de similaridade $(\mathcal{W} \times \mathcal{W})_{VEO}$



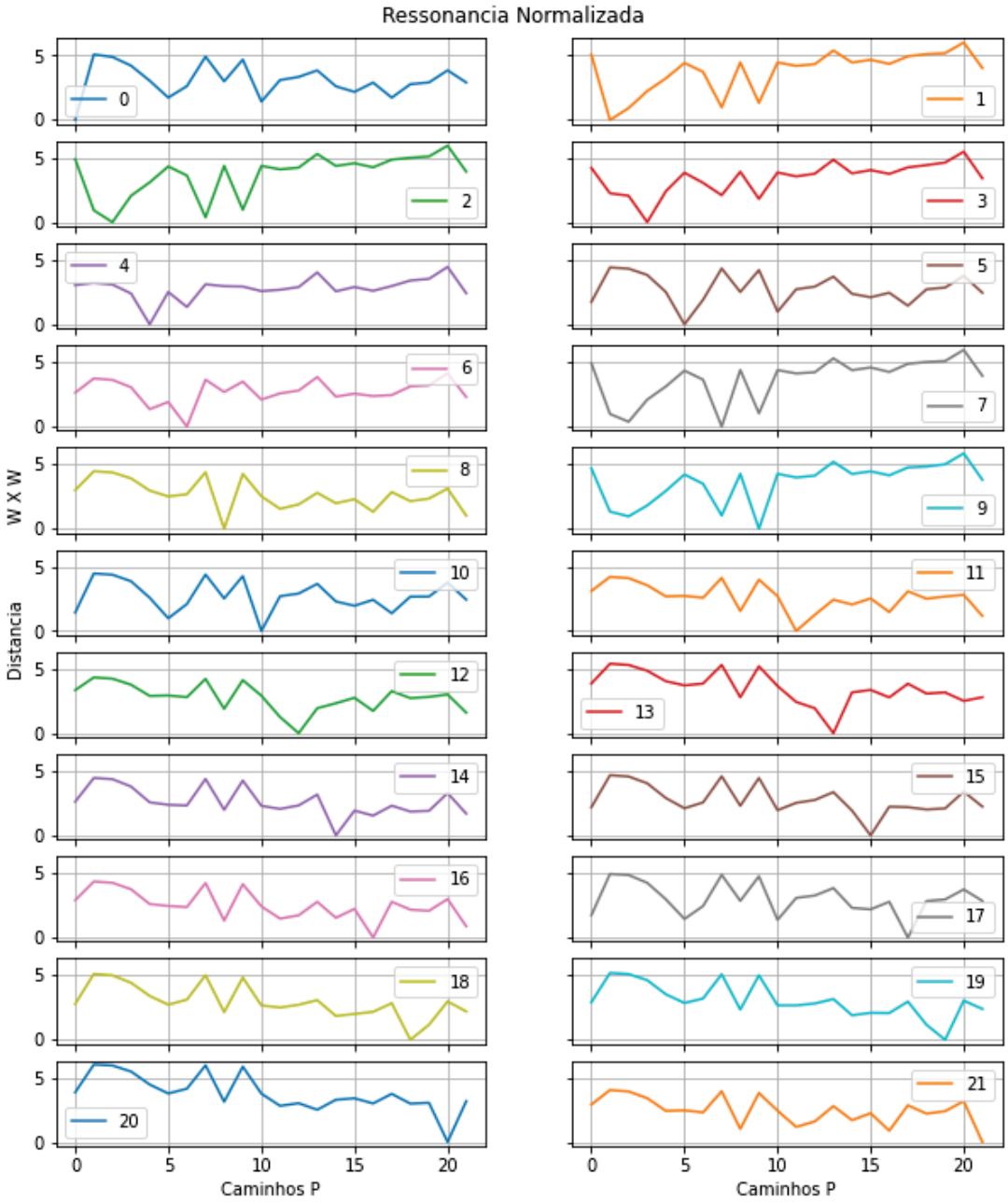
FONTE: O Autor

FIGURA 88 – Padrão de similaridade $(\mathcal{W} \times \mathcal{W})_{\delta-con}$



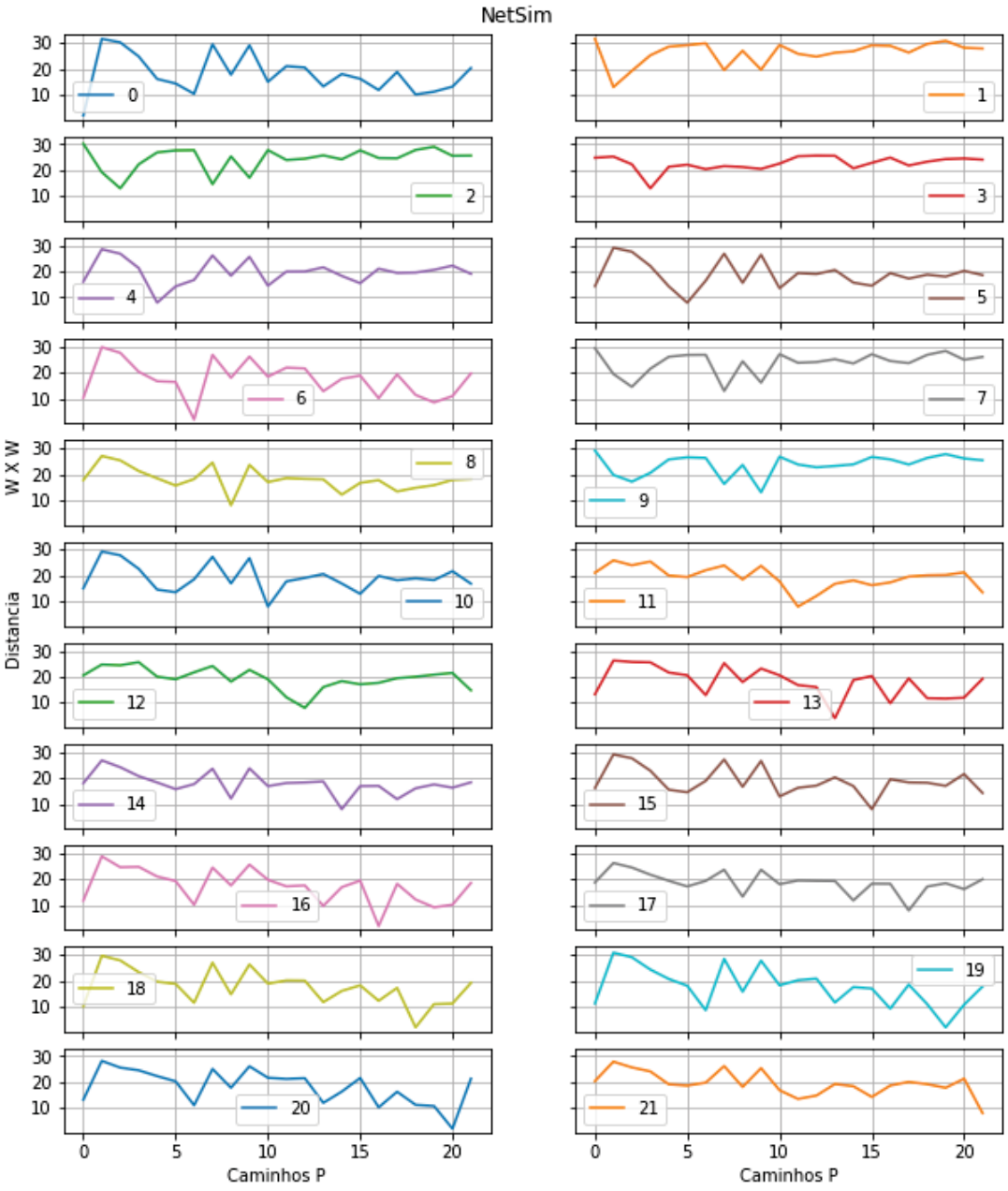
FONTE: O Autor

FIGURA 89 – Padrão de similaridade $(\mathcal{W} \times \mathcal{W})_{Res}$



FONTE: O Autor

FIGURA 90 – Padrão de similaridade $(\mathcal{W} \times \mathcal{W})_{Netsim}$

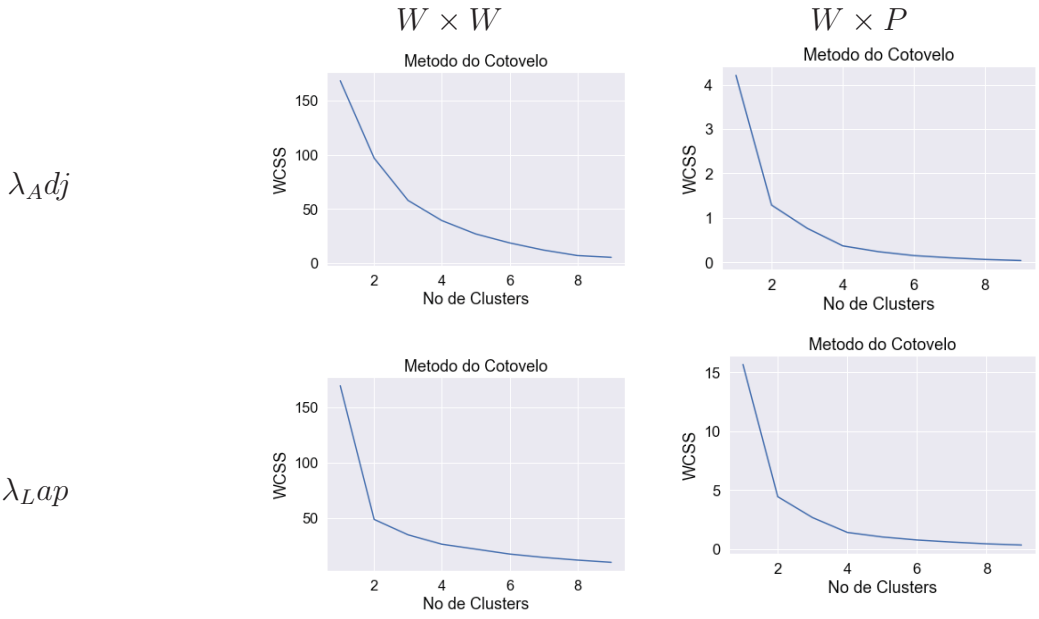


FONTE: O Autor

APÊNDICE 2 - ESTIMATIVA DO NÚMERO DE GRUPOS

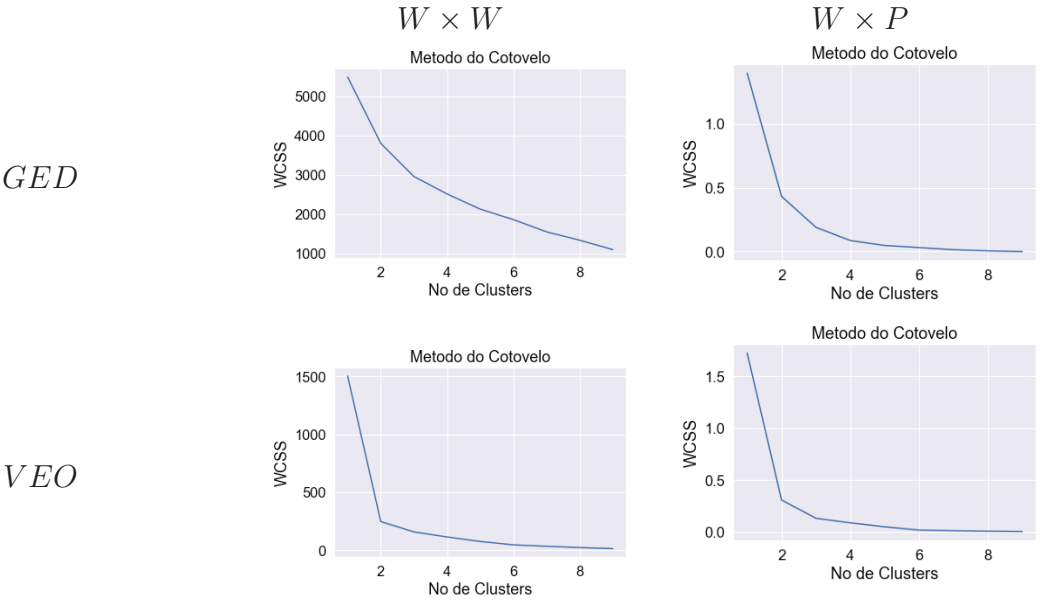
ESTIMATIVA DO NÚMERO DE GRUPOS

FIGURA 91 – Gráficos do método do cotovelo utilizados para estimar número k de grupos: medidas de similaridade espectrais.



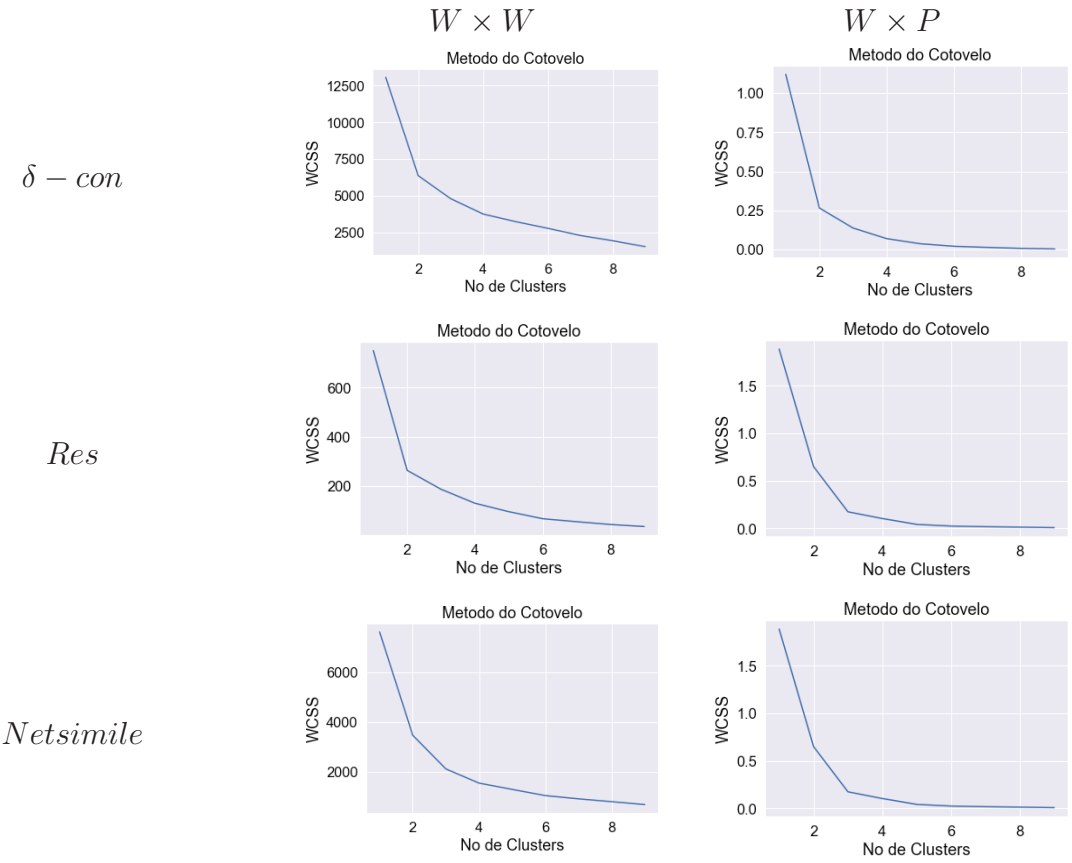
FONTE: O Autor

FIGURA 92 – Gráficos do método do cotovelo utilizados para estimar número k de grupos: medidas de similaridade matriciais *GED* e *VEO*



FONTE: O Autor

FIGURA 93 – Gráficos do método do cotovelo utilizados para estimar número k de grupos: medidas de similaridade Deltacon, Resitência e Netsim



FONTE: O Autor