



UNIVERSIDADE FEDERAL DE CAMPINA GRANDE

Programa de Pós-Graduação em Matemática

Mestrado Profissional - PROFMAT/CCT/UFCG



PROFMAT

Erivan Barbosa da Silva

***Clusters* de Unidades Federativas do Brasil,
segundo características socioeconômicas e
educacionais dos inscritos no Enem 2021**

Campina Grande - PB

Agosto/2023



UNIVERSIDADE FEDERAL DE CAMPINA GRANDE

Programa de Pós-Graduação em Matemática

Mestrado Profissional - PROFMAT/CCT/UFCG



PROFMAT

Erivan Barbosa da Silva

Clusters de Unidades Federativas do Brasil, segundo características socioeconômicas e educacionais dos inscritos no Enem 2021

Trabalho de Conclusão de Curso apresentado ao Corpo Docente do Programa de Pós-Graduação em Matemática - CCT - UFCG, na modalidade Mestrado Profissional, como requisito parcial para obtenção do título de Mestre.

Orientador: Dr. José Fernando Leite Aires

Coorientador: Dr. José Iraponil Costa Lima

Campina Grande - PB

Agosto/2023

S586c

Silva, Erivan Barbosa da.

Clusters de Unidades Federativas do Brasil, segundo características socioeconômicas e educacionais dos inscitos no Enem 2021 / Erivan Barbosa da Silva. - Campina Grande, 2023.

85 f. : il. color.

Dissertação (Mestrado em Matemática) - Universidade Federal de Campina Grande, Centro de Ciências e Tecnologia, 2023.

"Orientação: Prof. Dr. José Fernando Leite Aires, Prof. Dr. José Iraponil Costa Lima."

Referências.

1. *Clusters*. 2. Análise de Agrupamentos. 3. Unidades Federativas do Brasil. 4. Microdados do Enem 2021. 5. Questionário Socioeconômico. 6. Cotas na Educação. I. Aires, José Fernando Leite. II. Lima, José Iraponil Costa. III. Título.

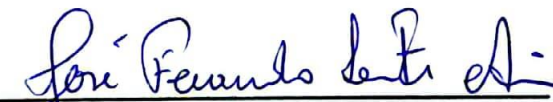
CDU 519.22(043)

Erivan Barbosa da Silva

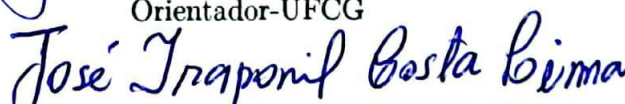
Clusters de Unidades Federativas do Brasil, segundo características socioeconômicas e educacionais dos inscritos no Enem 2021

Trabalho de Conclusão de Curso apresentado ao Corpo Docente do Programa de Pós-Graduação em Matemática - CCT - UFCG, na modalidade Mestrado Profissional, como requisito parcial para obtenção do título de Mestre.

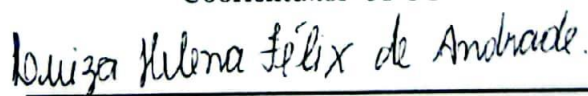
Trabalho aprovado. Campina Grande - PB, 11 de agosto de 2023:



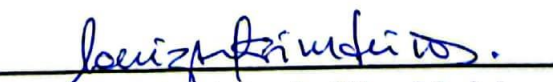
Dr. José Fernando Leite Aires
Orientador-UFCG



Dr. José Iraponil Costa Lima
Coorientador-UFCG



Dr.^a Luiza Helena Félix de Andrade
Membro Externo-UFERSA



Dr. Luiz Antônio da Silva Medeiros
Membro Interno-UFCG

Campina Grande - PB
Agosto/2023

Agradecimentos

A Deus por ter me concedido força e determinação para mais uma vitória em minha vida.

A minha esposa Rosiane por ter me incentivado e apoiado sempre, dando-me força e compreensão, fazendo com que eu sempre acreditasse nos meus sonhos e lutasse pelos meus objetivos, a quem sou muito grato pelo amor, carinho e paciência.

Ao meu filho Emanuel a quem amo e dedico como forma de incentivo e que somente a educação pode mudar vidas.

Ao meus pais Marluce e Nerivaldo, meus irmãos Nerivaldo Filho, Erivaldo e Eri-
vania e minha sobrinha Marilia Gabriela, minhas bases, na qual me fundamentou e fundamenta, com amor, dedicação e carinho, agradeço.

Aos meus avós paternos e maternos, em especial à Dona Irene e Zé Vaqueiro (In Memoriam), a quem estimo-os e sou grato pela educação e força que me foi concedido para me tornar a pessoa que sou hoje.

Aos meus tios e tias, primos e primas, amigos e amigas, que contribuíram direto e indiretamente e torceram para mais essa realização na minha vida.

A turma do mestrado, André Macedo, Andreson Alquino, Benildo Virgínio, Carlos Gonzaga, Cláudio Teodista, Eli Azevedo, Érico Felintro, Gilmar Verissimo, Gilvandro Melo, Idalice Maria, João Evayr, Rafael Augusto, Wellington Rodrigues e Wirander Oliveira que foram como irmãos nessa batalha em que o foco foi a vitória coletiva.

Ao meu amigo Jalmir que também foi e é um batalhador onde convivemos lutas diárias de estudos e compartilhamentos de conhecimentos.

Aos professores do Profmat, Dr.^a Deise Mara, Dr. Daniel Cordeiro, Dr. Jaime, Dr. José de Arimatéia, Dr. Leomarques, Dr. Luiz Antônio, Dr. Marcelo, Dr. Rodrigo Cohen e Dr. José Fernando à vocês a minha eterna gratidão, obrigado por toda dedicação que tiveram para com nós e pelo conhecimento transmitido com amor.

Ao meu orientador Dr. José Fernando Leite Aires e ao coorientador Dr. José Iraponil Costa Lima pela sua dedicação, paciência, respeito e amizade que sempre teve comigo, sou grato por todo conhecimento repassado e por sua disponibilidade.

Aos membros da banca examinadora Dr. Luiz Antônio da Silva Medeiros, Dr. Tiago Almeida de Oliveira, Dr. Rodrigo Cohen Mota Neme e a Dr.^a Luiza Helena Félix de Andrade por aceitarem o convite e contribuírem para o enriquecimento desse trabalho.

*“Sem sonhos, a vida não tem brilho.
Sem metas, os sonhos não tem alicerces.
Sem prioridades, os sonhos não se tornam reais.
Sonhe, trace metas, estabeleça prioridades
e corra riscos para executar seus sonhos.”*
(Augusto Cury)

Resumo

O presente trabalho tem por objetivo investigar a existência de agrupamentos (*clusters*) de Unidades Federativas (UFs) do Brasil, segundo características socioeconômicas e educacionais dos inscritos no Enem 2021. Para tal finalidade, empregamos as ferramentas da Análise de Agrupamentos (AA), implementando métodos hierárquicos e não hierárquicos (*k-means*) por meio do *Software R*. Nesse contexto, variáveis tais como grau de instrução dos pais, ocupação da mãe dos inscritos e a proporção de pessoas inscritas que se autodeclararam preta, parda ou indígena são utilizadas para caracterizar as UFs. Aplicada a metodologia da Análise de Agrupamentos à base de Microdados do Enem 2021, constatamos a presença de *clusters* de UFs em algumas das regiões geográficas brasileiras, além disso identificamos a presença de UFs isoladas.

Palavras-chave: Análise de Agrupamentos. Unidades Federativas do Brasil. *Clusters*. Enem 2021. Questionário Socioeconômico.

Abstract

The present work has to objective investigate the existence of groupings (clusters) of Federative Units (UF) of Brazil, according to socioeconomic and educational characteristics of the registered in Enem 2021. For this purpose, we used the Cluster Analysis, implementing hierarchical and non-hierarchical methods (k-means) for software **R**. In this context, variables such as education level of the parents, occupation of the mother and the proportion of people that declared themselves black, brown or indigenous are used to characterize the UF. Applying the methodology of Cluster Analysis based in microdata of Enem 2021, we found the presence of clusters of UF in some of the Brazilian geographic regions, in addition to identifying the presence of UF isolated.

Keywords: Cluster Analysis. Federative Units of Brazil. Clusters. Enem 2021. Socioeconomic Quiz.

Lista de ilustrações

Figura 1 – Interconexões entre competências e habilidades.	21
Figura 2 – Dados brutos e dados padronizados.	31
Figura 3 – Representação Gráfica em 3D das provas, segundo os valores padronizados das variáveis X_1 , X_2 e X_3	32
Figura 4 – Dendograma das provas, segundo o Método da Média das Distâncias.	38
Figura 5 – Dendograma das provas, segundo o Método da Centróide)	52
Figura 6 – Dendograma das provas, segundo o Método da Ligação Simples.	56
Figura 7 – Dendograma das provas, segundo o Método da Ligação Completa.	59
Figura 8 – Gráfico de Correlação de Pearson das Variáveis Socioeconômicas e Educacionais em Estudo.	66
Figura 9 – Dendograma das Unidades Federativas do Brasil, segundo o M.L.S.	67
Figura 10 – Dendograma das Unidades Federativas do Brasil, segundo o M.L.C.	68
Figura 11 – Dendograma das Unidades Federativas do Brasil, segundo o M.M.D.	69
Figura 12 – Gráfico do Nível de Fusão do Método da Média das Distâncias (M.M.D.).	70
Figura 13 – Dendograma das Unidades Federativas do Brasil segundo o Método da Média das Distâncias.	71
Figura 14 – Gráfico da Árvore Filogenética Aplicado ao M.M.D. após a interrupção do algoritmo no Passo 12.	72
Figura 15 – <i>Clusters</i> das Unidades Federativas do Brasil, segundo o <i>k-means</i>	74
Figura 16 – Rotina do R para o Exemplo Básico.	86
Figura 17 – Rotina do R para o Exemplo Básico.	87

Lista de tabelas

Tabela 1 – Áreas de conhecimento e suas respectivas disciplinas.	23
Tabela 2 – Relação entre competências, habilidades e eixos cognitivos da área de Matemática e suas tecnologias.	27
Tabela 3 – Provas do Enem (Matemática), segundo as variáveis X_1 , X_2 e X_3 .	30
Tabela 4 – Construção dos Valores Padronizados das Variáveis X_1 , X_2 e X_3 . . .	31
Tabela 5 – M.M.D. Passo 0.	34
Tabela 6 – M.M.D. Passo 1.	35
Tabela 7 – M.M.D. Passo 2.	35
Tabela 8 – M.M.D. Passo 3.	36
Tabela 9 – M.M.D. Passo 4.	36
Tabela 10 – Resumo do M.M.D. aplicado ao Exemplo Básico.	37
Tabela 11 – Dados das Variáveis Quantitativas do Exemplo Básico.	40
Tabela 12 – Relativização da Variáveis X_1 , X_2 e X_3	45
Tabela 13 – Resultados sobre a presença (1) ou ausência (0) de determinadas características a partir de sete variáveis X_n , com $n = 1, 2, 3, \dots, 7$.	46
Tabela 14 – Número observado de pares (1,1), (1,0), (0,1), (0,0).	47
Tabela 15 – Alguns Coeficientes de semelhança para variáveis dicotômicas(Baseado em (ROMESBURG, 1984)).	48
Tabela 16 – Variáveis Padronizados Z_1, Z_2, Z_3	49
Tabela 17 – M.C.	51
Tabela 18 – Resumo do processo hierárquico do M.C.	52
Tabela 19 – Resumo do processo hierárquico M.L.S.	55
Tabela 20 – Resumo do processo hierárquico M.L.C.	58
Tabela 21 – <i>Clusters</i> das provas do Enem 2021 (Matemática), segundo o método <i>k-means</i>	62
Tabela 22 – Número de Inscritos no Enem 2021 por Unidade Federativa (UF). . .	63
Tabela 23 – Variáveis Socioeconômicas e Educacionais dos Inscritos e suas respectivas definições e especificações para a Análise de Agrupamentos.	65
Tabela 24 – Resumo do M.L.S. Aplicado às Unidades Federativas do Brasil. . .	67
Tabela 25 – Resumo do M.L.C. Aplicado às Unidades Federativas do Brasil. . .	68
Tabela 26 – Resumo do M.M.D. Aplicado às Unidades Federativas do Brasil. . .	69
Tabela 27 – <i>Clusters</i> das Unidades Federativas do Brasil, segundo o M.M.D. . .	71
Tabela 28 – <i>Clusters</i> de Unidades Federativas do Brasil, segundo o método <i>k-means</i>	73

Tabela 29 – Proporção de inscritos por Unidade Federativa do Brasil segundo características socioeconômicas e educacionais dos inscritos no Enem 2021. (Parte 1)	83
Tabela 30 – Proporção de inscritos por Unidade Federativa do Brasil segundo características socioeconômicas e educacionais dos inscritos no Enem 2021. (Parte 2)	84

Lista de abreviaturas e siglas

AA	Análise de Agrupamentos
AC	Acre
AL	Alagoas
AM	Amazonas
AP	Amapá
BA	Bahia
CE	Ceará
DF	Distrito Federal
ENCCEJA	Exame Nacional para Certificação de Competências de Jovens e Adultos
ENEM	Exame Nacional do Ensino Médio
ES	Espírito Santo
GO	Goiás
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
LDB	Lei das Diretrizes e Bases da Educação Nacional
MA	Maranhão
MEC	Ministério da Educação
MC	Método da Centróide
MG	Minas Gerais
MLC	Método da Ligação Completa
MLS	Método da Ligação Simples
MMD	Método da Média das Distâncias
MS	Mato Grosso do Sul

MT	Mato Grosso
NSE	Nível Socioeconômico
PA	Pará
PB	Paraíba
PE	Pernambuco
PI	Piauí
PISA	Programme for International Student Assessment
PR	Paraná
PROUNI	Programa Unversidade para Todos
RJ	Rio de Janeiro
RN	Rio Grande do Norte
RO	Rondônia
RR	Roraima
RS	Rio Grande do Sul
SC	Santa Catarina
SE	Sergipe
SISU	Sistema de Seleção Unificada
SP	São Paulo
TO	Tocantins
TRI	Teoria de Resposta ao Item
UF	Unidade Federativa

Sumário

1	INTRODUÇÃO	15
1.1	Motivação e Justificativa	15
1.2	Trabalhos Relacionados	16
1.3	Objetivos	16
1.4	Organização	17
1.5	Suporte Computacional	17
2	O ENEM	18
2.1	Introdução	18
2.2	O Antigo Enem	19
2.3	O Novo Enem	21
2.4	Teoria de Resposta ao Item - TRI	27
2.5	Questionário Socioeconômico do Enem	28
3	ANÁLISE DE AGRUPAMENTOS	29
3.1	Introdução	29
3.2	Etapas da Análise de Agrupamentos	29
3.2.1	Introdução	29
3.2.2	Definição do Problema	30
3.2.3	Obtenção dos Dados	30
3.2.4	Tratamento dos Dados	31
3.2.5	Crítérios de Parecença (Semelhança ou Proximidade)	32
3.2.6	Aplicação da Técnica de Agrupamento	34
3.2.7	Apresentação dos Resultados	37
3.2.8	Interpretação dos Resultados	37
3.3	Medidas de Distância e Similaridade	38
3.3.1	Medidas de Similaridade e Dissimilaridade	38
3.3.2	Coeficientes de Parecença para Variáveis Quantitativas	40
3.3.2.1	Medidas Derivadas da Distância Euclidiana	40
3.3.2.2	Outros coeficientes	43
3.3.3	Coeficientes de Parecença para Variáveis Qualitativas Nominais	46
3.3.3.1	Coeficientes de Parecença para Variáveis Dicotômicas	46
3.4	Formando Agrupamentos	49
3.4.1	Introdução	49
3.4.2	Técnicas Hierárquicas de Agrupamento	50

3.4.2.1	Método da Centróide	50
3.4.2.2	Método da Ligação Simples ou do Vizinho mais Próximo. (M.L.S.)	53
3.4.2.3	Método da Ligação Completa ou do Vizinho mais Longe. (M.L.C.)	56
3.4.3	Métodos de Partição	59
3.4.3.1	Método das <i>k-means</i>	60
4	APLICAÇÃO	63
4.1	Objetivos	63
4.2	Caracterização da População	63
4.3	Seleção de Variáveis	63
4.4	Resultados	66
5	CONSIDERAÇÕES FINAIS	76
	REFERÊNCIAS	78
	ANEXOS	81
	ANEXO A –	82
	ANEXO B –	85

1 Introdução

1.1 Motivação e Justificativa

A influência dos fatores sociais no desempenho escolar foi confirmada pelo trabalho de (BOURDIEU; PASSERON, 1964). Nele, os autores afirmam que as famílias transmitem aos filhos, de maneira direta ou indireta, um conjunto de valores na forma de capital cultural. Sendo isso, uma hipótese imprescindível para explicar as disparidades no desempenho escolar entre estudantes de diferentes classes sociais.

No que tange ao desempenho dos estudantes no Exame Nacional do Ensino Médio (Enem), (JALOTO; PRIMI, 2021) identificaram que há associação entre o desempenho nas quatro provas do Enem 2018 e o atraso escolar, a dependência administrativa da escola do educando e o nível sócioeconômico dos alunos. Para (SIRIN, 2005) nível socioeconômico inclui como principais componentes a escolaridade dos pais, sua ocupação e a renda familiar. Além disso, (SANTANA, 2020) acentua que o grau de escolaridade dos pais ou dos responsáveis influencia o desempenho dos educandos na redação do Enem (2009-2018), (FEIJÓ; FRANÇA, 2021) destacam grandes diferenças entre alunos de escolas públicas e privadas. Para (ARISTIZABAL; CAICEDO; PARRA, 2017), (KARINO; LAROS, 2017), algumas variáveis tais como, escolaridade dos pais, atraso escolar, sexo, cor/raça e dependência administrativa se mostram estatisticamente importantes na predição do desempenho.

Segundo nossa Constituição Federal (BRASIL, 1988), a educação, direito de todos e dever do Estado e da família, visa o pleno desenvolvimento das pessoas, seu preparo para o exercício da cidadania e a qualificação profissional, tendo como um dos seus princípios a igualdade de condições para o acesso e permanência na escola. Neste trabalho, cabe-nos investigar quais Unidades Federativas (UFs) do nosso país oferecem condições similares aos estudantes que se inscrevem no Enem, levando em consideração as informações coletadas pelo questionário socioeconômico e disponíveis na base de Microdados 2021 do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep).¹

¹ O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) é um órgão federal vinculado ao Ministério da Educação, ele é responsável pelas evidências educacionais e atua em três esferas: avaliações e exames educacionais; pesquisas estatísticas e indicadores educacionais; e gestão do conhecimento e estudos.

1.2 Trabalhos Relacionados

Dentre os trabalhos que utilizam as técnicas da Estatística Multivariada para a análise de informações coletadas pelo questionário socioeconômico e presentes na base de Microdados do Enem destacamos (NASCIMENTO; CAVALCANTI; OSTERMANN, 2017) que recorreram a Análise de Correspondência para investigar explorativamente os Microdados do ENEM de 2014, levando em consideração variáveis tais como a dependência administrativa da escola na qual o estudante concluiu o Ensino Médio, a etnia autodeclarada pelo candidato, o desempenho obtido o exame; (LIMA et al., 2020) que investigaram o desempenho dos participantes do Enem de 2012 a 2017, originando agrupamentos de performance (baixa, mediana ou alta), relacionando-os com as 5 regiões geográficas brasileiras; e (MAIA; ANDRADE; FERNANDES, 2021) que utilizaram o método *k-means* para análise de características socioeconômicas de candidatos ao ensino superior inscritos no Enem 2018. Recentemente, (FAILLACE; BRITTO; COSTA, 2022) também aplicaram o método não-hierárquico *k-means* e identificaram seis grupos de participantes do Enem 2019, segundo o desempenho, relacionando-os com características socioeconômicas dos candidatos; (MELO et al., 2022) avaliaram o impacto das variáveis socioeconômicas no desempenho do Enem 2018, recorrendo a ferramentas da Estatística Espacial, e identificaram que as variáveis percentual de estudantes com bolsa, renda, raça, escolaridade e nível instrucional da mãe são fatores relevantes para o desempenho e a dispersão das notas dos estudantes de cada município.

1.3 Objetivos

Objetivo Geral:

- Investigar a existência de agrupamentos (*clusters*) de Unidades Federativas do Brasil, segundo características socioeconômicas e educacionais dos inscritos no Enem 2021.

Objetivos Específicos:

- Identificar variáveis e fixar critérios de parença que permitam mensurar a similaridade entre os objetos alvo do estudo;
- Implementar algoritmos hierárquicos e não hierárquicos da Análise de Agrupamentos;
- Obter uma estrutura de *clusters* de UFs do Brasil em que especifica-se o número de agrupamentos e determina-se os componentes de cada *clusters*.

1.4 Organização

A estrutura do trabalho será descrita abaixo em cinco capítulos.

Iniciamos o Capítulo 1 da introdução abordando a motivação e a justificativa para o desenvolvimento do trabalho. E em sequência, enfocamos o objetivo principal e os objetivos específicos, direcionando o leitor a entender melhor a nossa proposta.

O Capítulo 2 apresentamos alguns aspectos sobre o Enem, desde a sua criação em 1998, seus objetivos, as competências e habilidades, bem como tratamos sobre o Novo Enem a partir de 2009 com novos objetivos, competências e habilidades, além dos objetos de conhecimento, a divisão em quatro áreas de conhecimento e a criação de programas federais que visam selecionar alunos para ingressar em cursos superiores, como o PROUNI e SISU. Além de trazer uma seção relatando um pouco sobre o questionário socioeconômico, sua composição, objetivo e importância.

No Capítulo 3 apresentamos uma introdução à Análise de Agrupamentos, sua definição, as etapas da análise com exemplos para ilustrar, faremos uma alusão também às medidas de similaridade e dissimilaridade, finalizando o capítulo apresentando a teoria das técnicas hierárquicas e não-hierárquica exemplificando cada uma delas para uma melhor compreensão.

No Capítulo 4 é feita uma aplicação com uma seleção de variáveis escolhidas dentre o banco de dados do questionário socioeconômico do Enem 2021, onde usamos as técnicas hierárquicas e não-hierárquicas a fim de investigar a existência de *clusters* de Unidades Federativas do Brasil, segundo características socioeconômicas e educacionais, e em consequência disso é apresentado os resultados.

E no último Capítulo temos a conclusão do trabalho com as considerações finais e contribuições, como também algumas sugestões de trabalhos futuros. Além das referências bibliográficas para finalizar.

1.5 Suporte Computacional

Para a realização desta Análise de Agrupamentos foi necessário um suporte computacional, neste caso usamos o *Software R* versão 4.2.2, onde através dos comandos `hcluster` e `k-means` obtivemos as saídas na qual fizemos as devidas análises estatísticas por meio de dendogramas, tabelas e figuras.

2 O Enem

2.1 Introdução

No Brasil, o Enem é uma avaliação a nível nacional pelo qual pode-se avaliar a qualidade e aperfeiçoar os currículos do ensino médio. Ele tem sido a principal porta de acesso ao ensino superior no Brasil e foi criado em 1998 no governo de Fernando Henrique Cardoso com o objetivo de avaliar o desempenho dos estudantes ao final da educação básica, desde a sua criação o órgão responsável pela sua realização anual é o Inep. Mais especificamente, o exame foi criado por meio da Portaria do Ministério da Educação (MEC) nº 438, de 28 de maio de 1998, tendo os seguintes objetivos:

- I conferir ao cidadão parâmetro para auto-avaliação, com vistas à continuidade de sua formação e à sua inserção no mercado de trabalho;
- II criar referência nacional para os egressos de qualquer das modalidades do ensino médio;
- III fornecer subsídios às diferentes modalidades de acesso à educação superior;
- IV constituir-se em modalidade de acesso a cursos profissionalizantes pós-médio; (BRASIL, 1998).

De acordo com (BRASIL, 1998) o principal objetivo do Enem é “possibilitar uma referência para auto-avaliação, a partir das competências e habilidades que o estrutura”, nas quais são desenvolvidas e fortalecidas no ambiente escolar com a mediação do professor. Segundo Brasil(2015, p.63) para a criação do Enem tomou-se como referência outros documentos que regem a educação do nacional, tais como a Lei de Diretrizes e Bases da Educação Nacional 9.394/96 (LDB), os Parâmetros Curriculares Nacionais (PCNs), as Orientações Curriculares Nacionais para o Ensino Médio (OCNEM) e as Matrizes de Referência do Sistema de Avaliação da Educação Básica (SAEB).

Paralelo à realização do Enem, tem sido desenvolvidas políticas públicas com o objetivo de superar desigualdades educacionais existentes, dentre elas destaca-se a Lei de Cotas (BRASIL; INTERMEDIÁRIO, 2012). A Lei nº 12.711/2012, sancionada em agosto deste ano, garante a reserva de 50% das matrículas por curso e turno nas 59 universidades federais e 38 institutos federais de educação, ciência e tecnologia a alunos oriundos integralmente do ensino médio público, em cursos regulares ou da educação de jovens e adultos (BRAZ et al., 2022).

2.2 O Antigo Enem

A prova do Enem no período de 1998 à 2008 era composta de 63 itens (63 questões), sendo três itens por habilidade e uma redação que eram aplicadas em um único dia, nos quais os itens da prova eram elaborados a partir de uma Matriz de Referência que possuía 5 competências, assim como 21 habilidades, como podemos ver abaixo:

Competências (1998 à 2008)

I. Dominar a norma culta da Língua Portuguesa e fazer uso das linguagens matemática, artística e científica.

II. Construir e aplicar conceitos das várias áreas do conhecimento para a compreensão de fenômenos naturais, de processos histórico-geográficos, da produção tecnológica e das manifestações artísticas.

III. Selecionar, organizar, relacionar, interpretar dados e informações representados de diferentes formas, para tomar decisões e enfrentar situações-problema.

IV. Relacionar informações, representadas em diferentes formas, e conhecimentos disponíveis em situações concretas, para construir argumentação consistente.

V. Recorrer aos conhecimentos desenvolvidos na escola para elaboração de propostas de intervenção solidária na realidade, respeitando os valores humanos e considerando a diversidade sociocultural.

Habilidades (1998 à 2008)

1. Dada a descrição discursiva ou por ilustração de um experimento ou fenômeno, de natureza científica, tecnológica ou social, identificar variáveis relevantes e selecionar os instrumentos necessários para realização ou interpretação do mesmo.

2. Em um gráfico cartesiano de variável socioeconômica ou técnico-científica, identificar e analisar valores das variáveis, intervalos de crescimento ou decréscimo e taxas de variação.

3. Dada uma distribuição estatística de variável social, econômica, física, química ou biológica, traduzir e interpretar as informações disponíveis, ou reorganizá-las, objetivando interpolações ou extrapolações.

4. Dada uma situação-problema, apresentada em uma linguagem de determinada área de conhecimento, relacioná-la com sua formulação em outras linguagens ou vice-versa.

5. A partir da leitura de textos literários consagrados e de informações sobre concepções artísticas, estabelecer relações entre eles e seu contexto histórico, social, político ou cultural, inferindo as escolhas dos temas, gêneros discursivos e recursos expressivos dos autores.

6. Com base em um texto, analisar as funções da linguagem, identificar marcas de variantes linguísticas de natureza sociocultural, regional, de registro ou de estilo, e explorar as relações entre as linguagens coloquial e formal.

7. Identificar e caracterizar a conservação e as transformações de energia, em diferentes processos de sua geração e uso social, e comparar diferentes recursos e opções energéticas.

8. Analisar criticamente, de forma qualitativa ou quantitativa, as implicações ambientais, sociais e econômicas dos processos de utilização dos recursos naturais, materiais ou energéticos.

9. Compreender o significado e a importância da água e de seu ciclo para a manutenção da vida, em sua relação com condições socioambientais, sabendo quantificar variações de temperatura e mudanças de fase em processos naturais e de intervenção humana.

10. Utilizar e interpretar diferentes escalas de tempo para situar e descrever transformações na atmosfera, biosfera, hidrosfera e litosfera, origem e evolução da vida, variações populacionais e modificações no espaço geográfico.

11. Diante da diversidade da vida, analisar, do ponto de vista biológico, físico ou químico, padrões comuns nas estruturas e nos processos que garantem a continuidade e a evolução dos seres vivos.

12. Analisar fatores socioeconômicos e ambientais associados ao desenvolvimento e às condições de vida e saúde de populações humanas, por meio da interpretação de diferentes indicadores.

13. Compreender o caráter sistêmico do planeta e reconhecer a importância da biodiversidade para preservação da vida, relacionando condições do meio e intervenção humana.

14. Diante da diversidade de formas geométricas planas e espaciais, presentes na natureza ou imaginadas, caracterizá-las por meio de propriedades, relacionar seus elementos, calcular comprimentos, áreas ou volumes e utilizar o conhecimento geométrico para leitura, compreensão e ação sobre a realidade.

15. Reconhecer o caráter aleatório de fenômenos naturais ou não e utilizar em situações-problema processos de contagem, representação de frequências relativas, construção de espaços amostrais, distribuição e cálculo de probabilidades.

16. Analisar, de forma qualitativa ou quantitativa, situações-problema referentes a perturbações ambientais, identificando fonte, transporte e destino dos poluentes, reconhecendo suas transformações; prever efeitos nos ecossistemas e no sistema produtivo e propor formas de intervenção para reduzir e controlar os efeitos da poluição ambiental.

17. Na obtenção e produção de materiais e de insumos energéticos, identificar etapas, calcular rendimentos, taxas e índices e analisar implicações sociais, econômicas

e ambientais.

18. Valorizar a diversidade dos patrimônios etnoculturais e artísticos, identificando-a em suas manifestações e representações em diferentes sociedades, épocas e lugares.

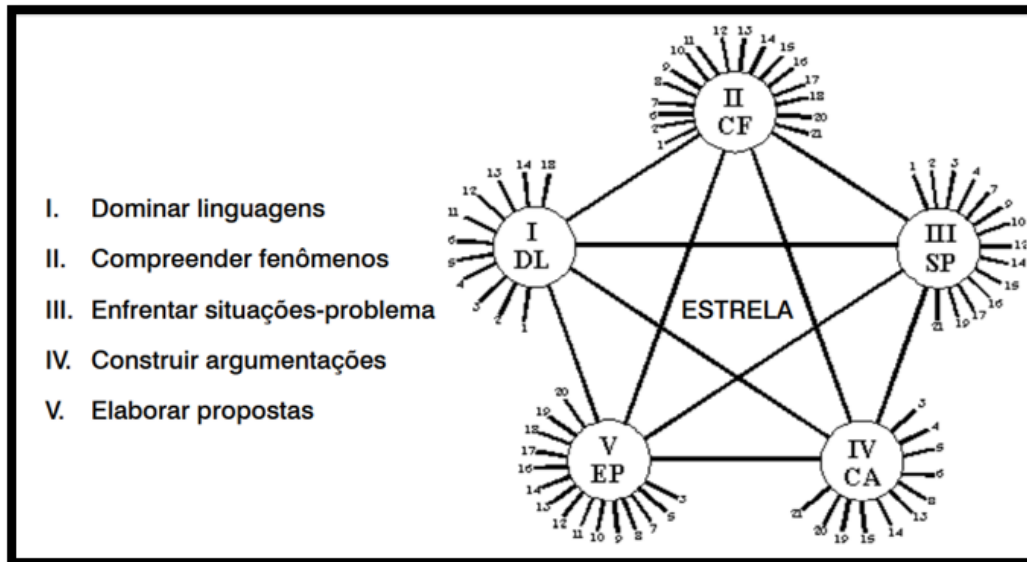
19. Confrontar interpretações diversas de situações ou fatos de natureza histórico-geográfica, técnico-científica, artístico-cultural ou do cotidiano, comparando diferentes pontos de vista, identificando os pressupostos de cada interpretação e analisando a validade dos argumentos utilizados.

20. Comparar processos de formação socioeconômica, relacionando-os com seu contexto histórico e geográfico.

21. Dado um conjunto de informações sobre uma realidade histórico-geográfica, contextualizar e ordenar os eventos registrados, compreendendo a importância dos fatores sociais, econômicos, políticos ou culturais.

As cinco competências e as vinte e uma habilidades citadas funcionam de forma orgânica e integrada entre si, em que cada competência possui suas habilidades específicas, estabelecendo assim um conjunto de interconexões, como podemos ver abaixo na Figura 1:

Figura 1 – Interconexões entre competências e habilidades.



Fonte: Relatório pedagógico Enem 2008.

2.3 O Novo Enem

No ano de 2004 foi criado o Programa Universidade para Todos (PROUNI) cujo objetivo é subsidiar alunos com bolsas de estudos parciais ou integrais selecionados

através da nota obtida no Enem para ingressar em cursos superiores. Oficialmente a partir de 2009 o exame passa a ser usado como forma de ingresso em instituições de educação de ensino superior. Após essa mudança em 2009 foi criado também o Sistema de Seleção Unificado (SISU) o qual foi criado a fim de selecionar alunos para cursos superiores de instituições federais e estaduais, tomando como base a nota obtida no Enem.

Por meio da Portaria MEC/INEP nº 109, de 27 de maio de 2009 o Enem passa por uma reformulação, a prova passa a ter 180 (cento e oitenta) itens sendo aplicada em dois dias, com 45 (quarenta e cinco) itens de cada área do conhecimento, além da redação no primeiro dia, assim como também foi inserido cinco itens sobre a disciplina de Língua Inglesa ou Espanhola, sendo o aluno responsável pela escolha da língua estrangeira no ato da inscrição. Nessa portaria fica estabelecido a sistemática para a realização do ENEM no exercício de 2009 e denomina como sendo “procedimento de avaliação do desempenho escolar e acadêmico dos participantes, para aferir o desenvolvimento das competências e habilidades fundamentais ao exercício da cidadania.” Nessa reformulação o Enem passa a ter outros objetivos, que são:

I - oferecer uma referência para que cada cidadão possa proceder à sua autoavaliação com vistas às suas escolhas futuras, tanto em relação ao mundo do trabalho quanto em relação à continuidade de estudos;

II - estruturar uma avaliação ao final da educação básica que sirva como modalidade alternativa ou complementar aos processos de seleção nos diferentes setores do mundo do trabalho;

III - estruturar uma avaliação ao final da educação básica que sirva como modalidade alternativa ou complementar aos exames de acesso aos cursos profissionalizantes, pós-médios e à Educação Superior;

IV - possibilitar a participação e criar condições de acesso a programas governamentais;

V - promover a certificação de jovens e adultos no nível de conclusão do ensino médio nos termos do artigo 38, §§ 1º e 2º da Lei nº 9.394/96 - Lei das Diretrizes e Bases da Educação Nacional (LDB);

VI - promover avaliação do desempenho acadêmico das escolas de ensino médio, de forma que cada unidade escolar receba o resultado global;

VII - promover avaliação do desempenho acadêmico dos estudantes ingressantes nas Instituições de Educação Superior;

A nova Matriz de Referência Enem foi criada a partir da matriz de competências e habilidades do Exame Nacional para Certificação de Competências de Jovens e Adultos (ENCCEJA) e da antiga matriz de referência do Enem, em que passa a ser dividida em quatro áreas de conhecimento e essas áreas abrangem os componentes curriculares

do ensino médio, conforme a Tabela 1:

Tabela 1 – Áreas de conhecimento e suas respectivas disciplinas.

Áreas de conhecimento	Disciplinas
Linguagens, códigos e suas Tecnologias	Língua Portuguesa, Literatura, Língua Estrangeira(Inglês ou Espanhol), Artes, Educação Física e Tecnologias da Informação e Comunicação.
Matemática e suas Tecnologias	Matemática.
Ciências da Natureza e suas Tecnologias	Química, Física e Biologia.
Ciências Humanas e suas Tecnologias	História, Geografia, Filosofia e Sociologia.

Fonte: Próprio autor

A nova Matriz de Referência do Enem é um documento composto de eixos cognitivos, competências, habilidades e objetos de conhecimento, no qual os eixos cognitivos, são comuns a todas as áreas do conhecimento, os demais são específicos de cada área e é por meio desse conjunto que são elaborados os itens da prova do Enem. São cinco os eixos cognitivos:

I. Dominar linguagens (DL): dominar a norma culta da Língua Portuguesa e fazer uso das linguagens matemática, artística e científica e das línguas espanhola e inglesa.

II. Compreender fenômenos (CF): construir e aplicar conceitos das várias áreas do conhecimento para a compreensão de fenômenos naturais, de processos histórico geográficos, da produção tecnológica e das manifestações artísticas.

III. Enfrentar situações-problema (SP): selecionar, organizar, relacionar, interpretar dados e informações representados de diferentes formas, para tomar decisões e enfrentar situações-problema.

IV. Construir argumentação (CA): relacionar informações, representadas em diferentes formas, e conhecimentos disponíveis em situações concretas, para construir argumentação consistente.

V. Elaborar propostas (EP): recorrer aos conhecimentos desenvolvidos na escola para elaboração de propostas de intervenção solidária na realidade, respeitando os valores humanos e considerando a diversidade sociocultural.

Além dos 5 (cinco) eixos, foram implementados 32 (trinta e duas) competências e 120 (cento e vinte) habilidades na Matriz de Referência do Enem, especificadamente na Matriz de Matemática são ao total, 7 (sete) competências e 30 (habilidades).

Competências da Matriz de Referência de Matemática e suas tecnologias

Competência de área 1 - Construir significados para os números naturais, inteiros, racionais e reais.

H1 - Reconhecer, no contexto social, diferentes significados e representações dos números e operações - naturais, inteiros, racionais ou reais.

H2 - Identificar padrões numéricos ou princípios de contagem.

H3 - Resolver situação-problema envolvendo conhecimentos numéricos.

H4 - Avaliar a razoabilidade de um resultado numérico na construção de argumentos sobre afirmações quantitativas.

H5 - Avaliar propostas de intervenção na realidade utilizando conhecimentos numéricos.

Competência de área 2 - Utilizar o conhecimento geométrico para realizar a leitura e a representação da realidade e agir sobre ela.

H6 - Interpretar a localização e a movimentação de pessoas/objetos no espaço tridimensional e sua representação no espaço bidimensional.

H7 - Identificar características de figuras planas ou espaciais.

H8 - Resolver situação-problema que envolva conhecimentos geométricos de espaço e forma.

H9 - Utilizar conhecimentos geométricos de espaço e forma na seleção de argumentos propostos como solução de problemas do cotidiano.

Competência de área 3 - Construir noções de grandezas e medidas para a compreensão da realidade e a solução de problemas do cotidiano.

H10 - Identificar relações entre grandezas e unidades de medida.

H11 - Utilizar a noção de escalas na leitura de representação de situação do cotidiano.

H12 - Resolver situação-problema que envolva medidas de grandezas.

H13 - Avaliar o resultado de uma medição na construção de um argumento consistente.

H14 - Avaliar proposta de intervenção na realidade utilizando conhecimentos geométricos relacionados a grandezas e medidas.

Competência de área 4 - Construir noções de variação de grandezas para a compreensão da realidade e a solução de problemas do cotidiano.

H15 - Identificar a relação de dependência entre grandezas.

H16 - Resolver situação-problema envolvendo a variação de grandezas, direta ou inversamente proporcionais.

H17 - Analisar informações envolvendo a variação de grandezas como recurso para a construção de argumentação.

H18 - Avaliar propostas de intervenção na realidade envolvendo variação de grandezas.

Competência de área 5 - Modelar e resolver problemas que envolvem variáveis socioeconômicas ou técnico-científicas, usando representações algébricas.

H19 - Identificar representações algébricas que expressem a relação entre grandezas.

H20 - Interpretar gráfico cartesiano que represente relações entre grandezas.

H21 - Resolver situação-problema cuja modelagem envolva conhecimentos algébricos.

H22 - Utilizar conhecimentos algébricos/geométricos como recurso para a construção de argumentação.

H23 - Avaliar propostas de intervenção na realidade utilizando conhecimentos algébricos.

Competência de área 6 - Interpretar informações de natureza científica e social obtidas da leitura de gráficos e tabelas, realizando previsão de tendência, extrapolação, interpolação e interpretação.

H24 - Utilizar informações expressas em gráficos ou tabelas para fazer inferências.

H25 - Resolver problema com dados apresentados em tabelas ou gráficos.

H26 - Analisar informações expressas em gráficos ou tabelas como recurso para a construção de argumentos.

Competência de área 7 - Compreender o caráter aleatório e não-determinístico dos fenômenos naturais e sociais e utilizar instrumentos adequados para medidas, determinação de amostras e cálculos de probabilidade para interpretar informações de variáveis apresentadas em uma distribuição estatística.

H27 - Calcular medidas de tendência central ou de dispersão de um conjunto de dados expressos em uma tabela de frequências de dados agrupados (não em classes) ou em gráficos.

H28 - Resolver situação-problema que envolva conhecimentos de estatística e probabilidade.

H29 - Utilizar conhecimentos de estatística e probabilidade como recurso para a construção de argumentação.

H30 - Avaliar propostas de intervenção na realidade utilizando conhecimentos de estatística e probabilidade.

Os objetos de conhecimentos associados a Matriz de Referência da área de Matemática e suas Tecnologias são 5 (cinco):

Conhecimentos numéricos: operações em conjuntos numéricos (naturais, inteiros, racionais e reais), desigualdades, divisibilidade, fatoração, razões e proporções, porcentagem e juros, relações de dependência entre grandezas, sequências e progressões, princípios de contagem.

Conhecimentos geométricos: características das figuras geométricas planas e espaciais; grandezas, unidades de medida e escalas; comprimentos, áreas e volumes; ângulos; posições de retas; simetrias de figuras planas ou espaciais; congruência e semelhança de triângulos; teorema de Tales; relações métricas nos triângulos; circunferências; trigonometria do ângulo agudo.

Conhecimentos de estatística e probabilidade: representação e análise de dados; medidas de tendência central (médias, moda e mediana); desvios e variância; noções de probabilidade.

Conhecimentos algébricos: gráficos e funções; funções algébricas do 1.^o e do 2.^o grau, polinomiais, racionais, exponenciais e logarítmicas; equações e inequações; relações no ciclo trigonométrico e funções trigonométricas.

Conhecimentos algébricos/geométricos: plano cartesiano; retas; circunferências; paralelismo e perpendicularidade, sistemas de equações.

Segundo (RABELO, 2013) há uma relação tridimensional entre os eixos cognitivos, “que são as ações e operações mentais que todos os jovens e adultos devem desenvolver como recursos mínimos que os habilitam a enfrentar melhor o mundo que os cercam,

com todas as suas responsabilidades”, e as competências e habilidades da área de Matemática e suas Tecnologias, como podemos ver na Tabela 2:

Tabela 2 – Relação entre competências, habilidades e eixos cognitivos da área de Matemática e suas tecnologias.

Competências de área	DL	CF	SP	CA	EP
C1	H1	H2	H3	H4	H5
C2	H6	H7	H8	H9	-
C3	H10	H11	H12	H13	H14
C4	-	H15	H16	H17	H18
C5	H19	H20	H21	H22	H23
C6	-	-	H24	H25	H26
C7	-	H27	H28	H29	H30

Fonte: (RABELO, 2013).

Para a correção das provas do novo Enem foi adotado a Teoria de Resposta ao Item (TRI), em que a nota do candidato é baseada no número de acertos de itens, que são classificados como fáceis, médias e difíceis e também de acordo com as competências e habilidades estabelecidas para cada item. “A estimação da proficiência está relacionada ao número de acertos, aos parâmetros dos itens e ao padrão de respostas. Apesar de não ser simples e exigir estimativas dos parâmetros realizada por métodos estatísticos avançados, o cálculo da proficiência é objetivo, e participantes com exatamente o mesmo padrão de respostas apresentam exatamente as mesmas proficiências”(KARINO; BARBOSA, 2012).

2.4 Teoria de Resposta ao Item - TRI

Segundo (BRASIL, 2021) a TRI é um conjunto de modelos matemáticos que busca representar a relação entre a probabilidade de o participante responder corretamente a uma questão, seu conhecimento na área em que está sendo avaliado e as características (parâmetros) dos itens. O Enem considera três parâmetros para caracterizar cada item (questão) que são: parâmetro de discriminação, parâmetro de dificuldade e o parâmetro de acerto casual.

- a) parâmetro de discriminação: avalia o desempenho de cada candidato na habilidade que compõe cada item, diferenciando os candidatos que dominam e os que não dominam a habilidade naquele item;

- b) parâmetro de dificuldade: associa a dificuldade da habilidade avaliada na questão, assim quanto maior seu valor, mais difícil é a questão. Dessa maneira os itens do Enem são distribuídos em diferentes níveis;
- c) parâmetro de acerto casual: representa a probabilidade do candidato acertar um item ao acaso não tendo domínio na habilidade exigida;

2.5 Questionário Socioeconômico do Enem

O questionário socioeconômico é um instrumento importante que todo estudante que se candidata ao Enem tem que preencher pelo qual o mesmo informa suas condições socioeconômicas, tais como, os aspectos sociais, econômicos e raciais.

No questionário socioeconômico do Enem podemos encontrar várias informações como, por exemplo, o município, o estado, a idade, o sexo, o tipo de escola em que o candidato frequentou durante os anos de escolarização no Ensino Médio, informações sobre a cor/raça, ou se o participante apresenta alguma necessidade educacional especial, ou deficiência. Assim como, informações sobre a escolaridade dos pais, sua ocupação, a renda familiar, a composição familiar (quantidade de habitantes que residem na casa, incluindo o candidato), e até mesmo a presença ou ausência de alguns eletrodomésticos como TVs, Microcomputadores e acesso a internet.

De acordo com (SANTANA, 2020) “o grau de escolaridade dos pais ou dos responsáveis influencia no desempenho dos candidatos” e “esta investigação tem como referencial a teoria sociológica de Bourdieu”. Assim, tem-se que quanto maior a escolaridade dos pais, maior é a nota na redação do Enem. É em consonância com esta citação que vemos a importância desse questionário socioeconômico, no qual quando preenchido com informações verídicas traz muitas informações importantes a respeito desses inscritos.

Em muitos casos, o preenchimento deste questionário implica em benefícios para si mesmo, como por exemplo a isenção na taxa de pagamento, além de proporcionar ao inscrito o direito a participar do sistema de cotas, ressaltando que o preenchimento do questionário não interfere na nota do inscrito.

Este estudo foi conduzido por meio do banco de microdados disponível no site do (INEP, 2021) especificadamente pelo questionário socioeconômico que os estudantes preencheram no momento da inscrição do Enem 2021. O órgão responsável pela obtenção e disponibilização dos microdados é o Instituto Nacional de Pesquisas Educacionais (INEP).

3 Análise de Agrupamentos

3.1 Introdução

A Análise de Agrupamentos (A.A.), também conhecida como Análise de Conglomerados, tem como objetivo aglomerar ou particionar os elementos de uma amostra, ou população, em grupos (*clusters*) de forma que os elementos pertencentes a um mesmo grupo sejam similares entre si com respeito às variáveis (características) que neles foram medidas, e os elementos em grupos diferentes sejam heterogêneos em relação a estas mesmas características (MINGOTI, 2013). A exposição teórica contida neste capítulo é um resumo daquela que consta em (BUSSAB; MIAZAKI; ANDRADE, 1990), mas com exemplos diferentes. Caso o leitor tenha interesse em referências bibliográficas complementares que, além da Análise de Agrupamentos, abordem outras ferramentas da Estatística Multivariada, citemos (JOHNSON; WICHERN et al., 2002) e (RENCHEER; CHRISTENSEN, 2012).

3.2 Etapas da Análise de Agrupamentos

3.2.1 Introdução

Neste tópico iremos ilustrar as sete etapas procedimentais de uma A.A., ressaltamos que cada etapa depende da anterior.

- (i) Definição de objetivos, critérios, escolha de variáveis e objetos;
- (ii) Obtenção dos dados;
- (iii) Tratamento dos dados;
- (iv) Escolha de critérios de similaridade ou dissimilaridade (parecença);
- (v) Adoção e execução de um algoritmo de A.A.;
- (vi) Apresentação dos resultados;
- (vii) Avaliação e interpretação dos resultados.

Com o desenvolvimento das etapas citadas acima pretende-se dar ao leitor um roteiro para que ele possa implementar a técnica de A.A.. Para um melhor entendimento,

ilustraremos por meio de um exemplo as principais ações necessárias à implementação de uma A.A..

3.2.2 Definição do Problema

Pretende-se investigar, exploratoriamente, a existência de similaridades entre as provas de Matemática do Enem (provas regulares e reaplicações) dos anos de 2019, 2020 e 2021, segundo características dos itens contidos em cada avaliação. Sabe-se que a cada item contido nas provas há três parâmetros característicos: o parâmetro de discriminação, que é o poder de discriminação do item para diferenciar os participantes que dominam daqueles que não dominam a habilidade avaliada; o parâmetro de dificuldade que está associado à dificuldade do item, sendo que quanto maior seu valor, mais difícil é o item; e o parâmetro de acerto ao acaso que é a probabilidade de um participante acertar o item não dominando a habilidade exigida. Sendo assim, consideramos as seguintes variáveis:

- Média aritmética dos parâmetros de discriminação associados aos itens (X_1);
- Média aritmética dos parâmetros de dificuldade associados aos itens (X_2);
- Média aritmética dos parâmetros de acerto ao acaso associados aos itens (X_3).

3.2.3 Obtenção dos Dados

Os dados obtidos referem-se as três variáveis que caracterizam as provas do Enem na área de Matemática, das edições de 2019, 2020 e 2021 da Prova Regular (A_{19} , C_{20} e E_{21}) e da Reaplicação (B_{19} , D_{20} e F_{21}), como vemos na Tabela 3.

Tabela 3 – Provas do Enem (Matemática), segundo as variáveis X_1 , X_2 e X_3 .

PROVAS	X_1	X_2	X_3
A_{19}	240.127	188.607	0.150
B_{19}	225.300	187.260	0.159
C_{20}	203.002	149.414	0.160
D_{20}	167.010	142.972	0.164
E_{21}	182.835	170.745	0.152
F_{21}	174.715	135.916	0.175

Fonte: Próprio autor.

A matriz de dados indica os valores das características por objetos de interesse, como podemos ver na Figura 2. E os dados podem ser brutos ou relativos (padronizados).

Figura 2 – Dados brutos e dados padronizados.

(a) Brutos.

(b) Relativos(Padronizados).

$$X = \begin{matrix} & X_1 & X_2 & \dots & X_p \\ \begin{matrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{matrix} & \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix} \end{matrix} \quad Z = \begin{pmatrix} Z_{11} & Z_{12} & \dots & Z_{1p} \\ Z_{21} & Z_{22} & \dots & Z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n1} & Z_{n2} & \dots & Z_{np} \end{pmatrix}$$

Fonte: (BUSSAB; MIAZAKI; ANDRADE, 1990).

Geralmente agrupamos objetos semelhantes de acordo com suas características (variáveis), mas isso não impede que possamos agrupar as variáveis segundo os valores obtidos pelos objetos.

3.2.4 Tratamento dos Dados

Os objetos que serão agrupados são as provas segundo as variáveis X_1 , X_2 e X_3 . Logo, iremos fazer a padronização estatística das três grandezas, conforme a fórmula:

$$Z = \frac{x - \bar{x}}{\sigma}, \tag{3.1}$$

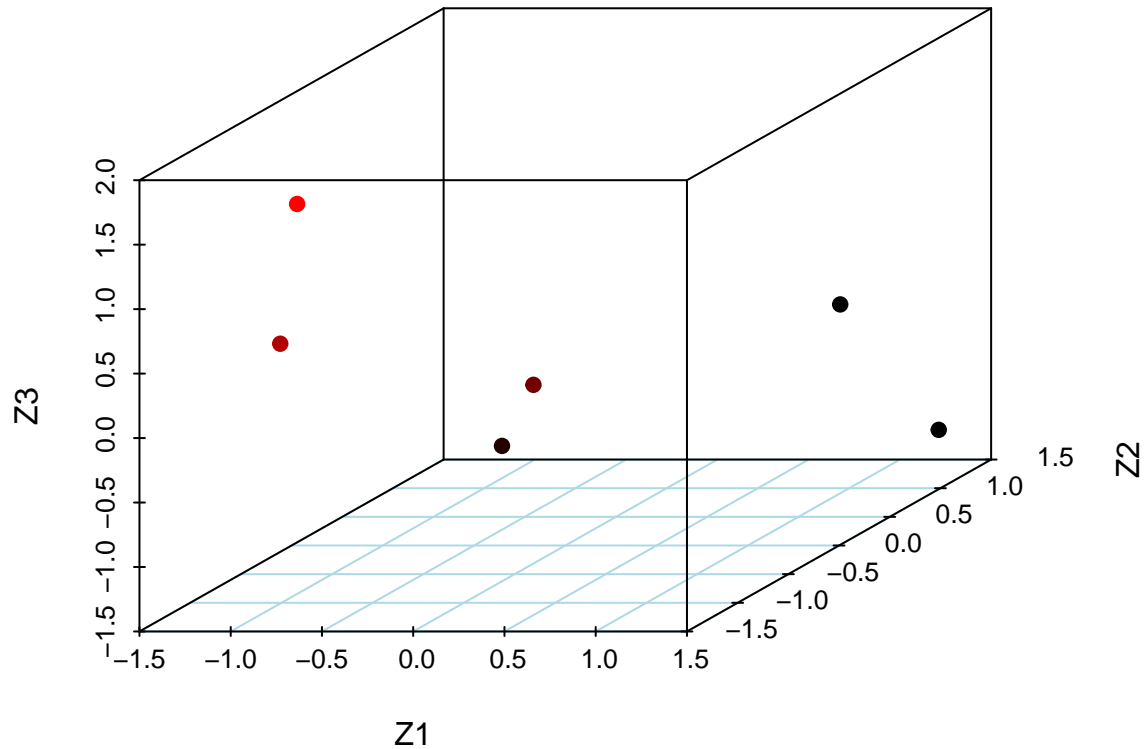
onde x é o valor de cada observação, \bar{x} é a média das observações e σ é o desvio-padrão das observações.

Tabela 4 – Construção dos Valores Padronizados das Variáveis X_1 , X_2 e X_3 .

PROVAS	X_1	X_2	X_3	Z_1	Z_2	Z_3
A ₁₉	240.127	188.607	0.150	1.413	1.141	-1.110
B ₁₉	225.300	187.260	0.159	0.905	1.082	-0.111
C ₂₀	203.002	149.414	0.160	0.143	-0.571	0.000
D ₂₀	167.010	142.972	0.164	-1.088	-0.852	0.444
E ₂₁	182.835	170.745	0.152	-0.547	0.361	-0.889
F ₂₁	174.715	135.916	0.175	-0.825	-1.160	1.665
MÉDIA	198.831	162.486	0.160	0.0	0.0	0.0
D.PADRÃO	29.235	22.899	0.009	1.0	1.0	1.0

Fonte: Próprio autor.

Figura 3 – Representação Gráfica em 3D das provas, segundo os valores padronizados das variáveis X_1 , X_2 e X_3



Fonte: Próprio autor.

Na Tabela 4 temos os valores padronizados das três variáveis, assim como a média aritmética e o desvio-padrão. E na Figura 3 temos um gráfico em 3D com as três variáveis, onde as três provas regulares e as três provas da reavaliação estão representadas por pontos.

3.2.5 Critérios de Parecência (Semelhança ou Proximidade)

Com o objetivo de verificar qual objeto é mais semelhante ou parecido com outro utilizaremos o conceito natural de distância, utilizando a distância euclidiana. Dados dois objetos J e K , caracterizados pelas grandezas Z_1 , Z_2 e Z_3 , a distância euclidiana entre J e K é definida por:

$$d = d(J, K) = [(z_1(J) - z_1(K))^2 + (z_2(J) - z_2(K))^2 + (z_3(J) - z_3(K))^2]^{\frac{1}{2}}, \quad (3.2)$$

onde $z_i(\cdot)$ denota o valor da variável Z_i para o objeto indicado.

A partir da Tabela 4 calcularemos a distância euclidiana entre algumas provas, a fim de quantificar essa proximidade. Por exemplo, a distância entre A_{19} e B_{19} :

$$\begin{aligned} d &= [(1.413 - 0.905)^2 + (1.141 - 1.082)^2 + (-1.110 + 0.111)^2]^{\frac{1}{2}} \\ d &= [(0.508)^2 + (0.059)^2 + (-0.999)^2]^{\frac{1}{2}} = 1.122. \end{aligned}$$

A distância entre C_{20} e D_{20} :

$$\begin{aligned} d &= [(0.143 + 1.088)^2 + (-0.571 + 0.852)^2 + (0.000 - 0.444)^2]^{\frac{1}{2}} \\ d &= [(1.231)^2 + (0.281)^2 + (-0.444)^2]^{\frac{1}{2}} = 1.339. \end{aligned}$$

E a distância entre C_{20} e E_{21} é:

$$\begin{aligned} d &= [(0.143 + 0.547)^2 + (-0.571 - 0.361)^2 + (0.000 + 0.889)^2]^{\frac{1}{2}} \\ d &= [(0.69)^2 + (0.932)^2 + (0.889)^2]^{\frac{1}{2}} = 1.460. \end{aligned}$$

Assim é possível construir a Matriz de Parecença ou Similaridade \mathbf{D} , dada por:

Matriz \mathbf{D} , usando a distância euclidiana:

$$D = \begin{pmatrix} & A_{19} & B_{19} & C_{20} & D_{20} & E_{21} & F_{21} \\ A_{19} & 0.000 & -- & -- & -- & -- & -- \\ B_{19} & 1.122 & 0.000 & -- & -- & -- & -- \\ C_{20} & 2.403 & 1.824 & 0.000 & -- & -- & -- \\ D_{20} & 3.555 & 2.833 & 1.339 & 0.000 & -- & -- \\ E_{21} & 2.121 & 1.798 & 1.460 & 1.881 & 0.000 & -- \\ F_{21} & 4.242 & 3.343 & 2.014 & 1.286 & 2.984 & 0.000 \end{pmatrix}$$

Em conformidade com (BUSSAB; MIAZAKI; ANDRADE, 1990), ao longo do nosso texto faremos uso da seguinte generalização da distância euclidiana para o cálculo da distância entre dois objetos:

$$d(A, B) = \left[\frac{\sum_{i=1}^p (z_i(A) - z_i(B))^2}{p} \right]^{\frac{1}{2}}, \quad (3.3)$$

que, por razões didáticas, chamaremos de distância euclidiana padronizada ou reduzida. Na definição da distância padronizada, p denota a dimensão do espaço das variáveis.

Matriz \mathbf{D} , usando a distância euclidiana padronizada:

$$D = \begin{pmatrix} & A_{19} & B_{19} & C_{20} & D_{20} & E_{21} \\ B_{19} & 0.648 & -- & -- & -- & -- \\ C_{20} & 1.387 & 1.053 & -- & -- & -- \\ D_{20} & 2.053 & 1.635 & 0.773 & -- & -- \\ E_{21} & 1.224 & 1.038 & 0.843 & 1.086 & -- \\ F_{21} & 2.449 & 1.930 & 1.163 & 0.743 & 1.723 \end{pmatrix}$$

3.2.6 Aplicação da Técnica de Agrupamento

Por meio do Exemplo Básico implementaremos o algoritmo do Método das Médias das Distâncias (M.M.D.), método aglomerativo hierárquico, em que a dimensão da matriz de parença da junção de pares semelhantes é reduzida a cada passo até o último passo que é a união de todos os pontos em um único grupo.

Mostraremos os seguintes passos para a aplicação do (M.M.D.).

(a) Passo 0

No Passo 0, iniciamos com a matriz de distância Reduzida que contém seis grupos (A_{19} , B_{19} , C_{20} , D_{20} , E_{21} e F_{21}).

Tabela 5 – M.M.D. Passo 0.

	A_{19}	B_{19}	C_{20}	D_{20}	E_{21}
B_{19}	0.648	-	-	-	-
C_{20}	1.387	1.053	-	-	-
D_{20}	2.053	1.635	0.773	-	-
E_{21}	1.224	1.038	0.843	1.086	-
F_{21}	2.449	1.930	1.163	0.743	1.723

Fonte: Próprio autor.

(b) Passo 1. Agrupar A_{19} com B_{19}

No Passo 0 verificamos que a menor distância é 0,648, entre A_{19} e B_{19} , logo iremos agrupar A_{19} com B_{19} . Para agrupar A_{19} com B_{19} é necessário calcular a distância entre $A_{19}B_{19}$ e os demais pontos C_{20} , D_{20} , E_{21} e F_{21} , originando uma nova matriz de similaridade. O M.M.D define a distância entre dois grupos como sendo a média entre os valores individuais dos objetos de um dos grupos com os do outro. Assim:

$$d(C_{20}, A_{19}B_{19}) = (d(C_{20}, A_{19}) + d(C_{20}, B_{19}))/2 = (1.387 + 1.053)/2 = 1.220,$$

$$d(D_{20}, A_{19}B_{19}) = (d(D_{20}, A_{19}) + d(D_{20}, B_{19}))/2 = (2.053 + 1.635)/2 = 1.844,$$

$$d(E_{21}, A_{19}B_{19}) = (d(E_{21}, A_{19}) + d(E_{21}, B_{19}))/2 = (1.224 + 1.038)/2 = 1.131,$$

$$d(F_{21}, A_{19}B_{19}) = (d(F_{21}, A_{19}) + d(F_{21}, B_{19}))/2 = (2.449 + 1.930)/2 = 2.189.$$

Logo, reduziremos a matriz de similaridade conforme Tabela 6:

Tabela 6 – M.M.D. Passo 1.

	C_{20}	D_{20}	E_{21}	F_{21}
D_{20}	0.773	-	-	-
E_{21}	0.843	1.086	-	-
F_{21}	1.163	0.743	1.723	-
$A_{19}B_{19}$	1.220	1.844	1.131	2.189

Fonte: Próprio autor.

(c) Passo 2. Agrupar D_{20} com F_{21}

No Passo 2 agruparemos D_{20} com F_{21} ao nível de 0.743, e determinaremos a distância entre $D_{20}F_{21}$ com os demais pontos C_{20} , E_{21} e $A_{19}B_{19}$ obtendo uma nova matriz de similaridade, então temos que:

$$d(C_{20}, D_{20}F_{21}) = (d(C_{20}, D_{20}) + d(C_{20}, F_{21}))/2 = (0.773 + 1.163)/2 = 0.968$$

$$d(E_{21}, D_{20}F_{21}) = (d(E_{21}, D_{20}) + d(E_{21}, F_{21}))/2 = (1.086 + 1.723)/2 = 1.404$$

$$\begin{aligned} d(A_{19}B_{19}, D_{20}F_{21}) &= (d(A_{19}, D_{20}) + d(A_{19}, F_{21}) + d(B_{19}, D_{20}) + d(B_{19}, F_{21}))/4 = \\ &= (2.053 + 2.449 + 1.635 + 1.930)/4 = 2.017 \end{aligned}$$

Assim, pela Tabela 7, temos uma nova matriz de similaridade:

Tabela 7 – M.M.D. Passo 2.

	C_{20}	E_{21}	$A_{19}B_{19}$
E_{21}	0.843	-	-
$A_{19}B_{19}$	1.220	1.131	-
$D_{20}F_{21}$	0.968	1.404	2.017

Fonte: Próprio autor.

(d) Passo 3. Agrupar C_{20} com E_{21}

No Passo 3 com uma nova matriz de similaridade em que agrupou-se D_{20} com F_{21} , teremos que reunir C_{20} com E_{21} ao nível de 0,843 de similaridade. Assim obtemos os grupos $A_{19}B_{19}$, $D_{20}F_{21}$ e $C_{20}E_{21}$, calculando as distâncias necessárias temos que:

$$d(A_{19}B_{19}, C_{20}E_{21}) = (d(A_{19}, C_{20}) + d(A_{19}, E_{21}) + d(B_{19}, C_{20}) + d(B_{19}, E_{21}))/4 = (1.387 + 1.224 + 1.053 + 1.038)/4 = 1.176$$

$$d(D_{20}F_{21}, C_{20}E_{21}) = (d(D_{20}, C_{20}) + d(D_{20}, E_{21}) + d(F_{21}, C_{20}) + d(F_{21}, E_{21}))/4 = (0.773 + 1.086 + 1.163 + 1.723)/4 = 1.186$$

Logo, pela Tabela 8 temos uma nova matriz de similaridade:

Tabela 8 – M.M.D. Passo 3.

	$C_{20}E_{21}$	$A_{19}B_{19}$
$A_{19}B_{19}$	1.176	-
$D_{20}F_{21}$	1.186	2.017

Fonte: Próprio autor.

(E) Passo 4. Agrupar $C_{20}E_{21}$ com $A_{19}B_{19}$

No quarto passo iremos reunir os grupos $C_{20}E_{21}$ com $A_{19}B_{19}$, a um nível de 1.175 de similaridade. Assim, teremos a distância entre os dois grupos $D_{20}F_{21}$ e $C_{20}E_{21}A_{19}B_{19}$, dada por:

$$d(D_{20}F_{21}, C_{20}E_{21}A_{19}B_{19}) = (d(D_{20}, C_{20}) + d(D_{20}, E_{21}) + d(D_{20}, A_{19}) + d(D_{20}, B_{19}) + d(F_{21}, C_{20}) + d(F_{21}, E_{21}) + d(F_{21}, A_{19}) + d(F_{21}, B_{19}))/8 = (0.773 + 1.086 + 2.053 + 1.635 + 1.163 + 1.723 + 2.449 + 1.930)/8 = 1.601$$

Tabela 9 – M.M.D. Passo 4.

	$C_{20}E_{21}A_{19}B_{19}$
$D_{20}F_{21}$	1.601

Fonte: Próprio autor.

(F) Passo 5

No último passo concluímos o processo unificando os grupos $D_{20}F_{21}$ e $C_{20}E_{21}A_{19}B_{19}$, que são similares a um nível 1.601 de parença.

3.2.7 Apresentação dos Resultados

Com os passos descritos acima é possível entender a técnica de agrupar, no entanto quanto a interpretação desses dados é mais complexo e necessita de instrumentos mais apropriados para essa interpretação, para isso temos o Resumo do M.M.D. na Tabela 10 que mostra as junções de cada passo e seu respectivo nível de semelhança.

Tabela 10 – Resumo do M.M.D. aplicado ao Exemplo Básico.

PASSO	JUNÇÃO	NÍVEL
1	A_{19}, B_{19}	0.648
2	D_{20}, F_{21}	0.743
3	C_{20}, E_{21}	0.843
4	$C_{20}E_{21}, A_{19}B_{19}$	1.176
5	$D_{20}F_{21}, C_{20}E_{21}A_{19}B_{19}$	1.601

Fonte: Próprio autor.

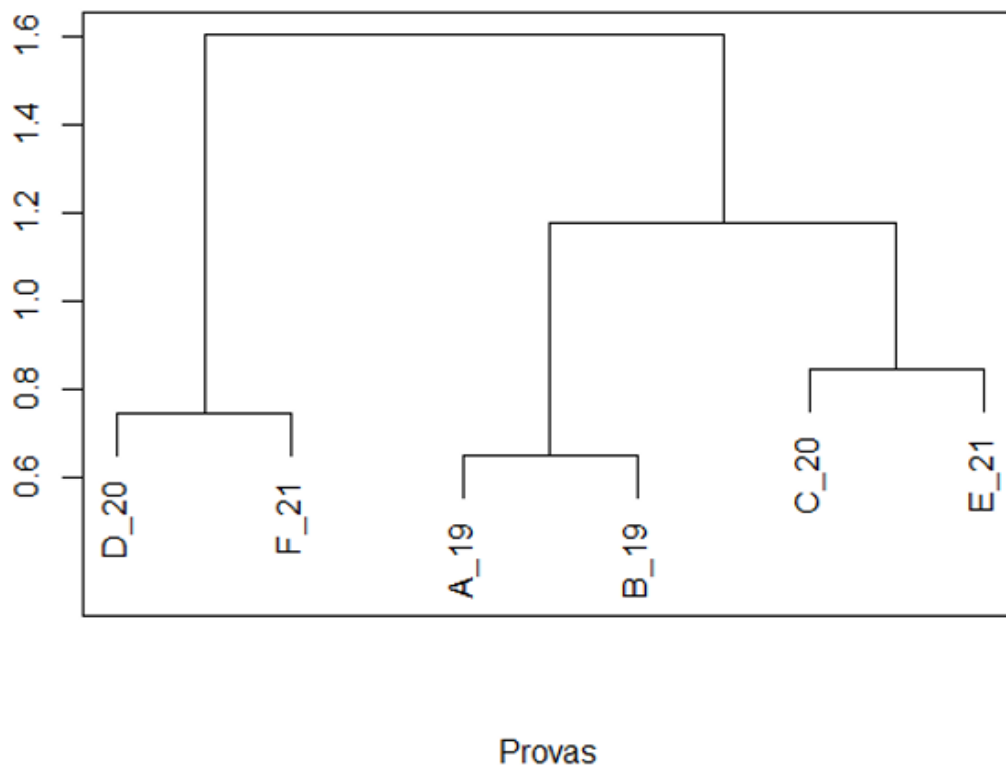
Para auxiliar ainda mais na interpretação e consequentemente na conclusão da técnica existe a representação gráfica conhecida como *Dendograma* (gráfico em forma de árvore), ilustrado na Figura 4. No eixo vertical à esquerda são marcados os níveis de similaridade, no eixo horizontal temos os objetos em estudo dispostos numa ordem conveniente e partindo deles temos as linhas verticais que possuem determinada altura e essa altura determina o quanto os objetos são considerados semelhantes. Por meio do dendograma observamos que no passo 5 haverá dois grupos homogêneos que são $(D_{20}F_{21})$ e $(C_{20}E_{21}A_{19}B_{19})$.

3.2.8 Interpretação dos Resultados

De acordo com o Dendograma da Figura 4 e a Tabela 10 a Análise de Agrupamentos pelo M.M.D. foi concluída em cinco passos, no passo 1 foi feita a junção da prova regular e a reaplicação de 2019 ($A_{19}B_{19}$) como sendo as mais semelhantes, no passo 2 a junção foi entre a reaplicação dos anos de 2020 e 2021 ($D_{20}F_{21}$), no passo 3 a união foi entre as provas regulares do ano de 2020 e 2021 ($C_{20}E_{21}$), já no passo 4 a junção foi entre os grupos obtidos no passo 3 e 1 ($C_{20}E_{21}$ e $A_{19}B_{19}$) e finalizando no passo 5 com a junção dos dois grupos de Provas ($D_{20}F_{21}$ e $C_{20}E_{21}A_{19}B_{19}$).

Se interrompêssemos o algoritmo no passo 3 ficaríamos com três grupos, são eles: grupo 1, $A_{19}B_{19}$, grupo 2 $D_{20}F_{21}$ e grupo 3 $C_{20}E_{21}$. No aspecto interpretativo o grupo 1 é a junção da prova regular e a reaplicação de 2019, ou seja, são mais semelhantes,

Figura 4 – Dendograma das provas, segundo o Método da Média das Distâncias.



seguidos dos grupos 2 e 3, no qual o grupo 2 é a reaplicação dos anos de 2020 e 2021 e o grupo 3 é a prova regular aplicada nos anos de 2020 e 2021. Assim, concluímos que a aplicação da prova regular tende a ser mais parecida do que a sua reaplicação, excetuando o caso do ano de 2019.

3.3 Medidas de Distância e Similaridade

3.3.1 Medidas de Similaridade e Dissimilaridade

Considere um conjunto de dados composto de n elementos amostrais, tendo-se medido p -variáveis aleatórias em cada um deles. O objetivo é agrupar em g grupos, para cada elemento amostral j , tem-se, portanto, o vetor de medidas X_j definido por:

$$X_j = [X_{1j}, X_{2j}, \dots, X_{pj}]', j = 1, 2, 3, \dots, n,$$

onde X_{ij} representa o valor observado da variável i medida no elemento j . Para agrupar os elementos é necessário escolher a medida de similaridade ou dissimilaridade a ser utilizada.

As medidas de dissimilaridade mais comuns são, distância euclidiana, distância generalizada ou ponderada e distância de minkowsky, nelas quanto menor for os seus valores, mais similares são os elementos que estão sendo comparados. Já as medidas de similaridade quanto maior o valor observado menos parecido será (mais dissimilares).

Exemplo 3.1

Neste exemplo podemos ver o comportamento de três variáveis X_1, X_2 e X_3 cujo coeficiente de parença (similaridade) foi o coeficiente de correlação.

$$S = \begin{pmatrix} & X_1 & X_2 & X_3 \\ X_1 & 1.000 & -- & -- \\ X_2 & 0.839 & 1.000 & -- \\ X_3 & -0.608 & -0.803 & 1.000 \end{pmatrix}$$

Note que as variáveis que possuem uma maior correlação são X_1 e X_2 , ou seja são mais parecidas. Enquanto X_2 e X_3 são menos similares pois possuem a menor correlação. Fazendo uma transformação $d(.,.) = 1 - corr(.,.)$ obtemos a matriz de similaridades.

$$D = \begin{pmatrix} & X_1 & X_2 & X_3 \\ X_1 & 0.000 & -- & -- \\ X_2 & 0.161 & 0.000 & -- \\ X_3 & 1.608 & 1.803 & 0.000 \end{pmatrix}$$

Na matriz de dissimilaridade verificamos que o maior valor observado é 1.803, ou seja a correlação entre X_2 e X_3 que são os objetos menos parecidos.

No caso em que o coeficiente de correlação é negativo, utilizamos a transformação $d(.,.) = 1 - |corr(.,.)|$, que pode ter o mesmo significado que o positivo, ou seja indica o mesmo grau de similaridade, como pode ser observado abaixo:

$$D = \begin{pmatrix} & X_1 & X_2 & X_3 \\ X_1 & 0.000 & -- & -- \\ X_2 & 0.161 & 0.000 & -- \\ X_3 & 0.392 & 0.197 & 0.000 \end{pmatrix}$$

Muitas vezes os coeficientes de parença não são definidos de um modo muito preciso, e não definem uma métrica sobre o espaço dos objetos. Isso pode levar a alguns problemas sérios de interpretação (BUSSAB; MIAZAKI; ANDRADE, 1990). Assim nesse estudo, faremos uma apresentação com as variáveis quantitativas e qualitativas nominais.

3.3.2 Coeficientes de Parecência para Variáveis Quantitativas

3.3.2.1 Medidas Derivadas da Distância Euclidiana

Seja X o vetor de coordenadas reais (X_1, X_2, \dots, X_p) como descritor dos objetos que serão investigados as similaridades. Utilizaremos a distância mais conhecida para indicar a proximidade de dois objetos A e B que é a distância euclidiana:

$$d(A, B) = \left[\sum_{i=1}^p (X_i(A) - X_i(B))^2 \right]^{\frac{1}{2}} \quad (3.4)$$

em linguagem matricial escrevemos da seguinte maneira:

$$d(A, B) = \left[(X(A) - X(B))'(X(A) - X(B)) \right]^{\frac{1}{2}}. \quad (3.5)$$

Derivada da expressão (3.5) temos a Distância Euclidiana Média, que é a raiz quadrada da razão entre a soma das diferenças ao quadrado e o número de coordenadas p , isto é:

$$d(A, B) = \left[\frac{\sum_{i=1}^p (X_i(A) - X_i(B))^2}{p} \right]^{\frac{1}{2}} \quad (3.6)$$

A distância definida em (3.6) possui as mesmas propriedades da distância (3.5) desse modo se submetido às técnicas de A.A. produzirão os mesmos resultados. Além disso, ela possui duas propriedades: A primeira é que ela pode ser usada na ausência de dados para algumas coordenadas (missing values) e a segunda é que ela permite acumular evidências empíricas sobre os níveis de parecência (BUSSAB; MIAZAKI; ANDRADE, 1990).

Exemplo 3.2

Tabela 11 – Dados das Variáveis Quantitativas do Exemplo Básico.

PROVAS	X_1	X_2	X_3	Z_1	Z_2	Z_3
A ₁₉	240.127	188.607	0.150	1.413	1.141	-1.110
B ₁₉	225.300	187.260	0.159	0.905	1.082	-0.111
C ₂₀	203.002	149.414	0.160	0.143	-0.571	0.000
D ₂₀	167.010	142.972	0.164	-1.088	-0.852	0.444
E ₂₁	182.835	170.745	0.152	-0.547	0.361	-0.889
F ₂₁	174.715	135.916	0.175	-0.825	-1.160	1.665
MÉDIA	198.831	162.486	0.160	0.0	0.0	0.0
D.PADRÃO	29.235	22.899	0.009	1.0	1.0	1.0

Fonte: Próprio autor.

(a) Distância Euclidiana

Calculando a Distância Euclidiana entre as Provas A_{19} e B_{19} :

$$\begin{aligned} d_2(A_{19}, B_{19}) &= [(240.127 - 225.300)^2 + (188.607 - 187.260)^2]^{\frac{1}{2}} = \\ &= (221.654338)^{\frac{1}{2}} = 14.8880602 \cong 14.888. \end{aligned}$$

A distância euclideana entre as variáveis X_1 , X_2 e X_3 das Provas A_{19} e B_{19} :

$$\begin{aligned} d_3(A_{19}, B_{19}) &= [(240.127 - 225.300)^2 + (188.607 - 187.260)^2 + (0.150 - 0.159)^2]^{\frac{1}{2}} = \\ &= (221.654419)^{\frac{1}{2}} = 14.888063 \cong 14.888. \end{aligned}$$

(b) Coeficiente Médio da Distância Euclideana

Pela expressão 3.6 podemos calcular a distância entre os pontos A_{19} e B_{19} , logo:

$$d_2(A_{19}, B_{19}) = \left(\frac{14.888}{2} \right)^{\frac{1}{2}} = 10.527$$

e

$$d_3(A_{19}, B_{19}) = \left(\frac{14.888}{3} \right)^{\frac{1}{2}} = 8.596$$

Note que d_2 e d_3 possuem um número de coordenadas diferentes porém a magnitude dos coeficientes são comparáveis.

(c) Distância Euclideana Padronizada

Vimos que utilizando os dados do exemplo 3.2 contidos na Tabela 11 podemos somar grandezas não comparáveis na distância euclidiana, fazendo isso podemos alterar completamente o significado e o valor do coeficiente. Em decorrência disso deve ser feita a padronização dessas variáveis, isto é:

$$Z_i = \frac{X_i(\cdot) - \bar{X}_i}{S_i}, \quad (3.7)$$

onde \bar{X} e S_i são respectivamente a média e o desvio padrão da i -ésima coordenada, passando a ser escrita como:

$$d(A, B) = \left[\sum_{i=1}^p (Z_i(A) - Z_i(B))^2 \right]^{\frac{1}{2}}. \quad (3.8)$$

Substituindo a expressão 3.7 na expressão 3.8 temos a soma dos desvios padronizados:

$$d(A, B) = \left[\sum_{i=1}^p \left(\frac{X_i(A) - X_i(B)}{S_i} \right)^2 \right]^{\frac{1}{2}}. \quad (3.9)$$

Em notação vetorial podemos escrever a expressão 3.9 da seguinte forma:

$$d(A, B) = [(X(A) - X(B))'D^{-1}(X(A) - X(B))]^{\frac{1}{2}}, \quad (3.10)$$

onde \mathbf{D} é uma matriz diagonal, tendo como p -ésimo componente a variância S_p^2 , isto é:

$$D = \text{diag}(s_1^2, s_2^2, \dots, s_p^2). \quad (3.11)$$

De modo análogo, definimos a distância euclidiana média como:

$$d(A, B) = \left[\frac{(X(A) - X(B))'D^{-1}(X(A) - X(B))}{p} \right]^{\frac{1}{2}}. \quad (3.12)$$

(d) Coeficiente da distância euclidiana padronizada

Pelos dados do Exemplo 3.2, temos que:

$$d_2(A_{19}, B_{19}) = [(1.413 - 0.905)^2 + (1.141 - 1.082)^2 + (0.150 - 0.159)^2]^{\frac{1}{2}} = 1.122,$$

ou

$$d_2(A_{19}, B_{19}) = \left[\left(\frac{240.127 - 225.300}{29.235} \right)^2 + \left(\frac{188.607 - 187.260}{22.899} \right)^2 + \left(\frac{0.150 - 0.159}{0.009} \right)^2 \right]^{\frac{1}{2}} = 1.122,$$

ou ainda,

$$\begin{aligned} d_2(A_{19}, B_{19}) &= \left[(240.127 - 225.300, 188.607 - 187.260, 0.150 - 0.159) \begin{pmatrix} 854.679 & 0 & 0 \\ 0 & 524.381 & 0 \\ 0 & 0 & 0.009 \end{pmatrix}^{-1} \begin{pmatrix} 14.827 \\ 1.347 \\ -0.009 \end{pmatrix} \right]^{\frac{1}{2}} \\ &= \left[(14.827, 1.347, -0.009) \begin{pmatrix} \frac{1}{854.679} & 0 & 0 \\ 0 & \frac{1}{524.381} & 0 \\ 0 & 0 & \frac{1}{0.009^2} \end{pmatrix} \begin{pmatrix} 14.827 \\ 1.347 \\ -0.009 \end{pmatrix} \right]^{\frac{1}{2}} = 1.122. \end{aligned}$$

(e) Distância Euclidiana Ponderada

Este coeficiente de parença é usado quando o pesquisador resolve ponderar as variáveis, ou seja, dar um peso maior a variável de interesse que julgar mais importante para definir semelhança. Assim, pode-se criar uma matriz \mathbf{D} baseada em critérios estatísticos ou criar pesos arbitrários para a diagonal da matriz \mathbf{D} .

$$d(A, B) = [(X(A) - X(B))'B(X(A) - X(B))]^{\frac{1}{2}}, \quad (3.13)$$

sendo \mathbf{B} a matriz de ponderação e $d(A, B)$ a distância ponderada por \mathbf{B} . Os casos particulares mais importantes são:

- (i) $B = I$, a ponderação é a matriz identidade, tem-se então a distância euclidiana usual.
- (ii) $B = [diag(s_1^2, s_2^2, \dots, s_p^2)]^{-1}$, e tem-se a distância das variáveis padronizadas.
- (iii) $B = V^{-1}$ é a matriz de covariâncias, tem-se então a “distância de Mahalanobis”.

A distância de Mahalanobis pondera pela variabilidade de cada uma das componentes, bem como leva em conta o grau de correlação entre elas, devido a isso torna muito difícil a interpretação de resultados.

3.3.2.2 Outros coeficientes

Os coeficientes de parença são criados com o intuito de moldar situações especiais de interesse do pesquisador, muitas dessas medidas são usadas em áreas específicas onde muito se utiliza a técnica da Análise de Agrupamento (taxonomia numérica, identificação de padrões, botânica, geologia, marketing, etc). Apresentaremos alguns coeficientes usados frequentemente, ou portadores de propriedades interessantes.

(a) Valor Absoluto

É muito comum usar o valor absoluto em vez de desvios quadráticos.

$$d(A, B) = \sum_{i=1}^p W_i |X_i(A) - X_i(B)|, \quad (3.14)$$

onde W_i 's representam as ponderações para as variáveis. As equiponderações mais usadas são $W_i = 1$ ou da média $W_i = \frac{1}{p}$. Utilizando as variáveis padronizadas do Exemplo 3.2, calculemos a distância entre os objetos A_{19} e B_{19} :

$$d(A_{19}, B_{19}) = \left\{ \frac{|1.413 - 0.905| + |1.141 - 1.082| + |-1.110 + 0.111|}{3} \right\} = 0.522.$$

(b) Distância de Minkowsky

A generalização da expressão 3.14 é dada por:

$$d(A, B) = \left[\sum_{i=1}^p W_i |X_i(A) - X_i(B)|^k \right]^{\frac{1}{k}}. \quad (3.15)$$

Para $k = 1$ e $k = 2$ passa a ser o caso anterior e a distância euclideana respectivamente, e para $k = 3$ e $W_i = \frac{1}{3}$ tem-se,

$$d(A_{19}, B_{19}) = \left\{ \frac{|1.413 - 0.905|^3 + |1.141 - 1.082|^3 + |-1.110 + 0.111|^3}{3} \right\}^{\frac{1}{3}} = 0.722.$$

(c) Coeficiente de Gower

Este coeficiente é baseado na proporção da variação em relação à maior discrepância possível:

$$d(A, B) = -\log_{10} \left[1 - \frac{1}{p} \sum_{i=1}^p \frac{|X_i(A) - X_i(B)|}{\max\{x_i\} - \min\{x_i\}} \right]. \quad (3.16)$$

Usando os dados do Exemplo 3.2, temos que:

$$d(A_{19}, B_{19}) = -\log_{10} \left[1 - \frac{1}{3} \left\{ \frac{|1.413-0.905|}{1.413-(-1.088)} + \frac{|1.141-1.082|}{1.141-(-1.160)} + \frac{|-1.110+0.111|}{1.665-(-1.110)} \right\} \right] = 0.095.$$

(d) Coeficiente de Similaridade de Cattell.

$$d(A, B) = \frac{2 \left(p - \frac{2}{3} \right) - d^2}{2 \left(p - \frac{2}{3} \right) + d^2}, \quad (3.17)$$

onde d^2 é a distância euclideana com variáveis padronizadas. Exemplificando tem-se:

$$d(A_{19}, B_{19}) = \frac{2 \left(3 - \frac{2}{3} \right) - 1.259^2}{2 \left(3 - \frac{2}{3} \right) + 1.259^2} = 0.575.$$

Derivada da expressão (3.17) temos o coeficiente de Cattell e Coulter (CORMACK, 1971):

$$d(A, B) = \frac{\sqrt{2p} - d^2}{\sqrt{2p} + d^2}. \quad (3.18)$$

Exemplificando tem-se:

$$d(A_{19}, B_{19}) = \frac{\sqrt{2p} - d^2}{\sqrt{2p} + d^2} = \frac{\sqrt{2 \cdot 3} - 1.259^2}{\sqrt{2 \cdot 3} + 1.259^2} = 0.321.$$

(e) Coeficiente para Variáveis Positivas

Alguns coeficientes são baseados no fato dos critérios de similaridade assumirem valores estritamente positivos e outros que foram ampliados para englobar valores negativos. No entanto, se as variáveis envolvidas assumirem apenas valores positivos fica mais fácil de compreender. Desse modo, temos abaixo alguns coeficientes que podem ser usados nesse contexto (BUSSAB; MIAZAKI; ANDRADE, 1990):

(e1) Coeficiente de Canberra.

$$d(A, B) = \frac{1}{p} \sum_{i=1}^p \frac{|X_i(A) - X_i(B)|}{X_i(A) + X_i(B)}. \quad (3.19)$$

(e2) Coeficiente de Bray-Curtis.

$$d(A, B) = \frac{\sum |X_i(A) - X_i(B)|}{\sum (X_i(A) + X_i(B))}. \quad (3.20)$$

(e3) Coeficiente de Sokal e Sneathurtis.

$$d(A, B) = \left\{ \frac{1}{p} \sum \left(\frac{X_i(A) - X_i(B)}{X_i(A) + X_i(B)} \right)^2 \right\}^{\frac{1}{2}}. \quad (3.21)$$

Exemplo 3.3 Usando os dados do Exemplo 3.2 podemos fazer outro tipo de relativização através da transformação

$$Z = \frac{X - X_{min}}{X_{max} - X_{min}}. \quad (3.22)$$

Obtendo-se a seguinte Tabela 12:

Tabela 12 – Relativização da Variáveis X_1, X_2 e X_3 .

PROVAS	X_1	X_2	X_3	Z_1	Z_2	Z_3
A ₁₉	240.127	188.607	0.150	1.000	1.000	0.000
B ₁₉	225.300	187.260	0.159	0.797	0.974	0.360
C ₂₀	203.002	149.414	0.160	0.492	0.256	0.400
D ₂₀	167.010	142.972	0.164	0.000	0.134	0.560
E ₂₁	182.835	170.745	0.152	0.216	0.661	0.080
F ₂₁	174.715	135.916	0.175	0.105	0.000	1.000
$X_{max} - X_{min}$	73.117	52.691	0.025	1.000	1.000	1.000

Fonte: Próprio autor.

(e1) Canberra.

$$d(A_{19}, B_{19}) = \frac{1}{3} \left\{ \frac{|1.000 - 0.797|}{1.000 + 0.797} + \frac{|1.000 - 0.974|}{1.000 + 0.974} + \frac{|0.000 - 0.360|}{0.000 + 0.360} \right\} = 0.375.$$

(e2) Bray-Curtis

$$d(A_{19}, B_{19}) = \frac{|1.000 - 0.797| + |1.000 - 0.974| + |0.000 - 0.360|}{(1.000 + 0.797) + (1.000 + 0.974) + (0.000 + 0.360)} = 0.143.$$

(e3) Sokal-Sneath.

$$d(A_{19}, B_{19}) = \left\{ \frac{1}{3} \left[\left(\frac{0.203}{1.797} \right)^2 + \left(\frac{0.026}{1.974} \right)^2 + \left(\frac{-0.360}{0.360} \right)^2 \right] \right\}^{\frac{1}{2}} = 0.581.$$

Exemplificamos acima os três coeficientes utilizando as provas A_{19} e B_{19} .

3.3.3 Coeficientes de Parecença para Variáveis Qualitativas Nominais

Quando estamos diante de variáveis qualitativas necessitamos de coeficientes que possam definir o grau de similaridade entre os objetos. De início apresentaremos o caso onde os critérios envolvidos são todos do tipo binário (sim ou não), em seguida enfatizaremos as variáveis com múltiplos atributos.

Na literatura são encontradas muitas propostas para esses tipos de coeficientes e praticamente qualquer medida de associação para as chamadas tabelas de contingência, pode ser usada como medida de parecença.

3.3.3.1 Coeficientes de Parecença para Variáveis Dicotômicas

Usaremos como exemplo as informações apresentadas nas Tabelas 13 e 14. No qual queremos medir a similaridade entre as provas de Matemática do Enem de 2019 (A_{19}) e 2021 (E_{21}) em sua aplicação regular, no qual a partir de sete variáveis obteremos como respostas a ausência ou presença de determinadas características.

Tabela 13 – Resultados sobre a presença (1) ou ausência (0) de determinadas características a partir de sete variáveis X_n , com $n = 1, 2, 3, \dots, 7$.

VARIÁVEL	X_1	X_2	X_3	X_4	X_5	X_6	X_7
A_{19}	0	1	0	1	1	0	1
E_{21}	0	1	1	0	1	0	0

onde X_1 é uma variável indicadora na qual $X_1 = 1$ indica a presença de todas as Habilidades presentes na Matriz de Referência e $X_1 = 0$ indica a ausência de pelo menos uma habilidade; X_2 é uma variável que indica a presença ($X_2 = 1$) de todas as Competências presentes na Matriz de Referência ou a ausência ($X_2 = 0$) de pelo menos uma das Competências; X_3 é uma variável indicadora da anulação ($X_3 = 1$) de pelo menos um item do Enem ou a não-anulação ($X_3 = 0$) de nenhum item Enem; a variável X_4 indica a presença ($X_4 = 1$) das Habilidades ($H5, H9, H14, H18, H23$) e ($X_4 = 0$) indica a ausência de pelo menos uma dessas Habilidades; a variável X_5 indica a presença ($X_5 = 1$) de todas as Habilidades ($H24, H25, H26, H27, H28, H29, H30$) e ($X_5 = 0$) indica a ausência de pelo menos uma dessas habilidades; a variável X_6 no qual $X_6 = 1$ indica a presença de quatorze letras repetidas ou mais no gabarito e $X_6 = 0$ indica a quantidade de letras repetidas menor ou igual a treze; já a variável X_7 indica a presença ($X_7 = 1$) de uma sequência de três letras ou mais repetidas e consecutivas no gabarito ou a ausência ($X_7 = 0$) dessa repetição de sequência de três letras ou mais.

A Tabela 14 foi criada a fim de facilitar os cálculos para exemplificarmos os coeficientes de semelhança para variáveis dicotômicas.

Tabela 14 – Número observado de pares (1,1), (1,0), (0,1), (0,0).

	A_{19}		TOTAL	
	1	0		
E_{21}	1	$a = 2$	$b = 1$	$a + b = 3$
	0	$c = 2$	$d = 2$	$c + d = 4$
TOTAL		$a + c = 4$	$b + d = 3$	$p = 7$

onde \mathbf{a} é a presença nas duas provas (A_{19} e E_{21}) de determinadas características, \mathbf{b} é a ausência na prova A_{19} e a presença na prova E_{21} de determinadas características, \mathbf{c} é a presença na prova A_{19} e a ausência na prova E_{21} de determinadas características e \mathbf{d} é a ausência nas duas provas (A_{19} e E_{21}) de determinadas características.

(a) Distância Euclidiana Média

Calculando a distância euclidiana entre os dois vetores da Tabela 13, tem-se:

$$d(A, B) = \left[\frac{1}{p} \sum_{i=1}^p (X_i(A) - X_i(B))^2 \right]^{\frac{1}{2}} = \left(\frac{b+c}{p} \right)^{\frac{1}{2}} = \left(\frac{b+c}{a+b+c+d} \right)^{\frac{1}{2}} \quad (3.23)$$

A distância de Sokal (Expressão 3.23) indica a proporção de atributos não coincidentes nos dois objetos. Sendo uma medida de dissimilaridade, quanto maior for esse número mais diferentes serão os objetos. Sua amplitude varia de 0 a 1, onde o valor nulo significa maior similaridade entre as cidades, esse valor é:

$$d(A_{19}, E_{21}) = \left[\frac{1+2}{2+1+2+2} \right]^{\frac{1}{2}} = \left[\frac{3}{7} \right]^{\frac{1}{2}} = 0.429.$$

(b) Coeficiente de Concordância Simples

Com o objetivo de construir um coeficiente de similaridade. A proposta mais usada é a proporção de coincidências, isto é:

$$s(A_{19}, E_{21}) = \frac{a+d}{a+b+c+d} = \frac{4}{7} = 0.571. \quad (3.24)$$

Esse coeficiente também varia de 0 a 1 e quanto maiores os valores mais similares os objetos serão.

(c) Coeficiente de Concordâncias Positivas

Para medir a similaridade entre objetos baseando-se na presença ou ausência de determinada característica, usamos o coeficiente de Russel e Rao (Expressão 3.25) ou o coeficiente de Jaccard (Expressão 3.26), conforme expressões abaixo, respectivamente.

$$s(A_{19}, E_{21}) = \frac{a}{a+b+c+d} = \frac{2}{7} = 0.286 \quad (3.25)$$

$$s(A_{19}, E_{21}) = \frac{a}{a + b + c} = \frac{2}{5} = 0.400 \tag{3.26}$$

(d) Outras Medidas

Muitos outros coeficientes foram criados oriundos dos anteriores, como podemos ver na Tabela 15.

Tabela 15 – Alguns Coeficientes de semelhança para variáveis dicotômicas (Baseado em (ROMESBURG, 1984)).

Nome	Expressão	Intervalo de Variação	Exemplo ilustrativo
Distância Binária de Sokal	$\left(\frac{b+c}{a+b+c+d}\right)^{\frac{1}{2}}$	(0,1)	0.655
Coincidência Simples	$\frac{a+d}{a+b+c+d}$	(0,1)	0.571
Rogers e Tanimoto	$\frac{a+d}{a+2(b+c)+d}$	(0,1)	0.400
Sokal e Sneath	$\frac{2(a+d)}{2(a+d)+b+c}$	(0,1)	0.727
Russel e Rao	$\frac{a}{a+b+c+d}$	(0,1)	0.286
Jaccard	$\frac{a}{a+b+c}$	(0,1)	0.400
Soreson	$\frac{2a}{2a+b+c}$	(0,1)	0.571
Ochiai	$\frac{a}{[(a+b)(a+c)]}$	(0,1)	0.167
Baroni-Urbani-Buser	$\frac{a+(ad)^{\frac{1}{2}}}{a+b+c+(ad)^{\frac{1}{2}}}$	(0,1)	0.571
Haman	$\frac{(a+d)-(b+c)}{a+b+c+d}$	(-1,1)	0.143
Yule	$\frac{ad-bc}{ad+bc}$	(-1,1)	0.333
ϕ	$\frac{ad-bc}{[(a+b)(a+c)(b+d)(c+d)]^{\frac{1}{2}}}$	(-1,1)	0.167
Ochiai II	$\frac{ad}{[(a+b)(a+c)(b+d)(c+d)]^{\frac{1}{2}}}$	(0,1)	0.333
	$\frac{1}{2} \left(\frac{a}{a+b} + \frac{a}{a+c} \right)$	(0,1)	0.583
	$\frac{1}{4} \left(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{d+b} + \frac{d}{d+c} \right)$	(0,1)	0.625

Fonte: (BUSSAB; MIAZAKI; ANDRADE, 1990)

Outros tópicos não vistos neste trabalho tais como: Os coeficientes de parença para variáveis qualitativas ordinais, coeficientes de parença para variáveis de diferentes tipos (coeficiente de combinado de semelhança, a transformação em variáveis binárias, e outros coeficientes) podem ser vistos em (BUSSAB; MIAZAKI; ANDRADE, 1990).

3.4 Formando Agrupamentos

3.4.1 Introdução

Existe um grande número de algoritmos para formar agrupamentos, haja vista que cada pesquisador pode ter uma maneira diferente para medir as duas idéias básicas que são: coesão interna dos objetos e isolamento externo entre os grupos (CORMACK, 1971). A classificação dos algoritmos de “Análise de Agrupamentos” segundo (CORMACK, 1971) é dividida em três:

- (I) Técnicas Hierárquicas: os objetos são classificados em grupos distintos em diferentes etapas sequenciais, de modo hierárquico, obedecendo ao critério de parença e produzindo uma árvore de classificação.
- (II) Técnicas de Partição: Os agrupamentos obtidos produzem uma partição do conjunto de objetos sob análise.
- (III) Técnicas de Cobertura: Os agrupamentos obtidos recobrem o conjunto de objetos, podendo haver sobreposição entre eles.

Através dos exemplos, iremos fazer uma aplicação de cada um dos métodos hierárquicos (Método da Centróide, Método da Média das Distâncias, Método da Ligação Simples ou *Single Linkage* e o Método da Ligação Completa ou *Complete Linkage*).

Exemplo 4.1

Continuaremos usando o Exemplo Básico contido na Tabela 11 em especial as variáveis padronizadas (Z_1, Z_2, Z_3) e usando como medida de parença a distância euclidiana, ou seja:

(a) Coordenadas

Tabela 16 – Variáveis Padronizados Z_1, Z_2, Z_3

PROVAS	Z_1	Z_2	Z_3
A ₁₉	1.413	1.141	-1.110
B ₁₉	0.905	1.082	-0.111
C ₂₀	0.143	-0.571	0.000
D ₂₀	-1.088	-0.852	0.444
E ₂₁	-0.547	0.361	-0.889
F ₂₁	-0.825	-1.160	1.665

Fonte: Próprio autor.

(b) Matriz de Distâncias

$$\begin{pmatrix} & A_{19} & B_{19} & C_{20} & D_{20} & E_{21} \\ B_{19} & 0.648 & -- & -- & -- & -- \\ C_{20} & 1.387 & 1.053 & -- & -- & -- \\ D_{20} & 2.053 & 1.635 & 0.773 & -- & -- \\ E_{21} & 1.224 & 1.038 & 0.843 & 1.086 & -- \\ F_{21} & 2.449 & 1.930 & 1.163 & 0.743 & 1.723 \end{pmatrix}.$$

3.4.2 Técnicas Hierárquicas de Agrupamento

As técnicas hierárquicas podem ser subdivididas em aglomerativas e divisivas. As aglomerativas caracterizam-se por fusões sucessivas de n objetos, que vai-se obtendo $n - 1, n - 2$, etc. grupos, até reunir todos os objetos em um único grupo. As divisivas partem de um único grupo, e por divisões sucessivas vai-se obtendo 2, 3, etc. grupos. A característica desse processo é que a reunião de dois agrupamentos numa certa etapa produz um dos agrupamentos da etapa superior, caracterizando o processo hierárquico.

Exemplificaremos cada método usando o exemplo contido na Tabela 11.

3.4.2.1 Método da Centróide

Como vimos em seções anteriores, o que caracteriza os algoritmos de produzir agrupamentos é o critério usado para definir a distância entre grupos. O Método da Centróide (M.C.) é mais direto, ele substitui cada fusão de objetos num único ponto representado pelas coordenadas de seu centro, a distância entre grupos é definida pela distância entre os centros e em cada etapa procura-se fundir em grupos que tenham a menor distância entre si.

Exemplo 4.2

a) 1º Passo. Inicia-se o processo com cada um objeto alocado a um grupo. E a distância entre os grupos é a distância entre os objetos, indicada no Exemplo 4.1, pela matriz de distâncias.

b) 2º Passo. Pela matriz das distâncias percebemos que os grupos A_{19} e B_{19} são parecidos, então iremos fundir dando origem ao grupo $A_{19}B_{19}$, cujas coordenadas do seu centro são:

$$Z_1(A_{19}B_{19}) = \frac{(1.413 + 0.905)}{2} = 1.159,$$

$$Z_2(A_{19}B_{19}) = \frac{(1.141 + 1.082)}{2} = 1.111,$$

e

$$Z_3(A_{19}B_{19}) = \frac{(-1.110 - 0.111)}{2} = -0.610.$$

Novas coordenadas, considerando as distâncias normalizadas com o novo agrupamento.

Tabela 17 – M.C.

PROVAS	Z ₁	Z ₂	Z ₃
C ₂₀	0.143	-0.571	0.000
D ₂₀	-1.088	-0.852	0.444
E ₂₁	-0.547	0.361	-0.888
F ₂₁	-0.825	-1.160	1.665
A ₁₉ B ₁₉	1.159	1.111	-0.610

Fonte: Próprio autor.

Agora podemos construir uma nova matriz de distância, no qual mudará apenas as distâncias envolvendo o grupo A₁₉B₁₉, assim:

$$d(C_{20}, A_{19}B_{19}) = \left\{ \frac{(0.143 - 1.159)^2 + (-0.571 - 1.111)^2 + (0.000 + 0.610)^2}{3} \right\}^{\frac{1}{2}} = 1.188.$$

De modo análogo temos:

$$d(D_{20}, A_{19}B_{19}) = 1.827 \quad d(E_{21}, A_{19}B_{19}) = 1.088 \quad d(F_{21}, A_{19}B_{19}) = 2.181.$$

Logo, a nova matriz de distâncias entre os 5 grupos é:

$$\begin{pmatrix} & C_{20} & D_{20} & E_{21} & F_{21} \\ D_{20} & 0.773 & --- & --- & --- \\ E_{21} & 0.843 & 1.086 & --- & --- \\ F_{21} & 1.163 & 0.743 & 1.723 & --- \\ A_{19}B_{19} & 1.188 & 1.827 & 1.088 & 2.181 \end{pmatrix}.$$

c) 3º Passo. De maneira semelhante ao passo anterior, faremos a fusão entre D₂₀ e F₂₁ ao nível de 0.743.

d) 4º Passo. Neste passo faremos a fusão entre C₂₀ e D₂₀F₂₁ ao nível de 0.782.

e) 5º Passo. Agora faremos a fusão entre os grupos E₂₁ e C₂₀D₂₀F₂₁ ao nível de 0.920.

f) 6º Passo. Para finalizar reuniremos os grupos A₁₉B₁₉ com E₂₁C₂₀D₂₀F₂₁ ao nível de 1.039, a fim de obtermos um único grupo contendo todos os objetos. Portanto, o processo hierárquico por ser resumido na Tabela 18.

Baseado na Tabela 18, construiremos o Dendograma, conforme Figura 5.

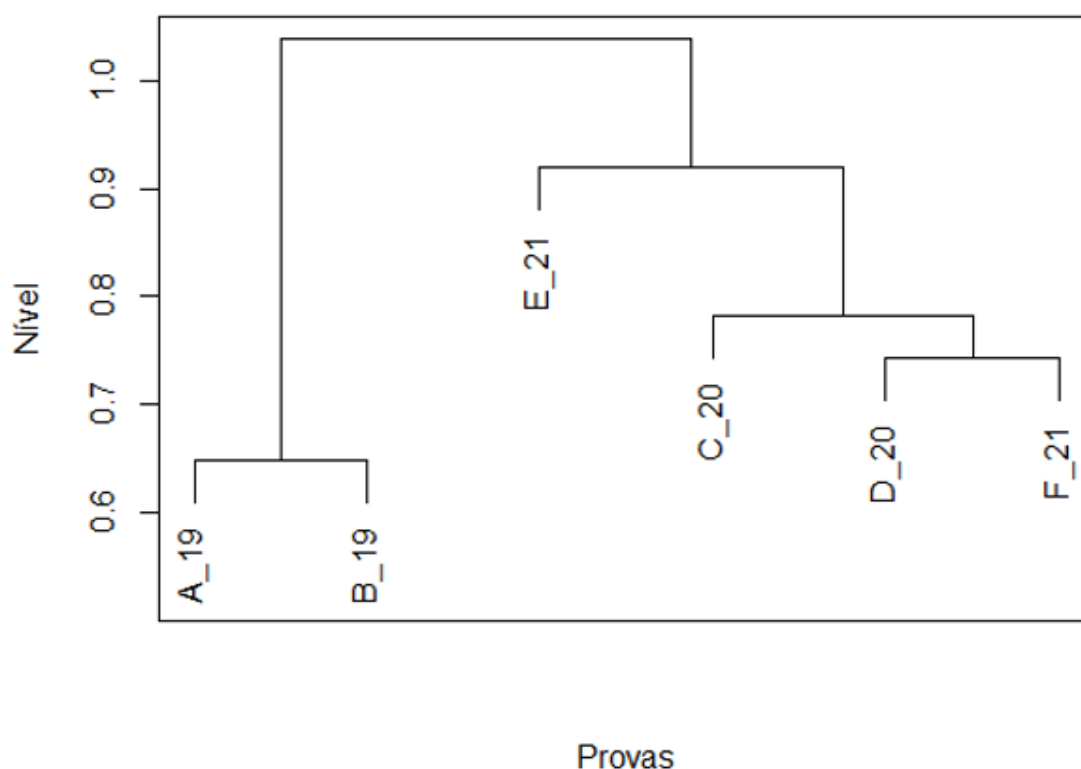
Ao usar essa técnica, há a necessidade de a cada passo voltar aos dados originais para recalcular as coordenadas e refazer as linhas da matriz de distâncias, há casos em que a matriz de parença é construída de tal maneira que não é possível voltar as

Tabela 18 – Resumo do processo hierárquico do M.C.

PASSO	JUNÇÃO	NÍVEL
1	A_{19} e B_{19}	0.648
2	D_{20} e F_{21}	0.743
3	C_{20} e $D_{20}F_{21}$	0.782
4	E_{21} e $C_{20}D_{20}F_{21}$	0.920
5	$A_{19}B_{19}$ e $E_{21}C_{20}D_{20}F_{21}$	1.039

Fonte: Próprio autor.

Figura 5 – Dendograma das provas, segundo o Método da Centróide)



coordenadas dos objetos. Os próximos processos hierárquicos vistos a seguir utilizam-se apenas da matriz de distâncias entre os objetos.

Atráves do Dendograma da Figura 5 observamos que no passo 1 as provas do ano de 2019 A_{19} e B_{19} (Prova regular e reaplicação) são mais parecidas de acordo com a Análise de Agrupamento (M.C.) realizada com os parâmetros que compõem cada item da prova. No passo 2 é feita a junção das provas da reaplicação de 2020 (D_{20}) e 2021 (F_{21}) sendo as mais parecidas, já no passo 3 esse grupo é unido à prova regular de 2020 (C_{20}). No quarto passo esse grupo formado com os objetos $C_{20}D_{20}F_{21}$ é unido com a prova regular de 2021 E_{21} . E no último passo é que temos a junção das provas realizadas em 2019 ($A_{19}B_{19}$) com o grupo do passo 4 ($E_{21}C_{20}D_{20}F_{21}$). Em resumo, a prova regular e a reaplicação de 2019 são mais dissimilares do que as demais provas se

considerarmos apenas dois grupos.

3.4.2.2 Método da Ligação Simples ou do Vizinho mais Próximo. (M.L.S.)

Neste método, a parença entre dois grupos é definido como sendo os dois membros mais parecidos, ou seja, entre todos os coeficientes de parença entre elementos de um grupo e de outro, escolhe-se o de maior parença como o coeficiente entre dois grupos. Logo, dados A e B , a distância entre eles será:

$$d(A, B) = \min\{d(i, j) : i \in A \text{ e } j \in B\}$$

ou no caso de similaridade

$$s(A, B) = \max\{s(i, j) : i \in A \text{ e } j \in B\}$$

Exemplo 4.3.

a) **1º Passo.** Inicialmente temos 6 grupos individuais e a matriz distância calculada no exemplo 4.1.

a) **2º Passo.** Os pontos A_{19} e B_{19} são os mais próximos, logo serão reunidos em um só grupo $A_{19}B_{19}$. Calculando a matriz das distâncias deste grupo aos demais, a partir da matriz inicial temos:

$$d(C_{20}, A_{19}B_{19}) = \min\{d(C_{20}, A_{19}), d(C_{20}, B_{19})\} = \min\{1.387; 1.053\} = 1.053,$$

$$d(D_{20}, A_{19}B_{19}) = \min\{d(D_{20}, A_{19}), d(D_{20}, B_{19})\} = \min\{2.053; 1.635\} = 1.635,$$

$$d(E_{21}, A_{19}B_{19}) = \min\{d(E_{21}, A_{19}), d(E_{21}, B_{19})\} = \min\{1.224; 1.038\} = 1.038,$$

e

$$d(F_{21}, A_{19}B_{19}) = \min\{d(F_{21}, A_{19}), d(F_{21}, B_{19})\} = \min\{2.449; 1.930\} = 1.930.$$

Então, a matriz de distâncias é dada por:

$$\begin{pmatrix} & C_{20} & D_{20} & E_{21} & F_{21} \\ D_{20} & 0.773 & --- & --- & --- \\ E_{21} & 0.843 & 1.086 & --- & --- \\ F_{21} & 1.163 & 0.743 & 1.723 & --- \\ A_{19}B_{19} & 1.053 & 1.635 & 1.038 & 1.930 \end{pmatrix}.$$

(c) **3º Passo.** Agruparemos D_{20} e F_{21} ao nível de 0.743, recalculando as demais distâncias tem-se:

$$d(C_{20}, D_{20}F_{21}) = \min\{d(C_{20}, D_{20}), d(C_{20}, F_{21})\} = \min\{0.773; 1.163\} = 0.773,$$

$$d(E_{21}, D_{20}F_{21}) = \min\{d(E_{21}, D_{20}), d(E_{21}, F_{21})\} = \min\{1.086; 1.723\} = 1.086,$$

e

$$\begin{aligned} d(A_{19}B_{19}, D_{20}F_{21}) &= \min\{d(A_{19}, D_{20}), d(A_{19}, F_{21}), d(B_{19}, D_{20}), d(B_{19}, F_{21})\} = \\ &= \min\{2.053; 2.449; 1.635; 1.930\} = 1.635. \end{aligned}$$

Estes resultados podem ser obtidos da matriz do passo anterior, as duas primeiras são as mesmas e a terceira pode ser escrita da seguinte maneira:

$$d(A_{19}B_{19}, D_{20}F_{21}) = \min\{d(A_{19}B_{19}, D_{20}), d(A_{19}B_{19}, F_{21})\} = \min\{1.635; 1.930\} = 1.635$$

A matriz resultante é:

$$\begin{pmatrix} & C_{20} & E_{21} & A_{19}B_{19} \\ E_{21} & 0.843 & -- & -- \\ A_{19}B_{19} & 1.053 & 1.038 & -- \\ D_{20}F_{21} & 0.773 & 1.086 & 1.635 \end{pmatrix}.$$

(d) **4º Passo.** Reunindo $D_{20}F_{21}$ com C_{20} , ao nível de 0.773 e recalculando as distâncias tem-se:

$$d(E_{21}, D_{20}F_{21}C_{20}) = \min\{d(E_{21}, D_{20}), d(E_{21}, F_{21}), d(E_{21}, C_{20})\} = \min\{1.086; 1.723; 0.843\} =$$

$$d(E_{21}, D_{20}F_{21}C_{20}) = 0.843.$$

$$d(A_{19}B_{19}, D_{20}F_{21}C_{20}) = \min\{d(A_{19}, D_{20}), d(A_{19}, F_{21}), d(A_{19}, C_{20}), d(B_{19}, D_{20}), d(B_{19}, F_{21}),$$

$$d(B_{19}, C_{20})\} = \min\{2.053; 2.449; 1.387; 1.635; 1.930; 1.053; \} = 1.053.$$

ou baseando-se na matriz anterior,

$$d(E_{21}, D_{20}F_{21}C_{20}) = \min\{d(E_{21}, F_{21}), d(E_{21}, C_{20})\} = \min\{1.086; 0.843\} = 0.843.$$

$$d(A_{19}B_{19}, D_{20}F_{21}C_{20}) = \min\{d(A_{19}B_{19}, D_{20}F_{21}), d(A_{19}B_{19}, C_{20})\} = \min\{1.635; 1.053; \} =$$

$$d(A_{19}B_{19}, D_{20}F_{21}C_{20}) = 1.053.$$

A matriz resultante é:

$$\begin{pmatrix} & E_{21} & A_{19}B_{19} \\ A_{19}B_{19} & 1.038 & -- \\ D_{20}F_{21}C_{20} & 0.843 & 1.053 \end{pmatrix}.$$

(e) **5º Passo.** Reunir E_{21} com $D_{20}F_{21}C_{20}$, ao nível de 0.843.

$$\begin{aligned} d(A_{19}B_{19}, D_{20}F_{21}C_{20}E_{21}) &= \min\{d(A_{19}, D_{20}), d(A_{19}, F_{21}), d(A_{19}, C_{20}), d(A_{19}, E_{21}), \\ & d(B_{19}, D_{20}), d(B_{19}, F_{21}), d(B_{19}, C_{20}), d(B_{19}, E_{21})\} = \\ &= \min\{2.053; 2.449; 1.387; 1.224; 1.635; 1.930; 1.053; 1.038\} = 1.038. \end{aligned}$$

ou ainda,

$$\begin{aligned} d(A_{19}B_{19}, D_{20}F_{21}C_{20}E_{21}) &= \min\{d(A_{19}B_{19}, E_{21}), d(A_{19}B_{19}, D_{20}F_{21}C_{20})\} = \\ &= \min\{1.038; 1.053\} = 1.038 \end{aligned}$$

Assim a matriz será:

$$\begin{matrix} & A_{19}B_{19} \\ D_{20}F_{21}C_{20}E_{21} & 1.038 \end{matrix}$$

(f) **6º Passo.** No último passo reuniremos em um único grupo contendo os seis objetos, ao nível de 1.038.

Tabela 19 – Resumo do processo hierárquico M.L.S.

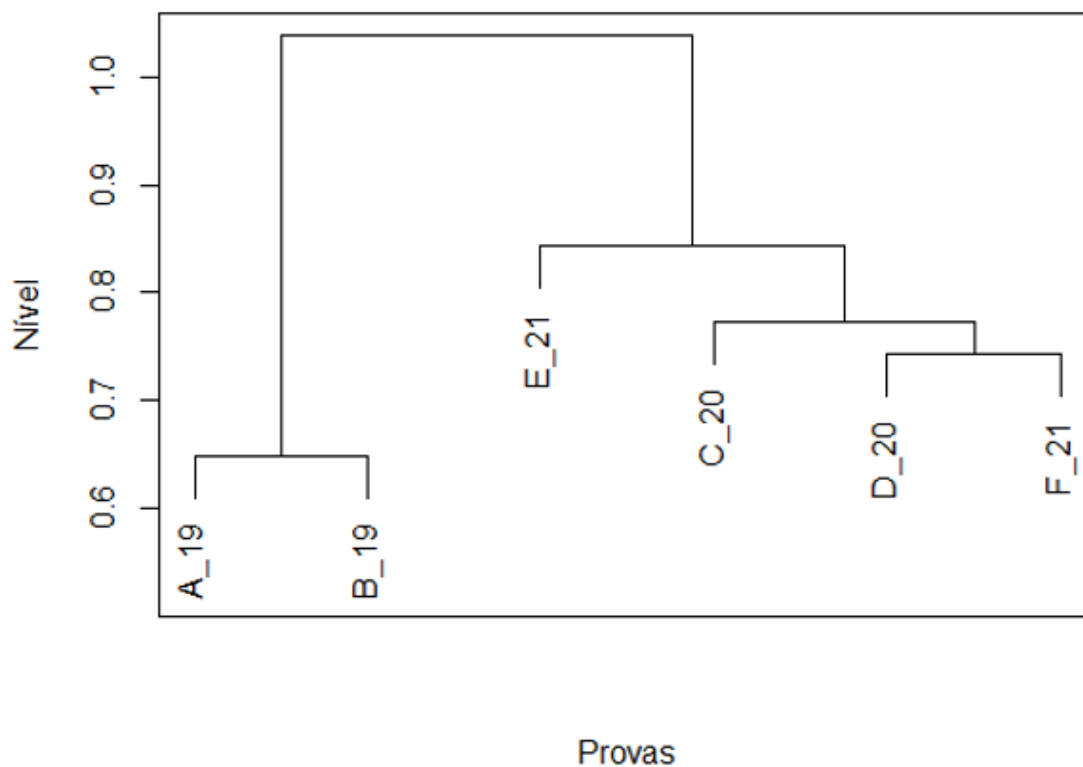
PASSO	JUNÇÃO	NÍVEL
1	A_{19} e B_{19}	0.648
2	D_{20} e F_{21}	0.743
3	$D_{20}F_{21}$ e C_{20}	0.773
4	$D_{20}F_{21}C_{20}$ e E_{21}	0.843
5	$D_{20}F_{21}C_{20}E_{21}$ e $A_{19}B_{19}$	1.038

Fonte: Próprio autor.

Em resumo de acordo com a Tabela 19, obtemos o seguinte Dendograma da Figura 6.

Neste método a formação dos grupos deu-se de maneira idêntica ao método estudado anteriormente (M.C.E.) mudando apenas os níveis em que cada objeto foi agrupado. Ou seja, possuindo uma mesma interpretação.

Figura 6 – Dendograma das provas, segundo o Método da Ligação Simples.



3.4.2.3 Método da Ligação Completa ou do Vizinho mais Longe. (M.L.C.)

Ao contrário do método anterior, a parecença entre dois grupos é definida pelos objetos de cada grupo que menos se parecem, ou seja, formam-se todos os pares com um membro de cada grupo e a parecença entre os grupos é definida pelo par que menos se parece. A parecença definida entre dois grupos A e B é dada por:

$$d(A, B) = \max\{d(i, j) : i \in A \text{ e } j \in B\}$$

Ressaltando que a fusão ainda é feita com os grupos mais parecidos, menor distância.

Exemplo 4.4.

(a) **1º Passo.** Iniciamos com seis grupos individuais e as distâncias da matriz do Exemplo 4.1.

(b) **2º Passo.** O mais parecidos são os grupos A_{19} e B_{19} , ao nível de 0.648 dando origem ao grupo $A_{19}B_{19}$. As distâncias ao novo grupo são calculadas, logo:

$$d(C_{20}, A_{19}B_{19}) = \max\{d(C_{20}, A_{19}), d(C_{20}, B_{19})\} = \max\{1.387; 1.053\} = 1.387,$$

$$d(D_{20}, A_{19}B_{19}) = \max\{d(D_{20}, A_{19}), d(D_{20}, B_{19})\} = \max\{2.053; 1.635\} = 2.053,$$

$$d(E_{21}, A_{19}B_{19}) = \max\{d(E_{21}, A_{19}), d(E_{21}, B_{19})\} = \max\{1.224; 1.038\} = 1.224,$$

$$d(F_{21}, A_{19}B_{19}) = \max\{d(F_{21}, A_{19}), d(F_{21}, B_{19})\} = \max\{2.449; 1.930\} = 2.449.$$

Obtendo-se a nova matriz de distâncias

$$\begin{pmatrix} & C_{20} & D_{20} & E_{21} & F_{21} \\ D_{20} & 0.773 & -- & -- & -- \\ E_{21} & 0.843 & 1.086 & -- & -- \\ F_{21} & 1.163 & 0.743 & 1.723 & -- \\ A_{19}B_{19} & 1.387 & 2.053 & 1.224 & 2.449 \end{pmatrix}.$$

(c) **3º Passo.** Os dois grupos mais parecidos são D_{20} e F_{21} ao nível de 0.743. As novas distâncias são:

$$d(C_{20}, D_{20}F_{21}) = \max\{d(C_{20}, D_{20}), d(C_{20}, F_{21})\} = \max\{0.743; 1.163\} = 1.163,$$

$$d(E_{21}, D_{20}F_{21}) = \max\{d(E_{21}, D_{20}), d(E_{21}, F_{21})\} = \max\{1.086; 1.723\} = 1.723,$$

$$\begin{aligned} d(A_{19}B_{19}, D_{20}F_{21}) &= \max\{d(A_{19}, D_{20}), d(A_{19}, F_{21}), d(B_{19}, D_{20}), d(B_{19}, F_{21})\} = \\ &= \max\{2.053; 2.449; 1.635; 1.930\} = 2.449. \end{aligned}$$

Esses resultados podem ser obtidos da matriz do passo anterior, por exemplo:

$$d(A_{19}B_{19}, D_{20}F_{21}) = \max\{d(A_{19}B_{19}, D_{20}), d(A_{19}B_{19}, F_{21})\} = \max\{2.053; 2.449\} =$$

$$d(A_{19}B_{19}, D_{20}F_{21}) = 2.449.$$

Em consequência disso, tem-se:

$$\begin{pmatrix} & C_{20} & E_{21} & A_{19}B_{19} \\ E_{21} & 0.843 & -- & -- \\ A_{19}B_{19} & 1.387 & 1.224 & -- \\ D_{20}F_{21} & 1.163 & 1.723 & 2.449 \end{pmatrix}.$$

(d) **4º Passo.** Reunindo C_{20} com E_{21} ao nível de 0.843. Calculando as distâncias tem-se:

$$\begin{aligned} d(A_{19}B_{19}, C_{20}E_{21}) &= \max\{d(A_{19}, C_{20}), d(A_{19}, E_{21}), d(B_{19}, C_{20}), d(B_{19}, E_{21})\} = \\ &= \max\{1.387; 1.224; 1.053; 1.038\} = 1.387 \end{aligned}$$

$$\begin{aligned} d(D_{20}F_{21}, C_{20}E_{21}) &= \max\{d(D_{20}, C_{20}), d(D_{20}, E_{21}), d(F_{21}, C_{20}), d(F_{21}, E_{21})\} \\ &= \max\{0.773; 1.086; 1.163; 1.723\} = 1.723 \end{aligned}$$

Obtendo-se:

$$\begin{pmatrix} & C_{20}E_{21} & A_{19}B_{19} \\ A_{19}B_{19} & 1.387 & -- \\ D_{20}F_{21} & 1.723 & 2.449 \end{pmatrix}.$$

(e) **5º Passo.** Reunindo $A_{19}B_{19}$ com $C_{20}E_{21}$, com as seguintes distâncias:

$$\begin{aligned} & d(D_{20}F_{21}, A_{19}B_{19}C_{20}E_{21}) = \\ & = \max\{d(D_{20}, A_{19}), d(D_{20}, B_{19}), d(D_{20}, C_{20}), d(D_{20}, E_{21}), \\ & d(F_{21}, A_{19}), d(F_{21}, B_{19}), d(F_{21}, C_{20}), d(F_{21}, E_{21})\} = \max\{1, 41; 2, 49\} = 2.449 \end{aligned}$$

$$\begin{array}{cc} D_{20}F_{21} & \\ A_{19}B_{19}C_{20}E_{21} & 2.449 \end{array}$$

(f) **6º Passo.** Reunimos os grupos $D_{20}F_{21}$ com $A_{19}B_{19}C_{20}E_{21}$ ao nível de 2.449. Logo, podemos ver de forma resumida os passos na Tabela 20:

Tabela 20 – Resumo do processo hierárquico M.L.C.

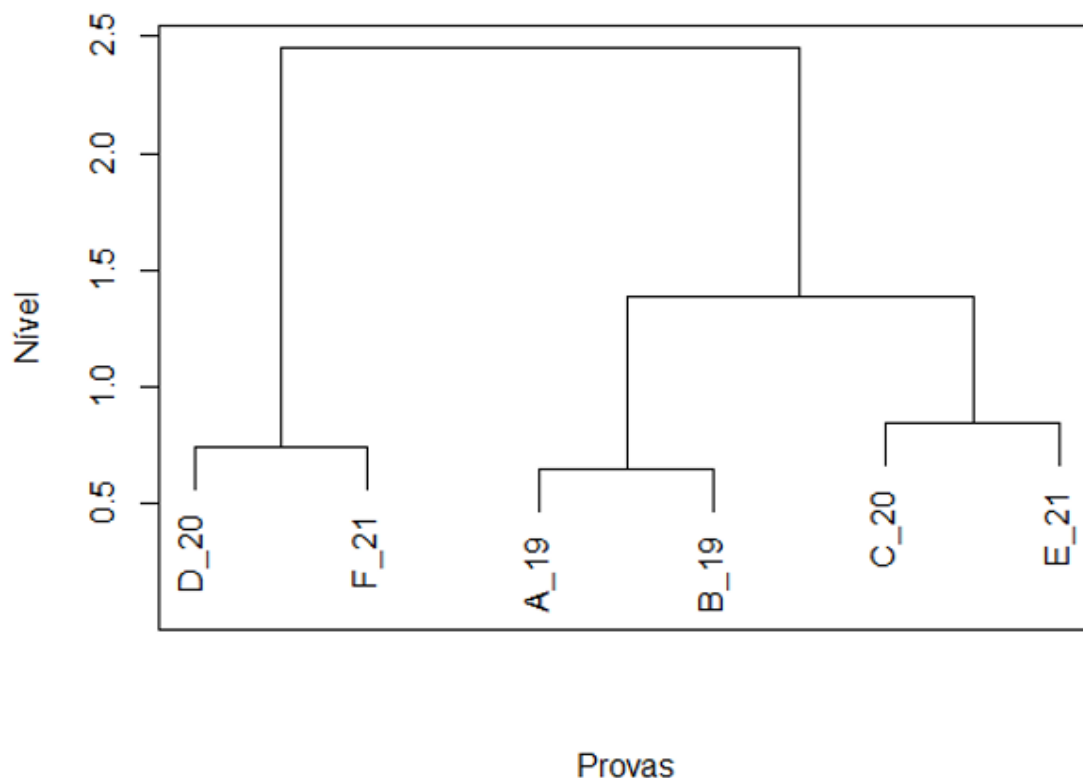
PASSO	JUNÇÃO	NÍVEL
1	A_{19} e B_{19}	0.648
2	D_{20} e F_{21}	0.743
3	C_{20} e E_{21}	0.843
4	$A_{19}B_{19}$ e $C_{20}E_{21}$	1.387
5	$D_{20}F_{21}$ e $A_{19}B_{19}C_{20}E_{21}$	2.449

Fonte: Próprio autor

Assim, com base na Tabela 20 e no Dendograma da Figura 7 vemos que no passo 1 houve a fusão da prova regular de 2019 (A_{19}) com a reaplicação de 2019 (B_{19}), no passo 2 é feita a junção das provas da reaplicação de 2020 (D_{20}) e 2021 (F_{21}) sendo as mais parecidas, no passo 3 a fusão foi entre a prova regular de 2020 (C_{20}) e a prova regular de 2021 (E_{21}). Já no passo 4 foi feita a fusão entre os passos 2 ($D_{20}F_{21}$) e 3 ($C_{20}E_{21}$), e finalizando o passo 5 com o agrupamento dos grupos $D_{20}F_{21}$ com $A_{19}B_{19}C_{20}E_{21}$.

Desse modo, concluímos que este método, o M.C e o M.L.S. mostraram que a prova regular e a reaplicação de 2019 são mais similares, além de haver uma similaridade entre as provas regulares de 2020 e 2021 e as reaplicações de 2020 e 2021, respectivamente. Já o que encontramos de diferença entre os três métodos estudados é que a reaplicação de 2020 e 2021 é mais dissimilar do que as demais provas no M.L.C., e no M.C. e no M.L.S. a prova regular e a reaplicação de 2019 é mais dissimilar do que as demais provas, se considerarmos dois grupos formados.

Figura 7 – Dendograma das provas, segundo o Método da Ligação Completa.



3.4.3 Métodos de Partição

Estes métodos objetivam encontrar uma partição de n objetos em k grupos (*clusters*), que satisfaçam às duas premissas básicas: “coesão” interna (ou “semelhança” interna) e isolamento (ou separação) dos *clusters* formados. Estas técnicas exigem a pré-fixação de critérios que produzam medidas sobre a qualidade da partição produzida, assim como esse método de partição pressupõe também o conhecimento do número k de partições desejadas, ao contrário das técnicas hierárquicas aglomerativas. Os métodos não hierárquicos mais conhecidos são o k -Médias (*k-means*) e o Fuzzy c -Médias (*Fuzzy c-Means*).

A diferença entre os algoritmos de partição depende da escolha diferente de um ou mais dos seguintes procedimentos abaixo:

- Método de iniciar os agrupamentos;
- Método de designar objetos aos agrupamentos iniciais;
- Método de redesignar um ou mais objetos já agrupados para outros agrupamentos.

3.4.3.1 Método das *k-means*

O Método das *k-means* (HARTIGAN; WONG et al., 1979) é o método não-hierárquico mais conhecido e mais utilizado em problemas práticos. Nele, cada elemento amostral é alocado no *cluster* cujo centróide é o mais próximo do seu vetor de valores observados.

As principais etapas deste método são:

Passo 1: Para iniciar o processo escolhem-se k centróides, chamados de sementes ou protótipos.

Passo 2: Cada elemento do conjunto de dados é comparado com cada centróide inicial através de uma medida de parença. O elemento é alocado no grupo correspondente à menor distância.

Passo 3: Após aplicar o passo 2 para todos os n elementos amostrais, atualiza-se os valores dos centróides de todos os grupos formados, e repete-se o passo 2 considerando os centróides desses novos grupos.

Passo 4: Os passos 2 e 3 são repetidos até que todos os elementos amostrais estejam alocados e nenhuma realocação de elementos seja necessária.

A escolha das sementes iniciais pode afetar o resultado final do algoritmo, sendo assim, cuidados são necessários na escolha das sementes. (MINGOTI, 2013) apresenta algumas sugestões a seguir:

Sugestão 1: Uso de técnicas hierárquicas aglomerativas

Para iniciar utilizamos algum dos métodos de agrupamento das técnicas hierárquicas aglomerativas para se obter k grupos iniciais. Logo após, calcula-se o centróide de cada grupo formado e inicia o método *k-means*.

Sugestão 2: Escolha aleatória

Escolhemos os k centróides iniciais aleatoriamente dentro do conjunto de dados a ser analisado por meio de uma amostragem aleatória simples sem reposição. Embora seja de execução simples, essa estratégia de escolha não é eficiente. Pode-se melhorar a eficiência, selecionando-se m amostras aleatórias constituídas de k sementes, com $m > 1$. Ou seja, o procedimento de amostragem aleatória simples é repetido m vezes. E depois, calculam-se as médias das m amostras, para cada grupo, que irão constituir os centróides de inicialização do processo de agrupamento.

Sugestão 3: Escolha via uma variável aleatória

Escolheremos a variável aleatória com maior variância dentre as p componentes do vetor aleatório X sob consideração, essa variável, por si só, já induz uma “partição natural” nos dados observados. O domínio (ou suporte) dessa variável é dividido em k sub-intervalos e as sementes consistem dos pontos médios desses sub-intervalos.

Sugestão 4: Observação dos valores discrepantes do conjunto de dados

Por meio de uma análise estatística preliminar, buscam-se k objetos discrepantes no conjunto de dados em que cada objeto discrepante passa a ser uma semente. Neste caso, a discrepância é um relação às p variáveis observadas, uma espécie de “discrepância conjunta”.

Sugestão 5: Escolha prefixada

O pesquisador escolhe arbitrariamente as sementes, esta escolha é indicada quando o pesquisador tem um grande conhecimento do problema ou está buscando validar uma solução já existente, que está sujeito a um alto grau de subjetividade.

Sugestão 6: Os k primeiros valores do banco de dados

Em boa parte dos *Softwares* a opção *default* consiste em utilizar as k primeiras observações do banco de dados como centróides iniciais. Nesse procedimento quando os k primeiros elementos amostrais são discrepantes entre si pode trazer bons resultados, no entanto não é recomendável quando são semelhantes entre si.

Quando se tem muitos objetos, o método do *k-means* é o mais usado em Análise de Agrupamentos. Primeiramente escolhe-se o critério de homogeneidade dentro do grupo e heterogeneidade entre os grupos. Inspirado em Análise da Variância, a soma de quadrados residual é o critério mais usado.

Podemos indicar uma partição de n objetos em k grupos, da seguinte maneira:

$$p(1) = \{o_i(1) : 1 \leq i \leq n_1\},$$

$$p(2) = \{o_i(2) : 1 \leq i \leq n_2\},$$

$$p(j) = \{o_i(j) : 1 \leq i \leq n_j\},$$

$$p(k) = \{o_i(k) : 1 \leq i \leq n_k\}.$$

O centro do grupo $p(j)$, ou seja, o ponto formado pela média das coordenadas de seus membros, será representada por $\bar{o}(j)$. Assim, a soma de quadrados residuais dentro de j -ésimo grupo será:

$$SQRes(j) = \sum d^2(o_i(j); \bar{o}(j)); \quad 1 \leq i \leq n_j. \quad (3.27)$$

onde d^2 representa o quadrado da distância euclideana do objeto i , dentro do grupo j , ao seu centro. Para toda a partição a soma de quadrados residual será:

$$SQRes = \sum SQRes(j), \quad 1 \leq j \leq k. \quad (3.28)$$

Quanto menor for este valor, mais homogêneos são os elementos dentro de cada grupo e "melhor" será a partição (BUSSAB; MIAZAKI; ANDRADE, 1990).

Para melhor entender o método do *k-means* utilizaremos o Exemplo Básico com o auxílio do comando `kmeans` do *Software R-Studio*, na escolha da semente escolhemos a sugestão 1 (Uso de técnicas hierárquicas aglomerativas), no caso com $k = 3$. Para mais detalhes deste exemplo encontra-se em Anexo B a rotina do R para a obtenção destes resultados.

Em consequência, obtemos a seguinte composição.

Tabela 21 – *Clusters* das provas do Enem 2021 (Matemática), segundo o método *k-means*.

Grupos	Soma de Quadrados	Provas
$n_1 = 2$	1.066	C_{20}, E_{21} .
$n_2 = 2$	0.629	A_{19}, B_{19} .
$n_3 = 2$	0.827	D_{20}, F_{21}

Fonte: Próprio autor

Assim, como podemos ver na Tabela 21, os *clusters* formados são: a prova regular de 2020 com a regular de 2021 ($C_{20}E_{21}$), a prova regular de 2019 com a reaplicação de 2019 ($A_{19}B_{19}$) e a reaplicação de 2020 com a reaplicação de 2021 ($D_{20}F_{21}$). A composição dos grupos formados no método *k-means* foi a mesma que nos métodos M.M.D. e no M.L.C.

4 Aplicação

4.1 Objetivos

Este capítulo possui como objetivo aplicar a técnica de Análise de Agrupamentos com o propósito de investigar a existência de *clusters* de Unidades Federativas do Brasil, segundo características socioeconômicas e educacionais dos inscritos do Enem 2021.

4.2 Caracterização da População

A obtenção dos dados para a Análise de Agrupamentos foi a partir do site (INEP, 2021). Utilizamos a base de microdados dos inscritos no Enem 2021 conforme Tabela 22. Ao total 3.389.832 se inscreveram para o Enem 2021. A partir desse banco de dados extraímos inicialmente quarenta e sete variáveis socioeconômicas e educacionais relativo aos participantes, no qual posteriormente foi reduzido para vinte e uma variáveis, conforme Seção 4.3.

Tabela 22 – Número de Inscritos no Enem 2021 por Unidade Federativa (UF).

UF	Inscritos	UF	Inscritos	UF	Inscritos
RO	33.016	CE	221.322	RJ	237.894
AC	20.059	RN	80.427	SP	507.983
AM	90.057	PB	103.091	PR	143.877
RR	8.079	PE	190.664	SC	79.484
PA	186.875	AL	57.663	RS	150.887
AP	21.666	SE	52.369	MS	42.237
TO	31.454	BA	267.698	MT	55.813
MA	128.477	MG	330.515	GO	138.338
PI	79.867	ES	64.177	DF	65.843

Fonte: INEP 2021

4.3 Seleção de Variáveis

Esta seção tratará da seleção das variáveis que possam contribuir para a explicação da variabilidade dos dados originais, para isto utilizaremos um critério teórico (referenciais teóricos) e um critério técnico (coeficiente de correlação de Pearson)

Inicialmente coletamos quarenta e sete variáveis socioeconômicas e educacionais do banco de microdados do Enem 2021 a respeito dos inscritos, no entanto com base em referenciais teóricos e critérios técnicos, essa quantidade foi reduzida para vinte e uma variáveis.

Uma das políticas públicas desenvolvidas a fim de diminuir as desigualdades educacionais foi a Lei de Cotas (12.711/2012). Com a sua implementação, “o percentual de negros era de 22% e em 2015 houve acréscimo para 44%. Para essa mensuração, também é relevante mencionar o aumento das pessoas que se declararam pretas ou pardas: em 2001, 46,1%, e, em 2015, 53,9%, variação de 17%” (BRAZ et al., 2022). A partir de 2012, verifica-se um aumento de estudantes egressos oriundos de escolas públicas nas instituições federais de ensino superior, em especial pessoas PPI (SENKEVICS; MELLO, 2019). Os fatores de contexto, grau de instrução do pai/mãe, ocupação da mãe, assim como, ser preto/pardo/indígena (PPI) e o tipo de instituição na qual cursou o ensino médio são extremamente importantes no que diz respeito à produção de políticas públicas afirmativas.

Em relação ao desempenho, (FEIJÓ; FRANÇA, 2021) destaca grandes diferenças entre alunos de escolas públicas e privadas. O nível de escolaridade da mãe e ser do sexo feminino estiveram positivamente associados ao desempenho nos testes de Leitura e Matemática (CO-OPERATION; DEVELOPMENT, 2013). Algumas variáveis sociais se mostram estatisticamente importantes na predição do desempenho como, escolaridade dos pais, atraso escolar, sexo, cor/raça e dependência administrativa (ARISTIZABAL; CAICEDO; PARRA, 2017), (KARINO; LAROS, 2017). O Nível socioeconômico tem incluído como principais componentes a escolaridade dos pais, sua ocupação e a renda familiar (SIRIN, 2005). Para (TRAVITZKI; FERRÃO; COUTO, 2016) uma variável de aproximação do NSE é a escolaridade dos pais. Segundo (JALOTO; PRIMI, 2021) há uma associação entre os fatores socioeconômicos e desempenho nas quatro provas do Enem 2018 e o atraso escolar, o NSE dos alunos e a dependência administrativa da escola estão associados a diferenças no desempenho nas quatro áreas. E ainda segundo (SANTANA, 2020) “o grau de escolaridade dos pais ou dos responsáveis influencia no desempenho dos candidatos ao Enem”.

Já as variáveis necessidade especial declarada, habitantes por residência, assim como algumas outras já citadas apresentam uma “correlação linear fraca ou moderada”. De acordo com (MINGOTI, 2013), o coeficiente de correlação de Pearson mede a similaridade de duas variáveis em relação a sua linearidade, esse coeficiente varia entre -1 a 1, e quanto maior for o valor do coeficiente de correlação em valor absoluto, maior serão os indícios de uma relação linear. Logo, mais similares serão, então tomando uma base técnica quando duas variáveis apresentavam uma correlação forte, optou-se por apenas uma delas.

Após a aplicação desses dois critérios das quarenta e sete variáveis iniciais, restaram apenas vinte e uma variáveis, descritas na Tabela 23.

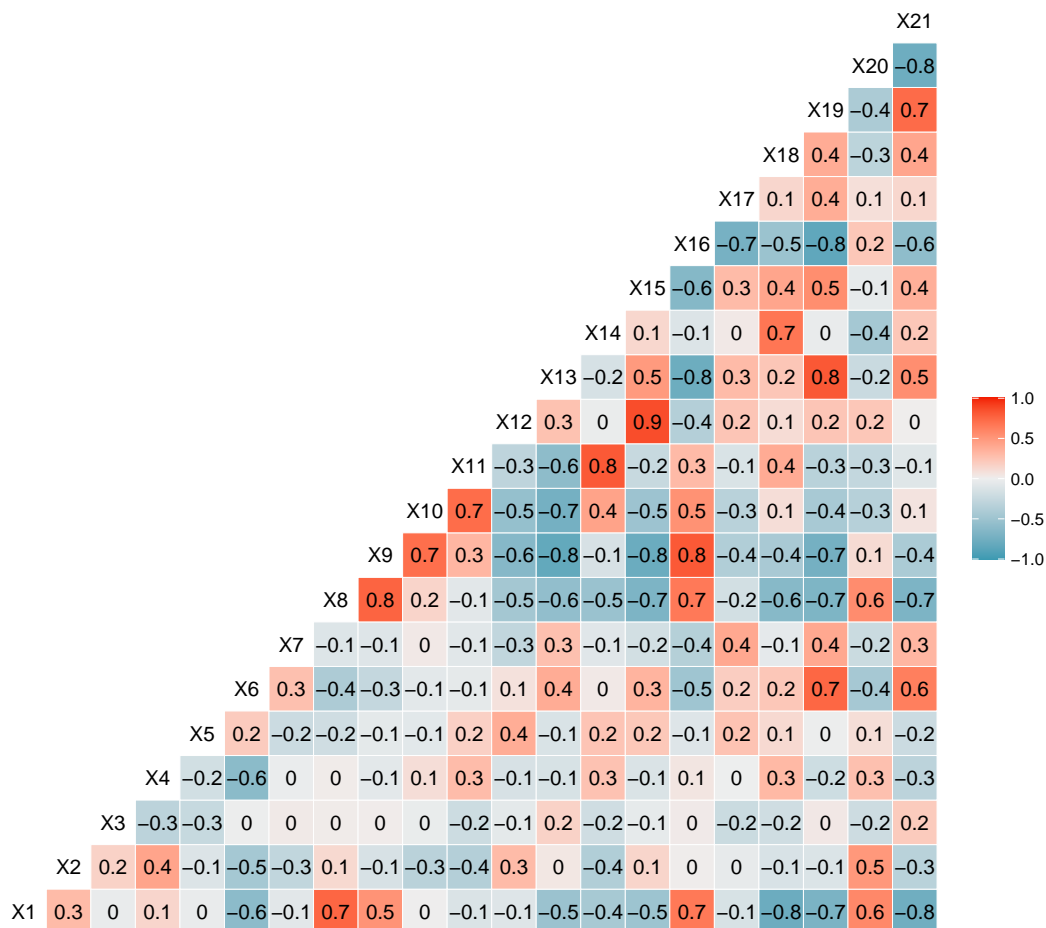
Tabela 23 – Variáveis Socioeconômicas e Educacionais dos Inscritos e suas respectivas definições e especificações para a Análise de Agrupamentos.

Variável	Definição	Especificação
X ₁	Preta/Parda	Proporção de inscritos por UF que se autodeclararam Pretos/Pardos.
X ₂	Indígena	Proporção de inscritos por UF que se autodeclararam Indígenas.
X ₃	Instituição Federal	Proporção de inscritos por UF oriundos de Instituição de Ensino Federal.
X ₄	Instituição Estadual	Proporção de inscritos por UF oriundos de Instituição de Ensino Estadual.
X ₅	Instituição Municipal	Proporção de inscritos por UF oriundos de Instituição de Ensino Municipal.
X ₆	Instituição Privada	Proporção de inscritos por UF oriundos de Instituição de Ensino Privada.
X ₇	Necessidade Especial Declarada	Proporção de inscritos por UF que se autodeclararam com alguma necessidade especial como: Baixa visão, cegueira, surdez, deficiência auditiva, surdo-cegueira, deficiência mental, déficit de atenção, dislexia, discalculia, autismo, visão monocular, outra deficiência ou condição especial
X ₈	Grau de Instrução da Mãe 1	Proporção de inscritos por UF que declararam que a mãe nunca estudou.
X ₉	Grau de Instrução da Mãe 2	Proporção de inscritos por UF que declararam que a mãe não completou a 4ª série/5º ano do Ensino Fundamental.
X ₁₀	Grau de Instrução da Mãe 3	Proporção de inscritos por UF que declararam que a mãe completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental.
X ₁₁	Grau de Instrução da Mãe 4	Proporção de inscritos por UF que declararam que a mãe completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio.
X ₁₂	Grau de Instrução da Mãe 5	Proporção de inscritos por UF que declararam que a mãe completou o Ensino Médio, mas não completou a Faculdade.
X ₁₃	Grau de Instrução da Mãe 6	Proporção de inscritos por UF que declararam que a mãe completou a faculdade ou Pós-graduação.
X ₁₄	Grau de Instrução do Pai 4	Proporção de inscritos por UF que declararam que o pai completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio.
X ₁₅	Grau de Instrução do Pai 5	Proporção de inscritos por UF que declararam que o pai completou o Ensino Médio, mas não completou a Faculdade.
X ₁₆	Ocupação da Mãe 1	Proporção de inscritos por UF que declararam que a mãe possui uma ou mais ocupações: lavradora, agricultora sem empregados, bóia fria, criadora de animais (gado, porcos, galinhas, ovelhas, cavalos etc.), apicultora, pescadora, lenhadora, seringueira, extrativista.
X ₁₇	Ocupação da Mãe 2	Proporção de inscritos por UF que declararam que a mãe possui uma ou mais ocupações: Diarista, empregada doméstica, cuidadora de idosos, babá, cozinheira (em casas particulares), motorista particular, jardineira, faxineira de empresas e prédios, vigilante, porteira, carteira, office-boy, vendedora, caixa, atendente de loja, auxiliar administrativa, recepcionista, servente de pedreiro, repositora de mercadoria.
X ₁₈	Ocupação da Mãe 3	Proporção de inscritos por UF que declararam que a mãe possui uma ou mais ocupações: padeira, cozinheira industrial ou em restaurantes, sapateira, costureira, joalheira, torneira mecânica, operadora de máquinas, soldadora, operária de fábrica, trabalhadora da mineração, pedreira, pintora, eletricista, encanadora, motorista, caminhoneira, taxista.
X ₁₉	Ocupação da Mãe 5	Proporção de inscritos por UF que declararam que a mãe possui uma ou mais ocupações: médica, engenheira, dentista, psicóloga, economista, advogada, juíza, promotora, defensora, delegada, tenente, capitã, coronel, professora universitária, diretora em empresas públicas ou privadas, política, proprietária de empresas com mais de 10 empregados.
X ₂₀	Habitantes por Residência	Proporção de inscritos por UF que declararam possuir de oito a vinte pessoas residindo em sua residência, incluindo com o próprio inscrito.
X ₂₁	Acesso a Internet	Proporção de inscritos por UF que declararam possuir acesso a internet.

Fonte: Próprio autor.

Neste caso, procuramos selecionar variáveis que possuem uma “baixa correlação”, ou seja, como podemos ver na Figura 8.

Figura 8 – Gráfico de Correlação de Pearson das Variáveis Socioeconômicas e Educa-
cionais em Estudo.



Próprio autor.

4.4 Resultados

Nesta seção apresentaremos os dendogramas obtidos por meio de métodos hierárquicos (Método da Ligação Simples, Método da Ligação Completa e o Método da Média das Distâncias). Para isso utilizaremos o comando `hcluster` do *software R-Studio* e através de tabelas e figuras faremos a interpretação.

Em seguida, iremos fazer uma análise do Método não-hierárquico K-means, no qual por meio do comando `kmeans` do *software* já citado iremos agrupar as UF's do Brasil de acordo com o seu nível de parecença, os resultados obtidos serão resumidos em tabelas e figuras.

No Método da Ligação Simples (M.L.S.) por intermédio do dendograma da Figura

9 os objetos foram agrupados, com seus respectivos passos, níveis e junção das unidades federativas conforme Tabela 24. Neste, através da análise do nível de fusão, o algoritmo será interrompido no passo 20 e será formado sete *clusters*. Os *clusters* formados são: CE, AM, PA, DF, RR, AP e AC/MA/RJ/TO/SP/RN/AL/PI//PB/SE/RO/PE/BA/RS/PR/SC/GO/MG/ES/MS/MT, sabendo que no próximo passo esse grupo maior será unido ao grupo AP, seguindo desse modo, a junção final será o CE com o grupo unificado nos passos anteriores.

Figura 9 – Dendograma das Unidades Federativas do Brasil, segundo o M.L.S.

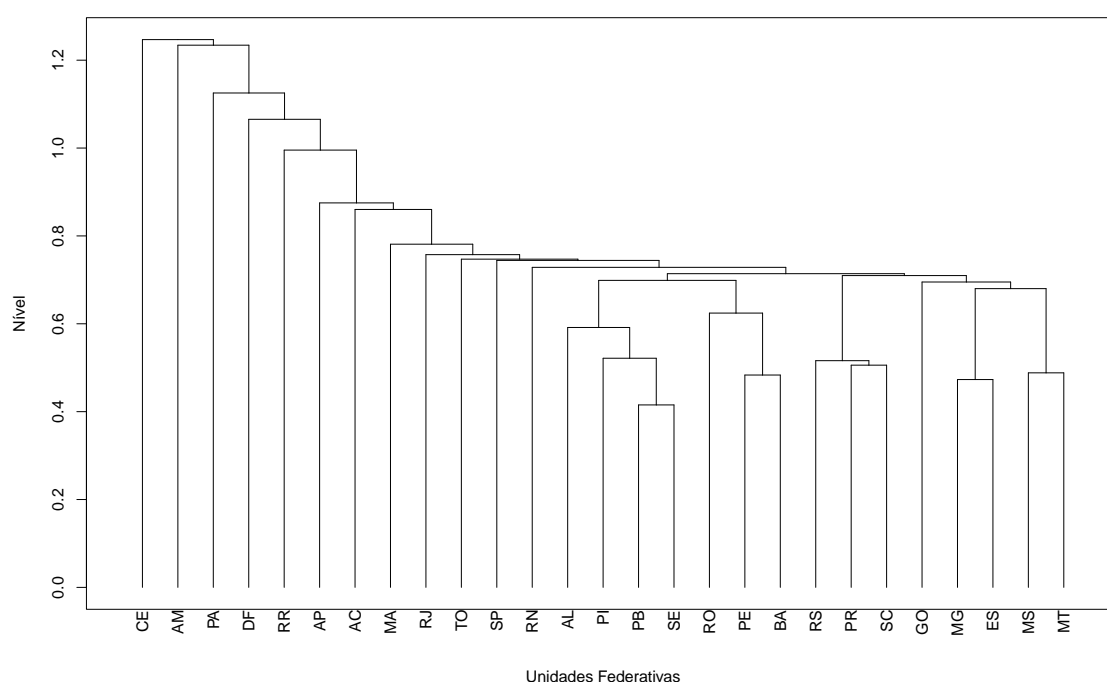


Tabela 24 – Resumo do M.L.S. Aplicado às Unidades Federativas do Brasil.

PASSO	NÍVEL	JUNÇÃO	PASSO	NÍVEL	JUNÇÃO
1	0.435	PB, SE	14	0.747	Grupos unidos nos passos 12 e 13
2	0.495	MG, ES	15	0.762	RN, grupo unido no passo 14
3	0.506	PE, BA	16	0.779	SP, grupo unido no passo 15
4	0.511	MS, MT	17	0.782	TO, grupo unido no passo 16
5	0.529	PR, SC	18	0.793	RJ, grupo unido no passo 17
6	0.540	RS, PR/SC	19	0.818	MA, grupo unido no passo 18
7	0.546	PI, PB/SE	20	0.900	AC, grupo unido no passo 19
8	0.619	AL, PI/PB/SE	21	0.916	AP, grupo unido no passo 20
9	0.653	RO, PE/BA	22	1.042	RR, grupo unido no passo 21
10	0.712	MG/ES, MS/MT	23	1.115	DF, grupo unido no passo 22
11	0.727	GO, MG/ES/MS/MT	24	1.178	PA, grupo unido no passo 23
12	0.731	AL/PI/PB/SE, RO/PE/BA	25	1.292	AM, grupo unido no passo 24
13	0.743	RS/PR/SC, GO/MG/ES/MS/MT	26	1.305	CE, grupo unido no passo 25

Fonte: Próprio autor

No Método da Ligação Completa (M.L.C.) através do dendograma da Figura 10 os objetos foram reunidos, com seus respectivos passos, níveis e junções das unidades federativas conforme Tabela 25. E conforme o nível de fusão utilizado nesse método o algoritmo de agrupamento foi interrompido no passo 16, formando onze *clusters*, que são: AM, PA/MA, CE, PI/PB/SE/RN/AL, RR, AC/AP, MS/MT/TO/RO/PE/BA, DF, GO/MG/ES, RS/PR/SC e RJ/SP.

Figura 10 – Dendograma das Unidades Federativas do Brasil, segundo o M.L.C.

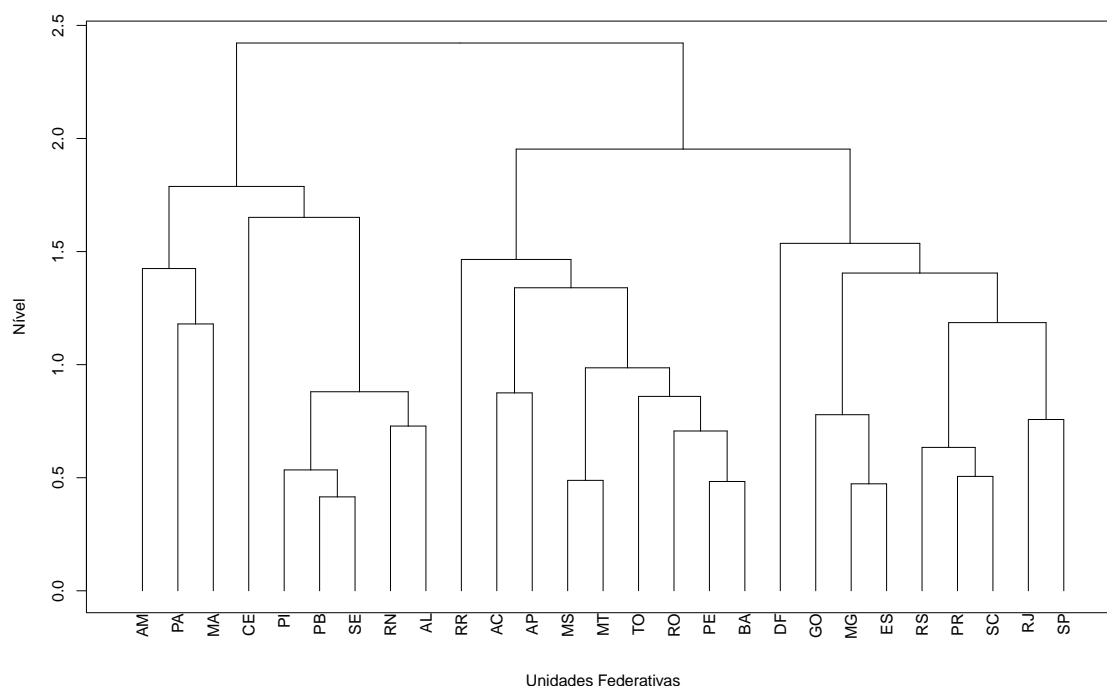


Tabela 25 – Resumo do M.L.C. Aplicado às Unidades Federativas do Brasil.

PASSO	NÍVEL	JUNÇÃO	PASSO	NÍVEL	JUNÇÃO
1	0.435	PB, SE	14	0.921	PI/PB/SE, RN/AL
2	0.495	MG, ES	15	1.032	MS/MT, TO/RO/PE/BA
3	0.506	PE, BA	16	1.235	PA, MA
4	0.511	MS, MT	17	1.241	RS/PR/SC, RJ/SP
5	0.529	PR, SC	18	1.402	AC/AP, MS/MT/TO/RO/PE/BA
6	0.559	PI, PB/SE	19	1.470	GO/MG/ES, RS/PR/SC/RJ/SP
7	0.664	RS, PR/SC	20	1.491	AM, PA/MA
8	0.740	RO, PE/BA	21	1.533	RR, AC/AP/MS/MT/TO/RO/PE/BA
9	0.762	RN, AL	22	1.608	DF, GO/MG/ES/RS/PR/SC/RJ/SP
10	0.793	RJ, SP	23	1.728	CE, PI/PB/SE/RN/AL
11	0.815	GO, MG/ES	24	1.871	AM/PA/MA, CE/PI/PB/SE/RN/AL
12	0.900	TO, RO/PE/BA	25	2.044	Grupos unidos nos passos 21 e 22
13	0.916	AC, AP	26	2.535	Grupos unidos nos passos 24 e 25

Fonte: Próprio autor

Os resultados obtidos pelo M.M.D. estão descritos na Tabela 26 e representados no dendograma da Figura 11.

Figura 11 – Dendrograma das Unidades Federativas do Brasil, segundo o M.M.D.

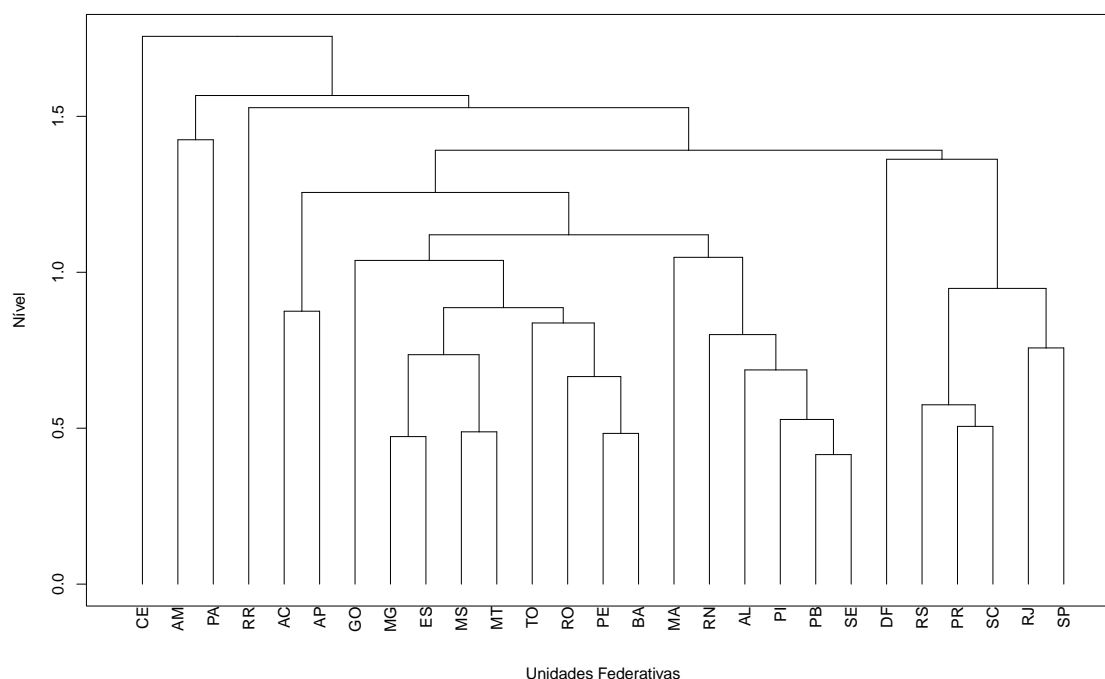


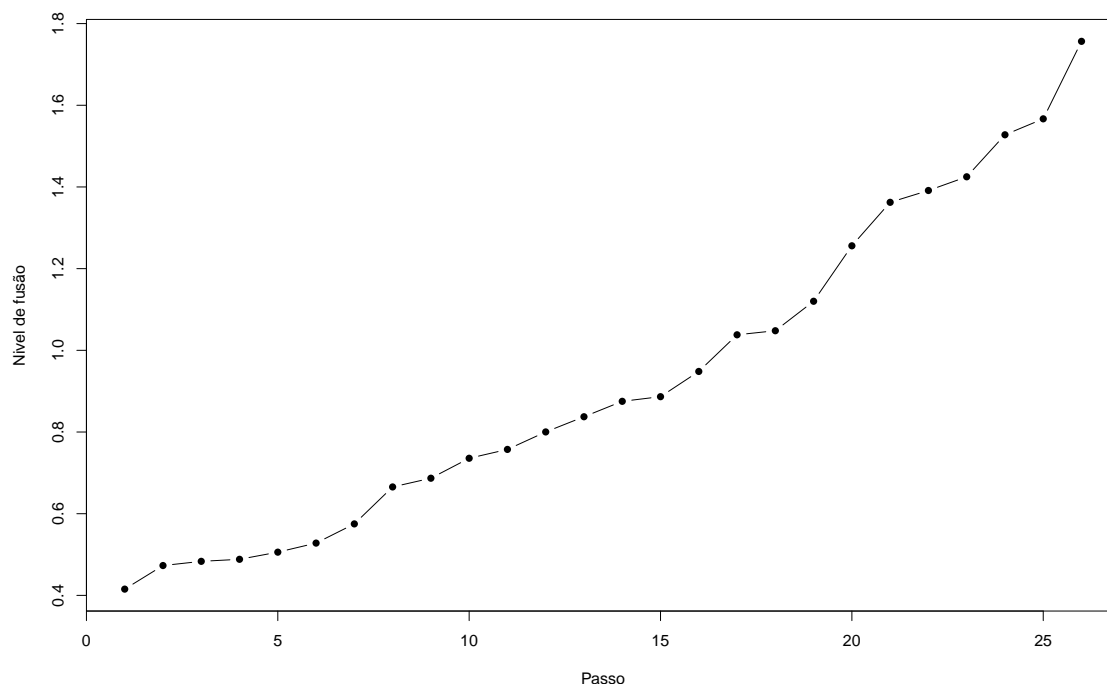
Tabela 26 – Resumo do M.M.D. Aplicado às Unidades Federativas do Brasil.

PASSO	NÍVEL	JUNÇÃO	PASSO	NÍVEL	JUNÇÃO
1	0.435	PB, SE	14	0.916	AC, AP
2	0.495	MS, MT	15	0.928	MG/ES/MS/MT, TO/RO/PE/BA
3	0.506	PE, BA	16	0.992	RS/PR/SC, RJ/SP
4	0.511	MG, ES	17	1.086	GO, MG/ES/MS/MT/TO/RO/PE/BA
5	0.529	PR, SC	18	1.097	MA, RN/AL/PI/PB/SE
6	0.553	PB/SE, PI	19	1.172	Grupos unidos nos passos 17 e 18
7	0.602	PR/SC, RS	20	1.314	AC/AP, Grupo unido no passo 19
8	0.696	RO, PE/BA	21	1.426	DF, RS/PR/SC/RJ/SP
9	0.719	AL, PI/PB/SE	22	1.456	Grupos unidos nos passos 20 e 21
10	0.770	MG/ES, MS/MT	23	1.491	AM, PA
11	0.793	RJ, SP	24	1.599	RR, grupo unido no passo 22
12	0.837	RN, AL/PI/PB/SE	25	1.640	AM/PA, grupo unido no passo 24
13	0.876	TO, RO/PE/BA	26	1.838	CE, grupo unido no passo 25

Fonte: Próprio autor

Usando o critério do nível de fusão e de acordo com a Figura 12, o algoritmo de agrupamento foi interrompido no passo 20, formando sete *clusters*, no entanto, desse modo teríamos um *cluster* com dezessete unidades federativas, logo achamos mais conveniente utilizar o próximo “ponto de salto”, que é no passo 12, formando assim quinze *clusters*. Os *clusters* formados são: CE, AM, PA, RR, AC, AP, GO, MG/ES/MS/MT, TO, RO/PE/BA, MA, RN/AL/PI/PB/SE, DF, RS/PR/SC e RJ/SP.

Figura 12 – Gráfico do Nível de Fusão do Método da Média das Distâncias (M.M.D.).



Fonte: Próprio autor

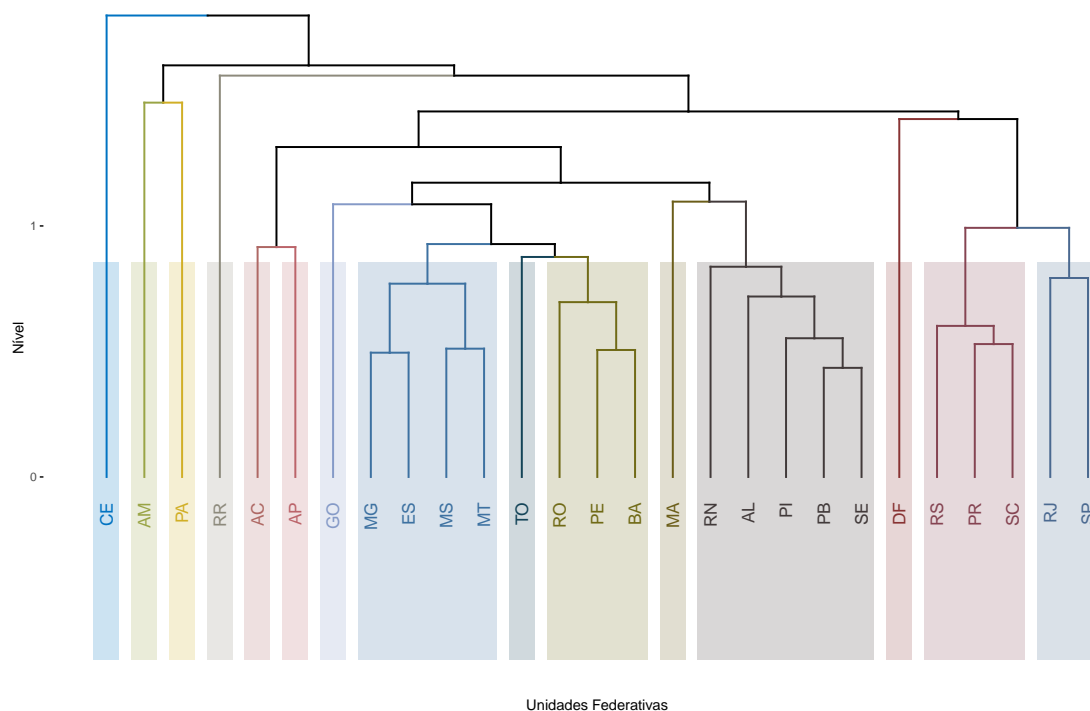
Assim, com base na Figura 13 podemos observar os quinze *clusters* formados, com suas respectivas Unidades Federativas, bem como construímos uma Tabela 27 para melhor representar os aglomerados.

Outra visualização da divisão das Unidades Federativas segundo o Método da Média das Distâncias pode ser visualizada na Figura 14.

Após a visualização e interpretação do dendograma com o auxílio da Tabela 24 do M.L.S verificamos que há um único *cluster* que reúne quinze UFs, dentro deste grupo há uma subdivisão de quatro grupos que são, AL/PI/PB/SE, RO/PE/BA, RS/PR/SC e GO/MG/ES/MS/MT. No qual esse grupo é unido com uma única UF de cada vez em cada passo, até o último passo que é a junção com a UF do CE. A UF do CE está em um único grupo, ela possui a maior proporção de inscritos no Enem 2021 que concluíram o ensino médio em dependência administrativa estadual, outro grupo que podemos fazer uma associação é o grupo formado por RS/PR/SC, estes possuem a menor proporção de inscritos que declararam o grau de instrução da mãe 1.

No M.L.C visualizamos a existência de alguns *clusters* com as UFs separadas em um único grupo, que são AM, CE, RR, DF. No caso de AM e RR deve-se ao fato de a UF do AM possuir a segunda maior proporção de inscritos pretos/pardos e indígenas, assim como possuir a menor proporção de inscritos oriundos de dependência admi-

Figura 13 – Dendrograma das Unidades Federativas do Brasil segundo o Método da Média das Distâncias.



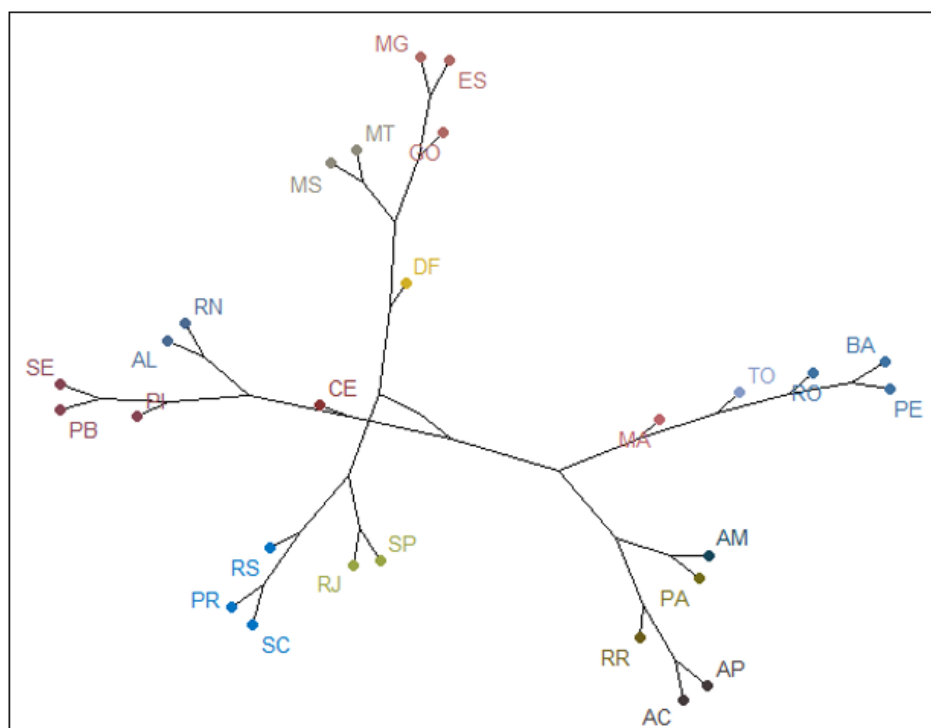
Fonte: Próprio autor

Tabela 27 – Clusters das Unidades Federativas do Brasil, segundo o M.M.D.

Grupos	Unidades Federativas
$n_1 = 1$	Ceará.
$n_2 = 1$	Amazonas.
$n_3 = 1$	Pará.
$n_4 = 1$	Roraima.
$n_5 = 1$	Acre.
$n_6 = 1$	Amapá.
$n_7 = 1$	Goiás.
$n_8 = 4$	Minas Gerais, Espírito Santo, Mato Grosso do Sul, Mato Grosso.
$n_9 = 1$	Tocantins.
$n_{10} = 3$	Rondônia, Pernambuco, Bahia.
$n_{11} = 1$	Maranhão.
$n_{12} = 5$	Rio Grande do Norte, Alagoas, Piauí, Paraíba, Sergipe.
$n_{13} = 1$	Distrito Federal.
$n_{14} = 3$	Rio Grande do Sul, Paraná, Santa Catarina.
$n_{15} = 2$	Rio de Janeiro, São Paulo.

Fonte: Próprio autor

Figura 14 – Gráfico da Árvore Filogenética recriada Aplicado ao M.M.D. após a interrupção do algoritmo no Passo 12.



Fonte: Próprio autor

nistrativa privada e possuir a menor proporção de inscritos com necessidade especial declarada. Em RR por possuir a menor proporção de inscritos que declararam o grau de instrução da mãe 5 e 6, além de possuir a maior proporção de inscritos no Enem que se autodeclararam indígena. E no DF por possui a maior proporção de inscritos que, declararam possuir alguma necessidade especial declarada e declarar o grau de instrução da mãe 6, bem como possuir a menor proporção de inscritos oriundos de dependência administrativa municipal e também a menor proporção que declararam o grau de instrução do pai 4.

E no M.M.D. podemos destacar a junção do RJ e SP em um único *cluster*, neste grupo é onde encontramos as maiores proporções de inscritos oriundos de dependência administrativa privada, que declararam que a mãe e o pai ocupam o grau de instrução 5, que declararam a ocupação da mãe 5, assim como possui a menor proporção de inscritos que declararam a ocupação da mãe 1. As UF's de GO, AP e AC também estão *clusters* separados, GO é a maior proporção de inscritos que declararam o grau de instrução do pai 4, como também possui a segunda maior proporção de inscritos que declararam a ocupação da mãe 2, no AP foi a maior proporção de inscritos que declararam haver entre 8 e 20 pessoas residindo na residência incluindo o inscrito e no AC houve a maior proporção de inscritos que declararam o grau de instrução da mãe

1.

Nota-se nos dendogramas obtidos que nos três métodos hierárquicos as Unidades Federativas do Amazonas, Ceará, Distrito Federal e Roraima estão em *clusters* separados, ou seja, estas unidades possuem uma distinção maior quando comparada com outras no que diz respeito às características socioeconômicas e educacionais. E ainda, que dois dentre os três métodos são mais semelhantes entre si, que são o Método da Ligação Completa e o Método da Média das Distâncias. Destes, temos sete *clusters* que possuem as mesmas unidades federativas, que são: Ceará, Amazonas, Roraima, Distrito federal, (Rio de Janeiro/São Paulo), (Rio Grande do Sul/Paraná/Santa Catarina) e (Rio Grande do Norte/Alagoas/Piauí/Paraíba/Sergipe). Pode-se perceber que nas três Tabelas 24,25, 26 as junções das unidades federativas são sempre as mesmas até o Passo 8.

Agora faremos uso do método não-hierárquico *k-means*. Iniciamos escolhendo o número de *clusters*, e de acordo com as sugestões de (MINGOTI, 2013) optamos pela sugestão 1 (Uso de técnicas hierárquicas aglomerativas), assim o método contará com quinze *clusters*, que são: CE, AM, PA, RR, AC, AP, GO, MG/ES/MS/MT, TO, RO/PE/BA, MA, RN/AL/PI/PB/SE, DF, RS/PR/SC e RJ/SP. Conforme ilustrado na Tabela 28.

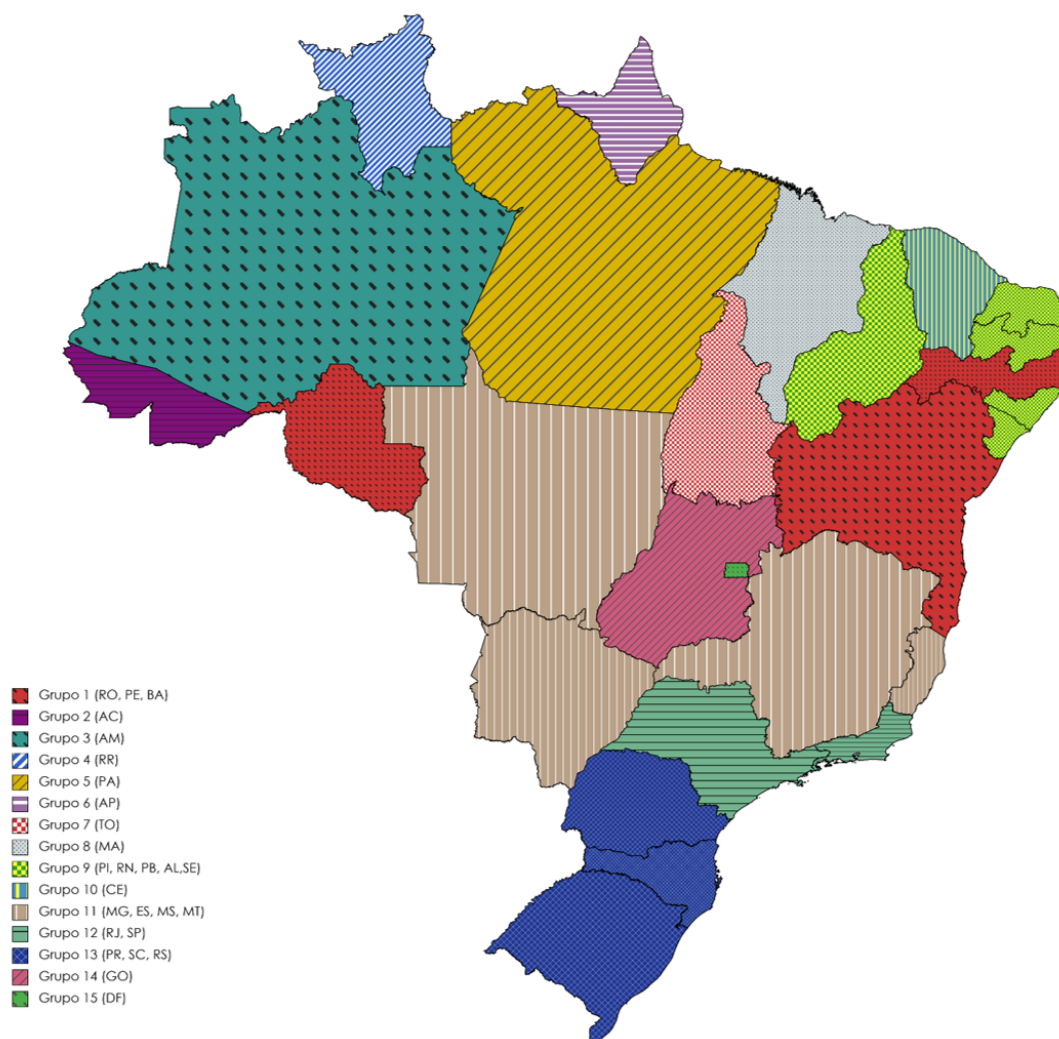
Tabela 28 – *Clusters* de Unidades Federativas do Brasil, segundo o método *k-means*.

Grupos	Soma de Quadrados	Unidades Federativas
$n_1 = 3$	8.608	Rondônia, Pernambuco, Bahia.
$n_2 = 1$	0.000	Acre.
$n_3 = 1$	0.000	Amazonas
$n_4 = 1$	0.000	Roraima.
$n_5 = 1$	0.000	Pará.
$n_6 = 1$	0.000	Amapá.
$n_7 = 1$	0.000	Tocantins.
$n_8 = 1$	0.000	Maranhão.
$n_9 = 5$	21.805	Piauí, Rio Grande do Norte, Paraíba, Alagoas, Sergipe.
$n_{10} = 1$	0.000	Ceará.
$n_{11} = 4$	15.145	Minas Gerais, Espírito Santo, Mato Grosso do Sul, Mato Grosso.
$n_{12} = 2$	6.597	Rio de Janeiro, São Paulo.
$n_{13} = 3$	7.085	Paraná, Santa Catarina, Rio Grande do Sul
$n_{14} = 1$	0.000	Goiás.
$n_{15} = 1$	0.000	Distrito Federal.

Fonte: Próprio autor

As mesmas informações contidas na Tabela 28, podem ser observadas também na Figura 15. No qual utilizamos o *site* (MAPCHART,)¹ para a criação do mapa do Brasil e a composição dos *clusters* das UFs do Brasil, segundo o *k-means*.

¹ Disponível em <https://www.mapchart.net/brazil.html>

Figura 15 – *Clusters* das Unidades Federativas do Brasil, segundo o *k-means*.

Fonte: Próprio autor

Assim como nos métodos hierárquicos estudados, o método *K-means*, mostra que as unidades federativas do Ceará, Distrito Federal e Roraima estão em *clusters* separados e que o Método da Média das Distâncias possui os mesmos *clusters* que no método *K-means*, indicando a ocorrência de resultados similares embora tenha se empregado métodos distintos.

A UF do PA foi onde houve a maior proporção de inscritos oriundos de dependência administrativa municipal, no MA houve a menor proporção de inscritos que declararam a ocupação da mãe 3 e 5, bem como foi a segunda maior proporção de inscritos que declararam a ocupação da mãe 1, além de ser a quarta maior proporção de inscritos que declararam ser pretos/pardos. No TO houve a segunda maior proporção de inscritos que declararam o grau de instrução da mãe 1. Nas UFs da PB, SE e RN, AL houve a maior proporção de inscritos que, declararam o grau de instrução da mãe 2 e declararam

ser oriundos de dependência administrativa federal, respectivamente. Desse modo, verificamos que todos os *clusters* formados pelo método *K-means* se relacionam com as UFs por meio das proporções das variáveis selecionadas.

Fazendo uma analogia dos *clusters* das unidades federativas formados no *k-means* com as cinco regiões brasileiras podemos verificar que: os *clusters* formados com as unidades federativas da região norte são todos diferentes, na região nordeste há um *cluster* com cinco unidades federativas (Piauí, Rio Grande do Norte, Paraíba, Alagoas e Sergipe), na região sudeste as unidades federativas do Rio de Janeiro e São Paulo estão no mesmo *cluster*, na região centro-oeste as unidades federativas do Mato Grosso e Mato Grosso Sul estão no mesmo *cluster* e na região sul o cluster formado coincide com as unidades federativas daquela região.

5 Considerações finais

Com o desenvolvimento deste trabalho, mais especificamente com o emprego de técnicas de agrupamentos da Estatística Multivariada, identificamos a existência de *clusters* de Unidades Federativas em nosso país, quando se considera características socioeconômicas e educacionais dos inscritos no Enem 2021.

A Análise de Agrupamentos, discutida no Capítulo 3, é um instrumento canônico utilizado para se investigar a existência de conglomerados de elementos de uma população caracterizados por variáveis. O exemplo básico, presente no citado capítulo, em que se avalia similaridades entre as provas de matemática do Enem, ilustrou a aplicabilidade das técnicas hierárquicas e não hierárquicas de agrupamentos.

Nossa investigação principal incidiu sobre a base de Microdados do Enem 2021. Inicialmente, selecionamos as variáveis de acordo com critérios estatísticos e com base em referenciais teóricos. Em seguida, fizemos o uso de técnicas hierárquicas (Método da Ligação Simples, Método da Ligação Completa e o Método da Média das Distâncias) e não-hierárquicas (Método *k-means*), implementadas com o auxílio do *Software R* por meio dos comandos `hcluster` e `kmeans`, possibilitando-nos a construção de dendogramas e a determinação de *clusters*.

No método hierárquico da média das distâncias podemos destacar os *clusters* formados por: Piauí, Rio Grande do Norte, Paraíba e Sergipe, estes possuem as maiores proporções de inscritos que declararam o grau de instrução da mãe 1, 2 e 3 e ocupação da mãe 1, assim como possuem as menores proporções de inscritos que declararam o grau de instrução da mãe e do pai 5 e ocupação da mãe 2. Já os dois *clusters* formados pelas UFs de Rondônia, Pernambuco e Bahia e UFs de Minas Gerais, Espírito Santo, Mato Grosso e Mato Grosso do Sul, são *clusters* cuja proporção de inscritos nas vinte e uma variáveis estudadas ficaram na parte intermediária das proporções, ou seja, não se destacam entre as maiores proporções e menores proporções de inscritos no Enem 2021. Por fim, enfatizamos a similaridade entre o Estado de Rondônia e algumas UFs da região Nordeste, bem como o posicionamento das UFs Minas Gerais e Espírito Santo das UFs em um agrupamento que não se encontram os Estados de Rio de Janeiro e São Paulo, ou seja, da região sudeste.

Na aplicação do *k-means* os resultados obtidos foram similares aos do método hierárquico da média das distâncias. Particularmente, destaquemos que as UFs do Rio de Janeiro e São Paulo formam um único *cluster*, assim como os Estados de Paraná, Santa Catarina e Rio Grande do Sul. Observa-se que os Estados de Rio de Janeiro e São Paulo possuem as maiores proporções de inscritos nas variáveis, alunos oriundos de dependência administrativa privada, grau de instrução da mãe e do pai 5 e ocupação

da mãe 5, enquanto a menor proporção foi na variável ocupação da mãe 1. Já a fusão entre as três UFs do Paraná, Santa Catarina e Rio Grande do Sul é uma consequência por possuir a maior proporção de inscritos de alunos que possuem acesso a internet e a menor proporção de inscritos nas variáveis, alunos autodeclarados pretos/pardos e indígenas e grau de instrução da mãe 1.

Vale ressaltar que, embora tenha-se empregado técnicas distintas, as UFs do Ceará e Distrito Federal formam *clusters* unitários isolados em ambas. Nota-se que o Ceará tem a maior proporção de inscritos nas variáveis, alunos oriundos de dependência administrativa estadual, grau de instrução da mãe 3 e 4, e é menor em proporção de inscritos em, alunos oriundos de dependência administrativa federal, grau de instrução da mãe 5 e 6 e o grau de instrução do pai 5. Por outro lado, o Distrito Federal possui a maior proporção de inscritos nas variáveis, necessidade especial declarada e grau de instrução da mãe 6, e é menor em proporção nas variáveis, alunos oriundos de dependência administrativa municipal e grau de instrução do pai 4.

Os resultados obtidos em nosso trabalho podem ser utilizados por gestores públicos para que possam planejar políticas públicas e executar ações, por meio de suas secretarias de educação ou desenvolvimento social, que visem melhorias nas condições socioeconômicas e educacionais dos estudantes, atuando de modo incisivo em fatores associados ao desempenho escolar, equiparando as condições civil, política e moral dos educandos e assegurando a estes condições igualitárias para o acesso ao ensino superior e inserção no mercado de trabalho.

Por fim, destaquemos duas linhas de ações pelas quais podemos expandir ou aprofundar nossa pesquisa, Primeiro, pode-se desenvolver uma Análise de Componentes Principais, com o propósito de investigar a estrutura da matriz de covariância das variáveis fixadas em nosso estudo, identificando quais variáveis têm maior peso na explicação da variabilidade da matriz, levando-nos, possivelmente, a uma redução no número de variáveis envolvidas e a construção de índices. Em um segundo momento, pode-se desenvolver uma análise similar a que realizada aqui, porém com a seleção de variáveis que refletem o desempenho escolar dos estudantes nas avaliações para caracterizar as Unidades Federativas, possibilitando-nos investigar associação entre os fatores socioeconômicos e o desempenho escolar dos estudantes em provas do Enem.

Referências

- ARISTIZABAL, G. C.; CAICEDO, M. C.; PARRA, J. C. M. Factores asociados a la adquisición de competencias en américa latina1. *Revista de ciencias sociales*, Facultad de Ciencias Sociales, v. 23, n. 4, p. 33–52, 2017. Citado 2 vezes nas páginas 15 e 64.
- BOURDIEU, P.; PASSERON, J. C. Les héritiers: les étudiants et la culture. (*No Title*), 1964. Citado na página 15.
- BRASIL, . Diário Oficial da U. *Portaria MEC Nº 438, de 28 de maio de 1998. Institui o Exame Nacional do Ensino Médio – ENEM*. 1998. Disponível em: <<https://www.legisweb.com.br/legislacao/?id=181748>>. Acesso em: 04 nov 2022. Citado na página 18.
- BRASIL, I.; INTERMEDIÁRIO, C. Instituto nacional de estudos e pesquisas educacionais anísio teixeira. *Exame Nacional do Ensino Médio*, 2012. Citado na página 18.
- BRASIL, I. N. d. E. e. P. E. A. T. I. Entenda a sua nota no enem: guia do participante. *Brasília, DF*, 2021. Citado na página 27.
- BRASIL, S. F. D. Constituição da república federativa do brasil. *Brasília: Senado Federal, Centro Gráfico*, 1988. Citado na página 15.
- BRAZ, M. M. A. et al. Políticas afirmativas no brasil: Análise do percurso de dez anos da lei 12.711/2012 (lei de cotas). *SciELO Preprints*, 2022. Citado 2 vezes nas páginas 18 e 64.
- BUSSAB, W. d. O.; MIAZAKI, É. S.; ANDRADE, D. F. d. Introdução à análise de agrupamento. In: *Introdução à análise de agrupamento*. [S.l.: s.n.], 1990. p. 105–105. Citado 8 vezes nas páginas 29, 31, 33, 39, 40, 44, 48 e 62.
- CO-OPERATION, O. for E.; DEVELOPMENT. *PISA 2012 results: Excellence through equity: Giving every student the chance to succeed (Volume II)*. [S.l.]: OECD publishing Paris, France, 2013. Citado na página 64.
- CORMACK, R. M. A review of classification. *Journal of the Royal Statistical Society: Series A (General)*, Wiley Online Library, v. 134, n. 3, p. 321–353, 1971. Citado 2 vezes nas páginas 44 e 49.
- FAILLACE, H. F. P.; BRITTO, I. L. M.; COSTA, F. da S. O desempenho dos alunos do ensino médio no enem 2019 e a desigualdade social: Um estudo utilizando análise exploratória e técnicas de agrupamento. *Cadernos do IME-Série Estatística*, v. 53, p. 33, 2022. Citado na página 16.
- FEIJÓ, J. R.; FRANÇA, J. M. S. d. Diferencial de desempenho entre jovens das escolas públicas e privadas. *Estudos Econômicos (São Paulo)*, SciELO Brasil, v. 51, p. 373–408, 2021. Citado 2 vezes nas páginas 15 e 64.

- HARTIGAN, J. A.; WONG, M. A. et al. A k-means clustering algorithm. *Applied statistics*, USA, v. 28, n. 1, p. 100–108, 1979. Citado na página 60.
- INEP. *Banco de Microdados Enem*. 2021. Disponível em: <<https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados/enem>>. Acesso em: 02 de fevereiro 2023. Citado 2 vezes nas páginas 28 e 63.
- JALOTO, A.; PRIMI, R. Fatores socioeconômicos associados ao desempenho no enem. *Em Aberto*, v. 34, n. 112, 2021. Citado 2 vezes nas páginas 15 e 64.
- JOHNSON, R. A.; WICHERN, D. W. et al. *Applied multivariate statistical analysis*. Prentice hall Upper Saddle River, NJ, 2002. Citado na página 29.
- KARINO, C.; BARBOSA, M. Nota técnica: procedimento de cálculo das notas do enem. *Brasília, DF: Inep*, 2012. Citado na página 27.
- KARINO, C. A.; LAROS, J. A. Estudos brasileiros sobre eficácia escolar: uma revisão de literatura. *Examen: Política, Gestão e Avaliação da Educação*, v. 1, n. 1, p. 32–32, 2017. Citado 2 vezes nas páginas 15 e 64.
- LIMA, A. et al. Analysis of enem's attendants between 2012 and 2017 using a clustering approach. *Journal of Information and Data Management*, v. 11, n. 2, 2020. Citado na página 16.
- MAIA, M. M.; ANDRADE, L. H. F. de; FERNANDES, S. K-means na análise de características socioeconômicas de candidatos ao ensino superior. *Anais do Encontro de Computação do Oeste Potiguar ECOP/UFERSA (ISSN 2526-7574)*, n. 5, 2021. Citado na página 16.
- MAPCHART. *Brazilian states map*. Disponível em: <<https://www.mapchart.net/brazil.html>>. Acesso em: 20 mai 2023. Citado na página 73.
- MELO, R. O. et al. Impacto das variáveis socioeconômicas no desempenho do enem: uma análise espacial e sociológica. *Revista de Administração Pública*, SciELO Brasil, v. 55, p. 1271–1294, 2022. Citado na página 16.
- MINGOTI, S. A. Análise de dados através de métodos estatística multivariada: uma abordagem aplicada. In: *Análise de dados através de métodos estatística multivariada: uma abordagem aplicada*. [S.l.: s.n.], 2013. Citado 4 vezes nas páginas 29, 60, 64 e 73.
- NASCIMENTO, M. M.; CAVALCANTI, C.; OSTERMANN, F. Análise de correspondência aplicada à pesquisa em ensino de ciências. *Enseñanza de las ciencias*, n. Extra, p. 1319–1324, 2017. Citado na página 16.
- RABELO, M. Avaliação educacional: fundamentos, metodologia e aplicações no contexto brasileiro. *Rio de janeiro: SBM*, v. 29, p. 30–31, 2013. Citado 2 vezes nas páginas 26 e 27.
- RENCHER, A. C.; CHRISTENSEN, W. F. *Methods of multivariate analysis*. Prentice hall Upper Saddle River, NJ, 2012. Citado na página 29.
- ROMESBURG, H. *Cluster analysis for researchers. Lifetime learning Pubi*. [S.l.]: Belmont California, 1984. Citado 2 vezes nas páginas 9 e 48.

SANTANA, I. J. *Análise do questionário socioeconômico do Enem: um olhar acerca do capital cultural como herança*. 2020. Disponível em: <<https://even3.blob.core.windows.net/processos/f12bf74bec94465084b5.pdf>>. Acesso em: 09 jul 2023. Citado 3 vezes nas páginas 15, 28 e 64.

SENKEVICS, A. S.; MELLO, U. M. O perfil discente das universidades federais mudou pós-lei de cotas? *Cadernos de Pesquisa*, Fundação Carlos Chagas, v. 49, n. 172, p. 184–208, 2019. Citado na página 64.

SIRIN, S. R. Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of educational research*, Sage Publications Sage CA: Thousand Oaks, CA, v. 75, n. 3, p. 417–453, 2005. Citado 2 vezes nas páginas 15 e 64.

TRAVITZKI, R.; FERRÃO, M. E.; COUTO, A. P. Desigualdades educacionais e socioeconômicas na população brasileira pré-universitária: Uma visão a partir da análise de dados do enem. *Education Policy Analysis Archives*, v. 24, p. 74–74, 2016. Citado na página 64.

Anexos

ANEXO A –

Tabela 29 – Proporção de inscritos por Unidade Federativa do Brasil segundo características socioeconômicas e educacionais dos inscritos no Enem 2021. (Parte 1)

UF	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₁₀	X ₁₁
RO	0.664042	0.010056	0.078714	0.002355	0.002355	0.067210	0.008238	0.036497	0.158226	0.118700	0.105161
AC	0.739469	0.007578	0.065608	0.476171	0.001238	0.076130	0.011317	0.065507	0.165512	0.090234	0.104043
AM	0.759364	0.037765	0.027792	0.713238	0.006268	0.041983	0.005918	0.044072	0.140600	0.109542	0.118369
RR	0.697487	0.046293	0.108429	0.531506	0.000409	0.094517	0.009160	0.028469	0.096670	0.075257	0.085654
PA	0.768289	0.005282	0.032109	0.406672	0.031044	0.153390	0.007738	0.035837	0.168910	0.111465	0.127144
AP	0.754639	0.007062	0.009775	0.462146	0.000832	0.123336	0.009093	0.060417	0.145943	0.087141	0.094849
TO	0.708018	0.007344	0.052577	0.473758	0.001310	0.080831	0.008743	0.025657	0.111019	0.088129	0.094169
MA	0.752921	0.003487	0.070869	0.454079	0.002476	0.092511	0.006133	0.038933	0.138663	0.111203	0.127322
PI	0.737626	0.003794	0.063750	0.401526	0.003526	0.174732	0.007675	0.048769	0.179686	0.125972	0.125171
CE	0.707408	0.004491	0.009696	0.761174	0.000738	0.065205	0.009773	0.044727	0.190713	0.151580	0.154707
RN	0.557835	0.003233	0.137996	0.339005	0.001070	0.186898	0.010345	0.037400	0.173375	0.133736	0.125393
PB	0.612391	0.009496	0.059936	0.420208	0.001609	0.164558	0.009370	0.045009	0.199271	0.140003	0.116683
PE	0.611815	0.013002	0.026797	0.501451	0.000617	0.136902	0.008476	0.034249	0.160329	0.128687	0.111384
AL	0.661637	0.007232	0.108644	0.375723	0.000591	0.220441	0.008411	0.064704	0.195238	0.124950	0.105926
SE	0.718555	0.004927	0.024188	0.388520	0.000578	0.198195	0.008746	0.048407	0.201608	0.135729	0.118658
BA	0.753114	0.006235	0.037359	0.368299	0.003673	0.127240	0.007284	0.039111	0.161989	0.112534	0.106362
MG	0.514679	0.002581	0.061000	0.404247	0.006571	0.193256	0.012114	0.019113	0.138472	0.127761	0.109587
ES	0.568895	0.004565	0.086184	0.537513	0.001520	0.138242	0.011718	0.024386	0.132477	0.122708	0.117082
RJ	0.497381	0.002455	0.056921	0.357025	0.007703	0.311671	0.010576	0.018735	0.107796	0.103336	0.122681
SP	0.356650	0.002660	0.011510	0.418750	0.015983	0.309747	0.007782	0.017150	0.094883	0.092086	0.105925
PR	0.280921	0.001585	0.027301	0.455748	0.001180	0.238505	0.008354	0.016208	0.109281	0.102574	0.120165
SC	0.212949	0.002403	0.071458	0.467830	0.002951	0.226031	0.007524	0.010681	0.106889	0.106525	0.117684
RS	0.190845	0.001518	0.057627	0.473732	0.006922	0.183384	0.008616	0.013109	0.125988	0.127625	0.127115
MS	0.467552	0.022397	0.067198	0.476860	0.002026	0.154007	0.006984	0.026801	0.122925	0.104861	0.113858
MT	0.606400	0.006468	0.090191	0.458406	0.002465	0.160887	0.008152	0.026374	0.110028	0.090749	0.108272
GO	0.570321	0.003658	0.022206	0.569564	0.001849	0.112477	0.012354	0.023631	0.118471	0.123719	0.136875
DF	0.558784	0.002977	0.038290	0.467145	0.000194	0.292827	0.015248	0.026882	0.111280	0.093526	0.096123

Próprio autor

Tabela 30 – Proporção de inscritos por Unidade Federativa do Brasil segundo características socioeconômicas e educacionais dos inscritos no Enem 2021. (Parte 2)

UF	X ₁₂	X ₁₃	X ₁₄	X ₁₅	X ₁₆	X ₁₇	X ₁₈	X ₁₉	X ₂₀	X ₂₁
RO	0.314878	0.239339	0.101345	0.245850	0.171614	0.422310	0.054277	0.040162	0.012176	0.909256
AC	0.296525	0.248118	0.091530	0.227130	0.161374	0.443940	0.046014	0.033102	0.027818	0.736178
AM	0.369755	0.167227	0.101702	0.293547	0.204248	0.406254	0.072943	0.030880	0.059773	0.681668
RR	0.375913	0.313034	0.090853	0.305360	0.113133	0.393118	0.045922	0.055452	0.036143	0.874489
PA	0.358940	0.169332	0.114226	0.263989	0.211794	0.435676	0.049509	0.026547	0.033894	0.729691
AP	0.339657	0.240515	0.089126	0.274578	0.142758	0.460399	0.045786	0.035909	0.072695	0.784686
TO	0.350607	0.304095	0.096999	0.292141	0.201882	0.363006	0.041712	0.047307	0.014497	0.856870
MA	0.370658	0.188220	0.113569	0.287810	0.307036	0.351837	0.039159	0.024775	0.026970	0.769056
PI	0.296105	0.198505	0.109194	0.210450	0.317478	0.348542	0.042383	0.027909	0.018005	0.819813
CE	0.251801	0.137420	0.127723	0.207268	0.278603	0.375363	0.070815	0.025551	0.013302	0.815445
RN	0.321372	0.189961	0.104318	0.261206	0.230271	0.373506	0.055951	0.038992	0.012061	0.919492
PB	0.291849	0.185021	0.094645	0.224229	0.280345	0.363659	0.043040	0.039615	0.012562	0.891659
PE	0.352951	0.186181	0.101975	0.281123	0.220834	0.391411	0.052186	0.038492	0.012540	0.909563
AL	0.284845	0.199088	0.093960	0.238142	0.209320	0.394707	0.042731	0.042714	0.015833	0.888993
SE	0.297848	0.172640	0.101663	0.248563	0.270351	0.378086	0.045886	0.034123	0.014703	0.847009
BA	0.363548	0.191305	0.101794	0.282143	0.209736	0.413787	0.047733	0.032260	0.013127	0.895797
MG	0.325029	0.260236	0.114788	0.272190	0.098691	0.435545	0.065546	0.053638	0.007301	0.942838
ES	0.332736	0.246288	0.113670	0.296804	0.111037	0.430840	0.063403	0.049410	0.006529	0.945931
RJ	0.380758	0.243966	0.109318	0.333611	0.028168	0.479815	0.058858	0.077316	0.007470	0.936505
SP	0.368564	0.304553	0.107482	0.329468	0.041881	0.427054	0.077162	0.077883	0.007262	0.965245
PR	0.333194	0.297692	0.119908	0.307652	0.081493	0.405499	0.083189	0.063547	0.004824	0.968007
SC	0.336885	0.303432	0.117923	0.303281	0.085149	0.364451	0.106713	0.068441	0.004944	0.972762
RS	0.339009	0.249253	0.121846	0.298866	0.088669	0.425126	0.083798	0.059323	0.005236	0.955105
MS	0.321235	0.288823	0.106849	0.279518	0.080261	0.449487	0.057201	0.056893	0.009139	0.915927
MT	0.327254	0.314568	0.110351	0.268629	0.099475	0.429004	0.056098	0.059144	0.009765	0.910791
GO	0.331724	0.228491	0.128482	0.264902	0.091341	0.471295	0.076747	0.044232	0.009318	0.915396
DF	0.314126	0.336315	0.087481	0.255137	0.064517	0.435642	0.044697	0.093283	0.015248	0.946129

ANEXO B –

Figura 16 – Rotina do R para o Exemplo Básico.

```

ROTINA "EXEMPLO BÁSICO"

provas<-matrix(c(240.127,225.300,203.002,167.010,182.835,174.715,
               188.607,187.260,149.414,142.972,170.745,135.916,
               0.150,0.159,0.160,0.164,0.152,0.175),nrow=6,ncol=3,
              dimnames = list(c("A_19", "B_19","C_20", "D_20","E_21",
                                "F_21"), c("X_1", "X_2", "X_3")))

#Métodos Hierárquicos

#Gráfico de levelplot

vec.medias<-matrix(c(rep(mean(provas[,1]),6),rep(mean(provas[,2]),6),
                  rep(mean(provas[,3]),6)), nrow=6,ncol=3)

invs<-matrix(c(1/sd(provas[,1]),0,0,0,1/sd(provas[,2]),0,0,0,1/sd(provas[,3])),
            nrow=3,ncol=3)

provasz<-(provas-vec.medias)%%invs

#Gráfico 3d

require(scatterplot3d)
library(scatterplot3d)
x<-provasz[,1]
y<-provasz[,2]
z<-provasz[,3]
scatterplot3d(x, y, z, highlight.3d=TRUE, col.axis="black",
              col.grid="lightblue", pch=19, xlab="z1",
              yla="z2",zlab="z3")

dis<-dist(provasz)
disp<-(1/sqrt(3))*dist(provasz)

##Entrada da função que calcula a distância padronizada

dist2full<-function(dis){
  n<-attr(dis,"size")
  full<-matrix(0,n,n)
  full[lower.tri(full)]<-dis
  full+t(full)}

dis.matrix<-dist2full(dis)

dist2fullp<-function(disp){
  n<-attr(disp,"size")
  full<-matrix(0,n,n)
  full[lower.tri(full)]<-disp
  full+t(full)}

disp.matrix<-dist2fullp(disp)

image(t(disp.matrix),axes=FALSE, col = hcl.colors(12, "y1orRd",
          rev = TRUE), oldstyle = TRUE)

hcs<-hclust((1/sqrt(3))*dist(provasz),method="single",members = NULL)
hcc<-hclust((1/sqrt(3))*dist(provasz),method='complete')
hca<-hclust((1/sqrt(3))*dist(provasz),method='average')
hcce<-hclust((1/sqrt(3))*dist(provasz),method='centroid')

```

Figura 17 – Rotina do R para o Exemplo Básico.

```
agrupamento=hclust((1/sqrt(3))*dist(provasz),method='single')
agrupamento1=hclust((1/sqrt(3))*dist(provasz),method='complete')
agrupamento2=hclust((1/sqrt(3))*dist(provasz),method='average')
agrupamento3=hclust((1/sqrt(3))*dist(provasz),method='centroid')

#Dendograma- Método da Ligacao Simples

plot(as.dendrogram(agrupamento), xlab = "Provas", horiz = FALSE,
     frame.plot = TRUE, main="Método da Ligação Simples")

#Dendograma - Método da Ligacao Completa

plot(as.dendrogram(agrupamento1), xlab = "Provas", horiz = FALSE,
     frame.plot = TRUE, main="Método da Ligação Completa")

#Dendograma - Método da Média das Distâncias

plot(as.dendrogram(agrupamento2), xlab = "Provas", horiz = FALSE,
     frame.plot = TRUE, main = "Método das Médias das Distâncias")

#Dendograma - Método da Centróide

plot(as.dendrogram(agrupamento3), xlab = "Provas", horiz = FALSE,
     frame.plot = TRUE, main="Método da Centróide")

#Método Não-Hierárquico

"k-means"

set.seed(2023)

agrupamentoskm=kmeans(provasz, centers =3)

agrupamentoskm
```