



**Universidade Federal de Uberlândia
Faculdade de Matemática
Bacharelado em Estatística**

Douglas Silva Oliveira

Fuzzy Clustering Multi-Critério para dados relacionados à genética

Uberlândia

2016

Douglas Silva Oliveira

FUZZY CLUSTERING MULTI-CRITÉRIO PARA DADOS RELACIONADOS À GENÉTICA

Monografia submetida à Faculdade de Matemática e à Coordenadoria do Curso de Bacharelado em Estatística da Universidade Federal de Uberlândia - Campus Santa Mônica, como requisito parcial para obtenção do grau de Bacharel em Estatística.

Área de pesquisa: Estatística Multivariada

Orientador: Prof^a. Dr^a. Priscila Neves Faria

Uberlândia
2016



Universidade Federal de Uberlândia

Faculdade de Matemática

Douglas Silva Oliveira

Esta Monografia foi julgada adequada para a obtenção do Grau de Bacharel em Estatística, sendo aprovada pela Faculdade de Matemática do Campus Santa Mônica da Universidade Federal de Uberlândia e pela banca examinadora:

Orientador: Prof^a. Dr^a. Priscila Neves Faria
Faculdade de Matemática - FAMAT/UFU

Prof. Dr. Lucio Borges de Araújo
Faculdade de Matemática - FAMAT/UFU

Prof. Dr. Edmilson Rodrigues Pinto
Faculdade de Matemática - FAMAT/UFU

Uberlândia, 8 de Dezembro de 2016

Dedico este trabalho primeiramente à Deus, pois sem ele eu não teria forças para essa longa jornada, e também aos meus professores, amigos e familiares que me ajudaram na conclusão desse trabalho.

Agradecimentos

Agradeço primeiramente à Deus que permitiu que tudo isso acontecesse ao longo da minha vida, não somente nestes anos como universitário, mas em todos os momentos.

A esta universidade, seu corpo docente, direção e administração que oportunizaram momentos de aprendizagem, descobrimento e amizades.

A todos os professores, em especial a minha orientadora Prof^a. Dr^a. Priscila Neves Faria, por me proporcionarem o conhecimento não apenas racional, mas também a manifestação do caráter profissional e social.

Aos meus pais Hamilton e Edileuza, ao meu irmão Lucas e a minha noiva Natalia, pelo amor, incentivo e apoio incondicional.

Aos meus amigos, companheiros de trabalho e irmãos na amizade que fizeram parte da minha formação e que continuarão presentes em minha vida com certeza.

A todos que direta ou indiretamente fizeram parte da minha formação, o meu muito obrigado.

“Sonhos determinam o que você quer. Ação determina o que você conquista”.

Aldo Novak

Resumo

Estudos a partir de dados genéticos são importantes para o conhecimento da variabilidade genética das populações, além de fornecer parâmetros para escolha de progenitores. Nos estudos para se analisar o quão distante geneticamente um genótipo é de outro são utilizados métodos da estatística multivariada, permitindo unificar múltiplas informações de um conjunto de caracteres. Assim, o presente trabalho tem como objetivo a utilização da metodologia *fuzzy c-means* aplicada a dados genéticos de uma determinada espécie, o que determinou distintos grupos genéticos com características distintas intergrupos. Para tal estudo foi utilizado a linguagem *R* com auxílio da biblioteca *fclust*.

Palavras-chave: dados genéticos, estatística multivariada, *fuzzy c-means*

Abstract

Studies based on genetic data are important for the knowledge of the genetic variability of the populations, besides providing parameters for parent choice. In the studies to analyze how far genetically one genotype is from another are used multivariate statistical methods, allowing to unify multiple information of a set of characters. Thus, the present work aims to use the *fuzzy c-means* methodology applied to genetic data of a given species, which determined different genetic groups with different intergroup characteristics. For this study, the *R* language was used with the help of the *fclust* library.

Keywords: genetic data, multivariate statistic, *fuzzy c-means*

Sumário

Lista de Figuras

Lista de Tabelas

1	Introdução	12
1.1	Estrutura do Trabalho	14
2	Revisão Literária	15
2.1	Métodos Hierárquicos	15
2.1.1	Métodos Aglomerativos	16
2.1.2	Métodos Divisivos	18
2.2	Métodos Não-Hierárquicos	19
2.3	Vantagens e Desvantagens	20
3	Materiais e Métodos	21
3.1	Base de Dados	21
3.2	Método <i>fuzzy c-means</i>	22
4	Resultados e Discussão	28
4.1	Análise Descritiva	28
4.2	Análise de agrupamento pelo método <i>fuzzy c-means</i>	32
5	Considerações Finais	39
	Referências Bibliográficas	40

Apêndice A – KNIME <i>Workflow</i> para tratamento e análise dos dados	42
Apêndice B – Algoritmo <i>R</i> para a análise descritiva	43
Apêndice C – Algoritmo <i>R</i> para descrição dos <i>clusters</i>	46

Lista de Figuras

1	Exemplo do algoritmo de clusterização pelos métodos hierárquicos aglomerativos e divisivos.	16
2	Distribuição das observações em relação à característica Altura da Planta (m)	30
3	Distribuição das observações em relação à característica Altura da Inserção da Espiga (m)	30
4	Distribuição das observações em relação à característica Grãos Ardidos (%)	31
5	Distribuição das observações em relação à característica Rendimento (Mg/ha)	31
6	Distribuição das observações em relação à característica Rendimento Real	31
7	Distribuição das observações em relação à característica Peso do Saco (kg)	31
8	Boxplot da característica Altura da Planta (m) por <i>cluster</i>	34
9	Boxplot da característica Altura de Inserção da Espiga (m) por <i>cluster</i>	34
10	Boxplot da característica Grãos Ardidos (%) por <i>cluster</i>	34
11	Boxplot da característica Rendimento (Mg/ha) por <i>cluster</i>	34
12	Boxplot da característica Rendimento Real por <i>cluster</i>	35
13	Boxplot da característica Peso do Saco (kg) por <i>cluster</i>	35
14	Boxplot do grau de pertinência dos indivíduos do <i>cluster 1</i>	36
15	Boxplot do grau de pertinência dos indivíduos do <i>cluster 2</i>	36
16	Boxplot do grau de pertinência dos indivíduos do <i>cluster 3</i>	36
17	Gráfico de barras empilhado da distribuição das UFs em relação aos <i>clusters</i>	37

Lista de Tabelas

1	Caracterização das variáveis estudadas.	21
2	Tabela de distribuição de frequências em relação ao <i>Gênero</i> do indivíduo observado.	28
3	Tabela de distribuição de frequências em relação ao <i>Município</i> de cultivo do indivíduo observado.	29
4	Tabela de distribuição de frequências em relação à <i>UF</i> de cultivo do indivíduo observado.	29
5	Estatísticas descritivas à respeito dos dados analisados.	30
6	Coeficientes de assimetria e curtose para as características estudadas.	31
7	Coeficiente de Correlação de Pearson entre as características estudadas.	32
8	Distribuição dos indivíduos entre os <i>clusters</i>	33
9	Descrição das características estudadas no <i>cluster 1</i>	33
10	Descrição das características estudadas no <i>cluster 2</i>	33
11	Descrição das características estudadas no <i>cluster 3</i>	34
12	Distribuição das UFs em relação aos <i>clusters</i>	37
13	Distribuição dos municípios em relação ao <i>clusters 1</i>	38
14	Distribuição dos municípios em relação ao <i>clusters 2</i>	38
15	Distribuição dos municípios em relação ao <i>clusters 3</i>	38

CAPÍTULO 1

INTRODUÇÃO

A cultura do milho abrangeu mais de 15,5 milhões de hectares na safra de 2013/2014 em território brasileiro, com produtividade média de 5,18 Mg/ha (ton/ha), sendo que, nos últimos anos, têm-se observado relevantes incrementos na produtividade de grãos dessa cultura. Isso se deve a mudanças tecnológicas recentes, tais como avanços no melhoramento genético, qualificação das adubações de base e de cobertura, modernização da mecanização agrícola, uso de irrigação e aprimoramento na gestão dos recursos agrícolas pela adoção da agricultura de precisão (VIAN *et al.*, 2016).

O potencial de produtividade de grãos de milho é determinado pela densidade de semeadura, população final de plantas, número de espigas por planta, comprimento de espigas, número de fileiras de grãos por espiga, número de grãos por fileira, massa média do grão e pelo genótipo utilizado (VIAN *et al.*, 2016).

Estudos a partir de dados genéticos são importantes para o conhecimento da variabilidade genética das populações, além de fornecer parâmetros para escolha de progenitores. Para que estes estudos sejam realizados, é necessário muitas vezes o acesso aos dados de algum banco de germoplasma, acesso esse muitas vezes difícil de obter. Estima-se que 0,5% dos acessos de um banco devem ser requisitados por ano considerando que este seria o uso mínimo desejável (BROWN, 1989). Porém, isto dificilmente acontece e os fatores que dificultam o uso do banco de germoplasma são a não disponibilidade dos acessos aos melhoristas e a falta de conhecimento sobre os acessos conservados (LI *et al.*, 1998). Além disso, um banco de germoplasma não pode ser como um museu, devendo-se estimular a avaliação dos acessos por meio de análises genéticas (HAMMER, 2003).

Nos estudos para se analisar o quão distante geneticamente um genótipo é de outro são utilizados métodos da estatística multivariada, permitindo unificar múltiplas informações de um conjunto de caracteres. A tarefa de agrupar ou classificar objetos em categorias é uma das atividades mais comuns e primitivas do homem e vem sendo intensificada em função do grande volume de informações disponíveis atualmente, sobre as mais diversas áreas (BACKER, 1995). Para realizar essa tarefa, pode ser empregado um mecanismo chamado análise de *cluster*, ou

clusterização, o qual é definido como um processo pelo qual procura-se classificar objetos em grupos.

O problema de clusterização pode ser tratado segundo diferentes abordagens, entre elas, a abordagem com métodos hierárquicos e a abordagem não hierárquica. Em muitas aplicações práticas, no entanto, uma clusterização hierárquica clássica pode ser muito restritiva e inviável, devido, muitas vezes, à imprecisão ou à não completeza dos dados. A imprecisão nos dados pode surgir de diversas fontes como, por exemplo, erros em medições ou ruídos no experimento que podem levar a valores parcialmente confiáveis de determinados atributos (BEZDEK; PAL, 1994). Em outros casos, pode ser difícil decidir quais são os atributos mais relevantes que caracterizam um indivíduo.

Nas técnicas de classificação convencionais, geralmente assume-se que um elemento pertence a apenas uma classe, o que nem sempre se verifica em domínios de dados reais, nos quais elementos pertencem a mais de uma classe, com diferentes graus. Por essas razões, torna-se conveniente o uso de uma metodologia com a propriedade de descrever os valores de atributos, em vez de tentar fornecer uma representação numérica exata para os dados com valores incertos dos atributos. No tipo de abordagem não hierárquica, temos essa propriedade na metodologia *fuzzy*, mais flexível, na qual um objeto pode ser classificado em várias categorias, com diferentes graus de associação a cada uma delas.

O uso da teoria de conjuntos *fuzzy* é conveniente uma vez que grande parte das categorias comumente encontradas e empregadas na clusterização possui limites vagos. No caso específico da clusterização *fuzzy* (ou *fuzzy clustering*), diferentes métodos podem ser utilizados, por exemplo, o algoritmo *fuzzy c-means*, desenvolvido por J. C. Bezdek e descrito neste trabalho.

Uma relevante propriedade da modelagem *fuzzy* é a capacidade de codificar conhecimentos inexatos, de tal forma que se aproxime dos processos decisórios (RUHOFF *et al.*, 2005). Os sistemas de inferência *Fuzzy* proporcionam a apreensão do conhecimento próximo ao modelo “cognitivo” muito aplicado na análise de problemas de previsão e classificação. Logo, o processo de obtenção do conhecimento é simplificado, mais preciso e com menor probabilidade de erros.

Considerando a importância da utilização de técnicas multivariadas em dados genéticos, justifica-se a utilização do método *fuzzy c-means* em uma base de germoplasma de determinada espécie de milho.

1.1 Estrutura do Trabalho

Esse trabalho está organizado da seguinte maneira: Inicialmente (capítulo 2) é realizada uma revisão literária abordando os principais métodos de agrupamento.

Em seguida (capítulo 3) é feita uma breve abordagem da base de dados utilizada nesse trabalho e um estudo mais aprofundado a respeito da metodologia utilizada para a segmentação dos indivíduos. Logo após esse estudo (capítulo 4) são apresentados os resultados obtidos tanto da base de dados (análise descritiva) quanto da metodologia utilizada (análise de segmentação).

Por fim (capítulo 5), são apresentadas as considerações finais obtidas durante e após o processo de agrupamento dos dados.

REVISÃO LITERÁRIA

Nesse capítulo, é realizada uma revisão a respeito das principais metodologias de clusterização. Inicialmente são mencionados os métodos hierárquicos, discorrendo sobre os métodos aglomerativos e divisivos. Por fim, é realizada uma breve discussão sobre os métodos não-hierárquicos, deixando para o próximo capítulo um foco especial para o método *fuzzy c-means*, visto que essa é a metodologia utilizada no trabalho em questão.

2.1 Métodos Hierárquicos

Os métodos hierárquicos de *cluster* consistem em uma série de sucessivos agrupamentos ou sucessivas divisões de elementos, em que tais elementos são agregados ou desagregados. Os métodos hierárquicos são subdivididos em métodos *aglomerativos* e *divisivos* (BEZDEK; EHRLICH; FULL, 1984).

Os grupos, nos métodos hierárquicos, são geralmente representados por um histórico de todo o processo de agrupamento hierárquico chamado de *dendograma* ou *diagrama de árvore*. Neste diagrama, cada ramo representa um elemento, enquanto a raiz representa o agrupamento de todos os elementos. Na Figura 1, são ilustrados os modos de funcionamento de um algoritmo aglomerativo e de um algoritmo divisivo com base em um dendograma. Observe que k representa o número do passo atual do algoritmo.

Para o algoritmo aglomerativo, a medida em que k aumenta, um único *cluster* vai sendo formado. De modo contrário, para o algoritmo divisivo a medida em que k aumenta, vários *clusters* vão sendo formados.

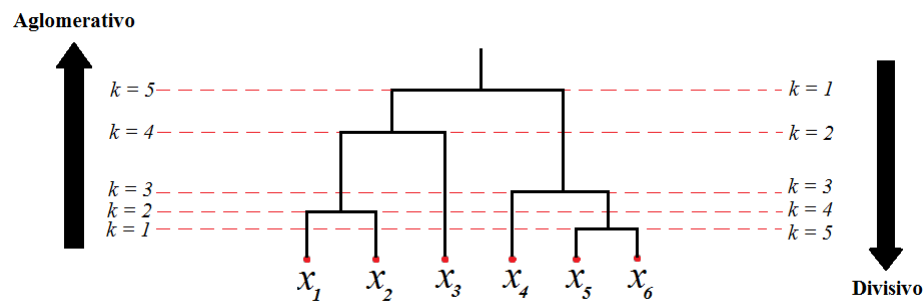


Figura 1: Exemplo do algoritmo de clusterização pelos métodos hierárquicos aglomerativos e divisivos.

2.1.1 Métodos Aglomerativos

No método aglomerativo, cada elemento inicia-se representando um grupo, e a cada iteração, um grupo ou elemento é ligado a outro de acordo com sua similaridade. Tal processo continua até que um único grupo contendo todos os elementos seja formado.

Existe uma variedade de métodos aglomerativos, que são caracterizados de acordo com o critério utilizado para definir as distâncias entre grupos. No entanto, a maioria dos métodos são formulações alternativas de três grandes conceitos de agrupamento aglomerativo (ANDERBERG, 2014):

1. Métodos de ligação (Vizinho Mais Próximo, Vizinho Mais Distante, Ligação pela Média, Ligação pela Mediana);
2. Método da Centroide;
3. Método de *Ward*.

Método do Vizinho Mais Próximo (*Single Linkage*)

O método de ligação por vizinho mais próximo, emprega a distância de valor mínimo. Algumas características desse método são: (ANDERBERG, 2014)

- Em geral, grupos muito próximos podem não ser identificados;
- Permite detectar grupos de formas não-elípticas;
- Apresenta pouca tolerância a ruído, pois tem tendência a incorporar os ruídos em um grupo já existente;
- Apresenta bons resultados tanto para distâncias Euclidianas quanto para outras distâncias;

- Tende a formar longas cadeias, ou seja, forma grupos com muitos indivíduos em relação à outros grupos.

Método do Vizinho Mais Distante (*Complete Linkage*)

O método de ligação por vizinho mais distante, emprega a distância máxima entre dois *clusters*. Algumas características desse método são: (KAUFMAN; ROUSSEUW, 2009)

- Apresenta bons resultados tanto para distâncias Euclidianas quanto para outras distâncias;
- Tendência a formar grupos compactos;
- Os ruídos demoram a serem incorporados ao grupo.

Os métodos de *ligação por vizinho mais próximo* e de *ligação por vizinho mais distante* trabalham em direções opostas. Contudo, se eles apresentam resultados semelhantes, significa que o grupo está bem definido no espaço, ou seja, o grupo existe. Porém, se não ocorre tal semelhança nos resultados, os grupos provavelmente não existem (ROMESBURG, 2004).

Método de Ligação pela Média (*Average Linkage*)

O método de ligação pela média, apresenta características como: (KAUFMAN; ROUSSEUW, 2009)

- Menor sensibilidade à ruídos que o os métodos de ligação por vizinho mais próximo e por vizinho mais distante;
- Apresenta bons resultados tanto para distâncias Euclidianas quanto para outras distâncias;
- Tendência a formar grupos com número de elementos similares.

Método de Ligação pela Mediana (*Median Linkage*)

O método de ligação pela mediana, apresenta características como:

- Resultado satisfatório quando os grupos possuem tamanhos diferentes;
- Pode apresentar resultado diferente quando permutado os elementos na matriz de similaridade;
- Robustez à presença de *outliers*.

Método da Centroide

No método da centroide, têm-se como principais características:

- Robustez à presença de ruídos;
- Devido ao fenômeno da reversão, o método não é muito utilizado.

O fenômeno da reversão ocorre quando a distância entre centroides é menor que a distância entre grupos já formados, isso fará com que os novos grupos se formem a um nível inferior aos grupos existentes, tornando o dendograma confuso.

Método de Ward

No método de *Ward*, têm-se como principais características:

- Apresenta bons resultados tanto para distâncias Euclidianas quanto para outras distâncias;
- Pode apresentar resultados insatisfatórios quando o número de elementos em cada grupo é praticamente igual;
- Tem tendência a combinar grupos com poucos elementos;
- Sensível à presença de *outliers*.

2.1.2 Métodos Divisivos

Os métodos divisivos trabalham na direção oposta dos métodos aglomerativos, ou seja, um grupo inicial contendo todos os elementos é dividido em dois subgrupos, de tal maneira que os elementos em um subgrupo estejam distantes dos elementos do outro subgrupo. Esses subgrupos são divididos em outros subgrupos dissimilares e o processo continua até cada elemento formar um grupo.

Os métodos divisivos são pouco mencionados na literatura, pois exigem uma maior capacidade computacional que os métodos aglomerativos (KAUFMAN; ROUSSEEUW, 2009).

Comparando apenas o primeiro passo dos métodos aglomerativos e divisivos, notamos que o método divisivo exige um maior número de iterações. No primeiro passo do método aglomerativo são consideradas todas as uniões possíveis de dois elementos, tendo a complexidade de tempo da ordem N^2 , sendo N o número de elementos. Este número cresce

quadraticamente à medida que N aumenta, ou seja, seu crescimento é acelerado. Mesmo assim a implementação do algoritmo é computacionalmente viável (DONI, 2004).

Já o método divisivo, baseado no mesmo princípio, começará considerando todas as divisões dos elementos em dois grupos, com pelo menos um elemento em cada grupo, tendo a complexidade de tempo da ordem de 2^N . Este número cresce exponencialmente à medida que N aumenta, dessa forma, para um grande número de elementos, torna-se inviável sua implementação computacional (DONI, 2004).

2.2 Métodos Não-Hierárquicos

Os métodos não-hierárquicos, também conhecidos como *métodos por particionamento*, foram desenvolvidos para agrupar elementos em K grupos, onde K é a quantidade de grupos definida previamente pelo usuário.

Nem todos valores de K apresentam grupos satisfatórios, sendo assim, aplica-se o método várias vezes para diferentes valores de K , escolhendo os resultados que apresentem melhor interpretação dos grupos ou uma melhor representação gráfica (BUSSAB, 1990).

A ideia central da maioria dos métodos por particionamento é escolher uma partição inicial dos elementos e, em seguida, alterar os membros dos grupos para obter-se a melhor partição (ANDERBERG, 2014).

Quando comparado com o método hierárquico, o método por particionamento é mais rápido porque não é necessário calcular e armazenar, durante o processamento, a matriz de similaridade (DONI, 2004).

De modo geral, os métodos por particionamento diferem entre si pela maneira que constituem a melhor partição. Os métodos por particionamento que serão discutidos são o Método *k-means* e o Método *Fuzzy c-means*, que por sua vez será utilizado como metodologia principal.

O método *k-means*, que não é o objetivo de estudo, é um algoritmo muito popular utilizado para agrupamento de dados. Além de ser bastante simples e funcionar bem na prática, serve como base para vários outros métodos de agrupamento. O objetivo deste método é agrupar as observações em uma quantidade k de agrupamentos, minimizando a distância intra-*clusters* e maximizando a distância inter-*clusters*.

Em contraste com a simplicidade de seu funcionamento, destacam-se algumas desvantagens. O método é altamente dependente da inicialização dos centroides, em que uma

determinada posição inicial pode levar a solução para um ótimo local. Podemos citar também o fato de que problemas podem ocorrer quando os agrupamentos possuem diferentes tamanhos, densidades e formas não-globulares no espaço de características (JUNIOR, 2006).

Observa-se que o algoritmo *k-means* realiza uma classificação *hard*, pois cada observação é classificada como membro de um único agrupamento. Assim, as observações de uma base de dados podem ser similares a um determinado centroide, ou podem se encontrar equidistantes entre dois centroides. Neste último caso, podem ocorrer erros de classificação, sendo adequado adotar métodos que abordam classificação *soft*, como a técnica *fuzzy c-means* (JUNIOR, 2006).

2.3 Vantagens e Desvantagens

Na análise de agrupamento não se tem um método ótimo de grupos, tanto as técnicas hierárquicas como as não-hierárquicas possuem suas vantagens e desvantagens. As vantagens das técnicas hierárquicas são baseadas na sua simplicidade, uso de diferentes medidas de similaridades e a rapidez. Já as desvantagens estão relacionadas à redução do impacto dos *outliers*, além de que não são indicadas para grandes conjuntos de dados (PRATA *et al.*, 2010).

Quanto às técnicas não-hierárquicas, as principais vantagens são referentes aos resultados que não são tão afetados por *outliers* e podem ser utilizados para grandes conjuntos de dados. No entanto, o uso aleatório de centróides iniciais pode ser uma desvantagem, pois faz com que o método seja inferior ao hierárquico. Mesmo não sendo as sementes aleatórias, a técnica não garante uma solução ótima. O método não é aconselhável em situações onde existem muitos agrupamentos (PRATA *et al.*, 2010).

MATERIAIS E MÉTODOS

Neste capítulo é realizada uma breve discussão acerca da base de germoplasma coletada por uma certa empresa, especificando a quantidade de indivíduos presentes no estudo e as características genéticas selecionadas para o estudo. Além disso, é realizado uma revisão mais aprofundada a respeito da metodologia *fuzzy c-means*.

3.1 Base de Dados

A base de germoplasma de certa cultura de milho foi cedida por uma determinada empresa e tal base inicialmente apresentava 29855 observações com 54 atributos/características.

Para o estudo em questão considerou-se as seguintes características germoplasmáticas, conforme Tabela 1.

Tabela 1: Caracterização das variáveis estudadas.

Variável	Tipo	Escala
Altura da Planta	Numérica	Metros
Altura da Inserção da Espiga	Numérica	Metros
Grãos Ardidos	Numérica	Porcentagem
Rendimento	Numérica	Quilogramas por hectare
Rendimento Real	Numérica	Quilogramas por hectare
Peso do Saco	Numérica	Quilogramas
Gênero	Categórica	Nominal
Município	Categórica	Nominal
UF	Categórica	Nominal

Apenas as variáveis de natureza numérica foram utilizadas para a seguimentação dos indivíduos. Quanto às demais variáveis, essas foram utilizadas para caracterização e análises dos grupos já consolidado conforme será apresentado adiante.

Outro ponto a se destacar quanto às variáveis em estudo, é referente ao *Rendimento Real*, esse que por sua vez é calculado pela seguinte expressão:

$$RR = RE - AR - DE$$

em que RR é o Rendimento Real, RE é o Rendimento Total, AR é a quantidade (em kg/ha) de Grãos Ardidos e DE é a quantidade (em kg/ha) de Descarte.

No entanto, para a execução do algoritmo *fuzzy c-means* é necessário que todas as observações não apresentem sequer alguma informação faltante. Desse modo, foi necessário um tratamento de maneira a omitir da base de germoplasma aqueles indivíduos que apresentaram pelo menos uma informação faltante. Por fim, o realizou-se o estudo com uma base de germoplasma de 13127 observações com 9 características, sendo que 6 dessas características foram utilizadas para a segmentação.

3.2 Método fuzzy c-means

O algoritmo de *clusterização fuzzy* mais popular é o *fuzzy c-means*. Esse algoritmo é melhor do que o *k-means* (que é um algoritmo *hard*), pois evita o mínimo local. No entanto, o *fuzzy c-means* pode ainda convergir para um mínimo local do critério baseado na soma de erros ao quadrado (JUNIOR, 2006). O desenvolvimento de funções de pertinência é o fator mais importante em *clusterização fuzzy*; diferentes escolhas incluem aquelas baseadas em decomposição de similaridade e centróides de *cluster* (JUNIOR, 2006).

Supondo que se tem um conjunto de n indivíduos $X = \{x_1, \dots, x_n\}$ em que cada indivíduo $x_k \in \mathbb{R}^p$, $k = 1, \dots, n$, e deseja-se organizá-los em c clusters, $C = \{C_1, \dots, C_c\}$. O algoritmo *fuzzy c-means* (BEZDEK; EHRLICH; FULL, 1984) é um conjunto não hierárquico de *clusterização* cujo objetivo é fornecer uma partição *fuzzy* de um conjunto de indivíduos em c clusters. Para tanto o *fuzzy c-means* define e minimiza uma função objetivo, a qual mede a adequação entre os indivíduos e os clusters. A função objetivo é definida como (JUNIOR, 2006):

$$J2(U, G) = \sum_{k=1}^n \sum_{i=1}^c (u_{ik})^m \phi(x_k, g_i) = \sum_{k=1}^n \sum_{i=1}^c (u_{ik})^m \sum_{l=1}^p (x_{kl} - g_{il})^2 \quad (3.1)$$

em que $U \in M_{fcn}$ é uma matriz de pertinência $\{u_{ik}\}$ ou apenas pertinência, do indivíduo k ao cluster C_i ; G é um vetor de protótipos dos clusters ou centróides $(g_1, \dots, g_K)$, $g_i = (g_{i1}, \dots, g_{ip})$; $m \in]1, +\infty[$ é um parâmetro que controla a fuzzificação da pertinência dos indivíduos; e

$\phi(x_k, g_i)$ é o quadrado da distância L_2 Minkowsky, ou Euclidiana, a qual mede a dissimilaridade entre um indivíduo k e um protótipo de cluster i .

Semelhante ao k -means, o fuzzy c-means apresenta problemas para lidar com ruídos e anomalias nos dados e, além disso, há a dificuldade de definir a partição inicial. Outro ponto de atenção referente ao fuzzy c-means é a sua complexidade computacional. O algoritmo do fuzzy c-means alterna o cálculo da matriz de pertinência (passo de *alocação*) e o cálculo da matriz de protótipos (passo de *representação*). Para o cálculo da matriz de pertinência, a ordem de complexidade é de nc^2p , enquanto que para calcular os protótipos a complexidade é de ordem npc , em que n é o quantidade de indivíduos, c é a quantidade de clusters e p é a quantidade de protótipos (JUNIOR, 2006).

O algoritmo começa a partir de uma matriz de pertinência inicial e a partir daí alterna um *passo de representação* e um *passo de alocação* até a convergência quando o critério $J2(U, G)$ atinge um valor estacionário que corresponde a um mínimo (geralmente local).

Proposição: Sejam os protótipos $g_i = (g_{i1}, \dots, g_{ip})$ do cluster C_i , com $i = 1, \dots, c$, eles minimizam o critério de clustering, em g_{il} , $J2(U, G)$ e são atualizados de acordo com a seguinte expressão:

$$g_{il} = \frac{\sum_{k=1}^n (u_{ik})^m x_{kl}}{\sum_{k=1}^n (u_{ik})^m}, \quad \text{para } l = 1, \dots, p \quad (3.2)$$

Demonstração

Uma vez que, no passo de representação, a pertinência de cada indivíduo k no cluster C_i e o parâmetro m estão fixos, pode-se reescrever o critério $J2(U, G)$ como

$$\sum_{i=1}^c \sum_{l=1}^p J2_{il}(g_{il})$$

em que

$$J2_{il}(g_{il}) = \sum_{k=1}^n (u_{ik})^m (x_{kl} - g_{il})^2$$

O critério $J2_{il}(g_{il})$ é aditivo e desta forma o problema torna-se achar o g_{il} que minimize $J2_{il}(g_{il})$. A solução deste problema é resolver a equação $\frac{\partial J2_{il}(g_{il})}{\partial g_{il}} = 0$. Desse modo,

$$\frac{\partial J_{2il}(g_{il})}{\partial g_{il}} = 0 \Rightarrow \frac{\partial}{\partial g_{il}} \left[\sum_{k=1}^n (u_{ik})^m (x_{kl} - g_{il})^2 \right] = 0 \Rightarrow -2 \sum_{k=1}^n (u_{ik})^m (x_{kl} - g_{il}) = 0$$

O resultado acima leva a

$$g_{il} = \frac{\sum_{k=1}^n (u_{ik})^m x_{kl}}{\sum_{k=1}^n (u_{ik})^m}$$

O resultado até agora mostra que atualizando-se g_{il} , pela equação 3.1 obtém-se um ponto extremo de $J_{2il}(g_{il})$. Para concluir se este é um ponto de mínimo usaremos o teste da segunda derivada.

$$\frac{\partial^2 J_{2il}(g_{il})}{\partial (g_{il})^2} = 0 \Rightarrow \frac{\partial}{\partial g_{il}} \left[-2 \sum_{k=1}^n (u_{ik})^m (x_{kl} - g_{il}) \right] = 2 \sum_{k=1}^n (u_{ik})^m > 0, \quad \forall g_{il}$$

A segunda derivada de $J_{2il}(g_{il})$ é positiva portanto pode-se concluir que calculando-se g_{il} conforme a equação 3.2 minimiza $J_{2il}(g_{il})$.

Proposição: Seja a pertinência u_{ik} de cada indivíduo k , com $k = 1, \dots, n$, ao *cluster* C_i , com $i = 1, \dots, c$, ela minimiza o critério de *clustering* $J_2(U, G)$ sob as seguintes restrições $u_{ik} \geq 0$ e $\sum_{i=1}^c u_{ik} = 1$ e é atualizada de acordo com a expressão a seguir:

$$u_{ik} = \left[\sum_{h=1}^c \left\{ \frac{\sum_{l=1}^p (x_{kl} - g_{il})^2}{\sum_{l=1}^p (x_{kl} - g_{hl})^2} \right\}^{\frac{1}{m-1}} \right]^{-1} \quad \text{para } i = 1, \dots, c \quad \text{e } k = 1, \dots, n \quad (3.3)$$

Demonstração

Considere que no passo de alocação, os protótipos g_i do *cluster* C_i e o parâmetro m estão fixos.

Uma vez que U é uma matriz não degenerada, suas colunas são independentes e a minimização de $J_2(U, G)$ pode ser obtida aplicando-se o método dos multiplicadores de Lagrange e cada termo u_k . Seja a função $\kappa(u_k)$

$$\kappa(u_k) = \sum_{i=1}^c (u_{ik})^m \sum_{l=1}^p (x_{kl} - g_{il})^2$$

Minimizar $J_2(U, G)$ implica em minimizar $\kappa(u_k)$ sob a restrição $\sum_{i=1}^c u_{ik} = 1$.

Seja o lagrangiano de $\kappa(u_k)$:

$$F(\mu, u_k) = \sum_{i=1}^c (u_{ik})^m \sum_{l=1}^p (x_{kl} - g_{il})^2 - \mu \left(\sum_{i=1}^c (u_{ik})^m - 1 \right)$$

O par (μ, u_k) é estacionário para a função F se o gradiente de F , ∇F , for zero. Fazendo o gradiente de F igual a zero tem-se:

$$\frac{\partial F}{\partial \mu}(\mu, u_k) = \left(\sum_{i=1}^c (u_{ik})^m - 1 \right) = 0 \quad (3.4)$$

$$\frac{\partial F}{\partial u_{ik}}(\mu, u_k) = \left[m(u_{ik})^{m-1} \sum_{l=1}^p (x_{kl} - g_{il})^2 - \mu \right] = 0 \quad (3.5)$$

A partir da equação 3.5, obtém-se:

$$u_{ik} = \left[\frac{\mu}{m \sum_{l=1}^p (x_{kl} - g_{il})^2} \right]^{\frac{1}{m-1}} \quad (3.6)$$

Usando os resultados da equação 3.4 e da equação 3.6, obtém-se:

$$\sum_{h=1}^c u_{hk} = \sum_{h=1}^c \left(\frac{\mu}{m} \right)^{\frac{1}{m-1}} \left[\frac{1}{\sum_{l=1}^p (x_{kl} - g_{hl})^2} \right]^{\frac{1}{m-1}} = \left(\frac{\mu}{m} \right)^{\frac{1}{m-1}} \left(\sum_{h=1}^c \left[\frac{1}{\sum_{l=1}^p (x_{kl} - g_{hl})^2} \right]^{\frac{1}{m-1}} \right) = 1$$

Disto conclui-se:

$$\left(\frac{\mu}{m} \right)^{\frac{1}{m-1}} = \frac{1}{\sum_{h=1}^c \left[\frac{1}{\sum_{l=1}^p (x_{kl} - g_{hl})^2} \right]^{\frac{1}{m-1}}} \quad (3.7)$$

Voltando à equação 3.6 e aplicando o resultado em 3.7 tem-se:

$$u_{ik} = \left(\frac{1}{\sum_{h=1}^c \left[\frac{1}{\sum_{l=1}^p (x_{kl} - g_{hl})^2} \right]^{\frac{1}{m-1}}} \right) \left[\frac{1}{\sum_{l=1}^p (x_{kl} - g_{il})^2} \right]^{\frac{1}{m-1}} =$$

$$= \frac{1}{\sum_{h=1}^c \left[\frac{\sum_{l=1}^p (x_{kl} - g_{il})^2}{\sum_{l=1}^p (x_{kl} - g_{hl})^2} \right]^{\frac{1}{m-1}}} = \left[\sum_{h=1}^c \left[\frac{\sum_{l=1}^p (x_{kl} - g_{il})^2}{\sum_{l=1}^p (x_{kl} - g_{hl})^2} \right]^{\frac{1}{m-1}} \right]^{-1}$$

O resultado acima permite concluir que calculando-se u_{ik} pela expressão 3.3 leva a um valor extremo de $J2(U, G)$. Usar-se-á o teste da segunda derivada para verificar se esta mesma expressão é um mínimo da função objetivo.

Minimizar $\kappa(u_k)$ leva à minimização de $J2(U, G)$. Já foi encontrado um valor extremo $\kappa(u_k)$ e conseqüentemente para $J2(U, G)$ sob a restrição $\sum_{i=1}^c u_{ik} = 1$. Será verificado agora pelo teste da segunda derivada de $\kappa(u_k)$ se este valor é um mínimo.

$$\frac{\partial \kappa}{\partial u_{ik}}(u_k) = \left[m(u_{ik})^{m-1} \sum_{d=1}^p (x_{kl} - g_{il})^2 \right]$$

$$\frac{\partial^2 \kappa}{\partial (u_{ik})^2}(u_k) = \left[m(m-1)(u_{ik})^{m-2} \sum_{d=1}^p (x_{kl} - g_{il})^2 \right]$$

Note que

$$\frac{\partial^2 \kappa}{\partial u_{zk} \partial u_{wk}}(u_k) = 0, \quad \forall z \neq w$$

O que leva a matriz Hessiana de $\kappa(u_k)$ ser definida como

$$H(\kappa(u_k)) = \begin{bmatrix} \left[m(m-1)(u_{1k})^{m-2} \sum_{d=1}^p (x_{kl} - g_{1l})^2 \right] & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \left[m(m-1)(u_{ck})^{m-2} \sum_{d=1}^p (x_{kl} - g_{cl})^2 \right] \end{bmatrix}$$

A matriz $H(\kappa(u_k))$ é diagonal e todos os seus termos são positivos. Isto permite concluir que ela é definida positivamente e portanto isso implica que atualizar u_{ik} pela expressão 3.3 leva a um valor mínimo de $J2(U, G)$.

O sumário dos passos do algoritmo *fuzzy c-means* é dado pelo algoritmo abaixo:

Algoritmo 1 Fuzzy c-means.

- 1: Dados c , $2 \leq c < n$; m , $1 < m < \infty$, T como limite máximo de iterações e um escalar ε , de modo que $0 < \varepsilon$ de modo que ε seja pequeno;
 - 2: Dado um método de inicialização, gere uma matriz de pertinência tal que a pertinência u_{ik} , $i = 1, \dots, c$ e $k = 1, \dots, n$, do indivíduo k no *cluster* C_i tal que $u_{ik} > 0$ e $\sum_{i=1}^c u_{ik} = 1$;
 - 3: **para** $t = 1$ até T **fazer**
 - 4: A pertinência u_{ik} do indivíduo k ao *cluster* C_i está fixa;
 - 5: Calcule os protótipos g_i do *cluster* C_i usando a equação 3.2;
 - 6: Os protótipos g_i dos *cluster* C_i estão fixos;
 - 7: Atualize o grau de pertinência u_{ik} do indivíduo k ao *cluster* C_i para $i = 1, \dots, c$ usando a equação 3.3;
 - 8: **se** $|J2^{(t+1)} - J2^{(t)}| \leq \varepsilon$ **então**
 - 9: pare;
 - 10: **fim se**
 - 11: **fim para**
-

RESULTADOS E DISCUSSÃO

Neste capítulo, é apresentado as principais características e como se comportam os indivíduos da base de germoplasma estudada. Logo em seguida é apresentado um estudo detalhado dos *clusters* formados a partir do método *fuzzy c-means*.

4.1 Análise Descritiva

O conjunto de dados referentes à características genéticas de várias espécies de milho apresentou 13127 observações. Essas observações representam um acumulado entre os anos de 2012 à 2016 de várias localidades do Brasil.

Para o trabalho em questão, foram consideradas seis características genéticas, sendo elas a *Altura da Planta (m)*, *Altura da Inserção da Espiga (m)*, *Grãos Ardidos (%)*, *Rendimento (Mg/ha)*, *Rendimento Real (RE – AR – DE)* e *Peso do Saco (kg)*.

Além dessas seis características, as Tabelas 2, 3 e 4 apresentam as demais características dos indivíduos que não foram utilizadas para a segmentação mas ajudam na caracterização dos mesmos.

Tabela 2: Tabela de distribuição de frequências em relação ao *Gênero* do indivíduo observado.

Gênero	Frequência	Freq. Relativa	Freq. Acumulada
F	12787	0,9741	0,9741
F - ESTERIL	273	0,0208	0,9949
F/M	39	0,0030	0,9979
M	28	0,0021	1

Tabela 3: Tabela de distribuição de frequências em relação ao *Município* de cultivo do indivíduo observado.

Município	Frequência	Freq. Relativa	Freq. Acumulada
PLANALTINA	3397	0,2588	0,2588
IRAI DE MINAS	2328	0,1773	0,4361
MORRINHOS	1560	0,1188	0,5549
JABOTICABAL	891	0,0679	0,6228
PIRACANJUBA	778	0,0593	0,6821
GUAIRA	647	0,0493	0,7314
CENTRALINA	570	0,0434	0,7748
MONTE ALTO	358	0,0273	0,8021
AGUA FRIA	357	0,0272	0,8293
JABORANDI	327	0,0249	0,8542
PARANAPANEMA	250	0,0190	0,8732
CAMPO ALEGRE	214	0,0163	0,8895
SILVANIA	197	0,0150	0,9045
PLANURA	176	0,0134	0,9179
VICENTINOPOLIS	146	0,0111	0,9290
PERDIZES	139	0,0106	0,9396
PATOS DE MINAS	135	0,0103	0,9499
GOIATUBA	133	0,0101	0,9600
PARACATU	124	0,0094	0,9694
CASCAVEL	116	0,0088	0,9782
CASA BRANCA	89	0,0068	0,9850
TAQUARITINGA	68	0,0052	0,9902
BURITIS	65	0,0050	0,9952
LUIZ EDUARDO MAGALHAES	62	0,0047	1

Tabela 4: Tabela de distribuição de frequências em relação à *UF* de cultivo do indivíduo observado.

UF	Frequência	Freq. Relativa	Freq. Acumulada
MG	3472	0,2645	0,2645
GO	3450	0,2628	0,5273
DF	3397	0,2588	0,7861
SP	2303	0,1754	0,9615
BA	389	0,0296	0,9911
PR	116	0,0088	1

Na Tabela 2, temos que as siglas *F*, *F - ESTERIL*, *F/M* e *M* representam, respectivamente os gêneros *Feminino*, *Feminino Estéril*, *Feminino/Masculino* e *Masculino*. Além disso, verifica-se que aproximadamente 97% dos indivíduos analisados são do gênero feminino (*F*). Outro ponto é que, em relação ao município, a maioria dos indivíduos, representados por 25,88% do total, são provenientes de Planaltina - DF, como se pode verificar na Tabela 3. Por outro lado, quando se analisa a distribuição dos indivíduos quanto à Unidade Federativa, verifica-se que aproximadamente de 78% dos indivíduos são provenientes dos estados de Minas Gerais, Goiás e Distrito Federal.

Em relação às características dos indivíduos, tem-se a Tabela 5 que apresenta as descrição das mesmas.

Tabela 5: Estatísticas descritivas à respeito dos dados analisados.

Característica	Mínimo	Média	Mediana	Máximo	Desvio Padrão
Altura da Planta (m)	0,31	1,75	1,76	2,9	0,27
Altura da Inserção da Espiga (m)	0,05	1,00	1,00	2,25	0,22
Grãos Ardidos (%)	0,02	2,93	1,87	50,24	3,59
Rendimento (Mg/ha)	30,56	5131,86	4907,19	19165,02	2486,16
Rendimento Real	24,23	4582,82	4352,61	17940,49	2454,70
Peso do Saco (kg)	3,12	14,23	14,08	57,47	2,97

Através das Figuras 2 à 7, é possível verificar a distribuição das observações em relação à cada uma das seis características de estudo.

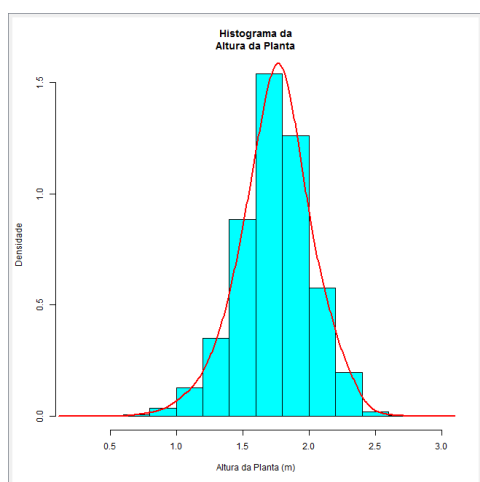


Figura 2: Distribuição das observações em relação à característica Altura da Planta (m)

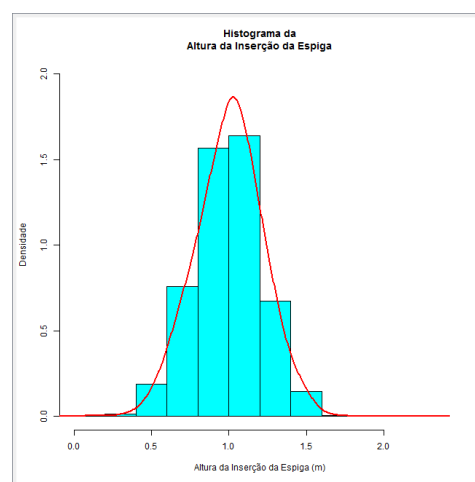


Figura 3: Distribuição das observações em relação à característica Altura da Inserção da Espiga (m)

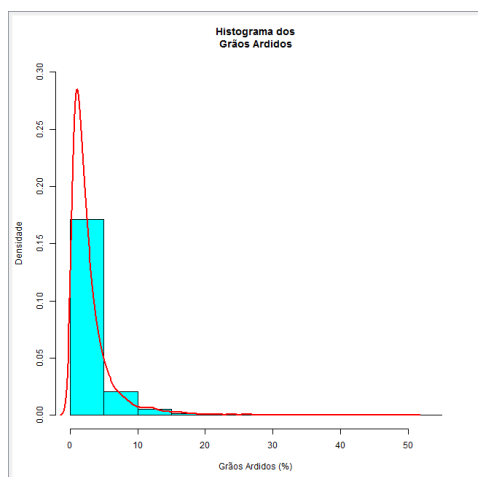


Figura 4: Distribuição das observações em relação à característica Grãos Ardidos (%)

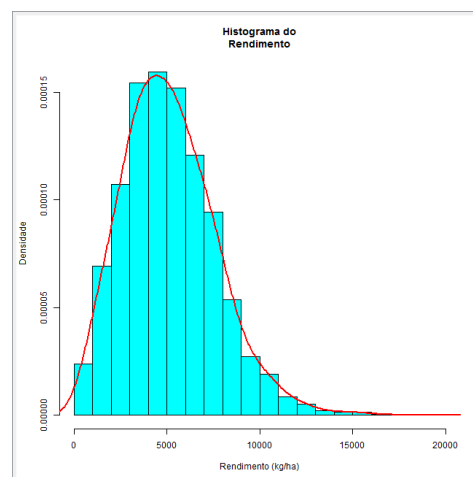


Figura 5: Distribuição das observações em relação à característica Rendimento (Mg/ha)

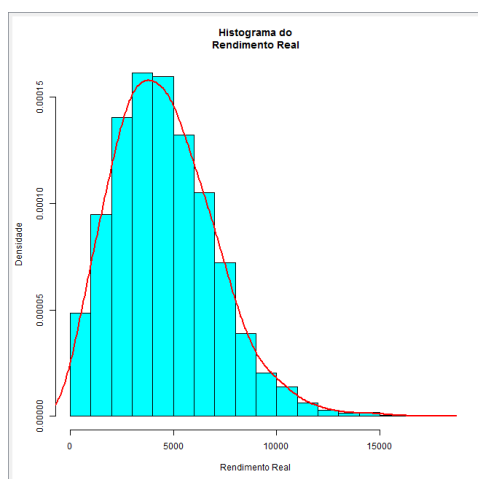


Figura 6: Distribuição das observações em relação à característica Rendimento Real

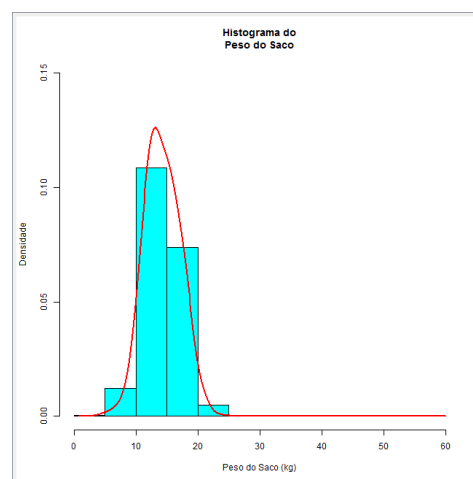


Figura 7: Distribuição das observações em relação à característica Peso do Saco (kg)

Visualmente, as características *Grãos Ardidos*, *Rendimento*, *Rendimento Real* e *Peso do Saco* apresentam assimetria. A Tabela 6 apresenta os coeficientes de assimetria e curtose das características estudadas.

Tabela 6: Coeficientes de assimetria e curtose para as características estudadas.

Característica	Assimetria	Curtose
Altura da Planta (m)	−0,255	0,468
Altura da Inserção da Espiga (m)	−0,051	0,363
Grãos Ardidos (%)	4,202	28,213
Rendimento (Mg/ha)	0,663	0,835
Rendimento Real	0,706	0,819
Peso do Saco (kg)	0,329	3,361

Verifica-se na Tabela 6 que a característica com maior assimetria é a *Grãos Ardidos* que, por sua vez, apresenta uma assimetria positiva (à direita). As características *Rendimento*, *Rendimento Real* e *Peso do Saco* apresentam leve assimetria à direita.

Através do Coeficiente de Correlação de Pearson é possível verificar a existência de associação linear entre as características estudadas. A Tabela 7 apresenta tais correlações.

Tabela 7: Coeficiente de Correlação de Pearson entre as características estudadas.

		1	2	3	4	5	6
1	Altura da Planta (m)	1					
2	Altura da Inserção da Espiga (m)	0,744	1				
3	Grãos Ardidos (%)	-0,097	-0,072	1			
4	Rendimento (Mg/ha)	0,445	0,485	-0,153	1		
5	Rendimento Real	0,414	0,468	-0,189	0,979	1	
6	Peso do Saco (kg)	0,120	0,211	0,016	0,425	0,490	1

Como se pode observar na Tabelas 7, algumas características apresentam correlação linear forte entre si como o *Rendimento Real* e o *Rendimento*, além da *Altura da Planta* e a *Altura de Inserção da Espiga*.

4.2 Análise de agrupamento pelo método fuzzy c-means

Observe que as características *Altura da Planta (m)*, *Altura da Inserção da Espiga (m)*, *Grãos Ardidos (%)*, *Rendimento (Mg/ha)*, *Rendimento Real (RE – AR – DE)* e *Peso do Saco (kg)* apresentam escalas distintas. Para evitar o efeito de determinadas escalas, padronizou-se os dados pela padronização Gaussiana.

Dessa maneira, com auxílio do pacote *fclust*, possibilitou-se a formação de três *clusters* pelo Método *fuzzy c-means* de maneira que não houvesse influência de determinada característica devido à escala. A quantidade de *cluster* foi determinada através de gerações sucessivas e comparações. Foram realizados experimentos considerando 2, 3, 4 e 5 *clusters*, sendo que o experimento de tamanho 2 não ajudou a diferenciar os indivíduos, enquanto que os experimentos de tamanho 4 e 5 não apresentou grandes diferenças do experimento de tamanho 3.

Por esse motivo, optou-se pelo experimento com divisão entre três *clusters*. A Tabela 8 apresenta a distribuição dos indivíduos em relação aos três *clusters*.

Tabela 8: Distribuição dos indivíduos entre os *clusters*.

<i>Cluster</i>	Frequência
1	3853
2	4881
3	4393

Essa se trata de uma clusterização *soft*, ou seja, uma clusterização em que os indivíduos não pertencem a um único *cluster* mas aos *c clusters* determinados pelo pesquisador, porém, com um grau de pertinência maior em um dos *clusters*.

Sendo assim, os indivíduos que foram agrupados do *cluster 1* apresentaram maior grau de pertinência nesse *cluster* em relação aos demais *clusters*. De maneira análoga podemos explicar o agrupamento dos demais indivíduos nos *clusters 2 e 3*.

As Tabelas 9, 10 e 11 descrevem as características dos *clusters 1, 2 e 3*.

Tabela 9: Descrição das características estudadas no *cluster 1*.

Característica	Média	Mediana	Desv. Pad.	Mínimo	Máximo
Altura da Planta (m)	1,915	1,900	0,226	1,000	2,670
Altura da Inserção da Espiga (m)	1,159	1,150	0,182	0,150	2,250
Grãos Ardidos (%)	2,255	1,580	2,258	0,020	24,300
Rendimento (Mg/ha)	7947	7611	1888,512	1593	19165
Rendimento Real	7431	7092	1834,177	1549	17940
Peso do Saco (kg)	16,677	16,892	2,676	4,579	57,471

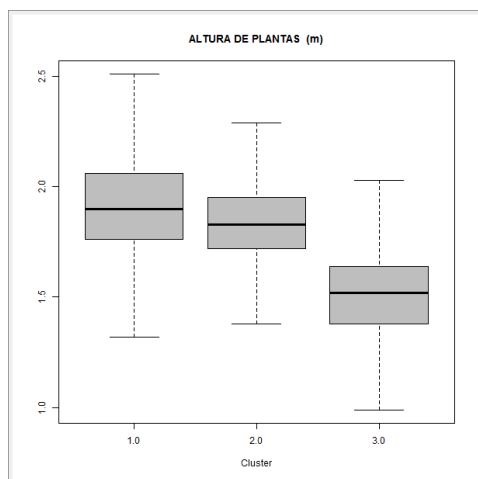
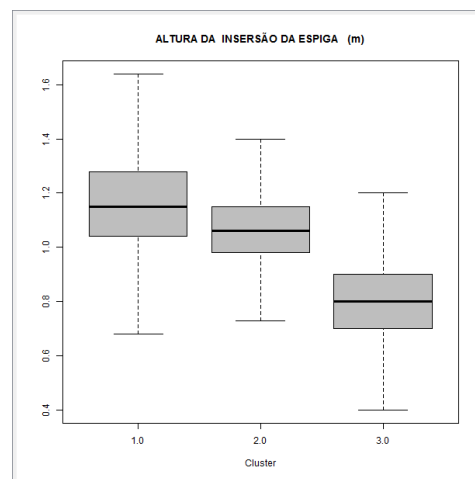
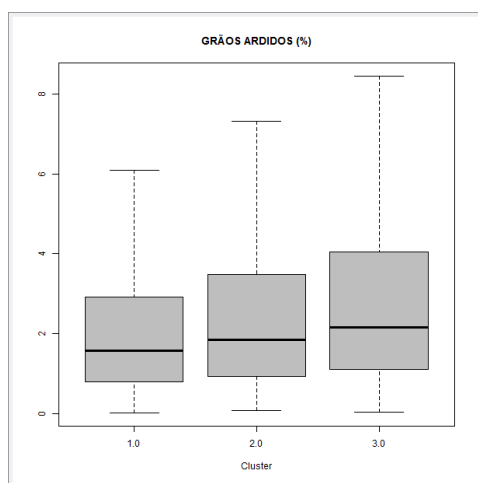
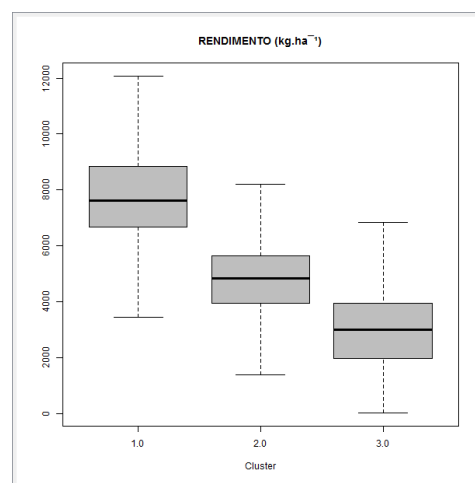
Tabela 10: Descrição das características estudadas no *cluster 2*.

Característica	Média	Mediana	Desv. Pad.	Mínimo	Máximo
Altura da Planta (m)	1,845	1,830	0,1823	1,050	2,900
Altura da Inserção da Espiga (m)	1,069	1,060	0,1391	0,510	1,700
Grãos Ardidos (%)	2,877	1,850	3,291	0,070	45,833
Rendimento (Mg/ha)	4795,400	4846,800	1294,650	30,570	9103,800
Rendimento Real	4150,800	4261,200	1259,485	147,500	8723,600
Peso do Saco (kg)	13,450	13,340	2,359	3,816	22,364

Tabela 11: Descrição das características estudadas no *cluster 3*.

Característica	Média	Mediana	Desv. Pad.	Mínimo	Máximo
Altura da Planta (m)	1,498	1,520	0,211	0,310	2,250
Altura da Inserção da Espiga (m)	0,792	0,800	0,152	0,050	1,440
Grãos Ardidos (%)	3,574	2,160	4,597	0,030	50,240
Rendimento (Mg/ha)	3037,020	2986,700	1420,245	30,570	9103,800
Rendimento Real	2564,560	2503,060	1376,842	24,230	8731,920
Peso do Saco (kg)	12,956	12,775	2,538	3,121	24,814

Para uma análise visual do comportamento das características dentro de cada um dos *clusters*, têm-se as Figuras 8, 9, 10, 11, 12 e 13.

**Figura 8:** Boxplot da característica Altura da Planta (m) por *cluster*.**Figura 9:** Boxplot da característica Altura de Inserção da Espiga (m) por *cluster*.**Figura 10:** Boxplot da característica Grãos Ardidos (%) por *cluster*.**Figura 11:** Boxplot da característica Rendimento (Mg/ha) por *cluster*.

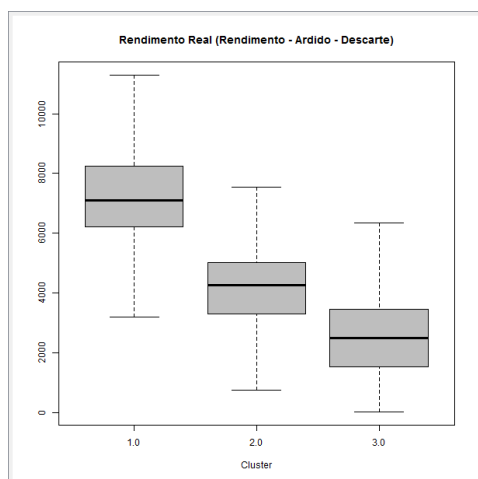


Figura 12: Boxplot da característica Rendimento Real por *cluster*.

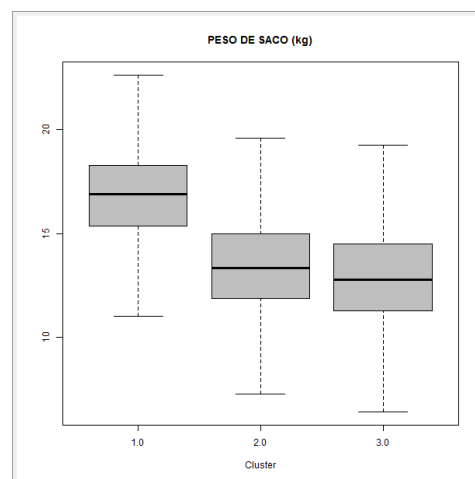


Figura 13: Boxplot da característica Peso do Saco (kg) por *cluster*.

Ao analisar cada um dos três *clusters*, com auxílio das Tabelas e Figuras apresentadas acima, verifica-se que os indivíduos com maior grau de pertinência no *cluster 1* apresentam características como o *Rendimento* (Mg/ha), o *Rendimento Real* ($RE - AR - DE$) e o *Peso do Saco* (kg) superiores aos indivíduos dos demais *clusters*.

Quanto às características *Altura da Planta* (m) e *Altura da Inserção da Espiga* (m), essas apresentaram valores menores para os indivíduos do *cluster 3*.

Outro ponto importante é que os indivíduos com maior grau de pertinência no *cluster 3* apresentam um comportamento quanto a quantidade de *Grãos Ardidos* (%) superior aos indivíduos dos demais *clusters*.

Por fim, os indivíduos que apresentaram maior grau de pertinência no *cluster 2*, destacaram-se por serem indivíduos que apresentam características intermediárias em relação aos indivíduos dos *clusters 1* e *3*.

As Figuras 14, 15 e 16 apresentam o comportamento dos indivíduos dos *clusters 1*, *2* e *3*, respectivamente, em relação ao grau de pertinência.

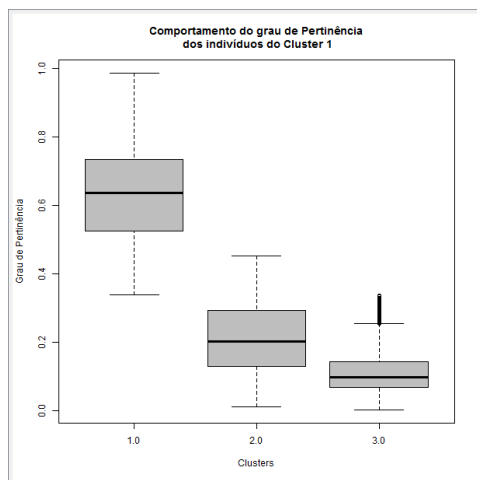


Figura 14: Boxplot do grau de pertinência dos indivíduos do *cluster 1*.

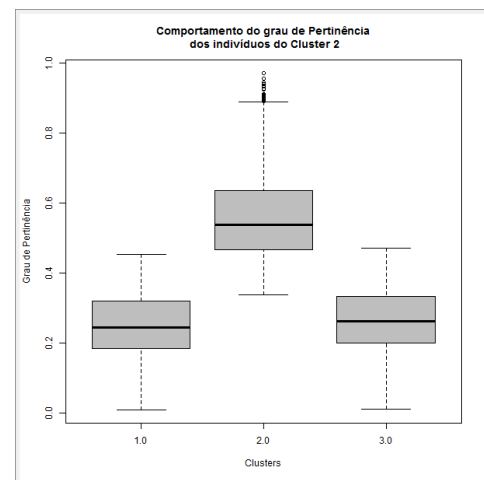


Figura 15: Boxplot do grau de pertinência dos indivíduos do *cluster 2*.

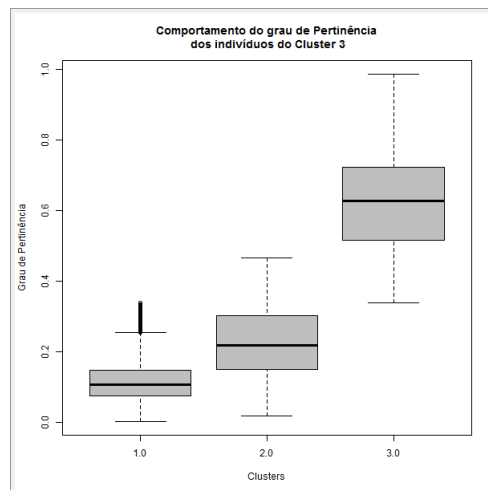


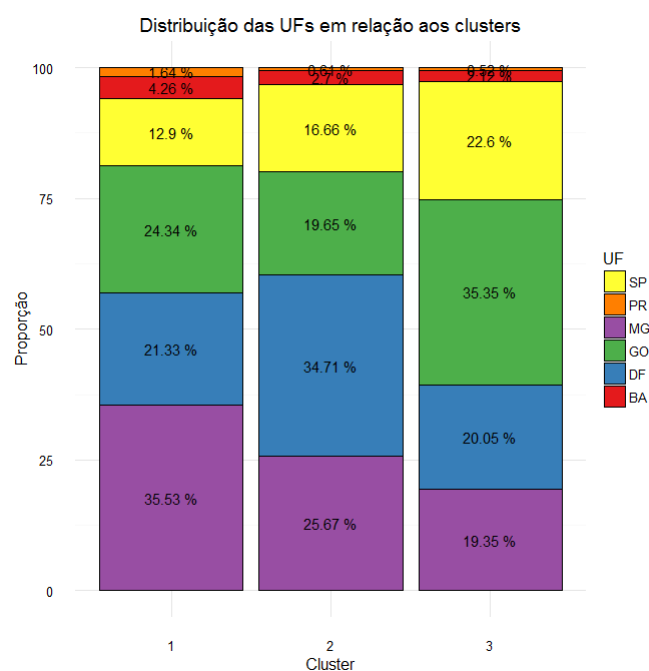
Figura 16: Boxplot do grau de pertinência dos indivíduos do *cluster 3*.

Observa-se que os indivíduos que apresentam maior grau de pertinência para o *cluster 1* também apresentam o menor grau de pertinência para o *cluster 3* e vice-versa. Quanto aos indivíduos que apresentam maior grau de pertinência para o *cluster 2*, estes são intermediadores entre os demais *clusters*.

Em relação ao local de origem dos indivíduos, a Tabela 12, com auxílio da Figura 17, mostra como está a distribuição dos *clusters* em relação às Unidades Federativas.

Tabela 12: Distribuição das UFs em relação aos *clusters*.

	<i>Cluster 1</i>	<i>Cluster 2</i>	<i>Cluster 3</i>
MG	1369	1253	850
DF	822	1694	881
GO	938	959	1553
SP	497	813	993
BA	164	132	93
PR	63	30	23

**Figura 17:** Gráfico de barras empilhado da distribuição das UFs em relação aos *clusters*.

Repare na Tabela 12 e na Figura 17 que os indivíduos do *cluster 1* são predominantes no estado de MG, enquanto que os indivíduos do *cluster 2* e *3*, apresentam predominância no DF e no estado de GO, respectivamente.

As Tabelas 13, 14 e 15 apresenta a distribuição dos cinco municípios mais frequentes.

Tabela 13: Distribuição dos municípios em relação ao *clusters 1*.

Município	Frequência	Freq. Relativa	Freq. Acumulada
IRAI DE MINAS	859	0,2229	0,2229
PLANALTINA	822	0,2133	0,4362
MORRINHOS	326	0,0846	0,5208
GUAIRA	277	0,0719	0,5927
CENTRALINA	272	0,0706	0,6633
DEMAIS LOCALIDADES	1297	0,3366	1

Tabela 14: Distribuição dos municípios em relação ao *clusters 2*.

Município	Frequência	Freq. Relativa	Freq. Acumulada
PLANALTINA	1694	0,3471	0,3471
IRAI DE MINAS	906	0,1856	0,5327
MORRINHOS	458	0,0938	0,6265
JABOTICABAL	341	0,0699	0,6964
GUAIRA	265	0,0543	0,7507
DEMAIS LOCALIDADES	1217	0,2493	1

Tabela 15: Distribuição dos municípios em relação ao *clusters 3*.

Município	Frequência	Freq. Relativa	Freq. Acumulada
PLANALTINA	881	0,2005	0,2005
MORRINHOS	776	0,1766	0,3771
IRAI DE MINAS	563	0,1282	0,5053
PIRACANJUBA	470	0,1070	0,6123
JABOTICABAL	443	0,1008	0,7131
DEMAIS LOCALIDADES	1260	0,2868	1

Aparentemente, ao analisarmos as Tabelas 13, 14 e 15, nenhuma região se destaca quanto ao *cluster*. Apesar de verificarmos, na Tabela 12 e Figura 17, certa predominância de determinadas UF, não é possível assumir que uma suposição a respeito da influência climática e geológica das regiões quanto à produção dessas culturas estudadas.

CONSIDERAÇÕES FINAIS

A análise de agrupamento é uma importante ferramenta de análise pois possibilita que o pesquisador conheça melhor os dados para os quais está analisando seu comportamento e características. Dessa maneira é possível realizar ações específicas para cada grupo de acordo com suas especificidades.

Diversos métodos de segmentação podem ser utilizados e cada um deles apresentam características específicas para determinadas situações. No caso da metodologia utilizada nesse trabalho, *fuzzy c-means*, essa apresenta uma peculiaridade em relação às demais metodologias comumente utilizadas (como a *k-means*) por não definir o indivíduo analisado em um único e exclusivo *cluster*, mas possibilitando tal indivíduo pertencer a diversos *clusters*.

Nesse estudo, em específico, foi possível avaliar o comportamento de certa cultivar de milho, proveniente de um banco de germoplasma cedido por uma determinada empresa, através do método de segmentação *fuzzy c-means*. Dessa maneira, foram identificados três *clusters* com características específicas.

Em destaque temos os indivíduos do *cluster 1* que apresentaram um comportamento superior aos indivíduos dos demais *clusters* em relação ao *Rendimento* e ao *Rendimento Real*.

Referências Bibliográficas

ANDERBERG, Michael R. **Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks**. [S.l.]: Academic press, 2014. v. 19.

BACKER, Eric. **Computer-assisted reasoning in cluster analysis**. [S.l.]: Prentice Hall International (UK) Ltd., 1995.

BERTHOLD, Michael R.; CEBRON, Nicolas; DILL, Fabian; GABRIEL, Thomas R.; KÖTTER, Tobias; MEINL, Thorsten; OHL, Peter; SIEB, Christoph; THIEL, Kilian; WISWEDEL, Bernd. KNIME: The Konstanz Information Miner. In: **Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)**. [S.l.]: Springer, 2007. ISBN 978-3-540-78239-1. ISSN 1431-8814.

BEZDEK, James C; EHRLICH, Robert; FULL, William. Fcm: The fuzzy c-means clustering algorithm. **Computers & Geosciences**, Elsevier, v. 10, n. 2-3, p. 191–203, 1984.

BEZDEK, James C; PAL, Sankar K. Fuzzy models for pattern recognition. 1994.

BROWN, Anthony Hugh Dean. **The use of plant genetic resources**. [S.l.]: Cambridge University Press, 1989.

BUSSAB, W de O. **Introdução à análise de agrupamentos**. [S.l.]: ABE, 1990. 105 p., 1990.

DONI, Marcelo Viana. **Análise de cluster: métodos hierárquicos e de particionamento**. São Paulo: Universidade Presbiteriana Mackenzie, São Paulo, 2004.

FERRARO, M.B.; GIORDANI, P. A toolbox for fuzzy clustering using the r programming language. **Fuzzy Sets and Systems**, v. 279, p. 1–16, 2015. Disponível em: <<http://dx.doi.org/10.1016/j.fss.2015.05.001>>.

HAMMER, Karl. A paradigm shift in the discipline of plant genetic resources. **Genetic Resources and Crop Evolution**, Springer, v. 50, n. 1, p. 3–10, 2003.

JUNIOR, Nicomedes Lopes Cavalcanti. **Clusterização baseada em algoritmos fuzzy**. Universidade Federal de Pernambuco, 2006.

KAUFMAN, Leonard; ROUSSEEUW, Peter J. **Finding groups in data: an introduction to cluster analysis**. [S.l.]: John Wiley & Sons, 2009. v. 344.

LI, Yu; WANG, Jie; CAO, Yongsheng; GAO, Weidong; FANG, Jiahe; LOU, Xizhi. The use of genetic resources in crop improvement: lessons from china. **Genetic Resources and Crop Evolution**, Springer, v. 45, n. 2, p. 181–186, 1998.

PRATA, Bruno de Athayde; FREITAS, Silvia Maria de; CASTRO, Bruno Monte de; EVANGELISTA, George Leitão. “raio de influência”: um método de agrupamento alternativo para análise de cluster. Simpósio Brasileiro de Probabilidade e Estatística, 2010.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2016. Disponível em: <<https://www.R-project.org/>>.

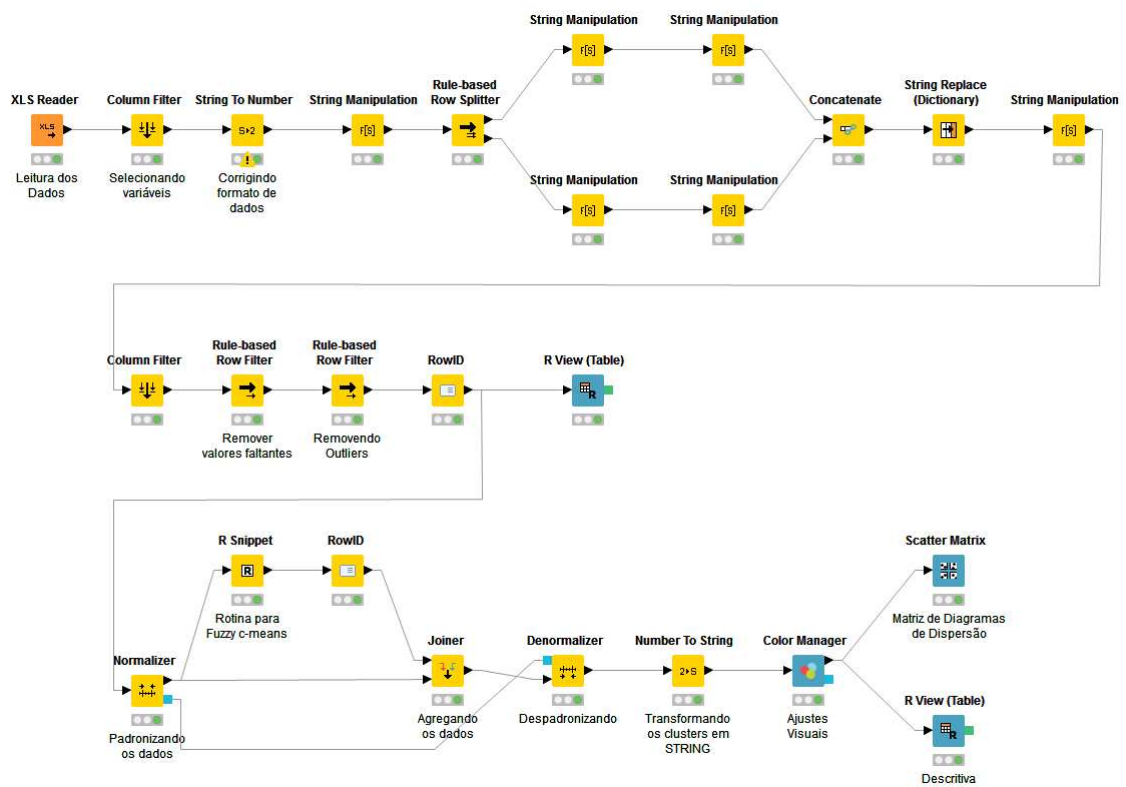
ROMESBURG, Charles. **Cluster analysis for researchers**. [S.l.]: Lulu. com, 2004.

RUHOFF, Anderson Luis; SOUZA, Bernardo Sayão Penna; GIOTTO, Enio; PEREIRA, Rudiney Soares. Lógica fuzzy e zoneamento ambiental da bacia do arroio grande. 2005.

VIAN, André Luis; SANTI, Antônio Luis; AMADO, Telmo Jorge Carneiro; CHERUBIN, Maurício Roberto; SIMON, Diego Henrique; DAMIAN, Junior Melo; BREDEMEIER, Christian. Variabilidade espacial da produtividade de milho irrigado e sua correlação com variáveis explicativas de planta. **Ciencia rural**, Universidade Federal de Santa Maria, v. 46, n. 3, p. 464–471, 2016.

YONAMINE, Frank Sussumu; SPECIA, Lucia; CARVALHO, VO; NICOLETTI, MC. Aprendizado não supervisionado em domínios fuzzy–algoritmo fuzzy c-means. **São Carlos: UFSCAR**, 2002.

APÊNDICE A – KNIME Workflow para tratamento e análise dos dados



APÊNDICE B – Algoritmo R para a análise descritiva

```
# LEITURA DOS DADOS
```

```
dados = data.frame(knime.in$"ALTURA DE PLANTAS (m)",
knime.in$"ALTURA DA INSERSÃO DA ESPIGA (m)",knime.in$"GRÃOS ARDIDOS (%)",
knime.in$"RENDIMENTO (kg.ha)",
knime.in$"Rendimento Real (Rendimento - Ardido - Descarte)",
knime.in$"PESO DE SACO (kg)")
```

```
colnames(dados) = c("AlturaPlanta","AlturaInsercaoEspiga","GraosArdidos",
"Rendimento","RendimentoReal","PesoSaco")
```

```
dados_cat = data.frame(knime.in$"F / M",knime.in$"MUNICIPIO",knime.in$"UF")
colnames(dados_cat) = c("Genero","Municipio","UF")
```

```
# ANÁLISE DOS DADOS QUANTITATIVOS
```

```
cor(dados)
```

```
hist(dados$AlturaPlanta, prob = TRUE, col = "cyan",
xlab = "Altura da Planta (m)", ylab = "Densidade",
main = "Histograma da\nAltura da Planta")
lines(density(dados$AlturaPlanta, adjust =2), col = "red", lwd = 2)
```

```
hist(dados$AlturaInsercaoEspiga, prob = TRUE, ylim = c(0,2), col = "cyan",
xlab = "Altura da Inserção da Espiga (m)", ylab = "Densidade",
main = "Histograma da\nAltura da Inserção da Espiga")
```

```
lines(density(dados$AlturaInsercaoEspiga, adjust =2), col = "red", lwd = 2)

hist(dados$GraosArdidos, prob = TRUE, ylim = c(0,0.3), col = "cyan",
xlab = "Grãos Ardidos (%)", ylab = "Densidade",
main = "Histograma dos\nGrãos Ardidos")
lines(density(dados$GraosArdidos, adjust =2), col = "red", lwd = 2)

hist(dados$Rendimento, prob = TRUE, col = "cyan", xlab = "Rendimento (kg/ha)",
ylab = "Densidade", main = "Histograma do\nRendimento")
lines(density(dados$Rendimento, adjust =2), col = "red", lwd = 2)

hist(dados$RendimentoReal, prob = TRUE, col = "cyan", xlab = "Rendimento Real",
ylab = "Densidade", main = "Histograma do\nRendimento Real")
lines(density(dados$RendimentoReal, adjust =2), col = "red", lwd = 2)

hist(dados$PesoSaco, prob = TRUE, ylim = c(0,0.15), col = "cyan",
xlab = "Peso do Saco (kg)", ylab = "Densidade",
main = "Histograma do\nPeso do Saco")
lines(density(dados$PesoSaco, adjust =2), col = "red", lwd = 2)

# ANÁLISE DOS DADOS QUALITATIVOS

cat("\n\nVARIÁVEIS CATEGÓRICAS\n\n")

dados_gen = data.frame(lapply(dados_cat,table)$Genero)
colnames(dados_gen) = c("Genero","Frequencia")

dados_mun = data.frame(lapply(dados_cat,table)$Municipio)
colnames(dados_mun) = c("Municipio","Frequencia")

dados_uf = data.frame(lapply(dados_cat,table)$UF)
colnames(dados_uf) = c("UF","Frequencia")

dados_gen
dados_mun
```

```
dados_uf
\end{verbatim}

\chapter{Algoritmo R$ usando o pacote \textit{fclust}}
\label{Apx:C}

\begin{verbatim}
install.packages('fclust')
library(fclust)

dados = data.frame(knime.in$"ALTURA DE PLANTAS (m)",
knime.in$"ALTURA DA INSERSÃO DA ESPIGA (m)",knime.in$"GRÃOS ARDIDOS (%)",
knime.in$"RENDIMENTO (kg.ha)",
knime.in$"Rendimento Real (Rendimento - Ardido - Descarte)",
knime.in$"PESO DE SACO (kg)")

cluster = data.frame(FKM(dados, k = 3)$clus)
graus = data.frame(FKM(dados, k = 3)$U)
table(cluster[1])

knime.out = as.data.frame(cbind(cluster[1],graus))
```

APÊNDICE C – Algoritmo R para descrição dos clusters

```
by(knime.in, knime.in$"Cluster", summary)
```

```
cat("\n\nALTURA DE PLANTAS (m)\n\n")
```

```
by(knime.in$"ALTURA DE PLANTAS (m)", knime.in$"Cluster", sd)
```

```
cat("\n\nALTURA DA INSERSÃO DA ESPIGA (m)\n\n")
```

```
by(knime.in$"ALTURA DA INSERSÃO DA ESPIGA (m)", knime.in$"Cluster", sd)
```

```
cat("\n\nGRÃOS ARDIDOS (%)\n\n")
```

```
by(knime.in$"GRÃOS ARDIDOS (%)", knime.in$"Cluster", sd)
```

```
cat("\n\nRENDIMENTO (kg.ha)\n\n")
```

```
by(knime.in$"RENDIMENTO (kg.ha)", knime.in$"Cluster", sd)
```

```
cat("\n\nRendimento Real (Rendimento - Ardido - Descarte)\n\n")
```

```
by(knime.in$"Rendimento Real (Rendimento - Ardido - Descarte)",  
knime.in$"Cluster", sd)
```

```
cat("\n\nPESO DE SACO (kg)\n\n")
```

```
by(knime.in$"PESO DE SACO (kg)", knime.in$"Cluster", sd)
```

```
boxplot(knime.in$"ALTURA DE PLANTAS (m)" ~ knime.in$"Cluster",
```

```
main = "ALTURA DE PLANTAS (m)", xlab = "Cluster", outline = FALSE,
```

```
col = "gray")
```

```
boxplot(knime.in$"ALTURA DA INSERSÃO DA ESPIGA (m)" ~ knime.in$"Cluster",
```

```
main = "ALTURA DA  INSERSÃO DA ESPIGA    (m)", xlab = "Cluster",
outline = FALSE, col = "gray")

boxplot(knime.in$"GRÃOS ARDIDOS (%)" ~ knime.in$"Cluster",
main = "GRÃOS ARDIDOS (%)", xlab = "Cluster", outline = FALSE, col = "gray")

boxplot(knime.in$"RENDIMENTO (kg.ha)" ~ knime.in$"Cluster",
main = "RENDIMENTO (kg.ha)", xlab = "Cluster", outline = FALSE, col = "gray")

boxplot(knime.in$"Rendimento Real (Rendimento - Ardido - Descarte)" ~
knime.in$"Cluster", main = "Rendimento Real (Rendimento - Ardido - Descarte)",
xlab = "Cluster", outline = FALSE, col = "gray")

boxplot(knime.in$"PESO DE SACO (kg)" ~ knime.in$"Cluster",
main = "PESO DE SACO (kg)", xlab = "Cluster", outline = FALSE, col = "gray")

boxplot(knime.in$"Clus.1" ~ knime.in$"Cluster",
main = "Comportamento do grau de Pertinência\n dos indivíduos do Cluster 1",
xlab = "Clusters", ylab = "Grau de Pertinência", col = "gray")

boxplot(knime.in$"Clus.2" ~ knime.in$"Cluster",
main = "Comportamento do grau de Pertinência\n dos indivíduos do Cluster 2",
xlab = "Clusters", ylab = "Grau de Pertinência", col = "gray")

boxplot(knime.in$"Clus.3" ~ knime.in$"Cluster",
main = "Comportamento do grau de Pertinência\n dos indivíduos do Cluster 3",
xlab = "Clusters", ylab = "Grau de Pertinência", col = "gray")
```