VINÍCIUS QUINTÃO CARNEIRO

APLICATIVOS COMPUTACIONAIS PARA O MELHORAMENTO GENÉTICO FUNDAMENTADOS EM ANÁLISE DE IMAGENS E INTELIGÊNCIA COMPUTACIONAL

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de Doctor Scientiae.

VIÇOSA
MINAS GERAIS - BRASIL
2018

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
C289a
2018

Carneiro, Vinícius Quintão, 1989-
Aplicativos computacionais para o melhoramento genético fundamentados em análise de imagens e inteligência computacional / Vinícius Quintão Carneiro. – Viçosa, MG, 2018.
ix, 127 f. : il. (algumas color.) ; 29 cm.

Texto em português e inglês.
Orientador: Cosme Damião Cruz.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Engenharia genética. 2. Melhoramento genético. 3. Software. 4. Melhoramento de cultivos agrícolas. I. Universidade Federal de Viçosa. Departamento de Biologia Geral. Programa de Pós-Graduação em Genética e Melhoramento. II. Título.

CDD 22. ed. 631.5233

VINÍCIUS QUINTÃO CARNEIRO

APLICATIVOS COMPUTACIONAIS PARA O MELHORAMENTO GENÉTICO FUNDAMENTADOS EM ANÁLISE DE IMAGENS E INTELIGÊNCIA COMPUTACIONAL

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de *Doctor Scientiae*.

APROVADA: 19 de outubro de 2018.

Antônio Carlos Baião de Oliveira

Francisco Charles do Santos Silva

Moysés Nascimento

Pedro Crescêncio Souza Carneiro

Cosme Damião Cruz
(Orientador)

A Deus,

Luz em meu caminho

OFEREÇO

Aos meus pais, José Eustáquio de Souza Carneiro e Beatriz Pereira Quintão, e ao meu irmão Carlos Eduardo Quintão Carneiro pelo incondicional apoio durante toda a minha vida.

DEDICO

AGRADECIMENTOS

Agradeço a Deus por iluminar o meu caminho durante toda esta jornada, dando-me saúde, conhecimento e força para conseguir superar os desafios da vida.

Aos meus pais, José Eustáquio e Beatriz pelo amor, carinho, dedicação e incentivo para que me permitisse alcançar mais este objetivo.

Ao meu irmão Carlos Eduardo pela amizade, compreensão, paciência e incentivo.

Aos meus familiares que sempre me apoiaram em minhas decisões, sobretudo aos meus avós exemplos de persistência, trabalho e doação.

À minha namorada Jussara Mencalha pela convivência, companheirismo, paciência e total apoio.

À Universidade Federal de Viçosa pela oportunidade de cursar o ensino médio no colégio de aplicação - Coluni e a graduação em Agronomia. Ao Programa de Genética e Melhoramento desta mesma instituição pela oportunidade de me tornar mestre e doutor em genética e melhoramento.

Aos secretários do Programa de Pós Graduação em Genética e Melhoramento, Marco Túlio e Odilon.

Aos funcionários do BIOAGRO pela disponibilidade e amizade, em especial ao senhor Paulo.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela concessão da bolsa de estudos.

Ao CNPq, FAPEMIG E CAPES pelo apoio financeiro durante a realização destes trabalhos.

Ao professor Cosme Damião Cruz, pela excelente orientação, pelos conhecimentos transmitidos, paciência e sobretudo pela amizade. Exemplo de dedicação, competência e humildade.

Ao professor Pedro Crescêncio Souza Carneiro, não somente pela coorientação, mas por todo apoio durante todo o período de minha formação acadêmica, pela dedicação, incentivo e amizade.

Ao professor José Eustáquio pelos conselhos transmitidos e disponibilidade. Sem o seu apoio seria impossível a realização deste sonho.

Ao professor Moysés, pela colaboração, ensinamentos, disponibilidade, sugestões e amizade.

Aos doutores Antônio Carlos Baião de Oliveira e Francisco Charles do Santos Silva pelas excelentes contribuições para este trabalho.

Aos professores dos departamentos de Biologia Geral, Fitotecnia e Estatística que colaboraram em minha formação acadêmica.

Aos amigos que fiz durante todo este período, em especial aos amigos do Laboratório de Bioinformática pela ótima convivência. Os quais levarei a amizade por toda a vida.

A todos aqueles que colaboraram de alguma forma, pelo incentivo, compreensão e amizade.

MUITO OBRIGADO!

BIOAGRAFIA

VINÍCIUS QUINTÃO CARNEIRO, filho de José Eustáquio de Souza Carneiro e Beatriz Pereira Quintão, nasceu em 03 de outubro de 1989, em Goiânia, estado de Goiás.

Em março de 2006, ingressou na Universidade Federal de Viçosa, formando-se no ensino médio pelo colégio de aplicação – COLUNI, em dezembro de 2008.

Em março de 2009, iniciou o curso de Agronomia na Universidade Federal de Viçosa, onde obteve o título em fevereiro de 2014.

Em março de 2014, iniciou o curso de Mestrado no Programa de Genética e Melhoramento na Universidade Federal de Viçosa, submetendo-se à defesa de dissertação em julho de 2015.

Em agosto de 2015, iniciou o Doutorado no Programa de Genética e Melhoramento na Universidade Federal de Viçosa, submetendo-se à defesa de tese em outubro de 2018.

SUMÁRIO

RESUMO

CARNEIRO, Vinícius Quintão, D.Sc., Universidade Federal de Viçosa, outubro de 2018. **Aplicativos computacionais para o melhoramento genético fundamentados em análise de imagens e inteligência computacional**. Orientador: Cosme Damião Cruz.

O melhoramento vegetal visa desenvolver cultivares altamente produtivas de alta qualidade física e nutricional. Cumprir esse objetivo não é processo simples, uma vez que é necessário reunir, no mesmo genótipo, elevado número de genes favoráveis para uma série de características de interesse, principalmente se considerar que o controle genético desses caracteres apresenta natureza poligênica. Portanto, para tornar o desenvolvimento de novas cultivares mais eficiente é necessário utilizar ferramentas tanto a nível de campo, laboratório e de análise de dados cada vez mais eficientes. Certas áreas tem ganhado elevado destaque no melhoramento genético como a inteligência artificial e a fenômica. A associação dos conhecimentos em fenômica e inteligência artificial podem auxiliar na solução dos principais desafios do melhoramento genético como a influência da interação genótipos por ambientes. Softwares são imprescindíveis para auxiliar nas análises por meio dessas abordagens. Portanto, o objetivo deste trabalho é disponibilizar softwares gratuitos e aplicações em inteligência artificial e fenômica com ênfase em redes neurais artificiais, lógica fuzzy e processamento digital de imagens. Com esse intuito foram desenvolvidos os softwares FENOM e BioFuzzy por meio do software Matlab em integração à linguagem Java. O software FENOM é subdividido em duas áreas de procedimentos: processamento digital de imagens e classificação por meio de redes neurais artificiais. Para processamento de imagens estão disponíveis procedimentos de aquisição, pré-processamento, segmentação e extração de características. Nos procedimentos de classificação estão disponíveis análises por redes neurais artificiais com arquitetura perceptron multicamadas. Já o software BioFuzzy disponibiliza procedimentos de sistemas de decisão fuzzy e de agrupamento fuzzy para auxiliar na recomendação de cultivares. Essas aplicações constituem em importante contribuição para o melhoramento vegetal, principalmente por visar a difusão de tecnologias como inteligência artificial, redes neurais artificiais, sistemas de tomada de decisão fuzzy e fenômica.

ABSTRACT

CARNEIRO, Vinícius Quintão, D.Sc., Universidade Federal de Viçosa, October, 2018. **Computational softwares for genetic breeding based on image analysis and computational intelligence**. Adviser: Cosme Damião Cruz.

Plant breeding aims to develop highly productive cultivars of high physical and nutritional quality. This process is not simples, since it is necessary to gather, in the same genotype, a high number of favorable genes for a series of characteristics of interest, especially considering that the genetic control of these characters has polygenic nature. Therefore, to make the development of new cultivars more efficient, it is necessary to use of new tools in field, laboratory and data analysis. Certain areas have gained high prominence in genetic breeding such as artificial intelligence and phenomics. The association of knowledge in phenomics and artificial intelligence can help in solving the main challenges of genetic breeding as the influence of genotypes by environments interaction. Softwares are essential to aid in the analysis by these approaches. Therefore, the objective of this work is to provide free softwares and applications in artificial intelligence and phenomics with emphasis on artificial neural networks, fuzzy logic and digital image processing. With this purpose, FENOM and BioFuzzy software were developed by Matlab software in Java language integration. The FENOM software is subdivided into two areas of procedures: digital image processing and classification by artificial neural networks. For image processing, acquisition, pre-processing, segmentation and feature extraction procedures are available. Artificial neural networks with architecture multilayer perceptron are available in classificatory procedures. BioFuzzy software provides procedures for fuzzy decision systems and fuzzy clustering to aid in the recommendation of cultivars. These applications constitute an important contribution to plant breeding, mainly for the diffusion of technologies such as artificial intelligence, artificial neural networks, fuzzy decision making systems and phenomics.

1. INTRODUÇÃO GERAL

A demanda por alimentos, fibras e combustíveis renováveis cresce cada vez mais no cenário mundial. Segundo Ray et al. (2013), a demanda em 2050 por esses recursos deverá ser o dobro da atual. Acredita-se que mais de 50% do aumento de produtividade obtido atualmente para diversas culturas agrícolas se deve a utilização de cultivares melhoradas. Portanto, não resta dúvidas que o melhoramento genético deverá ser ainda mais eficiente para suprir essa demanda nos próximos anos.

O melhoramento genético também desempenha importante função do ponto de vista de sustentabilidade agrícola, uma vez que contribui para menor utilização de insumos como fertilizantes e agrotóxicos e permite maior preservação do meio ambiente devido ao aumento da produtividade das lavouras. Além disso, o melhoramento de culturas que são fonte de combustíveis renováveis, a ponto de torná-las viáveis e rentáveis, desempenha importante função no cenário mundial que visa substituir qualquer fonte de combustível fóssil não renovável.

O principal objetivo do melhoramento vegetal é desenvolver cultivares altamente produtivas de alta qualidade física e nutricional (BORÉM, MIRANDA & FRITSCHE NETO, 2017). Entretanto, o desempenho desses genótipos para esses caracteres são principalmente limitados por estresses bióticos e abióticos. Portanto, desenvolver essas cultivares superiores não é processo simples, uma vez que é necessário reunir, no mesmo genótipo, elevado número de genes favoráveis para uma série de características de interesse, principalmente se considerar que o controle genético desses caracteres apresenta natureza poligênica. Devido a essa complexidade, para culturas anuais são necessários, pelos métodos convencionais, em torno de 6 a 7 anos até que a cultivar esteja disponível aos produtores (RAMALHO, SANTOS & ZIMMERMANN, 1993).

No intuito de tornar o desenvolvimento de novas cultivares mais eficiente houve, nos últimos 20 anos, elevado investimento em áreas da ciência relacionados à genética que ainda não eram empregada no melhoramento vegetal (CROSSA et al., 2017). Exemplo disso é a utilização de seleção assistida por marcadores moleculares (SAM) e, mais recentemente, dos estudos de seleção (GWS) ou associação genômica ampla (GWAS). Essas técnicas, que eram empregadas somente nas ciências médicas, atualmente são realidade no melhoramento vegetal (CROSSA et al., 2017). Com a utilização dessas técnicas moleculares é possível gerar elevado número de informações que têm auxiliado na seleção de genótipos superiores e em estudos de diversidade genética

O rápido avanço de técnicas genômicas gerou, como resultado, elevado volume de informações moleculares de alta qualidade, que tem se mostrado de grande utilidade ao melhoramento vegetal (CROSSA et al., 2017). Entretanto, somente essas informações não são suficientes para realizar estudos relacionados a SAM, GWS ou GWAS em plantas. Com esse intuito é necessário o uso associado de informações moleculares e fenotípicas (SCHUSTER & CRUZ, 2008). Observa-se que a atual dificuldade no melhoramento vegetal reside na grande disparidade de qualidade e volume entre as informações fenotípicas e moleculares dos genótipos avaliados. Enquanto os dados moleculares são obtidos por equipamentos com elevada acurácia e precisão, algumas das avaliações fenotípicas ainda são realizadas de forma manual, envolvendo análises visuais ou com equipamentos pouco acurados. Portanto, verifica-se que a utilização de tecnologias de fenotipagem de alta qualidade, denominada como fenômica, são cruciais para tornar mais eficiente o processo de identificação, seleção e recomendação de novas cultivares superiores (FRITSCHE NETO & BORÉM, 2015).

A manipulação de imagens digitais, à semelhança da genômica, era executada apenas por poucos especialistas que tinham acesso a equipamentos de alto custo (BURGER & BURGE, 2008). No entanto, a redução dos custos dos computadores com hardwares capazes de processar imagens e a facilidade de aquisição de imagens digitais por meio de dispositivos portáteis resultou em uma infinidade de informações úteis ao melhoramento vegetal. Portanto, o processamento digital de imagens se popularizou à semelhança do processamento de dados que é realizado convencionalmente no melhoramento vegetal.

A enorme quantidade de informações moleculares que são geradas já justifica considerar que o melhoramento vegetal, à semelhança de áreas como medicina, engenharia e mercado financeiro, também está na era dos conjuntos de dados com elevado volume, denominada Big Data (MA; ZHANG; WANG, 2014). Entretanto, os dados moleculares são somente uma parcela do volume de dados se considerar a fenotipagem de alta qualidade, principalmente se comparada às informações oriundas de utilização de imagens digitais. O grande desafio de processar e armazenar dados com elevado volume e variedade tem possibilitado o rápido avanço de áreas da ciência da computação como Data Science. Essa abordagem visa aplicar técnicas inovadores e novas estratégias para lidar com esse volume de informação, pois as técnicas estatísticas convencionais demonstraram ser pouco eficientes (SINGH et al., 2016).

O aprendizado de máquinas ou machine learning, área da inteligência artificial e de estudo do Data Science, oferece promissora solução computacional para analisar o elevado volume de dados que podem ser gerados pelo melhoramento vegetal. Essa abordagem multidisciplinar baseia-se em teoria de probabilidade, estatística, teoria de decisão e processos de otimização no intuito de identificar padrões que não são compreendidos pelas técnicas estatísticas convencionais (JAMES et al., 2013). Essa

abordagem tem sido aplicada para auxiliar de forma eficiente as tomadas de decisão em áreas como medicina (KOUROU et al., 2015) e mercado financeiro (CAVALCANTE et al., 2016), que requerem elevada acurácia em suas decisões. Apesar das diversas vantagens dessa abordagem, o aprendizado de máquinas ainda não tem sido amplamente utilizado para auxiliar as diversas etapas do melhoramento vegetal.

Outra dificuldade relacionada ao processamento de dados com grande volume e principalmente oriundos de avaliações fenotípicas de alta qualidade é a baixa quantidade de softwares interativos e gratuitos que tornem essas análises que são complexas em formas simples de serem utilizadas. Softwares como Matlab (MATLAB, 2018) e R (R Core Team, 2018), que são baseados em programação de matrizes de ordem elevada, têm sido utilizados para manipular e processar imagens digitais. Entretanto, esses softwares não dispõem de um arsenal de ferramentas para desenvolvimento de softwares com interfaces gráficas como linguagens similares a Java e Visual Basic. Portanto, no intuito de aproveitar as potencialidades dessas linguagens para desenvolver softwares destinados a análise de dados é necessário utilizá-las de forma integrada a softwares como o R e o Matlab.

Devido à natureza e complexidade dos dados provenientes do melhoramento vegetal, principalmente os oriundos de avaliações fenotípicas de alta qualidade, conclui-se que não basta um único software para auxiliar as decisões no melhoramento vegetal. Nesse intuito, seria necessário softwares que reúnam técnicas estatísticas e de aprendizado de máquinas com o objetivo de manipular e processar dados de diferentes naturezas, principalmente imagens digitais. Portanto, o objetivo deste trabalho é disponibilizar softwares e aplicações em inteligência artificial e fenômica com ênfase em redes neurais artificiais, lógica fuzzy e processamento digital de imagens.

2. REVISÃO BIBLIOGRÁFICA

2.1. Análises de dados aplicadas ao melhoramento vegetal

O melhoramento vegetal desempenha importante papel na sociedade moderna, pois visa principalmente, aumentar a produtividade de diversas culturas que são fontes de alimento para a população mundial. O melhorista de plantas é o pesquisador responsável por diversas atividades relacionadas ao desenvolvimento de novas cultivares. As tomadas de decisão dos melhoristas, durante todo esse processo, são embasadas principalmente nos conhecimentos em genética, biologia, fisiologia, entomologia, fitopatologia, fitotecnia e estatística (BORÉM, MIRANDA & FRITSCHE NETO, 2017). A genética voltada tanto para caracteres qualitativos como quantitativos assim como a genômica são áreas de extrema importância sob o aspecto do melhoramento vegetal. Uma vez que essas áreas são muito vastas, seria impossível um único pesquisador conseguir conduzir todo o programa de melhoramento sozinho. Portanto, verifica-se que trabalhos em parceria são essenciais para alcançar resultados nos programas de melhoramento.

A aplicação da estatística no melhoramento vegetal, denominada também como a base de desenvolvimento da Biometria, é uma das principais ferramentas para auxiliar os pesquisadores em suas tomadas de decisão. Várias são as adaptações de métodos estatísticos para auxiliar no melhoramento vegetal (CRUZ; REGAZZI; CARNEIRO, 2012; RAMALHO et al., 2012; CRUZ; CARNEIRO; REGAZZI, 2014). O ato de analisar dados oriundos do melhoramento vegetal e interpretar os resultados se tornou tão importante no melhoramento que vários pesquisadores, denominados como biometristas, se especializaram nessa área. Assim, hoje, identifica-se a área da Biometria como aquela capaz de abordar, ou modelar, processar, analisar e interpretar dados associados a

fenômenos de natureza biológica. A importância dos biometristas é facilmente constatada pelo número de análises propostas na literatura e de softwares para auxiliar nas diversas etapas dos programas de melhoramento. Se considerar a biometria atual, vários são os trabalhos clássicos propostos na literatura com o objetivo de auxiliar no estudo de populações oriundas de dialelos, na seleção de genótipos superiores ou na recomendação de cultivares.

Nos estudos relacionados a dialelos, as análises propostas por Griffing (1956), Gardner e Eberhart (1966) e Hayman (1954) constam de grandes contribuições na escolha de potenciais genitores nos programas de melhoramento. A seleção de genótipos superiores constitui outra etapa de grande importância e dificuldade no melhoramento vegetal, uma vez que novas cultivares devem agregar várias características de interesse simultaneamente (RAMALHO et al., 2012). Nesse sentido, os índices propostos por Pesek e Baker (1969) e Smith (1936) e Hazel (1943) auxiliaram em muito na identificação de genótipos que propiciem maior ganho pela seleção ao considerar um conjunto de caracteres. Já na etapa final dos programas de melhoramento, cujo o objetivo é recomendar novas cultivares, as análises propostas por Finlay e Wilkinson (1963), Eberhart e Russell (1966), Lin e Binns (1988), dentre outras, são importantes contribuições sob o aspecto de interação genótipos por ambientes. Todas essas análises são somente algumas, dentre a enormidade de propostas existentes na literatura (CRUZ; REGAZZI; CARNEIRO, 2012; RAMALHO et al., 2012; CRUZ; CARNEIRO; REGAZZI, 2014).

As propostas de análises constituem de contribuição relevante para o melhoramento vegetal. Entretanto, somente essas proposições não são suficientes para os melhoristas que tem interesse em aplicá-las. Para tornar realmente difundidas, estas devem estar disponíveis de forma simples para os pesquisadores analisar os experimentos,

interpretar os resultados e tomar decisões. Softwares como Genes (CRUZ, 2013, 2016), R (R CORE TEAM, 2018), Selegen (RESENDE, 2016), RBio (BHERING, 2017), dentre outros, são muito utilizados por melhoristas de todo o mundo. Alguns destes, como o software Genes, além de dispor de uma elevada gama de análises, são gratuitos e contam com interface gráfica interativa, que torna a sua difusão e utilização ainda maior. Entretanto, outros softwares como o R, que são amplamente difundidos nessa área, apesar de gratuito exige conhecimento em programação, o que torna mais complexa a sua utilização.

Em termos de formação complementar, atualmente, os melhoristas e, principalmente, os biometristas devem possuir conhecimento na área de linguagens de programação, seja para utilização de softwares como R, como para desenvolver novos aplicativos. Apesar de existir muitos softwares e esses disponibilizarem grande quantidade de análises, ainda existem áreas do conhecimento pouco abordadas sob o enfoque do melhoramento de plantas. A área de aprendizado de máquinas, muito aplicada em áreas como ciência da computação ainda não são difundidas no melhoramento vegetal. A falta de conhecimento a respeito dessa abordagem e de softwares para realizar essas análises são os principais motivos dessa baixa utilização. Outras áreas como as análises fenotípicas de alta qualidade, denominada como fenômica, apresentam similar desafio. Sob esse enfoque, já existe preocupação por parte dos biometristas em aprender linguagens de programação destinadas a análises de dados e também compreender novas áreas da ciência como aprendizado de máquina e fenômica afim de disponibilizar novos softwares para auxiliar no melhoramento vegetal.

2.2. Aprendizado de máquina

O aprendizado de máquina ou machine learning é área multidisciplinar da inteligência artificial que reúne conhecimentos de ciência da computação, estatística e teoria da informação com o intuito de desenvolver modelos computacionais capazes de extrair informações importantes presentes nos dados (LECUN; BENGIO; HINTON, 2015). Informações estas que não são identificadas por métodos convencionais e que propiciam melhorias na capacidade de reconhecimento de padrões. Devido a essa propriedade, técnicas de aprendizado de máquinas têm recebido grande atenção no cenário mundial, principalmente por auxiliar na solução de problemas de agrupamento, classificação e predição (SINGH et al., 2016). Essa área não está restrita somente à literatura especializada, mas também em aplicações rotineiras da sociedade moderna como em sites de busca, redes sociais, "e-commerce" e equipamentos como câmeras e smartphones.

O atual crescimento na utilização do aprendizado de máquinas se deve a um conjunto de fatores dos quais o principal é o aumento no volume de informação a ser processada, o que propiciou o surgimento de conhecimentos específicos em dados de elevado volume, comumente denominada de Big Data (MA; ZHANG; WANG, 2014). O aprendizado de máquinas se tornou essencial na análise de dados com elevado volume e variedade, uma vez que os métodos estatísticos utilizados em dados de maior dimensão não propiciam soluções eficientes. Portanto, estudos aplicados as áreas financeira, médica, genômica de qualquer natureza e processamento de sinais têm se baseado na utilização dessa abordagem.

No melhoramento genético, apesar de ainda pouco difundida, existem aplicações dessa abordagem tanto a nível fenotípico (BATCHELOR; YANG; TSCHANZ, 1997;

KAUL; HILL; WALTHALL, 2005; BARBOSA et al., 2011; CARNEIRO et al., 2017) como molecular (GIANOLA et al., 2011; EHRET et al., 2015; GONZÁLEZ-CAMACHO et al., 2016). Devido à natureza dos dados moleculares que se caracterizam como de elevado volume (Big data) e à necessidade de técnicas de elevado desempenho preditivo, vários são os esforços de utilização dessa abordagem para auxiliar em estudos de seleção e/ou associação genômica ampla (MA; ZHANG; WANG, 2014; CROSSA et al., 2017). Áreas de estudo mais recentes no melhoramento vegetal como o processamento de dados oriundos de avaliações fenotípicas de alta qualidade (fenômica) caracterizam-se ainda mais nessa abordagem de big data, o que torna ainda mais importante o entendimento das diferentes técnicas de aprendizado de máquinas.

A obtenção das estimativas dos parâmetros dos modelos de aprendizado de máquinas (calibração do modelo) consiste na etapa de treinamento na qual são utilizados algoritmos capazes de reconhecer determinados padrões por meio de exemplos representativos de uma situação (HAYKIN, 2008). Os atuais algoritmos podem ser divididos conforme o tipo de problema a ser solucionado em duas principais formas de aprendizado: supervisionado ou não supervisionado. Baseada nesses princípios de aprendizado de máquinas, uma nova frente de estudo mais recente e considerada por alguns como a terceira geração do aprendizado de máquinas é o comumente denominada deep learning (LECUN; BENGIO; HINTON, 2015; SCHMIDHUBER, 2015). Abordagem essa que utiliza uma nova forma de aprendizado mais complexa, porém com evidências de maior eficiência que os modelos de aprendizado de máquinas convencionais.

2.2.1. Aprendizado supervisionado

O aprendizado supervisionado consiste em realizar o treinamento dos modelos computacionais por meio de exemplos em que se conhece tanto as características (entradas) como o padrão de resposta (saída) das observações (KURTENER; DRAGAVTSEV, 2017). Entretanto, não basta somente estimar os parâmetros, o modelo gerado deve possuir elevada capacidade de generalização, portanto, a etapa de validação do modelo é de extrema importância, pois por meio dela será possível verificar se o modelo pode ser utilizado em novos conjuntos de dados com elevada eficácia. Dessa forma, verifica-se que existe a necessidade de subdividir o conjunto de dados em duas partes, uma para treinamento e outra para validação. Geralmente 75% dos dados são utilizados no treinamento e os 25% restantes são utilizados na validação. Na calibração do modelo é de interesse que, em ambas as etapas, seja obtida elevada capacidade preditiva. No intuito de constatar a eficiência dos modelos gerados, muitos estudos utilizam-se de outras estratégias de validação cruzada como a técnica de k-folds, que consiste na realização do treinamento e validação por várias vezes (k vezes), com variação dos dados utilizados no treinamento e validação (JAMES et al., 2013). Após o treinamento e validação, o modelo pode ser utilizado para predizer o padrão de resposta de novas observações em que ainda não se conhece esse padrão.

Aprendizagem supervisionada têm sido empregada em tarefas de classificação (SANT'ANNA et al., 2015; CARNEIRO et al., 2017) e predição (SILVA et al., 2014). Modelos dedicados à classificação visam predizer as classes das observações de interesse, ou seja, no melhoramento vegetal por exemplo seria a classificação de genótipos por severidade de doenças, por estrutura genética, entre outros padrões. Os modelos cujo o intuito é predição visam estimar tendências e a relação entre os valores reais e preditos

pelas técnicas de aprendizado de máquinas. Esse tipo de tarefa é de grande interesse no melhoramento vegetal, principalmente em estudos de seleção e/ou associação genômica ampla. Devido aos diferentes princípios abordados por esses algoritmos existem ampla gama de técnicas de aprendizado de máquinas como redes neurais artificiais, árvore de decisão, máquina de vetor de suporte, k-vizinhos mais próximos, entre outras (JANG; SUN; MIZUTANI, 2012).

2.2.1.1. Redes Neurais Artificiais

As redes neurais artificiais (RNA's) são técnicas computacionais baseadas em modelos matemáticos (HAYKIN, 2008), que apresentam funcionamento inspirado no cérebro humano. De forma geral, as RNA's apresentam estrutura e funcionamento semelhante à rede de neurônios biológicos do sistema nervoso humano, ou seja, são compostas por uma rede de unidades de processamento (neurônios artificiais) interconectadas responsável por aprender por meio de exemplos. Devido a essa propriedade, as RNA's são empregadas predominantemente na solução de problemas classificatórios, preditivos e de agrupamento.

O primeiro modelo artificial de um neurônio biológico foi proposto por McCulloch e Pitts (1943) (Figura 1). Apesar de simples, este já englobava as principais características de uma rede neural artificial. Os princípios propostos por McCulloch e Pitts (1943) serviram de base para o desenvolvimento de várias arquiteturas de RNA's que foram propostas posteriormente (ROSENBLATT, 1958; WIDROW; HOFF, 1960 GROSSBERG, 1980; HOPFIELD, 1982; RUMELHART; HINTON; WILLIAMS, 1986; KOHONEN, 1998).

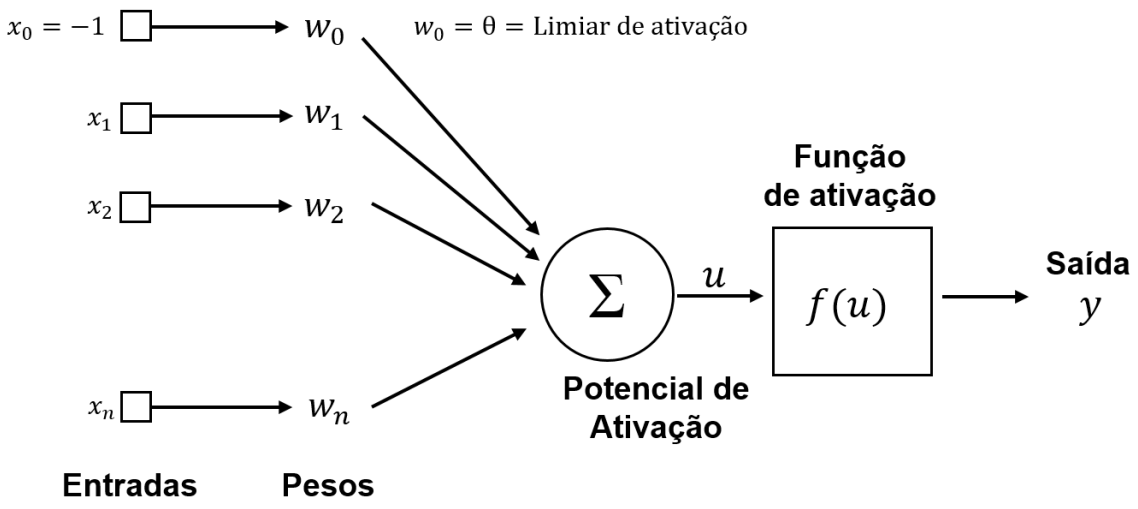


Figura 1. Modelo de neurônio artificial proposto por McCulloch e Pitts (1943)

Os neurônios artificiais apresentam quatro importantes componentes. O primeiro são as entradas (x) que constitui as variáveis preditoras de um problema. O segundo é o conjunto de pesos sinápticos (w) capaz de ponderar cada variável entrada. O terceiro é o combinador linear (Σ) capaz de agregar as entradas ponderadas pelos seus respectivos pesos em uma única medida denominada porta do limiar. O quarto são as funções de ativação (f(u)) capazes de regular o funcionamento dos neurônios de acordo com o potencial de ativação (u), que é a diferença entre a porta do limiar e um limiar de ativação (θ) próprio de cada neurônio conforme a expressão a seguir (HAYKIN, 2008):

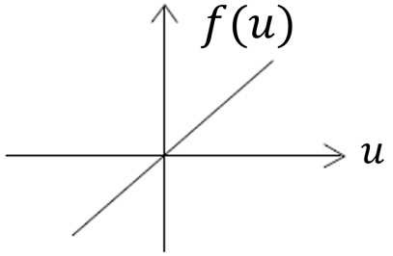
$$u = \sum_{i=0}^{n} w_i x_i \; = \; \sum_{i=1}^{n} w_i x_i \; - \; \theta$$

As funções de ativação são responsáveis por gerar um valor de saída ($y = f(u)$) em função de determinado conjunto de sinais de entrada no neurônio (BRAGA; CARVALHO; LUDEMIR, 2011). Existem diferentes funções de ativação e cada uma delas com suas respectivas particularidades. Uma característica que as diferencia é quanto ao valor de saída dos neurônios, que pode variar de [-∞, + ∞], [0, + ∞], [0,1], [-1,1], entre outros intervalos. As principais funções de ativação utilizadas são: linear, sigmoide,

tangente hiperbólica, unidade de retificação linear (Relu) e exponencial normalizada (softmax) (Figura 2).

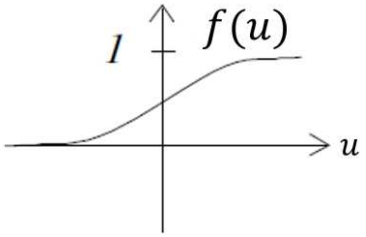
Função de Ativação Linear

$$f(u) = u$$



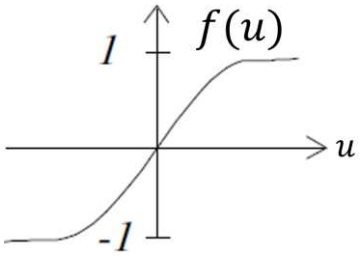
Função de Ativação Sigmoide

$$f(u) = \frac{1}{1 + e^{-u}}$$



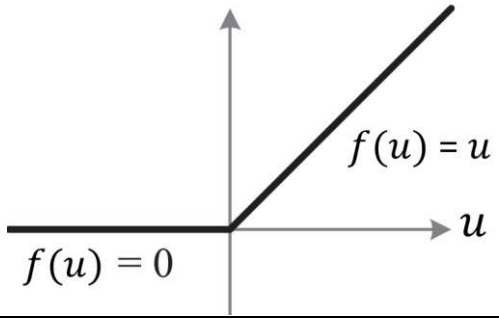
Função de Ativação Tangente Hiperbólica

$$f(u) = \frac{1 - e^{-2u}}{1 + e^{-2u}}$$



Função de Ativação Unidade de Retificação Linear (Relu)

$$f(u) = \max(0, u)$$



Função de Ativação Exponencial Normalizada (Softmax)

$$f(u)_j = \frac{e^{u_j}}{\sum_{k=1}^{K} e^{u_k}} \quad \text{para j = 1, ..., K}$$

Figura 2. Principais funções de ativação: linear, sigmoide, tangente hiperbólica, unidade de retificação linear e exponencial normalizada (Softmax).

O neurônio artificial proposto por McCulloch e Pitts (1943), apesar de muito rudimentar e capaz de solucionar somente problemas de natureza linear, é a base de

arquiteturas altamente sofisticadas como o perceptron múltiplas camadas (MLP) (Figura 3) (RUMELHART; HINTON; WILLIAMS, 1986). Essa arquitetura por sua vez é capaz de solucionar problemas classificatórios e preditivos de natureza não linear.

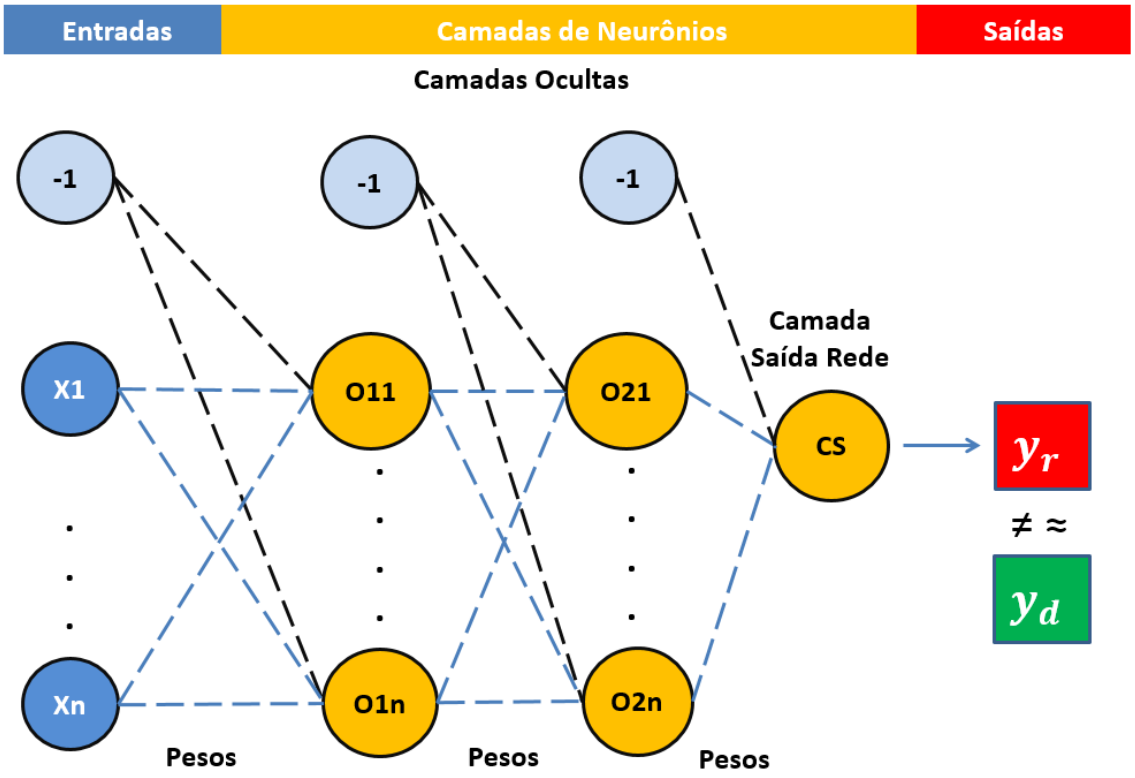


Figura 3. Rede neural artificial com arquitetura perceptron múltiplas camadas.

De forma geral, a arquitetura MLP pode ser dividida em três estruturas. A primeira delas é a camada de entradas (x) responsável por receber toda a informação relativa aos exemplos apresentados à RNA. As camadas ocultas (o), a segunda estrutura, são responsáveis por processar e extrair padrões presentes nos exemplos apresentados à entrada. O MLP pode ser constituído por uma ou múltiplas camadas ocultas de neurônios, assim como o número de neurônios e as funções de ativação em cada camada podem variar (RUMELHART ET AL., 1986). O número de camadas, de neurônios e quais funções de ativação são utilizadas definem a topologia de uma RNA MLP. A escolha correta da topologia é fundamental para o sucesso na solução tanto de problemas

classificatórios como preditivos. Essa topologia varia conforme a natureza e complexidade do padrão a ser reconhecido pela RNA.

A terceira estrutura é constituída por uma única camada de neurônios responsável por agregar toda a informação processada nas camadas ocultas e retornar os resultados da RNA, que é conhecida como camada de saída. O número de neurônios e o tipo de função de ativação dessa camada também podem variar. Por exemplo, em problemas classificatórios utiliza-se geralmente um número de neurônios na camada de saída igual ao número de classes do problema e se utiliza a função softmax (Figura 2), pois por meio dessa é possível estimar a probabilidade associada de classificação de cada observação em cada classe em estudo. Em problemas de predição é utilizado somente um único neurônio nesta camada com função linear (Figura 2).

O principal diferencial das RNA's em relação às técnicas estatísticas de classificação e predição é a forma com que os parâmetros (pesos) dos neurônios são estimados. Esse processo é realizado por meio de algoritmos de aprendizado que são capazes de reconhecer padrões presentes nos exemplos fornecidos no treinamento. As RNA’s MLP utilizam nessa etapa algoritmos de aprendizado supervisionado baseados no método do gradiente descendente (JANG; SUN; MIZUTANI, 2012). Um destes é o algoritmo backpropagation considerado base para todos os demais existentes na literatura (HAYKIN, 2008).

O algoritmo backpropagation é um processo iterativo inspirado no princípio de retroalimentação comumente realizado mediante duas fases bem específicas. A primeira fase denominada forward (propagação adiante) consiste em fornecer as entradas ao MLP para serem processadas de forma a obter as respectivas saídas preditas. Como o treinamento dessa arquitetura é supervisionado as saídas preditas pelo MLP são comparadas às saídas reais, o que permite a obtenção da estimativa de uma função de

perda. Uma das principais funções de perda é o erro quadrático médio conforme expressão a seguir:

$$EQM = \frac{1}{N}\sum_{i=1}^{N} 0.5 * (y_{ri} - y_{di})^2,$$

em que y_{ri} é a saída da RNA da i-ésima observação e y_{di} é saída real (desejada).

Em uma segunda fase conhecida como backward (propagação reversa), os pesos são ajustados com base no valor da função de perda obtido (SILVA; SPATTI; FLAUZINO, 2010). De forma geral, essas duas fases são processadas com o intuito de reduzir o valor da função de perda, de forma que cada reajuste dos pesos é conhecido como iteração ou época. Portanto, no treinamento inicialmente os pesos são gerados de modo aleatório e posteriormente são realizadas várias iterações com o intuito de ajustá-los até se obter um valor de interesse da função de perda definido a priori ou até um número limite de iterações.

Os estudos que aplicam MLP utilizam modificações do algoritmo de aprendizado backpropagation. Dentre esses destacam-se o algoritmo proposto por Levenberg–Marquardt (MARQUARDT, 1963) e o de regularização bayesiana (GIANOLA et al., 2011), uma vez que os resultados obtidos por meio desses algoritmos apresentam elevada eficiência. Devido a esse potencial, a arquitetura de MLP tem sido amplamente empregada na solução de problemas financeiros, médicos, de reconhecimento de sinais e principalmente voltados a visão computacional.

Nos últimos anos houve um aumento considerável da utilização das RNA's aplicadas tanto em estudos com caracteres fenotípicos quanto moleculares no melhoramento vegetal, uma vez que a maior parte dos problemas encontrados no melhoramento são de natureza classificatória ou de predição. Essa ferramenta já foi aplicada eficientemente para discriminar populações geneticamente muito próximas (SANT'ANNA et al., 2015), o que evidencia o potencial dessa abordagem em estudos de

diversidade genética. Carneiro et al. (2017) constataram que as RNA's MLP são eficientes para identificar genótipos superiores de feijão quanto a arquitetura de plantas. Alguns estudos têm proposto também a utilização das RNA's MLP para auxiliar os estudos de interação genótipos por ambientes e principalmente na recomendação de cultivares (TEODORO et al., 2015).

Sob o enfoque de predição, Silva et al. (2014) verificaram que as RNA's são ferramentas úteis para estimar valores genéticos e consequentemente selecionar genótipos superiores. Entretanto, as RNA's MLP tem sido utilizada principalmente em estudos relacionados a seleção e/ou associação genômica ampla, devido ao grande volume de dados moleculares a serem analisados e principalmente pela elevada capacidade preditiva dessa ferramenta (CROSSA et al., 2017; EHRET et al., 2015; GIANOLA et al., 2011). Apesar da seleção genômica ampla ser abordada como um problema de predição, outros autores têm utilizado RNA's MLP em problemas classificatórios de mesma natureza (GONZÁLEZ-CAMACHO et al., 2016).

As RNA's também foram empregadas com sucesso em estudos relacionados à fenômica. Nessa área, a abordagem de RNA's foi aplicada com sucesso na identificação e classificação de doenças de diversas culturas (SINGH et al., 2016). Entretanto, outras formas de RNA's baseadas em deep learning como as redes neurais convolucionais têm sido utilizadas com o intuito de tornar mais eficiente o processo de reconhecimento de padrões por meio de imagens digitais (CRUZ et al., 2017).

2.2.2. Aprendizado não supervisionado

Técnicas de aprendizado não supervisionado dispensam o conhecimento do padrão de respostas das observações. Devido a essa natureza, essa forma de aprendizagem

tem sido aplicada para agrupar observações por padrões mais similares. A técnica de k-Means é amplamente empregada em diversas áreas inclusive no melhoramento vegetal (CASTRO et al., 2017). Entretanto, esta é somente uma das diversas ferramentas utilizadas com essa finalidade. Dentre essas, ressalta-se o método de agrupamento fuzzy C-Means, que constitui uma modificação do método k-Means (BEZDEK; EHRLICH; FULL, 1984). Além desses, porém com ênfase em estrutura organizacional das observações se destaca os mapas auto organizáveis propostos por Kohonen (1982).

2.2.2.1. K-Means

O método k-Means proposto por Macqueen (1967) é um algoritmo de aprendizado não supervisionado cuja a finalidade é agrupar observações similares em um número k de grupos definido previamente. Com essa finalidade, esse algoritmo visa minimizar uma função critério baseada na distância euclidiana como medida de dissimilaridade (MIRANDA, 2011), conforme a seguinte expressão:

$$J(w,z) = \sum_{i=1}^{n}\sum_{j=1}^{k}\left\|x_i - z_j\right\|^2$$

em que n é o número de observações, x_i é a localização da i-ésima observação, z_j é o j-ésimo centroide.

O processo iterativo do algoritmo k-Means segue os passos: (i) Definir os k grupos e os seus respectivos centroides de forma aleatória ou com base em algum critério; (ii) atribuir cada observação ao grupo com centroide mais próximo; (iii) recalcular os centroides dos grupos e alocar novamente as observações; e (iv) determinar um critério de convergência e se este não for satisfeito realizar novamente a etapa ii.

O critério de convergência é de extrema importância no processo iterativo. Um critério típico é definir uma mínima re-atribuição de objetos para novos grupos, realizada na etapa iii. Além desse, pode-se determinar que a função objetivo seja menor que uma dada tolerância, conforme a seguinte expressão: $Se\ |J_1^m - J_2^m| < \varepsilon$, em que J_1^m é a função critério na iteração m calculada com base nos centroides da etapa ii e J_2^m é similar, entretanto é calculada após alocação das observações nos novos centroides da etapa iii (JANG; SUN; MIZUTANI, 2012).

Outro ponto importante de ser ressaltado é o número de grupos definidos previamente e os seus respectivos centroides. O número k de grupos pode ser definido pelo conhecimento prévio sobre as observações. Exemplo disso, pode-se utilizar a informação de localização onde essas observações foram obtidas ou alguma outra particularidade. Em termos de metodologias que auxiliam nessa escolha, poucas são as alternativas e as que existem também não são totalmente eficientes. Por isso, uma das grandes dificuldades enfrentadas ao utilizar esse método consiste nessa etapa.

Uma vez definido o número de grupos, é necessário também definir a posição inicial dos possíveis centroides. Essa etapa é de grande importância, pois os resultados de agrupamento por esse método são influenciados pelos centroides iniciais. Apesar disso, uma das formas mais utilizadas é escolher de forma aleatória os centroides (JANG; SUN; MIZUTANI, 2012). Outra estratégia mais elegante é executar esse algoritmo diversas vezes com partições iniciais aleatórias, o que melhora os agrupamentos das observações (MIRANDA, 2011).

A metodologia de k-Means têm sido utilizada para agrupar eficientemente grande volume de observações. Essa característica qualifica essa metodologia para ser utilizada em pesquisas que se utiliza imagens digitais, já que o número de pixels em imagens com elevada dimensão é muito grande. Métodos não supervisionados como k-Means são

muito utilizados como técnicas de segmentação de imagens de forma a agrupar pixels com mesmas intensidades de cor (DHANACHANDRA; MANGLEM; CHANU, 2015; LI; HE; WEN, 2015).

2.2.2.2. Lógica Fuzzy e o método fuzzy C-Means

A lógica fuzzy, proposta por Zadeh (1965), constitui uma subárea da inteligência artificial, inspirada na forma de raciocínio humano (ZADEH, 1975, 1997), que possibilita análises e interpretações de informações aproximadas (ZADEH, 1983). Esta técnica provê uma forma de subdividir o universo de discurso das variáveis de importância para a tomada de decisão em conjuntos que representam expressões verbais comuns na comunicação humana (ZADEH, 1996; SIMÕES E SHAW, 2007). A nível de teoria de conjuntos fuzzy, qualquer observação pode ser alocada a dois ou mais conjuntos diferentes simultaneamente com níveis de pertencimento distintos que variam entre 0 e 1 (ZADEH, 1965). Esses níveis são comumente denominados por pertinências e são definidos por funções matemáticas, o que constitui como o principal diferencial dessa abordagem (ZADEH, 1999). A lógica fuzzy, por apresentar essa propriedade de multivalência, permite converter a experiência humana em uma forma compreensível pelos computadores (ZADEH, 2004, 2008).

As particularidades da lógica fuzzy a qualificam como abordagem muito útil em diversas situações especialmente em tomadas de decisão (CARNEIRO et al., 2018) e como técnica de agrupamento (BEZDEK, 1984; GHOSH E DUBEY, 2013; DOGRUPARMAK et al., 2014). Essa abordagem é a base para os sistemas de tomada de decisão fuzzy, aplicados em algumas áreas como a medicina (KORENEVSKIY, 2015).

Os sistemas de tomada de decisão fuzzy utilizam toda a teoria da lógica fuzzy com o intuito de desenvolver sistemas capazes de auxiliar em tomadas de decisão em diferentes ocasiões. Para isso as variáveis que determinam as decisões são tratadas sob o enfoque dessa abordagem, ou seja, são consideradas como variáveis de entrada fuzzy do sistema. A partir desse momento, o universo de discurso de cada uma dessas variáveis é particionado em grupos conforme o conhecimento do desenvolvedor do sistema a respeito de cada variável (SIMÕES E SHAW, 2007). Por exemplo, a variável produtividade de grãos na cultura do feijão-comum pode ser particionada em quatro grupos conforme a Figura 4.

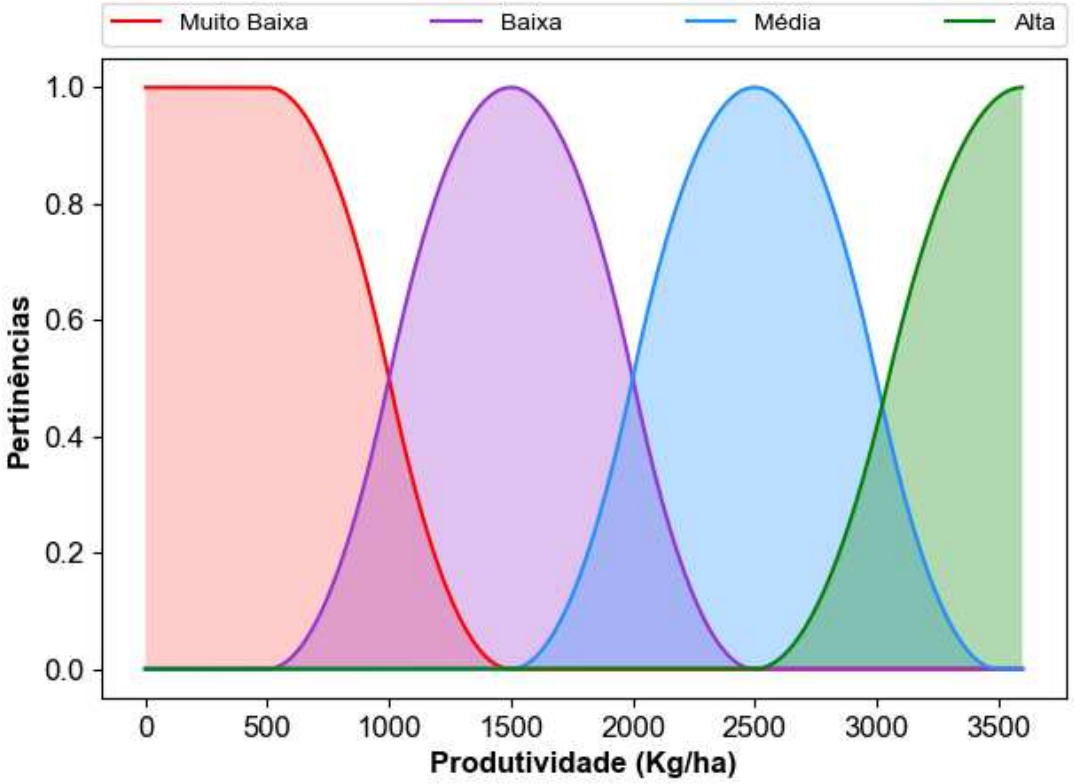


Figura 4. Representação da variável produtividade de grãos (Kg.ha^{-1}) de feijão-comum por meio da lógica fuzzy

Os conjuntos fuzzy podem ser representados por meio de diferentes funções de pertinências, que podem ser lineares ou não (JANG et al., 2012). De acordo com essas

funções de pertinência, cada valor do universo de discurso está associado a um grau de pertencimento (pertinência) aos grupos que compõem cada variável.

Além das variáveis de entrada, também existe uma variável de saída que reúne todo o conhecimento do sistema, que também é considerada sob o enfoque fuzzy. Portanto, existe uma relação de dependência entre a variável de saída e as de entrada. Essa relação é determinada pelas regras fuzzy que agregam informações das variáveis de entrada de forma a gerar uma saída. Essas regras são estabelecidas conforme o conhecimento empírico do desenvolvedor do sistema fuzzy sobre o problema em questão. Portanto, é imprescindível que esse desenvolvedor seja um especialista na área que deseja desenvolver esse sistema, uma vez que são essas regras que determinam a tomada de decisão.

Os dados coletados são processados nos sistemas de tomada de decisão fuzzy por uma série de etapas (Figura 5). A primeira delas é a fuzzyficação na qual os dados numéricos para cada variável são submetidos às suas respectivas funções de pertinência de modo a obter a pertinência das observações nos grupos que compõe cada uma das variáveis. Esses valores de pertinências são utilizados na etapa de implicação, que consiste no processamento de cada regra individualmente de modo a obter os valores de pertinências para a observação em questão na variável de saída. Portanto, para cada regra são obtidas as pertinências na variável de saída. A agregação consiste em combinar as saídas de todas as regras em um único resultado (JANG et al., 2012). Entretanto, esse resultado está na forma de pertinência e muitas vezes é necessário converte-lo em valores discretos, que é realizado na etapa de defuzzyficação. Todo o processamento dos dados em um sistema fuzzy é relativamente simples em que o custo computacional é dependente do número de variáveis de entrada e de funções de pertinência.

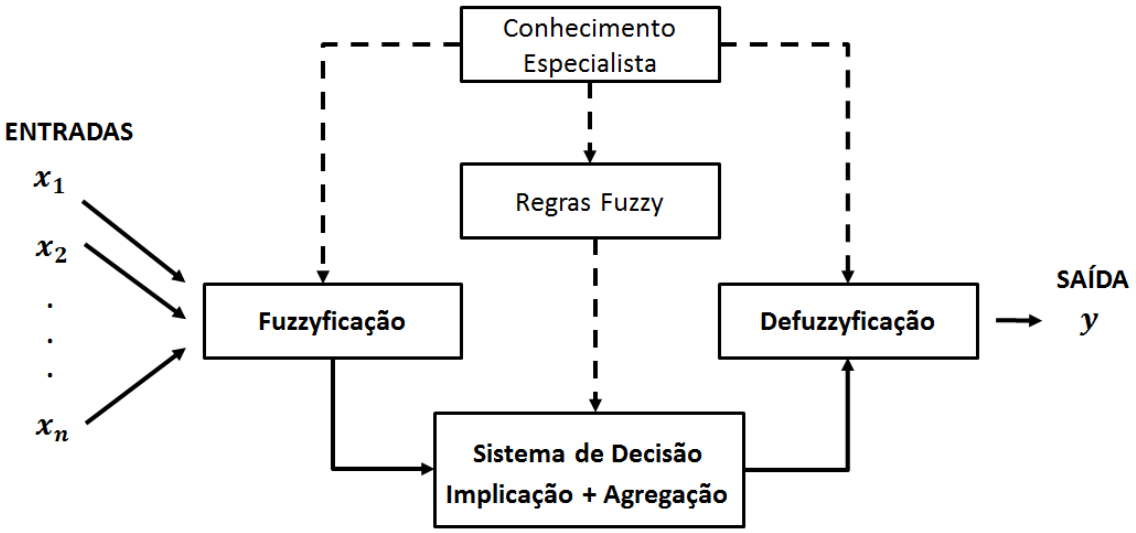


Figura 5. Estrutura de sistema de tomada de decisão fuzzy.

Sistemas dessa natureza, apesar de ainda não ser amplamente difundidos no melhoramento vegetal, podem ser adaptados para auxiliar nas tomadas de decisão rotineiras dos programas de melhoramento. Carneiro et al. (2018) propuseram a utilização de sistemas dessa natureza para auxiliar nas decisões de recomendação de cultivares.

A lógica fuzzy também é base para implementação de procedimentos de agrupamento, como fuzzy C-Means, muito utilizada no agrupamento de grandes conjuntos de dados e, principalmente, em segmentação de imagens (FENG ET AL., 2016). A semelhança do k-Means, o método de fuzzy C-Means proposto por Bezdek et al. (1984) é baseado no agrupamento de observações em um número de grupos pré-determinados, porém sob o enfoque da lógica fuzzy.

A alocação das observações em cada grupo pelo método fuzzy C-Means baseia-se na minimização da função custo:

$$\mathrm{J}(\mathrm{U}, \mathrm{c}_1, \ldots, \mathrm{c}_\mathrm{c}) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} d_{ij}^{2} = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} \left\| \mathrm{c_i} - \mathrm{x_j} \right\|^2,$$

que pode ser reescrita como:

$$\bar{J}(\mathrm{U}, \mathrm{c}_1, \ldots, \mathrm{c}_\mathrm{c}, \lambda_1, \ldots, \lambda_\mathrm{n}) = \sum_{\mathrm{i}=1}^{\mathrm{c}} \left(\sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{u_{ij}^{m}} \left\| \mathrm{c_i} - \mathrm{x_j} \right\|^2 \right) + \sum_{\mathrm{j}=1}^{\mathrm{n}} \lambda_\mathrm{j} \left(\sum_{\mathrm{i}=1}^{\mathrm{c}} \mathrm{u_{ij}} - 1 \right)$$ ao

considerar as n restrições $\sum_{\mathrm{i}=1}^{\mathrm{c}} \mathrm{u_{ij}} = 1, \ \ \forall \, \mathrm{j} = 1, \ldots, \mathrm{n}$. Nessas funções, U é a matriz de

pertinências ($u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}}$) das j observações nos i grupos; c_i é o vetor de informações do centroide do i-ésimo grupo ($c_i = \frac{\sum_j^n u_{ij}^m x_j}{\sum_j^n u_{ij}^m}$); m é o expoente de ponderação fuzzy (índice de difusividade), d_{ij}^2 é a distância euclidiana entre o vetor de informações de cada observação (x_j) e o i-ésimo centroide (c_i) e λ_j são os multiplicadores de Lagrange para as n restrições ($\sum_{i=1}^{c} u_{ij} = 1\ ,\ \ \forall\, j = 1, \dots, n$).

A técnica fuzzy C-Means baseia-se em processo iterativo para alocar as observações e determinar as pertinências destas em cada grupo conforme os seguintes passos: i. inicialização da matriz U com valores aleatórios de pertinências entre 0 e 1, de modo que a restrição $\sum_{i=1}^{c} u_{ij} = 1$ seja respeitada; ii. obtenção dos centroides dos c grupos por meio da expressão $c_i = \frac{\sum_j^n u_{ij}^m x_j}{\sum_j^n u_{ij}^m}$; iii. obtenção do valor da função custo e observação do critério de parada do processo iterativo; iv. em caso de não cumprimento do critério de parada, o processo iterativo reinicia de modo que se obtém uma nova matriz U de pertinências por meio da expressão $u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}}$ (MIRANDA, 2011; JANG et al., 2012; BEZDEK 1984).

2.2.2.3. Mapas auto organizáveis

O mapa auto organizável (self-organization map - SOM) proposto por Kohonen (1982) é uma RNA de arquitetura reticulada inspirada nos princípios biológicos do córtex cerebral, onde a ativação de uma região específica corresponde à resposta frente a determinado estímulo sensorial. Essa técnica apresenta enfoque de aprendizado não supervisionado e por isso é muito utilizada como técnica de agrupamento. Os mapas auto

organizáveis visam agrupar observações em classes por meio de similaridades e regularidades entre as observações em estudo.

A principal diferença das RNA's comumente utilizadas em relação aos mapas auto organizáveis está situada no seu tipo de treinamento que é de natureza competitiva (SILVA, SPATTI, FLAUZINO, 2010). O SOM é constituído de apenas uma camada de neurônios estruturados em mapas topológicos de uma (Figura 6A) ou duas dimensões (Figura 6B). Os mapas topológicos informam como os neurônios dessa camada estão organizados espacialmente frente ao comportamento de seus vizinhos.

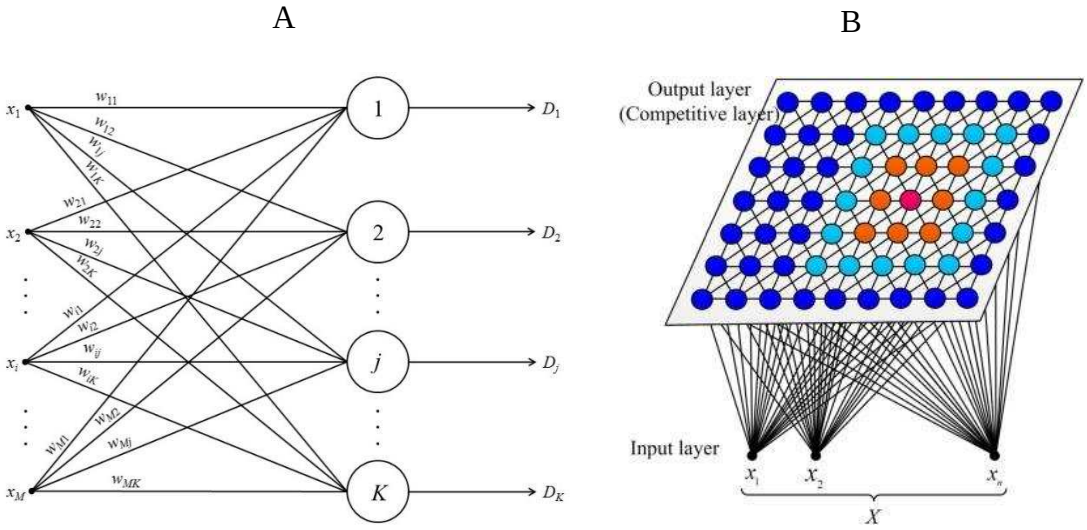


Figura 6. Mapas topológicos de um mapa auto organizável. A – Mapa topológico com uma dimensão (k x 1). B – Mapa topológico com duas dimensões (m x n) (LIAO et al., 2015).

No aprendizado competitivo admite-se que existe concorrência entre todos os neurônios do mapa topológico, de modo que um único seja considerado vencedor. Caso o neurônio vencedor não esteja conectado a nenhum outro, como na estratégia do vencedor leva tudo, somente o seu vetor de pesos é ajustado. Entretanto, se os neurônios estão conectados, os vetores de pesos dos neurônios vizinhos podem ser ajustados conforme o critério de vizinhança adotado.

A definição do neurônio vencedor é baseada na proximidade existente entre o vetor de pesos de cada neurônio e vetor de entradas para cada observação (JANG; SUN; MIZUTANI, 2012). O neurônio vencedor será aquele que possuir menor distância em relação a observação. A distância euclidiana tem sido utilizada para determinar essa proximidade. O vetor de pesos do neurônio vencedor poderá ser ajustado de modo que este se aproxime ainda mais da observação, conforme o seguinte ajuste:

$$w_{(vencedor)}^{i+1} = w_{(vencedor)}^{i} + \eta(x_j - w_{(vencedor)}^{i})$$

em que $w_{(vencedor)}^{i}$ é o vetor de pesos do neurônio vencedor na iteração i, η é a taxa de aprendizado e x_j é o vetor de entradas da j-ésima observação.

O treinamento competitivo é condicionado a estrutura dos mapas topológicos e do critério de vizinhança inter neurônios. O critério de vizinhança consiste em especificar um raio de abrangência em relação aos neurônios (SILVA, SPATTI, FLAUZINO, 2010). Valores diferentes de raio interferem na quantidade de vetores de pesos que serão ajustados. Maiores valores de raio proporcionam maior quantidade de vetores de pesos a serem ajustados. O ajuste dos pesos é realizado tanto no vetor de pesos do neurônio vencedor como também dos seus vizinhos. Entretanto, esse ajuste é proporcional à distância dos neurônios em relação ao neurônio vencedor. Uma estratégia utilizada quando se utiliza raio igual a uma unidade é assumir o ajuste dos neurônios vizinhos ao vencedor igual a metade da taxa de aprendizado.

$$w_{(vizinho)}^{i+1} = w_{(vizinho)}^{i} + 0{,}5 * \eta(x_j - w_{(vizinho)}^{i})$$

em que $w_{(vizinho)}^{i}$ é o vetor de pesos do neurônio vizinho ao vencedor na iteração i, η é a taxa de aprendizado e x_j é o vetor de entradas da j-ésima observação.

Outra estratégia utilizada principalmente em situações em que o raio é superior a uma unidade é aplicar a função gaussiana ao ajuste do vetor de pesos dos neurônios vizinhos ao vencedor conforme expressão a seguir:

$$w_{(vizinho)}^{i+1} = w_{(vizinho)}^{i} + \alpha^{(\Omega)} * \eta(x_j - w_{(vizinho)}^{i})$$

em que $\alpha^{(\Omega)}$ é o operador de vizinhança dado por:

$$\alpha^{(\Omega)} = e^{-\frac{\|w_{(vencedor)} - w_{(vizinho)}\|^2}{2\sigma^2}}$$

em que σ é a "largura efetiva" da função de vizinhança topológica gaussiana.

O SOM por ser baseado em aprendizado não supervisionado requer o conhecimento prévio do número de classes presentes para estruturar os mapas topológicos (KOHONEN, 1998). Portanto, o conhecimento a priori sobre as observações em estudo são fundamentais para auxiliar na definição dos mapas topológicos. Esses mapas podem conter o número de neurônios igual ou superior ao número de classes. Em situações que se utiliza mapas topológicos com número de neurônios superior ao número de classes, essa estrutura pode ser particionada de forma que dois ou mais neurônios constituem uma mesma classe, o que origina os mapas de contexto (SILVA; SPATTI; FLAUZINO, 2010). Após obtido um SOM, este pode ser utilizado para classificar novas observações não utilizadas no processo de treinamento.

2.3. Fenômica – Processamento digital de imagens

Imagens são informações visuais armazenadas sob diferentes condições. Até o advento da computação, as imagens eram armazenadas sob forma física como papel ou filmes fotográficos, por meio de pintura em tinta ou processos com uso de produtos químicos fotossensíveis (HUNT, 2010). Entretanto, essas formas sofrem grande influência das condições ambientais o que pode ocasionar em perda de qualidade ou até mesmo a perda total das imagens. Além disso, a cópias de imagens dessa natureza são pouco precisas e de baixa qualidade.

As imagens digitais, por sua vez, são informações visuais armazenadas em forma digital e, portanto, podem ser visualizadas em sistemas computacionais. Uma vez que imagens dessa natureza são armazenadas na forma binária em sistemas computacionais, estas não estão sobre a influência ambiental, ou seja, não estão condicionadas a perda de qualidade. Outra grande vantagem das imagens digitais é que estas podem ser copiadas rapidamente diversas vezes sem perda de qualidade e também enviadas para qualquer lugar do mundo com grande facilidade.

O número de dispositivos de aquisição de imagens digitais cresceu de forma muito rápida nos últimos anos. Essa tecnologia que até pouco tempo era restrita a uma parcela pequena da população, hoje está presente em uma ampla gama de celulares, câmeras digitais ou scanners, principalmente devido a redução dos custos para se obter esses dispositivos. Como resultado dessa tecnologia se tornar tão acessível se verificou um aumento enorme nas imagens digitais geradas tanto para fins pessoais como acadêmicos. Além do processamento, o grande volume de imagens com fins acadêmicos tornou a utilização com esse propósito em um problema de Big Data (CRUZ et al., 2017). Por isso, muitas das vezes esses dois temas são tratados associados.

A utilização de imagens digitais para fins acadêmicos, até pouco tempo, era restrita a um pequeno grupo de especialistas, principalmente, pois os equipamentos de aquisição de imagens e também de processamento dessas informações eram de elevado custo (BURGER & BURGE, 2008). Entretanto, a redução desses custos, a possibilidade de qualquer pessoa ter acesso a computadores com elevada capacidade de processamento e a disponibilidade de pacotes de análises tornaram possível o crescimento das áreas de processamento digital de imagens, visão computacional e da fenômica no melhoramento vegetal (SABROL, 2015).

O processamento digital de imagens é a área da ciência que visa analisar, processar ou melhorar imagens digitais para alcançar um resultado desejado. Diversas são as técnicas utilizadas com o objetivo de transformar uma imagem em outra melhorada ou para analisá-la para obter informações específicas. Entretanto, em algumas situações em que se tem interesse em informações mais complexas presentes nas imagens utilizar somente essas técnicas não é suficiente. Por isso, outro campo que cresceu muito nos últimos anos foi a visão computacional. Denominada em inglês como computer vision, essa área visa compreender e interpretar padrões complexos presentes nas imagens.

A visão computacional e a inteligência artificial estão muito associadas, pois ambas apresentam como principal objetivo o reconhecimento de padrões. A maioria das pesquisas que envolvem imagens digitais aplicam a abordagem de aprendizado de máquina para predizer padrões (SINGH et al., 2016). Várias são as razões para essa utilização, mas a elevada eficiência no reconhecimento de padrões e o elevado volume de informação presentes nas imagens são os dois principais motivos. Áreas como medicina (BAR et al., 2015) e de reconhecimento facial (PENG et al., 2015) tem utilizado com sucesso técnicas de visão computacional.

O melhoramento vegetal é uma área cujo os problemas são similares a áreas em que a visão computacional tem sido utilizada. Até pouco tempo poucas eram as pesquisas que visavam utilizar imagens para auxiliar na seleção e identificação de genótipos superiores. Outras áreas da genética como a genômica desenvolveram muito rápido, entretanto, as avaliações fenotípicas ficaram estagnadas por muito tempo (CROSSA et al., 2017). Porém, hoje já se sabe que somente as informações moleculares não são suficientes e, portanto, avaliações de alta qualidade são cruciais para acompanhar os avanços da genômica e consequentemente aumentar a acurácia dos processos seletivos. Com essa finalidade, houve um esforço muito grande de diversos laboratórios de

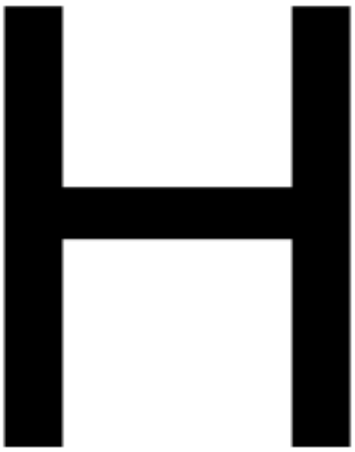
melhoramento vegetal no intuito de adquirir equipamentos para realizar avaliações dessa natureza assim como formar recursos humanos capazes de manipular e analisar imagens para auxiliar no melhoramento. As avaliações fenotípicas de alta qualidade associadas às técnicas de processamento digital de imagens, de visão computacional e de inteligência artificial propiciaram um campo fértil para o surgimento da fenômica (SINGH et al., 2016).

A fenômica é uma campo multidisciplinar do melhoramento vegetal que reúne tanto a parte vegetal como também computacional e de equipamentos (MAHLEIN, 2016). Devido a essa ampla gama de conhecimentos envolvidos, essa área muitas das vezes é realizada em parceria entre melhoristas e cientistas de dados. No intuito de compreender as pesquisas relacionadas a fenômica é necessário conhecer um pouco sobre o processamento digital de imagens, que pode ser subdividido em áreas específicas como: aquisição, pré-processamento, segmentação e extração de características (KHIRADE; PATIL, 2015).

A aquisição de imagens é realizada por meio de diferentes equipamentos que geram imagens distintas. Portanto, o primeiro passo consiste em compreender os diferentes tipos de imagens e identificar qual delas é a mais adequada para ser utilizada. Muitos dos estudos aplicados no melhoramento vegetal baseiam-se em imagens convencionais obtidas por meio de câmeras digitais, scanners ou celulares. Entretanto, para estudos mais complexos têm se utilizado câmeras térmicas ou multiespectrais capazes de captar outras informações que os dispositivos convencionais não conseguem (TATTARIS; REYNOLDS; CHAPMAN, 2016). Alguns estudos têm utilizado essas câmeras a nível de laboratório com uso de estúdios fotográficos, aliadas a drones ou plataformas de fenotipagem de forma a permitir obter informações a nível de experimentos de campo (WALTER; LIEBISCH; HUND, 2015).

2.3.1. Imagens digitais

Uma imagem digital é uma representação discreta de dados que contêm informação visual nas formas espaciais e de intensidades de cor. Essas são derivadas de uma distribuição contínua de luz, F(x,y), por meio do processo de discretização (BURGER & BURGE, 2008). Esse é um processo de amostragem de luz realizado por sensores que convertem um sinal contínuo F(x,y) em uma representação discreta de valores inteiros, I(m,n) (Figura 7).



F(x,y)

1	1	1	1	1	1	1	1	1	1
1	1	**0**	1	1	1	1	**0**	**1**	1
1	1	**0**	1	1	1	1	**0**	1	1
1	1	**0**	1	1	1	1	**0**	1	1
1	1	**0**	1	1	1	1	**0**	1	1
1	1	**0**	**0**	**0**	**0**	**0**	**0**	1	1
1	1	**0**	1	1	1	1	**0**	1	1
1	1	**0**	1	1	1	1	**0**	1	1
1	1	**0**	1	1	1	1	**0**	1	1
1	1	**0**	1	1	1	1	**0**	1	1
1	1	1	1	1	1	1	1	1	1

I(m,n)

Figura 7: Representações de uma imagem na forma visual e discreta.

A representação discreta de uma imagem digital é realizada pela função I(m,n), em que m pode variar de um a M linhas e n pode variar de um a N colunas (Figura 8). Portanto, imagem digital I sob o aspecto computacional é uma simples matriz bidimensional de valores inteiros que pode ser convertida para informação visual, ou seja,

cada valor em uma determinada posição da matriz de uma imagem pode ser convertido em uma intensidade de cor.

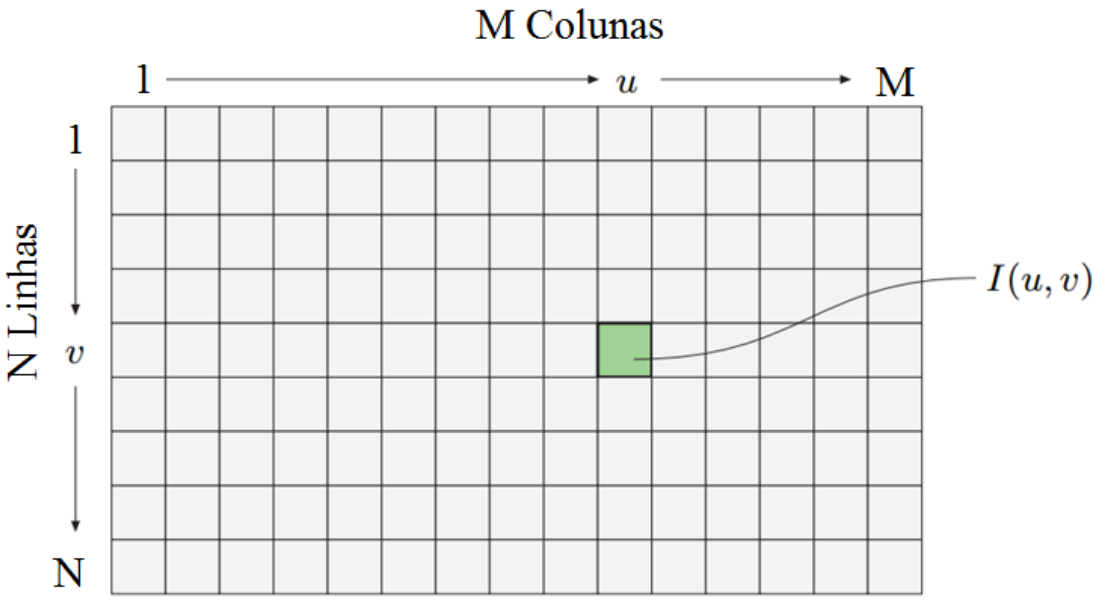


Figura 8: Plano discreto de uma imagem (BURGER & BURGE, 2008).

O tamanho de uma imagem é definido pelo número de linhas e colunas. Assim, as dimensões M x N de uma imagem define o tamanho e quantos pixels esta possui. Quanto maior é o tamanho de uma imagem maior é a quantidade de pixels e consequentemente a demanda computacional para armazenar e processar essas informações. Diferentemente do tamanho, a resolução de uma imagem é determinada pelo número de elementos da imagem por unidade de medida como por exemplo pontos por polegada (dots per inch - dpi) (HUNT, 2010). Devido os pixels possuírem formato de quadrado, a resolução das imagens são iguais tanto nas linhas quanto nas colunas. Em termos de processamento, a maioria dos algoritmos não dependem da resolução enquanto são mais dependentes das dimensões das imagens.

Existem diversos tipos de imagens e estas são definidas pelo número de canais que as compõe. Um canal é definido por uma matriz da imagem em que cada pixel é representado por um valor de intensidade. Em imagens com um único canal, cada pixel recebe um único valor que representa uma determinada intensidade na imagem. Já nas

imagens com mais canais que muitas das vezes são de natureza colorida, cada pixel recebe o número de valores igual ao número de canais. A combinação de valores em cada pixel é utilizada para representar uma determinada cor em cada pixel (SOLOMON & BRECKON 2013).

Os valores contidos em todos os pixels de uma imagem dependem do tipo de dado que será representado, ou seja, do tipo da imagem. Os valores que um pixel são convertidos em expressões binárias de comprimento k e, portanto, podem ser representados por 2^k valores (BURGER & BURGE, 2008). O valor de k representa o número de bits utilizados para representar a intensidade de cada pixel em cada canal.

Imagens binárias são um tipo especial de imagem com um único canal, capaz de representar presença ou ausência de um objeto. Para isso, nessa imagem utiliza-se somente um bit, ou seja, cada pixel pode assumir somente dois valores (0 ou 1). A Figura 7 é uma imagem binária em que k é igual a um, uma vez que cada pixel pode assumir somente os valores 0 (ausência) ou 1 (presença) que são convertidos em informações de intensidade iguais a preto ou branco, respectivamente.

Outro tipo de imagem com um único canal são as utilizadas para representar diferentes intensidades, brilho ou densidade de uma imagem. Imagens dessa natureza são consideradas de 8 bits pois cada pixel pode assumir valores de oito bits ($2^8 = 256$), que estão entre 0 e 255. Por exemplo, o valor 00000000 representa o valor 0 que é o de menor intensidade enquanto 11111111 define o valor 255 que é o mais intenso. Os outros valores nesse intervalo são representados pelas demais combinações de bits.

A conversão dos valores de cada pixel de uma imagem de intensidade em informação visual pode ser realizada por meio de mapas de cores, que alocam um tom específico de cor a cada nível numérico na imagem para a produção de uma representação visual dos dados (SOLOMON & BRECKON 2013). O mapa de cor mais comum é a

escala de cinza, que atribui ao valor menos intenso (0) a cor preta, enquanto que ao valor mais intenso (255) é atribuída a cor branca. Aos valores nesse intervalo de valores são atribuídos diferentes tons de cinza, conforme Figura 9A. Devido a dificuldade do olho humano de distinguir visualmente todos os diferentes tons de cinza, mapas de cores como Jet (Figura 9B) têm sido utilizados, especialmente em imagens térmicas. Nesse mapa de cor, o azul e o vermelho são atribuídos aos valores 0 e 255, respectivamente.

A

B

Figura 9. Mapas de cores: A – Escala de cinza e B - Jet

Apesar de muito informativas, as imagens de intensidade não contém informação a respeito de cores, uma vez que para representa-las é necessária a utilização de espaços de cor que dependem que a imagem apresente mais canais. Esses espaços são representações das cores por meio de intensidades de canais básicos, que podem conter informação de cor ou outras particularidades. Um dos sistemas mais utilizados é o RGB (Red, Green e Blue), que combina intensidades de vermelho, verde e azul para representar todas as possíveis cores da natureza em uma imagem. Nesse sistema, cada canal representado por uma cor básica comporta-se como uma imagem de intensidade, cujo os valores próximos a 255 apresentam maior intensidade no respectivo canal. Imagens dessa natureza são consideradas de 24 bits pois cada um dos três canais são imagens de intensidade de oito bits cada. Assim, cada pixel em uma imagem em RGB passa a receber

três valores, um para cada canal do sistema RGB, e a cor desse pixel será obtida pela combinação desses valores. Por exemplo, em uma imagem em RGB, o preto é representado pelos valores 0 nos três canais, enquanto que o branco é devido aos valores 255 nos três canais (Figura 10).

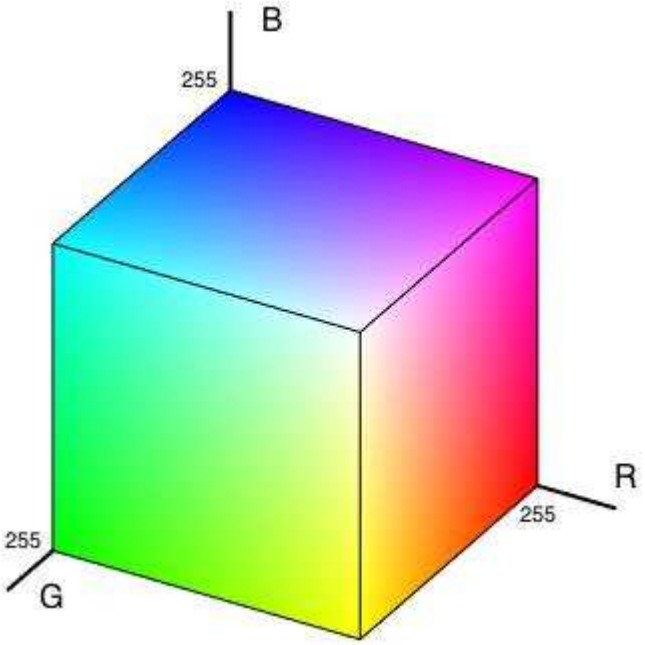


Figura 10. Espaço de cor RGB

Outros espaços de cores são utilizados para representar imagens coloridas. Dentre estes, os espaços HSV (hue, saturation, value) e os propostos pela Comissão Internacional de Iluminação (Comission Internationale de l'Eclairage - CIE) têm sido muito utilizados para auxiliar no processamento de imagens digitais. No HSV, a matiz ou tonalidade (H) determina o comprimento de onda dominante da cor, enquanto a saturação (S) e o valor (V) se referem a quantidade de luz branca e a luminosidade presente na imagem, respectivamente (Figura 11) (HUNT 2010). Esse espaço de cor perceptual é um modo alternativo de representar imagens em cores reais de forma mais natural a percepção humana que o RGB. Além disso, esse espaço de cor permite um elevado grau de

separação entre cor e iluminação, o que o qualifica para ser utilizado em diversas técnicas de processamento de imagens.

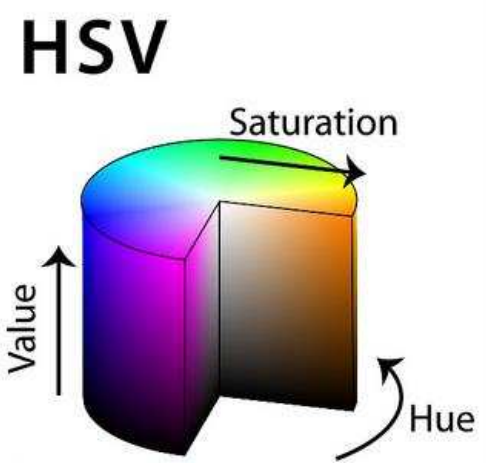


Figura 11. Espaço de cor HSV.

O espaço de cor Lab especificado pela CIE em 1976 foi desenvolvido com o objetivo de desenvolver um sistema de cores intuitivo pela linearização da representação de percepção de cor humana. Esse espaço é composto por um canal de luminosidade (L) e dois canais de cor (a e b), que especificam o matiz de cor e a saturação ao longo dos eixos verde-vermelho e azul-amarelo, respectivamente (Figura 12). Devido a essas propriedades, esse sistema tornou-se padrão em softwares de edição de imagens e auxiliado em muito no processamento de imagens digitais.

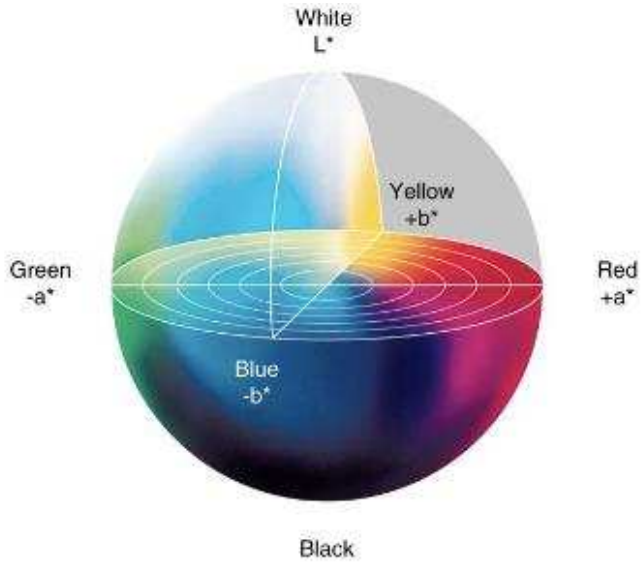


Figura 12. Espaço de cor Lab

A maioria dos dispositivos de aquisição de imagens utilizam o espaço de cor RGB, que muitas das vezes não facilita o processamento dessas imagens. Os espaços HSV e Lab apresentam varias particularidades que tornam o processamento dessas imagens mais simples. Assim, a conversão de imagens entre esses espaços de cor são essenciais para facilitar essa etapa. Outro aspecto essencial no processamento de imagens é a conversão das imagens coloridas em imagens em escala de cinza. A conversão RGB para escala de cinza é realizada em cada pixel nos três canais de cores, conforme a seguinte expressão:

$$I_{(\text{escala de cinza})}(n,m) = 0{,}2989*I(n,m,r) + 0{,}5870*I(n,m,g) + 0{,}1140*I(n,m,b)$$

em que n é a linha e m é a coluna de cada pixel; e r, g e b expressam os canais R, G e B, respectivamente (SOLOMON & BRECKON 2013). Apesar de útil para identificar objetos em imagens, essa operação não é invertível, uma vez após a conversão há a perda de informação relativa a cores.

Além da representação visual e matricial de imagens, outra forma de descrever uma imagem é a utilização de histogramas. Esses constituem de gráficos que representam a frequência dos valores de intensidade que ocorrem em cada canal de uma imagem (BURGER & BURGE, 2008), conforme Figura 13. O resultado de um histograma é um

vetor unidimensional de comprimento igual ao número de valores possíveis que o pixel da imagem em questão pode assumir. A soma de todos esses valores desse vetor corresponde ao total de pixels que uma imagem possui. Os histogramas são muito uteis para verificar se existem erros na aquisição das imagens assim como servir de base para realizar técnicas de segmentação de regiões de imagens.

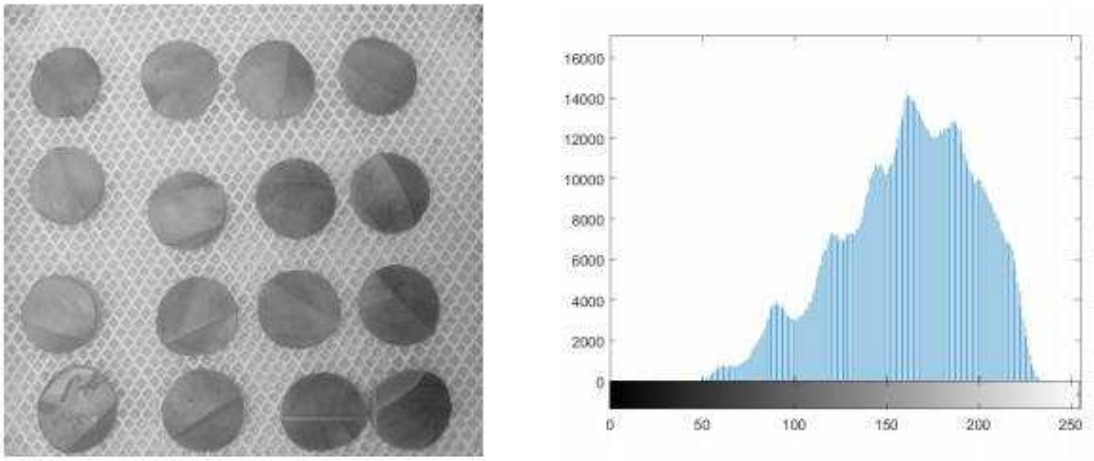


Figura 13. Representação visual e por histograma de uma imagem em escala de cinza.

Os histogramas podem ser obtidos tanto para imagens de intensidade como em imagens coloridas. Para imagens coloridas são obtidos histogramas para cada um dos canais da imagem (Figura 14). Os histogramas dos canais de uma imagem colorida são ferramentas muito úteis para verificar qual canal propicia maior discriminação dos objetos presentes em uma imagem. Canais como estes são muito úteis para identificar regiões de interesse em imagem. Por isso, o estudo das imagens por meio de histogramas constitui uma etapa preliminar que pode ser denominada como pré-processamento (KHIRADE; PATIL, 2015). Nessa etapa é possível visualizar possíveis erros de aquisição das imagens, presença de ruídos e também estudar a melhor forma de processar as imagens para se obter os melhores resultados.

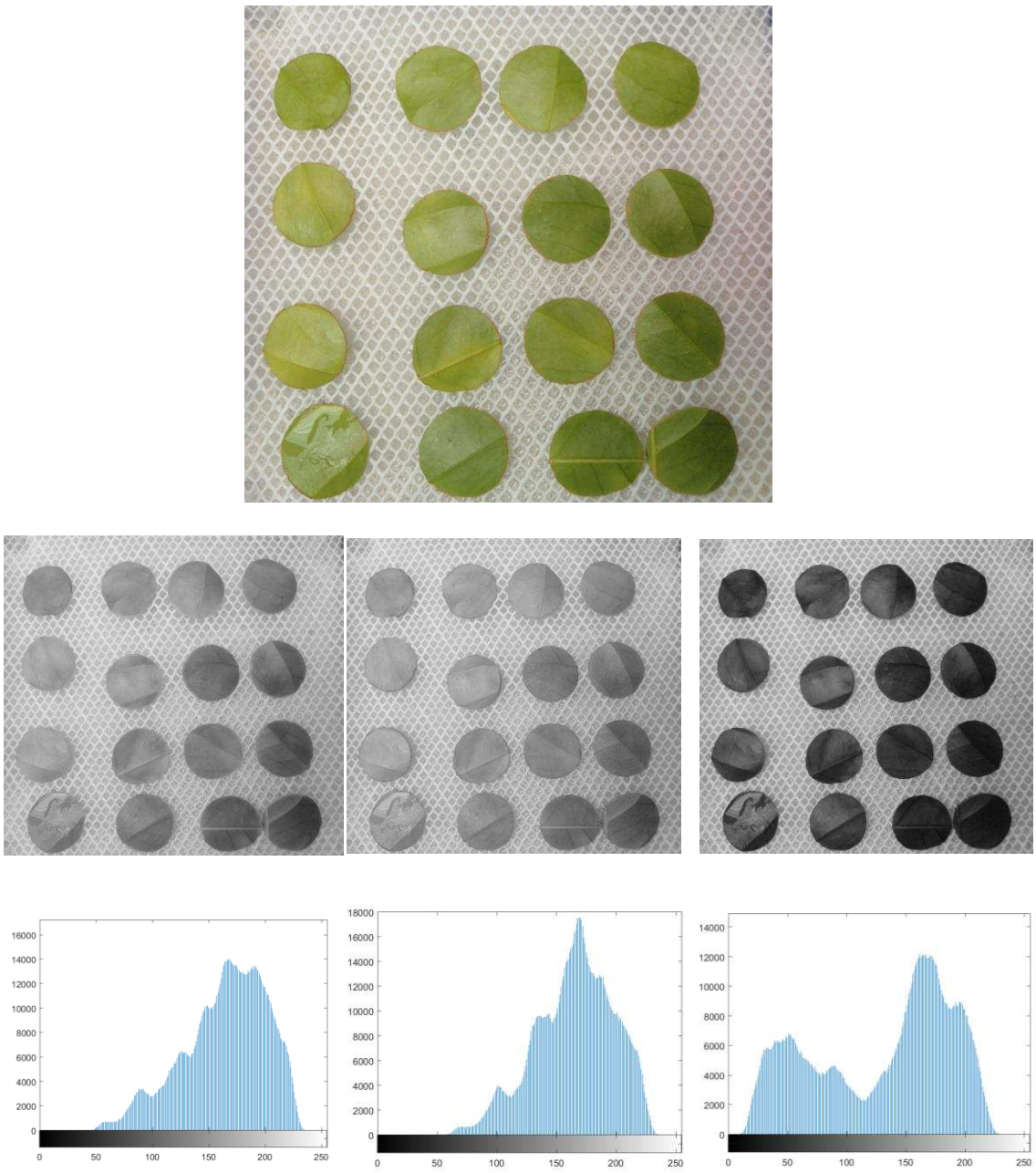


Figura 14. Representação visual e por histograma de uma imagem em RGB.

2.3.2. Processamento digital de imagens

O processamento de imagens constitui área que visa aplicar técnicas em imagens com o intuito de melhorar a sua visualização e destacar algumas particularidades (KHIRADE; PATIL, 2015). As técnicas com esse objetivo podem ser divididas em duas

estratégias, que se caracterizam pela forma com que atuam nas imagens. Essas estratégias são divididas em operações pontuais e regionais (SOLOMON & BRECKON 2013). As operações pontuais visam alterar os valores dos pixels sem alteração do tamanho, geometria e estrutura local das imagens. Essas técnicas aplicam transformações pontuais, que mapeia os valores dos pixels na imagem de entrada a pontos correspondentes na imagem de saída. Assim, os novos valores dependem exclusivamente dos valores anteriores. Além disso, essas transformações não são influenciadas por qualquer outro pixel, ou seja, a vizinhança de cada pixel não influencia no novo valor. As principais operações pontuais em uma imagem visam modificações no brilho ou contraste, transformações de intensidade e de cor e estudo de limiares em imagens.

Dentre as transformações pontuais mais básicas utilizadas para alterar o contraste de uma imagem, destacam-se operações de adição, subtração, multiplicação e divisão dos valores de cada pixel por determinado valor pré-estabelecido. Transformações logarítmica e exponencial assim como correção gama são alternativas mais eficientes para alterar o contraste de imagens. Além dessas, técnicas baseadas em manipulação de histogramas também são utilizados para modificar contraste e brilho em imagens. O alongamento de contraste ou normalização, denominada em inglês como contrast stretching, e a equalização de histogramas são técnicas com essa finalidade (SOLOMON & BRECKON 2013).

As operações pontuais apesar de eficientes são muito limitadas quando o interesse é visualizar certas informações nas imagens com maior clareza ou realizar tarefas de suavização de imagens. Entretanto, técnicas baseadas em operações regionais demonstram maior eficiência quando o interesse são tarefas dessa natureza. Diferentemente das operações pontuais, as técnicas regionais geram imagens resposta cujos os novos valores dos pixels são oriundos de transformações baseadas em um

conjunto de pixels da imagem original, que são localizados na vizinhança de cada pixel. Esse tipo de operação é o princípio do processo de convolução utilizado na filtragem de imagens digitais.

A convolução é um processo de mapeamento de imagens que empregam kernels para alterar os valores de pixels alvo da imagem original. Os kernels são matrizes retangulares de coeficientes, que são multiplicados pelos pixels da vizinhança do pixel alvo (HUNT, 2010), conforme Figura 15.

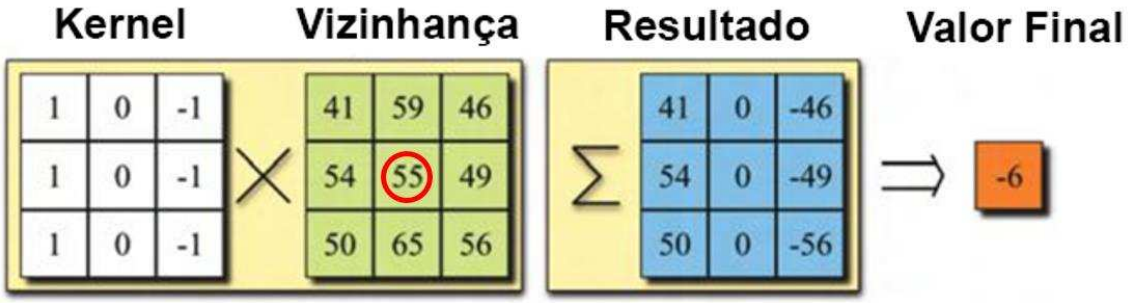


Figura 15. Processo de convolução em imagens (HUNT, 2010).

Essa transformação é realizada para cada pixel por meio do princípio de janela deslizante, ou seja, cada pixel na imagem é processado com base em uma operação realizada em sua vizinhança local de dimensão igual à do kernel adotado. Portanto, a convolução depende de três fatores: a vizinhança do pixel, os coeficientes do kernel e do deslocamento do kernel (janela deslizante).

A filtragem de imagens é uma das principais operações quando o intuito é melhorar a visualização do conteúdo de uma imagem. A remoção de ruído e detecção de bordas são os principais objetivos dessa forma de processamento de imagens. Diversos são os filtros utilizados com essa finalidade. O que os diferencia são os coeficientes dos kernels aplicados às imagens. Os filtros de média, mediana, moda e gaussiano são os principais utilizados para a remoção de ruído, enquanto que para detecção de bordas são utilizados os kernels de Roberts, de Prewitt e de Sobel (SOLOMON & BRECKON 2013).

Todos estes filtros auxiliam principalmente em etapas posteriores de processamento como a segmentação de imagens e também na extração de características. A convolução nessa última etapa é tão importante que é utilizada como base em técnicas de otimização como as redes neurais convolucionais que são uma das principais ferramentas de deep learning (LECUN; BENGIO; HINTON, 2015).

2.3.3. Segmentação e extração de características de imagens

A segmentação é o processo pelo qual regiões que compõem uma imagem são separadas por meio de métodos que se baseiam em particularidades como cor, textura ou movimento dos objetos (SABROL, 2015). Essa etapa é vital em projetos de visão computacional, uma vez que seria impossível extrair informações de regiões nas imagens e classificá-las se não fosse possível identificá-las corretamente. A presença de ruídos nas imagens a ser avaliadas pode comprometer essa segmentação. Por isso, etapas como filtragem e realce de imagens são imprescindíveis de ser realizadas antes da segmentação, o que evidencia que, apesar de muito útil, o processamento de imagens e estudos de visão computacional são muito complexos.

As técnicas de segmentação mais utilizadas visam agrupar os pixels mais similares em regiões específicas da imagem. Dentre as várias ferramentas de segmentação, as duas principais são a aplicação de limiar de intensidade ou de técnicas de aprendizado de máquina não supervisionado com o intuito de agrupar pixels similares (SOLOMON & BRECKON 2013). A primeira abordagem consiste em gerar uma imagem binária em que ao conjunto de pixels cuja a intensidade é superior ao limiar estabelecido previamente é atribuído o valor 1 (branco) e os pixels com intensidade inferior a esse limiar é atribuído valor zero (preto), conforme a expressão a seguir:

$$b(x,y) = \begin{cases} 1 \; se \; I(x,y) > T \\ 0 \; para \; todos \; os \; outros \; valores \; de \; I(x,y) \end{cases}'$$

em que b(x,y) é o novo valor gerado para o pixel original (I(x,y)) na linha x e coluna y após a aplicação do limiar (T).

A escolha do limiar pode ser baseada no estudo do histograma da imagem em questão. Imagens que apresentam histogramas com dois picos bem definidos facilitam a escolha visual do limiar. Entretanto, a escolha manual desse limiar não é a melhor estratégia. O método proposto por Otsu (1979) visa definir de forma automática esse limiar. Esse método consiste em identificar o valor de limiar que minimiza a variância dentro da classe dos pixels brancos e pretos após a aplicação dessa técnica. Imagens com problemas em sua aquisição e presença de ruídos dificultam a identificação desse limiar, uma vez que no histograma os picos podem não estar bem definidos. Portanto, a aplicação de limiar pode apresentar falhas na segmentação de imagens com esses problemas. Motivo pelo qual, outras técnicas mais eficientes baseadas em aprendizado de máquina são utilizadas com essa finalidade.

As técnicas de aprendizado de máquina não supervisionado cuja a finalidade é agrupamento de observações são muito utilizadas na segmentação de imagens coloridas. Exemplos dessas técnicas são: k-Means, fuzzy C-Means e os mapas auto organizáveis (MIRANDA 2011). A vantagem de utilizar essa abordagem é a possibilidade de identificar várias regiões da imagem por meio de uma única análise, já que é possível definir quantos grupos existem na imagem. Essas visam agrupar os pixels que apresentam similaridade quanto aos valores observados nos canais que compõe a imagem. Por isso é importante definir quais canais de cores são mais uteis para discriminar os pixels. Algumas das aplicações dessa técnica baseiam nos canais a e b da escala de cor Lab, uma vez que esses canais contém toda a informação de cor da imagem. Entretanto, outras escalas de cores como RGB e HSV podem ser utilizadas (SABROL, 2015).

O passo posterior à segmentação da imagem é a extração das características das regiões de interesse. Identificada uma região que pode ser folhas, plantas, sementes ou outras estruturas é possível quantificar medidas físicas como as dimensões, a área, o perímetro ou outras medidas associadas a morfometria geométrica dessas regiões. Além dessas, pode-se extrair informações relativas a cor dessas estruturas, assim como permitir estudos dos histogramas somente da região de interesse nos canais que compõem a imagem. Essa etapa é crucial, uma vez que essas informações podem ser utilizadas para auxiliar em estudos de classificação e predição na área de melhoramento de plantas, ou seja, exercem grande importância em estudos no melhoramento de plantas com enfoque na fenômica (SINGH et al., 2016).

Técnicas de aprendizado de máquina como redes neurais artificiais são muito utilizadas em estudos com imagens pois propiciam resultados muito eficazes (SABROL, 2015; SINGH et al., 2016). Porém para ser utilizadas as informações das imagens, estas precisam ser todas processadas para utilizar as informações em problemas de classificação e predição. Por isso, outra alternativa a todo esse processamento de imagens é a utilização das próprias imagens como entradas em redes neurais convolucionais. Essa abordagem de deep learning permite utilizar imagens como entradas no modelo e, por isso, tem enorme potencial para auxiliar em estudos de classificação aplicados ao melhoramento vegetal (CRUZ et al., 2017; LECUN; BENGIO; HINTON, 2015).

3. REFERÊNCIAS

BAR, Y. et al. Deep learning with non-medical training used for chest pathology identification. Medical Imaging 2015: Computer-Aided Diagnosis, v. 94140, n. March 2015, p. 94140V, 2015.

BARBOSA, C. D. et al. Artificial neural network analysis of genetic diversity in Carica papaya L . Crop Breeding and Applied Biotechnology, v. 11, n. 3, p. 224–231, 2011.

BATCHELOR, W. D.; YANG, X. B.; TSCHANZ, A. T. Development of a neural network for soybean rust epidemics. Transactions of the ASAE, v. 40, n. 1, p. 247–252, 1997.

BEZDEK, J. C. FCM : The Fuzzy C-Means Clustering Algorithm. Computers & Geosciences, v. 10, n. 2, p. 191–203, 1984.

BEZDEK, J. C.; EHRLICH, R.; FULL, W. FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences, v. 10, n. 2-3, p. 191–203, 1984.

BHERING, L. L. Rbio: A tool for biometric and statistical analysis using the R platform Rbio: A tool for biometric and statistical analysis using the R platform SOFTWARE/DEVICE RELEASE. Crop Breeding and Applied Biotechnology - Crop Breeding and Applied Biotechnology, v. 17, n. 17, p. 187–190, 2017.

BORÉM A.; MIRANDA G.V.; FRITSCHE-NETO R.. Melhoramento de Plantas 7ª ed. Viçosa: Editora UFV. 2017. 543p.

BRAGA, A. DE P.; CARVALHO, A. C. P. DE L. F. DE; LUDEMIR, T. B. Redes Neurais Artificiais - Teoria e Aplicações. 2. ed. Rio de Janeiro: LTC, 2011.

BURGER W.; BURGE M.J.. Digital Imaging Processing: An algorithmic introduction using Java. Verlag New York: Springer. 2008. 564p.

CARNEIRO, V. Q. et al. Artificial neural networks as auxiliary tools for the improvement of bean plant architecture. Genetics and Molecular Research, v. 16, n. 2, 2017.

CARNEIRO, V.Q., PRADO, A. L., CRUZ, C. D., CARNEIRO, P.C.S., Nascimento, M., CARNEIRO, J.E. de S. Fuzzy control systems for decision-making in cultivars recommendation. Acta Sci. Agron. V. 40, 1–8, 2018.

CASTRO, C. A. DE O. et al. High-performance prediction of macauba fruit biomass for agricultural and industrial purposes using Artificial Neural Networks. Industrial Crops & Products, v. 108, n. August, p. 806–813, 2017.

CAVALCANTE, R. C. et al. Computational Intelligence and Financial Markets: A Survey and Future Directions. Expert Systems with Applications, v. 55, p. 194–211, 2016.

CROSSA, J. et al. Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. Trends in Plant Science, v. 22, n. 11, p. 961–975, 2017.

CRUZ, A. C. et al. X-FIDO: An Effective Application for Detecting Olive Quick Decline Syndrome with Deep Learning and Data Fusion. Frontiers in Plant Science, v. 8, n. October, p. 1–12, 2017.

CRUZ, C. D. GENES - a software package for analysis in experimental statistics and quantitative genetics. Acta Scientiarum. Agronomy, v. 35, n. 3, p. 271–276, 2013.

CRUZ, C. D. Genes Software – extended and integrated with the R , Matlab and Selegen. Acta Scientiarum Agronomy, v. 38, n. 4, p. 547–552, 2016.

CRUZ, C. D.; CARNEIRO, P. C. S.; REGAZZI, A. J. Modelos biométricos aplicados ao melhoramento genético - Vol 2. 3. ed. Viçosa: Editora UFV, 2014.

CRUZ, C. D.; REGAZZI, A. J.; CARNEIRO, P. C. S. Modelos biométricos aplicados ao melhoramento genético. 4. ed. Viçosa: Editora UFV, 2012.

DHANACHANDRA, N.; MANGLEM, K.; CHANU, Y. J. Image Segmentation Using Kmeans Clustering Algorithm and Subtractive Clustering Algorithm. Procedia Computer Science, v. 54, p. 764–771, 2015.

DOGRUPARMAK, S.C., KESKIN, G.A., YAMAN, S., ALKAN, A., 2014. Using principal component analysis and fuzzy c–means clustering for the assessment of air quality monitoring. Atmos. Pollut. Res. v. 5, p. 656–663, 2014.

EBERHART, S. A.; RUSSELL, W. A. Stability Parameters for Comparing Varieties. Crop Science, v. 6, n. 1, p. 36, 1966.

EHRET, A. et al. Application of neural networks with back-propagation to genome-enabled prediction of complex traits in Holstein-Friesian and German Fleckvieh cattle. Genetics, selection, evolution : GSE, v. 47, p. 22, 2015.

FINLAY, K. W.; WILKINSON, G.. The Analysis of Adaptation in a Plant Breeding Programme. Australian Journal of Agricultural Research, v. 14, p. 742–754, 1963.

FRITSCHE-NETO R.; BORÉM A. Fenômica: como a fenotipagem de próxima geração está revolucionando o melhoramento de plantas. 1. ed. Visconde do Rio Banco, MG: Editora Suprema, 2015. 216p.

GARDNER, C.; EBERHART, S. Analysis and Interpretation of the Variety Cross Diallel and Related Populations. Biometrics, v. 22, n. 3, p. 439–452, 1966.

GIANOLA, D. et al. Predicting complex quantitative traits with Bayesian neural networks: A case study with Jersey cows and wheat. BMC Genetics, v. 12, n. 1, p. 87, 2011.

GONZÁLEZ-CAMACHO, J. M. et al. Genome-enabled prediction using probabilistic neural network classifiers. BMC Genomics, v. 17, n. 1, p. 1–16, 2016.

GHOSH, S., DUBEY, S.K.S. Comparative analysis of k-means and fuzzy c-means algorithms. Ijacsa v. 4, p. 35–38, 2013.

GRIFFING, B. Concept of General and Specific Combining Ability in Relation to Diallel Crossing Systems. Australian Journal of Biological Sciences, v. 9, n. 4, p. 463, 1956.

GROSSBERG, S. How does a brain build a cognitive code? Psychological review, v. 87, n. 1, p. 1–51, 1980.

HAYKIN, S. Neural Networks and Learning Machines. 3. ed. Hamilton: Pearson - Prentice Hall, 2008.

HAYMAN, B. I. The Theory and Analysis of Diallel Crosses. Genetics, v. 39, n. 6, p. 789– 809, 1954.

HAZEL, L. N. The Genetic Basis for Constructing Selection Indexes. Genetics, v. 28, n. 6, p. 476–490, 1943.

HOPFIELD, J. J. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences of the United States of America, v. 79, n. 8, p. 2554–2558, 1982.

HUNT K. A.. The Art of Image Processing with Java. A K Peters/CRC Press. 1 ed. 2010. 300p.

JAMES G.; WITTEN D.; HASTIE T.; TIBSHIRANI R. An Introduction to Statistical Learning: with Applications in R. 1 ed. Springer-Verlag New York. 2013. 426p.

JANG, J. S. R.; SUN, C. T.; MIZUTANI, E. Neuro-Fuzzy and Soft Computing - A computacional Approach to Learning and Machine Intelligence. New Delhi: PHI Learning Private Limited, 2012.

KAUL, M.; HILL, R. L.; WALTHALL, C. Artificial neural networks for corn and soybean yield prediction. Agricultural Systems, v. 85, n. 1, p. 1–18, 2005.

KHIRADE, S. D.; PATIL, A. B. Plant Disease Detection Using Image Processing. 2015 International Conference on Computing Communication Control and Automation, p. 768–771, 2015.

KOHONEN, T. Self-organized formation of topologically correct feature maps. Biological Cybernetics, 1982.

KOHONEN, T. The self-organizing map. Neurocomputing, v. 21, n. 1-3, p. 1–6, 1998.

KORENEVSKIY, N. A., 2015. Application of fuzzy logic for decision-making in medical expert systems. Biomed. Eng. (NY). v. 49, p. 33–35, 2015.

KOUROU, K. et al. Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, v. 13, p. 8–17, 2015.

KURTENER, D.; DRAGAVTSEV, V. Application of Soft Computing in Plant Breeding and Crop Production. European Agrphysical Journal, v. 4, n. 1, p. 10–24, 2017.

LECUN, Y. et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, v. 86, n. 11, p. 2278–2323, 1998.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. Nature, v. 521, n. 7553, p. 436–444, 2015.

LI, H.; HE, H.; WEN, Y. Optik Dynamic particle swarm optimization and K-means clustering algorithm for image segmentation. Optik - International Journal for Light and Electron Optics, v. 126, n. 24, p. 4817–4822, 2015.

LIAO, G. et al. Using SOM neural network for X-ray inspection of missing-bump defects in three-dimensional integration. Microelectronics Reliability, v. 55, n. 12, p. 2826–2832, 2015.

LIN, C. S.; BINNS, M. R. A superiority measure of cultivar performance for cultivar × location data. Canadian Journal of Plant Science, v. 68, n. 1, p. 193–198, 1988.

MA, C.; ZHANG, H. H.; WANG, X. Machine learning for Big Data analytics in plants. Trends in Plant Science, v. 19, n. 12, p. 798–808, 2014.

MACQUEEN.; J. B.. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Symposium on Math, Statistics, and Probability. Berkeley, CA: University of California Press. 1967. 281–297.

MAHLEIN, A.-K. Plant Disease Detection by Imaging Sensors - Parallels and Specific Demands for Precision Agriculture and Plant Phenotyping. Plant Disease, v. 100, n. 2, p. 1– 11, 2016.

MARQUARDT, D., "An Algorithm for Least-Squares Estimation of Nonlinear Parameters," SIAM Journal on Applied Mathematics, v. 11, n. 2, p. 431–441,1963.

MATLAB. (2018). R2017b. Natick, MA: The Math Works Inc.

MCCULLOCH, W. S.; PITTS, W. H. A Logical Calculus of the Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics, v. 5, p. 115–133, 1943.

MIRANDA. J. I.. Processamento de imagens digitais: métodos multivariados em Java. 1 ed. Editora: Embrapa. 2011. 400p.

OTSU, N., 1979. A Threshold Selection Method from Gray-Level Histograms. IEEE Trans. Syst. Man. Cybern. v. 9, p. 62–66, 1979.

PENG, Y. et al. Discriminative graph regularized extreme learning machine and its application to face recognition. Neurocomputing, v. 149, n. Part A, p. 340–353, 2015.

PESEK, J.; BAKER, R. J. Desired Improvement in Relation to Selection Indices. Canadian Journal of Plant Science, v. 804, n. 383, p. 803–804, 1969.

RAMALHO, M.A.P.; SANTOS, J.B.; ZIMMERMANN, M.J.O. Genética quantitativa de plantas autógamas: aplicações ao melhoramento do feijoeiro. Goiânia: UFG, 1993. 271p.

RAMALHO, M. A. P.; ABREU, A. F. B.; SANTOS, J. B.; NUNES, J. A. R.. Aplicações da Genética Quantitativa no Melhoramento de Plantas Autógamas. 1. ed. Lavras: Editora UFLA, 2012.

R CORE TEAM (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

RAY, D. K. et al. Yield Trends Are Insufficient to Double Global Crop Production by 2050. PLoS ONE, v. 8, n. 6, 2013.

RESENDE, M. D. V. DE. Software Selegen-REML / BLUP : a useful tool for plant breeding. Crop Breeding and Applied Biotechnology, v. 16, n. September, p. 330–339, 2016.

ROSENBLATT, F. The Perceptron: a probabilistic modelfor information storage and organization in the brain. Psycological Review, v. 65, n. 6, p. 386–408, 1958.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by backpropagating errors. Nature, 1986.

SABROL, H. Recent Studies of Image and Soft Computing Techniques for Plant Disease Recognition and Classification. International. Journal of Computer Applications, v. 126, n. 1, p. 44–55, 2015.

SANT'ANNA, I. C. et al. Superiority of artificial neural networks for a genetic classification procedure. Genetics and Molecular Research, v. 14, n. 3, p. 9898–9906, 2015.

SCHMIDHUBER, J. Deep learning in neural networks : An overview. Neural Networks, v. 61, p. 85–117, 2015.

SCHUSTER, I.; CRUZ, C. D. Estatística genômica. Aplicada a populações derivadas de cruzamentos controlados. Viçosa: Editora Universidade Federal de Viçosa, 2008. 568p.

SILVA, I. N. DA; SPATTI, D. H.; FLAUZINO, R. A. Redes neurais Artificiais para Engenharias e ciências Aplicadas. 1. ed. São Paulo: Artliber Editora Ltda, 2010.

SILVA, G. N. et al. Neural networks for predicting breeding values and genetic gains. Scientia Agricola, v. 71, n. 6, p. 494–498, 2014.

SIMÕES, M. G.; SHAW, I. S. Controle e Modelagem Fuzzy. 2. ed. São Paulo: EDGARD BLUCHER, 2007.

SINGH, A. et al. Machine Learning for High-Throughput Stress Phenotyping in Plants. Trends in Plant Science, v. 21, n. 2, p. 110–124, 2016.

SMITH, H. F. A discrimant function for plant selection. Annals of Eugenics, v. 7, p. 240–250, 1936.

SOLOMON C.; BRECKON T. Fundamentos de processamento digital de imagens: uma abordagem pratica com exemplos em Matlab. 1 ed. Editora LTC. 2013. 281p.

TATTARIS, M.; REYNOLDS, M. P.; CHAPMAN, S. C. A Direct Comparison of Remote Sensing Approaches for High-Throughput Phenotyping in Plant Breeding. Frontiers in Plant Science, v. 7, n. August, p. 1–9, 2016.

TEODORO, P. E.; BARROSO, L. M. A.; NASCIMENTO M.; TORRES, F. E.; SAGRILO, E.; SANTOS, A.; RIBEIRO, L. P.. Redes neurais artificiais para identificar genótipos de feijão-caupi semiprostrado com alta adaptabilidade e estabilidade fenotípicas. Pesquisa Agropecuaria Brasileira, v. 50, n. 11, p. 1054–1060, 2015.

WALTER, A.; LIEBISCH, F.; HUND, A. Plant phenotyping: From bean weighing to image analysis. Plant Methods, v. 11, n. 1, p. 1–11, 2015. WERBOS, P. J. Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. Cambridge, Massachusetts: Harvard University, 1974.

WIDROW, B.; HOFF, M. E. Adaptive switching circuits. IRE Western Electric Show and Convention Record, p. 96–104, 1960.

ZADEH, L. A. Fuzzy sets. Information and Control, v. 8, n. 3, p. 338–353, 1965.

ZADEH, L. A. The concept of a linguistic variable and its application to approximate reasoning—I. Information Sciences, v. 8, n. 4, p. 199–249, 1975.

ZADEH, L. A. A computational approach to fuzzy quantifiers in natural languages. Computers & Mathematics with Applications, v. 9, n. 1, p. 149–184, 1983.

ZADEH, L. A. Fuzzy logic equals Computing with words. Ieee Transactions on Fuzzy Systems, v. 4, n. 2, p. 103–111, 1996.

ZADEH, L. A. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets and Systems, v. 90, n. 2, p. 111–127, 1997.

ZADEH, L. A., 1999. A theory of possibility distributions. Fuzzy Sets Syst. v. 102, p. 135–155, 1999.

ZADEH, L. A. A theory of possibility distributions. Fuzzy Sets and Systems, v. 102, n. 2, p. 135–155, 1999. ZADEH, L. A. A note on web intelligence, world knowledge and fuzzy logic. Data and Knowledge Engineering, v. 50, n. 3 SPEC. ISS., p. 291–304, 2004.

ZADEH, L. A. Is there a need for fuzzy logic? Information Sciences, v. 178, n. 13, p. 2751– 2779, 2008.

CAPÍTULO 1 – FENOM: ANÁLISES FENOTÍPICAS DE ALTA QUALIDADE

FENOM: ANÁLISES FENOTÍPICAS DE ALTA QUALIDADE

RESUMO

No melhoramento vegetal, a grande disparidade na qualidade e volume entre as informações fenotípicas e moleculares dos genótipos avaliados tem dificultado a identificação e seleção de genótipos superiores. A fenômica, área cujo principal objetivo é o desenvolvimento e a utilização de tecnologias de fenotipagem de alta qualidade, é crucial para tornar o melhoramento genético mais eficiente. Os avanços nas tecnologias computacionais e a redução dos custos dos dispositivos de aquisição de imagens tornaram possíveis grandes avanços da fenômica. Entretanto, poucos ainda são os softwares interativos disponíveis aos programas de melhoramento vegetal para analisar dados dessa natureza. Portanto, o objetivo com este trabalho é apresentar o software FENOM voltado a análises científicas de dados com ênfase em fenotipagem de alta qualidade. O software FENOM é gratuito, compatível com sistema operacional Windows e está disponível para download no repositório https://github.com/VQCarneiro/Fenom. Esse software é subdividido em duas áreas de procedimentos: processamento digital de imagens e solução de problemas classificatórios por meio de redes neurais artificiais. Para processamento de imagens estão disponíveis procedimentos de aquisição, pré-processamento, segmentação e extração de características. Já nos procedimentos de classificação está disponível a técnica de redes neurais artificiais com arquitetura perceptron multicamadas. Essas aplicações constituem em importante contribuição para o melhoramento vegetal, principalmente por visar a difusão dessa tecnologia.

Palavras chave: fenômica, inteligência artificial, melhoramento genético

1. INTRODUÇÃO

No intuito de tornar o desenvolvimento de novas cultivares mais eficiente houve, nos últimos 20 anos, elevado investimento em áreas da ciência relacionadas à genética que ainda não eram empregadas no melhoramento vegetal (Crossa et al., 2017). Exemplo disso é a utilização de seleção assistida por marcadores moleculares (SAM) e, mais recentemente, dos estudos de seleção (GWS) e associação genômica ampla (GWAS). Com a utilização dessas técnicas moleculares é possível gerar elevado número de informações que têm auxiliado na seleção de genótipos superiores (Shikha et al., 2017) e em estudos de diversidade genética (Silva et al., 2017).

O rápido avanço de técnicas genômicas de alta qualidade gerou, como resultado, elevado volume de informações moleculares de alta qualidade e tem se mostrado de grande utilidade ao melhoramento vegetal (Desta e Ortiz, 2014; Crossa et al., 2017). Entretanto, somente essas informações não são suficientes para realizar estudos relacionados à SAM, GWS ou GWAS em plantas. Com esse intuito é necessário o uso associado de informações moleculares e fenotípicas (Schuster e Cruz, 2008; Houle et al., 2010).

Observa-se que o atual desafio em estudos genômicos reside na grande disparidade de qualidade e volume entre as informações fenotípicas e moleculares dos genótipos avaliados. Enquanto os dados moleculares são obtidos de forma rápida por equipamentos com elevada acurácia e precisão, algumas das avaliações fenotípicas ainda são realizadas de formas manuais (Lima et al., 2012) ou visuais (Oliveira et al., 2015; Moura et al., 2013). Portanto, verifica-se que a utilização de tecnologias de fenotipagem de alta qualidade, denominada como fenômica, são cruciais para tornar mais eficiente o

processo de identificação, seleção e recomendação de genótipos superiores (Fritsche Neto e Borém, 2015).

A manipulação de imagens digitais, à semelhança da genômica, era executada apenas por poucos especialistas que tinham acesso a equipamentos de elevado custo (Burger e Burge, 2008). No entanto, a redução dos custos dos computadores com hardwares capazes de processar imagens e a facilidade de aquisição de imagens digitais por meio de dispositivos portáteis resultou em uma infinidade de informações úteis ao melhoramento vegetal (Mutka e Bart, 2015; Tattaris, Reynolds e Chapman, 2016). Portanto, o processamento digital de imagens se popularizou à semelhança do processamento de dados que é realizado convencionalmente no melhoramento genético.

O avanço na área de inteligência artificial, principalmente em aprendizado de máquinas e deep learning, tornou ainda mais propícia a adoção da fenômica (Singh, 2016). No melhoramento vegetal, técnicas como redes neurais artificiais apresentaram elevada eficácia na solução de problemas classificatórios (Nascimento et al., 2013; Couto et al., 2015; Sant'Anna et al., 2015; Carneiro et al., 2017) e preditivos (Castro et al., 2017; Glória et al., 2016; Silva et al., 2014). O Deep Learning (Lecun et al., 2015), no qual a principal ferramenta são as redes neurais convolucionais (Angermueller et al., 2016; Ubbens e Stavness, 2017), também tem auxiliado em tarefas de avaliações fenotípicas de alta qualidade. As informações oriundas de imagens aliadas a técnicas de inteligência artificial são ferramentas cruciais para tornar o melhoramento genético mais eficiente.

Outra dificuldade relacionada ao processamento de dados com elevado volume e principalmente oriundos de avaliações fenotípicas de alta qualidade é a baixa quantidade de softwares interativos que tornem essas análises que são complexas em formas simples de serem utilizadas. Softwares como Matlab (The MathWorks, 2012) e R (R Core Team, 2018) têm sido utilizados para manipular e processar imagens digitais (Najafabadi e

Farahani, 2012). Entretanto, para utilização desses softwares existe a necessidade de conhecimento na área de programação, que muitos melhoristas ainda não detém. Além disso, existe grande necessidade de difundir o conhecimento em processamento digital de imagens no melhoramento vegetal, pois essa área é de especialidade restrita da ciência da computação e algumas engenharias. Portanto, o objetivo desse trabalho é apresentar o software FENOM voltado às análises científicas de dados com ênfase em fenotipagem de alta qualidade.

2. DESCRIÇÃO

O FENOM é desenvolvido em Matlab (https://www.mathworks.com/) em integração à linguagem de programação JAVA (https://www.java.com/pt_BR/). O Matlab foi utilizado para o desenvolvimento dos procedimentos destinados ao processamento digital de imagens e de redes neurais artificiais. Para isso foram utilizadas as bibliotecas de processamento de imagens (image processing toolbox) e de redes neurais artificiais (neural networks toolbox) deste software. A linguagem JAVA foi utilizada para desenvolver a interface gráfica do software. Portanto, para executar o FENOM é necessário que o computador possua o software Matlab com as bibliotecas instaladas e a linguagem JAVA.

O software FENOM é compatível com sistema operacional Windows e está disponível gratuitamente para download no repositório https://github.com/VQCarneiro/Fenom. Não há necessidade de realizar instalação deste software no computador. Basta fazer o download da pasta FENOM, que deve ser descompactada no diretório C: (por exemplo: C:\FENOM).

3. PROCEDIMENTOS

O software FENOM (Figura 1) é desenvolvido com o intuito de disponibilizar procedimentos que abordam todas as etapas de avaliação por meio de imagens de experimentos em campo ou laboratório. Portanto, os procedimentos envolvem as seguintes etapas do processamento digital de imagens: aquisição de imagens, pré-processamento, segmentação e extração de características. Além disso, este software dispõe de diversos exemplos práticos das etapas acima relatadas, uma vez que este software também é de natureza didática. O FENOM também dispõe de análises classificatórias por meio de redes neurais artificiais com arquitetura perceptron multicamadas (Jang et al., 2012).

Figura 1. Interface do software FENOM.

O FENOM apresenta sistema de leitura de imagens nos formatos: portable network graphics (png), joint photographic experts group (jpg ou jpeg) e tagged image file format (tiff). As imagens processadas durante as análises também são salvas nesses

formatos. Além dessas imagens, também são gerados relatórios e visualizações gráficas, que facilitam a interpretação dos resultados.

3.1. AQUISIÇÃO DE IMAGENS

O primeiro passo das avaliações fenotípicas de alta qualidade em experimentos é a captura das imagens. Essa consiste em etapa crucial em todo o processamento das imagens, uma vez que qualquer ruído ou imperfeição na imagem irá prejudicar todas as etapas seguintes. Outra questão de extrema importância são os equipamentos a serem utilizados. A avaliação deve ser a mais acurada e prática possível e, portanto, é necessário que o sistema apresente um nível adequado de automatização. Portanto, o ideal é que o equipamento de aquisição das imagens esteja acoplado ao computador e se possível a captura pelo software seja em tempo real, principalmente para evitar possíveis erros na aquisição.

No intuito de tornar automatizado todo o processo de captura das imagens foi implementado um procedimento de aquisição das imagens em tempo real no software FENOM. Nesse procedimento, o software já reconhece o dispositivo acoplado ao computador e realiza a aquisição das imagens em tempo real. Assim, torna-se possível identificar erros e ruídos presentes nas imagens, obter informações sobre as imagens capturadas e nomeá-las conforme identificação do experimento. Nesse mesmo procedimento é possível também carregar imagens disponíveis no computador e obter essas informações.

3.2. PRÉ-PROCESSAMENTO

Após a aquisição das imagens, essas podem ser submetidas a alguns pré-processamentos que permitem compreender melhor as imagens. Por exemplo, as imagens coloridas podem ser convertidas em imagens binárias (preto e branco) ou de intensidade (escala de cinza). Essa etapa é de grande importância, uma vez que é crucial identificar qual tipo de imagem é melhor para realizar etapas de processamento posteriores. Além disso, as imagens coloridas podem ser decompostas em canais de espaços de cores (Figura 2). A maioria dos dispositivos de aquisição de imagens utilizam o espaço de cor RGB (R - red, G -green, B - blue), que muitas das vezes não facilita o processamento dessas imagens. Entretanto, os espaços HSV (H - hue, S - saturation, V - value) e Lab (L – luminosidade, a – verde - vermelho e b – azul - amarelo) apresentam várias particularidades que tornam o processamento das imagens mais simples (Solomon e Breckon 2013).

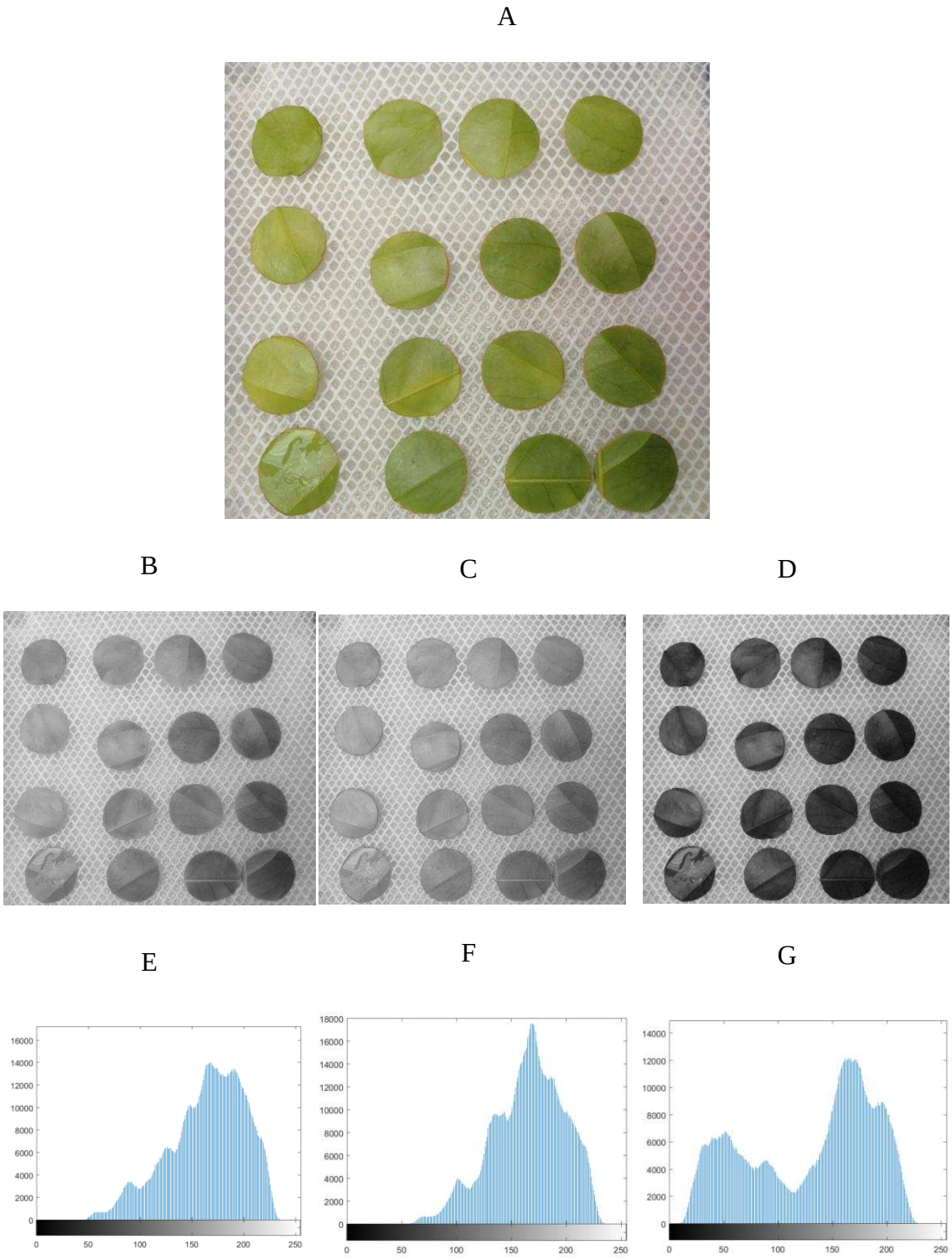


Figura 2. Representação visual e por histograma de uma imagem em RGB. A: imagem colorida; B: canal vermelho (R); C: canal verde (G); D: canal azul (B); E: histograma do canal vermelho, F: histograma do canal verde; G: histograma do canal azul.

Além da representação visual, outra forma de descrever uma imagem é a utilização de histogramas. Esses constituem de visualizações gráficas que representam a frequência dos valores de intensidade que ocorrem em cada canal de uma imagem (Figura 2 E, F, G) (Burger e Burge, 2008). Os histogramas podem ser obtidos tanto para imagens de intensidade como em imagens coloridas. Para imagens coloridas são obtidos histogramas para cada um dos canais da imagem. Os histogramas dos canais de uma imagem colorida são ferramentas muito úteis para verificar qual canal propicia maior discriminação dos objetos presentes em uma imagem (Burger e Burge, 2008). Canais como estes são muito úteis para identificar regiões de interesse em imagem. Além disso, nessa etapa é possível visualizar possíveis erros de aquisição das imagens, presença de ruídos e também estudar a melhor forma de processar as imagens para se obter os melhores resultados.

No intuito de facilitar o entendimento das imagens foi implementado no software FENOM procedimento de análise e conversão dos sistemas de cores. Nesse procedimento é possível visualizar imagens coloridas decompostas nos canais dos sistemas de cores RGB, HSV e Lab. Além das imagens em cada canal são obtidos os histogramas das imagens nos respectivos canais (Figura 2).

Outro aspecto importante da etapa de pré-processamento é o realce ou destaque de regiões específicas de uma imagem. O realce de imagens visa aplicar técnicas em imagens com o intuito de melhorar a sua visualização e destacar algumas particularidades (Solomon e Breckon 2013). O destaque de imagens apresenta o mesmo objetivo, porém consiste em realizar operações matemáticas para gerar uma imagem com um único canal que é oriunda da combinação linear dos pixels de todos, ou parte, dos canais que compõem uma imagem. No software FENOM foi implementado procedimento de destaque de imagens para cada uma das escalas de cor. Nesse mesmo procedimento

também é possível converter imagens coloridas em imagens de intensidade (escala de cinza).

3.3. SEGMENTAÇÃO DE IMAGENS

A segmentação é o processo pelo qual regiões que compõem uma imagem são separadas por meio de métodos que se baseiam em particularidades como cor, textura ou movimento dos objetos (Sabrol, 2015). Essa etapa é vital em projetos de visão computacional, uma vez que seria impossível extrair informações de regiões nas imagens e classificá-las se não fosse possível identificá-las corretamente. As técnicas de segmentação mais utilizadas visam agrupar os pixels mais similares em regiões específicas da imagem. Dentre as várias ferramentas de segmentação, as duas principais são a aplicação de limiar de intensidade ou de técnicas de aprendizado de máquina não supervisionado com o intuito de agrupar pixels similares (Solomon e Breckon 2013). A primeira abordagem consiste em gerar uma imagem binária que ao conjunto de pixels cuja a intensidade é superior ao limiar estabelecido previamente é atribuído o valor 1 (branco) e aos pixels com intensidade inferior a esse limiar é atribuído valor zero (preto).

A escolha do limiar pode ser baseada no estudo do histograma da imagem em questão. Imagens que apresentam histogramas que a distribuição dos pixels é bimodal facilitam a escolha visual do limiar. Entretanto, a escolha manual desse limiar não é a melhor estratégia. O método proposto por Otsu (1979) visa definir de forma automática esse limiar. Esse método consiste em identificar o valor de limiar que maximiza a variância entre classes dos pixels brancos e pretos após a aplicação dessa técnica (Burger e Burge, 2008). Imagens com problemas em sua aquisição e presença de ruídos dificultam a identificação desse limiar, uma vez que no histograma os picos podem não estar bem

definidos, como nas Figuras 2 (E e F). Portanto, a aplicação de limiar pode apresentar falhas na segmentação de imagens com esses problemas. Motivo pelo qual, outras técnicas mais eficientes baseadas em aprendizado de máquina são utilizadas com essa finalidade.

As técnicas de aprendizado de máquina não supervisionado, cuja finalidade é agrupamento de observações, são muito utilizadas na segmentação de imagens coloridas. Exemplos dessas técnicas são: k-Means (Macqueen, 1967), fuzzy C-Means (Bezdek, 1984) e os mapas auto organizáveis (Kohonen, 1982). A vantagem de utilizar essa abordagem é a possibilidade de identificar várias regiões da imagem por meio de uma única análise, já que é possível definir quantos grupos existem na imagem (Miranda, 2011). Essas visam agrupar os pixels que apresentam similaridade quanto aos valores observados nos canais que compõe a imagem. Por isso é importante definir quais canais de cores são mais úteis para discriminar os pixels. Algumas das aplicações dessa técnica baseiam nos canais a e b da escala de cor Lab, uma vez que esses canais contém toda a informação de cor da imagem. Entretanto, outras escalas de cores como RGB e HSV podem ser utilizadas (Sabrol, 2015).

No intuito de identificar certas regiões que compõe uma imagem foram implementados os procedimentos de segmentação baseados no método proposto Otsu (1979) e nas técnicas k-Means, fuzzy C-Means e mapas auto organizáveis. Exemplo da imagem 2A submetida a esses procedimentos é observada na Figura 3. Apesar da região da imagem caracterizada pelo fundo (região branca) oferecer problemas de distinção com os discos foliares verificou-se que é possível encontrar automaticamente os discos foliares na imagem 2A com rapidez e eficiência por meio de qualquer umas das técnicas de segmentação.

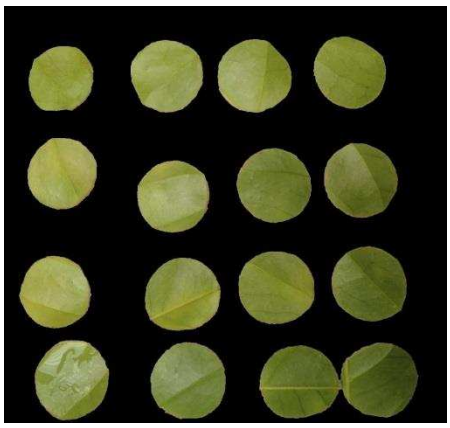


Figura 3. Imagem oriunda da segmentação da Figura 2A pelos métodos de Otsu, k-Means, fuzzy C-Means e mapas auto organizáveis.

Para o procedimento que utiliza o método proposto por Otsu (1979), além da imagem original, é necessário reconhecer qual escala de cor e o respectivo canal que permite a melhor discriminação das regiões de interesse da imagem. Portanto, o estudo das imagens coloridas por meio da sua decomposição nos respectivos canais, etapa esta realizada no procedimento de pré-processamento, é fundamental para o sucesso da segmentação. Nesse pré-processamento a parte visual já dá indícios de qual canal utilizar na segmentação, entretanto são os histogramas que definem qual canal utilizar na segmentação pelo método proposto por Otsu (1979). Por exemplo, se comparar os histogramas dos três canais da Figura 2A, percebe-se que o canal azul é o que propicia melhor discriminação do fundo da imagem dos discos foliares, uma vez que o histograma desse canal (Figura 2G) é o único que possui duas regiões distintas.

Enquanto o procedimento de segmentação baseado no método proposto por Otsu (1979) depende de duas imagens, o procedimento de segmentação pelas técnicas de aprendizado de máquinas requerem somente as imagens originais. Além de não necessitar de estudo prévio da imagem, esse procedimento permite identificar até três regiões de uma imagem. Esses métodos são mais eficientes em relação ao método proposto por Otsu (1979), entretanto requerem maior demanda computacional especialmente se as imagens

apresentarem elevada dimensão. Isso se deve ao fato desses métodos basearem-se em processo iterativo. Em todos os procedimentos de segmentação as imagens geradas podem ser salvas e utilizadas posteriormente para extração de características.

3.4. EXTRAÇÃO DE CARACTERÍSTICAS

O passo posterior à segmentação da imagem é a extração das características das regiões de interesse. Identificada uma região que pode ser folhas, plantas, sementes ou outras estruturas é possível quantificar medidas físicas como as dimensões, a área, o perímetro ou outras medidas associadas a morfometria geométrica dessas regiões. Além dessas, pode-se extrair informações relativas à cor dessas estruturas, assim como permitir estudos dos histogramas somente da região de interesse nos canais que compõem a imagem. Essa etapa é crucial, uma vez que essas informações podem ser utilizadas para auxiliar em estudos de classificação e predição na área de melhoramento de plantas, ou seja, exercem grande importância em estudos no melhoramento de plantas com enfoque na fenômica (Singh, 2016).

No software FENOM estão disponíveis dois procedimentos de extração de características. O primeiro permite determinar distâncias entre pontos das imagens e convertê-los em medidas reais. Nesse procedimento é possível selecionar a região onde serão obtidas as distâncias na imagem original (Figura 4) para permitir melhor visualização e identificação dos pontos na região selecionada (Figura 4). Este procedimento já foi utilizado para mensurar o comprimento, diâmetro de frutos de abóbora assim como espessura de polpa (Machado Junior, 2017). Nesse estudo, verificou-se elevada correlação entre as medidas realizadas manualmente e as obtidas por esse procedimento.

A B

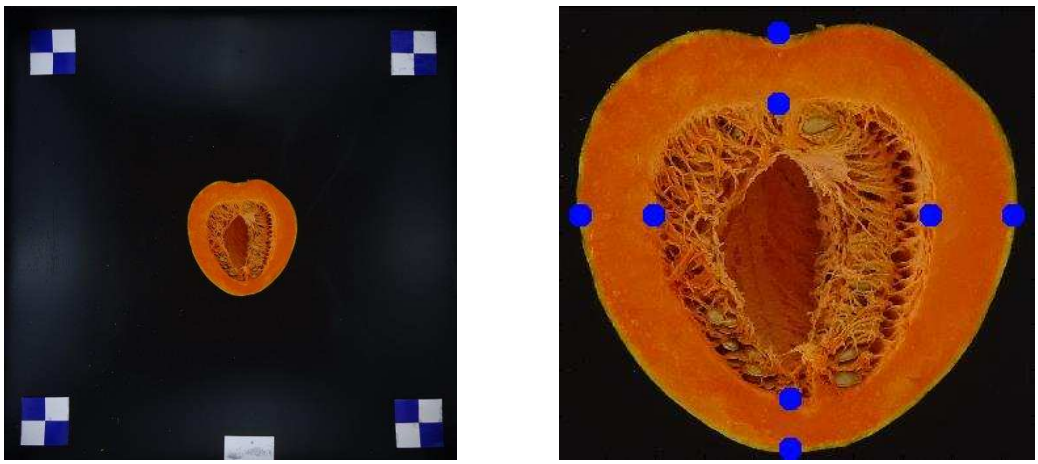


Figura 4. Imagens processadas no procedimento de mensuração de distâncias. A: Imagem original; B: Imagem selecionada com os respectivos pontos a serem medidos.

Outro procedimento utilizado para extração de características permite a contagem de objetos presentes nas imagens e a obtenção de diversas medidas de cada um destes. Dentre essas podem ser obtidas as medidas dos objetos em dois sentidos, a excentricidade, entre outras medidas que permitem distingui-los nas imagens. Esse procedimento consta de um exemplo onde é possível detectar quantas sementes estão presentes na imagem e medir cada uma delas (Figura 5).

Figura 5. Imagem segmentada utilizada no procedimento de detecção e mensuração de objetos.

3.5. CLASSIFICAÇÃO: REDES NEURAIS ARTIFICIAIS

O software FENOM pode ser subdividido em duas áreas de procedimentos: (1) processamento digital de imagens e (2) classificação por meio de redes neurais artificiais. Para processamento de imagens estão disponíveis procedimentos de aquisição, pré-processamento, segmentação e extração de características. Já nos procedimentos de classificação estão disponíveis procedimentos de redes neurais artificiais com arquitetura perceptron multicamadas. Assim, as medidas obtidas por meio dos procedimentos de extração de características e outras obtidas em campo ou laboratório, poderão ser utilizadas para classificar plantas ou genótipos por essa técnica de aprendizado de máquinas. Para cada procedimento foram disponibilizadas as estratégias de k-folds e jackknife para realizar amostragem dos dados nas etapas de treinamento e validação.

Em futuras versões serão implementados procedimentos de classificação e predição por meio de outras técnicas de aprendizado de máquinas. Além disso, no intuito de aumentar a eficácia na classificação com uso de imagens, também será implementada a abordagem de deep learning por meio de redes neurais convolucionais.

4. CONCLUSÃO

FENOM é um software com importantes aplicações no processamento digital de imagens e redes neurais artificiais que podem ser aplicados a avaliações fenotípicas de alta qualidade. Além da contribuição na área de pesquisa, esse software também é importante, pois visa difundir conhecimentos de processamento digital de imagens pouco explorados no melhoramento vegetal.

5. REFERÊNCIAS BIBLIOGRÁFICAS

Angermueller, C., Pärnamaa, T., Parts, L., Stegle, O., 2016. Deep learning for computational biology. Mol. Syst. Biol. 12, 878. doi:10.15252/msb.20156651

Bezdek, J.C., 1984. FCM : The Fuzzy C-Means Clustering Algorithm. Comput. Geosci. 10, 191–203.

Burger, W., Burge, M.J., 2008. Digital Image Processing - An Algorithmic Introduction Using Java., 1st ed. Springer, London. doi:10.1007/978-1-84628-968-2

Carneiro, V.Q., Silva, G.N., Cruz, C.D., Carneiro, P.C.S., Nascimento, M., Carneiro, J.E.S., 2017. Artificial neural networks as auxiliary tools for the improvement of bean plant architecture. Genet. Mol. Res. 16. doi:10.4238/gmr16029500.

Castro, C.A. de O., Resende, R.T., Kuki, K.N., Carneiro, V.Q., Marcatti, G.E., Cruz, C.D., Motoike, S.Y., 2017. High-performance prediction of macauba fruit biomass for agricultural and industrial purposes using Artificial Neural Networks. Ind. Crop. Prod. 108, 806–813. doi:10.1016/j.indcrop.2017.07.031.

Couto, M.F., Nascimento, M., do Amaral, A.T., e Silva, F.F., Viana, A.P., Vivas, M., 2015. Eberhart and Russel's Bayesian Method in the Selection of Popcorn Cultivars. Crop Sci. 55, 571. doi:10.2135/cropsci2014.07.0498.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín D., de los Campos, G., Burgueño, J., González-Camacho J.M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, Rutkoski, J., Varshney,

R.K., 2017. Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. Trends Plant Sci. 22, 961–975, doi: 10.1016/j.tplants.2017.08.011.

Desta, Z.A., Ortiz, R., 2014. Genomic selection: Genome-wide prediction in plant improvement. Trends Plant Sci. 19, 592–601. doi:10.1016/j.tplants.2014.05.006.

Fritsche-Neto, R., Borém, A., 2015. Phenomics: How next-generation phenotyping is revolutionizing plant breeding. Springer International Publishing. doi:10.1007/978-3-319-13677-6.

Glória, L.S., Cruz, C.D., Vieira, R.A.M., de Resende, M.D.V., Lopes, P.S., de Siqueira, O.H.G.B.D., Fonseca e Silva, F., 2016. Accessing marker effects and heritability estimates from genome prediction by Bayesian regularized neural networks. Livest. Sci. 191, 91–96. doi:10.1016/j.livsci.2016.07.015.

Houle, D., Govindaraju, D.R., Omholt, S., 2010. Phenomics: The next challenge. Nat. Rev. Genet. 11, 855–866. doi:10.1038/nrg2897.

Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. Biol. Cybern. doi:10.1007/BF00337288

Lecun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. doi:10.1038/nature14539.

Lima, M.S. De, Carneiro, J.E.D.S., Carneiro, P.C.S., Pereira, C.S., Vieira, R.F., Cecon, P.R., 2012. Characterization of genetic variability among common bean genotypes by morphological descriptors. Crop Breed. Appl. Biotechnol. 12, 76–84. doi:10.1590/S1984-70332012000100010.

Machado Junior, R., 2017. Fenômica aplicada à caracterização de frutos em germoplasma de abóbora. Dissertação. Universidade Federal de Viçosa - Viçosa.

Macqueen, J., 1967. Some methods for classification and analysis of multivariate observations. Proc. Fifth Berkeley Symp. Math. Stat. Probab. 1, 281–297. doi:citeulike-article-id:6083430

MathWorks, Inc. 2012. MATLAB R2012a, The MathWorks, Inc., Natick, Massachusetts, United States.

Nascimento, M., Peternelli, L.A., Cruz, C.D., Campana, A.C.N., Ferreira, R. de P., Bhering, L.L., Salgado, C.C., 2013. Artificial neural networks for adaptability and stability evaluation in alfalfa genotypes. Crop Breed. Appl. Biotechnol. 13, 152–156.

Miranda, J.I., 2011. Processamento de imagens digitais - Métodos multivariados em Java, 1st ed. Embrapa Informática Agropecuária, Campinas.

Moura, M.M., Carneiro, P.C.S., Carneiro, J.E. de S., Cruz, C.D., 2013. Potencial de caracteres na avaliação da arquitetura de plantas de feijão. Pesqui. Agropecuária Bras. 48, 417–425. doi:10.1590/S0100-204X2013000400010.

Mutka, A.M., Bart, R.S., 2015. Image-based phenotyping of plant disease symptoms. Front. Plant Sci. 5, 1–8. doi:10.3389/fpls.2014.00734.

Najafabadi, S.S.M., Farahani, L., 2012. Shape analysis of common bean (Phaseolus vulgaris L.) seeds using image analysis. Int. Res. J. Appl. Basic Sci. 3, 1619–1623.

Oliveira, A.M.C., Batista, R.O., Carneiro, P.C.S., Carneiro, J.E.S., Cruz, C.D., 2015. Potential of hypocotyl diameter in family selection aiming at plant architecture

improvement of common bean. Genet. Mol. Res. 14, 11515–11523. doi:10.4238/2015.September.28.3.

Otsu, N., 1979. A Threshold Selection Method from Gray-Level Histograms. IEEE Trans. Syst. Man. Cybern. 9, 62–66. doi:10.1109/TSMC.1979.4310076.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Sabrol, H., 2015. Recent Studies of Image and Soft Computing Techniques for Plant Disease Recognition and Classification. Int. J. Comput. Appl. 126, 44–55.

Sant'Anna, I.C., Tomaz, R.S., Silva, G.N., Nascimento, M., Bhering, L.L., Cruz, C.D., 2015. Superiority of artificial neural networks for a genetic classification procedure. Genet. Mol. Res. 14, 9898–9906. doi:10.4238/2015.August.19.24.

Schuster, I., Cruz, C.D., 2008. Estatística genômica aplicada a populações derivadas de cruzamentos controlados, 2nd ed. Editora UFV, Viçosa.

Shikha, M., Kanika, A., Rao, A.R., Mallikarjuna, M.G., Gupta, H.S., Nepolean, T., 2017. Genomic Selection for Drought Tolerance Using Genome-Wide SNPs in Maize. Front. Plant Sci. 8, 1–12. doi:10.3389/fpls.2017.00550.

Silva, G.N., Tomaz, R.S., Castro, I. De, Anna, S., Nascimento, M., Bhering, L.L., 2014. Neural networks for predicting breeding values and genetic gains. Sci. Agric. 71, 494–498.

Silva, M.J. da, Pastina, M.M., Souza, V.F. De, Schaffert, E., Carneiro, P.C.S, Noda, R.W., Carneiro, D.S., Maria, C., Damasceno, B., Augusto, R., Parrella, C., 2017. Phenotypic

and molecular characterization of sweet sorghum accessions for bioenergy production 1–19. doi:10.6084/m9.figshare.5212969.

Singh, A., Ganapathysubramanian, B., Singh, A.K., Sarkar, S., 2016. Machine Learning for High-Throughput Stress Phenotyping in Plants. Trends Plant Sci. 21, 110–124. doi:10.1016/j.tplants.2015.10.015.

Solomon, C., Breckon, T., 2013. Fundamentos de processamento digital de imagens: uma abordagem prática com exemplos em matlab, 1st ed. Editora LTC, Rio de Janeiro.

Tattaris, M., Reynolds, M.P., Chapman, S.C., 2016. A Direct Comparison of Remote Sensing Approaches for High-Throughput Phenotyping in Plant Breeding. Front. Plant Sci. 7, 1–9. doi:10.3389/fpls.2016.01131.

Ubbens, J.R., Stavness, I., 2017. Deep Plant Phenomics: A Deep Learning Platform for Complex Plant Phenotyping Tasks. Front. Plant Sci. 8. doi:10.3389/fpls.2017.01190.

CAPÍTULO 2 - BIOFUZZY: LÓGICA FUZZY APLICADA AO MELHORAMENTO VEGETAL

BIOFUZZY: LÓGICA FUZZY APLICADA AO MELHORAMENTO VEGETAL

RESUMO

A eficiência do melhoramento vegetal está intimamente relacionada às tomadas de decisão que são realizadas por meio de interpretações dos resultados obtidos nas análises dos dados experimentais. Diversas são as proposições de metodologias genéticas e estatísticas que visam auxiliar nessas tomadas de decisão. Novas abordagens, como inteligência artificial e especialmente a lógica fuzzy, oferece promissora solução computacional para aumentar a eficiência do melhoramento vegetal. Inspirada na forma de raciocínio humano, a lógica fuzzy possibilita interpretações e tomadas de decisão baseadas em informações aproximadas. Apesar dos problemas que a lógica fuzzy pode solucionar serem muito comuns no melhoramento vegetal, as técnicas baseadas nessa abordagem não têm sido aplicadas com esse objetivo. O desconhecimento e a falta de softwares interativos que contemplem essas análises são os principais motivos pelos quais essa abordagem não é difundida no melhoramento vegetal. Portanto, o objetivo deste trabalho é apresentar o software BioFuzzy, que contempla análises fuzzy biométricas com ênfase no melhoramento vegetal. BioFuzzy está disponível gratuitamente para download no repositório https://github.com/VQCarneiro/BioFuzzy. Esse software disponibiliza procedimentos estatísticos como análises de variância e análises de adaptabilidade e estabilidade. Sob o enfoque da lógica fuzzy, esse software apresenta procedimentos de sistemas de decisão fuzzy e de agrupamento fuzzy C-Means para auxiliar na identificação de genótipos superiores.

Palavras chave: inteligência artificial, software e melhoramento genético

1. INTRODUÇÃO

O melhoramento vegetal desempenha importante função tanto do ponto de vista sócio-econômico como em ações relativas à sustentabilidade agrícola (Borém et al., 2017). A identificação e posterior recomendação de genótipos superiores são algumas das principais atividades dos melhoristas. A eficiência dessas práticas está intimamente relacionada às tomadas de decisão baseadas nas interpretações dos resultados obtidos nas análises dos dados experimentais (Nass, 2001).

Diversas são as proposições de metodologias genéticas e estatísticas que visam auxiliar nas tomadas de decisão no melhoramento vegetal (Cruz et al., 2012, 2014). Não existem dúvidas que a aplicação das técnicas biométricas, para auxiliar nas tomadas de decisão, é uma contribuição indispensável ao melhoramento vegetal. Novas abordagens e metodologias estão cada vez mais em destaque (Silva et al., 2014; Sant'Anna et al., 2015; Carneiro et al., 2017; Carneiro et al., 2018), uma vez que existe grande necessidade em tornar o melhoramento ainda mais eficiente frente às demandas que se apresentam.

O aprofundamento do conhecimento sobre novas abordagens, como a inteligência artificial, oferece promissora solução computacional para auxiliar nessas tomadas de decisão. Existem relatos de que tais técnicas tem sido aplicadas para auxiliar de forma eficiente nas tomadas de decisão em áreas como medicina (Kourou et al., 2015) e mercado financeiro (Cavalcante et al., 2016), que requerem elevada acurácia em suas decisões. Apesar das diversas vantagens dessa abordagem, a inteligência artificial ainda não tem sido amplamente utilizada para auxiliar as diversas etapas de um programa de melhoramento genético.

A lógica fuzzy, proposta por Zadeh (1965), constitui uma subárea da inteligência artificial, inspirada na forma de raciocínio humano (Zadeh, 1975, 1997), que possibilita

análises e interpretações de informações aproximadas (Zadeh, 1983). Esta técnica provê uma forma de subdividir o universo de discurso das variáveis de importância para a tomada de decisão em conjuntos que representam expressões verbais comuns na comunicação humana (Zadeh, 1996; Simões e Shaw, 2007). A nível de teoria de conjuntos fuzzy, qualquer observação pode ser alocada a dois ou mais conjuntos diferentes simultaneamente com níveis de pertencimento distintos (Zadeh, 1965). Esses níveis são comumente denominados por pertinências e são definidos por funções matemáticas, o que constitui como o principal diferencial dessa abordagem (Zadeh, 1999). A lógica fuzzy, por apresentar essa propriedade de multivalência, permite converter a experiência humana em uma forma compreensível pelos computadores (Zadeh, 2008).

As particularidades da lógica fuzzy a qualificam como abordagem muito útil em diversas situações especialmente em tomadas de decisão e como técnica de agrupamento (Bezdek, 1984; Ghosh e Dubey, 2013; Dogruparmak et al., 2014). Essa abordagem é a base para a implementação de procedimentos de agrupamento, como fuzzy C-Means (Bezdek, 1984), muito utilizada no agrupamento de grandes conjuntos de dados e, principalmente, em segmentação de imagens (Feng et al., 2016). Já em relação a técnicas de tomada de decisão, destacam-se os sistemas de tomada de decisão fuzzy, aplicados em algumas áreas como a medicina (Korenevskiy, 2015). No melhoramento vegetal, Carneiro et al. (2018) propuseram a utilização de sistemas dessa natureza para auxiliar nas decisões de recomendação de cultivares.

As técnicas baseadas na lógica fuzzy têm potencial para auxiliar em diversos estudos no melhoramento vegetal. Entretanto, o desconhecimento por parte dos melhoristas a respeito da lógica fuzzy e a falta de softwares interativos que contemplem essas análises são os principais motivos pelos quais essa abordagem não é difundida no

melhoramento vegetal. Portanto, o objetivo deste trabalho é apresentar o software BioFuzzy, que contempla análises fuzzy biométricas com ênfase no melhoramento vegetal. Este software além de caráter de pesquisa, aborda a lógica fuzzy sob o aspecto de ensino, uma vez que um de seus objetivos é difundir o conhecimento a respeito dessa abordagem.

2. DESCRIÇÃO

O software BioFuzzy é compatível com sistema operacional Windows. Este software é desenvolvido por meio da integração do software Matlab com as linguagens Python e JAVA. Portanto, para o funcionamento deste software é necessário que o usuário tenha instalado o software Matlab (https://www.mathworks.com/), as linguagens Python (https://www.python.org/) e Java (https://www.java.com/pt_BR/) no computador. O Matlab, por meio da fuzzy logic toolbox, foi utilizado para o desenvolvimento de procedimentos que utilizam a lógica fuzzy. Na linguagem Python foram implementados procedimentos estatísticos. Já a linguagem JAVA foi utilizada para desenvolver a interface gráfica do software de forma a tornar o acesso aos procedimentos em Python e em Matlab mais simples e prático.

O software BioFuzzy está disponível gratuitamente para download no repositório https://github.com/VQCarneiro/BioFuzzy. A descrição a respeito deste software pode ser encontrada nesse repositório. Não há necessidade de realizar instalação deste software no computador. Basta fazer o download da pasta BioFuzzy no diretório C: (por exemplo: C:\BioFuzzy). Os pré-requistos para utilização do BioFuzzy é a instalação no computador do software Matlab com a fuzzy logic toobox, as linguagens Python e JAVA.

3. PROCEDIMENTOS

No intuito de tornar o software BioFuzzy (Figura 1) capaz de realizar procedimentos de tomada de decisão que permitissem identificar e recomendar genótipos superiores foram implementados procedimentos estatísticos de análises de variância e de adaptabilidade e estabilidade. Sob o aspecto da lógica fuzzy, foram implementados sistemas de tomada de decisão para auxiliar nas interpretações dos resultados das análises de adaptabilidade e estabilidade. Além desses procedimentos, foi implementada a análise de agrupamento fuzzy C-Means. Esse software ainda dispõe de apresentações e procedimentos de ensino para melhor entendimento da lógica fuzzy disponíveis na seção Ensino.

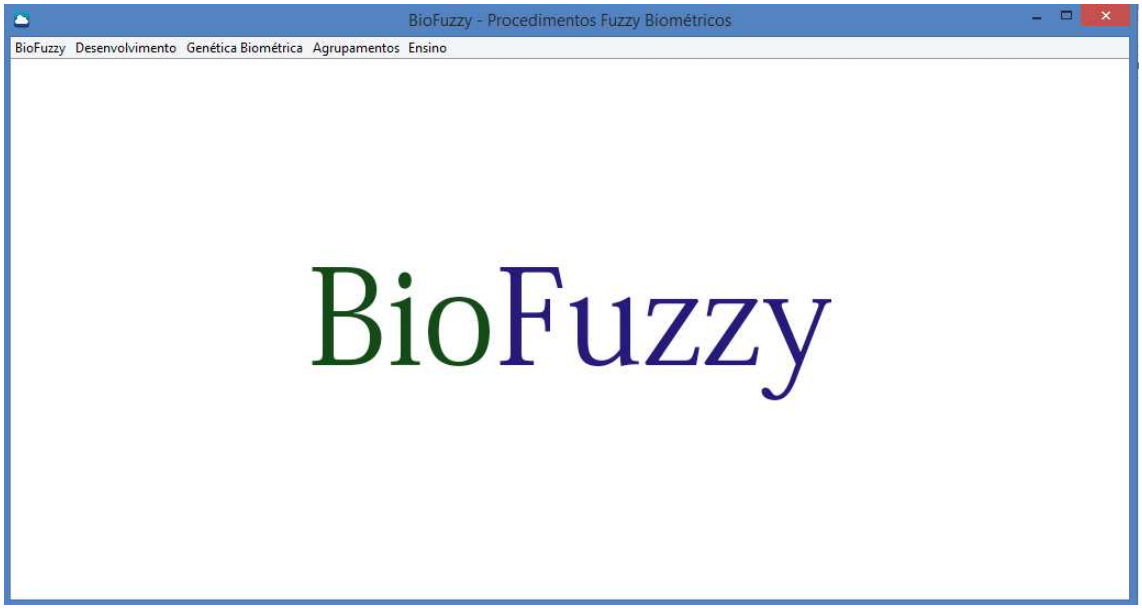


Figura 1. Interface do software BioFuzzy.

Todos os procedimentos contam com sistema de leitura de dados em formato texto (extensão .txt ou compatível), facilmente editados e acessados em bloco de notas. Para tornar mais simples e didática a utilização desse software foram implementados exemplos para cada um dos procedimentos. A declaração de informações pertinentes às análises são realizadas na própria interface gráfica do software. Além disso, o software conta com

sistema de diagnóstico do arquivo de dados para cada análise afim de facilitar a realização das análises e identificação de possíveis erros. Uma vez realizadas as análises, são gerados relatórios para facilitar a interpretação dos resultados. Adicionalmente, para cada análise são gerados arquivos contendo valores de médias no formato de bloco de notas (.txt) que podem ser utilizados para análises posteriores ou para gerar visualizações gráficas.

3.1. ANÁLISES ESTATÍSTICAS

I. Fundamentação Teórica

A estatística têm sido utilizada no melhoramento vegetal com o intuito de permitir compreender melhor os fenômenos biológicos. A associação dos conhecimentos genéticos com os métodos estatísticos constitui em um dos principais pilares dos modelos biométricos adotados no melhoramento vegetal (Cruz et al., 2012, 2014). Dentre as diversas técnicas estatísticas, as análises de variância são aplicadas rotineiramente nos programas de melhoramento, especialmente quando o intuito é testar hipóteses relativas a existência de variabilidade genética, quantificar componentes de variância e obter estimativas de parâmetros para fins de selecionar genótipos superiores (Alves et al., 2015) ou recomendar cultivares (Barili et al., 2015).

As análises de variância podem ser subdivididas em individual ou conjunta conforme a natureza dos experimentos e dos modelos adotados para analisar os dados experimentais (Ramalho et al., 2012). No contexto do melhoramento vegetal, as análises individuais de variância são realizadas quando genótipos são avaliados em um único ambiente. Os modelos adotados para essas análises variam conforme o delineamento estatístico utilizado nos experimentos. O modelo para delineamento inteiramente

casualizado considera $Y_{ij} = \mu + G_i + \varepsilon_{ij}$, em que Y_{ij} é o valor observado do i-ésimo genótipo na j-ésima repetição, μ é a média geral, G_i é o efeito do i-ésimo genótipo e ε_{ij} é o efeito do erro aleatório associado à observação de ordem ij (Cruz et al., 2012). Para delineamento em blocos casualizados adota-se o modelo $Y_{ij} = \mu + B_j + G_i + \varepsilon_{ij}$, em que Y_{ij} é o valor observado do i-ésimo genótipo no j-ésimo bloco, μ é a média geral, B_j é o efeito do j-ésimo bloco, G_i é o efeito do i-ésimo genótipo e ε_{ij} é o erro aleatório associado à observação de ordem ij.

As análises de natureza conjunta são aplicadas quando os mesmos genótipos são avaliados em dois ou mais experimentos implantados em ambientes (locais, anos ou safras) distintos. O modelo para delineamento em blocos casualizados considera: $Y_{ijk} = \mu + B/A_{jk} + G_i + A_j + GA_{ij} + \varepsilon_{ijk}$, em que Y_{ijk} é o valor observado do i-ésimo genótipo no j-ésimo ambiente e no k-ésimo bloco, μ é a média geral, B/A_{jk} é o efeito do k-ésimo bloco dentro do j-ésimo ambiente, G_i é o efeito do i-ésimo genótipo, A_j é o efeito do j-ésimo ambiente, GA_{ij} é o efeito da interação do i-ésimo genótipo com o j-ésimo ambiente e ε_{ijk} é o erro aleatório associado à observação de ordem ijk (Van Eeuwijk et al., 2016).

As análises conjuntas de variância são muito utilizadas na etapa de recomendação de cultivares, especialmente em dados oriundos de ensaios de valor de cultivo e uso (VCU), nos quais o intuito é avaliar o desempenho dos genótipos ao longo dos ambientes. Entretanto, a análise conjunta de variância proporciona informação somente a respeito da existência da interação genótipos por ambientes (Cruz et al, 2012). Assim, análises complementares como de adaptabilidade e estabilidade são essenciais para obter informações detalhadas a respeito dos genótipos nos diferentes ambientes de avaliação dos VCU's, especialmente quando se identifica a existência de interação de genótipos por ambientes.

Diversas são as propostas de métodos para avaliar a adaptabilidade e estabilidade de genótipos conduzidos em vários ambientes (Cruz et al., 2012, 2014). Métodos baseados em regressão linear como o proposto por Eberhart e Russell (1966) estão entre os mais adotados. Essa metodologia permite avaliar a resposta dos genótipos por meio de uma regressão das médias dos genótipos em cada ambiente em função do índice ambiental proposto por Finlay e Wilkinson (1963) e modificado por Eberhart e Russell (1966). O método proposto por Cruz et al. (1989) apresenta princípio similar ao proposto por Eberhart e Russell (1966), entretanto a regressão adotada é bissegmentada. Ambos os métodos são de natureza paramétrica e dispõe de elevado número de parâmetros a ser interpretados para compreender a resposta dos genótipos nos ambientes de avaliação. Essa constitui a principal dificuldade e limitação desses métodos, pois exige grande expertise dos melhoristas para interpretar os resultados.

Os métodos não paramétricos foram propostos com o intuito de facilitar as interpretações das análises de adaptabilidade e estabilidade. Métodos baseados em ideótipos como o proposto por Lin e Binns (1988) e suas modificações (Cruz et al., 2012) são muito utilizados devido a facilidade de interpretação. Relata-se que essa metodologia assim como a proposta por Eberhart e Russell (1966) são as mais utilizadas no melhoramento do feijoeiro (Backes et al., 2005; Melo et al., 2007; Ribeiro et al., 2008, 2009).

II. Procedimentos

No software Biofuzzy, as análises de variância foram implementadas como o primeiro passo para identificar potenciais genótipos a serem selecionados e posteriormente recomendados. As análises de variância, tanto individual como conjunta,

foram implementadas em linguagem Python, devido a facilidade que esta linguagem possui no processamento de dados. Nesses procedimentos além de relatórios são gerados arquivos em extensão .txt (bloco de notas) com médias dos genótipos em análises individuais e médias de genótipos por ambientes na análise conjunta. Esses arquivos são utilizados para realizar análises de adaptabilidade e estabilidade ou em sistemas de tomada de decisão fuzzy.

Os métodos de adaptabilidade e estabilidade propostos por Eberhart e Russell (1966), Cruz et al. (1989) e Lin e Binns (Lin e Binns, 1988; Cruz et al., 2012) também foram implementados no BioFuzzy. Os parâmetros ou medidas oriundos dessas análises são armazenados em arquivos em extensão .txt (bloco de notas) para serem utilizados como entradas em sistemas de tomada de decisão fuzzy.

3.2. LÓGICA FUZZY

3.2.1. SISTEMAS DE TOMADA DE DECISÃO FUZZY

I. Fundamentação Teórica

Os sistemas de tomada de decisão fuzzy utilizam toda a teoria da lógica fuzzy com o intuito de desenvolver sistemas capazes de auxiliar em tomadas de decisão em diferentes ocasiões. Para isso as variáveis que determinam as decisões são tratadas sob o enfoque dessa abordagem, ou seja, são consideradas como variáveis de entrada fuzzy do sistema. A partir desse momento, o universo de discurso de cada uma dessas variáveis é particionado em grupos conforme o conhecimento do desenvolvedor do sistema a respeito de cada variável (Simões e Shaw, 2007). Diferentemente da teoria de conjuntos clássicos,

os conjuntos fuzzy podem ser representados por meio de diferentes funções de pertinências, que podem ser lineares ou não (Jang et al., 2012). De acordo com essas funções de pertinência, cada valor do universo de discurso está associado a um grau de pertencimento (pertinência) aos grupos que compõem cada variável.

Além das variáveis de entrada, também existe uma variável de saída que reúne todo o conhecimento do sistema, que também é considerada sob o enfoque fuzzy. Portanto, existe uma relação de dependência entre a variável de saída e as de entrada. Essa relação é determinada pelas regras fuzzy que agregam informações das variáveis de entrada de forma a gerar uma saída. Essas regras são estabelecidas conforme o conhecimento empírico do desenvolvedor do sistema fuzzy sobre o problema em questão. Portanto, é imprescindível que esse desenvolvedor seja um especialista na área que deseja desenvolver esse sistema, uma vez que são essas regras que determinam a tomada de decisão.

Os dados coletados são processados nos sistemas de tomada de decisão fuzzy por uma série de etapas (Figura 2). A primeira delas é a fuzzyficação na qual os dados numéricos para cada variável são submetidos às suas respectivas funções de pertinência. Esses valores de pertinências são utilizados na etapa de implicação, que consiste no processamento de cada regra individualmente de modo a obter os valores de pertinências para a observação em questão na variável de saída. Portanto, para cada regra são obtidas as pertinências na variável de saída. A agregação consiste em combinar as saídas de todas as regras em um único resultado (Jang et al., 2012). Entretanto, esse resultado está na forma de pertinência e muitas vezes é necessário converte-lo em valores discretos, que é realizado na etapa de defuzzyficação. Todo o processamento dos dados em um sistema fuzzy é relativamente simples em que o custo computacional é dependente do número de

variáveis de entrada e de funções de pertinência. Sistemas dessa natureza podem ser adaptados para auxiliar nas tomadas de decisão no melhoramento vegetal.

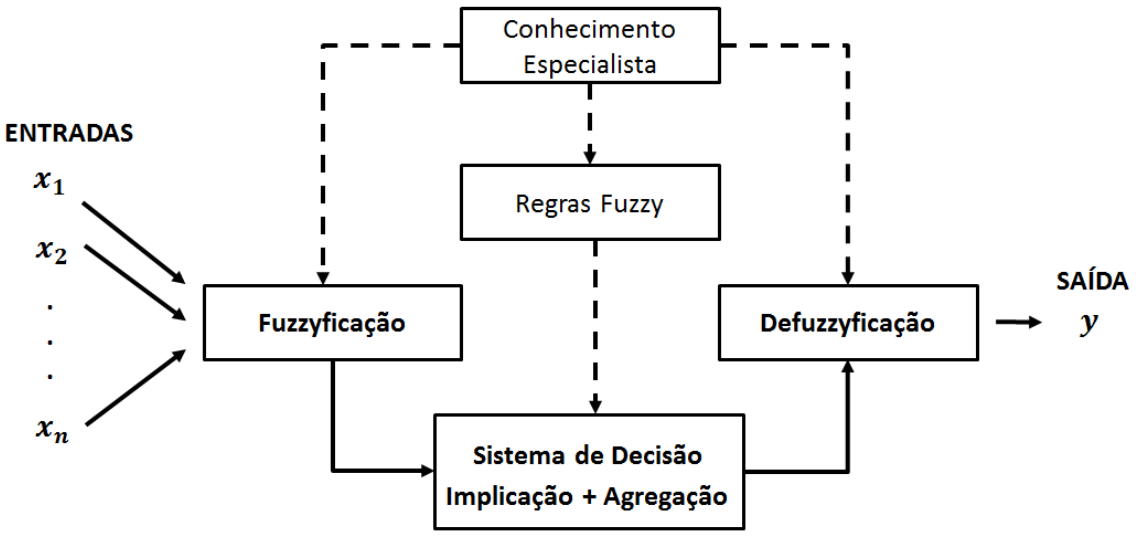


Figura 2. Estrutura de sistema de tomada de decisão fuzzy.

II. Procedimentos

No BioFuzzy foi implementado um procedimento que permite o próprio usuário desenvolver sistemas de tomada de decisão fuzzy. Esse procedimento utiliza a própria interface de lógica fuzzy do Matlab. Nesse procedimento é possível criar sistemas fuzzy com uso de diferentes funções de pertinências assim como métodos de implicação, agregação e defuzzyficação. As funções de pertinência disponíveis no software podem ser visualizadas na seção ensino (conjuntos fuzzy) do software BioFuzzy.

Sistemas de tomada de decisão fuzzy específicos também foram implementados no software BioFuzzy com o objetivo de auxiliar a identificação de genótipos superiores e também na recomendação de cultivares. Para tanto, foi proposto um sistema fuzzy com o intuito de seleção fenotípica de genótipos avaliados em caráter multivariado e oito sistemas de recomendação de cultivares. No sistema fuzzy de seleção de genótipos, cada variável avaliada é considerada sob o aspecto fuzzy e, portanto, particionada em dois

grupos (Médias, Baixas ou Altas). Como variável fuzzy de saída foi proposta uma variável que é particionada em dois grupos (Selecionados ou Excluídos). Para um genótipo ser selecionado todas as médias das variáveis devem pertencer predominantemente ao grupo de médias altas.

Os sistemas fuzzy de recomendação de cultivares apresentam como objetivo tornar a recomendação de cultivares mais simples e prática, uma vez que algumas metodologias (Eberhart e Russell, 1966, Cruz et al., 1989) dispõe de elevado número de parâmetros a serem interpretados. Característica essa que torna a recomendação complexa, especialmente em experimentos que são avaliados grande número de genótipos. Para as metodologias propostas por Eberhart e Russell (1966), Cruz et al. (1989) e Lin e Binns (1988) modificado (Cruz et al., 2012) foram implementados sistemas fuzzy semelhantes aos propostos por Carneiro (2018). Nesses sistemas, as entradas fuzzy são os parâmetros e/ou medidas oriundas dessas metodologias, enquanto que a saída é uma variável fuzzy particionada em quatro grupos (Adaptabilidade Geral, Adaptabilidade a ambientes favoráveis, Adaptabilidade a ambientes desfavoráveis e não recomendado). Além desses sistemas, também foi implementado um sistema fuzzy híbrido, cujo as entradas são parâmetros dos métodos propostos por Eberhart e Russell (1966) e Lin e Binns (1988) modificado (Cruz et al., 2012), semelhante ao proposto por Carneiro (2018).

Além dos sistemas fuzzy baseados em metodologias estatísticas, foram propostos outros dois sistemas fuzzy de recomendação de cultivares. Um desses considera como entradas as médias dos genótipos em cada ambiente, enquanto que o outro considera como entrada as médias de cada genótipo em ambientes favoráveis e desfavoráveis. A classificação dos ambientes em favoráveis e desfavoráveis é realizada por meio do índice ambiental proposto por Finlay e Wilkinson (1963) e modificado por Eberhart e Russell

(1966). A variável de saída desses sistemas são similares a variável de saída dos sistemas baseados em metodologias estatísticas (sistemas fuzzy de recomendação de cultivares).

3.2.2. AGRUPAMENTO FUZZY C-MEANS

I. Fundamentação Teórica

A técnica de agrupamento fuzzy C-Means proposta por Bezdek et al. (1984) é uma modificação sob o enfoque fuzzy do método k-Means (Macqueen, 1967) que se baseia no agrupamento de observações em um número pré-determinado de grupos. Diferentemente da técnica k-Means, essa abordagem de agrupamento fuzzy considera que uma observação pode pertencer simultaneamente a todos os grupos, porém com graus de pertencimento (pertinências) distintos que variam entre 0 e 1 (Ghosh e Dubey, 2013). A soma das pertinências de cada observação contabiliza uma unidade ($\sum_{i=1}^{c} u_{ij} = 1, \ \forall j = 1, \dots, n$) e, portanto, verifica-se que essa restrição é considerada para as n observações em estudo.

A alocação das observações em cada grupo pelo método fuzzy C-Means baseia-se na minimização da função custo:

$$J(U, c_1, \dots, c_c) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} d_{ij}^{2} = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} \left\| c_i - x_j \right\|^2,$$

que pode ser reescrita como:

$$\bar{J}(U, c_1, \dots, c_c, \lambda_1, \dots, \lambda_n) = \sum_{i=1}^{c} \left(\sum_{j=1}^{n} u_{ij}^{m} \left\| c_i - x_j \right\|^2 \right) + \sum_{j=1}^{n} \lambda_j \left(\sum_{i=1}^{c} u_{ij} - 1 \right)$$ ao considerar as n restrições $\sum_{i=1}^{c} u_{ij} = 1, \ \forall j = 1, \dots, n$. Nessas funções, U é a matriz de pertinências ($u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)}}$) das j observações nos i grupos; c_i é o vetor de

informações do centroide do i-ésimo grupo ($c_i = \frac{\sum_j^n u_{ij}^m x_j}{\sum_j^n u_{ij}^m}$); m é o expoente de ponderação fuzzy (índice de difusividade), d_{ij}^2 é a distância euclidiana entre o vetor de informações de cada observação (x_j) e o i-ésimo centroide (c_i) e λ_j são os multiplicadores de Lagrange para as n restrições ($\sum_{i=1}^c u_{ij} = 1\ ,\ \ \forall\, j = 1, \ldots, n$).

A técnica fuzzy C-Means baseia-se em processo iterativo para alocar as observações e determinar as pertinências destas em cada grupo conforme os seguintes passos: i. inicialização da matriz U com valores aleatórios de pertinências entre 0 e 1, de modo que a restrição $\sum_{i=1}^c u_{ij} = 1$ seja respeitada; ii. Obtenção dos centroides dos c grupos por meio da expressão $c_i = \frac{\sum_j^n u_{ij}^m x_j}{\sum_j^n u_{ij}^m}$; iii. Obtenção do valor da função custo e observação do critério de parada do processo iterativo; iv. em caso de não cumprimento do critério de parada, o processo iterativo reinicia de modo que se obtém uma nova matriz U de pertinências por meio da expressão $u_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}}$ (Miranda, 2011; Jang et al., 2012; Bezdek 1984).

II. Procedimentos

No software BioFuzzy foram implementados dois procedimentos que abordam a técnica fuzzy C-Means. Um destes é a aplicação desse método não supervisionado para agrupar observações em número pré-determinado de grupos. O outro procedimento consiste na adaptação desse método para auxiliar na recomendação de cultivares, semelhante ao método centroide proposto por Rocha et al. (2005), ou seja, identificar as respostas dos genótipos por meio da proximidade destes a ideótipos previamente

estabelecidos. A grande vantagem dessa adaptação é a utilização das pertinências fuzzy para distinguir as respostas dos genótipos.

4. CONCLUSÃO

BioFuzzy é um software com importantes aplicações no processamento de dados experimentais aplicados ao melhoramento vegetal. Esse por contar com procedimentos fuzzy biométricos é uma importante contribuição à análise de dados dessa área, especialmente por não existirem softwares com procedimentos dessa natureza. Além da contribuição na área de pesquisa, esse software também é importante, pois visa difundir conhecimentos de lógica fuzzy pouco explorados no melhoramento vegetal.

5. REFERÊNCIAS BIBLIOGRÁFICAS

Alves, A.F., Menezes Júnior, J.A.N. De, Menezes, V.M.P.S., Carneiro, J.E.D.S., Carneiro, P.C.S., Alves, A.F., 2015. Genetic progress and potential of common bean families obtained by recurrent selection. Crop Breed. Appl. Biotechnol. 15, 218–226. doi:10.1590/1984-70332015v15n4a38

Backes, R.L., Elias, H.T., Hemp, S., Nicknich, W., 2005. Análise de estabilidade de genótipos de feijoeiro no Estado de Santa Catarina. Acta Sci. Agron. 27, 623–628. doi:10.4025/actasciagron.v27i4.1675

Barili, L.D., Vale, N.M. Do, Prado, A.L. Do, Carneiro, J.E.D.S., Silva, F.F.E., Nascimento, M., 2015. Genotype-environment interaction in common bean cultivars with

carioca grain, recommended for cultivation in Brazil in the last 40 years. Crop Breed. Appl. Biotechnol. 15, 244–250. doi:10.1590/1984-70332015v15n4a41

Bezdek, J.C., 1984. FCM : The Fuzzy C-Means Clustering Algorithm. Comput. Geosci. 10, 191–203.

Borém, A., Miranda, G. V., Fritsche-Neto, R., 2017. Melhoramento de Plantas, 7th ed. Editora UFV, Viçosa.

Carneiro, V.Q., Adalgisa Leles do Prado, Cosme Damião Cruz, Carneiro, P.C.S., Nascimento, M., Carneiro, J.E. de S., 2018. Fuzzy control systems for decision-making in cultivars recommendation. Acta Sci. Agron. 40, 1–8. doi:10.4025/actasciagron.v40i1.39314

Carneiro, V.Q., Silva, G.N., Cruz, C.D., Carneiro, P.C.S., Nascimento, M., Carneiro, J.E.S., 2017. Artificial neural networks as auxiliary tools for the improvement of bean plant architecture. Genet. Mol. Res. 16. doi:10.4238/gmr16029500

Cavalcante, R.C., Brasileiro, R.C., Souza, V.L.F., Nobrega, J.P., Oliveira, A.L.I., 2016. Computational Intelligence and Financial Markets: A Survey and Future Directions. Expert Syst. Appl. 55, 194–211. doi:10.1016/j.eswa.2016.02.006

Cruz, C.D., Carneiro, P.C.S., Regazzi, A.J., 2014. Modelos biométricos aplicados ao melhoramento genético - Vol 2, 3rd ed. Editora UFV, Viçosa.

Cruz, C.D., Regazzi, A.J., Carneiro, P.C.S., 2012. Modelos biométricos aplicados ao melhoramento genético – Vol 1, 4th ed. Editora UFV, Viçosa.

Cruz, C.D., Torres, R.A. de A., Vencovsky, R., 1989. An alternative approach to the stability analysis proposed by Silva and Barreto. Rev. Bras. Genética 12, 567–580.

Dogruparmak, S.C., Keskin, G.A., Yaman, S., Alkan, A., 2014. Using principal component analysis and fuzzy c–means clustering for the assessment of air quality monitoring. Atmos. Pollut. Res. 5, 656–663. doi:10.5094/APR.2014.075

Eberhart, S. A., Russell, W.A., 1966. Stability Parameters for Comparing Varieties. Crop Sci. 6, 36. doi:10.2135/cropsci1966.0011183X000600010011x

Feng, Y., Guo, H., Zhang, H., Li, C., Sun, L., Mutic, S., 2016. A modified fuzzy C-means method for segmenting MR images using non-local information 24, 785–793. doi:10.3233/THC-161208

Finlay, K., Wilkinson, G., 1963. The analysis of adaptation in a plant-breeding programme. Aust. J. Agric. Res. doi:10.1071/AR9630742

Ghosh, S., Dubey, S.K.S., 2013. Comparative analysis of k-means and fuzzy c-means algorithms. Ijacsa 4, 35–38. doi:10.14569/IJACSA.2013.040406

Jang, J.S.R., Sun, C.T., Mizutani, E., 2012. Neuro-Fuzzy and Soft Computing - A computacional Approach to Learning and Machine Intelligence. PHI Learning Private Limited, New Delhi.

Korenevskiy, N. a., 2015. Application of fuzzy logic for decision-making in medical expert systems. Biomed. Eng. (NY). 49, 33–35. doi:10.1007/s10527-015-9494-X

Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M. V., Fotiadis, D.I., 2015. Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. 13, 8–17. doi:10.1016/j.csbj.2014.11.005

Lin, C.S., Binns, M.R., 1988. A superiority measure of cultivar performance for cultivar × location data. Can. J. Plant Sci. 68, 193-198. doi:10.4141/cjps88-018

Macqueen, J. B., 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Symposium on Math, Statistics, and Probability. Berkeley, CA: University of California Press. 281–297.

Melo, L.C., Melo, P.G.S., De Faria, L.C., Diaz, J.L.C., Del Peloso, M.J., Rava, C.A., Da Costa, J.G.C., 2007. Interação com ambientes e estabilidade de genótipos de feijoeiro-comum na Região Centro-Sul do Brasil. Pesqui. Agropecu. Bras. 42, 715–723. doi:10.1590/S0100-204X2007000500015

Miranda, J.I., 2011. Processamento de imagens digitais - Métodos multivariados em Java, 1st ed. Embrapa Informática Agropecuária, Campinas.

Nass, L.L., 2001. Recursos genéticos e melhoramento de plantas., 1st ed. Fundação MT, Rondonópolis.

Ramalho, M.A.P., Ferreira, D.F., Oliveira, A.C. de., 2012. Experimentação em genética e melhoramento de plantas. 3st ed. Lavras: UFLA.

Ribeiro, N.D., Antunes, I.F., Souza, J.F. de, Poersch, N.L., 2008. Adaptação e estabilidade de produção de cultivares e linhagens-elite de feijão no Estado do Rio Grande do Sul. Ciência Rural. doi:10.1590/S0103-84782008005000018

Ribeiro, N.D., Souza, J.F. de, Antunes, I.F., Poersch, N.L., 2009. Estabilidade de produção de cultivares de feijão de diferentes grupos comerciais no Estado do Rio Grande do Sul. Bragantia 68, 339–346. doi:10.1590/S0006-87052009000200007

Rocha, R.B., Muro-Abad, J.I., Araújo, E.F., and Cruz, C.D., 2005. Evaluation of the centroid method for study of environment adaptability of clones of Eucalyptus grandis. Cienc. Florestal. 15: 255–266.

Sant'Anna, I.C., Tomaz, R.S., Silva, G.N., Nascimento, M., Bhering, L.L., Cruz, C.D., 2015. Superiority of artificial neural networks for a genetic classification procedure. Genet. Mol. Res. 14, 9898–9906. doi:10.4238/2015.

Silva, G.N., Tomaz, R.S., Castro, I. De, Anna, S., Nascimento, M., Bhering, L.L., 2014. Neural networks for predicting breeding values and genetic gains. Sci. Agric. 71, 494–498.

Simões, M.G., Shaw, I.S., 2007. Controle e Modelagem Fuzzy, 2nd ed. Edgard Bulcher, São Paulo.

Van Eeuwijk, F.A., Bustos-Korts, D. V., Malosetti, M., 2016. What should students in plant breeding know about the statistical aspects of genotype × Environment interactions? Crop Sci. 56, 2119–2140. https://doi.org/10.2135/cropsci2015.06.0375

Zadeh, L. A., 1965. Fuzzy sets. Inf. Control 8, 338–353. doi:10.1016/S0019-9958(65)90241-X

Zadeh, L. A., 1975. The concept of a linguistic variable and its application to approximate reasoning—II. Inf. Sci. (Ny). 8, 301–357. doi:10.1016/0020-0255(75)90046-8

Zadeh, L. A., 1983. A computational approach to fuzzy quantifiers in natural languages. Comput. Math. with Appl. 9, 149–184. doi:10.1016/0898-1221(83)90013-5

Zadeh, L. A., 1996. Fuzzy logic equals Computing with words. Ieee Trans. Fuzzy Syst. 4, 103–111.

Zadeh, L. A., 1997. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets Syst. 90, 111–127. doi:10.1016/S0165-0114(97)00077-8

Zadeh, L. A., 1999. A theory of possibility distributions. Fuzzy Sets Syst. 102, 135–155. doi:10.1016/S0165-0114(97)00102-4

Zadeh, L. A., 2008. Is there a need for fuzzy logic? Inf. Sci. (Ny). 178, 2751–2779. doi:10.1016/j.ins.2008.02.012

CAPÍTULO 3 – APPLICATION OF FUZZY CLUSTERING IN CULTIVARS RECOMMENDATION

APPLICATION OF FUZZY CLUSTERING IN CULTIVARS RECOMMENDATION

ABSTRACT

The genotypes x environments interaction is the main factor that influences the response of evaluated genotypes in trials of value for cultivation and use. Adaptability and stability analyses are fundamental to understand the performance of genotypes in a growing region. Some of these methodologies incorporate previous information about recommending of an extra group of genotypes denominated as specific ideotypes to certain cultivation conditions. Based on this strategy, the centroid method and its modifications have been widely used due to the simplicity of classification of the evaluated genotypes. However, these methodologies present problems to identify adaptability patterns of some genotypes. Artificial intelligence techniques as fuzzy C-Means can be an alternative to reduce these difficulties, since its uses, besides distance information between genotypes, the memberships (measures that quantify how much an observation belongs to a particular class) to increase discriminatory power. Therefore, our aim was to propose and evaluate the phenotypic adaptability by fuzzy clustering method to assist the cultivars recommendation. The adaptation of the fuzzy C-Means method to classify the genotypes was implemented in BioFuzzy software. The grain yield data of black common bean genotypes were used in order to evaluate the potential of the method. The results obtained by this method were compared with those obtained by the centroid method. The phenotypic adaptability method by clustering fuzzy was effective to identify the adaptability pattern of common bean genotypes. Moreover, the discriminative power was higher than the observed in the centroid method.

Keywords: Artificial intelligence, fuzzy logic, plant breeding.

1. INTRODUCTION

The evaluation of elite genotypes in trials of value for cultivation and use (VCU) is essential to identify potential genotypes that will be recommended to growers. In order to verify how these genotypes respond to the environmental variations, these trials are performed in a set of environments that represents the growing regions of each crop.

The genotypes x environments interaction (GE interaction) is the main factor that influences the response of evaluated genotypes in VCU trials (Malosetti et al., 2013; Van Eeuwijk et al., 2016). Besides of studies about the nature of GE interaction (Robertson, 1959; Cruz & Castoldi, 1991), adaptability and stability analyses are fundamental to understanding the genotype performances in a growing region. These analyses can be based on different statistical principles such as regression (Finlay & Wilkinson, 1963; Eberhart & Russell, 1966; Nascimento et al., 2011), mixed models (Resende, 2004), principal components (Gauch & Zobel, 1996; Yan et al., 2000) or factor analysis (Murakami & Cruz, 2004). For recommending cultivars purpose, these analyses involve classification procedures in which the genotypes are allocated in classes of adaptability and stability according to the performance in the evaluated environments. Although these methods are widely used, some present a high number of parameters to be interpreted, which makes the cultivars recommendation even more complex (Cruz et al., 2012).

In statistical approaches, there are adaptability and stability methodologies (Lin & Binns, 1988; Rocha et al., 2005; Vasconcelos et al., 2011; Nascimento et al., 2015) that incorporate in the analysis, previous information about recommending of an extra group of genotypes denominated as specific ideotypes to certain cultivation conditions. Based on this strategy, the centroid method (Rocha et al., 2005) and its modifications (Vasconcelos et al.,

2011; Nascimento et al., 2015) have been widely used by breeding programs due to the simplicity of classification of the evaluated genotypes. This classification is based in the Cartesian distance of the evaluated genotypes to four pre-established ideotypes. Despite providing practical results, these methods present difficulties in identifying the adaptability of genotypes with similar proximity to two or more ideotypes (Vasconcelos et al., 2011).

Artificial intelligence techniques such as artificial neural networks (ANN) have also been used in cultivars recommendation (Barroso et al., 2013; Nascimento et al., 2013; Teodoro et al., 2015). These methodologies present recognized potential solving classificatory and predictive problems (Silva et al., 2014; Sant'Anna et al., 2015; Glória et al., 2016; Carneiro et al., 2017). This approach applied to classificatory problems requires the prior knowledge of the classifications of the set of observations in the training process (Ma et al., 2014; Singh et al., 2016). Therefore, for using ANN in the cultivars recommendation, which in statistical approach is a classificatory problem, is necessary the previously known about the adaptability and/or stability of a dataset for train an artificial neural network. Despite efficient, these techniques require high processing time and computational demand.

Unsupervised artificial intelligence techniques are not used yet for assist in recommending cultivars. These techniques dispense the previous known about the classification of the observations and require which the algorithm itself to find unknown patterns in the data (Schmidhuber, 2015). This is the principle of clustering techniques such as k-Means (Macqueen, 1967), fuzzy C-Means (FCM) (Bezdek, 1984) and self-organizing map, (Kohonen, 1982) used to organize and categorize data (Jang, Sun, & Mizutani, 2012).

The FCM can be an alternative to reduce the difficulties of identifying grouping patterns regarding phenotypic adaptability, since its uses, besides distance information between genotypes, the memberships (measures that quantify how much an observation belongs to a particular class) to increase discriminatory power. Therefore, our aim is to propose and evaluate the phenotypic adaptability method by fuzzy clustering to assist in the cultivars recommendation.

2. MATERIALS AND METHODS

The grain yield data ($kg.ha^{-1}$) of black common bean genotypes described in detail by Oliveira et al. (2006) were used in order to evaluate the potential of the phenotypic adaptability method by fuzzy clustering. In a succinct way, 18 lines and the two cultivars, Ouro Negro and Valente, (Table 1) were evaluated in 12 experiments conducted in a randomized block design with three replications. These experiments were carried out in six cities of Minas Gerais State (Viçosa, Coimbra, Ponte Nova, Leopoldina, Florestal and Capinópolis) in the fall-winter seasons in 2002; and in the dry and fall-winter seasons in 2003.

Table 1. Identification of black common bean genotypes.

Identification	Genotypes	Identification	Genotypes
1	VP1	11	VP11
2	VP2	12	VP12
3	VP3	13	VP13
4	VP4	14	Vi 5700
5	VP5	15	Vi 5500
6	VP6	16	Vi 7800
7	VP7	17	CNFP 9346
8	VP8	18	CNFP 7988
9	VP9	19	Ouro Negro
10	VP10	20	Valente

The data were submitted to individual analysis of variance, which was performed assuming fixed effect for genotypes. The grouping of means was performed according to the Scott-Knott test (Scott & Knott, 1974) for the experiments in which a significant effect of genotypes was verified. Subsequently, a joint analysis of variance was performed in which the following statistical model was adopted:

$$y_{ijk} = \mu + G_i + B/E_{jk} + E_k + GE_{ik} + \varepsilon_{ijk},$$

in which, y_{ijk} is the observed value of ith genotype evaluated in k^{th} environment and in j^{th} block; μ is the general mean; G_i is the fixed effect of i^{th} genotype (i = 1, 2, 3, ..., 20); B/E_{jk} is the random effect of j^{th} block in k^{th} environment; E_k is random effect of k^{th} environment; GE_{ik} is the random effect of genotypes and environments interaction between the i^{th} genotype and the k^{th} environment and ε_{ijk} is the experimental error associated with y_{ijk}.

Detailed studies about the phenotypic adaptability of genotypes belong to the evaluated environments were realized by fuzzy clustering method, as detailed below.

2.1. Phenotypic adaptability by fuzzy clustering

The phenotypic adaptability by fuzzy clustering method is based on FCM approach (Bezdek, 1984) which is a technique that aims to group observations into a number of pre-determined clusters. The allocation of individuals in each cluster is given by minimizing the following objective function:

$$J(U, c_1, \dots, c_c, \lambda_1, \dots, \lambda_n) = \sum_{i=1}^{c} \left(\sum_{j}^{n} u_{ij}^{m} \left\| c_i - x_j \right\|^2 \right) + \sum_{j=1}^{n} \lambda_j \left(\sum_{i=1}^{c} u_{ij} - 1 \right) \quad (1),$$

in which, U is the membership matrix ($u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{id}}{d_{kj}} \right)^{2/(m-1)}}$) of j observations in i predefined clusters; c_i is the centroid of i^{th} cluster ($c_i = \frac{\sum_{j}^{n} u_{ij}^{m} x_j}{\sum_{j}^{n} u_{ij}^{m}}$); $m \in [1, \infty]$ is the fuzzy weighting exponent and λ_j are the Lagrange multipliers for the n constraints ($\sum_{i=1}^{c} u_{ij} = 1 , \ \forall j = 1, \dots, n$).

The objective function minimization (1) is performed by an iterative process, according to the following steps: i. initialization of U matrix with random values between zero and one, so that the restriction $\sum_{i=1}^{c} u_{ij} = 1$ is respected. ii. obtaining the centroids of clusters defined a priori; iii. obtaining the objective function according to equation 1; iv. initialization of the iterative process and definition of the stop criterion, which is determined between two consecutive iterations and obtaining the new matrix U.

In order to apply the FCM algorithm in studies of phenotype adaptability pattern identification, 1000 ideotypes were added to the original data set of the average values of the genotypes in each environment. These genotypes were partitioned in four classes (wide adaptability, adaptability to favorable environments, adaptability to unfavorable environments and unadapted) according Rocha et al. (2005). That is, 250 ideotypes are

denoted for each class. The ideotypes for wide adaptability (cluster I) were defined by the maximum values in all environments evaluated, while unadapted ideotypes presented minimum values in all environments. The ideotypes for adaptability to favorable environments presented maximum values in favorable enviroments and minimum values in unfavorable environments. The ideotypes of adaptability to unfavorable environments presented opposite values.

The mean values of genotypes and ideotypes in each environment, that is, a matrix of dimensions $M_{(1000+20)\times 12}$, were adopted to verify the phenotypic adaptability pattern by the fuzzy clustering method. The influence of fuzzy weighting exponent value (m) on the membership of each genotype in clusters of phenotypic adaptability was also evaluated. The values considered for this exponent were 1.5 and 2.0.

The classification of each genotype was performed considering the greater membership observed among the four clusters established a priori. After identifying in which cluster each genotype presented greater pertinence, this result was compared with the classification of the already known ideotypes, which allowed identifying the patterns of phenotypic adaptability. The results obtained by this method were compared with the obtained by centroid method (Rocha et al., 2005).

The variance analyzes and the centroid method were performed in the GENES software (Cruz, 2016). The phenotypic adaptability by fuzzy clustering method was developed with the aid of Matlab software's fuzzy logic toolbox and implemented in BioFuzzy software, available at https://github.com/VQCarneiro/BioFuzzy.

3. RESULTS

There was a significant effect of the genotypes in all experiments ($p < 0.01$), except for the experiment in Ponte Nova during the dry season of 2003 (Table 2). The coefficients of variation were lower than 20%, indicating high experimental accuracy and reliability of results. The overall mean was 2389 $kg.ha^{-1}$, however, it was observed a variation between 1436 $kg.ha^{-1}$ and 3665 $kg.ha^{-1}$ for the experiments in Viçosa and Coimbra during the fall-winter season of 2003, respectively.

Table 2. Grain yield (kg.ha^{-1}) of the black common bean genotypes (GEN) evaluated in the fall-winter/2002, dry/2003 and fall-winter/2003 seasons in the municipalities of Viçosa (VI), Coimbra (CB), Ponte Nova (PN), Leopoldina (LP), Florestal (FT) and Capinópolis (CP) in the Minas Gerais State.

GEN	Fall - Winter Season/2002				Dry Season/2003				Fall - Winter Season/ 2003			
	VI	CB	PN	LP	VI	CB	PN	FT[5]	VI	CB	PN	CP
VP1	1863 c*	3061 a	2588 d	2171 a	3475 a	1656 f	2701	712 f	1230 d	3891 b	2850 d	2858 d
VP2	1483 h	2250 b	2485 d	1982 b	3005 b	1774 f	3081	1551 c	1623 c	3322 c	2585 h	2508 g
VP3	2133 b	2955 a	2878 b	2043 b	2860 b	2065 d	2704	1052 e	2215 a	4047 b	2760 e	2480 g
VP4	1839 c	2782 a	2429 e	1995 b	2689 b	1897 e	2203	1571 c	1612 c	3367 c	2695 f	2464 g
VP5	1741 d	2411 b	2790 b	2209 a	3326 a	2017 e	2828	2479 a	1495 c	3403 c	2902 c	3058 b
VP6	1782 d	2548 b	2836 b	2009 b	3007 b	1364 g	2837	1632 c	1410 d	3669 c	2817 e	2436 h
VP7	1491 h	2176 b	2111 g	1811 b	2774 b	2516 b	2886	2476 a	1565 c	3531 c	2496 i	2564 f
VP8	1544 g	2370 b	2668 c	2285 a	3164 a	2099 d	2853	1566 c	1346 d	3516 c	2581 h	2619 e
VP9	1416 i	2186 b	2431 e	1795 b	3364 a	2705 a	2912	1969 b	1027 e	3494 c	2065 m	2480 g
VP10	1481 h	2243 b	2390 e	1796 b	2746 b	2251 c	2572	1412 d	1368 d	3548 c	2358 k	2181 j
VP11	1475 h	2365 b	2375 e	1943 b	2983 b	2471 b	2577	1719 c	1343 d	3879 b	2345 k	2664 e
VP12	1424 i	2302 b	2210 f	1928 b	3279 a	2475 b	2452	1359 d	1247 d	3596 c	2242 l	2430 h
VP13	1635 f	2376 b	2349 e	2161 a	3098 a	1974 e	2798	1469 d	1353 d	3669 c	2436 j	2597 f
Vi 5700	1861 c	3110 a	2563 d	2375 a	3488 a	2586 a	2753	1689 c	1626 c	3856 b	3000 b	2969 c
Vi 5500	1687 e	3044 a	2820 b	2421 a	2977 b	2110 d	3307	1992 b	1843 b	3902 b	3050 a	2358 i
Vi 7800	1899 c	2735 a	2403 e	1892 b	2535 b	2182 c	2724	1452 d	1679 c	3445 c	2641 g	2419 h
CNFP 9346	1308 j	2288 b	2214 f	1456 b	3261 a	2177 c	2963	2051 b	1253 d	3478 c	2486 i	2447 g
CNFP 7988	1087 k	2311 b	2248 f	1710 b	3307 a	1972 e	2405	639 f	1026 e	3097 c	2085 m	2541 f
Ouro Negro	2260 a	3006 a	3230 a	2096 a	3110 a	2217 c	3094	1659 c	1937 b	4405 a	2804 e	3285 a
Valente	1531 g	3049 a	2376 e	2009 b	3504 a	2389 b	2702	1686 c	530 f	4179 a	2785 e	2680 e
Means	1647	2578	2520	2004	3098	2145	2768	1607	1436	3665	2599	2602
Ej †	-742	189	131	-385	709	-244	379	-782	-953	1276	210	213
CV (%)	12	8	12	10	10	14	14	17	20	6	11	11
p-value	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	>0.05	<0.01	<0.01	<0.01	<0.01	<0.01

† E_j – Environmental index; * Values followed by a same letter in column belong to a same group, Scott and Knott test at significant at level of 5% probability.

The experiments carried out in the fall-winter season of 2002 in Viçosa and Leopoldina presented averages lower than the overall mean (2389 kg.ha^{-1}). Therefore, these environments were considered representative of unfavorable environments, due to its negative environmental indexes (Table 2). The experiments installed in Coimbra and Florestal during the dry season of 2003 and in Viçosa during the fall-winter season of 2003 were also classified as unfavorable environments. The others were considered representative of favorable environments since they presented averages higher than the overall average and positive environmental indexes (Table 2).

The cultivars Ouro Negro and Valente were allocated, by Scott Knott test (Scott & Knott, 1974), as belonging to the group of the highest-yielding genotypes for 58% and 30% of the environments, respectively (Table 2). While Ouro Negro stood out in both favorable and unfavorable environments, Valente cultivar excelled in favorable environments. The higher averages (4405 kg.ha^{-1} e 4179 kg.ha^{-1}) were observed in Coimbra for these cultivars during the fall-winter season of 2003, where the highest environmental average was observed. The lines Vi 5700 e Vi 5500 presented averages as high as the highest-yielding genotypes at 30% and 25% of the experiments, respectively. These lines also stood out in both classes of specific environments. The line CNFP 9346, although allocated as one of the highest-yielding in the experiment during the dry season of 2003 in Viçosa, presented low averages for the other regions.

The joint analysis of the experiments showed that the effects of genotypes and environments were significant ($p < 0.05$) (Table 3), as described by Oliveira et al. (2006). In addition, there was a significant interaction between genotypes and environments. The coefficient of variation was 11.64%, that is, low value for this characteristic.

Table 3. Summary of joint variance analysis of black common bean genotypes evaluated for grain yield (kg.ha^{-1}) in 12 environments in the Minas Gerais State.

Sources of Variation	DF	Mean Square
Block/E	24	427258.22
Environments (E)	11	25596941.77*
Genotypes (G)	19	1019567.40*
GE interaction	209	230746.91**
Error	456	77373.15
Mean	2389.49	
CV (%)	11.64	

**, * Significant according to a F-test at the 0.01 and 0.05 probability level, respectively; DF: degrees of freedom; CV: coefficient of variation

Using the centroid method, it was verified that the lines VP3 (29,1%), VP5 (31,9%), Vi 5700 (36,5%), Vi 5500 (35,3%) and the cultivar Ouro Negro (44,6%) presented wide adaptability (Table 4; figure 1). Seven lines were classified as adapted to unfavorable environments (Table 4). The line VP7 stood out among the others in this cluster with a probability of 36.7%. The line VP1 and the Valente cultivar presented phenotypic adaptability to favorable environments with probabilities equals to 33.8% and 29.7%, respectively.

Table 4. Memberships, spatial probabilities and phenotypic adaptability classification of black common bean genotypes according the centroid method (C_0) and FCM (C_1) with m values equals to 1.5 or 2.0.

Genotypes	Memberships (%)												Class†
	I			II			III			IV			
	C_0	C_1		C_0	C_1		C_0	C_1		C_0	C_1		
		1.5	2.0		1.5	2.0		1.5	2.0		1.5	2.0	
VP1	22.5	12.4	19.3	**33.8**	**63.0**	**43.5**	19.1	6.5	13.9	24.7	18.1	23.3	II
VP2	22.4	15.1	19.8	22.8	16.1	20.6	27.0	32.5	28.9	**27.8**	**36.2**	**30.6**	IV
VP3	**29.1**	**43.0**	**33.6**	23.5	18.0	21.8	25.7	26.1	26.1	21.6	12.9	18.5	I
VP4	21.6	11.6	18.1	20.2	8.7	15.7	**31.1**	**50.4**	**37.5**	27.2	29.3	28.7	III
VP5	**31.9**	**52.8**	**39.1**	20.9	9.5	16.7	27.7	30.2	29.5	19.6	7.4	14.7	I
VP6	23.8	20.3	22.6	26.2	29.4	27.3	23.8	20.5	22.7	**26.2**	**29.9**	**27.4**	IV
VP7	23.9	12.9	21.0	17.8	3.9	11.7	**36.7**	**74.4**	**50.1**	21.6	8.7	17.3	III
VP8	24.5	22.6	23.9	23.4	18.6	21.8	**26.8**	**32.6**	**28.6**	25.4	26.2	25.7	III
VP9	23.3	17.1	21.4	21.4	12.1	18.0	**29.4**	**44.4**	**34.2**	25.9	26.3	26.4	III
VP10	19.2	6.6	14.0	20.1	7.9	15.4	28.7	33.7	31.5	**32.0**	**51.8**	**39.1**	IV
VP11	23.8	18.8	22.3	21.5	12.4	18.2	**29.3**	**44.2**	**34.0**	25.4	24.5	25.5	III
VP12	20.6	9.7	16.4	21.4	11.2	17.7	27.9	33.7	30.4	**30.1**	**45.4**	**35.4**	IV
VP13	22.6	15.8	20.2	23.0	16.8	20.9	26.8	32.0	28.6	**27.6**	**35.5**	**30.2**	IV
Vi 5700	**36.5**	**73.9**	**49.7**	21.9	9.3	17.8	23.6	12.7	20.7	17.9	4.1	11.8	I
Vi 5500	**35.3**	**69.3**	**47.0**	21.7	9.5	17.6	24.7	16.3	22.8	18.3	4.9	12.6	I
Vi 7800	22.8	14.7	20.2	20.4	9.3	16.2	**31.1**	**52.0**	**37.8**	25.7	24.0	25.8	III
CNFP 9346	22.6	15.7	20.2	22.6	15.5	20.2	**27.4**	**34.6**	**29.8**	27.4	34.3	29.8	III
CNFP 7988	15.1	1.2	7.5	21.5	4.8	15.2	19.1	3.0	12.0	**44.4**	**91.0**	**65.4**	IV
Ouro Negro	**44.6**	**91.2**	**65.7**	21.6	4.8	15.3	18.8	2.8	11.7	15.0	1.1	7.4	I
Valente	25.3	23.9	25.1	**29.7**	**45.2**	**34.7**	21.3	12.1	17.9	23.8	18.7	22.3	II

† Phenotypic adaptability classification of genotypes by centroid method (C_0) and FCM (C_1); I – Wide adaptability; II – Adaptability to favorable environments; III – Adaptability to unfavorable environments; IV – Unadapted.

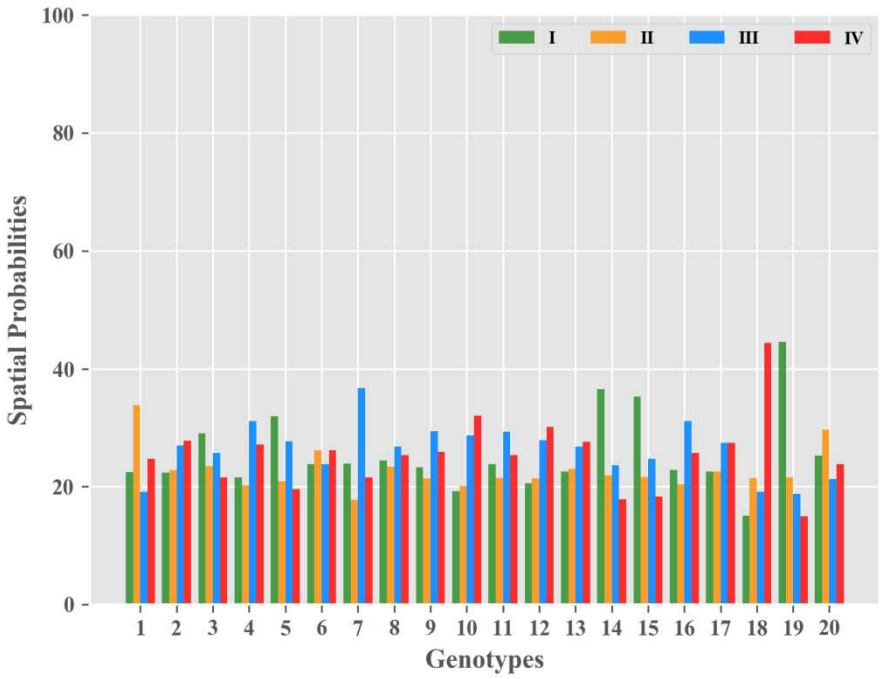


Figure 1. Spatial probabilities of genotypes obtained in phenotypic adaptability classes (I – Wide adaptability; II – Adaptability to favorable environments; III – Adaptability to unfavorable environments; IV – Unadapted genotypes) by centroid method.

All genotypes showed probabilities, by centroid method (Rocha et al., 2005), ranging between 15.0% and 44.6% for all clusters of phenotypic adaptability. This variation was observed for the cultivar Ouro Negro, in which 44.6% was the probability in cluster I and 15% in cluster IV. However, most of the genotypes presented very close probabilities between two clusters, especially VP 6 (26.2%) and CNFP 9346 (27.4%). The difference between the probabilities of CNFP 9346 in clusters III and IV was less than 0.1% (Table 4). Similar difference was also observed when considering the probabilities of line VP 6 in clusters II and IV.

The phenotypic adaptability pattern of black common bean genotypes regardless of the m value adopted in the FCM methodology, presented 100% similarity to the centroid method (Table 4). The cultivar Ouro Negro presented better performance in cluster I with membership equals 91.2% and 65.7% when adopting m values of 1.5 and 2, respectively (Table 4 and Figure 2). Considering these m values, the line VP 1 was the

one that most stood out in cluster II with membership equals 63% and 43.5%. On the other hand, the line VP 7 was highlighted in cluster III, with membership equals 74.4% and 50.1% considering these same m values. The lines VP 8, VP 9, VP 11, CNFP 9346 and cultivar Valente did not present a predominant phenotypic adaptability pattern when adopting m equal 1.5 since they did not have greater membership in a cluster of more than 50%. For these lines, the second largest membership occurred in cluster IV. However, the second largest membership of cultivar Valente occurred in cluster I.

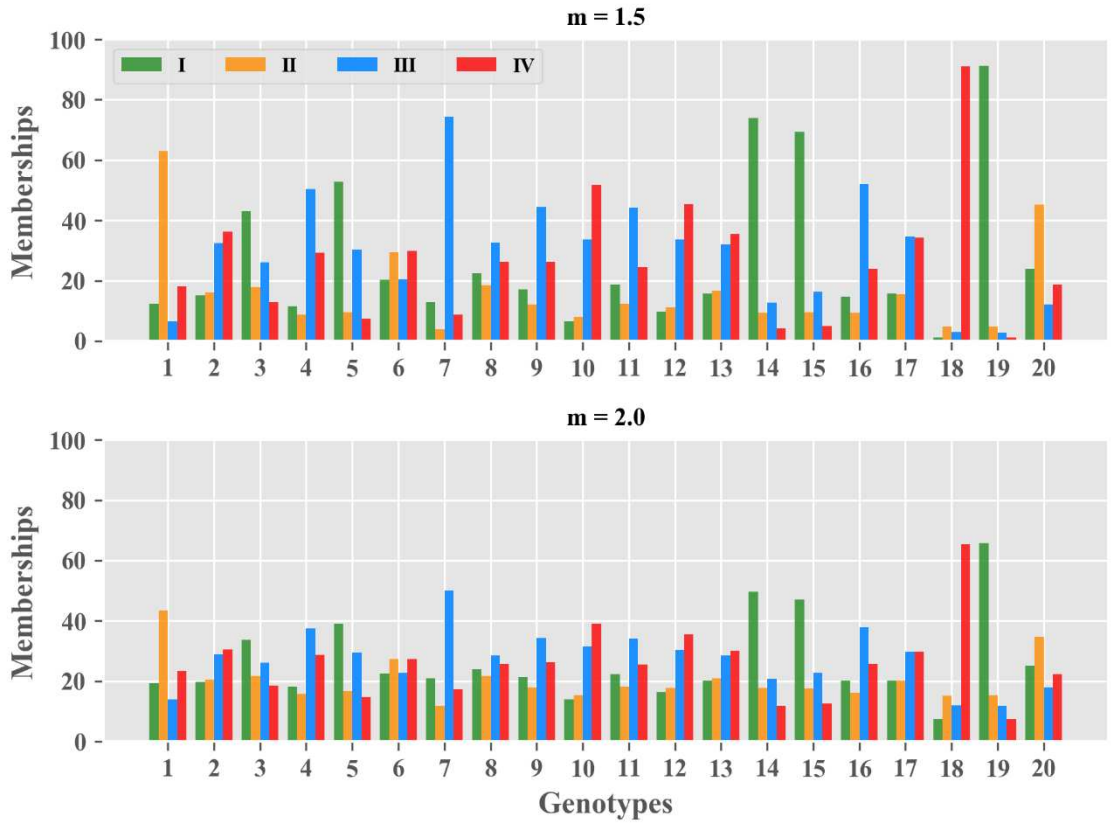


Figure 2. Memberships of genotypes in phenotypic adaptability classes (I – Wide adaptability; II – Adaptability to favorable environments; III – Adaptability to unfavorable environments; IV – Unadapted genotypes) obtained by fuzzy clustering method with m values equals to 1.5 and 2.0 .

The memberships of the genotypes in the predominant clusters were higher for the fuzzy clustering method when compared to the centroid method (Figures 1 and 2). It was also observed that, for m equal to 1.5, the memberships values were higher than the

memberships considering m equal to 2 (Figure 2). These facts could be verified by the response of the cultivar Ouro Negro, which presented the greatest discrepancy between the memberships in clusters I and IV. Adopting m values equal to 1.5 and 2, the membership ranged from 1.1% to 91.2% and from 7.4% to 65.7%, respectively. In addition, adopting m equal to 1.5 provides a better discrimination of genotypes within the same group.

Adopting the centroid method, the ideotype of cluster I was characterized by having, in all environments, maximum values of grain yield ranging from 2215 $kg.ha^{-1}$ to 4405 $kg.ha^{-1}$. On the other hand, the ideotype of cluster IV presented minimum values for this characteristic ranging between 530 $kg.ha^{-1}$ and 3097 $kg.ha^{-1}$. In the FCM method, the centroid values of cluster I range from 2207 kg ha^{-1} to 4398 kg. ha^{-1} and from 2211 $kg.ha^{-1}$ to 4401 $kg.ha^{-1}$ when using m equals to 1.5 and 2.0, respectively (Table 5). For these values of m, in the centroid of cluster IV there was variation from 540 $kg.ha^{-1}$ to 3102 kg ha^{-1} and from 535 $kg.ha^{-1}$ to 3099 $kg.ha^{-1}$.

Table 5. Centroids (CE) of phenotypic adaptability clusters (I – Wide adaptability; II – Adaptability to favorable environments; III – Adaptability to unfavorable environments; IV – Unadapted) obtained by centroid method and FCM.

Centroid												
CE	Fall-Winter / 2002				Dry Season / 2003				Fall-Winter / 2003			
	VI[1]	CB[2]	PN[3]	LP[4]	VI	CB	PN	FT[5]	VI	CB	PN	CP[6]
I	2260	3110	3230	2421	3504	2705	3307	2479	2215	4405	3050	3285
II	1087	3110	3230	1456	3504	1364	3307	639	530	4405	3050	3285
III	2260	2176	2111	2421	2535	2705	2203	2479	2215	3097	2065	2181
IV	1087	2176	2111	1456	2535	1364	2203	639	530	3097	2065	2181
FCM (m = 1,5)												
I	2254	3106	3224	2417	3499	2698	3302	2469	2207	4398	3047	3278
II	1091	3108	3226	1460	3502	1368	3303	644	535	4401	3048	3281
III	2251	2180	2115	2414	2541	2698	2211	2469	2204	3104	2072	2186
IV	1092	2179	2114	1461	2542	1373	2209	648	540	3102	2069	2185
FCM (m = 2,0)												
I	2257	3108	3227	2419	3501	2701	3304	2474	2211	4401	3048	3282
II	1089	3109	3227	1458	3503	1366	3305	642	533	4403	3049	3283
III	2256	2178	2113	2418	2538	2702	2207	2474	2210	3100	2068	2183
IV	1089	2178	2113	1459	2539	1369	2206	643	535	3099	2067	2183

[1] VI – Viçosa; [2] CB – Coimbra; [3] PN – Ponte Nova; [4] LP – Leopoldina;[5] FT – Florestal;
[6] CP – Capinópolis

4. DISCUSSION

The black common bean genotypes presented a differential response to environmental variations, as emphasized by the response of the cultivar Valente, which presented the highest average in Viçosa during the dry season of 2003 and the lowest mean considering same site and year during the fall-winter season. This result corroborates with Oliveira et al. (2006), which affirms that the season is the condition that contributes most to the genotypes and environments interaction. Therefore, detailed studies using the phenotypic adaptability method by fuzzy clustering and centroid method (Rocha et al., 2005) were performed in order to identify the phenotypic adaptability pattern of the genotypes evaluated.

The phenotypic adaptability method by fuzzy clustering presented similarity of 100% in the phenotypic adaptability pattern when compared to the centroid method. However, the fuzzy approach provided higher discrimination power of the genotypes, since the membership obtained in the four clusters demonstrated a higher discrepancy in relation to those obtained by the centroid method. Studies have reported that the spatial probabilities obtained by the centroid method are very close to each other, which makes it difficult to recommend cultivars (Amorim et al., 2011; Batista et al., 2015). Therefore, it was verified that this proposed method allowed identifying the phenotypic adaptability pattern of the black common bean genotypes with high discrimination power, consequently, it helps in the cultivars recommendation in breeding programs.

The concept of adaptability based on ideotypes given by the fuzzy clustering method is similar to that proposed by Rocha et al. (2005). However, these methodologies differ in some particularities related to the information processing of genotypes. While the centroid method adopts only four ideotypes that represent the phenotypic adaptability pattern (Rocha et al., 2005), the fuzzy clustering method adopts 250 ideotypes for each of these patterns. It is highlighted in this work that the use of the 1000 ideotypes was effective in identifying of the phenotypic adaptability of the 20 genotypes.

The FCM, despite presenting a similar classification to the centroid method, provided more discrepant membership values in the clusters. This high discrepancy made it possible to verify that there were genotypes classified as recommendable (wide or specific adaptability) that still have high membership values in the cluster of the unadapted genotypes. These genotypes presented a predominant membership below 50% in the recommended clusters (wide or specific adaptability) and the second biggest pertinence in the cluster of the unadapted genotypes. Therefore, we recommend that lines such as VP 8, VP 9, VP 11, CNFP 9346 be considered as unadapted since they presented

predominant membership in the cluster of genotypes adapted to unfavorable environments below 50% and the second higher membership in the cluster of unadapted genotypes. According to Scott-Knott's grouping pattern, in at least 50% of the evaluated environments, these lines were allocated in groups with low means (Table 2). That is, to classify these adaptability lines as to unfavorable environments as suggested by the centroid method constitutes a risk for the cultivars recommendation.

The value of m equal to 1,5 provided a greater discrimination of the genotypes by their respective membership. In these conditions, the best discrimination of the Vi 5700, Vi 5500 and Ouro Negro cultivar was observed in the wide adaptability cluster, as the membership of these genotypes in this cluster were equal to 73.9%, 69.3% and 91.2% while when adopting m equal to 2, the membership was 49.7%, 47% and 65.7% respectively (Table 5). The literature reports that the increase in the m value provides a greater fuzzy effect in the grouping by the FCM approach (Pimentel and Souza, 2013).

Despite the great potential, fuzzy logic has not been sufficiently explored by breeding programs. This approach has been used as a principle of grouping techniques such as FCM (Rundo et al., 2017) and also as fuzzy decision-making systems (Mardani et al., 2015, Carneiro et al., 2018), which is not yet explored in plant breeding. The FCM method allows to observe the groupings and better discriminates them through memberships (Pimentel and Souza, 2013), which is not possible in other fuzzy techniques such as k-Means and self-organizing map. Thus, it was verified with this work that the fuzzy logic, especially the phenotypic adaptability method by fuzzy clustering, presents high potential to be adopted by breeding programs in genotype selection studies and especially in the recommendation of cultivars.

5. CONCLUSIONS

The phenotypic adaptability method by fuzzy clustering was effective to identify the adaptability pattern of black common bean genotypes. Moreover, the discriminative power of this method obtained by membership in each cluster was higher than the observed in results of the centroid method.

6. REFERENCES

Amorin, B.S., G.I. Souza, M.A. Silveira, I.R. Nascimento, and T.A. Ferreira. 2011. Phenotypic adaptability of sweet potato genotypes of botanical seeds in the region South of the State of Tocantins. (In Portuguese) Appl. Res. & Agrotec. 4: 31–50. doi: 10.5777/PAeT.V4.N3.02.

Arunaghalam, V. 1981. Genetic Distance in Plant Breeding. Indian J. Genet. 41: 226-236.

Barroso, L.M.A., M. Nascimento, A.C.C. Nascimento, F.F. Silva, and R.P. Ferreira. 2013. Use of the Eberhart and Russell method as a priori information for the application of Artificial Neural Networks and discriminant analysis aiming at the classification of alfalfa genotypes for adaptability and stability. (In Portuguese) Rev. Bras. Biom. 31: 176–188.

Batista, R.O., R.L. Hamawaki, L.B. Sousa, A.P.O. Nogueira, and O.T. Hamawaki. 2015. Adaptability and stability of soybean genotypes in off-season cultivation. Genet. Mol. Res. 14: 9633–9645. doi: 10.4238/2015.

Bezdek, J.C., R. Ehrlich, and W. Full. 1984. FCM: The fuzzy c-means clustering algorithm. Comput. Geosci. 10: 191–203. doi: 10.1016/0098-3004(84)90020-7.

Carneiro, V.Q., G.N. Silva, C.D. Cruz, P.C.S. Carneiro, M. Nascimento, and J.E.S. Carneiro. 2017. Artificial neural networks as auxiliary tools for the improvement of bean plant architecture. Genet. Mol. Res. 16: 2. doi: 10.4238/gmr16029500.

Carneiro, V.Q., A.L. Prado, C.D. Cruz, P.C.S. Carneiro, M. Nascimento, and J.E.S. Carneiro. 2018. Fuzzy control systems for decision-making in cultivars recommendation. Acta Sci. Agron. 40: 1–8. doi: 10.4025/actasciagron.v40i1.39314.

Cruz, C.D. 2016. Genes Software – extended and integrated with the R , Matlab and Selegen. Acta Sci. Agron. 38: 547–552. doi: 10.4025/actasciagron.v38i4.32629.

Cruz, C.D., and F.L. Castoldi. 1991. Performance of the genotypes x environments interaction in simple and complex parts. (In Portuguese, with English abstract) Rev. Ceres. 38: 422-430.

Cruz, C.D., A.J. Regazzi, and P.C.S. Carneiro. 2012. Modelos biométricos aplicados ao melhoramento genético. 4th ed. Editora UFV, Viçosa.

Dogruparmak, S.C., G.A. Keskin, S. Yaman, and A. Alkan. 2014. Using principal component analysis and fuzzy c–means clustering for the assessment of air quality monitoring. Atmos. Pollut. Res. 5: 656–663. doi: 10.5094/APR.2014.075.

Eberhart, S.A., and W.A. Russell. 1966. Stability Parameters for Comparing Varieties. Crop Sci. 6: 36. doi: 10.2135/cropsci1966.0011183X000600010011x.

Finlay, K., and G. Wilkinson. 1963. The analysis of adaptation in a plant-breeding programme. Aust. J. Agric. Res. 14: 742-754. doi: 10.1071/AR9630742.

Gauch, H.G., and R.W. Zobel. 1966. AMMI analysis of yield trials. In: M.S. Kang and H.G. Gauch, editors, Genotype by environment interaction. CRC Press, Boca Raton. p. 85–122.

Glória, L.S., C.D. Cruz, R.A.M. Vieira, M.D.V. Resende, P.S. Lopes, et al. 2016. Accessing marker effects and heritability estimates from genome prediction by Bayesian regularized neural networks. Livest. Sci. 191: 91–96. doi: 10.1016/j.livsci.2016.07.015.

Jang, J.S.R., C.T. Sun, and E. Mizutani. 2012. Neuro-Fuzzy and Soft Computing - A computacional Approach to Learning and Machine Intelligence.PHI Learning Private Limited, New Delhi.

Kohonen, T. 1982. Self-organized formation of topologically correct feature maps. Biol. Cybern. 43: 55-69. doi: 10.1007/BF00337288.

Lin, C.S., and M.R. Binns. 1988. A superiority measure of cultivar performance for cultivar × location data. Can. J. Plant Sci. 68: 193–198. doi: 10.4141/cjps88-018.

Ma, C., H.H. Zhang, and X. Wang. 2014. Machine learning for Big Data analytics in plants. Trends Plant. Sci. 19: 798–808. doi: 10.1016/j.tplants.2014.08.004.

Macqueen, J. 1967. Some methods for classification and analysis of multivariate observations. Proc. Fifth Berkeley Symp. on Math. Statist. and Prob. 1: 281–297. doi: citeulike-article-id:6083430.

Malosetti, M., J.M. Ribaut, and F.A. van Eeuwijk. 2013. The statistical analysis of multi-environment data: Modeling genotype-by-environment interaction and its genetic basis. Front. Physiol. 4: 1–17. doi: 10.3389/fphys.2013.00044.

Mardani, A., A. Jusoh, and E.K. Zavadskas. 2015. Fuzzy multiple criteria decision-making techniques and applications - Two decades review from 1994 to 2014. Expert Syst. Appl. 42: 4126–4148. doi: 10.1016/j.eswa.2015.01.003.

Murakami, D.M., and C.D. Cruz. 2004. Proposal of methodologies for environment stratification and analysis of genotype adaptability. Crop Breed. Appl. Biot. 4: 7–11.

Nascimento, M., C.D. Cruz, A.C.M. Campana, R.S. Tomaz, C.C. Salgado, et al. 2009. Alteration of the centroid method to evaluate genotypic adaptability. (In Portuguese, with English abstract) Pesq. Agropec. Bras. 44: 263–269. doi: 10.1590/S0100-204X2009000300007.

Nascimento, M., L.A. Peternelli, C.D. Cruz, A.C.M. Campana, R.P. Ferreira, et al. 2013. Artificial neural networks for adaptability and stability evaluation in alfalfa genotypes. Crop Breed. Appl. Biotech. 13: 152–156.

Nascimento, M., A. Ferreira, A.C.C. Nascimento, F.F. Silva, R.P. Ferreira, et al. 2015. Multiple centroid method to evaluate the adaptability of alfalfa genotypes. Rev. Ceres. 62: 30–36. doi: 10.1590/0034-737X201562010004.

Oliveira, G.V., P.C.S. Carneiro, J.E.S. Carneiro, and C.D. Cruz. 2006. Adaptability and stability of common bean in Minas Gerais State, Brazil. (In Portuguese, with English abstract) Pesq. Agropec. Bras. 41: 257–265. doi: 10.1016/j.vetimm.2006.05.008.

Pimentel, B.A., and R.M.C.R. Souza. 2013. A multivariate fuzzy c-means method. Appl. Soft Comput. 13: 1592–1607. doi: 10.1016/j.asoc.2012.12.024.

Resende, M.D.V. 2004. Métodos Estatísticos Ótimos na Análise de Experimentos de Campo. Doc. 100. Embrapa Florestas, Colombo.

Robertson, A. 1959. The Sampling Variance of the Genetic Correlation Coefficient. Biometrics. 15: 469–485. doi: 10.2307/2527750.

Rocha, R.B., J.I. Muro-Abad, E.F. Araújo, and C.D. Cruz. 2005. evaluation of the centroid method for study of environment adaptability of clones of Eucalyptus grandis. (In Portuguese, with English abstract) Cienc. Florestal. 15: 255–266.

Rundo, L., C. Militello, G. Russo, A. Garufi, S. Vitabile, et al. 2017. Automated prostate gland segmentation based on an unsupervised fuzzy C-means clustering technique using multispectral T1w and T2w MR imaging. Information. 8: 1–28. doi: 10.3390/info8020049

Sant'Anna, I.C., R.S. Tomaz, G.N. Silva, M. Nascimento, L.L. Bhering, et al. 2015. Superiority of artificial neural networks for a genetic classification procedure. Genet. Mol. Res. 14: 9898–9906. doi: 10.4238/2015.August.19.24.

Scott, A.J., and M. Knott. 1974. A Cluster Analysis Method for Grouping Means in the Analysis of Variance. Biometrics. 30: 507–512. doi: 10.2307/2529204.

Schmidhuber, J. 2015. Deep learning in neural networks : An overview. Neural Networks. 61: 85–117. doi: 10.1016/j.neunet.2014.09.003.

Silva, G.N., R.S. Tomaz, Sant'Anna I.C., M. Nascimento, and L.L. Bhering. 2014. Neural networks for predicting breeding values and genetic gains. Sci. Agric. 71: 494–498. doi: 10.1590/0103-9016-2014-0057.

Singh, A., B. Ganapathysubramanian, A.K. Singh, and S. Sarkar. 2016. Machine Learning for High-Throughput Stress Phenotyping in Plants. Trends Plant. Sci. 21: 110–124. doi: 10.1016/j.tplants.2015.10.015

Valle, S., W. Li, and S.J. Qin. 1999. Selection of the Number of Principal Components: The Variance of the Reconstruction Error Criterion with a Comparison to Other Methods. Ind. Eng. Chem. Res. 38: 4389- 4401. doi: 10.1021/ie990110i.

Van Eeuwijk, F.A., D.V. Bustos-Korts, and M. Malosetti. 2016. What should students in plant breeding know about the statistical aspects of genotype × environment interactions?. Crop Sci. 56: 2119–2140. doi: 10.2135/cropsci2015.06.0375.

Vasconcelos, E.S., M.S. Reis, C.D. Cruz, T. Sediyama, and A.C. Scapim. 2011. Integrated method for adaptability and phenotypic stability analysis. Acta Sci. Agron. 33: 251–257. doi: 10.4025/actasciagron.v33i2.8272.

Yan, W., L.A. Hunt, Q. Sheng, and Z. Szlavnics. 2000. Cultivar evaluation and mega-environment investigation based on the GGE biplot. Crop Sci. 40: 597-605. doi: 10.2135/cropsci2000.403597x.

Teodoro, P.E., F.E. Torres, L.P. Ribeiro, L.M.A. Barroso, M. Nascimento, et al. 2015. Yield adaptability and stability of semi‐prostrate cowpea genotypes. (In Portuguese, with English abstract) Pesq. Agropec. Bras. 50: 1054–1060. doi: 10.1590/S0100-204X2015001100008.

4. CONSIDERAÇÕES FINAIS

Os softwares desenvolvidos e apresentados neste trabalho constituem importante contribuição para o melhoramento vegetal, pois constituem tecnologias gratuitas que incorporam modelagens baseadas em inteligência artificial, redes neurais artificiais, sistemas de tomada de decisão fuzzy e fenômica. Além da contribuição na área de pesquisa, esses softwares também são importantes como veículos para difundir conhecimentos de processamento digital de imagens pouco explorados no melhoramento vegetal.

O FENOM é um software com importantes aplicações no processamento digital de imagens e redes neurais artificiais que podem ser aplicados a avaliações fenotípicas de alta qualidade. Esse software é subdividido em duas áreas de procedimentos: processamento digital de imagens e classificação por meio de redes neurais artificiais. Para processamento de imagens estão disponíveis procedimentos de aquisição, pré-processamento, segmentação e extração de características. Nos procedimentos de classificação estão disponíveis análises por redes neurais artificiais com arquitetura perceptron multicamadas.

O BioFuzzy é um software com aplicações no processamento de dados experimentais aplicados ao melhoramento vegetal. Esse software disponibiliza procedimentos de sistemas de decisão fuzzy e de agrupamento fuzzy para auxiliar na recomendação de cultivares. Trata-se de importante contribuição à análise de dados, focados em lógica fuzzy, especialmente por não existirem softwares com procedimentos dessa natureza. Exemplo desses procedimentos foi apresentado por meio da proposição do método de adaptabilidade fenotípica por agrupamento fuzzy baseado no método fuzzy C-Means. Esse método apresentou elevada eficácia na identificação da adaptabilidade de

genótipos de feijão comum do grupo comercial preto, além de apresentar elevada capacidade discriminatória dos genótipos avaliados.