RAPHAEL MILLER DE SOUZA CALDAS

INTELIGÊNCIA ARTIFICIAL APLICADA AO ESTUDO DA DIVERSIDADE GENÉTICA DO BANCO DE GERMOPLASMA DE VIDEIRA DA EMBRAPA SEMIÁRIDO

RECIFE – PE
AGOSTO DE 2021

RAPHAEL MILLER DE SOUZA CALDAS

INTELIGÊNCIA ARTIFICIAL APLICADA AO ESTUDO DA DIVERSIDADE GENÉTICA DO BANCO DE GERMOPLASMA DE VIDEIRA DA EMBRAPA SEMIÁRIDO

Dissertação apresentada ao Programa de Pós graduação em Agronomia “Melhoramento Genético de Plantas”, da Universidade Federal Rural de Pernambuco, como parte dos requisitos para obtenção do grau de Mestre em Agronomia – Melhoramento Genético de Plantas.

ORIENTAÇÃO:
Profa. Dra. Rosimar dos Santos Musser (UFRPE-SEDE)
COORIENTAÇÃO:
Dra. Patricia Coelho de Souza Leão (Embrapa Semiárido)
Prof. Dr. André Câmara Alves do Nascimento (UFRPE-SEDE)

RECIFE – PE
AGOSTO DE 2021

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

C145i Caldas, Raphael Miller de Souza
INTELIGÊNCIA ARTIFICIAL APLICADA AO ESTUDO DA DIVERSIDADE GENÉTICA DO BANCO DE GERMOPLASMA DE VIDEIRA DA EMBRAPA SEMIÁRIDO / Raphael Miller de Souza Caldas. - 2021.
69 f. : il.

Orientadora: Rosimar dos Santos Musser.
Coorientador: Andre Camara Alves do Nascimento.
Inclui referências e apêndice(s).

Dissertação (Mestrado) - Universidade Federal Rural de Pernambuco, Programa de Pós-Graduação em Agronomia - Melhoramento Genético de Plantas, Recife, 2021.

1. Redes neurais artificiais. 2. Machine learning. 3. Vittis spp. 4. Rede SOM. 5. Emergent self-organizing maps. I. Musser, Rosimar dos Santos, orient. II. Nascimento, Andre Camara Alves do, coorient. III. Título

CDD 581.15

RAPHAEL MILLER DE SOUZA CALDAS

INTELIGÊNCIA ARTIFICIAL APLICADA AO ESTUDO DA DIVERSIDADE GENÉTICA DO BANCO DE GERMOPLASMA DE VIDEIRA DA EMBRAPA SEMIÁRIDO

Dissertação defendida e aprovada pela banca examinadora em:

Banca examinadora:

_____

Profª Dra. Rosimar dos Santos Musser

(DEPA – UFRPE)

_____

Prof. Dr. Ricardo Bastos Cavalcante Prudêncio

(CIN – UFPE)

_____

Profª Dra. Angélica Virgínia Valois Montarroyos

(DEPA – UFRPE)

RECIFE – PE

2021

OFEREÇO

Ao Senhor Jesus Cristo, que cuida de mim e está comigo todos os dias, em todos os momentos;

A minha esposa, Joyce, pela paciência, carinho, cuidado e amor;

Aos meus pais, que sempre me deram todo o amor que um filho precisa para ser feliz e sempre acreditaram em mim;

A toda família da Igreja Universal do Reino de Deus, que sempre me deu suporte espiritual para enfrentar as batalhas dessa vida.

“Assim diz o Senhor: Não se glorie o sábio na sua sabedoria, nem se glorie o forte na sua força, não se glorie o rico nas suas riquezas. Mas o que se gloriar, glorie-se nisto: em Me entender e Me conhecer.”

Jeremias 9 : 23-24

“Os fortes não têm medo de encarar o pior: os fracos fogem dele porque sua mera visão os esmaga.”

Olavo de Carvalho

AGRADECIMENTOS

Agradeço a Deus pela sua misericórdia e amor incondicionais. Pela força dada nos momentos de fraqueza e pela Sua intervenção divina nos momentos de maior necessidade.

Agradeço a minha esposa, Joyce, por ser uma parceira em todos os momentos e por acreditar sempre, não importando o momento ou a dificuldade, sendo um exemplo de fé para mim.

Agradeço aos meus pais, que sempre sacrificaram suas necessidades e até mesmo os seus sonhos para sonhar os meus. O amor mais puro desse mundo vem de vocês.

Agradeço aos meus orientadores por toda a paciência, auxílio, direcionamento e instrução. Sem a ajuda de vocês, esse trabalho não seria possível.

Agradeço ao trabalho espiritual incansável da Igreja Universal do Reino de Deus, que sempre me acolheu de portas abertas nos momentos que mais precisei de uma orientação de fé.

Agradeço ao bispo Edir Macedo, líder da Igreja Universal do Reino Deus, pelo seu esforço em levar a Palavra de Deus aos aflitos e necessitados nos quatro cantos desse mundo, apesar de ser diariamente vilipendiado pela mídia. Em momentos de aflição, suas mensagens me alcançaram, fortalecendo minha fé.

Agradeço a todas as instituições públicas que me proporcionaram a realização desse curso de mestrado.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE *BOXPLOTS*

SUMÁRIO

RESUMO

O objetivo deste trabalho foi aplicar a inteligência artificial, através do uso do algoritmo de agrupamento *Emergent self-organizing maps* (ESOM), ao estudo da diversidade genética de acessos de uvas de mesa do Banco Ativo de Germoplasma de videira da Embrapa Semiárido. Foram avaliadas as seguintes variáveis agronômicas quantitativas de 93 acessos de uva de mesa: produção (kg/planta), número de cachos por planta, peso do cacho (g), comprimento e largura dos cachos (cm), peso da baga (g), comprimento e diâmetro das bagas (mm), teor de sólidos solúveis (ºBrix), acidez titulável (g/100ml) e relação de sólidos solúveis/acidez titulável. O experimento foi realizado no Campo Experimental de Mandacaru, Juazeiro-BA, sendo analisados os dados referentes a quatro safras (2018.1, 2018.2, 2019.1 e 2019.2). O agrupamento realizado pelo algoritmo ESOM foi capaz de descobrir padrões genéticos e diferenças entre os acessos de uva de mesa estudados, permitindo a formação de 10 grupos heteróticos. O grupo 0 apresentou os maiores valores máximos para as variáveis produção (8,44 kg), peso do cacho (520,19 g), comprimento do cacho (22,30 cm), largura do cacho (15,10 cm), peso da baga (8,00 g), comprimento da baga (26,84 mm), diâmetro da baga (22,17 mm) e sólidos solúveis (23,90 °Brix). Também apresentou os maiores valores médios para o peso do cacho (331,20 g), comprimento do cacho (16,6 cm), largura do cacho (10,5 cm), peso da baga (5,78 g), comprimento da baga (23,5 mm) e diâmetro da baga (19,7 mm). A presença de variabilidade genética para as variáveis analisadas foi evidenciada pela formação dos mapas de variabilidade genética, demonstrando ampla base genética para os 93 acessos analisados. A matriz ESOM de similaridade genética indicou os cruzamentos mais promissores entre os 10 grupos heteróticos de uva de mesa com base na divergência genética. O cruzamento de genótipos do grupo 0, composto por genótipos, em sua maioria, pirênicos e com bagas de tamanho grande, com genótipos dos grupo 1, 2, 4, 5 e 7, que sejam apirênicos e de bagas de tamanho menor são os mais indicados. Cruzamentos entre genótipos dos grupos 0 e 9 não são indicados. Apesar de serem os grupos mais distantes geneticamente, todos os indivíduos do grupo 9 possuem um tamanho pequeno de baga e sementes, características consideradas indesejáveis em um programa de melhoramento genético de uvas de mesa. Os ESOM se mostraram promissores na análise da diversidade genética e consequente formação de grupos heteróticos, além de indicar quais cruzamentos mais promissores. Estudos futuros sobre a validação dos ESOM como um método de agrupamento eficiente no melhoramento genético de plantas são indicados.

Palavras-chave: *Vitis* spp., agrupamento, variabilidade genética, grupos heteróticos, algoritmos inteligentes, *emergent self-organizing maps*.

ABSTRACT

The objective of this work was to apply artificial intelligence, through the use of the ESOM clustering algorithm, to the study of the genetic diversity of genotypes of table grapes from the Embrapa Semiárido Vine Germplasm Active Bank. The following quantitative agronomic variables of 93 table grape genotypes were evaluated: yield (kg/plant), number of clusters per plant, cluster weight (g), cluster length and width (cm), berry weight (g), length and diameter of berries (mm), soluble solids content (ºBrix), titratable acidity (g/100ml) and soluble solids/titratable acidity ratio. The experiment was carried out in the Experimental Field of Mandacaru, Juazeiro-BA, and data referring to four harvesting seasons (2018.1, 2018.2, 2019.1 and 2019.2) were analyzed. The grouping performed by the ESOM network was able to discover genetic patterns and differences between the studied table grape genotypes, allowing the formation of 10 heterotic groups. Group 0 had the highest maximum values for the variables yield (8.44 kg), cluster weight (520.19 g), cluster length (22.30 cm), cluster width (15.10 cm), weight of berry (8.00 g), berry length (26.84 mm), berry diameter (22.17 mm) and soluble solids (23.90 °Brix) and also presented the highest mean values for cluster weight (331.20 g), cluster length (16.6 cm), cluster width (10.5 cm), berry weight (5.78 g), berry length (23.5 mm) and berry diameter (19.7 mm). The presence of genetic variability for the analyzed variables was evidenced by the formation of genetic variability maps, demonstrating a broad genetic base for the 93 genotypes analyzed. The variables with less variability were cluster length and berry length. The ESOM matrix of genetic similarity indicated the most promising crosses among the 10 heterotic table grape groups based on genetic divergence. The crossing of genotypes from group 0, mostly composed of genotypes with seeds and large berries, with genotypes from group 1, 2, 4, 5 and 7, which are seedless and have smaller berries being the most indicated. Crosses between genotypes from groups 0 and 9 are not indicated. Despite being the most genetically distant groups, all individuals in group 9 have a small berry size and have seeds, characteristics considered undesirable in a table grape breeding program. ESOM proved to be promising in the analysis of genetic diversity and consequent formation of heterotic groups, in addition to indicating which crosses are more promising. Future studies on the validation of ESOM as an efficient grouping method for plant genetic improvement are indicated.

Keywords: *Vitis* spp., clustering, genetic variability, heterotic clusters, intelligent algorithms, *emergent* self-organizing maps.

CAPÍTULO I

CONSIDERAÇÕES GERAIS

1. INTRODUÇÃO

A videira (*Vitis* spp.) ocupa lugar de destaque entre as mais importantes espécies vegetais, sendo considerada a planta frutífera de domesticação mais antiga que se tem conhecimento (Radmann and Bianchi 2008). O sucesso dos programas de melhoramento genético da videira depende da diversidade de *Vitis* spp., que normalmente são conservadas em bancos ativos de germoplasma, no campo ou *in vitro*. Esses bancos de germoplasma devem ser continuamente avaliados e sua ampliação deve ser feita de forma criteriosa, evitando-se duplicatas e erros de identificação. Os estudos de diversidade genética são fundamentais para o manejo racional das coleções de germoplasma, como também para fornecer informações necessárias sobre os genótipos mais divergentes, orientando cruzamentos e aumentando a eficiência dos programas de melhoramento (Leão 2008).

A diversidade genética em bancos de germoplasma pode ser estimada com base em caracteres morfológicos qualitativos e/ou quantitativos através da utilização de técnicas multivariadas (Bertan et al. 2006) e moleculares (Poyraz 2016). Vários métodos multivariados podem ser aplicados no estudo da diversidade genética, como a análise de componentes principais, de variáveis canônicas e os métodos de agrupamentos, que podem ainda ser assistidos pela utilização de marcadores moleculares. As estratégias multivariadas são úteis na caracterização e classificação de caracteres de importância agronômica dos genótipos avaliados (Kumar et al. 2020). No entanto, um novo paradigma pode ser empregado no melhoramento genético para fins de seleção e estudos de diversidade genética que envolvam princípios de aprendizagem em uma abordagem de inteligência computacional: a inteligência artificial (Silva et al. 2014).

A inteligência artificial é um conjunto de tecnologias que permitem aos computadores perceber, aprender, raciocinar e auxiliar na tomada de decisões para resolução de problemas, semelhante ao que um ser humano faz (Smith and Shum, 2018). A inteligência artificial tem sido utilizada no contexto da agricultura de diversas formas, como por exemplo na previsão da produtividade das culturas, detecção de doenças, detecção de plantas daninhas, padrão de qualidade dos produtos agrícolas, reconhecimento de espécies vegetais, manejo e bem estar animal, manejo da água e

do solo, estudos meteorológicos e no melhoramento genético de plantas (Liakos et al. 2018, Harfouche et al. 2019, Salehnia et al. 2019).

Os métodos de análise de agrupamento tem inúmeras aplicações em várias áreas da ciência, como engenharia, ciências médicas, ciências da terra, economia e muitas outras (Robeva and Macauley 2018). A análise de agrupamentos é uma aplicação típica do modelo de aprendizado de máquina não supervisionado, normalmente usado para encontrar agrupamentos naturais de dados (*clusters*). Algumas técnicas de aprendizado de máquina não supervisionado são o aprendizado hierárquico, agrupamento de dados, modelos de variáveis latentes, redução da dimensionalidade e detecção de *outliers* (Liakos et al. 2018, Usama et al. 2019). Essas técnicas de aprendizado de máquina não supervisionado facilitam a análise de conjuntos de dados brutos, ajudando a gerar um entendimento intuitivo analítico a partir de dados não rotulados. Por exemplo, o agrupamento hierárquico é uma estratégia bem conhecida na mineração de dados e na análise estatística, na qual os dados são agrupados em uma hierarquia de *clusters* usando uma abordagem aglomerativa ou divisiva (Usama et al. 2019).

Os recentes avanços dos algoritmos de agrupamento, em particular, as técnicas baseadas em redes neurais artificiais, ajudaram a avançar significativamente o estado da arte em técnicas de aprendizado de máquina não supervisionado, facilitando o processamento de dados brutos sem exigir uma engenharia cuidadosa e conhecimento de domínio para a criação de recursos. Uma rede neural artificial é um processador paralelo distribuído massivamente, composto por unidades de processamento simples chamadas “neurônios”, uma função de ativação não linear, uma função de custo e um algoritmo de retropropagação. É uma técnica que modela a abstração de alto nível nos dados usando uma, algumas ou várias camadas de transformações lineares e não lineares. Com uma grande quantidade dessas camadas de transformação, uma máquina pode aprender automaticamente um modelo ou representação de dados bastante complexos (Haykin 2008, Deng et al. 2014, LeCun et al. 2015, Usama et al. 2019).

O objetivo deste trabalho é aplicar a inteligência artificial, através do uso do algoritmo de agrupamento ESOM , ao estudo da diversidade genética de acessos de uvas de mesa do Banco Ativo de Germoplasma (BAG) de videira da Embrapa Semiárido. Este o primeiro trabalho sobre o uso de inteligência artificial realizado no

BAG. Espera-se que este estudo apresente mais uma ferramenta a ser incorporada ao acervo do melhorista de plantas, auxiliando no manejo racional do banco de germoplasma de videira através do fornecimento de novas informações disponibilizadas pelo advento da inteligência artificial na formação de grupos heteróticos e indicação de cruzamentos mais promissores para os genótipos de uva de mesa disponíveis no BAG da Embrapa Semiárido.

2. REFERENCIAL TEÓRICO

2.1. Origem e domesticação, classificação botânica e características morfológicas da videira

Entre as culturas domesticadas, poucas são tão importantes quanto a videira (*Vitis vinifera* spp. *sativa*). As uvas são usadas como fonte de comida e vinho há séculos e têm um significado particular nos rituais e na religião. O Antigo Testamento, por exemplo, menciona as uvas em seu primeiro livro (Gênesis 9:20), detalhando o plantio de uma videira por Noé após o dilúvio (Zhou et al. 2019). Acredita-se que as espécies de videira tenham três centros de origem: a região entre o Mar Negro e o Mar Cáspio - o "Crescente Fértil", a América do Norte e a Ásia. As evidências mais antigas sugerem que o centro de origem mais provável da videira seja a região da Cordilheira do Cáucaso, onde hoje está localizado o país da Geórgia. Sua domesticação, associada à descoberta dos locais onde foram encontrados os primeiros vestígios da fabricação e produção de vinho, por volta de 6.000 anos a.C., ocorreu no Oriente Próximo.

Com base em evidências micro botânicas, foi descoberto que videiras cresciam em alguns sítios da Geórgia, possivelmente dentro das aldeias, e que seus frutos eram usados como fonte de alimento. Essas evidências, associadas às evidências químicas de um produto feito de uva dentro de vários frascos, que teriam servido como recipientes de líquidos, permitiu chegar a conclusão de que o vinho de uva era provavelmente um dos produtos fabricados (This et al. 2006, McGovern et al. 2017).

Um recente estudo realizado por Riaz et al. (2018), através do uso de microssatélites (SSR) para análise da diversidade genética em amostras de *Vitis vinifera* spp. *sylvestris*, também indica a Geórgia como principal centro de origem da videira e fonte de sua diversidade genética. A domesticação da videira possibilitou a

difusão da “Cultura do Vinho” por todo o Oriente Próximo e Egito, e, mais tarde, sua dispersão por todo o mundo (McGovern et al. 2017).

A videira pertence ao Reino Plantae, Subreino Tracheobionta, Superdivisão Spermatophyta, Divisão Magnoliophyta, Classe Magnoliopsida, Subclasse Rosidae, Ordem Rhamnales e Família Vitaceae (USDA 2020). Segundo Keller (2015), a família Vitaceae contém aproximadamente 1000 espécies designadas para 17 gêneros que são tipicamente arbustos ou cipós lenhosos que escalam por meio de suas gavinhas opostas às folhas - daí o nome Vitaceae (do latim *viere* – que siginifica ‘para anexar’).

Estudos mais recentes, como o de Lu et al. (2018), indicam algumas pequenas diferenças, atualizando que a família Vitaceae está organizada em 16 gêneros e contém cerca de 950 espécies. Wen et al. (2018) concluíram que muitos gêneros da família Vitaceae precisam ser revisados taxonomicamente através de novas ferramentas de bioinformática para uma correta classificação. As raízes das plantas da família Vitaceae são geralmente fibrosas e bem ramificadas, e podem crescer até vários metros de comprimento. As folhas são alternadas, exceto durante a fase juvenil de plantas cultivadas a partir de sementes, e podem ser simples ou compostas. Os frutos são geralmente bagas que possuem de uma a quatro sementes (Keller 2015).

Todas as uvas cultivadas pertencem ao gênero Muscadínia (2n = 40 cromossomos) ou ao gênero Vitis (2n = 38 cromossomos). Porém, alguns autores dividem o gênero Vitis em dois subgêneros: Vitis e Muscadínia, enquanto outros preferem manter os dois táxons como gêneros separados. As análises das diferenças genômicas entre Vitis e Muscadinia separam claramente os dois táxons, mas não na medida em que outros gêneros em Vitaceae divergem. Por causa dos diferentes números de cromossomos, cruzamentos entre esses dois gêneros raramente produzem híbridos férteis. As principais características morfológicas dos dois gêneros incluem: folhas simples, gavinhas simples ou bifurcadas, geralmente flores unissexuais, pétalas de flores fundidas que se separam na base, formando uma caliptra ou “tampa”, frutos em forma de bagas, macias e com polpa (Keller 2015, Walker et al. 2019).

Muitas espécies de Vitis são simpátricas a uma ou mais espécies e todas são totalmente férteis ao cruzarem, com exceção do táxon Muscadínia, que forma híbridos estéreis com as espécies de Vitis. Ao coletar uvas na natureza, pode ser difícil identificar uma espécie, dada a variação natural e o alto potencial para produção de

híbridos naturais, em locais onde duas ou mais espécies se sobrepõem. As espécies de Vitis, quando encontradas na natureza, são mantidas separadas por diferenças em suas preferências de habitat, barreiras geográficas e diferenças fenológicas nas datas de floração. Embora a maioria das espécies da família Vitaceae resida nos trópicos e subtrópicos, uma única espécie das zonas temperadas, a *Vitis vinifera* spp. *sativa*, se tornou a principal espécie frutífera em quase 90 países para a produção de vinho e suco, ou como uvas frescas de mesa ou uvas secas (passas) (Keller 2015, Walker et al. 2019).

2.2. Importância econômica da vitivinicultura

O Submédio do Vale do São Francisco destaca-se como uma das mais importantes regiões do Brasil no que diz respeito à fruticultura. De acordo com dados da Companhia de Desenvolvimento dos Vales do São Francisco e do Parnaíba (Codevasf), em 2019, o Submédio do Vale do São Francisco teve um valor bruto de produção de mais de 3 bilhões de reais (Figura 1). As culturas da uva, manga, cana-de-açúcar e goiaba representaram 88% desse valor, com destaque para a cultura da videira, que representou 41% do valor bruto total de produção, conforme Figura 2.

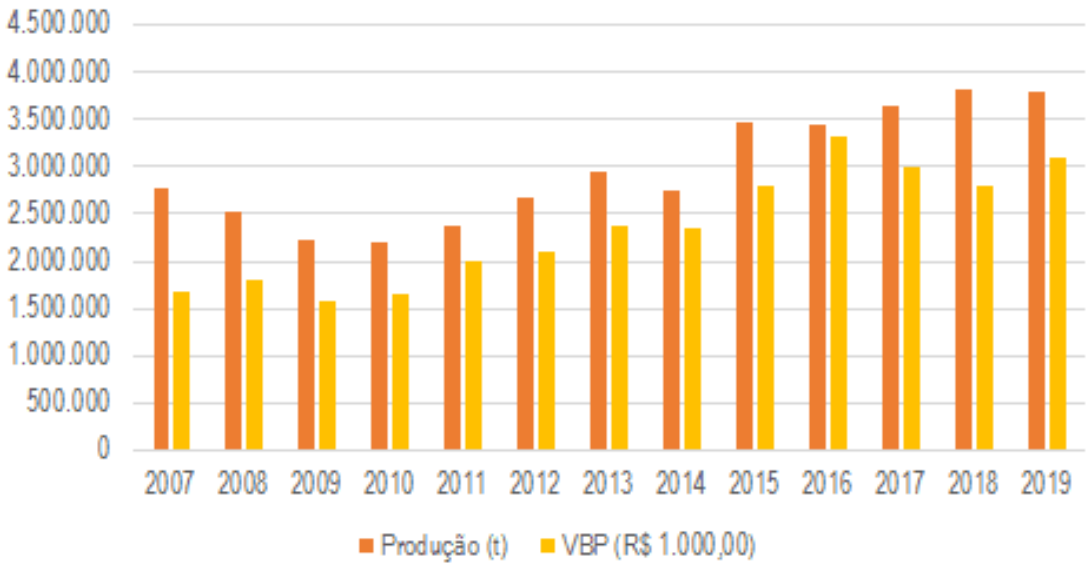


Figura 1. Evolução dos dados de produção dos projetos públicos de irrigação da Codevasf, 2007-2019. Fonte: Codevasf, 2020.

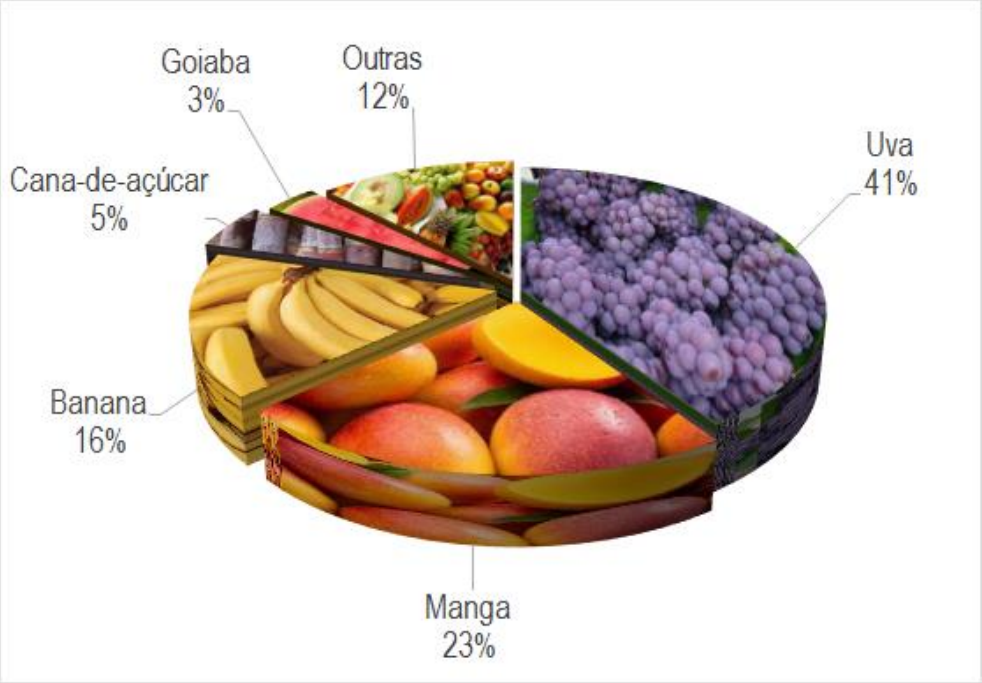


Figura *2*. Principais culturas produzidas nos projetos públicos de irrigação da Codevasf de acordo com o Valor Bruto de Produção (VBP), em 2019. Fonte: Codevasf, 2019.

Segundo dados da Organisation Internationale de la Vigne et du Vin (OIV), no ano de 2018, a superfície mundial plantada com videiras foi de cerca de 7,4 milhões de hectares, com cerca de 41% dos vinhedos de todo o mundo concentrados em cinco países: Espanha (13%), China (12%), França (11%), Itália (9%) e Turquia (6%). A produção mundial foi de 77,8 milhões de toneladas de uvas, sendo 57% de uvas de vinho, 36% de uvas de mesa e 7% de uvas secas. Os cinco principais países produtores são: China (11,7 milhões de toneladas), Itália (8,6 milhões de toneladas), Estados Unidos (6,9 milhões de toneladas), Espanha (6,9 milhões de toneladas) e França (6,2 milhões de toneladas). O Brasil ocupa a 15ª colocação, tendo produzido em 2020 cerca de 1.416.398 milhões de toneladas de uvas, sendo 48,61% dessa produção de uvas de mesa e 51,39% de uvas para processamento de vinhos e sucos (IBGE 2020, Kist et al. 2020).

A Vitivinicultura é um dos mais representativos setores da fruticultura brasileira em relação a geração de emprego, renda e valor bruto, pelos múltiplos usos da uva, que é utilizada desde o consumo in natura até a produção de sucos, vinhos, espumantes e doces (Kist et al. 2020). Está difundida desde o Rio Grande do Sul (a 31°S de latitude) até o Rio Grande do Norte e Ceará (a 05°S de latitude). A variação de altitude também é grande, havendo considerável diversidade ambiental entre as

zonas de produção, incluindo regiões de clima temperado, subtropical e tropical (Camargo et al. 2011).

Segundo dados do IBGE, em 2020, a área plantada com videiras no Brasil foi de 74.826 hectares, com destaque para a região Sul (54.716 ha) e Nordeste (10.429 ha), que produzem 853.300 e 387.662 toneladas, respectivamente. Os maiores estados produtores são o Rio Grande do Sul, com uma produção de 735.356 toneladas, e Pernambuco, com 338.837 toneladas. Aproximadamente metade da produção de uvas brasileiras atende às demandas do mercado de consumo *in natura*, enquanto a outra metade é explorada para processamento. A concentração da produção de uvas de mesa ocorre na região Nordeste, enquanto a produção de uvas para processamento está concentrada no Sul do país (Maia et al. 2018).

A Vitivinicultura tropical é típica de regiões onde as temperaturas mínimas não são suficientemente baixas para induzir a videira à dormência. A videira cresce continuamente e através de tecnologia apropriada é possível a obtenção de duas ou mais colheitas por ano, no mesmo parreiral. Os principais pólos de Vitivinicultura tropical no Brasil são o Submédio do Vale do São Francisco, o noroeste paulista e o norte de Minas Gerais (Camargo et al. 2011). As uvas para exportação são produzidas no Submédio do Vale do São Francisco, localizado nos estados de Pernambuco e Bahia, onde prevalece a atuação de médias e grandes empresas agrícolas (Maia et al. 2018).

De 1998 até 2017, as exportações brasileiras de uvas de mesa tiveram um crescimento expressivo de 909,9%, apresentando um incremento médio anual de 7,6% nesse período. Essas taxas expressivas de crescimento devem-se à ampliação da produção no Submédio do Vale do São Francisco, direcionada principalmente para o consumo *in natura*. Cerca de 10% da produção é destinada para o mercado externo, concentrada principalmente no segundo semestre (meses de setembro, outubro e novembro), pois se trata de um período de entressafra mundial de uvas de mesa e os preços pagos pela fruta atingem valores mais elevados (Maia et al. 2018).

A área de vinhedos destinada à elaboração de vinhos finos na região do Submédio do Vale do São Francisco totaliza aproximadamente 500 hectares. A região produz os chamados vinhos tropicais, com originalidade e identidade própria da região tropical, como vinhos finos tranquilos e espumantes, vinho licoroso e brandy. Anualmente são produzidos aproximadamente 4 milhões de litros de vinhos finos

(além dos vinhos comuns, que também começam a ganhar espaço na região), empregando direta ou indiretamente cerca de 3.000 pessoas (Embrapa 2020). A produção de suco de uva é vista como uma nova alternativa de geração de renda para os viticultores locais.

2.3. Banco Ativo de Germoplasma de Videira da Embrapa Semiárido

Ao adquirir o hábito da agropecuária, o homem foi domesticando e selecionando as espécies para o meio-ambiente em que se fixava. Com o domínio desses recursos genéticos, surgiu a necessidade de colecionar espécies vegetais, animais e microorganismos para uso presente e futuro. Segundo Costa et al. (2012), um Banco Ativo de Germoplasma (BAG) se caracteriza por manter um número grande de acessos, os quais são representativos da variabilidade genética da espécie e gênero e é mantido com fins conservacionistas, sem efetuar descartes de seus acessos. Esses acessos — que podem ser o próprio indivíduo representativo da espécie, ou partes dele, como estacas (no caso de vegetais), sementes, gametas, embriões, tecidos e DNA — são conservados, regenerados, caracterizados e disponibilizados para o uso e intercâmbio (Costa et al. 2012). Quando o melhorista necessita de mais genes que confiram outras características de interesse às suas novas pesquisas, ele faz uso dos acessos de um BAG (Costa et al. 2012, Machado et al. 2016).

O avanço da agropecuária sobre áreas de vegetação que antes não possuíam valor agrícola é considerado um dos agentes causadores da erosão da biodiversidade no planeta. Hoje, os BAGs transformaram-se nos principais refúgios de recursos genéticos, pois são neles que se encontram armazenados os genes que podem devolver parte da vida ao planeta. Desses BAGs saem os genes das novas cultivares, das espécies alternativas, das espécies do repovoamento e revegetação de áreas degradadas, entre outros fins (Costa et al. 2012). Segundo dados do INRAE (2020), a maioria dos países com tradição em vitivinicultura mantém bancos ou coleções de germoplasma de uva, sendo contabilizados em todo o mundo cerca de 130 repositórios e o maior deles preserva mais de 7.000 acessos. Esses bancos e coleções são mantidos normalmente no campo, para avaliação do material e fácil distribuição dos genótipos. Em um BAG de videira são mantidas vivas, normalmente, de duas a dez plantas de cada genótipo (espécies, híbridos, cultivares e clones)

propagados vegetativamente. Os BAGs no campo tem manutenção e manejo trabalhosos, além de estarem expostos aos riscos de perda por calamidades naturais, pragas e doenças (Leão 2008).

O Banco Ativo de Germoplasma mantido pela Embrapa Uva e Vinho, possui 1.400 acessos e é considerado o maior acervo de germoplasma de videira de toda a América Latina, bem como nas demais unidades parceiras da Embrapa espalhadas pelo país que fazem parte do Programa de Melhoramento Genético 'Uvas do Brasil', caso da Embrapa Semiárido.

Ao longo de sua história, a Embrapa Semiárido vem executando ações de pesquisa e desenvolvimento nesse espaço, mantendo um abrangente programa de geração de conhecimentos, de tecnologias e de inovação para as áreas secas do Nordeste, com foco na sustentabilidade da agropecuária, preservação ambiental e a melhoria dos índices sociais do Semiárido brasileiro (Embrapa 2020).

O BAG de Videira da Embrapa Semiárido (Figura 3), está localizado no Campo Experimental de Mandacaru, na cidade de Juazeiro, Estado da Bahia (9°24"S, 40°26"O e 365,5m de altitude) e possui acessos que incluem um grande número de variedades das espécies cultivadas (*Vitis vinifera* e *Vitis labrusca*), variedades híbridas interespecíficas e espécies silvestres americanas. A maioria dos acessos destina-se ao consumo *in natura* e para a elaboração de vinhos, mas existem ainda variedades para a elaboração de sucos, com potencial para produção de uva passa e porta-enxertos. As atividades referentes a recursos genéticos de videira, como coleta, intercâmbio, conservação, documentação, caracterização e avaliação de germoplasma são realizadas no BAG de Videira da Embrapa Semiárido para fornecer informações valiosas para subsidiar o melhoramento genético e o desenvolvimento de novas variedades de videira para o semiárido brasileiro.

Leão (2008) avaliou a diversidade genética presente em uma coleção do BAG de videira da Embrapa Semiárido com base em características morfoagronômicas de variação contínua e discreta, através de técnicas multivariadas, análise dos componentes principais, método de otimização de Tocher, UPGMA e projeção gráfica das distâncias, concluindo que estas foram eficientes no agrupamento dos genótipos mais similares, de acordo com as suas características fenotípicas, ou com base em sua genealogia e origem.

Leão et al. (2011) avaliaram a diversidade genética presente em 136 acessos de uvas de mesa do BAG da Embrapa Semiárido, com base em características morfoagronômicas de variação contínua e discreta. A análise de agrupamento pelo método de Tocher resultou na formação de 30 grupos utilizando-se descritores morfoagronômicos de variação contínua e 9 grupos, com base em caracteres multicategóricos. Não houve concordância entre os grupos obtidos pela análise de descritores fenotípicos contínuos e discretos, independente do método de agrupamento utilizado e foi detectada a existência de variabilidade genética satisfatória entre os acessos de uvas de mesa avaliados.

Leão & Motoike (2011) analisaram a diversidade genética de 47 acessos de uvas de mesa, procedentes do BAG de Videira da Embrapa Semiárido, por meio de 20 marcadores moleculares RAPD e sete marcadores microssatélites, e obtiveram as distâncias genéticas entre pares de acessos com base no índice de similaridade de Jaccard para marcadores RAPD e no complemento aritmético do índice ponderado para dados de microssatélites. Os grupos foram formados de acordo com a análise de agrupamento de Tocher e com o método de agrupamento não ponderado (UPGMA), concluindo que os marcadores microssatélites foram mais eficientes do que os RAPD na identificação das relações de parentesco e que as informações de distância genética, baseadas em características moleculares e aliadas ao desempenho agronômico das cultivares, permitiram a recomendação de parentais para cruzamentos, para a obtenção de híbridos superiores nas populações segregantes do programa de melhoramento de videira da Embrapa Semiárido.

Batista et al. (2015) avaliaram a divergência genética entre 31 variedades de videira do BAG da Embrapa Semiárido, considerando características de qualidade, compostos de importância funcional e atividade antioxidante dos frutos. Foram avaliadas nas bagas as seguintes características: teor de sólidos solúveis, acidez titulável, relação sólidos solúveis/acidez titulável, resistência da baga à força de compressão, teor de taninos (dímeros, oligoméricos e poliméricos), polifenóis extraíveis totais, antocianinas totais, flavonoides amarelos e atividade antioxidante. Foi utilizada a distância generalizada de Mahalanobis para quantificar a divergência genética entre as variedades. Como estratégias de agrupamento, foram empregados os métodos UPGMA e a análise de variáveis canônicas. A variabilidade genética para teores de compostos fenólicos, determinados por meio de taninos poliméricos, de

polifenóis extraíveis totais e de antocianinas totais, existentes em variedades de videira do Banco Ativo de Germoplasma da Embrapa Semiárido permitiu a identificação de genótipos divergentes com potencial aproveitamento em futuras ações de melhoramento voltadas para a melhoria das propriedades funcionais.

Figura 3. Banco Ativo de Germoplasma de Videira - Embrapa Semiárido - Juazeiro, BA. Fonte: Autor, 2021.

A contribuição mais recente do BAG de Videira da Embrapa Semiárido para o agronegócio brasileiro e, em especial, do Submédio do Vale do São Francisco, ocorreu em 2020, com o lançamento da primeira cultivar de uva de mesa 100% nordestina: a BRS Tainá (MAPA 2020).

2.4. Diversidade Genética

Uma condição básica da evolução biológica é a existência da diversidade genética. No passado, a natureza dióica das espécies de videira selvagem ajudou a manter a heterozigose e essa diversidade. Também aumentaram as chances de hibridação e desenvolvimento de formas híbridas onde existiam espécies simpátricas. As uvas também fazem parte de um processo de conservação e aumento da diversidade genética, pois quando as bagas amadureciam, eram consumidas por pássaros que espalhavam suas sementes na natureza. Essas sementes podiam

originar indivíduos únicos e introduzir seus alelos às populações pelas quais passavam (Walker et al. 2019, Yue et al. 2019).

O estudo da diversidade genética permite o aprimoramento genético das populações de plantas e o fornecimento de ricos recursos de germoplasma para a criação de novas variedades (Yue et al. 2019). Estresses bióticos e abióticos provocados por novos patógenos, pragas e mudanças climáticas estimularam a criação de variedades melhor adaptadas. A diversidade genética adequada é a chave para o melhoramento de culturas capazes de resistir a esses desafios (Riaz et al. 2018). Os BAGs, devidamente constituídos, são ferramentas valiosas à preservação e à conservação do máximo possível de diversidade e variabilidade genética existente na natureza. A importância da diversidade genética consiste em introduzir novos caracteres no contexto genotípico cultivado já existente. A variabilidade genética é muito importante porque sem ela não pode haver ganho com a seleção (Costa et al. 2012, Brown et al. 2014).

A análise da diversidade genética nos bancos e coleções de germoplasma facilita a classificação de acessos e a identificação de subconjuntos de acessos principais com possível utilidade para fins específicos de melhoramento. O estudo da diversidade genética é o processo pelo qual a variação entre indivíduos ou grupos de indivíduos ou populações é analisada por um método específico ou uma combinação de métodos. Os dados geralmente envolvem medições numéricas e, em muitos casos, combinações de diferentes tipos de variáveis (Mohammadi and Prasanna 2003).

A diversidade genética pode ser estudada pelo uso de várias técnicas estatísticas que podem prever a diversidade. Essas técnicas se baseiam em dados de linhagem, dados morfológicos, dados de desempenho agronômico, dados bioquímicos e dados moleculares (Mohammadi and Prasanna 2003, Barbosa et al. 2011). Vários são os métodos disponíveis para a análise da diversidade genética em acessos de germoplasma, linhagens e populações, como a análise de componentes principais, de variáveis canônicas e os métodos de agrupamentos (Cruz 1994, Mohammadi and Prasanna 2003). Atualmente, com o avanço de recursos computacionais e técnicas matemáticas em análise biológica, novas técnicas e métodos de avaliação da diversidade genética têm sido propostos (Barbosa et al. 2011).

2.5. Inteligência artificial

A inteligência artificial (IA) é uma área da ciência e engenharia. Pode ser definida como a ciência que estuda diferentes tipos de algoritmos, também chamados de algoritmos inteligentes, que podem ser implementados em máquinas, permitindo a essas máquinas raciocinar, aprender e agir de forma inteligente, funcionando de maneira autônoma em ambientes complexos e em constante mudança (Russell and Norvig 2016, Husain 2017).

O seu estudo teve início após a Segunda Guerra Mundial, e o termo "inteligência artificial" foi criado em 1956, por John McCarthy, que fundou formalmente essa área de estudo em parceria com Marvin Minsky. O primeiro trabalho reconhecido a tratar sobre IA foi realizado por Warren McCulloch e Walter Pitts, em 1943. Eles propuseram um modelo de neurônios artificiais, chamados de *perceptrons*, com base em uma análise detalhada dos neurônios originais biológicos, em que cada neurônio é caracterizado como "ligado" ou "desligado", com uma mudança para "ligado" ocorrendo em resposta a um estímulo por um número suficiente de neurônios vizinhos. Dessa forma, qualquer função computável poderia ser calculada por alguma rede de neurônios conectados e todos os conectivos lógicos (e, ou, não, etc.) podem ser implementados por estruturas de rede simples. McCulloch e Pitts também sugeriram que redes adequadamente definidas poderiam aprender, e, portanto, mudar sua ação em relação ao tempo, tornando-se agentes inteligentes. Idealmente, um agente inteligente executa a melhor ação possível em uma situação (Russell and Norvig 2016, Warwick 2013).

O aprendizado de máquina (*machine learning*) é a parte da IA dedicada ao desenvolvimento e estudo de algoritmos que aprendem com os dados. Ele pode ser definido como uma subárea científica da IA que confere às máquinas a habilidade de aprender sem serem estritamente programadas. As metodologias do aprendizado de máquina envolvem um processo de treinamento com o objetivo de aprender com a “experiência” (dados de treinamento) para executar uma tarefa. Geralmente, um exemplo individual é descrito por um conjunto de atributos, também conhecidos como características ou variáveis. O desempenho do modelo de aprendizado de máquina em uma tarefa específica é medida através de experimentos empíricos sobre os dados. Para calcular o desempenho dos algoritmos de aprendizado de máquina, vários modelos estatísticos e matemáticos são usados. Após o final do processo de

aprendizagem, o modelo treinado pode ser usado para classificar, prever ou agrupar novos exemplos usando a experiência obtida durante o processo de treinamento (Liakos et al. 2018, Husain 2017).

Técnicas de aprendizado de máquina não supervisionado facilitam a análise de conjuntos de dados brutos, ajudando a gerar compreensões analíticas a partir de dados não rotulados. Os recentes avanços no aprendizado hierárquico, agrupamento de dados, modelos de variáveis latentes, redução de dimensionalidade e detecção de outliers, ajudaram a avançar significativamente o estado da arte em técnicas de aprendizado de máquina não supervisionada. Por exemplo, a introdução recente de métodos de aprendizagem de representação baseada em redes neurais profundas (*deep learning*) têm facilitado o processamento de dados brutos, sem a necessidade de uma cuidadosa engenharia e conhecimento do domínio para a criação de recursos (LeCun et al. 2015, Usama et al. 2019).

A IA vem avançando rapidamente nos últimos anos, tanto em termos da quantidade de recursos dedicados a ela como também em termos de seus resultados. O crescimento do investimento nessa área recentemente foi impulsionado e também contribuiu para o rápido aumento nas capacidades técnicas da inteligência artificial (Furman and Seamans 2019). Atualmente, a IA fornece revolucionárias oportunidades em uma grande diversidade de aplicações que substituem o trabalho humano ou apoiam os seres humanos em seu trabalho, sob a forma de robótica, como parte de uma variedade de dispositivos inteligentes (Jaakkola et al. 2019), apresentando inovações em todos os setores da economia (Smith 2020), com aplicação nas mais diversas áreas: na indústria do petróleo (Rahmanifard and Plaksina 2018), no setor da contabilidade (Shi 2019) e gestão financeira (Polak et al. 2019), na educação e sistemas educacionais (Vincent-Lancrin and Vlies 2020), na medicina (Winkler-Schwartz et al. 2019, Mei et al. 2020), no comércio (Song 2019), na biotecnologia (Kim 2019), na agricultura (Liakos et al. 2018), entre várias outras áreas.

2.6. Inteligência artificial aplicada ao melhoramento genético de plantas

A agricultura desempenha um papel crítico na economia global. A pressão no sistema agrícola aumentará com a expansão contínua da população humana. Novas tecnologias, capazes de lidar com uma grande quantidade de dados para impulsionar a produtividade agrícola, minimizando o impacto ambiental, são necessárias (Liakos

et al. 2018). A IA pode fornecer informações mais precisas sobre situações de campo, permitindo que o responsável técnico analise, pense e decida sobre as ações de gerenciamento. Monitorar os sistemas agrícolas com maior precisão promove um aumento das habilidades de investigação, permitindo entender as razões pelas quais vários fenômenos estão ocorrendo. Um desafio ao trabalhar com sistemas biológicos é que eles podem não se comportar conforme o esperado, porque são influenciados por fatores conhecidos ou desconhecidos de maneiras que não entendemos. A coleta mais refinada de dados via IA permite análises aprimoradas de como os diferentes fatores interagem para influenciar os sistemas agrícolas (Smith 2020).

O aperfeiçoamento dos algoritmos de IA e do entendimento biológico aumentaram a capacidade de prever o desenvolvimento da produção agrícola. O melhoramento genético de plantas é uma área que pode ser beneficiada com grandes avanços através de previsões mais precisas sobre o desempenho das diferentes variedades em diferentes circunstâncias. A consideração clássica da produção em função dos efeitos genotípicos, ambientais e de manejo, ou seja, a equação onde o fenótipo é o resultado da interação genótipo-ambiente, poderia ser resolvida com maior precisão através do uso da IA para a recomendação de novas variedades (Anderson 2010, Smith 2020).

Khaki et al. (2020) utilizaram a IA na previsão da performance produtiva dos parentais de milho no melhoramento genético de plantas. Essas previsões foram feitas com base nos desempenhos históricos de linhas puras e testadores, suas informações de agrupamento genético e locais de plantio. Foi concluído que o modelo proposto pôde estimar o desempenho de qualquer combinação entre as linhas puras e os testadores antes dos cruzamentos reais. Isso auxiliaria os melhoristas de plantas a se concentrarem nas melhores combinações possíveis, podendo também estender esse método para incluir outras variáveis importantes, como componentes climáticos e condições do solo, para melhorar o desempenho da previsão.

Uzal et al. (2018) realizaram a fenotipagem de plantas de soja para estimar o número de grãos por vagem para o melhoramento genético através do uso de imagens aplicadas à uma rede neural convolucional e uma abordagem clássica usualmente utilizada para tais estudos. Concluíram que a rede neural convolucional superou a abordagem clássica, obtendo uma acurácia de 86.2 % frente a 50.4% da abordagem clássica. Além disso, a rede neural convolucional ofereceu um comportamento mais

robusto diante das mudanças sazonais e também aprendeu a detectar formas específicas no contorno das vagens.

Xavier et al. (2017), utilizaram a IA para avaliar a interdependência dos caracteres agronômicos da soja e os componentes de rendimento para entender a dependência mútua desses caracteres através de sua correlação fenotípica, genotípica e ambiental usando um conjunto de dados coletados em vários anos. Como resultado do estudo, concluíram que o uso de métodos de aprendizado de máquina não supervisionada fornece uma boa estrutura para investigar interações entre vários caracteres quantitativos, definindo assim quais deles seriam os alvos para o melhoramento genético.

Corrêa et al. (2016) utilizaram redes neurais artificiais na seleção de genótipos de feijoeiro com alta adaptabilidade e estabilidade fenotípica e para verificar sua consistência com o método de Eberhart e Russell. Eles observaram que houve alta concordância entre as metodologias avaliadas para discriminação da adaptabilidade fenotípica de genótipos de feijoeiro, indicando que redes neurais artificiais podem ser utilizadas em programas de melhoramento genético para a seleção de genótipos de feijoeiro com alta adaptabilidade e estabilidade fenotípica.

Ampatzidis et al. (2019) desenvolveram uma metodologia para a rápida avaliação de novas cultivares de porta-enxertos de citros em um campo comercial em larga escala utilizando uma nova técnica de alto rendimento baseada no uso de um veículo aéreo não tripulado (drone), relatando que os dados coletados pelo drone e analisados por redes neurais artificiais mostraram uma forte correlação com os dados coletados manualmente. Os autores citam que a correlação entre o tamanho da copa das árvores medido manualmente e medido pelo drone foi alta (R = 0,84), concluindo que a técnica baseada no uso do drone reduz o trabalho, é mais econômica e consistente do que o método manual, e que a técnica proposta pode ser utilizada para avaliar com precisão e rapidez as cultivares e práticas de manejo, auxiliando os melhoristas a identificar cultivares de porta-enxertos de citros com melhor desempenho.

Vários são os estudos disponíveis na literatura que utilizam a inteligência artificial para fins de agrupamento ou classificação e análise de diversidade genética. Gutiérrez et al. (2015) propuseram em seu estudo um novo método de classificação para variedades de videiras através do uso de máquinas de vetor de suporte e redes

neurais artificiais utilizando um espectrofotômetro infravermelho portátil nas folhas das plantas em situação de campo. As máquinas de vetores de suporte e as redes neurais artificiais mostraram uma alta confiabilidade na criação de modelos de classificação varietal de folhas de videira.

Marques et al. (2019) propuseram um método automático para a classificação de variedades de videira com base nas características das folhas da planta utilizando o aprendizado de máquina. Foram analisadas as imagens de folhas de três variedades de videira, adquiridas em um vinhedo, onde foram avaliados seis algoritmos de aprendizado de máquina normalmente utilizados para fins de classificação. Os resultados mostraram que o método proposto pode ser usado como uma alternativa eficaz ao procedimento manual para classificação da videira com base nas características morfológicas da folha.

Barbosa et al. (2011) avaliaram a viabilidade de redes neurais artificiais como uma técnica de análise da diversidade genética de *Carica papaya* L., implementando uma rede neural artificial, conforme modelo proposto por Kohonen (1982), na tentativa de propor uma classificação e a formação de grupos divergentes dos acessos, com base no uso de um banco de dados de caracteres agronômicos, como o peso do fruto, comprimento e diâmetro, espessura da polpa, firmeza, sólidos solúveis, entre outros. Os resultados mostraram que 91.9% dos acessos foram classificados corretamente nos grupos recomendados pela rede neural artificial, que mostrou ser uma técnica viável na classificação dos acessos.

Campos et al. (2016) avaliaram a formação de grupos heteróticos em goiaba, com base em descritores quantitativos e utilizando uma rede neural artificial tipo *Self-organizing-maps* (SOM). Essa técnica determinou o número ideal de três grupos, onde a consistência do agrupamento foi determinada por análise discriminante linear, que obteve um percentual de classificação dos grupos de 86%, concluindo-se que o método da rede neural artificial é eficaz para detectar a divergência genética e formação de grupos heteróticos.

Costa et al. (2018) avaliaram a diferenciação genética das variedades de porta-enxertos de videira através do algoritmo SOM baseada na combinação de marcadores moleculares RAPD e SSR para determinar a caracterização entre as variedades de porta-enxertos 420-A, Schwarzmann, IAC-766 Campinas, Traviú, Kober 5BB e IAC-572 Jales e concluíram que apesar das semelhanças morfológicas das variedades

420-A e Kober 5BB, que compartilham a mesma origem genética, foram geradas duas novas variedades geneticamente divergentes e que mostram diferenças no desempenho.

O desenvolvimento de novas abordagens para alcançar um melhoramento genético de plantas de alto nível é imperativo. Trabalhos sobre a associação da genômica e fenômica à IA estão fornecendo novos conhecimentos sobre os complexos mecanismos biológicos da interação genótipo-ambiente, como por exemplo, às funções da planta em resposta a perturbações ambientais. Nos próximos anos, não apenas o melhoramento genético de plantas, mas toda a agricultura, confiará nos métodos de IA que tomam decisões e fazem recomendações eficientes baseadas em enormes quantidades de dados altamente heterogêneos e complexos (Harfouche et al. 2019).

2.7. Agrupamento de dados (*data clustering*)

A classificação, uma das atividades mais primitivas do homem, desempenha um papel importante e indispensável na longa história do desenvolvimento humano. Para aprender um novo objeto ou entender um novo fenômeno, os seres humanos sempre tentam procurar os caracteres que podem descrevê-lo e compará-lo com outros objetos ou fenômenos conhecidos, com base na semelhança, dissimilaridade ou proximidade, segundo certos padrões ou regras (Anderberg 1974, Xu and Wunsch 2005).

O agrupamento de dados (*data clustering* ou análise de cluster) é uma técnica de aprendizado não supervisionado que visa organizar uma coleção de itens de dados em clusters (grupos, subconjuntos ou categorias), de modo que os itens de um cluster sejam mais semelhantes entre si do que os itens dos outros clusters .

A maioria dos pesquisadores descreve um cluster considerando a homogeneidade interna e a separação externa, ou seja, os padrões no mesmo cluster devem ser semelhantes entre si, enquanto os padrões em diferentes clusters devem ser diferentes. Essa noção de similaridade pode ser expressa de maneiras muito diferentes, de acordo com o objetivo do estudo, com premissas específicas do domínio e com o conhecimento prévio do problema (Grira et al. 2005, Xu and Wunsch 2005, Everitt et al. 2011).

O agrupamento é tradicionalmente visto como parte do aprendizado não supervisionado, pois é realizado quando não há informações disponíveis sobre a associação de itens de dados a classes pré-definidas, visto que nenhum dado rotulado está disponível. O objetivo do agrupamento em cluster é separar um conjunto de dados finito não rotulado em um conjunto finito e discreto de dados que possuem padrões ocultos "naturais" entre si, ao invés de fornecer uma caracterização precisa de amostras não observadas geradas a partir da mesma distribuição de probabilidade (Grira et al. 2005, Xu and Wunsch 2005, Everitt et al. 2011).

Segundo Xu and Wunsch (2005), uma análise de cluster tem basicamente quatro etapas: seleção ou extração de caracteres, design ou escolha do algoritmo de cluster, validação do cluster e interpretação dos resultados. A seleção escolhe os caracteres diferenciados de um conjunto de candidatos, enquanto a extração utiliza algumas transformações para gerar caracteres úteis e novos a partir dos originais. Ambos são cruciais para a eficácia de aplicação do cluster. O design ou escolha do algoritmo de cluster é uma etapa geralmente combinada com a seleção de uma medida de proximidade correspondente, e a construção de uma função de critério e a medida de proximidade afeta diretamente a formação dos clusters. Quase todos os algoritmos de agrupamento estão explícita ou implicitamente conectados a alguma definição de medida de proximidade. Uma vez escolhida uma medida de proximidade, a construção de uma função de critério de agrupamento torna a partição de clusters um problema de otimização, que é bem definido matematicamente e possui várias soluções na literatura. O agrupamento é uma técnica muito difundida no mundo todo, e existe uma enorme variedade de algoritmos de clustering, desenvolvidos para resolver problemas diferentes em campos específicos. Ainda segundo Xu and Wunsch (2005), não existe um algoritmo de clustering que pode ser usado universalmente para resolver todos os problemas, sendo muito importante investigar cuidadosamente as características do problema em questão para selecionar ou projetar uma estratégia de clustering apropriada.

Na etapa de validação do clustering deve-se ter em mente que em um conjunto de dados qualquer, cada algoritmo de agrupamento sempre pode gerar uma divisão, independentemente da estrutura existir ou não. Além disso, abordagens diferentes geralmente levam a grupos diferentes; e mesmo para o mesmo algoritmo, a identificação de parâmetros ou a ordem de apresentação dos padrões de entrada

podem afetar os resultados finais (Xu and Wunsch 2005). Sendo assim, os padrões e critérios efetivos de avaliação são importantes para fornecer um grau de confiança nos resultados de cluster derivados dos algoritmos usados. Essas avaliações devem ser objetivas e não ter preferência para nenhum algoritmo. Elas devem ser úteis para responder perguntas como quantos clusters estão ocultos nos dados, se os clusters obtidos são significativos ou apenas um artefato dos algoritmos, ou por que escolhemos um algoritmo em vez de outro. Algumas métricas podem ser utilizadas para a validação do cluster, como por exemplo a métrica de silhueta, que obtém o número de agrupamento ideal pela diferença entre a distância média dentro do agrupamento e a distância mínima entre os agrupamentos, ou seja, o efeito de agrupamento ideal (Wang and Xu 2019). O objetivo final do clustering é fornecer informações significativas a partir dos dados originais, para que eles possam resolver efetivamente os problemas propostos. Na última etapa da análise de cluster, que é a interpretação dos resultados, especialistas nos campos relevantes interpretam a separação dos dados. Análises adicionais, inclusive experimentos, podem ser necessárias para garantir a confiabilidade do conhecimento extraído (Bishop 1995, Jain et al. 2000, Xu and Wunsch 2005).

Os algoritmos de clustering separam os dados em um certo número de clusters. Segundo Charikar et al. (2019), o agrupamento hierárquico é um popular algoritmo exploratório de análise de dados com uma variedade de aplicações, sendo utilizado para o agrupamento de imagens e textos até a análise de redes sociais e mercados financeiros. Uma de suas principais aplicações é na área da filogenética, onde são usados os padrões de similaridade ou dissimilaridade genômica, com a finalidade de criar taxonomias de organismos, objetivando lançar luz sobre a evolução de espécies pela compreensão de uma árvore ancestral da vida. Os SOM também são usados para mapear e organizar populações geneticamente similares ou distintas. Como não é possível comparar seus resultados com os resultados obtidos por técnicas convencionais, estudos de simulação onde a estrutura genética da população já é conhecida foram importantes para demonstrar a eficiência dos self organizing maps, revelando que os resultados por eles obtidos têm sentido biológico (Oliveira et al. 2019).

2.8. Redes Neurais Artificiais

Segundo Haykin (2008), uma rede neural artificial (RNA) é um processador paralelo (ou seja, um sistema de processamento de dados) distribuído massivamente, composto de unidades de processamento simples (os neurônios) com aptidão natural para armazenar conhecimento experimental e disponibilizá-lo para uso. Assemelha-se ao cérebro em dois aspectos: 1) O conhecimento é adquirido pela rede a partir de seu ambiente, através de um processo de aprendizado; 2) Os pontos fortes da conexão entre os neurônios, conhecidos como pesos sinápticos, são usados para armazenar o conhecimento necessário. A modificação dos pesos sinápticos fornece o método tradicional para o design de redes neurais. No entanto, também é possível para uma rede neural modificar sua própria topologia, baseada no fato de que os neurônios, no cérebro humano, podem morrer e novas conexões sinápticas podem surgir.

Uma RNA padrão consiste em componentes simples conectados chamados de neurônios, que produzem saídas com valores reais de acordo com os pesos atribuídos às conexões e suas funções de ativação (Sharma and Singh, 2017). Um neurônio básico de uma RNA é considerado uma unidade de processamento de dados que é conectada a outros neurônios por meio de pesos. Cada conexão tem seu próprio peso, que geralmente é apenas um número aleatório no início (Van Veen, 2017).

Haykin (2008) e Goodfellow et al. (2016) explicam em detalhes que um neurônio artificial é constituído por três elementos básicos: 1) um conjunto de sinapses ou elos de conexão, cada um dos quais é caracterizado por um peso ou força própria. Especificamente, um sinal xj na entrada da sinapse j conectado ao neurônio k e é multiplicado pelo peso sináptico wkj. O primeiro subscrito em wkj refere-se ao neurônio em questão e o segundo subscrito ao final da entrada da sinapse à qual o peso se refere. Ao contrário do peso de uma sinapse no cérebro, o peso sináptico de um neurônio artificial pode estar no intervalo que inclui valores negativos e positivos; 2) um somador, que soma os sinais de entrada, ponderados pelos respectivos pesos sinápticos do neurônio; 3) e uma função de ativação, para limitar a faixa de amplitude de saída de um neurônio para algum valor finito.

As RNAs podem possuir poucas ou muitas camadas e diferentes tipos, arranjos e quantidade de neurônios. Essas diferentes formas de organização dizem respeito a topologia da RNA. Pode-se então dizer que existem diferentes tipos de RNAs, pois

elas possuem topologias diferentes, e, portanto, diferentes aplicabilidades (Van Veen, 2017). Tch (2017) comenta sobre alguns tipos de RNAs em seu artigo: o Perceptron é considerado o modelo mais simples e antigo de RNA e possui apenas uma camada de entrada e um neurônio de saída. A RNA *Feed Forward* também é uma RNA antiga que apresenta uma topologia que contém apenas uma camada entre a camada de entrada e de saída. A RNA MLP (*Multiple Layer Perceptron*) possui três tipos de camadas: a camada de entrada (*input layer*), camadas ocultas (*hidden layers* ou *black box*) e a camada de saída (*output layer*), enquanto as RNAs *Deep Feed Forward* possuem várias camadas ocultas. As RNAs Convolucionais são um tipo de RNA que tem sido bastante utilizadas atualmente para a solução de problemas através do reconhecimento de imagens. A dimensionalidade das camadas de uma RNA determina a amplitude do modelo. Existem RNAs que apresentam topologias bem características, desenvolvidas para soluções de problemas específicos, como, por exemplo, os *Autoencoders*, que são um tipo de RNA usada nas tarefas de classificação, agrupamento e redução de caracteres, e a RNA de Kohonen, que possui como característica o conceito de “distância para o neurônio”, onde esses neurônios tentam ajustar seus pesos sinápticos ao máximo em relação às entradas que são apresentadas à rede. A RNA de Kohonen (*self-organizing maps*) é bastante utilizada para tarefas de agrupamento (Van Veen 2017, Usama et al. 2019).

Aprender é o processo de atribuir parâmetros ideais de ativação que permitem à RNA executar os dados na camada de entrada de modo que sejam mapeados para um padrão na camada de saída. Para resolver um determinado problema, uma RNA pode, ou não, exigir múltiplas camadas ocultas (Usama et al. 2019). O processo de treinamento de uma RNA é supervisionado quando o resultado ideal na camada de saída é especificado. Caso o resultado ideal não seja fornecido, trata-se de um treinamento não supervisionado. O treinamento não supervisionado geralmente ensina a RNA a colocar os dados em vários grupos definidos pela contagem de neurônios de saída, tratando-se de um processo iterativo. Porém, calcular o erro não é tão fácil, pois não há um resultado esperado. Dessa forma, não é possível mensurar o quão distante está o resultado do treinamento não supervisionado da RNA do resultado ideal. Então, deve ser repetido um número fixo de iterações. Caso a RNA precise de mais treinamento, ele será fornecido pelo programa (Heaton 2015).

Uma RNA obtém seu poder computacional através de sua estrutura paralela massivamente distribuída e de sua capacidade de aprender e, portanto, generalizar. A generalização refere-se à produção de resultados razoáveis da RNA para informações (dados) não encontrados durante o treinamento (aprendizado). Algumas outras vantagens das RNAs são: 1) ser uma estrutura não linear, formada por uma interconexão de neurônios lineares e não lineares, distribuída através da rede. A não linearidade é uma propriedade altamente importante, principalmente se as condições físicas subjacentes ao mecanismo responsável pela geração do sinal de entrada (por exemplo, um sinal de fala) é inerentemente não linear; 2) realizar o mapeamento de entrada-saída de dados. Para isso, é apresentado à rede um conjunto de dados com um exemplo escolhido aleatoriamente desse conjunto, e os pesos sinápticos (parâmetros livres) da rede são modificados para minimizar a diferença entre a resposta desejada e a resposta real da rede produzida pelo sinal de entrada de acordo com um critério estatístico adequado. O treinamento da rede é repetido para muitos exemplos no conjunto, até que a rede atinja um estado estável, onde não há mais alterações significativas nos pesos sinápticos. Assim, a rede aprende com os exemplos construindo um mapeamento de entrada e saída para o problema em questão. Essa abordagem lembra o estudo da estatística não paramétrica; 3) possuir adaptatividade, que é a capacidade incorporada das RNAs para adaptar seus pesos sinápticos às mudanças no ambiente; 4) proporcionar uma resposta evidencial, que no contexto da classificação de padrões, significa dizer que uma RNA pode ser projetada para fornecer informações não só sobre qual padrão particular selecionar, mas também sobre a confiança na decisão tomada. Esta última informação pode ser usada para rejeitar padrões ambíguos que devem surgir, melhorando assim o desempenho da classificação da rede; 5) ser tolerante à falhas. Uma RNA implementada em forma de hardware tem potencial para ser inerentemente tolerante à falhas, ou capaz de computação robusta, no sentido de que seu desempenho diminui muito pouco sob condições de operações adversas; 6) ser capaz de produzir informação contextual, pois todo neurônio na rede é potencialmente afetado pela atividade global de todos os outros neurônios. Consequentemente, as informações contextuais são tratadas naturalmente por uma RNA (Haykin 2008).

As técnicas de *deep learning* vêm revolucionando o aprendizado de máquina e a IA, e estão sendo cada vez mais usadas em diversas configurações, como, por

exemplo, na identificação e classificação de objetos através de imagens, na transcrição da fala em texto, entre outros. Devido a esses avanços, vários tipos de RNAs têm sido aperfeiçoadas e outros novos tipos estão surgindo (Usama et al. 2019).

2.8.1. *Self - Organizing Maps*

As RNAs de aprendizado competitivo não supervisionado são uma estrutura de neurônios do tipo *winner-takes-all* (o vencedor fica com tudo) onde cada neurônio compete pelo direito da resposta a um subconjunto dos dados de entrada. Esse esquema é usado para remover as redundâncias dos dados não estruturados. Duas técnicas principais de RNAs de aprendizado competitivo não supervisionado são os *self-organizing maps* e as redes ART (*Adaptive Resonance Theory*) (Usama et al. 2019).

Os *self-organizing maps* (SOM), que também são conhecidos como mapas auto-organizáveis ou *Kohonen's Maps* (Mapas de Kohonen) foram criados pelo professor finlandês Teuvo Kohonen. Em sua descoberta, Kohonen (1986) propôs que, em uma rede simples de elementos físicos adaptativos que recebem sinais de um espaço de evento primário, as representações desses sinais são mapeadas automaticamente para um conjunto de respostas de saída, de tal maneira que as respostas adquiram a mesma ordem topológica dos eventos primários, conforme representado na Figura 4. Em outras palavras, foi descoberto um princípio que facilita a formação automática de mapas de caracteres topologicamente corretos de eventos observáveis.

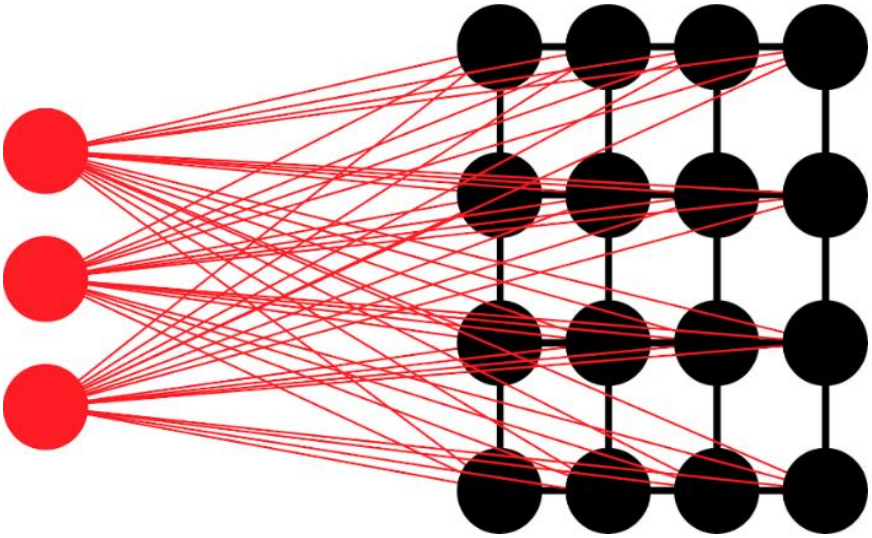


Figura 4. Representação dos sinais de entrada mapeados para um conjunto de respostas de saída. Adaptado de Kohonen, 1983. Fonte: Autor, 2021.

Auto-organização significa a habilidade de um sistema biológico ou técnico de adaptar sua estrutura interna às estruturas percebidas na entrada do sistema. Essa adaptação deve ser realizada de forma que, primeiro, nenhuma intervenção do ambiente seja necessária (aprendizagem não supervisionada) e, em segundo lugar, a estrutura interna do sistema auto-organizado representa as características dos dados de entrada que são relevantes para o sistema. Um exemplo biológico de auto-organização é o aprendizado de línguas pelas crianças. Este processo pode ser realizado por todas as crianças desde a mais tenra idade, em diferentes línguas e em diferentes culturas (Ultsch 1999).

A rede SOM é uma RNA bidimensional que organiza dados de um processo de aprendizado não supervisionado através da noção de vizinhança (função neighborhood) e usando a distância euclidiana. O aprendizado começa com a atribuição de pesos sinápticos e, em seguida, é iniciado um processo de competição no qual cada amostra de dados é alocada ao neurônio que melhor o representa. Este neurônio é chamado de "vencedor". Então, tem início a cooperação, na qual o neurônio vencedor determina a aproximação dos outros neurônios na ordem de proximidade. Finalmente, os neurônios que estabelecem sua vizinhança vão para a fase de adaptação, onde há ajustes de peso. Após todas as iterações, o mapa é organizado em uma estrutura topológica que reflete a proximidade dos elementos em estudo. A rede SOM pode apresentar topologia hexagonal, onde cada neurônio tem no máximo seis vizinhos diretos ou topologia quadrática com no máximo quatro

vizinhos diretos. Além disso, são estabelecidos diferentes arranjos que definem o número de neurônios disponíveis no mapa. A rede é forçada a representar dados de alta dimensão em uma representação de menor dimensão, preservando as propriedades topológicas dos dados de entrada, enquanto transforma os dados em um espaço topológico no qual as posições dos neurônios são representadas por características estatísticas intrínsecas que fazem parte da natureza inerentemente não linear das redes SOM (Rosenblatt 1958, Usama et al. 2019, Santos et al. 2019).

As redes SOM têm sido bastante utilizadas no melhoramento genético de plantas, em especial nos estudos de diversidade genética. Mancuso (2001) utilizou a rede SOM para realizar o agrupamento de 20 variedades de uvas de vinho com base em imagens das folhas dessas variedades, que foram convertidas em variáveis contínuas, concluindo que a rede SOM foi eficiente no agrupamento, permitindo elucidar as complexas relações entre os dados ampelográficos que não poderiam ser detectadas por meio de métodos tradicionais.

Oliveira et al. (2020) utilizaram a rede SOM para mapear os efeitos de deriva genética, seleção, migração e consanguinidade ao longo de gerações a partir das frequências alélicas e genotípicas, concluindo que a rede SOM é eficiente para capturar padrões de diversidade genética de populações sujeitas a processos que reduzem a variabilidade, como deriva, consanguinidade e seleção, e a processos que aumentam a variabilidade genética, como a migração. Também podem ser encontrados na literatura outros trabalhos utilizando as redes SOM para estudos de diversidade genética na cultura do arroz (Santos et al. 2019), do mamão (Barbosa et al. 2011), da soja (Sá 2015) e da goiaba (Campos et al. 2015).

2.8.2. *Emergent Self-Organizing Maps*

As redes SOM são complexas na forma como são compostas por vários elementos interconectados: neurônios e vizinhanças. Os SOM podem ser considerados como um sistema multi-agente, que seria um sistema composto por vários agentes que colaboram para atingir um objetivo. Esses agentes podem ser identificados como os neurônios e a colaboração é a modificação do peso do neurônio em uma vizinhança. O objetivo dos SOM é a adaptação à estrutura dos dados de entrada (Ultsch 2007).

Segundo Ultsch (1999), emergência é a capacidade de um sistema de produzir um fenômeno em um novo nível, superior. Esta mudança de nível é denominada pela física de "modo" ou "mudança de fase" e é produzida pela cooperação de muitos processos elementares. A emergência ocorre tanto em sistemas naturais quanto em sistemas artificiais ou técnicos. Até mesmo multidões de seres humanos podem produzir fenômenos emergentes. Um exemplo é o chamado "Ola" em estádios de futebol. Os seres humanos funcionam como processos elementares que, por cooperação, produzem uma onda em grande escala ao se elevar de seus lugares jogando seus braços para o ar. Esta onda pode ser observada em escala macroscópica e poderia, por exemplo, ser descrita em termos de comprimento de onda, velocidade e taxa de repetição. Sistemas técnicos importantes que são capazes de mostrar a emergência são, em particular, o laser e o maser. Nesses sistemas técnicos, bilhões de átomos (processos elementares) produzem um feixe de radiação coerente. A existência e cooperação de um grande número de processos elementares é necessária para alcançar a emergência.

Na rede SOM tradicional, o número de neurônios é muito pequeno para alcançar a emergência. Uma forma comum de uso da rede SOM é que o número de neurônios é aproximadamente igual ao número de clusters que se espera encontrar no conjunto de dados. Um único neurônio é normalmente considerado um cluster, e todos os dados cujas melhores correspondências caem nesse neurônio, são membros deste cluster (Ultsch and Mörchen 2005). Ultsch (1995) afirma em seu estudo que esse tipo de mapa de características de Kohonen realiza o agrupamento de uma maneira que é semelhante a outro algoritmo de agrupamento estatístico chamado *k-means*.

Os *emergent self-organizing maps* (ESOM), também chamados de mapas auto-organizáveis emergentes, contém um número muito maior de neurônios do que uma rede SOM tradicional. Usando a cooperação de muitos neurônios, os ESOM são capazes de construir estruturas que permitem a visualização de dados de alta dimensão em mapas de baixa dimensão, que, de outra forma, seriam invisíveis (Ultsch 1999). Esse grande número de neurônios pode representar clusters de dados individualmente, o que facilita sua detecção. Dessa forma, uma rede ESOM pode capturar a topologia das dimensões originais com maior acurácia (Ultsch and Kämpf 2004, Ultsch 2007).

Uma condição absolutamente necessária para obter a emergência é a cooperação de muitos processos elementares. Portanto, espera-se que a emergência aconteça apenas em mapas de características auto-organizáveis com um grande número de neurônios. Esses mapas de características emergentes normalmente têm pelo menos alguns milhares, senão dezenas de milhares de neurônios. O número de neurônios pode ser muito maior do que o número de dados de entrada. Os clusters são detectados nos ESOM não considerando neurônios individuais, mas considerando a estrutura geral de todo o mapa de características (Ultsch 1999). Através da emergência, os ESOM são diferentes e geralmente superiores aos algoritmos clássicos de agrupamento. Segundo Ultsch (1995) um exemplo canônico onde essa superioridade pode ser vista é em um conjunto de dados que consiste em dois subconjuntos diferentes. Usando um mapa de características emergentes de uma dimensão de 64 por 64 (4096 neurônios), os dois subconjuntos foram facilmente distinguidos. Porém, muitos algoritmos estatísticos, em particular o algoritmo *k-means*, foram incapazes de produzir uma classificação correta. A propriedade de emergência permite visualizar o surgimento de novas estruturas em um nível de abstração diferente e coincide bem com a ideia de realizar novas descobertas.

Haddad et al. (2009) utilizaram os ESOM com sucesso para realizar a identificação e o agrupamento de centenas de metabólitos de um banco de dados, visando encontrar diferenças ou semelhanças nas constituições metabólicas de um organismo entre diferentes condições ambientais.

Rimet et al. (2009) utilizaram os ESOM para o agrupamento de oito comunidades de algas presentes em fitoplâncton e realizaram a comparação desse agrupamento em relação a outras técnicas multivariadas, concluindo que os ESOM tiveram a vantagem de ordenar e agrupar os dados em uma única análise e apresentar os resultados de uma forma auto evidente e intuitiva, ao contrário das técnicas multivariadas, que geraram gráficos indecifráveis quando a quantidade de dados era enorme.

Apesar da aplicação dos *self-organizing maps* em estudos de diversidade genética de plantas e a consequente formação de grupos heteróticos já ser consagrada na literatura, não há, até o presente momento, estudos que utilizaram os *emergent self-organizing maps* para a análise da diversidade genética.

3. REFERÊNCIAS BIBLIOGRÁFICAS

Anderson WK (2010) Closing the gap between actual and potential yield of rainfed wheat. The impacts of environment, management and cultivar. **Field Crops Research 116**: 14-22.

Barbosa CD, Viana AP, Quintal SSR and Pereira MG (2011) Artificial neural network analysis of genetic diversity in Carica papaya L. **Crop Breeding and Applied Biotechnology** 11: 224-231.

Bertan I, Carvalho FIF, Oliveira AC, Vieira EA, Hartwig I, Silva JAG, Shimidt DAM, Valério IP, Busato CC and Ribeiro G (2006) Comparação de métodos de agrupamento na representação da distância morfológica entre genótipos de trigo. **Current Agricultural Science and Technology** 12: 1-8.

Bishop CM (1995) **Neural networks for pattern recognition**. Oxford University Press, New York, 457p.

Brown J, Caligari P and Campos H (2014) **Plant breeding**. Wiley-Blackwell, West Sussex, 296p.

Camargo UA, Tonietto J, Hoffmann A (2011) Progressos na viticultura brasileira. **Revista Brasileira de Fruticultura** 33: 144-149.

Campos B, Viana AP, Quintal SSR, Barbosa CD and Daher RF (2016) Heterotic group formation in Psidium guajava L. by artificial neural network and discriminant analysis. **Revista Brasileira de Fruticultura** 38: 151-157.

Charikar M, Chatziafratis V and Niazadeh R (2019) Hierarchical clustering better than average-linkage. In **Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms**. Society for Industrial and Applied Mathematics, p. 2291-2304.

CODEVASF – **Companhia de Desenvolvimento dos Vales do São Francisco e do Parnaíba** (2019) Notícias. Disponível em: <https://www.codevasf.gov.br/linhas -de-negocio/irrigacao/projetos-publicos-de-irrigacao/boletim-informativo-dos-projetos-da-codevasf/bip-22/resultados-economicos-dos-projetos-publicos-de-irrigacao-em-2019./> Acesso em: jun 2020.

Costa AM, Spehar CR and Sereno JR (2012) **Conservação de recursos genéticos no Brasil**. Embrapa, Brasília, 628p.

Corrêa AM, Teodoro PE, Gonçalves MC, Barroso LMA, Nascimento M, Santos A and Torres FE (2016) Artificial intelligence in the selection of common bean genotypes with high phenotypic stability. **Genetics and Molecular Research** 15: 1-7.

Cruz CD, Carvalho SP and Vencovsky R (1994). Estudo sobre divergência genética I. Fatores que afetam a predição do comportamento de híbridos. **Revista Ceres 41**: 178-182.

Embrapa - **Empresa Brasileira de Pesquisa Agropecuária** (2020) Notícias. Disponível em:<https://www.embrapa.br/busca-de-noticias/-/noticia/24428422/producao-nacional>. Acesso em: 05 mai. 2020.

Everitt BS, Landau S, Leese M and Stahl D (2011) **Cluster analysis**. Wiley, London, 346p.

Furman J and Seamans R (2019). AI and the Economy. **Innovation Policy and the Economy 19**: 161-191.

Goodfellow I, Bengio Y and Courville A (2016) **Deep learning**. MIT Press, Massachusetts, 775p.

Grira N, Crucianu M and Boujemaa N (2005) Unsupervised and semi-supervised clustering: a brief survey. **A review of machine learning techniques for processing multimedia content** 1: 9-16.

Gutiérrez S, Tardaguila J, Fernández-Novales J and Diago MP (2015). Support vector machine and artificial neural network models for the classification of grapevine varieties using a portable NIR spectrophotometer. **Plos One** 10: e0143197.

Haddad I, Hiller K, Frimmersdorf E, Benkert B, Schomburg D and Jahn D (2009). An emergent self-organizing map based analysis pipeline for comparative metabolome studies. **In silico biology** 9 (4): 163-178.

Harfouche AL, Jacobson DA, Kainer D, Romero JC, Harfouche AH, Mugnozza GS, Moshelion M, Tuskan GA, Keurentjes JJB and Altman A (2019) Accelerating climate

resilient plant breeding by applying next-generation artificial intelligence. **Trends in Biotechnology** 37: 1217-1235.

Haykin S (2008). **Neural Networks and Learning Machines**. Pearson, Canada, 906p.

Heaton J (2015) **Artificial Intelligence For Humans - Volume 3: Deep Learning and Neural Networks**. Heaton Research, USA, 374p.

Husain A (2017) **The sentient machine: The coming age of artificial intelligence**. Simon and Schuster, New York, 224p.

IBGE - **Instituto Brasileiro de Geografia e Estatística** (2020) Levantamento Sistemático da Produção Agrícola. Disponível em: <https://sidra.ibge.gov.br/tabela/1618>. Acesso em: 07 abr. 2020.

INRAE - **L'institut National de Recherche pour l'agriculture, l'alimentation et l'environnement**. Disponível em: <https://www6.inrae.fr/iggp/Resources/Grape - Genetic-Resources>. Acesso em: 07 abr. 2020.

Jaakkola H, Henno J, Mäkelä J and Thalheim B (2019). Artificial intelligence yesterday, today and tomorrow. In **2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics.** IEEE, Croatia, p. 860-867.

Keller M (2015) **The science of grapevines: Anatomy and Physiology**. Academic Press, Washington, 554p.

Khaki S, Khalilzadeh Z and Wang L (2020) Predicting Yield Performance of Parents in Plant Breeding: A Neural Collaborative Filtering Approach. **Plos One** 15: e0233382.

Kim H (2019) AI, big data, and robots for the evolution of biotechnology. **Genomics & Informatics** 17: e44.

Kist BB, Santos CE, Carvalho C and Beling RR (2020) **Anuário brasileiro de horti & fruti 2019**. Editora Gazeta Santa Cruz, Santa Cruz do Sul, 96p.

Kohonen T (1982) Self-organized formation of topologically correct feature maps. **Biological Cybernetics** 43: 59-69.

Kumar Y, Niwas R, Nimbal S and Dalal MS (2020) Hierarchical cluster analysis in barley genotypes to delineate genetic diversity. **Electronic Journal of Plant Breeding 11** (3): 742-748.

Leão PCS (2008) **Recursos Genéticos de Videira (*Vitis* spp.): análise da diversidade e caracterização da coleção de germoplasma da Embrapa Semiárido.** Tese de Doutorado**,** 126f. Universidade Federal de Viçosa. Genética e Melhoramento de Plantas, Viçosa-MG.

LeCun Y, Bengio Y and Hinton G (2015) Deep learning. **Nature** 521: 436-444.

Liakos KG, Busato P, Moshou D, Pearson S and Bochtis D (2018) Machine learning in agriculture: a review. **Sensors** 18: 1-29.

Lu LM, Ickert-Bond S and Wen J (2018) Recent advances in systematics and evolution of grape family Vitaceae. **Journal of Systematics and Evolution** 56: 259–261.

Maia JDG, Ritschel PS and Lazzarotto JJ (2018) A viticultura de mesa no Brasil: produção para o mercado nacional e internacional. **Territoires du Vin** 9: 1-9. Disponível em: < https://www.alice.cnptia.embrapa.br/alice/bitstream/doc/1103185/1/AViticulturadeMesanoBrasil.pdf>. Acesso em: 07 abr. 2020.

Machado LC, Oliveira VC, Paraventi MD, Cardoso RN, Martins DS and Ambrósio CE (2016) Maintenance of brazilian biodiversity by germplasm bank. **Pesquisa Veterinária Brasileira** 36: 62-66.

Mancuso S (2001) Clustering of grapevine (*Vitis vinifera* L.) genotypes with Kohonen neural networks. **Vitis-Geilweilerhof** 40: 59-64.

Marques P, Pádua L, Adão T, Hruška J, Sousa J, Peres E, Sousa JJ, Morais R and Sousa A (2019) Grapevine Varieties Classification Using Machine Learning. **EPIA Conference on Artificial Intelligence**. Springer, Cham, p.186-199.

Mei X, Lee HC, Diao K, Huang M, Lin B, Liu C, Xie Z, Ma Y, Robson PM, Chung M, Bernheim A, Mani V, Calcagno C, Li K, Li S, Shan H, Lv J, Zhao T, Xia J, Long Q, Steinberger S, Jacobi A, Deyer T, Luksza M, Liu F, Little BP, Fayad ZA and Yang Y (2020) Artificial intelligence for rapid identification of the coronavirus disease 2019 (COVID-19). **medRxiv**: 1-28.

McGovern P, Jalabadze M, Batiuk S, Callahan MP, Smith KE, Hall GR, Kvavadze E, Maghradze D, Rusishvili N, Bouby L, Failla O, Cola G, Mariani L, Boaretto E, Bacilieri R, This P, Wales N and Lordkipanidze D (2017) Early neolithic wine of Georgia in the South Caucasus. **Proceedings of the National Academy of Sciences** 114: e10309-e10318.

Mohammadi SA and Prasanna BM (2003) Analysis of genetic diversity in crop plants - salient statistical tools and considerations. **Crop Science** 43: 1235-1248.

MAPA – **Ministério da Agricultura, Pecuária e Abastecimento** (2020) Embrapa apresenta primeira cultivar de uva 100% nordestina. Disponível em: https://www.gov.br/agricultura/pt-br/assuntos/noticias/embrapa-apresenta-primeira-cultivar-de-uva-100-nordestina-1. Acesso em: 02 fev 2021.

Odong TL, Van Heerwaarden J, Jansen J, van Hintum TJ and Van Eeuwijk FA (2011) Determination of genetic structure of germplasm collections: are traditional hierarchical clustering methods appropriate for molecular marker data. **Theoretical and Applied Genetics** 123: 195-205.

Oliveira MS, Dos Santos IG and Cruz CD (2020) Self-organizing maps: a powerful tool for capturing genetic diversity patterns of populations. **Euphytica** 216: 1-9.

OIV - **Organização Internacional da Vinha e do Vinho (Office Internationale de la Vigne et du Vin)** 2018. Disponível em: <http://www.oiv.int/>. Acesso em: Fev. 2020.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M and Duchesnay E (2011) Scikit-learn: Machine Learning in Python**. Journal of Machine Learning Research** 12: 2825-2830.

Polak P, Nelischer C, Guo H and Robertson DC (2019) “Intelligent” finance and treasury management: what we can expect. **AI & Society**: 1-12.

Poyraz I (2016) Comparison of ITS, RAPD and ISSR from DNA-based genetic diversity techniques. **Comptes Rendus Biologies** 339: 171-178.

Radmann EB and Bianchi VJ (2008) Uva: da antiguidade a mesa de nossos dias. In Barbieri RL and Stumpf ER (eds) **Origem e evolução de plantas cultivadas**. Embrapa, Brasília, p. 891-909.

Rahmanifard H and Plaksina T (2018) Application of artificial intelligence techniques in the petroleum industry: a review. **Artificial Intelligence Review** 52: 2295-2318.

Riaz S, De Lorenzis G, Velasco D, Koehmstedt A, Maghradze D, Bobokashvili Z, Musayev M, Zdunic G, Laucou V, Walker MA, Failla O, Preece JE, Aradhya M, Arroyo-Garcia R (2018) Genetic diversity analysis of cultivated and wild grapevine (Vitis vinifera L.) accessions around the Mediterranean basin and Central Asia. **BMC Plant Biology** 18: 1-14.

Rimet F, Druart JC and Anneville O (2009) Exploring the dynamics of plankton diatom communities in Lake Geneva using emergent self-organizing maps (1974–2007). **Ecological informatics** 4 (2): 99-110.

Robeva R and Macauley M (2018) **Algebraic and Combinatorial Computational Biology**. Academic Press, London, 419p.

Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological Review** 65: 386.

Russell SJ and Norvig P (2016) **Artificial intelligence: a modern approach**. Prentice Hall, New Jersey, 1132p.

Salehnia N, Ansari H, Kolsoumi S and Bannayan M (2019) Climate data clustering effects on arid and semi-arid rainfed wheat yield: a comparison of artificial intelligence and K-means approaches. **International Journal of Biometeorology 63**: 861-872.

Santos IGD, Carneiro VQ, Silva Junior ACD, Cruz CD and Soares PC (2019) Self-organizing maps in the study of genetic diversity among irrigated rice genotypes. **Acta Scientiarum. Agronomy** 41: 1-9.

Sharma P and Singh A (2017) Era of deep neural networks: A review. In **2017 8th International Conference on Computing, Communication and Networking Technologies**. IEEE, p. 1-5.

Shi Y (2019) The Impact of Artificial Intelligence on the Accounting Industry. In **The International Conference on Cyber Security Intelligence and Analytics**. Springer, Cham, p. 971-978..

Smith B and Shum H (2018) **The future computed: artificial Intelligence and its role in society**. Microsoft Corporation, Redmond, 143p.

Song X, Yang S, Huang Z and Huang T (2019) The Application of Artificial Intelligence in Electronic Commerce. **Journal of Physics: Conference Series** 302: 1-7.

Tch A (2017) **The mostly complete chart of neural networks, explained**. Towards Data Science. Disponível em: < https://towardsdatascience.com/the-mostly-complete-chart-of-neural-networks-explained-3fb6f2367464>. Acesso em: 17 out. 2020.

This P, Lacombe T and Thomas MR (2006) Historical origins and genetic diversity of wine grapes. **Trends in Genetics** 22**:** 511-519.

Ultsch A (1995) Self-Organizing Neural Networks perform different from statistical k-means clustering. **Gesellschaft f. Klassifikation**: 1-13.

Ultsch A (1999) Data mining and knowledge discovery with emergent self-organizing feature maps for multivariate time series. In **Kohonen Maps**. Elsevier Science, Amsterdam, 400p.

Ultsch A (2007) Emergence in self organizing feature maps. In **International Workshop on Self-Organizing Maps: Proceedings**. Data Bionics Research Group of University of Marburg, p.7.

Ultsch A and Kampf D (2004) Knowledge discovery in DNA microarray data of cancer patients with emergent self organizing maps. In: **ESANN**. Belgium, Bruges, p. 501-506.

Ultsch A and Mörchen F (2005) ESOM-Maps: Tools for clustering, visualization, and classification with emergent SOM. **Technical Report - Data Bionics Research Group of University of Marburg**: 1-7.

Usama M, Qadir J, Raza A, Arif H, Alvin Yau K, Elkhatib Y, Hussain A and Al-fuqahausama A (2019) Unsupervised machine learning for networking: Techniques, applications and research challenges. **IEEE Access 7**: 65579-65615.

USDA - **United States Department of Agriculture** (2020) The PLANTS Database. Disponível em: <https://plants.usda.gov/core/profile?symbol=VITIS>. Acesso em: 03 abr. 2020.

Uzal LC, Grinblat GL, Namías R, Larese MG, Bianchi JS, Morandi EN and Granitto P M (2018) Seed-per-pod estimation for plant breeding using deep learning. **Computers and Electronics in Agriculture** 150: 196-204.

Van Veen F (2017) **Neural Network Zoo Prequel: Cells and Layers**. The Asimov Institute. Disponível em: <https://www.asimovinstitute.org/author/fjodorvanveen/>. Acesso em 17 out. 2020.

Vincent-Lancrin S and van der Vlies R (2020) Trustworthy artificial intelligence (AI) in education: Promises and challenges. **OECD Education Working Papers** 218: 1-18.

Walker MA, Heinitz C, Riaz S and Uretsky J (2019) Grape Taxonomy and Germplasm. In **The Grape Genome**. Springer, Cham, p. 25-38.

Wang X and Xu Y (2019). An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. In **IOP Conference Series: Materials Science and Engineering**. IOP Publishing, p. 5.

Warwick K (2013) **Artificial intelligence: the basics**. Routledge, London, 175p.

Wen J, Lu LM, Nie ZL, Liu XQ, Zhang N, Ickert-Bond S, Gerrath J, Manchester SR, Boggan J and Chen ZD (2018) A new phylogenetic tribal classification of the grape family (Vitaceae). **Journal of Systematics and Evolution** 56: 262-272.

Winkler-Schwartz A, Bissonnette V, Mirchi N, Ponnudurai N, Yilmaz R, Ledwos N, Siyar S, Azarnoush H, Karlik B and Del Maestro RF (2019) Artificial intelligence in medical education: best practices using machine learning to assess surgical expertise in virtual reality simulation. **Journal of Surgical Education** 76: 1681-1690.

Xu R and Wunsch D (2005) Survey of clustering algorithms. **IEEE Transactions on Neural Networks** 16: 645-678.

Yue Q, Zhang C, Wang Q, Wang W, Wang J and Wu Y (2019) Analysis on genetic diversity of 51 Grape germplasm resources. **Ciência Rural** 49: 1-10.

Zhou Y, Muyle A, Gaut BS (2019) Evolutionary Genomics and the Domestication of Grapes. In **The Grape Genome**. Springer, Cham, p. 39-55.

CAPÍTULO II

EMERGENT SELF-ORGANIZING MAPS APLICADOS AO ESTUDO DA DIVERSIDADE GENÉTICA DE ACESSOS DE UVA DE MESA DO BANCO DE GERMOPLASMA DE VIDEIRA (*Vitis* spp.) DA EMBRAPA SEMIÁRIDO

1. INTRODUÇÃO

As uvas alcançaram, em 2016, o valor bruto de produção mundial de quase US$ 68 bilhões, sendo considerada uma das culturas frutíferas de maior importância comercial do mundo (FAO 2020). Nos últimos 19 anos, as exportações brasileiras de uvas de mesa obtiveram a expressiva marca de 909,9% de crescimento, devido à expansão da produção no Submédio do Vale do São Francisco. Cerca de 10% da produção nesta região é destinada ao mercado internacional, concentrando-se principalmente no segundo semestre, pois é um período de entressafra global das uvas de mesa e os preços pagos pelas frutas atingem valores mais elevados (Maia et al. 2018). Em termos de qualidade dos frutos destinados ao consumo *in natura*, características como apirenia (ausência de sementes), aparência do cacho, sabor da baga (neutro, moscatel e foxado), baixo desgrane, consistência e textura da polpa, e resistência pós-colheita, são as de maior interesse (Leão and Borges 2009).

O crescimento da população e da renda *per capita* provocam o aumento da demanda por uvas e por todos os seus produtos derivados. Cultivares com maior rendimento, mais resilientes à pragas, doenças e estresses ambientais, que apresentem características agronômicas desejáveis para o mercado, são procuradas pelos produtores. O Banco de Germoplasma de Videira da Embrapa Semiárido se destaca por ser o único da região Nordeste do país, constituindo-se em um recurso estratégico fundamental para a sustentabilidade da vitivinicultura tropical no Submédio do Vale do São Francisco. O desafio do melhoramento genético da videira é encontrar maneiras de introduzir efetivamente novos caracteres desejáveis sem abrir mão de muitas características já existentes que os produtores também valorizam, de modo que seja rentável adotar novas cultivares (Leão and Borges 2009; Alston and Sambucci 2019).

Os cruzamentos biparentais são os mais utilizados em *Vitis vinifera* e servem como base para os processos de seleção, cuja finalidade é a obtenção de híbridos planejados através de polinização controlada, que apresentem grande potencial para posterior seleção de clones superiores. O sucesso dos cruzamentos biparentais está relacionado, entre outros fatores, à escolha dos genitores (Leão and Borges 2009).

Os métodos preditivos baseados em características morfológicas, agronômicas, fisiológicas ou genéticas dos genitores, determinadas antes dos cruzamentos, podem ajudar os melhoristas a concentrarem seus esforços em

combinações promissoras (Cruz et al. 2006). A heterose, expressa em híbridos, está diretamente relacionada à diversidade genética entre seus genitores (Falconer 1989).

No melhoramento genético, os estudos de diversidade genética são muito importantes, pois permitem a diferenciação de acessos e auxiliam na identificação de genótipos contrastantes para a realização de cruzamentos promissores. Por meio de métodos clássicos, ferramentas biotecnológicas ou novas abordagens, é gerada uma população com variabilidade genética na qual a seleção pode ser praticada a fim de se obter um ou mais indivíduos que reúnam caracteres agronômicos de interesse (Leão and Borges 2009; Campos et al. 2016).

A diversidade genética pode ser estimada através de várias técnicas estatísticas e a escolha do método mais adequado será determinada de acordo com os objetivos do pesquisador, pela facilidade da análise e pela forma como os dados foram obtidos. Métodos multivariados, como a análise de componentes principais, variáveis canônicas e métodos de agrupamento (k-means, DBSCAN, agrupamento hierárquico da soma de quadrados do erro de Ward, entre outros) tem sido bastante utilizados em estudos de diversidade genética (Miranda et al. 1988; Cruz 1990; Cruz et al. 1994; Silva et al. 2007; Murtagh et al. 2011).

Outras abordagens para o estudo da diversidade genética também tem sido adotadas, como a utilização de *self-organizing maps* (SOM), que são um tipo de rede neural artificial bidimensional que organiza dados de um processo de aprendizagem não supervisionado por meio da noção de vizinhança e usando a distância euclidiana. O aprendizado começa com a atribuição de pesos sinápticos. Um processo de competição tem início e cada amostra de dados é alocada para o neurônio que melhor a representa, denominado *Best Matching Unit* (BMU) ou neurônio vencedor (Santos et al. 2019).

Os SOM são uma ferramenta de visualização que preserva a topologia dos dados originais, fornecendo uma representação visual de grupos de instâncias de dados semelhantes através da geração de mapas. Eles apresentam grande complexidade, pois são compostos por vários elementos interconectados: neurônios e vizinhanças (Kohonen 2012, Ultsch and Mörchen 2005). Seu uso como ferramenta para o estudo da diversidade genética tem se tornado cada vez mais frequente (Barbosa et al. 2011; Sá 2015, Campos et al. 2016; Santos et al. 2019; Oliveira et al. 2020).

Emergent self-organizing maps (ESOM) são um tipo de SOM que possuem como principal característica um grande número de neurônios em relação a uma rede SOM comum. Esse grande número de neurônios, que pode variar de 1.000 a 1.000.000, é capaz de produzir o fenômeno da “emergência”. Emergência é a capacidade de um sistema de produzir um fenômeno em um nível novo e superior. Através da cooperação desse número maior de neurônios, a captura da topologia do espaço original (dados) ocorre com maior acurácia, já que esse grande número de neurônios pode representar clusters de dados individualmente, o que facilita sua detecção (Ultsch 2007, Ultsch and Kampf 2004). Até o presente momento, não existem relatos na literatura sobre a utilização dos ESOM para estudos de diversidade genética em culturas agrícolas.

Sendo assim, o objetivo deste estudo é apresentar a técnica dos ESOM como método alternativo para o estudo da diversidade genética em programas de melhoramento genético de plantas, através da sua aplicação à análise da diversidade genética de acessos de uva de mesa do Banco de Germoplasma de Videira da Embrapa Semiárido.

2. MATERIAL E MÉTODOS

2.1. Localização e Manejo do Banco de Germoplasma de Videira

O Banco de Germoplasma de Videira da Embrapa Semiárido está localizado no Campo Experimental de Mandacaru, Juazeiro-BA (9 ° 24 ”S, 40 ° 26” O e 365,5 m de altitude), conforme pode ser visualizado na Figura 5:

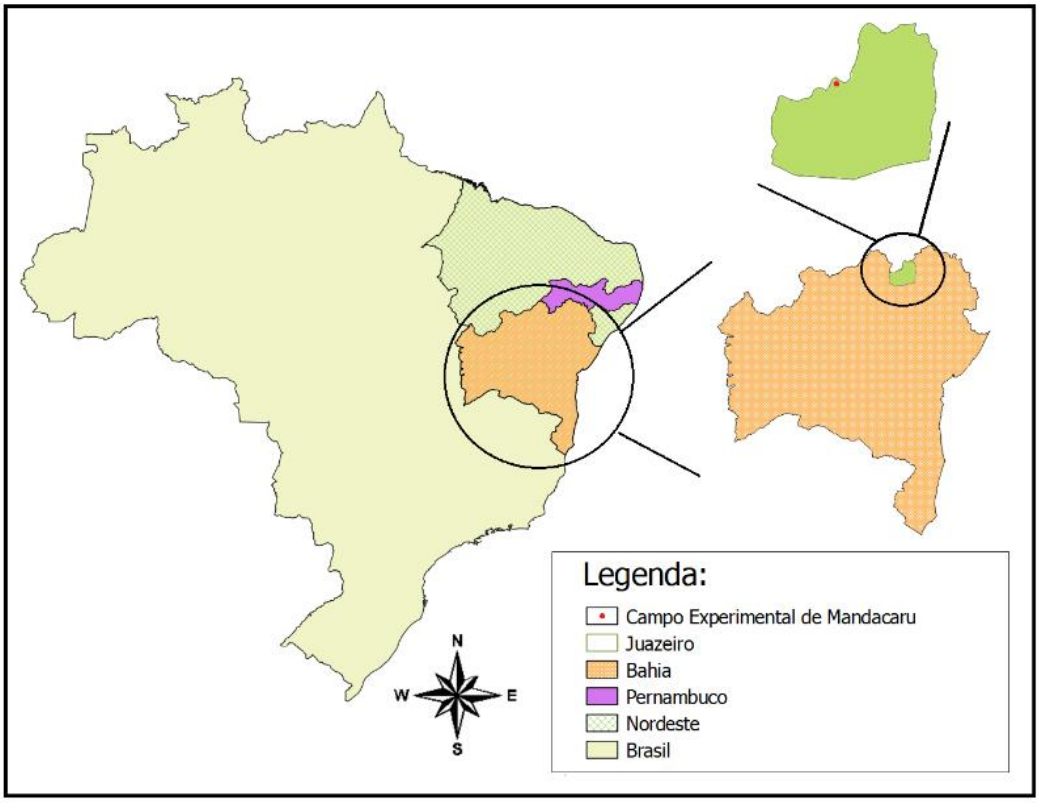


Figura 5. Localização do Banco de Germoplasma de Videira da Embrapa Semiárido, na Estação Experimental de Mandacaru em Juazeiro, BA.

As plantas são conduzidas em sistema de espaldeira, utilizando um espaçamento de 3,0 metros entre filas de plantio e 2,0 metros entre plantas, com irrigação localizada por gotejamento. Cada acesso é composto por quatro plantas formadas em cordão bilateral que são podadas em duas épocas alternadas ao longo do ano, sendo uma poda curta no 1º semestre e uma poda média com 6 a 8 gemas no 2º semestre. O manejo da copa consiste na poda, desbrota e desponte de ramos, além do controle de plantas daninhas com aplicação de herbicida e roço mecanizado e controle químico de pragas e doenças.

O estudo foi realizado durante os anos de 2018 e 2019, utilizando as médias de quatro ciclos de produção (2018.1, 2018.2, 2019.1 e 2019.2) de 93 acessos de uva de mesa (Tabela 1).

Tabela. Códigos de identificação dos 93 acessos de uva de mesa do Banco de Germoplasma de Videira da Embrapa Semiárido utilizados para a análise de diversidade genética.

Genótipo	Código de Identificação	Genótipo	Código de Identificação	Genótipo	Código de Identificação	Genótipo	Código de Identificação
A1105	0	**Christmas Rose**	24	**IAC Madalena**	47	**Moscatel de Jundiaí**	70
A1118	1	**CNPUV 8**	25	**Igawa**	48	**Moscatel Grega**	71
A1581	2	**Concord Clone**	26	**Impero**	49	**Moscatel Nazareno**	72
Ângelo Pirovano	3	**Crimson Seedless**	27	**Isaura**	50	**Muscat Noir**	73
August Giant	4	**Dattier de Beirut**	28	**Itália**	51	**Neo Muscat**	74
Aurora	5	**Dawn Seedless**	29	**Itália Clone I**	52	**Niágara Rosada**	75
Baresana	6	**Don Mariano**	30	**Itália Melhorada**	53	**Patrícia**	76
Benifuji	7	**Dona Maria**	31	**Júpiter**	54	**Perlette**	77
Benitaka	8	**Early Muscat**	32	**Kagina**	55	**Perlona**	78
Blue Lake	9	**Emerald Sedless**	33	**Kyoho**	56	**Piratininga**	79
Branca Salitre	10	**Estevão Marinho**	34	**Lake Emerald**	57	**Portuguesa Blanes**	80
Brasil	11	**Eumelan**	35	**Lakemont Seedless**	58	**Queen**	81
Bronx Seedless	12	**Feal**	36	**Liberty**	59	**Red Globe**	82
BRS Clara	13	**Ferral**	37	**Madeleine Royal**	60	**Regina Roma**	83
BRS Isis	14	**Fiesta**	38	**Marengo**	61	**Reliance**	84
BRS Linda	15	**Flame Seedless**	39	**Maria**	62	**Rosaky Rosada**	85
BRS Morena	16	**Flame Tokay**	40	**Marroo Seedless**	63	**Ruby Seedless**	86
BRS Vitória	17	**Frakenthal**	41	**Michele Paglieri**	64	**Seyve Villard 12327**	87
Califórnia	18	**Golden Queen**	42	**Monte Serrat**	65	**Soraya**	88
Ceilad	19	**H 449100**	43	**Moscatel Branca**	66	**Sovrana Pirovano**	89
Centenial Seedless	20	**Himoront**	44	**Moscatel Caillaba**	67	**Tardia de Caxias**	90
CG 351	21	**IAC 13822**	45	**Moscatel de Alexandria**	68	**Thompson Seedless**	91
CG 38049	22	**IAC 77526**	46	**Moscatel de Hamburgo**	69	**Vênus**	92
CG 4113	23						

2.2. Variáveis agronômicas analisadas

Foram analisadas as seguintes variáveis agronômicas quantitativas:

a. Produção: avaliada na colheita por meio da pesagem de todos os cachos colhidos utilizando em balança eletrônica digital, expressa em quilogramas (kg);

b. Número de cachos: obtido pela contagem de todos os cachos;

c. Peso dos cachos: determinado pela divisão do peso total dos cachos pelo número de cachos por videira, expresso em gramas (g);

d. Comprimento e largura do cacho: medidos em uma amostra de cinco cachos por acesso, por meio de régua e expressos em centímetros (cm);

e. Peso da baga: determinado numa amostra de 10 bagas colhidas em cada cacho, num total de 50 bagas por acesso, através de balança eletrônica digital, expresso em gramas (g);

f. Comprimento e diâmetro das bagas: avaliados na mesma amostra de bagas do item anterior, por meio de régua, expressos em milímetros (mm);

g. Teor de sólidos solúveis totais: leituras feitas no mosto extraído de 50 bagas por parcela, em refratômetro digital com ajuste automático de temperatura (ATAGO, Digital Pocket Refractometer, modelo PAL-1) e expresso em ° Brix;

h. Acidez titulável: através da diluição de 5 ml de polpa de uva em 50 ml de água destilada juntamente com solução de NaOH 0,1 N utilizando um titulador automático, marca Metrohm (modelo 848 Titrino plus) (AOAC, 2010), e os resultados foram expressos em g de ácido tartárico a 100 mL -1;

i. Relação teor de sólidos solúveis totais/acidez titulável.

2.3. *Emergent self-organizing maps*

Para estudar a diversidade genética de 93 acessos de uva de mesa do Banco de Germoplasma de Videira da Embrapa Semiárido, foram gerados ESOM de configuração matricial de 16.000 neurônios (100 linhas x 160 colunas). Para a geração dos ESOM, os dados foram padronizados e foi utilizado o pacote Somoclu (Wittek et al. 2013) em linguagem de programação *Python*.

O agrupamento formado pela ESOM pode ser executado da seguinte forma: os *Best Matching Units* (BMU) são identificados e seus pontos de dados correspondentes podem ser agrupados manualmente em vários grupos. A associação dos grupos pode ser visualizada pela coloração dos BMU. Ou seja, os BMU que possuem cores iguais pertencem ao mesmo grupo (Ultsch and Mörchen, 2005).

Foram gerados onze *boxplots* afim de facilitar a discussão ao permitir a visualização dos valores médios, máximos, mínimos para cada variável analisada, bem como os *outliers* de cada grupo.

Também foram gerados os mapas de variabilidade genética, que permitem visualizar a variabilidade genética existente entre os grupos formados para cada variável analisada.

Por fim, foi gerada a matriz ESOM de similaridade genética, que indica quais os cruzamentos mais promissores entre os grupos heteróticos formados pela rede ESOM com base na distância Euclidiana entre os grupos.

3. RESULTADOS E DISCUSSÃO

3.1. Agrupamento da rede ESOM

O mapa de agrupamento pode ser interpretado da seguinte forma: cada ponto colorido corresponde a um genótipo que possui sua respectiva identificação numérica. Pontos da mesma cor representam os genótipos que pertencem ao mesmo grupo genético, ou seja, genótipos similares geneticamente de acordo com as variáveis analisadas. Dessa forma, é possível identificar os grupos formados pela rede ESOM, conforme Figura 6.

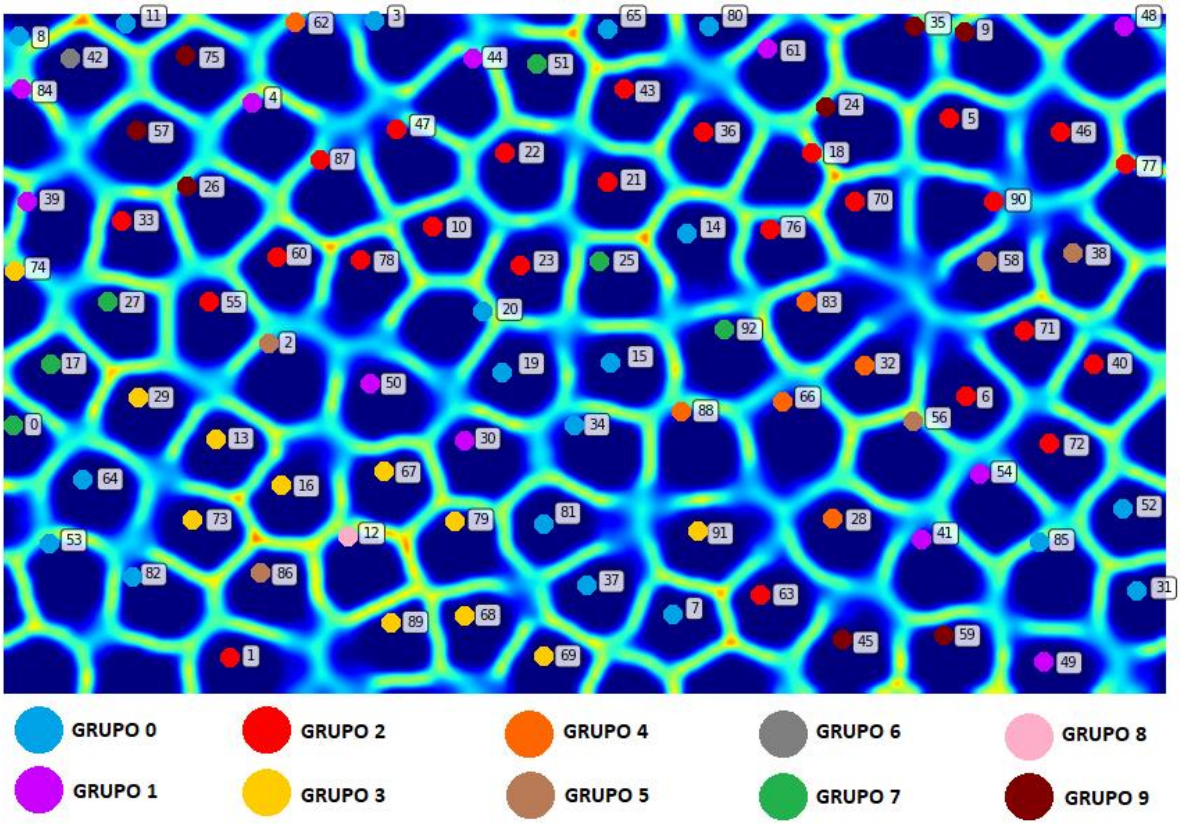


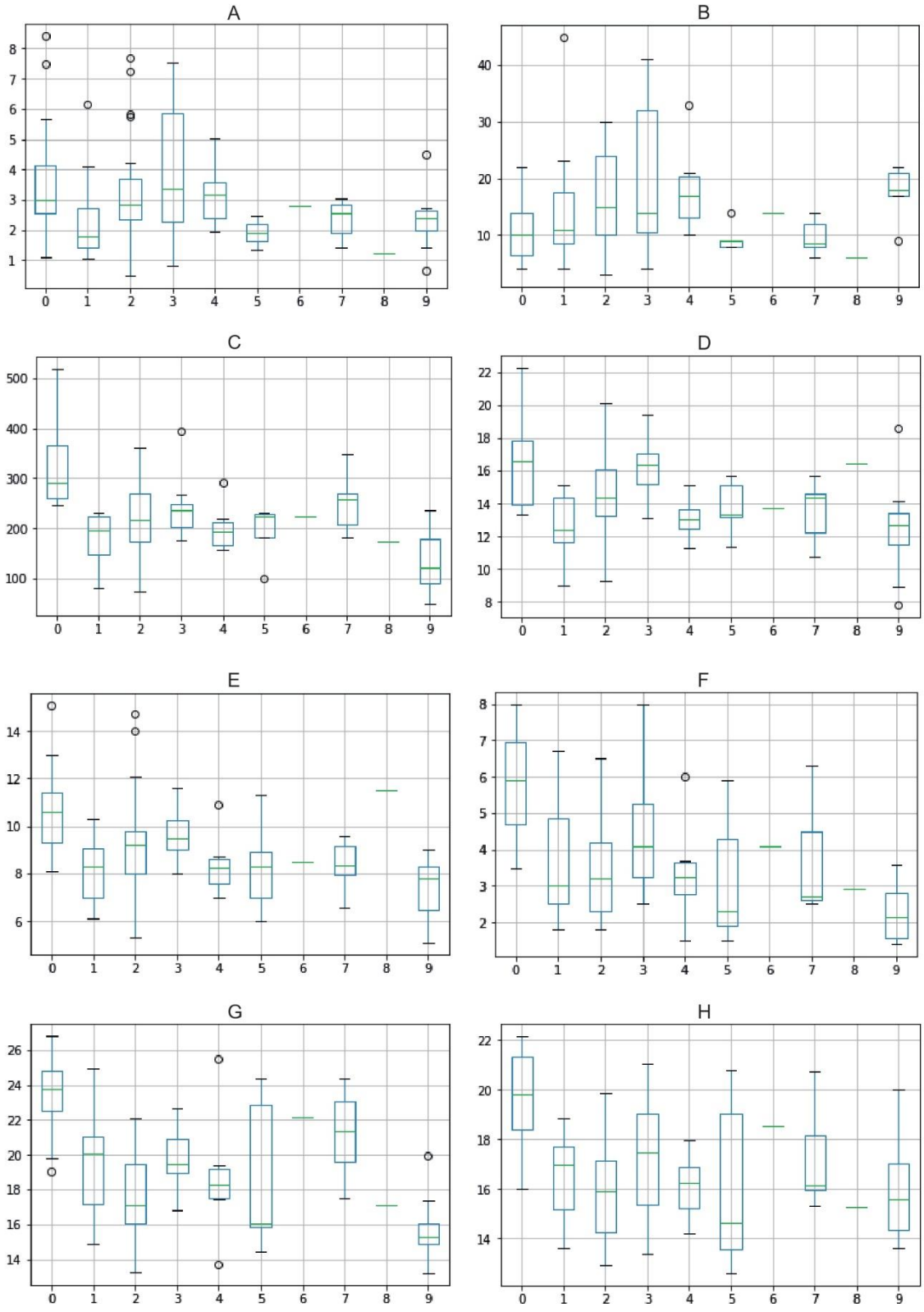
Figura 6. *Emergent self-organizing map* do agrupamento realizado pela rede neural.

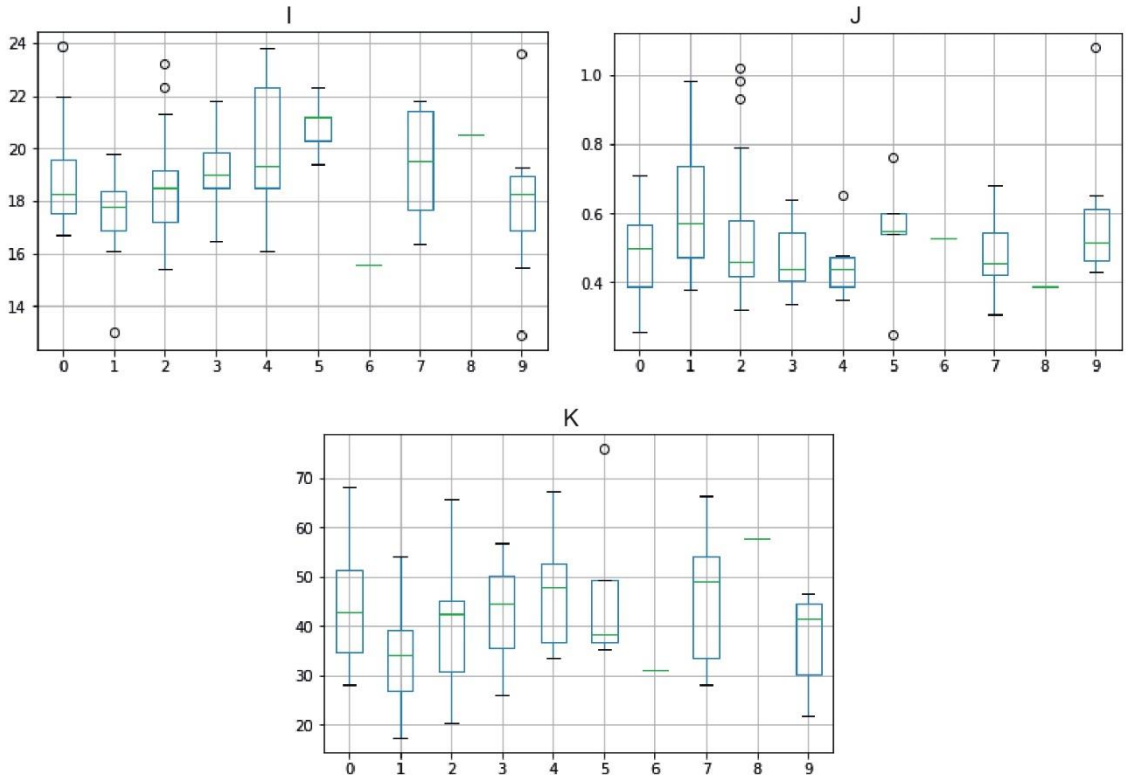
A porcentagem de acessos alocados em cada grupo foi: 20,43 % no grupo 0; 11,82 % no grupo 1; 26,88 % no grupo 2; 11,82 % no grupo 3; 6,45 % no grupo 4; 5,37 % no grupo 5; 1,07 % no grupo 6; 6,45 % no grupo 7; 1,07 % no grupo 8 e 8,60 % no grupo 9.

Leão (2008), ao analisar a diversidade genética em 136 acessos de uva de mesa pertencentes ao BAG da Embrapa Semiárido através da análise de componentes principais, concluiu que não foi possível identificar uma tendência na formação dos grupos, tais como uma característica comum, genealogia ou origem geográfica. Ainda segundo Leão (2008), alguns grupos formados agruparam cultivares tão distintas quanto aquelas de diferentes origens geográficas; espécies, tais como *Vitis vinifera*, *Vitis labrusca* e híbridos interespecíficos; cultivares com sementes e sem sementes. Resultados semelhantes foram obtidos no presente estudo. Apesar da predominância de cultivares sem sementes e com sementes em alguns grupos, como no grupo 7, onde apenas a cultivar Itália possui sementes, e do grupo 9 ser formado apenas por cultivares com sementes, os demais grupos formados por mais de um genótipo possuem cultivares com e sem sementes.

3.2. *Boxplots* dos grupos formados pela rede ESOM

Os valores médios, máximos e mínimos para as variáveis estudadas em cada grupo, bem como os *outliers*, podem ser visualizados nos *boxplots* A ao K.





*Boxplot*s. Produção em kg/planta (A), número de cachos por planta (B), peso do cacho em gramas (C), comprimento do cacho em centímetros (D), largura do cacho em centímetros (E), peso da baga em gramas (F), comprimento da baga em milímetros (G) diâmetro da baga em milímetros (H), teor de sólidos solúveis em °Brix (I), acidez titulável (J) e relação sólidos solúveis/acidez titulável (K) para os 10 grupos formados pelos *Emergent self-organizing maps*.

O grupo 3 apresenta o maior valor médio de produção por planta (3,97 kg/planta) e também número de cachos (20 cachos/planta). O grupo 0 apresenta os maiores valores máximos para as variáveis produção (8,44 kg), peso do cacho (520,19 g), comprimento do cacho (22,30 cm), largura do cacho (15,10 cm), peso da baga (8,00 g), comprimento da baga (26,84 mm), diâmetro da baga (22,17 mm) e sólidos solúveis (23,90 °Brix) e também apresenta os maiores valores médios para o peso do cacho (331,20 g), comprimento do cacho (16,6 cm), largura do cacho (10,5 cm), peso da baga (5,78 g), comprimento da baga (23,5 mm) e diâmetro da baga (19,7 mm). O grupo 4 apresentou o maior valor médio de sólidos solúveis (20,8 °Brix). O menor valor médio de produção (1,22 kg) e acidez titulável (0,39 g/ml) e o maior valor médio de sólidos solúveis / acidez titulável (57,6) pertence ao grupo 8, que é formado por apenas 1 acesso. O grupo 9 possui os menores valores médios para os caracteres peso do cacho (131,84 g), comprimento do cacho (12,52 cm), largura do cacho (7,30 cm), peso da baga (2,27 g) e comprimento da baga (15,74 cm). O grupo 6 possui o

menor valor médio de sólidos solúveis (15,6 °Brix) e sólidos solúveis / acidez titulável (31,13).

A formação de grupos entre os acessos de um banco de germoplasma é muito importante, pois permite ao melhorista a possibilidade de formação de grupos heteróticos. Um grupo heterótico pode ser definido como um grupo de genótipos, relacionados ou não, de uma mesma ou diferentes populações, que exibe capacidade de combinação e resposta heterótica quando cruzado com genótipos de outros grupos de germoplasma geneticamente distintos (Melchinger and Gumber 1998). De acordo com Campos et al. (2016), a formação de grupos heteróticos facilita a seleção de genótipos divergentes para o melhoramento por meio da geração de híbridos, visto que permite a seleção de genótipos adequados para cruzamentos entre diferentes grupos heteróticos, aumentando as chances de obtenção de genótipos superiores.

Os grupos heteróticos têm um forte impacto na melhoria das culturas que apresentam o fenômeno da heterose porque eles pré determinam o tipo de germoplasma que será usado em um programa de melhoramento genético para seleção de híbridos por um longo período de tempo (Melchinger and Gumber 1998). Leão et al. (2011) afirmam que, por ser uma espécie altamente heterozigótica, espera-se obter o efeito heterótico máximo nas gerações segregantes da videira a partir do cruzamento de pais divergentes.

Semelhantemente aos resultados obtidos por Leão (2008) houve a formação de um grupo composto por clones da cultivar Itália de bagas vermelhas como ‘Benitaka’ e ‘Brasil’, que possuem teores de sólidos solúveis e sabor moscatel mais acentuado que a cultivar Itália, bem como as cultivares ‘Itália Melhorada’ e ‘Itália Clone 1’, todas inseridas no grupo 0. Nesse grupo também predominam cultivares que se destacam pelo tamanho de suas bagas: ‘Dona Maria’, ‘Queen’ e ‘Red Globe’. Também foram inseridas no grupo 0 as cultivares ‘BRS Linda’ e ‘BRS Isis’, duas importantes cultivares comerciais desenvolvidas pela Embrapa.

O grupo 3 se destacou por possuir o maior valor para as variáveis produção e número de cachos, sendo composto também por duas cultivares comerciais desenvolvidas pela Embrapa: ‘BRS Clara’ e ‘BRS Morena'.

O grupo 7 é composto por duas importantes cultivares comerciais que bem representam o passado e presente do melhoramento genético das uvas de mesa: a cultivar ‘Itália’, uma das primeiras cultivares comerciais do Submédio Vale do São

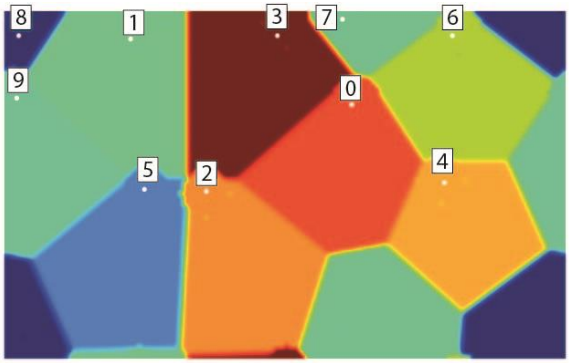
Francisco, e a cultivar ‘BRS Vitória’, uma das cultivares mais comercializadas no Submédio do Vale do São Francisco atualmente.

As cultivares ‘Aurora’ e ‘IAC 77526’, consideradas sinonímias, foram alocadas no grupo 2. Leão (2008) obteve resultado semelhante com relação a essas duas cultivares, porém, o mesmo resultado não foi obtido com relação a outras cultivares (‘Dattier de Saint Vallier’, ‘Seyve Villard 20365’, ‘Emperatriz’ e ‘CG 28467’), chegando a conclusão que a técnica multivariada de agrupamento utilizando variáveis morfoagronômicas não foi eficiente em agrupar genótipos idênticos. No presente estudo, não foi possível concluir se o agrupamento realizado pela rede ESOM foi eficiente ao agrupar sinonímias devido ao pequeno número de sinonímias utilizadas na a análise (apenas duas).

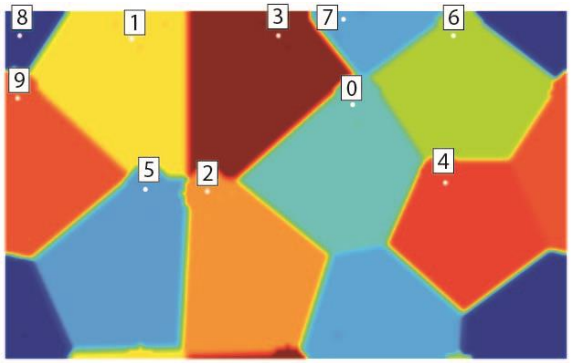
O agrupamento realizado pela rede ESOM foi capaz de descobrir padrões genéticos e diferenças entre os acessos de uva de mesa estudados, permitindo a formação de grupos heteróticos.

3.3. Mapas de variabilidade genética

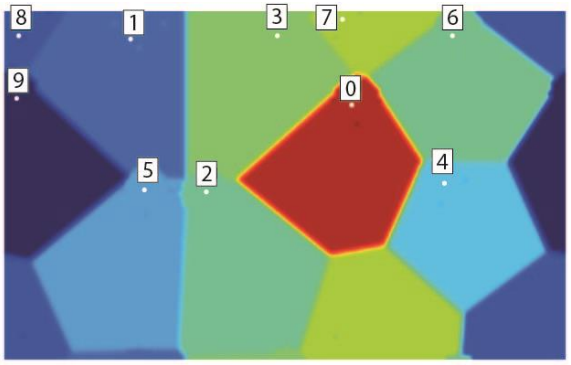
Os ESOM de variabilidade genética dos grupos formados também foram gerados e podem ser visualizados na Figura 7.



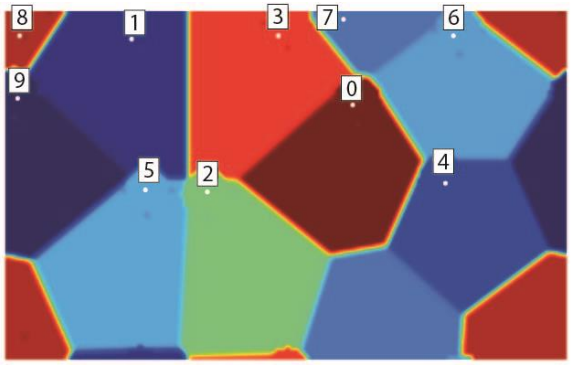
Mapa A



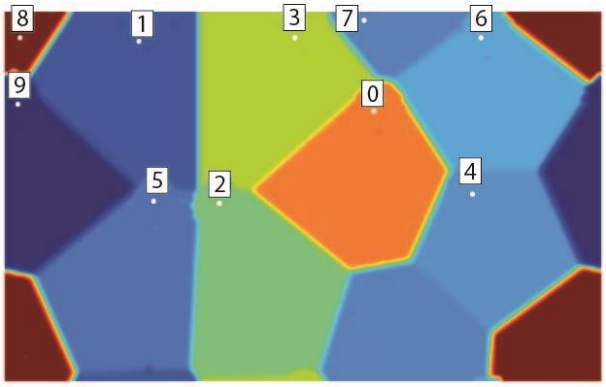
Mapa B



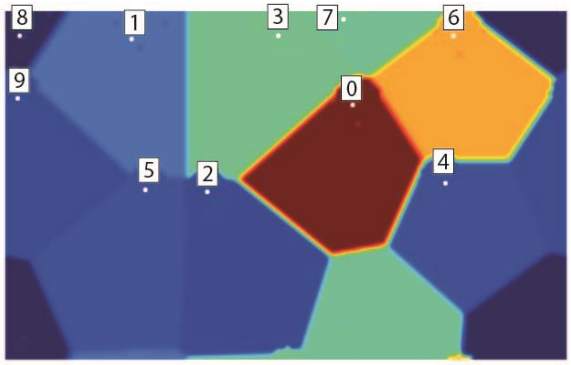
Mapa C



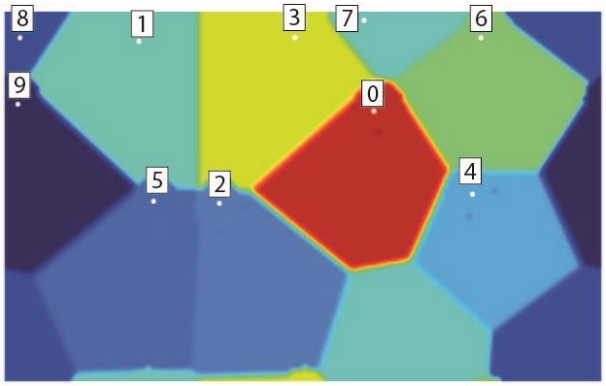
Mapa D



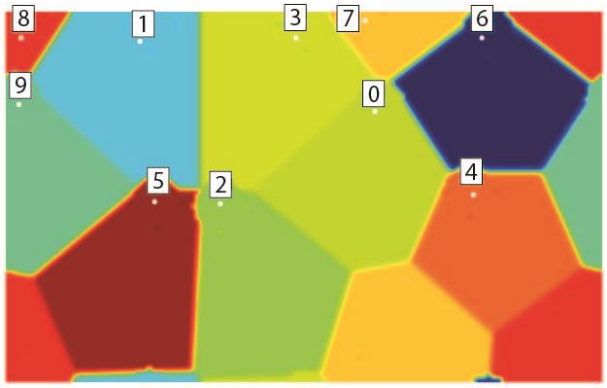
Mapa E



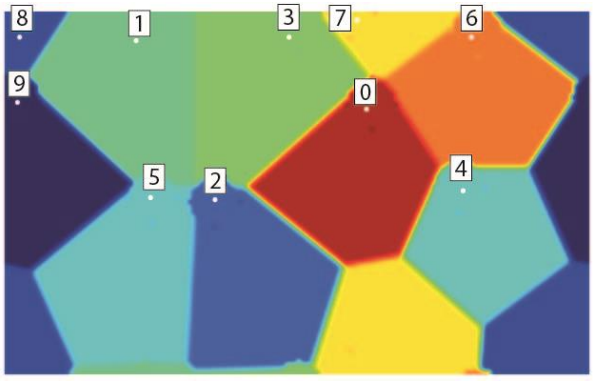
Mapa H



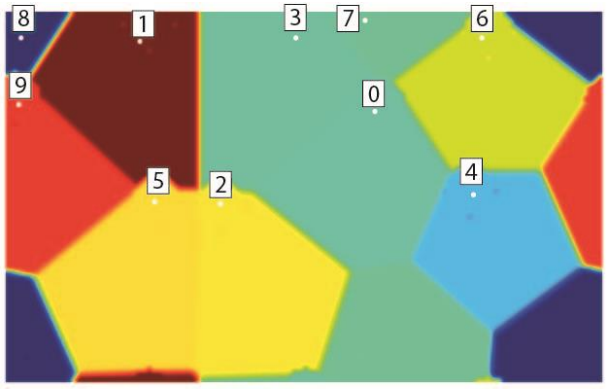
Mapa F



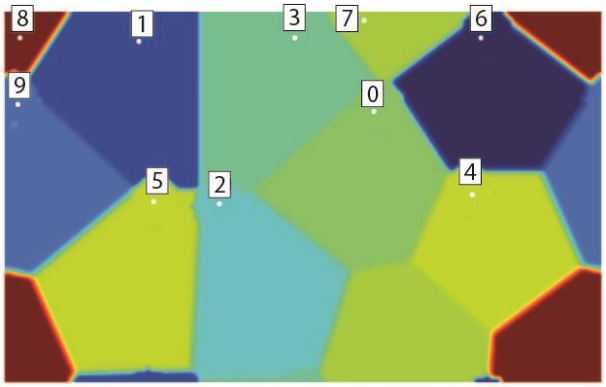
Mapa I



Mapa G



Mapa J



Mapa K

Figura 7. *Emergent self-organizing maps* das variáveis produção (A), número de cachos (B), peso do cacho (C) e comprimento do cacho (D), largura do cacho (E), peso da baga (F), diâmetro da baga (G), comprimento da baga (H), sólidos solúveis (I), acidez titulável (J) e relação sólidos solúveis / acidez titulável (K) para os 10 grupos heteróticos.

A interpretação dos mapas de variabilidade genética é bastante simples, sendo análoga à interpretação de um mapa de calor: o tom em vermelho mais escuro representa o valor de maior magnitude, enquanto o tom em azul mais escuro representa o valor de menor magnitude. Os valores de variação dentro desse intervalo são representados por cores intermediárias aos tons vermelho/azul escuro. Dentro das regiões coloridas, que representam os acessos, existem pontos de cor branca e seus respectivos rótulos numéricos de identificação. Como os mapas gerados têm a forma de um retângulo, deve-se perceber que o quadrilátero inferior é uma continuação do quadrilátero superior ou vice versa. O mesmo vale para o quadrilátero da direita e da esquerda.

As variáveis que possuem maior variabilidade são a produção, o número de cachos, o peso e o comprimento da baga, o teor de sólidos solúveis e a acidez titulável. Leão (2008), em seu estudo de diversidade genética deste mesmo BAG, também concluiu que o número de grupos formados e a distribuição satisfatória dos acessos nestes grupos possibilitou a identificação de genitores para a formação de populações segregantes com ampla base genética.

3.4. Matriz ESOM de similaridade genética

A matriz ESOM de similaridade genética (Figura 8) indica quais os cruzamentos mais promissores entre os 10 grupos heteróticos de uva de mesa com base na divergência genética.

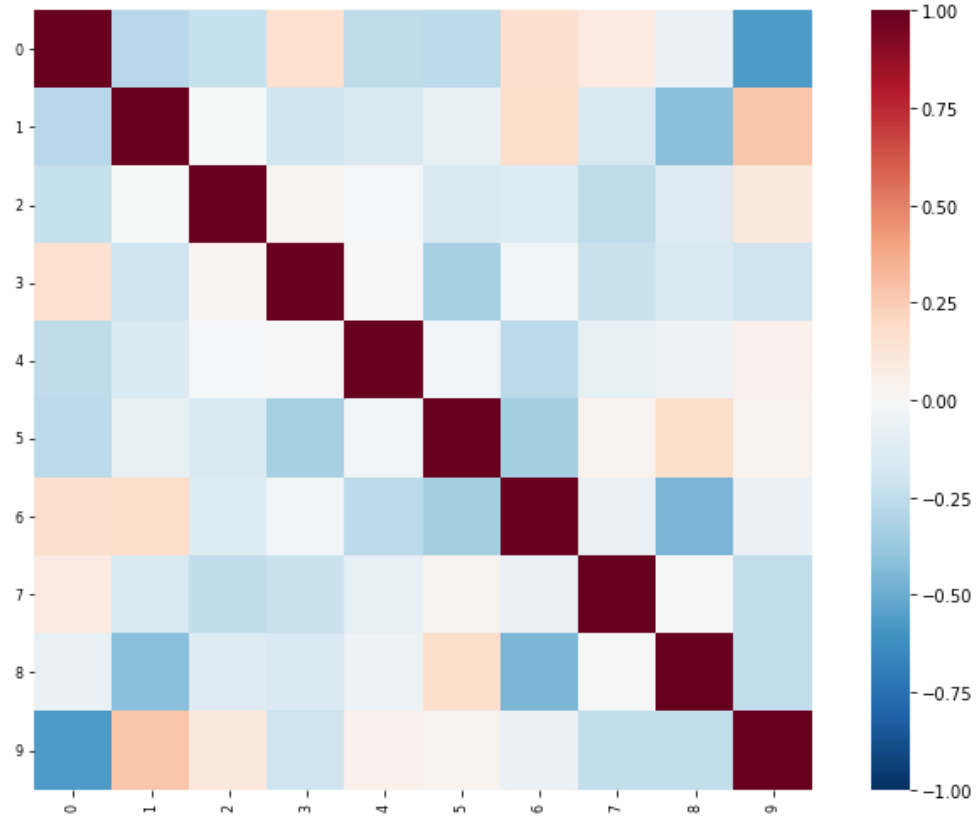


Figura 8. Matriz *Emergent self-organizing maps* de similaridade genética.

Visando a qualidade dos frutos destinados ao consumo *in natura*, características como apirenia (ausência de sementes), aparência do cacho, sabor da baga (neutro, moscatel, especial e foxado), baixo desgrane, consistência e textura da polpa, e resistência pós-colheita, são as de maior interesse (Leão and Borges, 2009), sendo consideradas características que norteiam os programas de melhoramento genético da videira. Entre essas características, a apirenia é de grande importância, pois as uvas sem sementes são geralmente as preferidas pelo consumidor para o consumo *in natura*, e essa demanda tem aumentado nos últimos anos (Pearl et al. 2003). Sendo assim, a produção de novas cultivares de uvas sem sementes e com bagas grandes, adequadas ao consumo de mesa, constitui um dos principais objetivos dos estudos de melhoramento de videira.

Segundo Leão (2008), a seleção de genótipos com base somente na divergência genética, sem considerar seus próprios desempenhos, não é uma boa estratégia em um programa de melhoramento. Dessa forma, a recomendação de cruzamentos em programas de melhoramento entre genótipos divergentes, mas que apresentem desempenho superior nas principais características de importância agronômica seria a mais indicada. O cruzamento de genótipos do grupo 0,

característico por possuir genótipos com bagas de tamanho grande com genótipos sem semente do grupo 1, são indicados. O genótipo ‘Flame Seedless’ pode ser utilizado como progenitor masculino e cruzado com o genótipo ‘Itália Melhorada’, utilizado como progenitor feminino. Também são indicados cruzamentos entre os genótipos ‘Reliance’ e ‘Júpiter’ (grupo 1), utilizados como progenitores masculinos, com os genótipos ‘Itália Clone 1’ e ‘Michele Paglieri’ (grupo 0), respectivamente, utilizados como progenitores femininos.

O grupo 0 também apresenta divergência genética em relação aos grupos 2, 4 e 5. Os genótipos sem sementes e de bagas pequenas ‘Emerald Seedless’, ‘CG351’ e ‘Marroo Seedless’ (grupo 2), podem ser usados como progenitores masculinos e cruzados com os genótipos ‘Dona Maria’ e ‘Monte Serrat’ (grupo 0), que seriam os progenitores femininos. Os genótipos sem sementes do grupo 5, ‘Lakemont Seedless’, ‘Ruby Seedless’ e ‘Fiesta’ (progenitores masculinos) podem ser cruzados com ‘Itália Melhorada’, ‘Dona Maria’ e ‘Michele Paglieri’ (progenitores femininos) do grupo 0.

O grupo 7, no qual predominam acessos de uvas sem sementes e no qual está inserida a cultivar comercial ‘BRS Vitória’ pode ser cruzada com os grupos 2, 3 e 9. Cruzamentos entre a ‘BRS Vitória’, progenitor masculino e ‘Moscatel de Hamburgo’ ou ‘Moscatel de Alexandria’ (grupo 3, progenitores femininos) são indicados, visto que pode-se aliar o sabor moscatel e maior tamanho de bagas a uma cultivar comercial de menor tamanho de bagas.

Os grupos 0 e 9 são os mais distantes geneticamente. Porém, considerando que o objetivo principal do programa de melhoramento genético para uvas de mesa é a obtenção de novas cultivares apirênicas, seus cruzamentos não são indicados. Todos os genótipos alocados no grupo 9 possuem sementes e bagas de tamanho pequeno.

4. CONCLUSÃO

O agrupamento realizado pela rede ESOM permitiu a formação de 10 grupos heteróticos. Não foi possível identificar uma tendência na formação desses grupos, como uma característica comum, genealogia ou origem geográfica. O grupo 0 se destacou por alocar genótipos com maiores valores médios para as variáveis peso do cacho, comprimento do cacho, largura do cacho, peso da baga, comprimento da baga

e diâmetro da baga. Alguns genótipos importantes fazem parte desse grupo, como a 'Itália Melhorada', 'Benitaka', 'Red Globe' e as cultivares comerciais 'BRS Linda' e 'BRS Isis', desenvolvidas pela Embrapa. Com exceção do genótipo 'Itália', todos os outros genótipos alocados no grupo 7 são apirênicos. Porém, existem genótipos com e sem sementes alocados em um mesmo grupo, indicando que o método da rede ESOM não conseguiu detectar, a partir das variáveis analisadas, genótipos pirênicos e apirênicos.

Os ESOM (ou mapas) de variabilidade genética evidenciaram a grande variabilidade genética presente nas variáveis analisadas entre os grupos formados, indicando uma distribuição satisfatória dos acessos nesses grupos. Uma das vantagens dos mapas de variabilidade consistem em permitir a visualização em 2D da variabilidade genética de uma população ou coleção em estudo a partir de variáveis quantitativas, que seriam mais difíceis de serem visualizadas através de uma tabela ou de uma lista gerada a partir das técnicas multivariadas tradicionais. Em estágios iniciais de programas de melhoramento genético de plantas, onde normalmente se dispõe de um grande número de genótipos, ou no caso de estudos com populações grandes, os mapas de variabilidade podem ser tornar uma ferramenta valiosa, indicando ao melhorista qual a base genética de sua população ou plantel: se é uma base genética ampla ou estreita.

A matriz ESOM de similaridade genética indicou os cruzamentos mais promissores baseados na divergência genética. Porém, esses cruzamentos devem ser orientados aliados ao desempenho superior de cada grupo para as principais variáveis de importância agronômica. O cruzamento de genótipos do grupo 0, composto por genótipos, em sua maioria, pirênicos e com bagas de tamanho grande, com genótipos dos grupo 1, 2, 4, 5 e 7, que sejam apirênicos e de bagas de tamanho menor são os mais indicados.

Cruzamentos entre genótipos dos grupos 0 e 9 não são indicados. Apesar de serem os grupos mais distantes geneticamente, todos os indivíduos do grupo 9 possuem um tamanho pequeno de baga e possuem sementes, características consideradas indesejáveis em um programa de melhoramento genético de uvas de mesa. Estudos futuros sobre a validação dos ESOM como um método de agrupamento eficiente no melhoramento genético de plantas são indicados.

5. REFERÊNCIAS BIBLIOGRÁFICAS

Alston JM, Sambucci O (2019) Grapes in the World Economy. In **The Grape Genome**. Springer, Cham, p. 1-24.

Barbosa CD, Viana AP, Quintal SSR and Pereira MG (2011) Artificial neural network analysis of genetic diversity in Carica papaya L. **Crop Breeding and Applied Biotechnology** 11: 224-231.

Campos B, Viana AP, Quintal SSR, Barbosa CD and Daher RF (2016) Heterotic group formation in Psidium guajava L. by artificial neural network and discriminant analysis. **Revista Brasileira de Fruticultura** 38: 151-157.

Cruz CD (1990) **Aplicação de algumas técnicas multivariadas no melhoramento de plantas**.Tese de Doutorado, 188f. Universidade de São Paulo. ESALQ, Genética e Melhoramento de Plantas, Piracicaba-SP.

Cruz CD, Carvalho SP and Vencovsky R (1994). Estudo sobre divergência genética I. Fatores que afetam a predição do comportamento de híbridos. **Revista Ceres** 41: 178-182.

Cruz CD, Regazzi AJ and Carneiro PCS (2006) **Modelos Biométricos aplicados ao Melhoramento Genético**. Editora UFV, Viçosa, 585p.

Cruz CD, Ferreira FM and Pessoni LA (2011) **Biometria aplicada ao estudo da diversidade genética**. Editora Suprema, Visconde do Rio Branco, 620p.

Falconer DS (1989) **Introduction to Quantitative Genetics**. Longman Scientific & Technical, New York, 438p.

FAO - **Food and Agriculture Organization of the United Nations Statistics** (FAOSTAT) (2018) Disponível em: <http://www.fao.org/faostat/en/#data>. Acesso em: 10 jun. 2020.

Kohonen T (2012) **Self-Organizing Maps**. Springer Science & Business Media, Cham, 373p.

Leão PCS (2008) **Recursos Genéticos de Videira (*Vitis* spp.): análise da diversidade e caracterização da coleção de germoplasma da Embrapa Semiárido.** Tese de Doutorado**,** 126f. Universidade Federal de Viçosa. Genética e Melhoramento de Plantas, Viçosa-MG.

Leão PCS and Borges RME (2009) **Melhoramento Genético da Videira**. Embrapa Semiárido-Documentos (INFOTECA-E. Disponível em: < https://www.infoteca.cnptia.embrapa.br/bitstream/doc/748878/1/SDC224.pdf>. Acesso em: 15 abr. 2020.

Leão PCS, Cruz CD and Motoike SY (2011) Genetic diversity of table grape based on morphoagronomic traits. **Scientia Agricola** 68: 42-49.

Maia JDG, Ritschel PS and Lazzarotto JJ (2018) A viticultura de mesa no Brasil: produção para o mercado nacional e internacional. **Territoires du Vin** 9: 1-9.

Melchinger AE and Gumber RK (1998) Overview of heterosis and heterotic groups in agronomic crops. **Concepts and Breeding of Heterosis in Crop Plants** 25: 29-44.

Miranda JEC, Cruz CD and Costa CP (1988) Predição do comportamento de híbridos de pimentão (*Capsicum annum* L.) pela divergência genética dos progenitores. **Revista Brasileira de Genética** 11: 929-937.

Murtagh F and Legendre P (2011) Ward's hierarchical clustering method: Clustering criterion and agglomerative algorithm. **arXiv**: 1111.6285.

Murtagh F and Legendre P (2014) Ward’s hierarchical agglomerative clustering method: which algorithms implement Ward’s criterion. **Journal of Classification** 31: 274-295.

Oliveira MS, Dos Santos IG and Cruz CD (2020) Self-organizing maps: a powerful tool for capturing genetic diversity patterns of populations. **Euphytica** 216: 1-9.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M and Duchesnay E (2011) Scikit-learn: Machine Learning in Python**.** **Journal of Machine Learning Research** 12: 2825-2830.

Santos IGD, Carneiro VQ, Silva Junior ACD, Cruz CD and Soares PC (2019) Self-organizing maps in the study of genetic diversity among irrigated rice genotypes. **Acta Scientiarum. Agronomy** 41: 1-9.

Silva CM, Karasawa MMG, Vencovsky R and Veasey EA (2007) Elevada diversidade genética interpopulacional em Oryza glumaepatula Steud. (Poaceae) avaliada com microssatélites. **Biota Neotropica** 7: 165-171.

Ultsch A and Kampf D (2004) Knowledge discovery in DNA microarray data of cancer patients with emergent self organizing maps. In: **ESANN**. Belgium, Bruges, p. 501-506.

Ultsch A and Mörchen F (2005) ESOM-Maps: Tools for clustering, visualization, and classification with emergent SOM. **Technical Report - Data Bionics Research Group of University of Marburg**: 1-7.

Ultsch A (2007) Emergence in self organizing feature maps. In **International Workshop on Self-Organizing Maps: Proceedings**. Data Bionics Research Group of University of Marburg, p.7.

Wittek P, Gao SC, Lim IS and Zhao L (2013) Somoclu: An efficient parallel library for self-organizing maps. **arXiv**:1305.1422.