

UNIVERSIDADE ESTADUAL PAULISTA “JÚLIO DE MESQUITA FILHO”

FACULDADE DE CIÊNCIAS E ENGENHARIA

Programa de Pós-Graduação em Agronegócio e Desenvolvimento

PATRÍCIA DE FREITAS PELOZO HESPANHOL

**ANÁLISE DE PADRÕES NA PRODUÇÃO DE CANA DE AÇÚCAR UTILIZANDO
APRENDIZADO DE MÁQUINA**

TUPÃ - SP

2019

PATRÍCIA DE FREITAS PELOZO HESPANHOL

**ANÁLISE DE PADRÕES NA PRODUÇÃO DE CANA DE AÇÚCAR UTILIZANDO
APRENDIZADO DE MÁQUINA**

Dissertação apresentada ao Programa de Pós-Graduação em Agronegócio e Desenvolvimento da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Câmpus de Tupã, como requisito para a obtenção do título de Mestre em Agronegócio e Desenvolvimento.

Área de concentração: Agronegócio e Desenvolvimento

Linha de pesquisa: Competitividade de Sistemas Agroindustriais

Orientador: Prof. Dr. Luís Roberto Almeida Gabriel Filho

Co-orientadores: Prof. Dr. Luiz Fernando Sommaggio Coletta

Profa. Dra. Camila Pires Cremasco Gabriel

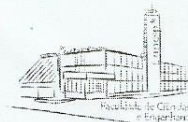
TUPÃ - SP

2019

Ficha catalográfica elaborada pela Seção Técnica de Biblioteca e Documentação da UNESP, Câmpus de Tupã:

H462a	<p>Hespanhol, Patrícia de Freitas Pelozo. Análise de padrões na produção de cana de açúcar utilizando aprendizado de máquina / Patrícia de Freitas Pelozo Hespanhol. - Tupã, 2019. 124 f.</p> <p>Dissertação (Mestrado em Agronegócio e Desenvolvimento) – Faculdade de Ciências e Engenharia – Universidade Estadual Paulista “Júlio de Mesquita Filho”, 2019.</p> <p>Orientador: Luis Roberto Almeida Gabriel Filho Coorientadora: Camila Pires Cremasco Gabriel Coorientador: Luiz Fernando Sommaggio Coletta</p> <p>1. Agronegócio. 2. Reconhecimento de Padrões. 3. Floresta de Caminhos Ótimos. 4. K-Média. 5. Fuzzy C-means. I. Título. II. Autor.</p>
-------	---

Fonte: Eliana Kátia Pupim Bibliotecária CRB 8 – 6202. Essa ficha não pode ser modificada.

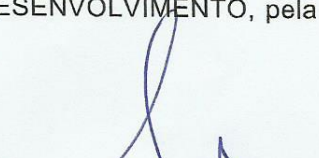


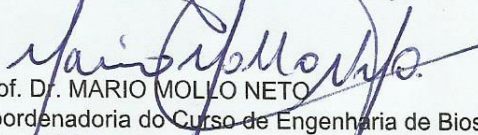
CERTIFICADO DE APROVAÇÃO

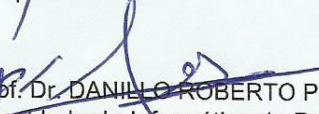
**TÍTULO DA DISSERTAÇÃO: ANÁLISE DE PADRÕES NA PRODUÇÃO DE CANA DE AÇÚCAR
UTILIZANDO APRENDIZADO DE MÁQUINA**

**AUTORA: PATRÍCIA DE FREITAS PELOZO HESPANHOL
ORIENTADOR: LUÍS ROBERTO ALMEIDA GABRIEL FILHO
COORIENTADOR: LUIZ FERNANDO SOMMAGGIO COLETTA
COORIENTADORA: CAMILA PIRES CREMASCO GABRIEL**

Aprovada como parte das exigências para obtenção do Título de Mestra em AGRONEGÓCIO E DESENVOLVIMENTO, pela Comissão Examinadora:


Prof. Dr. LUÍS ROBERTO ALMEIDA GABRIEL FILHO
Coordenadoria do Curso de Administração / Faculdade de Ciências e Engenharia - FCE - UNESP - Tupã/SP


Prof. Dr. MARIO MOLLO NETO
Coordenadoria do Curso de Engenharia de Biosistemas / Faculdade de Ciências e Engenharia - FCE - UNESP - Tupã/SP


Prof. Dr. DANILLO ROBERTO PEREIRA
Faculdade de Informática de Presidente Prudente / Universidade do Oeste Paulista - UNOESTE - Presidente Prudente/SP

Tupã, 10 de junho de 2019

Dedico,

*Aos meus pais, Luis e Maria Inês, e ao meu esposo
Rafael que contribuíram para esta conquista.*

AGRADECIMENTOS

Agradeço a Deus pelo dom da vida e por tudo que conquistei.

Aos meus pais, por todo suporte e apoio incondicional durante toda minha vida.

Ao meu esposo, Rafael Hespanhol, pelas inúmeras orientações essenciais no desenvolvimento deste trabalho. Agradeço por todo suporte, paciência e ensinamentos, mas principalmente pelo carinho e amor dedicados.

A Raquel, Rogério, Renata e Juninho, que além de todo apoio, me ajudaram nos momentos em que precisei, sempre com conselhos preciosos que levarei comigo por toda vida.

A todos os professores da Pós-graduação em Agronegócio e Desenvolvimento da UNESP, por todo conhecimento adquirido, especialmente ao professor Luís Roberto Almeida Gabriel Filho, por todo auxílio e dedicação no desenvolvimento do trabalho.

Aos meus coorientadores, o professor Luiz Fernando Sommaggio Coletta, que passei a admirar por sua inteligência e paciência ao passar seus ensinamentos, que foram fundamentais para o desenvolvimento desse trabalho e que me ensinou muito sobre Inteligência Artificial e a professora Camila Pires Cremasco Gabriel por todo suporte no desenvolvimento do trabalho.

Aos membros da banca de Qualificação e Defesa, que foram essenciais com sugestões, auxiliando no aprimoramento do trabalho.

Aos meus amigos Jéssica e Murilo por todos os conselhos, conversas e por estar ao meu lado em todos os momentos. Vocês foram essenciais para minha formação profissional e pessoal.

A todos os meus colegas de sala de aula, com os quais eu aprendi muito não apenas academicamente, mas também com ensinamentos que levarei para vida.

Às Famílias Pelozo, Freitas e Hespanhol, por todo aconchego e paz que me proporcionam e por todos os conselhos relacionados ao meu desenvolvimento profissional e pessoal.

A todos aqueles que, de algum modo, contribuíram para realização deste trabalho.

HESPANHOL, Patrícia Freitas Pelozo. Análise de padrões na produção de cana de açúcar utilizando Aprendizado de Máquina. 124 p. Dissertação (Mestrado em Agronegócio e Desenvolvimento) – Faculdade de Ciências e Engenharia, Universidade Estadual Paulista “Júlio de Mesquita Filho”, Tupã, 2019.

Resumo

O presente trabalho buscou identificar padrões na produção de cana de por meio da utilização de Inteligência Artificial. Para tanto, foi realizada coleta de informações de fontes secundárias, com dados estatísticos fornecidos por órgãos públicos sobre a área cultivada e a produção de cana de açúcar, índices como pluviométricos e de temperatura e o tipo de solo dos municípios do estado de São Paulo, no ano de 2017, por meio de pesquisa documental. Com a utilização dos métodos Floresta dos Caminhos Ótimos (OPF), *K-means* e *Fuzzy C-means* (FCM) buscou-se identificar *clusters*, ou padrões, que representem essas características produtivas. Além disso, o trabalho testou a utilização do algoritmo OPF como ferramenta de apoio à decisão no setor agroindustrial e fez a comparação do método com os agrupadores de padrões *K-means* e FCM. Após o processamento dos dados foi possível identificar padrões na produção de cana de açúcar pelos três algoritmos, sendo que o OPF proporcionou resultados muito parecidos com o *K-means* e FCM, confirmando a eficiência do método. Além disso, foi possível identificar, no ano de 2017, um padrão de produção com municípios com alta produtividade, grandes áreas destinadas a produção de cana de açúcar e produção da cultura, com temperatura média alta e índices pluviométricos baixos. Os municípios que possuem pequenas áreas com plantação de cana de açúcar possuem uma variabilidade muito grande em resultados de produtividade. O padrão de município com baixa produtividade é acompanhado por temperatura média muito baixa, índices pluviométricos muito altos e solos do tipo Cambissolos, Neossolos e Espodossolos. O padrão do tipo de solo que proporcionou maior produtividade para os municípios foi o Latossolo.

Palavras-chave: Agronegócio. Reconhecimento de padrões. Floresta de Caminhos Ótimos. *K-médias*. *Fuzzy C-means*.

HESPANHOL, Patrícia Freitas Pelozo. Pattern Analysis on sugar cane production using Machine Learning. 124 p. Dissertação (Mestrado em Agronegócio e Desenvolvimento) – Faculdade de Ciências e Engenharia, Universidade Estadual Paulista “Júlio de Mesquita Filho”, Tupã, 2019.

Abstract

The present work sought to identify patterns in sugarcane production through the use of Artificial Intelligence. For this purpose, information was collected from secondary sources, with statistical data provided by public agencies on cultivated area and sugarcane production, rainfall and temperature indices, and the soil type of the municipalities of the State of São Paulo, in the year 2017, through documentary research. Using Optimum-Path Forest (OPF), K-means and Fuzzy C-means (FCM) methods, the aim was to identify clusters, or patterns, that represent these productive characteristics. In addition, the work tested the use of OPF algorithm as a decision support tool in the agribusiness sector and compared the method with the K-means and FCM standards groupers. After data processing, it was possible to identify patterns in sugarcane production by the three algorithms, and OPF provided results very similar to K-means and FCM, confirming the efficiency of the method. In addition, it was possible to identify, in the year 2017, a production pattern of municipalities with high productivity, large areas destined to the production of sugar cane and crop production, with high average temperature and low rainfall. Municipalities that have small areas with sugar cane plantation have a very large variability in productivity results. The municipal pattern with low productivity is accompanied by very low average temperature, very high rainfall rates and soils of the type Cambisols, Neosols and Spodosols. The soil type pattern that provided the highest productivity for the municipalities was the Oxisol.

Keywords: Agribusiness. Pattern Recognition. Optimum-Path Forest. K-means. Fuzzy C-means.

LISTA DE FIGURAS

Figura 1 -	Aumento do PIB brasileiro de 2017 em relação a 2016 e a contribuição dos principais setores	16
Figura 2 -	Série histórica da área com produção de cana de açúcar colhida no Brasil	19
Figura 3 -	Série histórica da produção de cana de açúcar no Brasil.....	20
Figura 4 -	Produção de cana de açúcar das regiões brasileiras em 2017.....	20
Figura 5 -	EDRs do estado de São Paulo com maior produção de cana de açúcar.....	22
Figura 6 -	Produção de cana de açúcar em alguns municípios do estado de São Paulo em 2017.....	23
Figura 7 -	Tipos de algoritmos de reconhecimento de padrões.....	30
Figura 8 -	Variações do algoritmo OPF.....	39
Figura 9 -	Ilustração da base de dados Z	41
Figura 10 -	Amostras na dimensão espaço.....	41
Figura 11 -	Mapa pedológico do estado de São Paulo.....	47
Figura 12 -	Resultados obtidos com o K -means.....	52
Figura 13 -	<i>BoxPlot</i> da produtividade obtida pelo K -means para os seis <i>clusters</i>	55
Figura 14 -	<i>BoxPlot</i> da produção obtida pelo K -means para os seis <i>clusters</i> ..	56
Figura 15 -	<i>BoxPlot</i> da área obtida pelo K -means para os seis <i>clusters</i>	58
Figura 16 -	<i>BoxPlot</i> do índice pluviométrico obtido pelo K -means para os seis <i>clusters</i>	59
Figura 17 -	<i>BoxPlot</i> da temperatura média obtida pelo K -means para os seis <i>clusters</i>	60
Figura 18 -	<i>BoxPlot</i> do tipo de solo obtido pelo K -means para os seis <i>clusters</i>	62
Figura 19 -	Gráficos de dispersão relacionando os atributos à produtividade, obtidos com o K -means.....	64

Figura 20 - Representação dos municípios pertencentes aos <i>clusters</i> um e seis do <i>K-means</i>	66
Figura 21 - Resultados obtidos com o FCM.....	68
Figura 22 - <i>BoxPlot</i> da produtividade obtida pelo FCM para os seis <i>clusters</i> ..	70
Figura 23 - <i>BoxPlot</i> da produção obtida pelo FCM para os seis <i>clusters</i>	71
Figura 24 - <i>BoxPlot</i> da área obtida pelo FCM para os seis <i>clusters</i>	72
Figura 25 - <i>BoxPlot</i> do índice pluviométrico obtido pelo FCM para os seis <i>clusters</i>	74
Figura 26 - <i>BoxPlot</i> da temperatura média obtida pelo FCM para os seis <i>clusters</i>	75
Figura 27 - <i>BoxPlot</i> do tipo de solo obtido pelo FCM para os seis <i>clusters</i>	76
Figura 28 - Gráficos de dispersão relacionando os atributos à produtividade, obtidos com o FCM.....	78
Figura 29 - Representação dos municípios pertencentes aos <i>clusters</i> um e seis do FCM.....	80
Figura 30 - Resultados obtidos com o OPF.....	81
Figura 31 - <i>BoxPlot</i> da produtividade obtida pelo OPF para os quatro <i>clusters</i>	83
Figura 32 - <i>BoxPlot</i> da produção obtida pelo OPF para os quatro <i>clusters</i>	84
Figura 33 - <i>BoxPlot</i> da área obtida pelo OPF para os quatro <i>clusters</i>	86
Figura 34 - <i>BoxPlot</i> do índice pluviométrico obtido pelo OPF para os quatro <i>clusters</i>	87
Figura 35 - <i>BoxPlot</i> da temperatura média obtida pelo OPF para os quatro <i>clusters</i>	88
Figura 36 - <i>BoxPlot</i> do tipo de solo obtido pelo OPF para os quatro <i>clusters</i> ..	90
Figura 37 - Gráficos de dispersão relacionando os atributos à produtividade, obtidos com o OPF.....	92
Figura 38 - Representação dos municípios pertencentes aos <i>clusters</i> um e seis do OPF.....	93

LISTA DE QUADROS

Quadro 1 -	Exemplos de aplicações do <i>K-means</i> nos últimos cinco anos....	33
Quadro 2 -	Exemplos de aplicações do FCM nos últimos cinco anos.....	37
Quadro 3 -	Exemplos de aplicações do OPF nos últimos cinco anos.....	42
Quadro 4 -	Municípios do estado de São Paulo que não fazem parte desta pesquisa.....	45
Quadro 5 -	Classificação dos tipos de solos.....	48
Quadro 6 -	Coeficiente de silhueta.....	49
Quadro 7 -	Base de dados Z.....	50
Quadro 8-	Tipo de solo predominante de cada <i>cluster</i> do <i>K-means</i>	54
Quadro 9-	Tipo de solo predominante de cada <i>cluster</i> do FCM.....	69
Quadro 10-	Tipo de solo predominante de cada <i>cluster</i> do OPF.....	82

LISTA DE TABELAS

Tabela 1 -	Produção de cana de açúcar nos estados da região sudeste brasileira.....	21
Tabela 2 -	Análise estatística da produtividade obtida pelo <i>K-means</i> para os seis <i>clusters</i>	56
Tabela 3 -	Análise estatística da produção obtida pelo <i>K-means</i> para os seis <i>clusters</i>	57
Tabela 4 -	Análise estatística da área obtida pelo <i>K-means</i> para os seis <i>clusters</i>	58
Tabela 5 -	Análise estatística do índice pluviométrico obtido pelo <i>K-means</i> para os seis <i>clusters</i>	60
Tabela 6 -	Análise estatística da temperatura média obtida pelo <i>K-means</i> para os seis <i>clusters</i>	61
Tabela 7 -	Análise estatística do tipo de solo obtido pelo <i>K-means</i> para os seis <i>clusters</i>	63
Tabela 8 -	Análise estatística da produtividade obtida pelo FCM para os seis <i>clusters</i>	70
Tabela 9 -	Análise estatística da produção obtida pelo FCM para os seis <i>clusters</i>	72
Tabela 10 -	Análise estatística da área obtida pelo FCM para os seis <i>clusters</i>	73
Tabela 11 -	Análise estatística do índice pluviométrico obtido pelo FCM para os seis <i>clusters</i>	74
Tabela 12 -	Análise estatística da temperatura média obtida pelo FCM para os seis <i>clusters</i>	76
Tabela 13 -	Análise estatística do tipo de solo obtido pelo FCM para os seis <i>clusters</i>	77
Tabela 14 -	Análise estatística da produtividade obtida pelo OPF para os quatro <i>clusters</i>	84
Tabela 15 -	Análise estatística da produção obtida pelo OPF para os quatro <i>clusters</i>	85

Tabela 16 -	Análise estatística da área obtida pelo OPF para os quatro <i>clusters</i>	86
Tabela 17 -	Análise estatística do índice pluviométrico obtido pelo OPF para os quatro <i>clusters</i>	88
Tabela 18 -	Análise estatística da temperatura média obtida pelo OPF para os quatro <i>clusters</i>	89
Tabela 19 -	Análise estatística do tipo de solo médio obtido pelo OPF para os quatro <i>clusters</i>	90

SUMÁRIO

1	INTRODUÇÃO.....	11
1.1	Objetivo Geral.....	13
1.1.1	Objetivos específicos.....	13
2	REVISÃO BIBLIOGRÁFICA.....	14
2.1	O Agronegócio.....	14
2.1.1	Cana de açúcar.....	17
2.2	Inteligência Artificial (IA).....	23
2.2.1	História da IA.....	25
2.2.2	Aprendizado de Máquina (AM).....	28
2.3	Reconhecimento de Padrões.....	29
2.3.2	<i>K-means</i>	32
2.3.3	<i>Fuzzy C-means</i> (FCM).....	35
2.3.1	Floresta de Caminhos Ótimos (OPF).....	39
3	METODOLOGIA.....	44
4	RESULTADOS E DISCUSSÕES.....	52
4.1	Resultados e análises obtidos por meio do <i>K-means</i>	52
4.2	Resultados e análises obtidos por meio do FCM.....	67
4.3	Resultados e análises obtidos por meio do OPF.....	80
5	CONSIDERAÇÕES FINAIS.....	95
	REFERENCIAS.....	97

1 INTRODUÇÃO

O agronegócio é considerado como uma das mais importantes fontes de riqueza do Brasil por colocar o país entre as nações mais competitivas do mundo na produção de *commodities* agroindustriais. O setor engloba atividades como fornecimento de insumos, pecuária, lavoura, extração vegetal, processo agroindustrial e todas as áreas que dão suporte ao fluxo de produtos até o consumidor final (SILVA; MONTEIRO; LIMA, 2015; JANK; NASSAR; TACHINARDI, 2005).

A cana de açúcar é uma das principais culturas do agronegócio brasileiro, sendo o país o maior produtor mundial da cultura. São vários os motivos que levam a esse cenário, mas podem-se citar, como principais, as condições climáticas e de solo, que são favoráveis ao cultivo da cana de açúcar e a grande extensão de terras disponíveis (CONAB, 2018a; KORB *et al.*, 2016; MIRANDA; VASCONCELOS; LANDELL, 2010; FERNANDES 1984). Em 2017, a área colhida de cana de açúcar foi de 8.729,5 mil hectares, sendo que a produção atingiu 633.261,9 mil toneladas (CONAB, 2018a).

A região do país com maior área colhida e maior volume produção de cana de açúcar é o Sudeste, que em 2017, foi sozinho responsável por 65,9% da produção total do país, sendo que nesta região, o estado de São Paulo é o maior produtor, atingindo uma produção de 349.200,5 mil toneladas em 2017 (CONAB, 2018a).

A cana-de-açúcar é o principal produto da agropecuária do estado de São Paulo, que em 2017 atingiu uma produção de 349.200,5 mil toneladas, com uma participação de 35,8% (R\$ 28,07 bilhões) no valor da produção agropecuária e florestal total do estado em 2016 (CONAB, 2018a; IEA, 2018a).

Foram utilizados algoritmos de Aprendizado de Máquina para identificar padrões na cadeia produtiva da cana de açúcar no estado de São Paulo. Essas informações são importantes para órgãos que dão suporte aos produtores rurais, pois a identificação dos possíveis padrões por município ou agrupamentos de características semelhantes, pode vir a servir como base para a elaboração de políticas públicas e direcionamento focado de recursos (humanos, financeiros ou materiais), atualmente realizado de maneira não sistemática, bem como de guia para possíveis investidores na cadeia produtiva da cana de açúcar.

Portanto, considerando área, produção e produtividade da cana de açúcar, índices pluviométricos e de temperatura e o tipo de solo dos municípios do estado de

São Paulo, procurar-se-á responder a seguinte pergunta: é possível identificar padrões na produção de cana de açúcar no estado de São Paulo por meio da utilização de algoritmos de agrupamento de dados?

Para encontrar os padrões nas características de produção, adotar-se-á o método da Floresta dos Caminhos Ótimos (OPF), desenvolvido por Papa (2008), por meio da utilização do método não supervisionado, visando o agrupamento dos dados da produção de cana de açúcar. O OPF é fundamentado na Teoria dos Grafos, e é considerado um método de abordagem matemática simples, conectando amostras modeladas de um grafo completo, em que elementos mais representativos de cada classe são escolhidos como sendo fronteiras entre as classes (SOUZA; LOTUFO; RITTNER, 2012; HESPANHOL; PEREIRA; FORTES, 2015).

Pretende-se também testar a eficiência do método, comparando-o com o *K-means*, que possui como diferença o fato de suas iterações acontecerem após a designação de um conjunto inicial de partições estabelecidas por meio de seu algoritmo, para posteriormente, a partir do centro de cada uma destas partições, calcular a similaridade desses elementos com os outros a serem agrupados (WIVES, 2004). Ainda com o objetivo de testar a eficiência do OPF, o método *Fuzzy C-means* foi escolhido por ser uma extensão *fuzzy* do *K-means*, permitindo assim a análise de sobreposição de dados nos diferentes agrupamentos gerados. Esta característica não está presente nos métodos OPF e *K-means* e permite um mapeamento mais realista dos dados.

Esta pesquisa apresenta como hipóteses ao problema proposto, as seguintes proposições:

- Existirão regiões que formarão *clusters* com características de alta produção, em áreas extensas, com baixas temperaturas e baixo índice pluviométrico e baixa produtividade;
- Pode haver regiões com uma faixa de índices pluviométricos e de temperatura parecidos com a hipótese anterior, com baixa produção e área destinada para a produção da cultura, porém alta produtividade,
- Pode-se ainda haver regiões em que os índices pluviométricos são altos, temperaturas altas, com grandes extensões de área destinada à cultura e alta produtividade.

Desta forma, os índices deverão variar, proporcionando diferentes resultados e formando diferentes *clusters* com características semelhantes.

1.1 Objetivo Geral

O trabalho possui o objetivo de identificar padrões na produção de cana de açúcar no estado de São Paulo por meio da utilização de algoritmos de agrupamento de dados. Desta forma, acredita-se que será possível extrair conhecimentos e fornecer indicativos de fatores que contribuam para um melhor padrão de produção.

1.1.1 Objetivos específicos

- Avaliar as ferramentas Aprendizado de máquina *K-means*, *Fuzzy C-means* e OPF no reconhecimento de padrões no agronegócio;
- Analisar os padrões encontrados e verificar as ameaças e oportunidades na produção de cana de açúcar;
- Fornecer mapeamentos e indicativos de representatividade da produção de cana de cana de açúcar nos municípios analisados.

2 REVISÃO BIBLIOGRÁFICA

2.1 O Agronegócio

A agricultura mundial passou por um processo de expansão a partir da Segunda Guerra Mundial, principalmente relacionado ao comércio agrícola e industrialização da agricultura mundial, que acarretou na disseminação internacional do sistema de produção denominado agronegócio (MENDONÇA, 2015).

John Davis e Ray Goldberg publicaram o livro “*A Concept of Agribusiness*” em 1957, na *School of Business Administration* da Universidade de Harvard, que deu origem ao termo *agribusiness*, Davis e Goldberg ressaltaram, em seu livro, que o campo estaria vivendo uma revolução tecnológica, dando origem as chamadas “fazendas modernas”. Para os autores, as fazendas não mais podiam se sustentar sozinhas e passaram a ter função comercial, com sua produção baseada em monocultivos. Outras empresas passaram a operar atividades como armazenamento, processamento e distribuição e também produzir produtos industriais utilizados neste modelo agrícola, como tratores, caminhões, combustível, fertilizantes, ração, pesticidas, entre outros. Desta forma, o termo agronegócio torna-se adequado para acompanhar o progresso que estava ocorrendo no campo (MENDONÇA, 2015; SILVA; MONTEIRO; LIMA, 2015).

No Brasil, o termo agronegócio agrega as atividades agroquímicas, industriais e comerciais aos cálculos econômicos da agricultura, sendo um termo utilizado para justificar a criação das chamadas cadeias produtivas (MENDONÇA, 2015). A cadeia produtiva é a representação de atividades integradas em um conjunto produtivo, tratando-se de uma corrente que vem desde a extração e manuseio da matéria prima até a distribuição. Essa abordagem possibilita ter ampla visualização do processo produtivo a partir da formação de uma cadeia produtiva, permitindo a análise sistêmica das atividades no agronegócio (MONFORT, 1983). Batalha (2007) conceitua cadeia produtiva como a relação de compra e venda em um sistema composto por um conjunto de setores econômicos, organizado sequencialmente no processo produtivo, em que está envolvida toda a atividade de produção e comercialização de um produto que possuir valor agregado conforme se avança na cadeia.

Os componentes da cadeia produtiva agroindustrial abrangem o âmbito “antes da porteira”, que costuma ser a atividade econômica do fornecedor de insumos para a agricultura, “dentro da porteira”, que é a responsável pela produção rural, e “depois da porteira”, que envolve a industrialização e distribuição (PASSADOR J., ROSA, PASSADOR C., 2004).

O conceito de agronegócio estabelece relações entre a indústria a montante, que é caracterizada pelas que produzem bens de capital e insumos básicos para o campo, estabelecimentos rurais e indústrias a jusante, que são os processadores de alimentos, as empresas logísticas e o mercado consumidor. Estão inclusas ainda as influências governamentais, mercados futuros e associações comerciais. Desta forma, a construção da ideia do que seria o agronegócio possui ampla aplicação, perpassando pelo desenho de políticas públicas, arquitetura de organizações e elaboração de estratégias corporativas (ZYLBERSZTAJN, 1995). Sendo assim, todas as atividades, como fornecimento de insumos, pecuária, lavoura, extração vegetal, processo agroindustrial e todas as áreas que dão suporte ao fluxo de produtos até o consumidor final estão englobadas no que se chama de agronegócio (SILVA; MONTEIRO; LIMA, 2015).

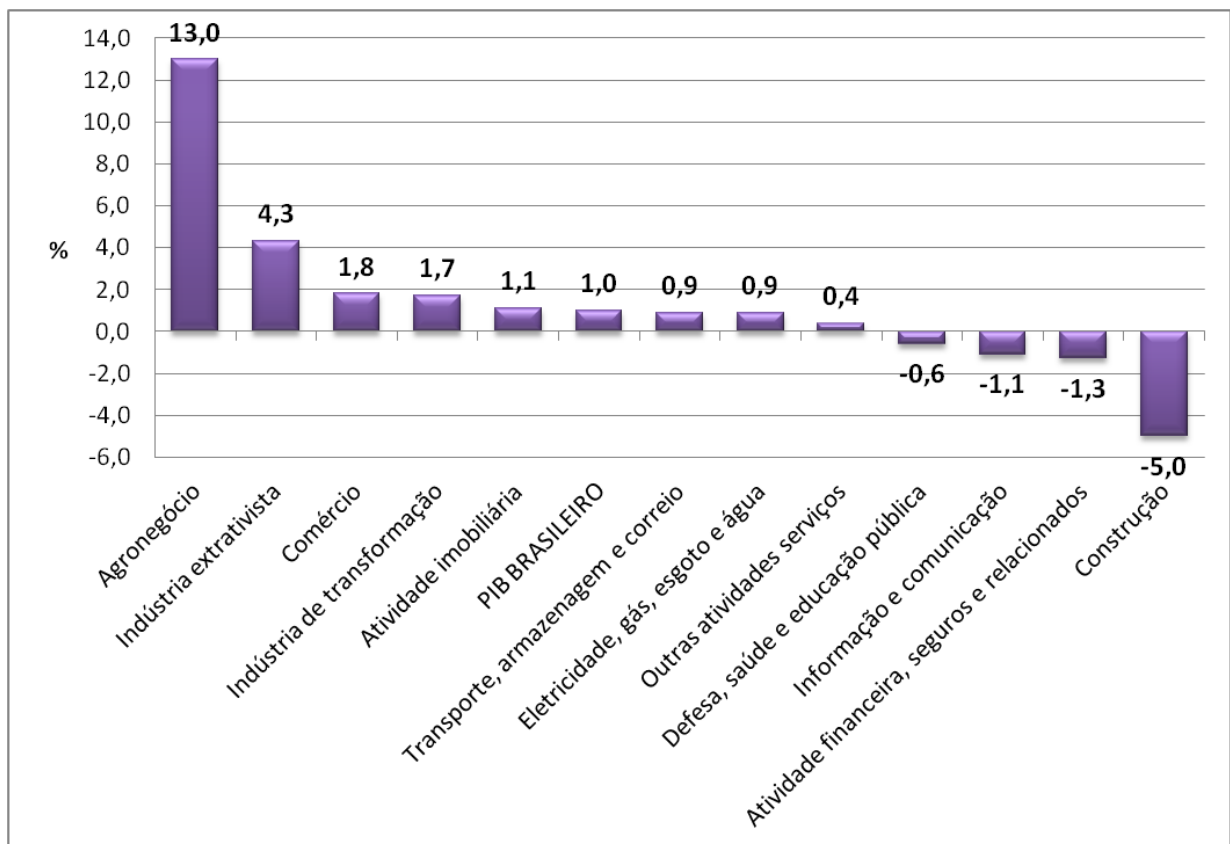
Em 1990, a Embrapa incorporou esta perspectiva sobre o termo agronegócio com o objetivo de incluir como clientes setores denominados “de fora da porteira da fazenda”. Esta estratégia passou a ser utilizada no Brasil inicialmente por meio do termo complexo agroindustrial e atualmente é chamado de agronegócio (MENDONÇA, 2015; CASTRO; LIMA; CRISTO, 2002).

O agronegócio é considerado como uma das mais importantes fontes de riqueza do Brasil, sendo importante para colocar o país entre as nações mais competitivas do mundo na produção de *commodities* agroindustriais. Tudo isso acontece devido a uma combinação de fatores, sendo os principais: investimentos em tecnologia e pesquisa, desregulamentação dos mercados com a redução da intervenção do governo, a abertura comercial e a estabilização da economia após o Plano Real (JANK; NASSAR; TACHINARDI, 2005).

A Confederação da Agricultura e Pecuária do Brasil (CNA, 2018) divulgou que o Produto Interno Bruto (PIB) brasileiro a preços de mercado cresceu 1% na comparação com 2016, o que corresponde a R\$ 6,6 trilhões. A agropecuária foi o setor que mais contribuiu para esse resultado positivo, apresentando 13% de aumento em relação a 2016, como demonstrado na Figura 1, sendo o maior crescimento dentre

todos os segmentos e também o maior já registrado na história, registrando o valor de R\$ 299,5 bilhões. Torna-se necessário ressaltar que esse crescimento robusto da safra 2017 se deve, além do aumento de produtividade, de estar partindo de uma base comparativa de produção inferior, visto que as condições climáticas em 2017 foram favoráveis, contrariando a safra de grãos de 2015/2016 que sofreu fortes quedas devido a adversidades climáticas (CNA, 2018).

Figura 1 - Aumento do PIB brasileiro de 2017 em relação a 2016 e a contribuição dos principais setores



Fonte: Adaptada de CNA (2018).

A perspectiva da CNA para 2018 é de uma safra também positiva considerando que o clima afetou negativamente apenas algumas áreas na região sul do Rio Grande do Sul, e algumas regiões no Paraná e no Mato Grosso, ocasionando perdas pontuais em algumas regiões.

2.1.1 Cana de açúcar

Linneu descreveu a cana de açúcar, em 1753, como *Saccharum officinarum* e *Saccharum spicatum*. Essa classificação, porém, sofreu algumas modificações com o passar dos anos, sendo atualmente classificada de acordo com cada espécie, sendo elas: *Saccharum Barberi*, *Saccharum Officinarum*, *Saccharum Robustum*, *Saccharum Sinensis*, e *Saccharum Spontaneum* (CESNIK, 2004).

A cana de açúcar é uma planta pertencente ao gênero *Saccharum*, sendo cultivada um híbrido multiespecífico, recebendo a designação "*Saccharum spp*". É uma planta da família *Poaceae* (representada pelo milho, sorgo, arroz e outras gramíneas), anuais ou perenes, da família das angiospermas da classe *Liliopsida*, também conhecidas como capins, gramas ou relvas, sendo plantas floríferas, frequentemente com forma de vida em arbusto, árvore, bambu, erva, trepadeira ou subarbusto, vivendo sob substrato aquático. As principais características de plantas da família *Poaceae* são as formas da inflorescência (espiga), o crescimento do caule ou colmos, e as folhas com laminas de sílica em suas bordas e bainha aberta (DIOLA E SANTOS, 2011).

A origem da cana de açúcar é um assunto divergente na literatura, porém existe um consenso de que, 325 anos a. C., Alexandre, o Grande, foi o primeiro a descobrir a cana de açúcar na Índia Ocidental e o responsável por levá-la a Pérsia, atual Irã, que em suas conquistas a territórios e trajetórias comerciais, expandiram o cultivo para o Oriente Médio, no Egito e Síria. Existem vários relatos na bibliografia de que o cultivo da cana de açúcar foi expandido por várias regiões, como Espanha, Sicília, regiões do Mediterrâneo e Zanzibar. O século XV foi uma fase de grande expansão da cultura para ilhas como Canárias, Cabo Verde e São Tomé (MIRANDA; VASCONCELOS; LANDELL, 2010; ARBEX 2001).

Foi em 1493, quando Cristovão Colombo, que era genro de um grande produtor de açúcar, viajava pela segunda vez para o continente americano, que a cana de açúcar foi introduzida na América, em um local onde hoje é a República Dominicana (MIRANDA; VASCONCELOS; LANDELL, 2010; CESNIK, 2004; ARBEX, 2001).

A cana de açúcar chegou no Brasil em 1502, trazida por Martim Afonso de Souza e que há registros na alfândega de Lisboa de entrada de açúcar brasileiro nos anos de 1520 e 1526 (CESNIK, 2004). Miranda, Vasconcelos e Landell (2010) e Arbex

(2001) acreditam foi a partir de 1502 que os portugueses trouxeram mudas de cana de açúcar da Ilha da Madeira para o Brasil. A cultura demorou a se consolidar no país, pois havia pouca mão de obra e a maior preocupação dos habitantes, na época, era o extrativismo de madeira e a descoberta de minas.

Em 1532, relata-se o início da indústria açucareira, quando Martim Afonso de Souza criou o primeiro engenho em São Vicente (MIRANDA; VASCONCELOS; LANDELL, 2010; CESNIK, 2004). Porém, no final do século XVI, São Vicente diminuía cada vez mais sua produção de cana de açúcar, enquanto Bahia e Pernambuco tornavam-se regiões prósperas na produção de cana e açúcar. O Brasil foi considerado o maior produtor de açúcar do mundo na década de 1580 (FERNANDES, 1984; MIRANDA; VASCONCELOS; LANDELL, 2010; CESNIK, 2004; ARBEX, 2001). Fernandes (1984) e Miranda, Vasconcelos e Landell (2010) afirmam que as condições climáticas e do solo brasileiro eram favoráveis ao cultivo da cana de açúcar, sendo esse um dos fatores para a grande expansão da cultura no país.

Logo após o período da expansão, nos séculos XVIII, XIX e XX, o cultivo da cana de açúcar passou diferentes fases no Brasil. No final do século XX, o país estava vulnerável no campo energético, após abalos na balança comercial brasileira devido a participação das importações de petróleo sobre o total das importações do país, que passaram de cerca de 10%, em 1973, para 57%, em 1983. Com a probabilidade maior de um colapso da economia caso acontecesse uma interrupção no fornecimento de petróleo, outras fontes de energia passaram a ser estudadas. Foi neste período que surgiram estudos voltados a utilização de combustível alternativo e fontes de energias renováveis como a biomassa (ARBEX, 2001; FERNANDES, 1984).

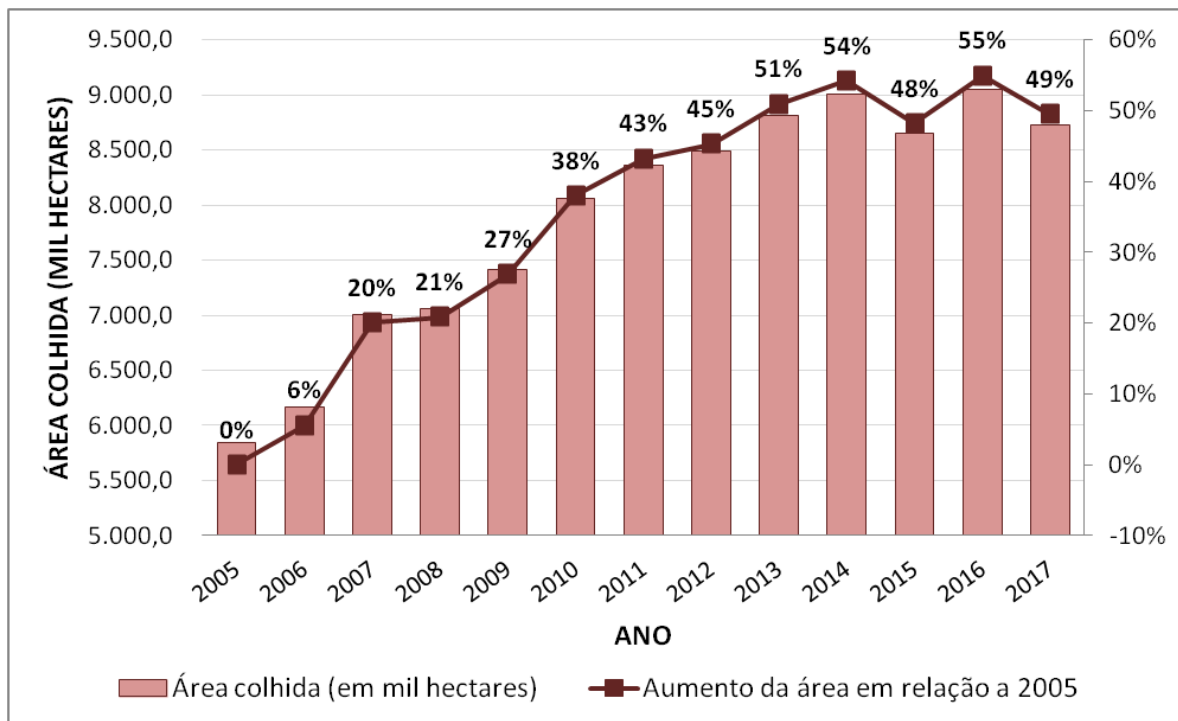
O Brasil, sendo um dos maiores produtores mundiais durante séculos e com abundância de terras disponíveis para plantio, iniciou um processo de união de destilarias a usinas. Em 1978 o governo brasileiro passou a apoiar o desenvolvimento de novas tecnologias no setor, o que possibilitou o surgimento dos primeiros veículos movido exclusivamente a álcool (ARBEX, 2001). A partir de então, a cultura da cana de açúcar atingiu praticamente todo o território brasileiro, sendo o Brasil, atualmente, considerado mundialmente como expoente na produção de biocombustíveis (CONAB, 2018a; VILELA *et al*, 2015).

O Brasil é um país considerado como promissor para a exportação de cana de açúcar, sendo atualmente o maior produtor mundial da cultura, tendo grande importância para o agronegócio brasileiro. Essa consideração se dá pelo fato de que

cada vez aumenta-se a demanda mundial por etanol, principalmente aqueles que são oriundos de fontes renováveis e também pelas grandes áreas cultiváveis e condições edafoclimáticas¹ favoráveis (CONAB, 2018a).

Dados divulgados pela CONAB (2018a) revelam que em 2016 o Brasil atingiu a maior área colhida de cana de açúcar, com 9.049,2 mil hectares, tendo assim um aumento de 55% de área colhida quando comparado com 2005. Em 2017, a área colhida de cana de açúcar foi de 8.729,5 mil hectares, como demonstrado na Figura 2.

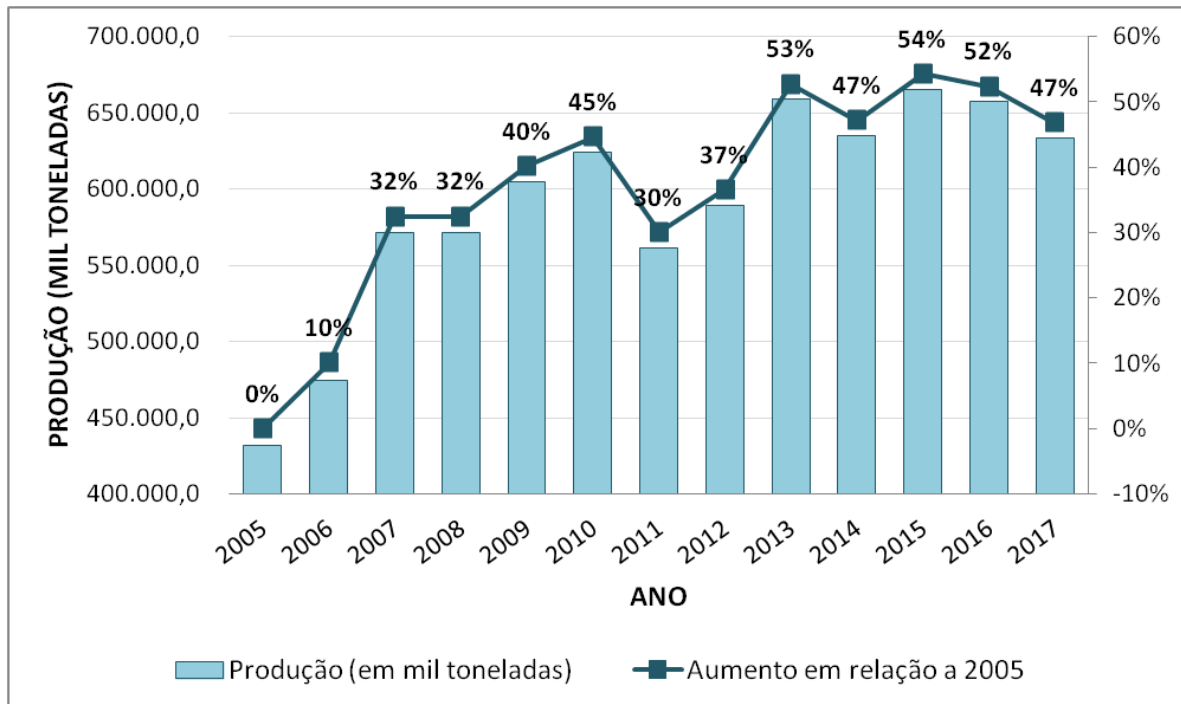
Figura 2 - Série histórica da área com produção de cana de açúcar colhida no Brasil



Fonte: Elaborada pela autora com base nos dados da CONAB (2018b)

Em decorrência do aumento da área destinada a cana de açúcar, o impacto na produção também foi grande. Dados divulgados pela CONAB (2018a) revelam que, em 2015, o Brasil obteve uma produção recorde de cana de açúcar, atingindo 665.586,2 mil toneladas, sendo uma produção 54% maior que o ano de 2005. Nos anos de 2016 e 2017 o Brasil alcançou uma produção de 657.184,0 e 633.261,9 mil toneladas, respectivamente, como demonstrado na Figura 3.

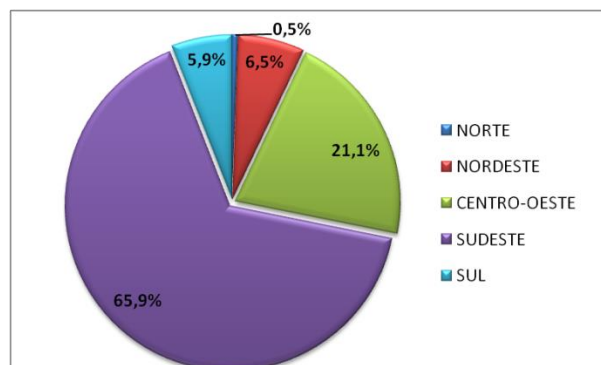
¹ Edafoclimáticas: relativo ao solo e ao clima

Figura 3 - Série histórica da produção de cana de açúcar no Brasil

Fonte: Elaborada pela autora com base nos dados da CONAB (2018b).

A CONAB (2018a) estima que a área colhida no Brasil em 2018 seja um pouco menor do que o ano anterior, atingindo 8.613,6 mil hectares. Além disso, a produção também terá um pequeno decréscimo em relação ao ano anterior, estimando-se 625.963,0 mil toneladas colhidas.

A região do país com maior área colhida e produção de cana de açúcar é o sudeste, que em 2017, foi responsável por 65,9% da produção total do país. A região centro-oeste é a segunda maior produtora sendo responsável por 21,1% da produção, seguida pela região nordeste com 6,5%, sul com 5,9% e por último a região norte com 0,5%, como demonstrado na Figura 4.

Figura 4 - Produção de cana de açúcar das regiões brasileiras em 2017

Fonte: Elaborada pela autora com base nos dados da CONAB (2018a).

Entre os séculos XVI a XVIII, o nordeste brasileiro era uma região tradicional na produção de cana de açúcar, porém, com a migração da cultura para o sul do país desde os anos 1990, o estado de São Paulo tem chamado a atenção por sua forte representatividade nesse segmento, sendo o estado produtor e processador (CAMARA; CALDARELLI, 2016), sendo que 41,7% das usinas produtoras de açúcar e álcool estão situadas no estado de São Paulo (IEA, 2018a)

Na região sudeste, o estado de São Paulo atingiu em 2017 uma produção de 349.200,5 mil toneladas, como demonstrado na Tabela 1, o que representou aproximadamente 85% da produção total da região.

Tabela 1 - Produção de cana de açúcar nos estados da região sudeste brasileira.

Ano	Minas Gerais	Espírito Santo	Rio de Janeiro	São Paulo	Total
2005	27.557,1	4.243,4	7.576,4	265.543,3	304.920,2
2006	33.558,0	3.967,1	6.853,5	284.825,6	329.204,2
2007	44.120,0	4.419,0	3.556,3	340.510,4	392.605,7
2008	41.461,4	4.419,0	3.556,3	345.657,7	395.094,4
2009	49.923,4	4.009,6	3.260,0	362.664,7	419.857,7
2010	56.013,6	3.524,8	2.537,8	361.723,3	423.799,5
2011	50.241,8	4.003,8	2.207,9	305.636,4	362.089,9
2012	51.208,0	3.431,6	1.893,8	330.694,9	387.228,3
2013	60.759,5	3.770,0	2.007,6	372.805,9	439.343,0
2014	59.528,7	3.191,7	1.586,4	341.589,7	405.896,5
2015	64.932,4	2.809,6	1.066,2	367.587,6	436.395,8
2016	63.670,3	1.356,9	1.005,2	369.925,1	435.957,5
2017	65.017,4	2.380,7	872,1	349.200,5	417.470,7

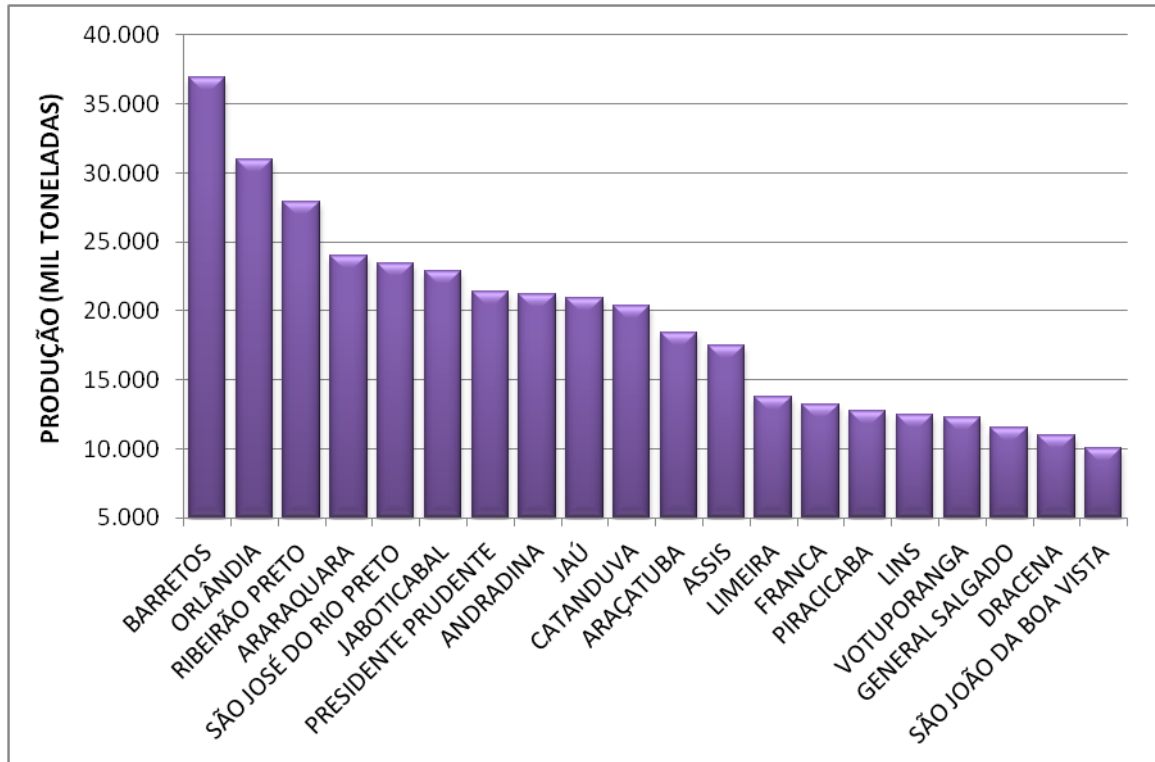
Fonte: Elaborada pela autora com base nos dados divulgados pela CONAB (2018a).

A cana-de-açúcar é o principal produto da agropecuária do estado de São Paulo, possuindo uma participação de 35,8% (R\$ 28,07 bilhões) no valor da produção agropecuária e florestal total do estado em 2016. A cultura ocupa cerca de 5,88 milhões ha no estado, aproximadamente 100 mil Unidades de Produção Agropecuária (UPAs), e as maiores regiões produtoras são os Escritórios de Desenvolvimento Rural (EDRs) de Barretos, Orlandia e Ribeirão Preto (IEA, 2018a).

Dados divulgados pelo IEA (2018b) mostram que, em 2017, o EDR de Barretos atingiu uma produção de 36.896 mil toneladas de cana de açúcar, seguido por Orlandia que produziu 30.957 mil toneladas e Ribeirão Preto com 27.874 mil

toneladas. A Figura 5 demonstra os 20 EDRs do estado de São Paulo com maior produção de cana de açúcar em 2017.

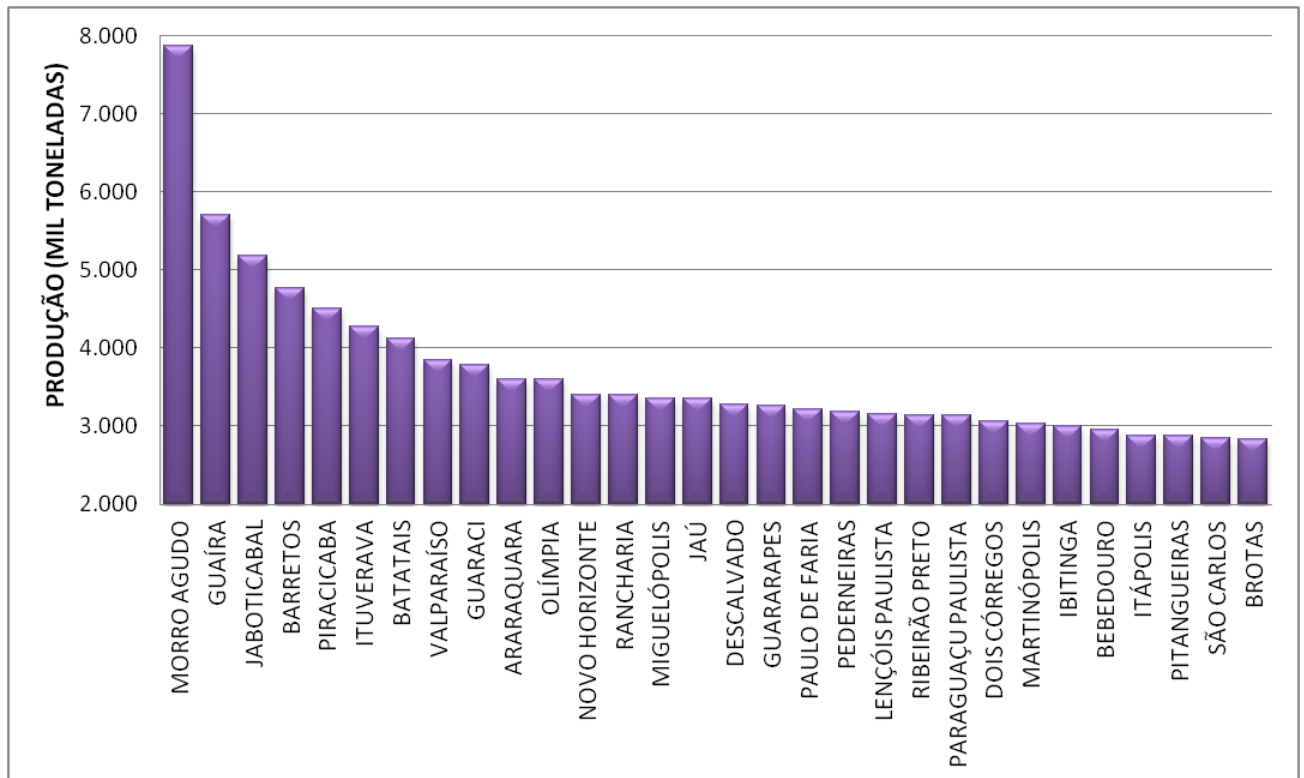
Figura 5 - EDRs do estado de São Paulo com maior produção de cana de açúcar



Fonte: Elaborada pela autora com base nos dados do IEA (2018b).

Levando em consideração os municípios do estado de São Paulo, Morro Agudo, que pertence ao EDR de Ribeirão Preto, é o maior produtor atingindo aproximadamente 7900 mil toneladas em 2017. Com uma área de 138 mil hectares, em 2017, Morro Agudo possuiu cerca de 70% de sua área destinada a produção de cana de açúcar, o que corresponde a 96 mil hectares. Guaira ocupou em 2017 o segundo lugar na produção da cana de açúcar com 5700 mil toneladas e Jaboticabal foi o terceiro com 5179 mil toneladas, como demonstrado na Figura 6 (IEA, 2018b).

Figura 6 - Produção de cana de açúcar em alguns municípios do estado de São Paulo em 2017



Fonte: Elaborada pela autora com base nos dados do IEA (2018b).

Levando em consideração a importância da cana de açúcar no Brasil e no estado de São Paulo, demonstrada neste tópico, e com o objetivo de utilizar da Inteligência Artificial e de algoritmos de Aprendizado de Máquina para identificar padrões na cadeia produtiva da cana de açúcar no estado de São Paulo, nos próximos tópicos, o foco da revisão bibliográfica será em discutir Inteligência Artificial, Aprendizado de Máquina e Reconhecimento de Padrões.

2.2 Inteligência Artificial

A Inteligência Artificial (IA) surge na metade do século XX com o propósito de conferir às máquinas características da inteligência humana, como a capacidade de aprender e analisar as relações entre os dados (PEREIRA, 2017). A espécie humana busca continuamente entender a forma como se pensa, compreende e manipula o mundo. O campo da IA busca, além de compreender, construir entidades inteligentes (RUSSELL; NORVIG, 2013).

Sistemas inteligentes são sistemas computacionais e máquinas que, ligados às pessoas, dados e conhecimentos específicos, desenvolvam um

comportamento inteligente, auxiliando na tomada de decisões (SILVA; SPRITZER; OLIVEIRA, 2004). Os autores completam ainda que esses sistemas não apenas armazenam e manipulam dados, como também possuem a capacidade de deduzir novos conhecimentos e novos fatos a partir dos existentes, visando, normalmente, solucionar problemas.

Russell e Norvig (2013) tratam da definição de Inteligência Artificial a partir de quatro dimensões. A primeira dimensão trata da IA pensando como o ser humano, a segunda refere-se a IA pensando racionalmente, como terceira dimensão os autores tratam da IA agindo como o ser humano e a última agindo racionalmente.

Ao tratar da Inteligência Artificial como “pensando como o ser humano”, os autores acreditam que para considerar que um programa computacional pense como o ser humano, deva-se determinar como os seres humanos pensam. A Inteligência Artificial, neste caso, está relacionada ao esforço em fazer computadores pensarem, sendo as máquinas a representação da mente humana. Desta forma, a evidência de que determinado programa pensa como o ser humano acontece quando os comportamentos atividades de “entrada”, “processamento” e “saída” ocorrem de forma coincidente aos comportamentos do ser humano. O campo interdisciplinar da ciência cognitiva reúne modelos computacionais da IA e técnicas experimentais da psicologia para tentar construir teorias precisas e verificáveis a respeito dos processos de funcionamento da mente humana (RUSSELL; NORVIG, 2013).

Russell e Norvig (2013) tratam também da definição de Inteligência Artificial “agindo como seres humanos”. O teste de Turing é tratado pelos autores como uma das formas de identificar a eficiência de um computador ao agir como um ser humano. No teste de Turing, um ser humano faz perguntas por escrito e o computador passa a ser considerado eficiente quando o interrogador não consegue identificar se as respostas vêm de um computador ou de outro ser humano. Russell e Norvig (2013) consideram a IA, nesta dimensão, como uma arte capaz de criar máquinas que realizam funções que, ao ser executadas por seres humanos, exigem inteligência.

A terceira dimensão tratada por Russell e Norvig (2013) refere-se as máquinas “pensando racionalmente” que se trata do uso de modelos computacionais no estudo das faculdades mentais. Neste caso, existe o chamado “pensamento correto”, que são aqueles raciocínios considerados irrefutáveis, estruturados de tal forma que, partindo de premissas corretas, sempre resultam em conclusões corretas. O exemplo abordado pelos autores trata-se da famosa frase: “Sócrates é um homem;

todos os homens são mortais; então, Sócrates é mortal”. A IA, relacionada ao estudo da lógica, busca desenvolver programas para criar sistemas inteligentes.

A última dimensão para definir a IA, trata das máquinas “agindo racionalmente”, em que um agente é algo que age, mas espera-se que o agente computacional além de agir, opere autonomamente, compreenda seu ambiente, adapte-se a mudanças e seja capaz de criar e perseguir metas. Desta forma, o agente computacional é aquele que age para alcançar o melhor resultado esperado (RUSSELL; NORVIG, 2013).

Guimarães (2005) explica que a IA busca imitar a forma como o homem resolve problemas estatísticos, mas utilizando-se dos fundamentos da heurística. O autor ressalta que o termo Inteligência Artificial possui uma magnitude muito grande, envolvendo até mesmo fundamentos filosóficos.

2.2.1 História da IA

O primeiro estudo reconhecido como da área de IA foi realizado em 1943 por Warren McCulloch e Walter Pitts, que sugeriram um modelo de neurônios artificiais, no qual eles mostravam que qualquer função computável podia ser calculada por certa rede de neurônios conectados, sendo que redes definidas adequadamente seriam capazes de aprender. Em 1949, Donald Hebb criou a regra que hoje é chamada de “aprendizado de Hebb”, em que demonstrava um mecanismo simples de atualização para modificar as intensidades de conexão entre neurônios. O primeiro computador de rede neural foi criado por Marvin Minsky e Dean Edmonds, dois alunos de Harvard, em 1950 (RUSSELL; NORVIG, 2013).

Foram muitos os trabalhos desta época que hoje são reconhecidos como do campo da Inteligência Artificial, porém, o mais influente até hoje foi de Alan Turing, em 1950, intitulado “*Computing Machinery and Intelligence*”, em que foram apresentados assuntos como o teste de Turing, algoritmos genéticos, aprendizagem por reforço e aprendizado de máquina (RUSSELL; NORVIG, 2013).

Em 1956, John McCarthy foi quem organizou o primeiro seminário, no *Dartmouth College*, em Hanover, que reunia 10 pesquisadores dos Estados Unidos interessados no estudo da inteligência, sendo Minsky, Clude Shannon e Nathaniel Rochester os primeiros a aceitarem a proposta. A proposta do seminário era discutir aspectos da aprendizagem ou características inteligentes que possam ser simuladas

por uma máquina. Também participaram do seminário os pesquisadores Trenchard More, Arthur Samuel, Ray Solomonoff, Oliver Selfridge, Allen Newell e Hebert Simon (RUSSELL; NORVIG, 2013).

Foi neste período que ficou claro que a IA não poderia ser um ramo da pesquisa operacional ou da matemática, mas sim se tornar um campo separado, pois seus objetivos eram diferentes de qualquer outro existente até então, que era reproduzir faculdades humanas como criatividade, autoaperfeiçoamento e uso de linguagem. Outro ponto que diferencia a IA de outros campos do conhecimento é sua metodologia, que busca construir máquinas que funcionem de forma autônoma em ambientes complexos e mutáveis (RUSSELL; NORVIG, 2013).

Em 1957, Newell e Simon, que foram os grandes destaques do primeiro seminário de Inteligência Artificial, desenvolveram o “*General Problem Solver*” (GPS), que era um programa de computador criado para solucionar problemas gerais imitando protocolos humanos, incorporando a abordagem desejada de “pensar de forma humana”. Em 1958, John McCarthy definiu a linguagem de alto nível “Lisp”, que se tornou, nos próximos 30 anos, a linguagem de programação dominante na IA e também descreveu um programa de computador denominado de “*Advice Taker*”, que foi considerado o primeiro sistema de IA completo. O primeiro laboratório de IA foi criado em 1963, em Stanford, também por McCarthy (RUSSELL; NORVIG, 2013).

A IA se tornou uma indústria a partir do ano de 1980. A empresa *Digital Equipment Corporation* comercializou em 1982 o primeiro sistema especialista. Em 1988, a empresa Du Pont já tinha 100 desses sistemas em uso e desenvolvia mais 500 deles (RUSSELL; NORVIG, 2013).

Até o século XX, a maior preocupação dos pesquisadores na área da IA estava voltada para o tipo de algoritmo que seria utilizado para resolver os problemas existentes. Porém, no século XXI percebeu-se que mais importante do que o algoritmo a ser utilizado, eram os dados que seriam trabalhados nesses algoritmos, visto que a disponibilidade desses dados era cada vez mais crescente (RUSSELL; NORVIG, 2013).

Dentre as diversas áreas de aplicações que o campo gigantesco da IA possui, Luger (2014) aborda as mais importantes, sendo elas:

- Jogos: as pesquisas iniciais de IA foram realizadas utilizando jogos de tabuleiro comuns, como damas, xadrez e quebra-cabeças;

- Raciocínio automatizado e prova de teoremas: foram responsáveis por trabalhos iniciais na IA, nas pesquisas desenvolvidas por Newell e Simon em 1963, e formalização de algoritmos de busca e desenvolvimento de linguagens de representações formais;
- Sistemas especialistas: a estratégia para resolver problemas, utilizada em sistemas especialistas, é dependente do conhecimento de um especialista humano no domínio que é responsável por fornecer conhecimento necessário do domínio do problema;
- Compreensão da linguagem natural e modelagem semântica: programas que sejam capazes de entender e gerar a linguagem humana;
- Modelando o desempenho humano: sistemas que modelam explicitamente algum aspecto do desempenho humano;
- Planejamento e robótica: projetos de robôs que sejam capazes de realizar tarefas com algum grau de flexibilidade e resposta em relação ao mundo externo;
- Linguagens e ambientes para IA: linguagens de programação e ambientes de desenvolvimento de *software*;
- Aprendizado de Máquina: sistemas capazes de adquirir conhecimento de forma automática seja por experiência, exemplos anteriores, analogia ou por um especialista que diga o que fazer. Porém, ao contrário de um ser humano, que ao resolver um problema uma vez, pode não se lembrar de como resolvê-lo novamente em outra oportunidade, o sistema baseado em Aprendizado de Máquina realizará a mesma tarefa sem se “esquecer”;
- Representações alternativas (redes neurais e algoritmos genéticos): utiliza modelos que imitam a estrutura de neurônios do cérebro humano para construir programas inteligentes.

Essa pesquisa envolveu, em particular, algoritmos de Aprendizado de Máquina sendo, então, o assunto abordado na próxima seção.

2.2.2 Aprendizado de Máquina (AM)

Desde que os computadores foram inventados, se questionou se eles poderiam aprender. Então, buscou-se entender se seria possível programá-los para aprender e melhorar automaticamente com experiências adquiridas. O campo de Aprendizado de Máquina (AM) é o ramo da Inteligência Artificial que está preocupado com a questão de como construir programas de computador, algoritmos e técnicas que permitam o computador aprender automaticamente com experiências acumuladas para, então, tomar decisões (PEREIRA, 2017; MONARD; BARANAUSKAS, 2003; MITCHELL, 1997).

O AM tipicamente busca auxiliar na análise de conjuntos de dados extensos (LIBBRECHT; NOBLE, 2015; GUIMARÃES, 2005), pois assume que máquinas aprenderam com estes dados e, a partir de experiências, ajustaram-se e adaptaram-se às estruturas existentes. Considera-se que o AM está relacionado, principalmente, a três características: aprender, adaptar e generalizar. Ao modelar as ações, aprender e adaptar em um computador, o Aprendizado de Máquina permite que, conforme dados são inseridos e experiências adquiridas, o computador adapte suas ações para que essas sejam mais precisas. A generalização é reconhecer a similaridade entre situações diferentes, de modo que as coisas aplicadas em um lugar possam ser usadas em outro. O principal fator que torna o Aprendizado de Máquina útil é a generalização, que torna possível utilizar o conhecimento adquirido em lugares e situação diferentes (MARSLAND; 2015).

Considerando como exemplo uma pessoa jogando contra um computador, é possível que a pessoa vença o computador por diversas vezes nos primeiros jogos, mas ao aumentar a quantidade de partidas disputadas, o computador irá começar a derrotar a pessoa, até que ela não ganhe mais nenhuma vez. Ao aprender a derrotá-lo, o computador continuará a utilizar as estratégias contra outros jogadores, adaptando-se e reconhecer a similaridade entre situações diferentes, que é uma forma de generalização (MARSLAND; 2015).

Do ponto de vista computacional, o AM tenta fazer com que os programas de computador “aprendam” com os dados que eles “estudam”, tal que esses programas tomem decisões diferentes baseadas nas características dos dados estudados, usando a estatística para os conceitos fundamentais e adicionando

heurística avançada da Inteligência Artificial aos algoritmos para alcançar seus objetivos (PEREIRA, 2017; GUIMARÃES, 2005).

Computadores são capazes de aprender, a partir de registros médicos, os tratamentos que são mais eficazes para novas doenças. O AM vem sendo muito utilizado em programas de mineração de dados que aprendem a detectar uso suspeito de cartões de crédito, a sistemas de arquivamento de informações que aprendem a ler preferências, a veículos autônomos que aprendem a dirigir em rodovias públicas (MITCHELL, 1997).

O Aprendizado de Máquina, definido de forma ampla, inclui qualquer programa de computador que melhore seu desempenho em alguma tarefa por meio da experiência. Um programa de computador aprende com a experiência E em relação a alguma classe de tarefas T e medida de desempenho P , se seu desempenho em tarefas em T , conforme medido por P , melhora com a experiência E (MITCHELL, 1997).

Um algoritmo de Aprendizado de Máquina pode descobrir por si próprio uma função que conecta um conjunto de entradas X a um conjunto de saídas Y dado um conjunto suficientemente grande de exemplos rotulados mapeando algumas das entradas para saídas. O Aprendizado de Máquina transformou as habilidades das máquinas para executar uma série de tipos básicos de percepção que permitem um conjunto mais amplo de aplicativos. Considera-se a visão de máquina como a capacidade de reconhecer objetos, rotulá-los em fotos e interpretar padrões vídeos (BRYNJOLFSSON; ROCK; SYVERSON, 2017)

2.3 Reconhecimento de Padrões

Enquanto o Aprendizado de Máquina teve sua origem na IA, o Reconhecimento de Padrões cresceu a partir da engenharia. Porém, existem muitos trabalhos nos últimos anos que envolvem as duas áreas, de tal forma que o Aprendizado de Máquina é utilizado como ferramenta para o Reconhecimento de Padrões (BISHOP, 2006).

Padrões são os meios pelos quais o mundo é interpretado, sendo que para reconhecer esses padrões, o indivíduo precisa estar vinculado a estímulos aos quais foi exposto anteriormente. O Reconhecimento de Padrões e o Aprendizado de

Máquina associam-se com o objetivo de desenvolver máquinas que sejam capazes de reconhecer esses padrões (CASTRO; PRADO, 2002).

Os padrões permitem a extração de características relevantes de objetos, de tal modo que seja possível agrupar ou classificar características semelhantes dentro de uma determinada categoria, mediante a interpretação de dados de entrada. O reconhecimento de padrões é um procedimento de identificação de estruturas nos dados de entrada, que são comparados com estruturas já conhecidas e, posteriormente, agrupados ou classificados dentro de categorias, de modo que o nível de associação entre estruturas de mesma categoria seja maior e entre as categorias de estruturas diferentes seja menor (CASTRO; PRADO, 2002).

O reconhecimento de padrões é definido por Bishop (2006) como uma busca automática por descobrir regularidades em dados por meio de algoritmos de computador, sendo que essas regularidades são usadas para realizar ações como classificar dados em diferentes categorias ou agrupá-los.

Métodos e algoritmos de AM para reconhecimento de padrões são capazes de adquirir conhecimento a partir de dois tipos elementares de aprendizado: o aprendizado supervisionado e o não-supervisionado. Conforme a Figura 7, estes dois paradigmas de aprendizado oferecem, respectivamente, fundamentos para o estudo e desenvolvimento dos chamados classificadores e agrupadores (WITTEN, FRANK, 2005).

Figura 7 – Tipos de algoritmos de reconhecimento de padrões.



Fonte: Elaborada pela autora.

Em uma abordagem supervisionada, os dados passam por uma fase inicial de treinamento. Tomando como exemplo o reconhecimento de padrões em imagens, o resultado da execução do algoritmo de Aprendizado de Máquina pode ser expresso como uma função $f(x)$ que recebe como entrada uma nova imagem, por exemplo, de um objetivo qualquer x , e que gera uma saída y , que indica a classe a qual a imagem

ou objeto pertence. A forma precisa da função $f(x)$ é determinada durante a fase de treinamento, também conhecida como fase de aprendizagem, com base em dados de treinamento. Uma vez que o modelo é treinado, ele pode determinar a identidade de novas imagens digitais, que dizem incluir um conjunto de testes. A capacidade de categorizar corretamente novos exemplos que diferem daqueles usados para treinamento é conhecida como generalização (BISHOP, 2006). Portanto, a classificação de dados é um processo que consiste de duas etapas, em que inicialmente, um modelo de classificação é induzido a partir do conjunto de treinamento, no qual cada objeto é rotulado de acordo com a classe à qual pertence. Na segunda etapa, o classificador obtido na primeira etapa é utilizado para prever a classe de novos objetos não rotulados - e não observados durante o treinamento (ALPAYDIN, 2010).

Considerando a abordagem não supervisionada, os dados de treinamento consistem em um conjunto de vetores de entrada x sem nenhum valor alvo correspondente (BISHOP, 2006). A ausência de supervisão se refere ao fato de que, diferentemente de um problema de classificação, os rótulos de classes não estão disponíveis para se construir o modelo. Nestes casos, o conceito de conjunto de treinamento tem pouca importância prática, e pode-se denotar o conjunto de dados simplesmente por X , o qual é particionado de tal forma a se obter um conjunto de grupos (*clusters*), também chamado de agrupamento (*clustering*). A forma pela qual os grupos são induzidos depende do algoritmo de agrupamento (agrupador) usado. Em termos gerais, usualmente procura-se induzir uma partição cujos grupos possuam objetos similares (JAIN; DUBES, 1988). Mais especificamente, e considerando-se um contexto de otimização, tipicamente se busca maximizar a homogeneidade entre os objetos de um mesmo grupo e, concomitantemente, maximizar a heterogeneidade entre objetos de grupos distintos.

A abordagem não supervisionada é a utilizada neste trabalho, portanto os dados disponíveis não possuem classes a priori. Os métodos de agrupamento buscarão relacionar padrões segundo o critério de similaridade e o viés assumido pelo algoritmo empregado. Em particular, foram abordados três algoritmos não supervisionados, dos quais serão descritos nas próximas seções.

2.3.1 *K-means*

O *K-means* (ou K-médias) é um algoritmo de agrupamento de dados proposto pela primeira vez em 1957 por Stuart Lloyd, em uma técnica de modulação por código de pulso, porém a proposta foi publicada apenas em 1982. O termo *K-means* foi utilizado pela primeira vez por James MacQueen em 1967, mas foi no período de 1975 e 1979 que Hartigan e Wong publicaram uma proposta de versão mais eficiente do método (SHAFEEQ; HAREESHA, 2012).

O objetivo do algoritmo é dividir n objetos em k grupos, buscando a minimização da distância total entre os elementos de um grupo. O algoritmo atribui a k grupos, n elementos, e calcula as médias dos vetores de cada grupo. Neste método, k é o número de *clusters* definido pelo usuário, em seguida, os centroides se deslocam de acordo com cada grupo de vetor médio correspondente do qual ele esteja mais próximo (PIMENTEL; FRANÇA; OMAR, 2003).

O centro do *cluster* é formado a partir dos dados mais próximos, e comparados com os outros *clusters* formados, seguido por um processo contínuo de atualização e de iteração que torna possível encontrar os centros finais dos *clusters* (SILVA; PORTUGAL; CECHIN, 2001). Este processo continua até todos os elementos se encontrarem nos seus vetores médios mais próximos (PIMENTEL; FRANÇA; OMAR, 2003).

O *K-means* é um método de particionamento iterativo, em que o conjunto de dados é particionado criando conglomerados e fazendo várias iterações nesse conjunto (WIVES, 2004). O algoritmo designa um conjunto inicial de partições a partir da quantidade de *clusters* indicado pelo usuário. A partir do centro de cada uma destas partições, o computador calcula a similaridade desses elementos com os outros a serem agrupados e o processo iterativo continua até que os centros não mudem mais de posição.

No entanto, esse algoritmo é sensível à escolha dos centroides iniciais e dependente da informação do número de grupos (k), o qual deve ser fornecido pelo usuário. Além disso, o *K-means* produz bons resultados apenas quando não há sobreposição de grupos, i.e., quando os grupos são disjuntos (compactos e bem separados). Para tentar superar essas limitações, diversas variantes do *K-means* têm sido desenvolvidas (COLETTA, 2011).

O algoritmo *K-means* busca minimizar o erro quadrático (soma das distâncias médias entre os centroides e objetos dos grupos correspondentes). Nesse algoritmo, os centroides são os pontos médios de cada grupo que são recalculados a cada iteração. O *K-means* determina uma partição com k grupos globulares utilizando-se da medida de distância euclidiana (COLETTA, 2011). Os principais passos do algoritmo *K-means* são (COLETTA, 2011):

1. Seja k o número de grupos;
2. Selecione os centroides de grupos iniciais, (v_1, v_2, \dots, v_k) ;
3. Compute as distâncias $\|x_j - v_i\|$ entre os objetos e os centroides;
4. Atribua cada objeto ao grupo de centroide mais próximo;
5. Recalcule os centroides dos grupos de acordo com os seus objetos;
6. Pare se o critério de convergência¹ foi atingido ou o número de iterações excedeu um dado limite. Caso contrário, volte ao passo 3;

A complexidade do algoritmo *K-means* é estimada em $O(k \cdot N \cdot d \cdot n_t)$, em que k é o número de grupos, N é o número de objetos, d é o número de atributos e n_t o número de iterações (COLETTA, 2011).

A principal vantagem deste método está no fato das diversas iterações realizadas no conjunto de dados corrigirem possíveis problemas de alocação (ALDENDERFER; BLASHFIED, 1984). Como desvantagem, os autores citam o fato do número dos aglomerados ter que ser especificado pelo usuário, já que não é possível prever, para determinados conjuntos de dados, qual seria o mais adequado. De acordo com os autores, uma das formas de minimizar esta desvantagem é executar o método por diversas vezes, testando todas as configurações possíveis de aglomerados.

Nos últimos cinco anos o método *K-means* foi muito aplicado, sendo que algumas dessas foram publicadas em periódico de elevado fator de impacto, estão sendo citadas no Quadro 1.

Quadro 1 - Exemplos de aplicações do *K-means* nos últimos cinco anos.

Autor(es) (ano da publicação)	Aplicação do método <i>K-means</i>
AY; KISI (2014)	Este artigo propõe a integração dos métodos <i>K-means clustering</i> e <i>multi-layer perceptron (K-means-MLP)</i> na modelagem da demanda química

	de oxigênio (DQO). Este método proposto foi testado usando dados medidos diários de sólidos em suspensão, pH, temperatura, descarga e concentração de COD a montante do sistema de tratamento de águas residuais municipais na província de Adapazari, na Turquia.
NALDI; CAMPELLO (2014)	Propuseram o uso de algoritmos evolutivos para superar as limitações de <i>K-means</i> e, ao mesmo tempo, lidar com dados distribuídos em repositórios separados, pois a maioria das técnicas de <i>cluster</i> requer que os dados sejam centralizados.
ZUPERKU; PRKIC; STUCKE; MILLER HOPP; STUTH (2015)	Este relatório apresenta um método que classifica automaticamente os neurônios de acordo com seus padrões de descarga e deriva um contorno médio de subgrupo de cada classe. Ele é baseado na técnica de clusterização <i>K-means</i> .
GOUVA; KOTROTSIOU; GOURGOULIANNIS; SKENTERIS (2015)	Foi utilizada a análise de agrupamentos <i>K-means</i> para identificar a percepção de pessoas que vivem na área central da Grécia em relação ao sistema público de saúde
LI; SUN; JIA; CAI; WANG (2016)	Foi estabelecido um modelo integrado baseado na análise <i>K-means</i> de agrupamento e análise de pares, com o objetivo de avaliar os riscos associados à poluição da água em áreas de água de nascente na região Shiyan, na China.
JAVADI; HASHEMY; MOHAMMADI; HOWARD; NESHAT (2017)	Neste estudo, é introduzida uma técnica de <i>clustering</i> que classificação da vulnerabilidade do aquífero no Irã usando análise de <i>cluster K-means</i> .
BORGWARDT; BRIEDEN; GRITZMANN (2017)	Foi apresentada e analisada a generalização do <i>K-means</i> , que é capaz de lidar com conjuntos de pontos ponderados e limites inferiores e superiores prescritos em os tamanhos de <i>cluster</i> , que é chamada K-balanceados por peso.

SANTOS; GALO; TACHIBANA (2018)	Foi proposto o uso de métricas estimadas a partir do método K-médias para classificação na extração de primitivas geométricas sobre dados LiDAR (Detecção e Alcance de Luz)
LIU; YANG; HAO; ZHANG (2018)	Foi realizada uma avaliação de eficiência energética informada por grandes volumes de dados dos setores industriais da China com base no agrupamento <i>K-means</i>

Fonte: Elaborado pela autora.

2.3.2 Fuzzy C-means (FCM)

O método *Fuzzy C-means* (FCM) foi desenvolvido por James Bezdek em 1981, como um algoritmo de agrupamento não supervisionado que é aplicado a uma ampla gama de problemas, principalmente nas áreas da em engenharia agrícola, astronomia, química, geologia, análise de imagem, diagnóstico médico, análise de formas e reconhecimento de metas, padrões e mineração de dados (GOSH; DUBEY, 2013; PIMENTEL; SOUZA, 2013).

A aplicação da lógica difusa a métodos de agrupamento resultou em muitas técnicas amplamente utilizadas, como o método *Fuzzy C-means* (FCM) (ZARINBAL; ZARANDI; TURKSEN, 2014). Bezdek, Ehrlich e Full (1984) explicam que o FCM gera partições difusas e protótipos para qualquer conjunto de dados numéricos, sendo que essas partições são úteis para confirmar subestruturas conhecidas ou sugerir subestrutura em dados que ainda não foram explorados.

Considera-se que a análise de *cluster* se refere a um amplo espectro de métodos que tentam subdividir um conjunto de X dados em C subconjuntos, que são os chamados *clusters*. Os aglomerados são então denominados de uma partição C dura, neste caso, ainda não pode ser chamado de *Fuzzy*. O fato é que algoritmos que formam partições “duras” possuem uma limitação, pois cada ponto em X pode ser inequivocamente agrupado com outros membros do *cluster* sem necessariamente apresentar similaridade aparente com outros membros de X (BEZDEK; EHRLICH; FULL, 1984).

Uma dessas maneiras de demonstrar a semelhança existente entre um ponto individual com outros *clusters* formados em determinado agrupamento foi introduzida em 1965 por Zadeh. Zadeh buscou representar a similaridade que um ponto compartilha com cada *cluster* a partir de uma função (denominada função de associação) cujos valores (chamados de associações) variam entre zero e um. Desta forma, cada amostra dos dados disponíveis terá uma associação em cada *cluster*, as associações próximas à ao número um implicam um alto grau de similaridade entre a amostra e um *cluster*, enquanto as associações próximas a zero significam pouca similaridade entre a amostra e esse *cluster*. Sendo assim, é produzidas partições *C* difusas (*fuzzy*) de um determinado conjunto de dados. Além disso, a soma das associações para cada ponto de amostra deve ser igual a um (BEZDEK; EHRLICH; FULL, 1984).

A partir de uma base de dados de entrada, o FCM forma os *clusters* e os centros de todos os *clusters*, a partir da análise da distância entre os vários pontos dos dados de entrada (GOSH; DUBEY, 2013). Um conjunto de dados é agrupado em “*C*” *clusters*, sendo que cada ponto terá um grau de pertinência ou conexão com o *cluster* a que ele pertence. Quanto mais próximo o ponto está do centro do *cluster*, maior é o grau de conexão com ele, sendo que quanto mais distante, menor será a conexão entre o ponto e seu *cluster* (GOSH; DUBEY, 2013).

Em contraste com o método de agrupamento de *K-means*, o *Fuzzy C-means* envolve o parâmetro adicional, que é chamado de “fuzzificador”. Um ponto de dados não é diretamente atribuído a um único *cluster*, sendo possível que um ponto de dado seja membro de todos os *clusters*. Isso permite diminuir o efeito de objetos de dados que não pertencem a um *cluster* específico, por exemplo, objetos localizados entre *clusters* sobrepostos (SCHWAMMLE; JENSEN, 2010).

A maioria dos modelos convencionais não possui mecanismo natural para identificar a sobreposição de dados e que o método de partição *fuzzy* foi introduzido como um meio de alterar os axiomas básicos subjacentes aos modelos de agrupamento e classificação, com o objetivo de acomodar essa necessidade (BEZDEK; EHRLICH; FULL, 1984). Os autores explicam que uma amostra pode pertencer inteiramente a um único *cluster*, geralmente, ela é capaz de se tornar membro parcial de vários *clusters fuzzy*.

O algoritmo FCM é uma extensão do algoritmo *K-means* para o contexto *fuzzy*. Enquanto o algoritmo *K-means* assume que cada objeto deve pertencer

exclusivamente a um único grupo, o algoritmo FCM permite relaxar essa restrição, possibilitando que cada objeto pertença, sob graus diferentes, a cada um dos grupos. Dessa maneira, esse algoritmo tem como principal vantagem a identificação de grupos sobrepostos. O algoritmo FCM usa protótipos como representantes dos grupos. Tais protótipos guardam informações análogas àquelas armazenadas pelos centroides computados pelo *K-means*. No entanto, cada protótipo é computado levando-se em conta todos os objetos e suas respectivas pertinências aos grupos.

Os principais passos do Fuzzy C-Means (FCM) (Bezdek, 1981; Dunn, 1973) estão sumarizados nos seguintes passos:

1. Seja c o número de grupos;
2. Selecione os protótipos de grupos iniciais, (v_1, v_2, \dots, v_c) ;
3. Compute as distâncias $\|x_j - v_i\|$ entre os objetos e os protótipos;
4. Compute os elementos (pertinências) da matriz de partição *fuzzy*:

$$u_{ij} = \left[\sum_{p=1}^c \left(\frac{\|X_j - V_i\|}{\|X_j - V_p\|} \right)^{2/(m-1)} \right]^{-1}$$

em que, u_{ij} é um elemento da matriz de partição fuzzy $U \in R^{c \times N}$.

5. Compute os protótipos de grupos:

$$v_i = \frac{\sum_{j=1}^N u_{ij}^m X_j}{\sum_{j=1}^N u_{ij}^m}$$

6. Pare se o critério de convergência foi atingido ou o número de iterações excedeu um dado limite. Caso contrário, volte ao passo 3;

A sua complexidade é estimada em $O(c^2 \cdot N \cdot d \cdot n_t)$, onde c é o número de grupos, N é o número de objetos, d é o número de atributos e n_t o número de iterações.

Muitas foram as pesquisas e aplicações do método *Fuzzy C-means* nos últimos cinco anos, sendo que algumas delas, publicadas em periódico de elevado fator de impacto, estão sendo citadas no Quadro 2.

Quadro 2 - Exemplos de aplicações do FCM nos últimos cinco anos

Autor(es) (ano da publicação)	Aplicação do método FCM
YIN; SUN; YANG; GUO (2014)	Foi realizada uma comparação do desempenho dos algoritmos <i>K-means</i> e <i>Fuzzy C-means</i> para

	determinação automatizada da função de entrada arterial
ZARINBAL; ZARANDI; TURKSEN (2014).	Apresentaram um novo método de agrupamento <i>fuzzy</i> baseado no FCM com o objetivo de definir uma função de regularização para maximizar a dissimilaridade entre <i>clusters</i> .
STETCO; ZENG; KEANE (2015)	Propuseram o método <i>Fuzzy C-means ++</i> para melhorar a eficácia e a velocidade do algoritmo <i>Fuzzy C-means</i>
TANG; ISA; CH'NG (2015)	Este estudo empregou o agrupamento <i>Fuzzy C-means</i> como técnica de segmentação Para quantificar os diferentes graus de segmentação da cromatina em imagens de amostras de células escamosas não neoplásicas em testes de Papanicolau.
BAI; DHAVALI; SARKIS (2016)	O estudo propôs uma metodologia baseada no FCM para ajudar a gerenciar investimentos e ajudar as organizações a integrar atividades para melhorar o desempenho ambiental natural de suas cadeias de suprimentos
PARASTAR; BAZRAFESHAN (2016)	O agrupamento <i>Fuzzy C-means</i> é proposto como um método promissor para o agrupamento de impressões digitais cromatográficas de amostras complexas, como óleos essenciais.
MAHELA; SHAIK (2017)	Este artigo apresenta uma técnica para reconhecer os distúrbios de qualidade de energia associados à penetração da energia solar na rede de distribuição com a ajuda do agrupamento <i>Fuzzy C-means</i> .
BASER; GOKTEN S.; GOKTEN P. (2017)	Foi sugerida a utilização do algoritmo de clusterização FCM para produzir pontuações de saúde financeira de empresas, especialmente para decisões de investimento de curto prazo, usando números contábeis recentemente anunciados.

XUE; ZHOU; KOJIMA; MUCHANGOS; MACHIMURA; TOKAI (2018)	Aplicação de aglomeração <i>Fuzzy C-means</i> em produtos químicos PRTR descobrindo suas características de liberação e transferência de poluentes e toxicidade.
WANG; WU; STEIN; ZHU; ZENG (2018)	Este artigo apresenta um método para prever a profundidade do solo por meio da construção de uma função de associação baseada em FCM.

Fonte: Elaborado pela autora.

2.3.3 Floresta de Caminhos Ótimos (OPF)

O OPF (*Optimum Path Forest*) se consolidou em 2008 na tese de doutorado de João Paulo Papa, baseando-se na *Image Foresting Transform (IFT)* - Transformada Imagem Floresta, criada por Falcão, Stolfi e Lotufo (2004), que é uma ferramenta que reduz problemas de processamento da imagem ao cálculo de uma floresta de caminhos de custo ótimo em um grafo derivado da imagem. Enquanto a IFT se limita ao domínio de imagens, o OPF pode ser usado para diferentes tipos de dados (HESPANHOL; PEREIRA; FORTES, 2015).

O OPF pode ser utilizado nas versões supervisionadas e não supervisionadas, de acordo com seu algoritmo de aprendizado, como demonstrado na Figura 8 (PAPA, 2008).

Figura 8 – Variações do algoritmo OPF



Fonte: Elaborada pela autora.

Neste trabalho foi adotado o método não supervisionado, buscando o agrupamento dos dados, de acordo com seus padrões, na produção de cana de açúcar. É possível encontrar, por meio do OPF não supervisionado, diversos agrupamentos em um determinado conjunto de dados, sendo dispensável definir a sua quantidade e tornando possível encontrar assim os elementos mais significativos (HESPANHOL, 2016). O método utiliza-se de amostras não rotuladas, considerando a distância entre as características, em que se têm os nós de um grafo no espaço, que são as amostras de dados e os arcos são definidos por uma condição do que é próximo (semelhante).

O algoritmo fundamenta-se na Teoria dos Grafos e em algumas técnicas deste contexto. (SOUZA; LOTUFO; RITTNER, 2012). A Teoria dos Grafos surgiu no século XVIII com o matemático Leonard Euler, e após longo período sem utilização, teve impulso com as aplicações voltadas a problemas de otimização, no século XX, compondo atualmente um conjunto de técnicas com aplicação em diferentes áreas do conhecimento (BOVO, 2004). Um grafo G pode ser compreendido como um par (V, A) , em que V é um conjunto não vazio de vértices $\{v_1, v_2, \dots, v_n\}$ e A uma família (a_1, a_2, \dots, a_m) de elementos pertencentes ao produto cartesiano $V \times V$, chamados de arestas.

O problema de reconhecimento de padrões é modelado como um problema de floresta de caminhos ótimos em um grafo definido no espaço de atributos, onde os nós são as amostras, as quais são representadas pelos seus respectivos vetores de atributos, e os arcos são definidos de acordo com uma relação de adjacência pré-estabelecida (PAPA, 2008). Na abordagem não supervisionada, as amostras são modeladas como sendo os nós de um grafo, cujos arcos conectam os K -vizinhos mais próximos no espaço de atributos. O grafo é ponderado nos nós por valores de densidades originando, assim, uma função de densidade de probabilidade (fdp), a qual é calculada levando-se em consideração as distâncias (peso dos arcos) entre os vetores de atributos de amostras adjacentes. O valor do melhor k é encontrado minimizando uma medida de corte em grafo e a maximização de uma função de valor de caminho origina uma floresta de caminhos ótimos, onde cada árvore (*cluster*) é enraizada em um máximo da fdp (PAPA, 2008).

O OPF fundamenta-se na teoria de que, seja Z uma base de dados, ilustrada na Figura 9, tal que, para toda amostra $s, t \in Z$, existe um vetor de atributos $\vec{v}(s)$, que é representado por uma linha dessa base de dados. Seja $d(s, t)$ uma

distância entre s e t no espaço de atributos. Sendo que, cada linha da base de dados Z , representa um vetor de atributo.

Figura 9 – Ilustração da base de dados Z .

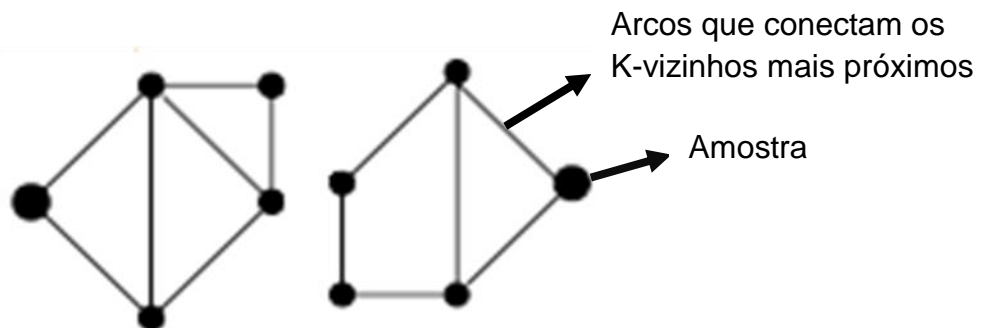
	Atributo 1	Atributo 2	Atributo 3	Atributo 4	Atributo 4
Amostra s	0000	0000	0000	0000	0000
Amostra t	0000	0000	0000	0000	0000

} **Base de dados Z**

Fonte: Elaborada pela autora.

Em uma dimensão no espaço, as amostras da base de dados Z é representada pelos nós da Figura 10, que representam uma linha da base de dados, ou seja, os vetores de atributos.

Figura 10 – Amostras na dimensão espaço



Fonte: Elaborada pela autora.

O problema fundamental na área de *clustering* é identificar grupos de amostras em Z , sendo que amostras de um mesmo grupo deveriam representar algum nível de semelhança de acordo com algum significado semântico. Uma amostra t é adjacente a uma amostra s quando alguma relação de adjacência é satisfeita (PAPA, 2008). Desta forma, t e s pertencem ao mesmo *cluster* se t é K-vizinho mais próximo de s no espaço de atributos.

Nos últimos cinco anos o método Floresta de Caminhos Ótimos foi muito aplicado, sendo que algumas dessas foram publicadas em periódico de elevado fator de impacto, estão sendo citadas no Quadro 3.

Quadro 3 – Exemplos de aplicações do OPF nos últimos cinco anos

Autor(es) (ano da publicação)	Aplicação do método OPF
PISANI; NAKAMURA; RIEDEL; ZIMBACK; FALCAO; PAPA (2014)	Utilizaram o OPF, supervisionado e não supervisionado, para classificação da cobertura do solo. Os resultados foram comparados com outras técnicas de reconhecimento de padrões.
NUNES; COELHO; LIMA; PAPA; ALBUQUERQUE (2014)	Este estudo testou o método OPF na classificação do sinal eletroencefalograma (EEG) para o diagnóstico de epilepsia.
KUANAR; RANGA; CHOWDHURY (2015)	Foi proposta uma solução teórica de gráficos para os problemas de sumarização de vídeos para que com várias visualizações seja possível representar com eficiência as informações mais significativas de um conjunto de vídeos capturados por um determinado período de tempo por várias câmeras.
COSTA; PEREIRA; NAKAMURA; PEREIRA; PAPA; FALCÃO (2015)	Foi proposta uma abordagem inspirada na natureza para estimar a função de densidade de probabilidade (pdf) usada para agrupamento de dados com base no algoritmo de floresta de caminho ótimo (PFC).
PASSOS JÚNIOR; RAMOS; RODRIGUES; PEREIRA; SOUZA; COSTA; PAPA (2016)	O algoritmo de agrupamento de floresta de caminhos ótimos (OPF) foi empregado para identificar perfis irregulares e regulares de consumidores comerciais e industriais obtidos de uma companhia brasileira de energia elétrica.
PAPA; FERNANDE S; FALCÃO (2017)	Foi realizada uma explicação teórica profunda sobre o classificador de OPF supervisionado e a proposta de dois algoritmos de treinamento e classificação diferentes que permitem que o OPF trabalhe mais rápido.
MARTINS; PEREIRA; ALMEIDA; PAPA (2018)	Este trabalho lida com o problema de sumarização de vídeo estático usando o OPF não supervisionado.

CHEN; SUN; YANG; SUN; GUAN (2018)	Este artigo propõe um novo algoritmo de agrupamento de Floresta de Caminhos Ótimos (OPF) que pode ser usado para segmentação por sensoriamento remoto. O método utiliza o princípio de que os centros de <i>cluster</i> são caracterizados com base em suas densidades e as distâncias entre os centros e as amostras com densidades mais altas.
--------------------------------------	--

Fonte: Elaborado pela autora.

3 METODOLOGIA

O trabalho foi conduzido de maneira quantitativa, que Dalfovo, Lana e Silveira (2008) definem como tudo que pode ser medido em números, classificados e analisados, utilizando-se, por exemplo, de técnicas estatísticas. A partir de dados quantitativos, o pesquisador possui uma base mais segura para tirar conclusões, sendo esta tarefa intimamente ligada às ferramentas de Inteligência Artificial (IA) que buscam analisar grandes quantidades de dados para extrair conhecimento útil e auxiliar na tomada de decisões (MASCARENHAS, 2012).

Inicialmente, para a coleta de dados, foi realizada uma pesquisa documental, que busca identificar informações sobre os fatos a partir de documentos (LUDKE; ANDRÉ, 1986). Gauthier (1984) complementa que esse tipo de pesquisa elimina qualquer influência do pesquisador sobre o conjunto dos acontecimentos, impossibilitando o fornecedor da informação de reagir à maneira como os dados são obtidos.

O levantamento de dados secundários foi realizado apenas por meio de instituições oficiais do estado de São Paulo. Primeiramente junto à CATI e ao IEA, sendo possível obter informações sobre a área cultivada (hectare), produção de cana de açúcar (tonelada) e produtividade no ano de 2017 em todos os municípios do estado de São Paulo que apresentaram produção neste período. A CATI é um órgão da Secretaria de Agricultura e Abastecimento do Governo do estado de São Paulo, criada em 1967, e trabalha na prestação de serviços de suporte ao produtor rural de maneira sustentável, por meio de ações e programas que envolvem a sociedade e entidades parceiras (CATI, 2018). O IEA é o braço econômico da Agência Paulista de Tecnologia dos Agronegócios (APTA), pertencente a Secretaria de Agricultura e Abastecimento do estado de São Paulo (IEA, 2019).

O estado de São Paulo possui 645 municípios, sendo que na primeira coleta de dados, junto a CATI e ao IEA, referente a área, produção e produtividade, foram identificados 535 municípios com produção de cana de açúcar no estado de São Paulo nos últimos 10 anos. Exatamente estes municípios fizeram parte desta pesquisa (ficaram de fora da pesquisa os municípios listados no Quadro 4).

Quadro 4 – Municípios do estado de São Paulo que não fazem parte desta pesquisa

Águas de São Pedro	Guarulhos	Ribeirão Grande
Alfredo Marcondes	Ibiúna	Ribeirão Pires
Aumínio	Iguape	Rio Grande da Serra
Álvaro de Carvalho	Ilha Comprida	Salesópolis
Alvinlândia	Itanhaém	Salto de Pirapora
Aparecida	Itapecerica da Serra	Santa Isabel
Apiaí	Itapevi	Santana de Parnaíba
Araçariguama	Itapirapuã Paulista	Santo André
Arujá	Itaquaquecetuba	Santo Antônio do Jardim
Barão de Antonina	Itararé	Santo Antônio do Pinhal
Barra do Chapéu	Itariri	Santos
Barueri	Jambeiro	São Bento do Sapucaí
Bertioga	Jandira	São Bernardo do Campo
Biritibamirim	Juquiá	São Caetano do Sul
Bom Jesus dos Perdões	Juquitiba	São José dos Campos
Bom Sucesso de Itararé	Lavrinhas	São Lourenço da Serra
Caieiras	Lorena	São Miguel Arcanjo
Cajamar	Lupércio	São Paulo
Campo Limpo Paulista	Mairinque	São Roque
Campos do Jordão	Mairiporã	São Vicente
Canas	Mauá	Sete Barras
Capão Bonito	Miracatu	Suzano
Caraguatatuba	Mogi das Cruzes	Taboão da Serra
Carapicuíba	Mongaguá	Tapiraí
Cotia	Monteiro Lobato	Taquarivaí
Cruzeiro	Morungaba	Taubaté
Cubatão	Nova Campina	Torre de Pedra
Cunha	Osasco	Tremembé
Diadema	Pedro de Toledo	Tuiuti
Embu	Pilar do Sul	Ubatuba
Embu-Guaçu	Pinhalzinho	Valinhos
Ferraz de Vasconcelos	Piquete	Vargem Grande Paulista
Francisco Morato	Pirapora do Bom Jesus	Várzea Paulista
Franco da Rocha	Poá	Vera Cruz
Garça	Potim	Vinhedo
Guapiara	Praia Grande	Votorantim
Guarujá	Ribeirão Branco	

Fonte: Elaborado pela autora.

Logo após a coleta e tabulação dos dados obtidos por meio da CATI, foram coletados também os índices pluviométricos (mm) e de temperatura de cada município obtidos por meio do Sistema de Monitoramento Agrometeorológico (Agritempo) disponibilizados pelo Governo Federal. O Agritempo é um sistema online criado e

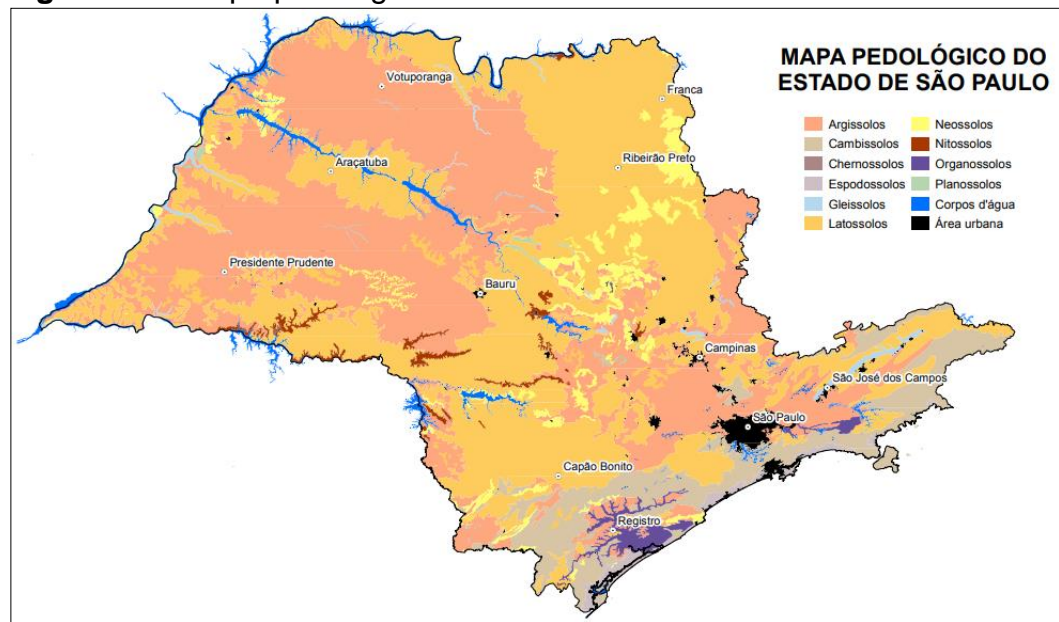
mantido pela Embrapa e pelo Centro de Pesquisas Meteorológicas e Climáticas Aplicadas à Agricultura (Cepagri), vinculado à Universidade Estadual de Campinas (Unicamp). O objetivo do sistema é realizar monitoramento climatológico e meteorológico e assim, produzindo e disponibilizando boletins e mapas com informações sobre estiagem agrícola, precipitação acumulada, necessidade de irrigação, tratamentos fitossanitários, condições de manejo do solo e de aplicação de defensivos agrícolas (EMBRAPA, 2018).

Para a coleta de dados referente aos índices pluviométricos e de temperatura, junto ao Agritempo, dos 535 municípios estudados nem todos possuem estações meteorológicas, sendo que são 255 os que possuem e estão listados no APÊNDICE A. Para os municípios que não possuem estações meteorológicas, foram pesquisados municípios limítrofes a eles com estação, para que assim, os índices obtidos na pesquisa fossem os mais próximos da realidade possível. Desta forma, os municípios que não possuem estações meteorológicas foram listados e estão disponíveis no APÊNDICE B, com os respectivos municípios limítrofes utilizados.

Também foi coletado como dado o tipo de solo predominante de cada município por meio do Instituto Agrônomo (IAC) e da CATI. De acordo com o IAC (2018), o tipo de solo pode variar de acordo com a localização, que pode ser influenciado pela paisagem em relação direta com o relevo. Existem solos mais intemperizados, como Latossolos e Nitossolos, que são características típicas de relevos mais suavizados e com menores taxas de erosão natural (IAC, 2018). Além disso, solos do tipo Cambissolos e de Neossolos são características de solos mais rasos e com maiores teores de minerais e planícies de inundações frequentes (IAC, 2018). Os solos do tipo Gleissolo, Espodossolos e Organossolos são característicos de costeiras ou do interior, planícies, com lençol freático mais elevado. Localizações com relevos mais ondulados normalmente são Argissolos e Planossolos (IAC, 2018). O IAC (2018) ressalta ainda que essas tendências podem variar em decorrência de fatores como rochas ou camadas de rochas muito resistentes à alteração ou condição microclimática superúmida.

Para extrair o dado do tipo de solo predominante de cada município, o mapa demonstrado na Figura 11 foi analisado pela ferramenta do Google Earth, que permite pesquisar a localização exata de cada município no mapa. Além disso, os dados do mapa foram confrontados com o Plano Municipal de Desenvolvimento Rural Sustentável, elaborado pela CATI em 2013 para cada município do estado.

Figura 11 – Mapa pedológico do estado de São Paulo



Fonte: IAC (2018).

Estes dados foram organizados em planilha eletrônica. Os dados foram organizados por municípios, por ordem alfabética, sendo desta forma a primeira coluna destinada para o município, a segunda destinada para a área cultivada, a terceira para a produção e a quarta para a produtividade de cada município, que foi representada pela divisão da produção pela área. A quinta coluna correspondente ao índice pluviométrico (mm), a sexta representa a temperatura média (°C) e a última ao valor numérico atribuído ao tipo de solo.

Para utilizar o dado do tipo de solo nos algoritmos de reconhecimento de padrões, torna-se necessário atribuir valores numéricos para cada um deles. Como apenas cinco tipos de solos, dos nove classificados pelo IAC, foram identificados como predominantes nos municípios estudados, apenas eles foram classificados numericamente. A classificação variou de um a cinco, levando em consideração que o solo com classificação um é menos indicado para o plantio de cana de açúcar e o solo com classificação cinco, sendo mais indicado para cultura. Essa classificação foi feita com base em dados existentes na bibliografia e também em um questionário aplicado com um especialista que contribuiu com essa classificação, demonstrada no Quadro 5.

Quadro 5 – Classificação dos tipos de solos.

Tipo de solo	Classificação
Latossolos	5
Argissolos	4
Cambissolos	3
Neossolos	2
Espodossolos	1

Fonte: Elaborado pela autora

O questionário foi aplicado a dois engenheiros agrônomos da CATI Regional de Dracena-SP e, além da contribuição com a classificação do solo, o questionário também objetivou coletar possíveis contribuições do engenheiro agrônomo com relação ao que se espera como características para possíveis padrões obtidos nos resultados. O questionário utilizado está apresentado no APÊNDICE C.

O objetivo do questionário é que os especialistas indiquem qual a influência dos fatores edafoclimáticos na produtividade da cana de açúcar, sendo as questões referentes à:

- Quais os tipos do solo influenciam para uma alta produtividade e para baixa produtividade da cana de açúcar;
- Qual a faixa de chuva (muita, média ou baixa) influencia para maior produtividade da cana de açúcar;
- Qual a faixa de temperatura (alta, média ou baixa) influencia para maior produtividade da cana de açúcar.

Logo após a tabulação dos dados e a aplicação do questionário com os especialistas, os dados serão processados e análises descritivas serão conduzidas por meio do agrupamento de dados.

Para processamento dos dados no *K-means*, foi utilizado o programa de computador *Elki-Data Mining Framework*, em que foi utilizado como documento de entrada a planilha de dados em formato csv.

Como o algoritmo *K-means* possui como pré-requisito o estabelecimento do número de *clusters* do que será obtido com o processamento, foram realizados testes em busca de identificar a melhor quantidade de *cluster* para a base de dados disponível, por meio do coeficiente de silhueta. O coeficiente de silhueta avalia a coesão dos *clusters* por meio da proximidade dos pontos do *cluster*. O valor deste coeficiente varia entre $[-1, 1]$, sendo que quanto maior este valor, maior é a qualidade

do agrupamento (NUNES, 2016). O cálculo da silhueta já está presente nas configurações do *Elki-Data Mining Framework*.

Desta forma, foi escolhido o resultado com seis *cluster*, devido a ele ter obtido o maior coeficiente de silhueta, como demonstrado no Quadro 6.

Quadro 6 – Coeficiente de silhueta.

Número de <i>clusters</i>	Coeficiente de Silhueta
4	0,5880
5	0,5909
6	0,5979
7	0,5962

Fonte: Elaborado pela autora.

Tipicamente, o *K-means* pode não produzir bons resultados quando há sobreposição de grupos, ou seja, quando os grupos não são disjuntos (com pouca compactação e separabilidade). Para tentar superar essa limitação, diversas variantes do *K-means* têm sido desenvolvidas (COLETTA, 2011). O *Fuzzy C-means*, usado neste trabalho, é uma delas.

Após processamento dos dados no *K-means*, os mesmos dados foram processados no *Fuzzy C-means*. Assim como o *K-means*, o FCM é sensível à escolha dos protótipos iniciais, além de depender da informação do número de grupos (*c*). Desta forma, com o objetivo de padronizar os resultados e facilitar as análises foi estabelecido que o número de grupos do FCM será o mesmo do *K-means*.

Para processamento dos dados no *Fuzzy C-means*, também foi utilizado o programa de computador *Elki-Data Mining Framework*, tendo como documento de entrada a planilha de dados em formato csv. Após o processamento dos dados e obtenção dos resultados, foram elaborados gráficos para auxiliar na comparação dos resultados de cada *cluster* e realizadas as mesmas análises estatísticas do método anterior.

O último processamento de dados foi realizado no algoritmo de reconhecimento de padrões Floresta de Caminhos Ótimos, ou *Optimum-Path Forest* (OPF). O OPF foi utilizado como principal procedimento de investigação, sendo a ferramenta utilizada para o estudo é denominada *LibOPF*, na versão 2.1, disponível em novembro de 2016, no *website* do *software LibOPF*, do Instituto de Computação da Universidade de Campinas (UNICAMP, 2018).

Diferentemente de outros métodos, no OPF o próprio algoritmo agrupa os dados determinando a quantidade de *clusters* existentes naquela base de dados. Ao processar os dados na ferramenta *LibOPF*, foram encontrados 4 *clusters* e estes foram demonstrados em gráficos e foram realizadas análises estatísticas buscando extrair maiores conhecimentos.

Partindo da teoria proposta pelo OPF, em que seja Z uma base de dados tal que, para toda amostra $s, t \in Z$, existe um vetor de atributos $\vec{v}(s)$. Todos os dados coletados e tabelados neste trabalho dão origem a base de dados Z que será estudada, e está exemplificado no Quadro 7.

Quadro 7 – Base de dados Z .

Municípios	Área	Produtividade	Índices Pluviométricos	Temperatura média	Amplitude térmica	Tipo de solo
Adamantina	1300	80	1137,03	24,19		4
...
...
...
...
Zacarias	6500	80	1258,79	24,06		5

Fonte: Elaborado pela autora.

As amostras s, t são representadas por cada linha da base de dados Z , sendo Adamantina uma amostra, Zacarias outra amostra, e assim por diante, com cada município de estudo representando uma amostra na base de dados. Cada município, ou linha da base de dados, possui um vetor de atributos. Na prima linha do Quadro 6, podemos identificar os atributos desses municípios, que neste trabalho são a área de produção da cana de açúcar, produtividade, índices pluviométricos, temperatura, amplitude térmica e tipo de solo. Assim, uma amostra (um município) possui seu vetor de atributos representados por esses dados coletados.

Após o processamento dos dados, os resultados foram descritos em três seções, sendo que cada seção foi destinada a descrição dos resultados de um algoritmo de reconhecimento de padrões utilizado neste trabalho. Para cada algoritmo, foram elaborados gráficos para auxiliar na comparação dos resultados de cada *cluster* e realizadas análises.

A primeira análise estatística realizada com os resultados foi a elaboração de gráficos *boxplot*, utilizando o programa MiniTab. No gráfico *boxplot* é identificada a mediana de cada *clusters*, que é representada pela linha no interior na caixa, sendo uma medida comum do centro de seus dados, a caixa da amplitude interquartílica, que representa 50% dos dados, mostra a distância entre o primeiro e o terceiro quartis, e os traços que se estendem de ambos os lados da caixa representam as amplitudes para o fundo de 25% e o topo de 25% dos valores de dados, excluindo *outliers*.

Logo depois, foram tabulados os centroides obtidos no agrupamento juntamente com seus respectivos desvios padrões. Com isso, foi possível determinar se as diferenças entre as médias dos grupos são estatisticamente significativas, sendo que as médias que não compartilham uma letra são significativamente diferentes. As letras para demonstrar se há diferença significativa entre os centróides foram encontradas a partir do Teste de Turkey, que é um método usado em ANOVA para criar intervalos de confiança para todas as diferenças pareadas entre as médias dos níveis dos fatores controlando a taxa de erro global para um nível de significância especificado. O Teste de Turkey também foi realizado utilizando o programa MiniTab.

Após análise estatística, foram elaborados três mapas, cada um deles destinado a um algoritmo usado neste trabalho. Os mapas foram elaborados utilizando o programa de computador Qgis e identificam, no estado de São Paulo, os *clusters* um e seis. O objetivo deste mapeamento é identificar se há diferença de localização entre os municípios que ficaram inseridos no *cluster* com maior média produtiva e os municípios do *clusters* com menor média.

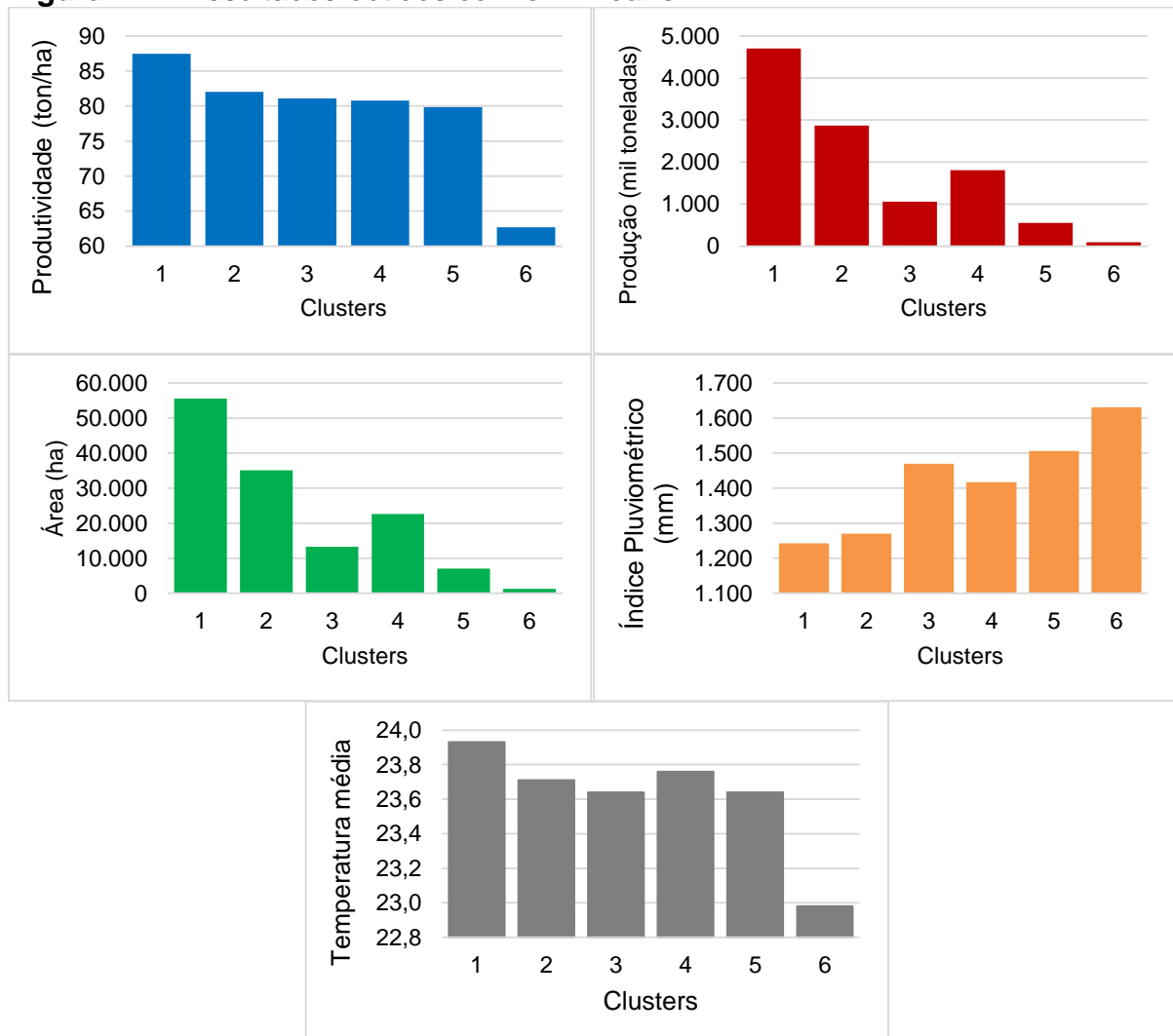
4 RESULTADOS E DISCUSSÕES

Os resultados serão descritos com uma seção destinada a cada algoritmo utilizado nesta pesquisa.

4.1 Resultados e análises obtidas por meio do *K-means*

Os resultados obtidos nos seis *clusters* do *K-means* estão demonstrados na Figura 12. Nesta Figura, cada coluna de cada gráfico refere-se a um *cluster* encontrado pelo algoritmo. O eixo vertical de cada gráfico mostra a média intragrupo de cada um dos atributos da base de dados (tais valores compõem o vetor médio de cada grupo no espaço geométrico formados pelos atributos, ou seja, os centroides dos *clusters*).

Figura 12 – Resultados obtidos com o *K-means*.



Fonte: Elaborada pela autora

Em uma discussão preliminar, observa-se que o *cluster* um, que foi o com maior produtividade, aglomerou também municípios com alta produção e área destinada a cana de açúcar e maior temperatura média, porém com menor índice pluviométrico registrado no ano. Conforme a produtividade diminui nos *clusters* dois, três, quatro e cinco, diminui-se também a produção, área e temperatura média, enquanto o índice pluviométrico aumenta. No *cluster* seis, que foi o com menor centroide de produtividade, estão concentrados municípios com produções e áreas baixíssimas, índice pluviométrico que ultrapassa 1600mm no ano e uma temperatura bem abaixo das registradas nos demais *clusters*.

Analisando as variáveis (atributos) isoladamente no gráfico de barras, percebe-se que o gráfico de área destinada ao plantio de cana de açúcar e o de produção da cultura possuem médias com tendências semelhantes para cada *clusters*. Naturalmente, em grandes áreas foram identificados grandes índices de produção, o que inicialmente demonstra que o algoritmo encontrou *clusters* que condizem com a realidade.

Considerando o gráfico da produtividade, o *cluster* um e dois, que são os com maiores índices de produtividade, são formados por municípios que também possuem grandes áreas e produções. Isso pode significar que a quantidade produzida pode influenciar na obtenção de grandes índices de produtividade.

A relação das variáveis produtividade e índices pluviométricos foi inversamente proporcional. Os *clusters* compostos por municípios com maiores índices de produtividades também são aqueles que obtiveram menor índice pluviométrico no ano de 2017.

A cana de açúcar necessita de quantidade de chuva razoavelmente alta, atingindo pelo menos 1200 mm no ano. Ao analisar os *clusters* formados em relação ao índice pluviométrico, pode-se observar que o *clusters*, apesar de obter o menor índice pluviométrico, alcançou em 2017 uma quantidade de chuva superior a 1200mm. A boa distribuição das chuvas durante o período de crescimento da cana de açúcar é importante, porém a cultura precisa também de uma época de seca para que ocorra a sua maturação. O clima no estado de São Paulo favorece esse cenário, pois em geral o estado tem um período de chuvas bem distribuído (de setembro ou outubro até março ou abril), que coincide com período quente (altas temperaturas) e de crescimento dos colmos da cana.

Essa colocação em relação a necessidade de período quente para crescimento dos colmos da cana vai ao encontro dos resultados obtidos nos *clusters* para a temperatura média. Pode-se observar que o *cluster* 1, com maior índice de produtividade, também obteve maior temperatura média no ano de 2017 e o *clusters* 6, com menor índice de produtividade, também foi aquele com menor temperatura média. Em relação à temperatura média anual a cana é uma planta tropical, portanto um clima mais quente favorece o crescimento da planta.

O tipo de solo predominante em cada um dos *clusters* é apresentado no Quadro 8 o qual mostra que este tipo de variável não apresentou grande relação com a produtividade. Segundo o especialista, os solos Latossolos e Argissolos apresentam menos problemas para cultivo da cana de açúcar, podendo apresentar uma fertilidade de média boa e geralmente relevos que podem ser trabalhados sem grandes problemas, bem como poucos problemas com erosão. No entanto, o estado de São Paulo possui a maioria de seus municípios com predominância desses dois tipos de solo, sendo um fator positivo para o cultivo da cultura em todo o estado, o que dificulta que qualquer outro tipo de solo possua predominância dentro de um *clusters*, por ser mais raro.

Quadro 8 – Tipo de solo predominante de cada *cluster* do *K-means*.

Cluster	Tipo de solo
1	Latossolos
2	Argissolos
3	Latossolos
4	Argissolos
5	Argissolos
6	Argissolos

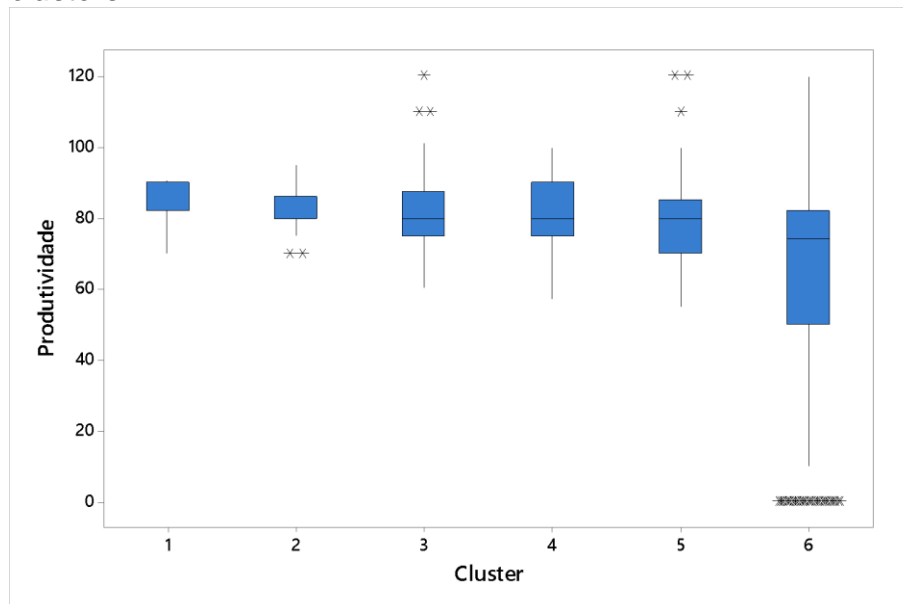
Fonte: Elaborado pela autora.

Com o objetivo de aprofundar a discussão a respeito dos resultados obtidos pelo *K-means*, foram realizadas análises estatísticas e gráficos *boxplot* identificando a mediana de cada *cluster*, a caixa de amplitude interquartilica e os traços de fundo e topo, bem como *outliers*.

Ao analisar a Figura 13, que representa a análise estatística dos resultados obtidos para os seis *clusters* em relação a produtividade, pode-se observar que o *cluster* 1 possui mediana maior que 80, com esse valor caindo de forma sutil nos

outros *clusters*. O *cluster 1* apresenta apenas um traço que se estende para baixo da caixa. Isso demonstra uma amplitude representada por uma observação dentro de *cluster* que apresentou produtividade igual a 70 ton/ha. No *cluster 1*, não houverem *outliers* que demonstrassem uma maior dispersão dos dados.

Figura 13 – *BoxPlot* da produtividade obtida pelo *K-means* para os seis *clusters*.



Fonte: Elaborada pela autora.

Observa-se que o *cluster 6* apresenta mediana abaixo de 80, com a maioria das observações também abaixo desse valor, traço grande que se estendem para baixo e uma quantidade grande de *outliers* para baixo. Isso demonstra que esse *cluster* é formado por vários elementos com produtividade baixa (entre 50 e 70 ton/ha) e por todos os elementos da base de dados com valor igual a 0. Com isso, pode-se concluir que o agrupamento dos dados foi eficiente em agrupar municípios com produtividade 0, a municípios com baixa produtividade.

Os municípios que apresentaram produtividade zero em 2017 foram Águas de Prata, Águas de Lindóia, Avaí, Boreri, Buri, Cabralia Paulista, Cajati, Divinolândia, Eldorado, Fernão, Hortolândia, Igaratá, Itaporanga, Jacupiranga, Joanópolis, Paraibuna, Parquera-açu, Paulínia, Paulistânia, Pedreira, Piedade, Pindamonhangaba, Presidente Alves, Queluz, Sarapuí, Taguaí, Timburi e Vargem.

Em análise estatística, também foi possível identificar o valor médio das observações pertencentes a cada *cluster* (centroides), demonstrado na Tabela 2.

Tabela 2 – Análise estatística da produtividade obtida pelo *K-means* para os seis *clusters*.

Cluster	Produtividade centroide	Desvio padrão
1	87,50a	6,807
2	82,02a	5,819
3	81,11a	9,855
4	80,79a	8,588
5	79,84a	11,154
6	62,69b	32,525

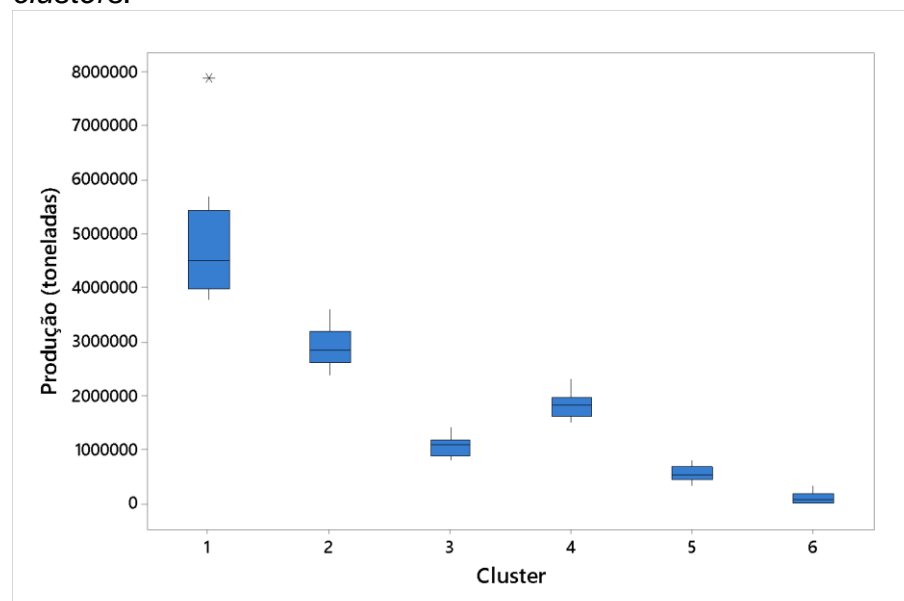
Fonte: Elaborada pela autora.

Pode-se observar que o a produtividade média dos dados do *cluster* 6 não compartilha da mesma letra que os outros *clusters*, isso significa que ele possui uma diferença significativa em relação aos outros.

O desvio padrão dos primeiros *clusters*, com maior média de produtividade, é baixo em relação ao desvio padrão dos *clusters* com menor média. A explicação para o desvio padrão alto do *cluster* 6 está na quantidade de *outliers* demonstrados no Gráfico 1, com elementos de produtividade igual a zero.

Na Figura 14, que representa a análise estatística dos resultados obtidos para os seis *clusters* em relação a produção, observa-se que a mediana de cada *clusters* é bastante distinta, variando desde uma mediana que se aproxima de 5 milhões de toneladas até uma mediana menor que 1 milhão de toneladas.

Figura 14 – *BoxPlot* da produção obtida pelo *K-means* para os seis *clusters*.



Fonte: Elaborado pela autora.

Observa-se também que os traços que se estendem de ambos os lados da caixa, representado a amplitude dos dados, são pequenos, o que em um primeiro momento permite concluir que os dados são concisos e com desvio padrão baixo. Apenas o *cluster* 1 apresentou um elemento *outlier*, que representa o município de Morro Agudo, que é o maior produtor de cana de açúcar do estado de São Paulo.

A Tabela 3, demonstra que no desvio padrão, com exceção do *cluster* 1 que possui um *outlier*, confirma-se a concisão dos dados pertencentes a cada *clusters*, demonstrando eficácia do algoritmo no agrupamento dos dados.

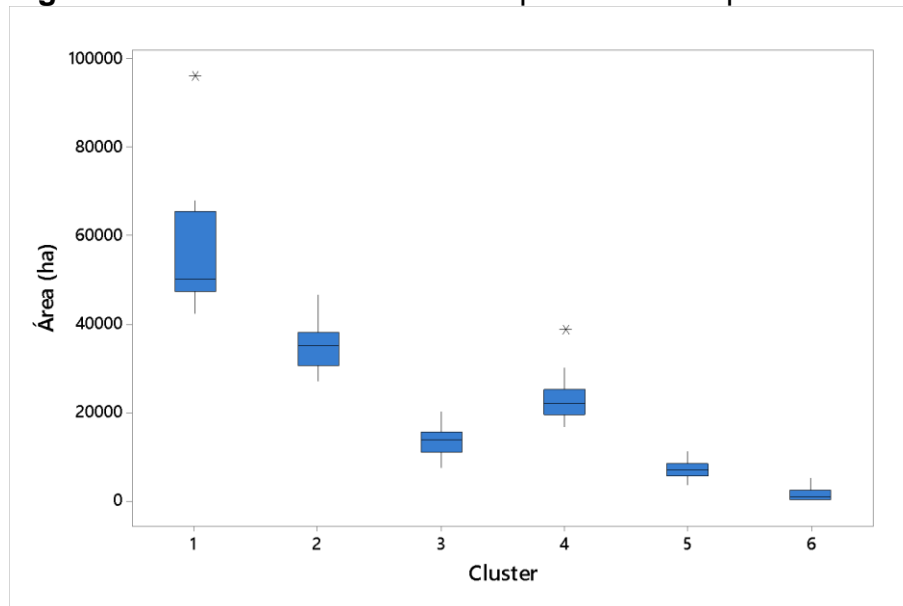
Tabela 3 – Análise estatística da produção obtida pelo *K-means* para os seis clusters.

Cluster	Produção centroide	Desvio padrão
1	4.692.629a	1.279.106
2	2.883.990b	369.012
3	1.063.983d	180.807
4	1.815.680c	234.045
5	556.689e	131.411
6	97.320f	103.270

Fonte: Elaborada pela autora.

Ressalta-se também que o agrupamento do método *K-means* para 6 *clusters*, causada por diferenças no conjunto das variáveis analisadas, teve grande influência da variável produção, uma vez que a análise estatística revelou diferença significativa nas médias de todos *clusters*, pois a produção média dos dados de cada *cluster* que não compartilham uma mesma letra são significativamente diferentes.

Ao analisar a Figura 15, que representa a análise estatística dos resultados obtidos para os seis *clusters* em relação a área, observa-se que, assim como na produção, os dados os *clusters* possuem medianas bem distintas, com baixa amplitude dos dados.

Figura 15 – *BoxPlot* da área obtida pelo *K-means* para os seis *clusters*.

Fonte: Elaborada pela autora.

O *cluster* 1, possui Morro Agudo como um de seus elementos, o município com maior produção e área do estado de São Paulo, que está sendo representado pelo *outlier* no mapa e como consequência eleva o valor do desvio padrão do *cluster*. No *BoxPlot* da área foi identificado também um *outlier* no *cluster* 4, que é o município de Teodoro Sampaio, que possui área destinada a produção de cana de açúcar de aproximadamente 38 mil hectares, enquanto a média do *cluster* no qual ele está inserido é em torno de 22 mil hectares, o que eleva o desvio padrão do *cluster* 4, demonstrado na Tabela 4.

Tabela 4 – Análise estatística da área obtida pelo *K-means* para os seis *clusters*.

<i>Cluster</i>	Área centroide	Desvio padrão
1	55.714a	16.626
2	35.321b	5.131
3	13.305d	2.706
4	22.765c	4.168
5	7.071e	1.793
6	1.284f	1.387

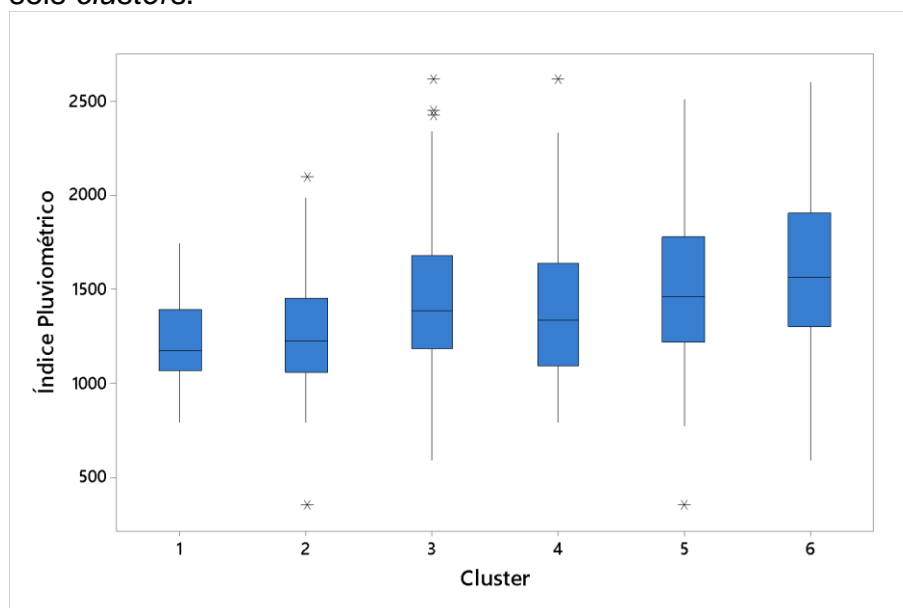
Fonte: Elaborada pela autora.

Novamente, o agrupamento do método *K-means* é evidenciada agora pela variável área, em que a média dos dados mostra-se também com diferenças

significativas de todos *clusters* entre si, tendo assim, grande influência sobre o agrupamento.

Na Figura 16, que representa a análise estatística dos resultados obtidos para os seis *clusters* em relação ao índice pluviométrico, observa-se que o primeiro *clusters* apresentou a menor mediana dos seus elementos, porém os dados são mais concisos e os traços que representam amplitude são menor, quanto o *cluster* seis é aquela com maior mediana, porém também com maior amplitude de dados.

Figura 16 – *BoxPlot* do índice pluviométrico obtido pelo *K-means* para os seis *clusters*.



Fonte: Elaborada pela autora.

No *cluster* 1, a maior amplitude representada pelo traço superior deu-se devido ao município de Guaíra, que foi o único do grupo com índice superior a 1500mm, com 1745 mm de chuva em 2017.

No *cluster* 6, vários municípios foram responsáveis pelo elevado índice pluviométrico, entre eles estão Vitória Brasil, Três Fronteiras, São Francisco, Santa Selete, Santa Fé do Sul, Santa Clara d'Oeste, Rubinéia e Nova Canaã Paulista, que apresenta índice superior a 2500 mm. Os municípios pertencentes ao *cluster* 6, que foram responsáveis pela grande amplitude do traço inferior foram Santo Antônio da Alegria, Cássia dos Coqueiros e Americana que não atingiram 1000 mm. Essa grande dispersão dos dados dentro de seus *clusters* é representada também pelo desvio padrão da Tabela 5.

Tabela 5 – Análise estatística do índice pluviométrico obtido pelo *K-means* para os seis *clusters*

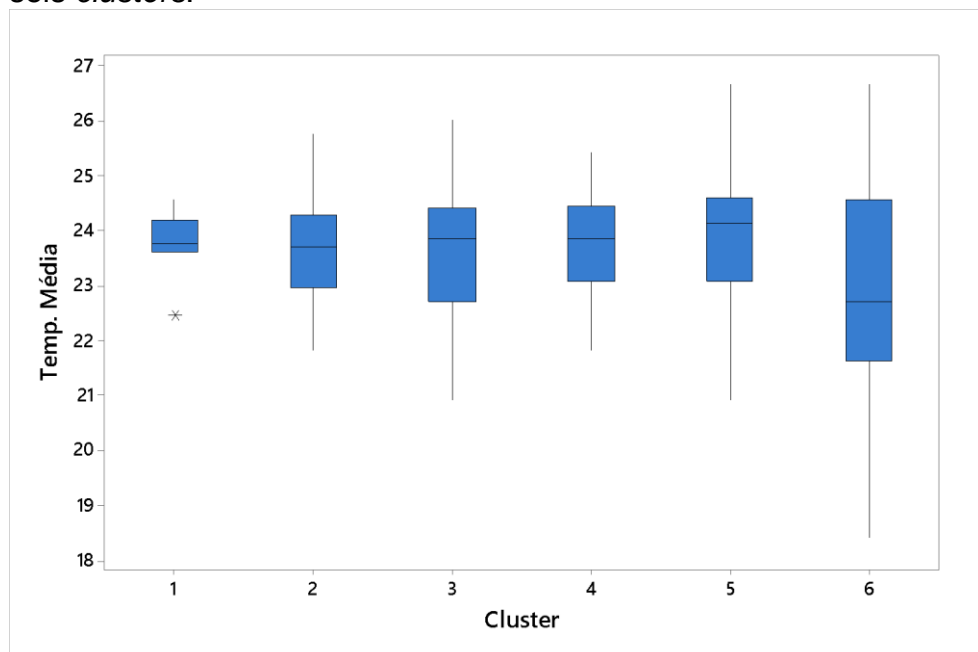
<i>Cluster</i>	Índice pluviométrico centroide	Desvio padrão
1	1239bc	277,5
2	1267c	338,7
3	1470bc	401,5
4	1417bc	398,3
5	1506ab	392,8
6	1631a	421,5

Fonte: Elaborada pela autora.

Observa-se que o índice pluviométrico do *cluster* 6 compartilha apenas de uma letra com o *cluster* 5, demonstrando uma diferença significativa em relação aos outros *clusters*. O *cluster* 5, possui algum grau de semelhança com o *cluster* 6, ao mesmo tempo que compartilha semelhança com os *clusters* 1, 3 e 4, diferenciando-se significativamente do *cluster* 2. Os *clusters* que apresentam maior semelhança estatística entre si, são o 1, 3 e 4 que compartilham exatamente das mesmas letras.

A análise estatística para a temperatura média está representada na Figura 17.

Figura 17 – *BoxPlot* da temperatura média obtida pelo *K-means* para os seis *clusters*.



Fonte: Elaborada pela autora.

Nota-se que, apesar de crescente, a mediana dos quatro primeiros *clusters* são bem parecidas, variando de menos que meio grau, enquanto que a mediana do *cluster* 6 está bem abaixo das restantes. Levando em consideração que a produtividade deste último *cluster* é bem inferior da dos restantes, pode-se concluir preliminarmente que temperatura média muito baixo pode afetar a produtividade da cana de açúcar.

No *clusters* 6, encontram-se municípios como Roseira, Pedra Bela e Lagoinha que apresentaram temperatura média menor que 20° e Socorro, Serra Negra, Monte Alegre do Sul e Lindóia com temperatura média menor que 20,5°.

Na Tabela 6, percebe-se que o desvio padrão dos *cluster* 3, 4 e 5 são relativamente altos, com seus elementos variando mais que um grau, enquanto que o *cluster* 1 tem pequena amplitude dos dados, exceto por um *outlier* que é o município de Piracicaba, que foi o único do *cluster* que apresentou temperatura média inferior a 23°.

Tabela 6 – Análise estatística da temperatura média obtida pelo *K-means* para os seis *clusters*.

Cluster	Temperatura centroide	Desvio padrão
1	23,87ab	0,601
2	23,71a	0,925
3	23,64a	1,220
4	23,77a	0,952
5	23,86a	1,192
6	22,98b	1,653

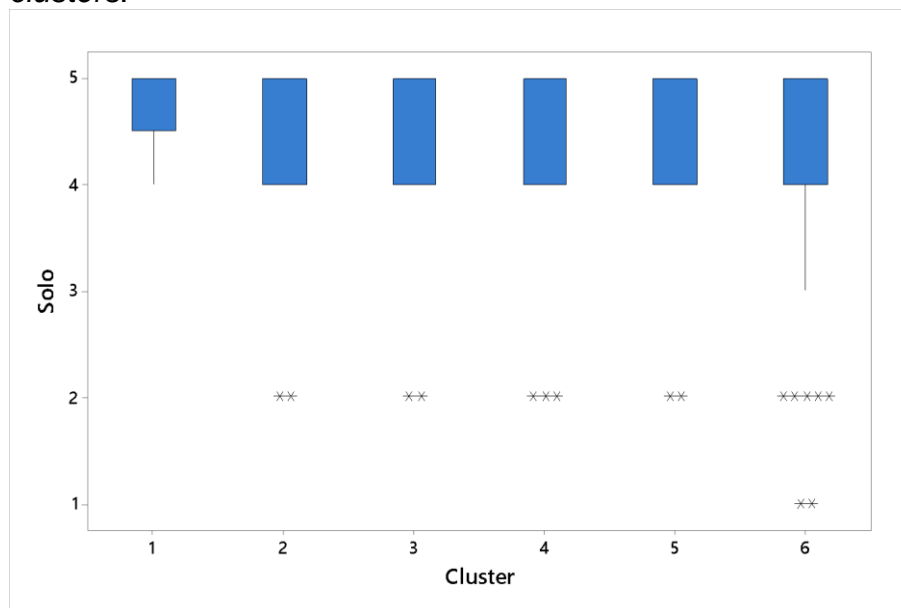
Fonte: Elaborada pela autora.

Ressalta-se que a temperatura média dos dados dos *clusters* 2, 3, 4 e 5 apresentam a mesma letra, não demonstrando diferença significativa, diferenciando-se de maneira mais significativa dos *clusters* 1 e 6. O *cluster* 6 apresenta um grau de semelhança com o 1, porém ele é o que mais apresenta diferença significativa por não ser representado pela letra a.

Ao analisar a Figura 18, que representa a análise estatística dos resultados obtidos para os seis *clusters* em relação ao tipo de solo, observa-se que, apesar do algoritmo *K-means* ter identificado apenas dois tipos de solo como predominantes (Latosolos e Argissolos), alguns *clusters* são compostos por elementos que

apresentam outros tipos de solos menos indicados para a produção de cana de açúcar. O *cluster* 1, com maior índice de produtividade, é composto em sua grande maioria pelo Latossolos, que foi numerado como 5, com algumas ocorrências do solo 4 (Argissolos), lembrando que, segundo o especialista entrevistado, os solos Latossolos e Argissolos são os mais indicados para o plantio da cana de açúcar e solos como Neossolos e Espodossolos são menos indicados.

Figura 18 – *BoxPlot* do tipo de solo obtido pelo *K-means* para os seis *clusters*.



Fonte: Elaborada pela autora.

O *cluster* 2, 3, 4 e 5, apesar de também serem compostos basicamente pelos solos 4 e 5, possuem alguns *outliers* do solo numerado como 2, que é o Neossolos. O *cluster* 6 é o com maior número de *outliers* e o único que, além de ter elementos com solo Neossolos, tem Espodossolos.

Apesar do algoritmo *K-means* não ter possibilitado uma análise mais aprofundada da relação entre o tipo de solo e a produtividade obtida pelos *clusters*, a análise estatística realizada por meio do gráfico *BoxPlot* permitiu concluir que o tipo de solo possui influência na produtividade.

Na Tabela 7, pode-se verificar o baixo desvio padrão do *cluster* um, por ser formado apenas por elementos com os tipos de solo 1 e 2, enquanto que os outros *clusters* apresentam valores maiores para o desvio padrão.

Tabela 7 – Análise estatística do tipo de solo obtido pelo *K-means* para os seis *clusters*.

Cluster	Solo centroide	Desvio padrão
1	4,7a	0,4410
2	4,4a	0,7564
3	4,5a	0,6232
4	4,4a	0,7663
5	4,3a	0,5559
6	4,1b	0,7294

Fonte: Elaborada pela autora.

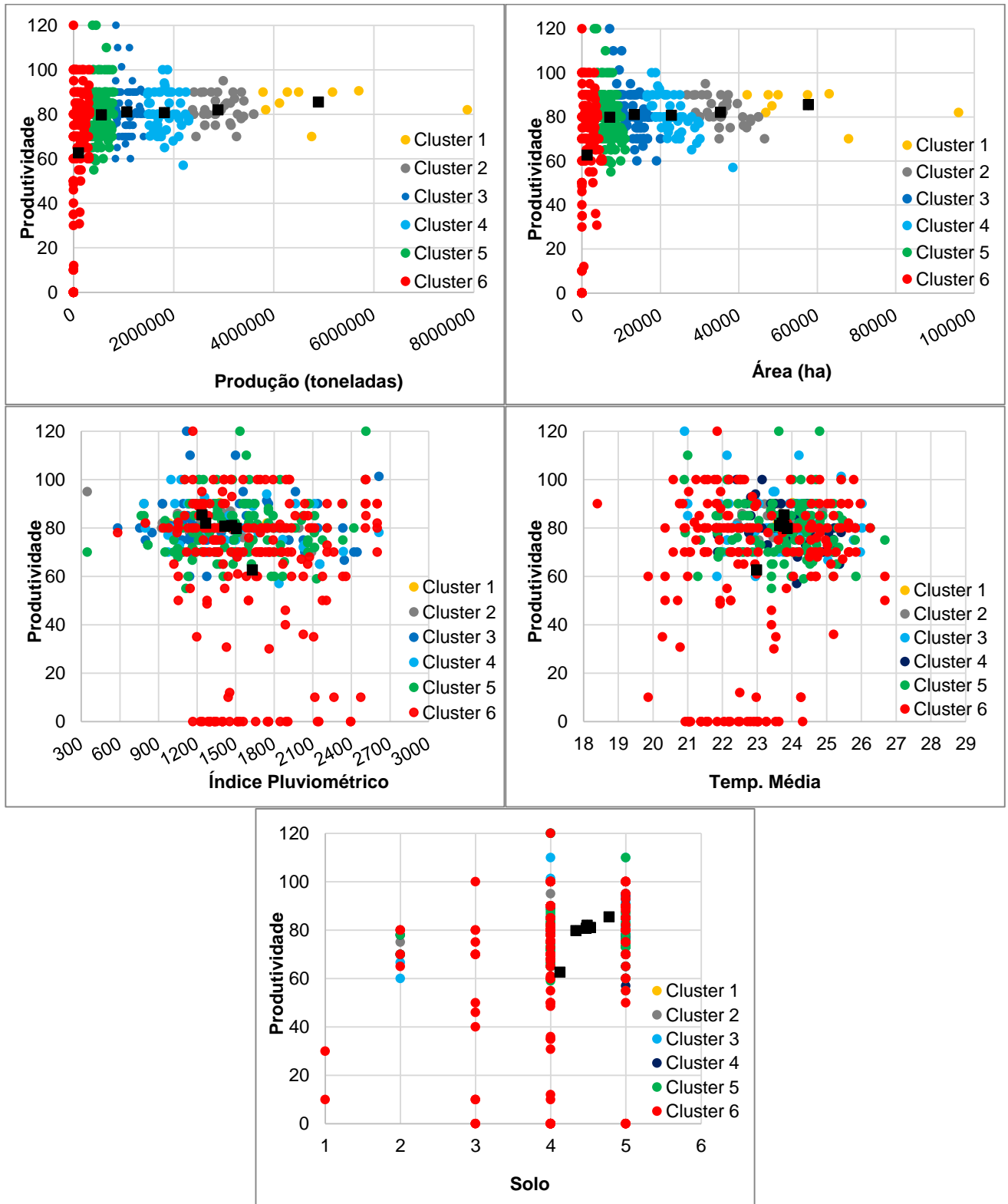
Observa-se que o *cluster* seis apresenta uma diferença estatisticamente diferente dos outros *clusters* por não compartilhar uma mesma letra, sendo assim, entende-se que o tipo de solo influenciou no agrupamento de municípios com baixa produtividade.

Com intuito de avaliar a relação de cada atributo discutido neste trabalho com a produtividade, foram elaborados gráficos de dispersão, demonstrados na Figura 19.

Nota-se que, quanto menor a produção, menor o controle sobre a variável produtividade, ou seja, a produtividade em pequenas produções varia muito, desde zero até 120 toneladas por hectare. É preciso considerar que o município que ficou inserido no *cluster* 4 com produtividade igual a 120 é Itupeva. Itupeva é um município localizado na região metropolitana de Jundáí, que é muito desenvolvida tecnologicamente, o que pode ter influência positiva para a questão da produtividade em pequenas áreas.

Por outro lado, percebe-se que a produtividade em grandes áreas normalmente mante-se acima de 80 ton/ha, com exceção de Barretos que em uma área de 68 mil hectares produziu 4,7 milhões de hectares, com produtividade de 70 ton/hectare. Assim, pode-se concluir que é mais fácil controlar a produtividade e atingir níveis mais altos em grandes áreas destinada a produção de cana de açúcar.

Figura 19 – Gráficos de dispersão relacionando os atributos à produtividade, obtidos com o *K-means*.



Fonte: Elaborada pela autora

No gráfico que relaciona a área e produtividade, assim como na produção, o *cluster* 1 é formado por elementos com desvio padrão baixo em relação a

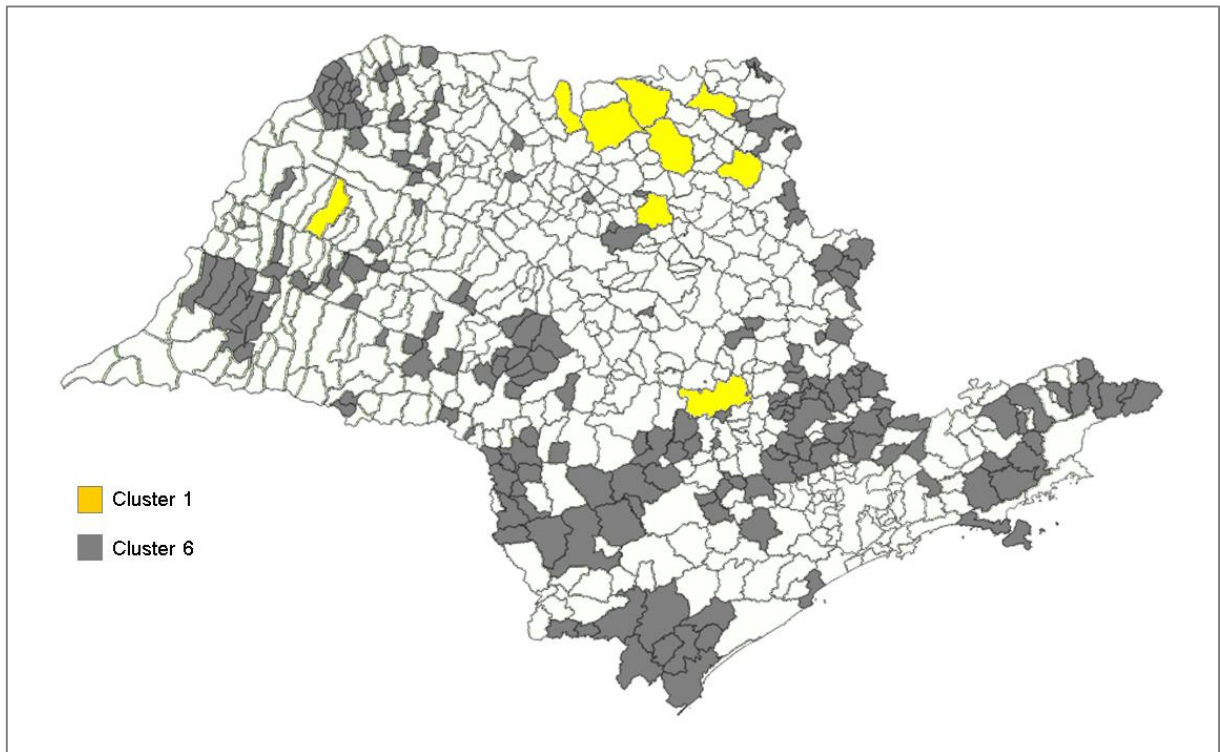
produtividade, ou seja, os elementos pertencentes ao grupo 1, com maior área destinada a produção de cana de açúcar, todos possuem produtividade relativamente alta. Em contrapartida, a produtividade do *cluster* 1, com pequenas áreas destinadas a produção de cana de açúcar varia em níveis elevados, desde 0 até 120 toneladas por hectare, necessitando assim de um estudo aprofundado, focando nesses municípios, com o intuito de identificar outras variáveis que influenciam na produtividade.

Ao relacionar em um gráfico de dispersão o índice pluviométrico e a produtividade, torna-se difícil identificar uma possível influência da temperatura média sobre a produtividade, assim como demonstrado na análise estatística feita anteriormente.

Na relação entre o tipo de solo e a produtividade acontece a mesma situação, porém, percebe-se que as duas ocorrências de solo do tipo um, inseridas no *cluster* seis, possuem produtividade abaixo de 40 toneladas por hectare. Esses municípios de solo tipo um, Espodossolos, são Cananéia e Peruíbe. Os dois municípios pertencem ao litoral Sul do estado de São Paulo.

Para encerrar as discussões envolvendo os resultados do agrupamento obtidos pelo algoritmo *K-means* o mapa demonstrado na Figura 20 representa os municípios que foram agrupados no *cluster* 1, ou seja, o *cluster* com maior índice de produtividade.

Figura 20 – Representação dos municípios pertencentes aos *clusters* um e seis do *K-means*.



Fonte: Elaborada pela autora

Percebe-se que a os municípios pertencentes ao *cluster* um estão concentrados no Nordeste, Oeste e no Centro paulista. Os municípios pertencentes ao *cluster* 1, com maior média de produtividade são: Barretos, Batatais, Guaiúra, Guaraci, Ituverava, Jaboticabal, Morro Agudo, Piracicaba e Valpaíso.

A expansão da cana de açúcar no município de Barretos transformou seu espaço geográfico e sua economia, trocando outras atividades agrícolas, como a laranja que era a principal cultura. Em 2017 Barretos destinou uma área de 68 mil hectares para produção de cana de açúcar (IEA, 2019). O IDHM de Barretos evoluiu nos últimos vinte anos, mudando a posição no ranking estadual e nacional, subindo na classificação, demonstrando avanço de desenvolvimento, influenciado pela cultura da cana de-açúcar e a produção de seus derivados para. (ÁVILA, 2014).

Guaraci, devido a decadência da citricultura, expandiu gradativamente a produção de cana de açúcar no município após os anos 2000, atingindo uma área maior que 42 mil hectares em 2017 (IEA, 2019), que foi quando se instalou no município uma usina de cana de açúcar tornando-se a principal atividade econômica e acarretando no crescimento da migração de trabalhadores. A Usina Vertente,

localizada em Guaraci, é hoje a maior fonte de arrecadação do município e maior geradora de empregos.

O cultivo e industrialização de cana de açúcar é a principal atividade econômica também de Valparaíso, que possui uma área de aproximadamente 47 mil hectares destinados a cana de açúcar, atingindo em 2017 uma produção de 3,8 milhões de toneladas (IEA, 2019).

Morro Agudo, que é o maior produtor de cana de açúcar do Brasil, com uma área de aproximadamente 96 mil hectares destinados a produção de cana de açúcar, alcançando em 2017 uma produção de aproximadamente 8 milhões de toneladas (IEA, 2019). O município sempre teve como base econômica a agricultura, porém, as culturas eram soja, milho, sorgo e grãos. Na década de 1980 produtores rurais começaram a enfrentar grandes dificuldades em manter suas terras e em encontrar mercado para suas lavouras de soja, e foi quando a mesorregião de Ribeirão Preto acabou se tornando de grande importância às atividades sucroalcooleiras. Neste período, Morro Agudo expandiu o cultivo de cana de açúcar, que tornou-se a atividade agrícola de maior expressividade principal fonte empregatícia do município. Morro Agudo possui cerca de 25% dos empregos relacionados ao agronegócio, enquanto a média do estado de São Paulo é de 3,3% (COSTA, CLEPS; 2014).

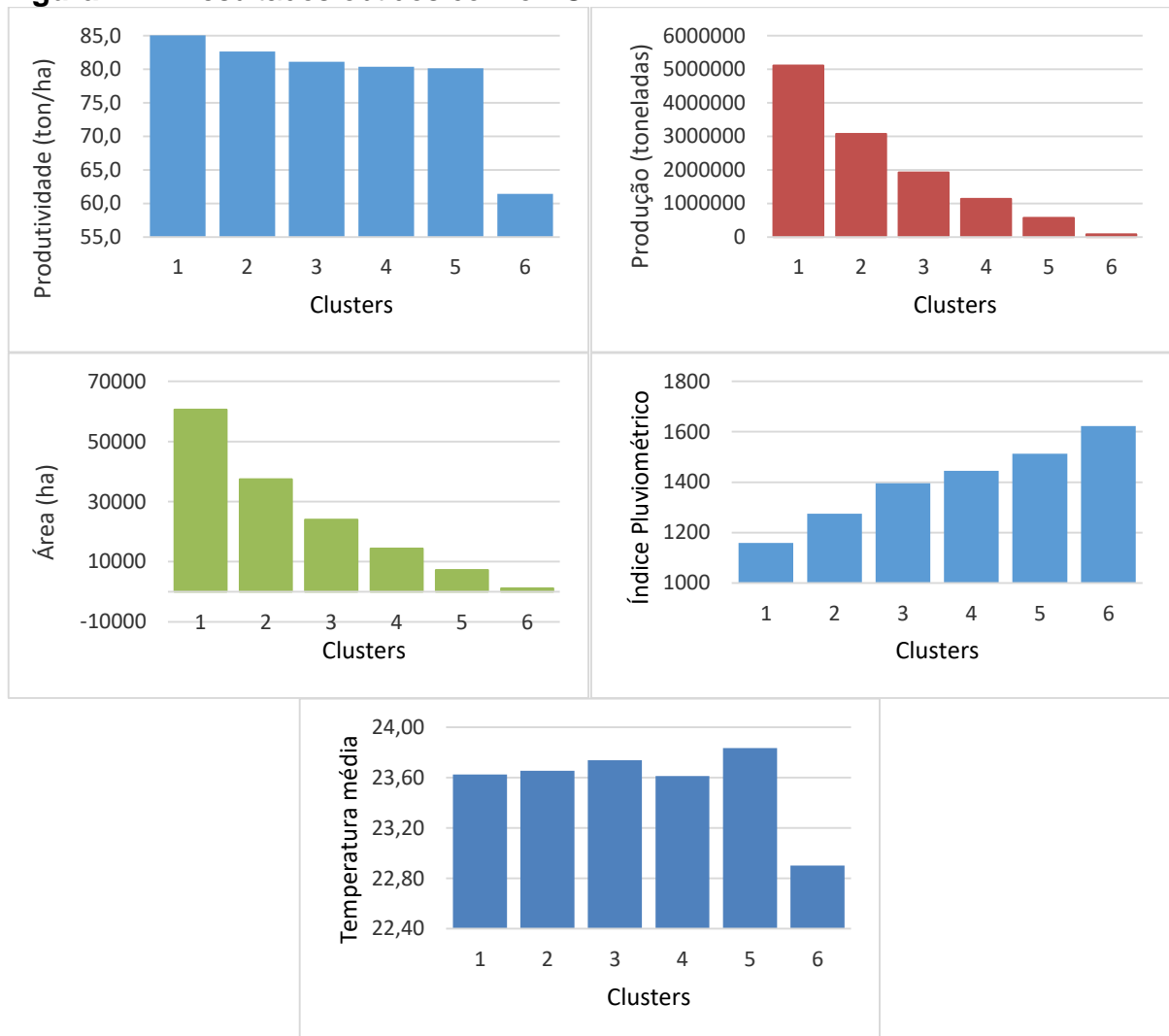
Piracicaba desenvolveu a agricultura do município no decorrer do século XIX com a produção de café e cana de açúcar. No século XX veio o fim do ciclo do café com a estagnação da economia do município e o aumento da industrialização, tornando-se um dos primeiros municípios do país a se industrializar e inaugurar plantas fabris. Mesmo com a mudança da agricultura para o setor de industrialização, Piracicaba ainda possui uma área de aproximadamente 50 mil hectares destinados a cana de açúcar, com uma produtividade média de 90 toneladas por hectare, sendo uma das maiores produtividades do estado. (IEA, 2019).

4.3 Resultados e análises obtidas por meio do FCM

Os resultados obtidos nos seis *clusters* do *Fuzzy C-means* estão demonstrados na Figura 21. Nesta Figura, cada coluna de cada gráfico refere-se a um *cluster* encontrado pelo algoritmo. O eixo vertical de cada gráfico mostra a média intragrupo de cada um dos atributos da base de dados (tais valores compõem o vetor

médio de cada grupo no espaço geométrico formados pelos atributos, ou seja, os centroides dos *clusters*).

Figura 21 – Resultados obtidos com o FCM.



Fonte: Elaborada pela autora.

Os resultados obtidos pelo FCM possuem grande similaridade com os resultados do *K-means*. Percebe-se que no gráfico da produtividade, os *cluster* com maior índice, também são aqueles com maior área e produção, assim como nos resultados obtido pelo algoritmo *K-means*.

Os *clusters* compostos por municípios com maiores índices de produtividades também são aqueles que obtiveram menor índice pluviométrico no ano de 2017, confirmando a relação inversamente proporcional dos dados.

Para a temperatura média, os *clusters* identificados não possuem rótulo que demonstrem uma tendência de forma clara. Porém, observa-se que o *clusters*

seis, que é o mesmo que obteve menor índice de produtividade, apresentou uma temperatura média bem abaixo dos demais, o que vai ao encontro do resultado obtido pelo *K-means*.

O resultado obtido pela *clusterização* para o tipo de solo predominante em cada um dos *clusters*, apresentado no Quadro 9, demonstra que os *clusters* 1, 2, 3 e 4 possuem predominância de solo Latossolos, enquanto que o 5 e 6, com menor produtividade, são predominados pelo Latossolos.

Quadro 9 – Tipo de solo predominante de cada *cluster*.

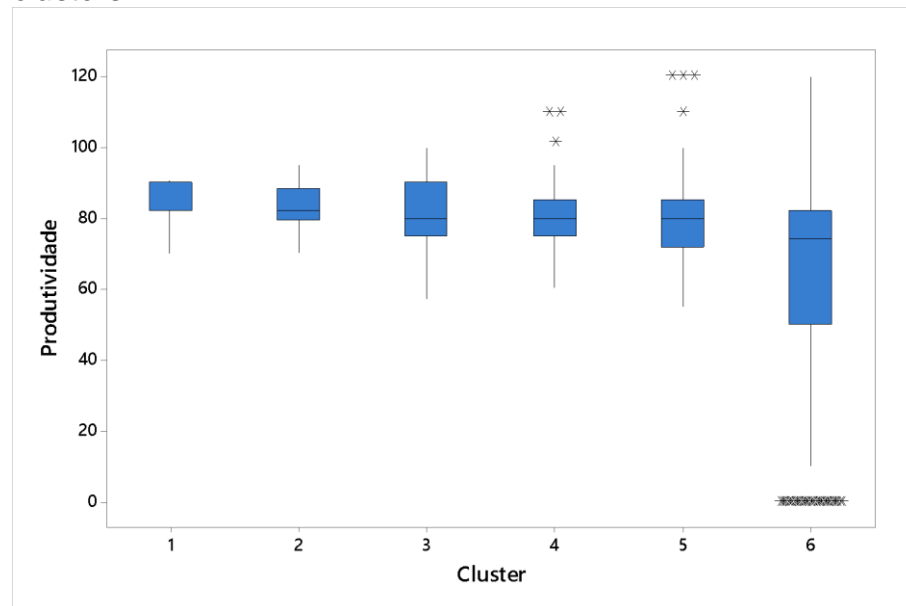
Cluster	Tipo de solo
1	Latossolos
2	Latossolos
3	Latossolos
4	Latossolos
5	Argissolos
6	Argissolos

Fonte: Elaborado pela autora.

Aprofundando a discussão a respeito dos resultados obtidos pelo *Fuzzy C-means*, foram realizadas análises estatísticas e elaborado gráficos *BoxPlot* identificando a mediana de cada *cluster*, a caixa de amplitude interquartílica e os traços que se estendem de ambos os lados da caixa.

Ao analisar a Figura 22, que representa a análise estatística dos resultados obtidos para os seis *clusters* em relação a produtividade, observa-se que uma mediana maior para o *cluster* 1, com uma leve tendência de queda até o *cluster* 6. O *cluster* 1 apresenta apenas um traço que se estende para baixo da caixa, demonstrando uma pequena amplitude dos dados para baixo. Assim como no *K-means*, o *cluster* 1 não apresentou *outliers* que demonstrassem uma maior dispersão dos dados. Nota-se de forma clara que quanto menor a produtividade média do *cluster*, maior a amplitude dos dados.

Figura 22 – *BoxPlot* da produtividade obtida pelo FCM para os seis *clusters*.



Fonte: Elaborada pela autora.

O *cluster* 6 apresenta mediana abaixo de 80, com a maioria das observações também abaixo desse valor, traço grande que se estendem para baixo e uma quantidade grande de *outliers* para baixo. Isso demonstra que esse *cluster* é formado por vários elementos com produtividade baixa, assim como no *K-means*, e por todos os elementos da base de dados com valor igual a 0. Com isso, pode-se concluir que o agrupamento dos dados no FCM também foi eficiente em agrupar municípios com produtividade 0, a municípios com baixa produtividade.

A análise estatística com apresentação do valor médio das observações pertencentes a cada *cluster* e os desvios padrões estão demonstrados na Tabela 8,

Tabela 8 – Análise estatística da produtividade obtida pelo FCM para os seis *clusters*.

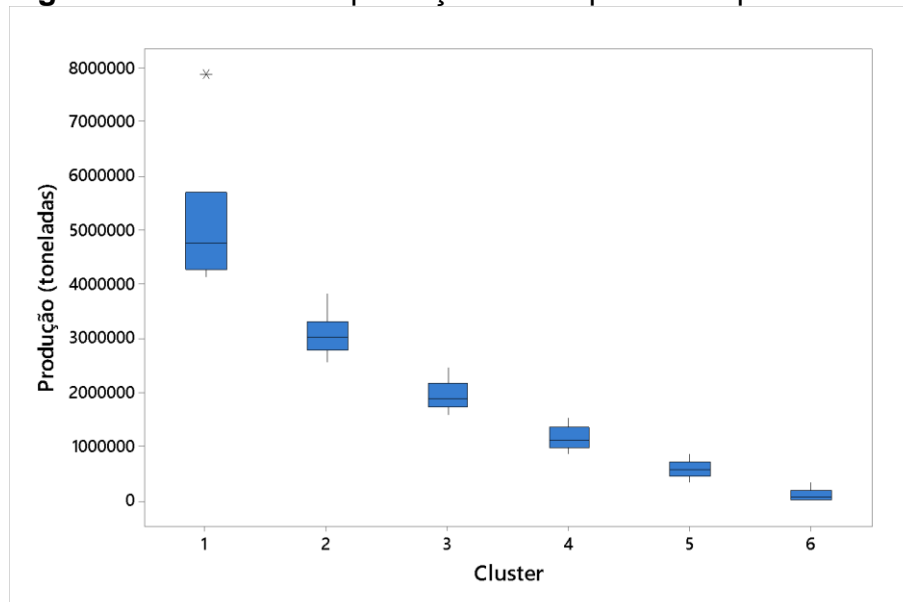
<i>Cluster</i>	Produtividade centroide	Desvio padrão
1	85,35ab	7,506
2	82,70a	5,881
3	80,58a	8,605
4	80,67a	8,812
5	80,22a	11,325
6	62,69b	32,525

Fonte: Elaborada pela autora.

O desvio padrão dos *clusters* ressalta a maior amplitude dos dados, já demonstrada na Figura 22, para os *clusters* com menor média produtiva. O *clusters* 6 apresentou um desvio padrão de 32,525, o que representa quatro vezes mais do que o desvio do *cluster* 1.

A Figura 23, que representa a análise estatística dos resultados obtidos para os seis *clusters* em relação a produção, observa-se que a mediana de cada *clusters* é bastante distinta, variando desde uma mediana que se aproxima de 5 milhões de toneladas até uma mediana menor que 1 milhão de toneladas.

Figura 23 – *BoxPlot* da produção obtida pelo FCM para os seis *clusters*.



Fonte: Elaborada pela autora.

No FCM, o comportamento do agrupamento em relação a produção comportou-se de maneira bem similar ao *K-means* analisado anteriormente, com pequenos traços que se estendem de ambos os lados da caixa demonstrando o quanto os dados são concisos e apenas o *cluster* 1 apresentou um elemento *outlier*, que representa o município de Morro Agudo, que é o maior produtor de cana de açúcar do estado de São Paulo. Nos dois algoritmos, Morro Agudo ficou inserido no *cluster* 1, que é o representado por maior média de produtividade.

Os baixos desvios padrões apresentados na Tabela 9, com exceção do *cluster* 1 que possui um *outlier*, confirma-se a concisão dos dados pertencentes a cada *clusters*, demonstrando eficácia do algoritmo no agrupamento dos dados

Tabela 9 – Análise estatística da produção obtida pelo FCM para os seis *clusters*.

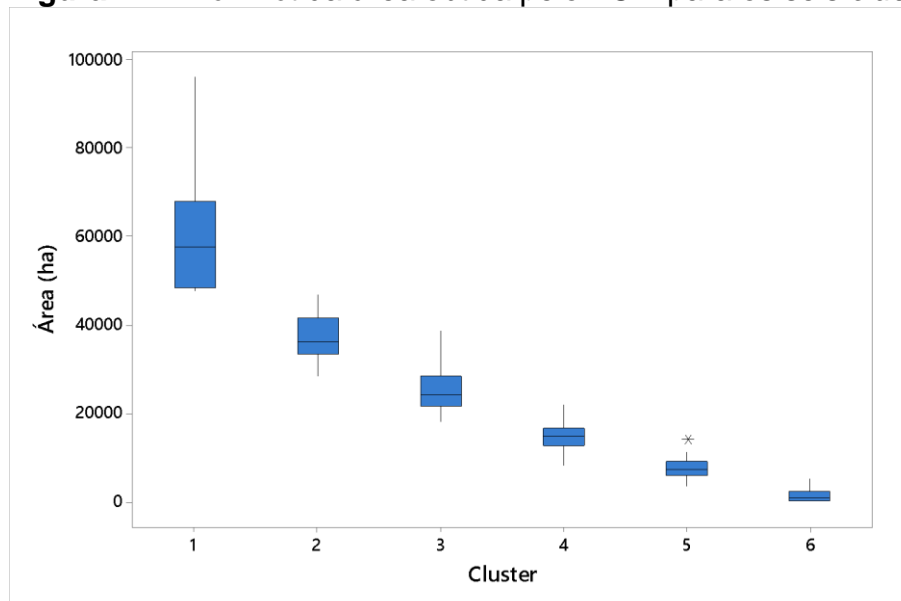
<i>Cluster</i>	Produção centroide	Desvio padrão
1	5.200.424a	1.297.644
2	3.057.309b	356.034
3	1.962.045c	264.293
4	1.161.953d	197.891
5	584.579e	150.124
6	97.320f	103.270

Fonte: Elaborada pela autora.

O agrupamento do método FCM para seis *clusters*, causada por diferenças no conjunto das variáveis analisadas, teve grande influência da variável produção, uma vez que a análise estatística revelou diferença significativa nas médias de todos *clusters*, pois a produção média dos dados de cada *cluster* que não compartilham uma mesma letra são significativamente diferentes.

Ao analisar a Figura 24, que representa a análise estatística dos resultados obtidos para os seis *clusters* em relação a área, observa-se que, assim como na produção, os dados os *clusters* possuem medianas bem distintas, com baixa amplitude dos dados.

Figura 24 – *BoxPlot* da área obtida pelo FCM para os seis *clusters*.



Fonte: Elaborada pela autora.

A Figura 24 ressalta a importância da variável área no agrupamento dos dados no FCM. Percebe-se que os grupos são formados por áreas bem distintas umas das outras, com pequena amplitude dos dados, confirmando a diferença entre os grupos e a eficácia do método no agrupamento. Confirmando essa baixa amplitude dos dados dentro de cada *cluster*, a Tabela 10 demonstra que, com exceção do *cluster* 1, os grupos, possuem dados internos com baixo desvio padrão.

Tabela 10 – Análise estatística da área obtida pelo FCM para os seis *clusters*.

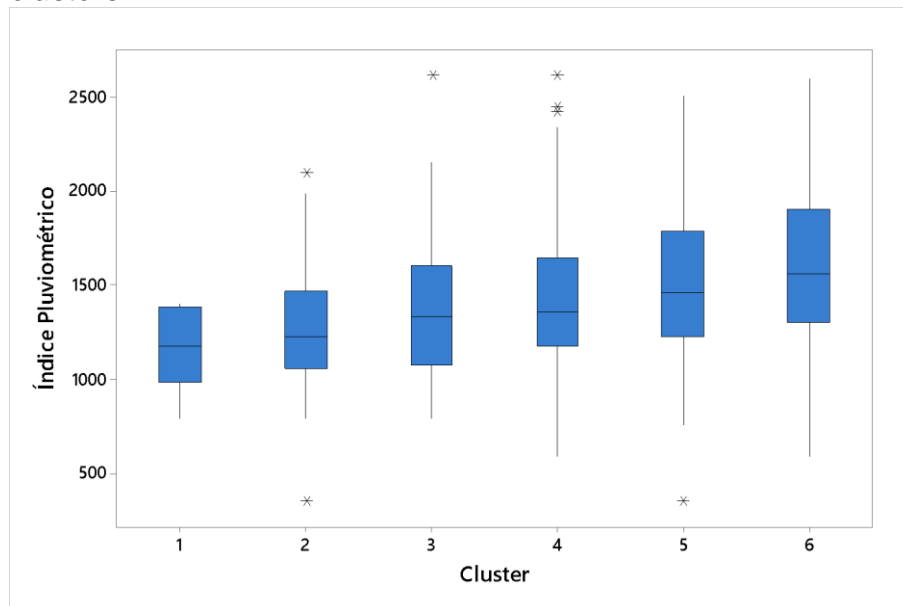
Cluster	Área centroide	Desvio padrão
1	61.497a	17.076
2	37.150b	5.036
3	24.645c	4.430
4	14.553d	2.779
5	7.382e	1.988
6	1.284f	1.387

Fonte: Elaborada pela autora.

O mesmo não acontece quando a área média de cada grupo é comparada. Nota-se que o agrupamento do método FCM para seis *clusters* teve grande influência da variável produção, uma vez que a análise estatística revelou diferença significativa nas médias de todos *clusters*, pois a área média dos dados de cada *cluster* que não compartilham uma mesma letra são significativamente diferentes.

A Figura 25, que representa a análise estatística dos resultados obtidos para os seis *clusters* em relação ao índice pluviométrico no FCM comportou-se de maneira muito parecida com o *K-means*, em que o primeiro *clusters* apresentou a menor mediana dos seus elementos, porém os dados são mais concisos e os traços que representam amplitude são menores, quanto o *cluster* seis é aquela com maior mediana, porém também com maior amplitude de dados.

Figura 25 – *BoxPlot* do índice pluviométrico obtido pelo FCM para os seis *clusters*.



Fonte: Elaborada pela autora.

A grande dispersão dos dados demonstrada pelos traços do gráfico *BoxPlot* são expressados numericamente na Tabela 11.

Tabela 11 – Análise estatística do índice pluviométrico obtido pelo FCM para os seis *clusters*.

<i>Cluster</i>	Índice pluviométrico centroide	Desvio padrão
1	1.180a	233,2
2	1.287b	365,8
3	1.392bc	388,3
4	1.453bc	403,9
5	1.507c	389,5
6	1.631bc	421,5

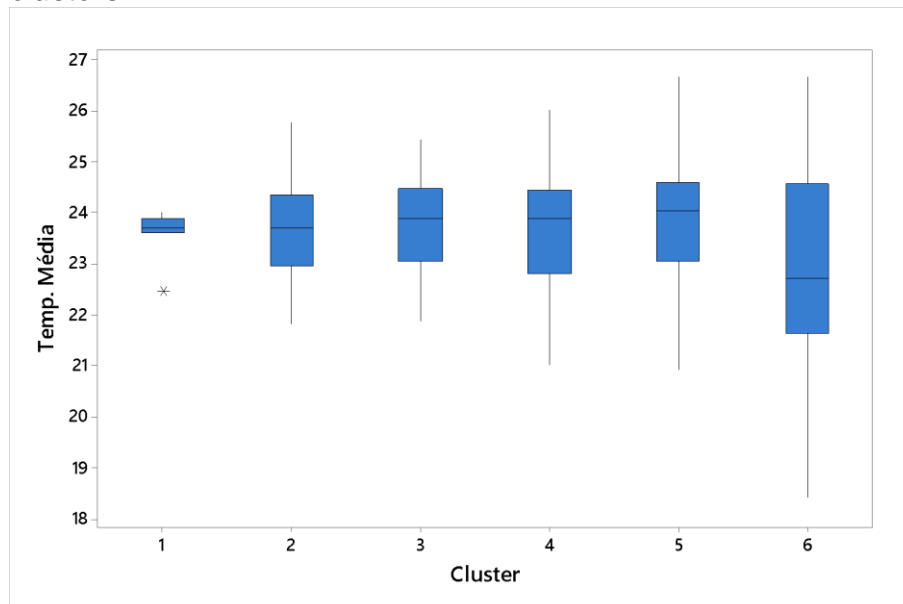
Fonte: Elaborada pela autora.

O índice pluviométrico do *cluster* 1 foi o único que não compartilhou da mesma letra com nenhum outro *cluster*, demonstrando uma diferença estatística significativa. Isso significa que o baixo índice pluviométrico dos municípios no ano de 2017 influenciou no agrupamento dos dados do *cluster* 1. Em entrevista, o especialista relatou a importância de volumes de chuva maiores que 2000mm no ano para o bom desenvolvimento da cultura, porém, no caso dos municípios inseridos no *cluster* 1, o índice pluviométrico médio ficou em 1180mm no ano, muito abaixo do esperado. Isso

pode significar que, muito melhor do que o volume de chuva, é importante que essa chuva ocorra no período certo, que é o de crescimento da planta.

A análise estatística para a temperatura média está representada na Figura 26.

Figura 26 – *BoxPlot* da temperatura média obtida pelo FCM para os seis *clusters*.



Fonte: Elaborada pela autora.

Nota-se que, assim como ocorreu no agrupamento dos dados no *K-means*, apesar de crescente, a mediana dos cinco primeiros *clusters* são bem parecidas, variando de menos que meio grau, enquanto que a mediana do *cluster* 6 está bem abaixo das restantes. Confirma-se que a temperatura muito baixa influencia negativamente na produtividade da cana de açúcar.

Na Tabela 12, estão representados os desvios padrões percebe-se que os desvios padrões dos *clusters*. Quanto maior a temperatura média do *cluster*, menor seu desvio padrão. O *cluster* seis que apresentou temperatura média baixa, foi também o com maior desvio padrão.

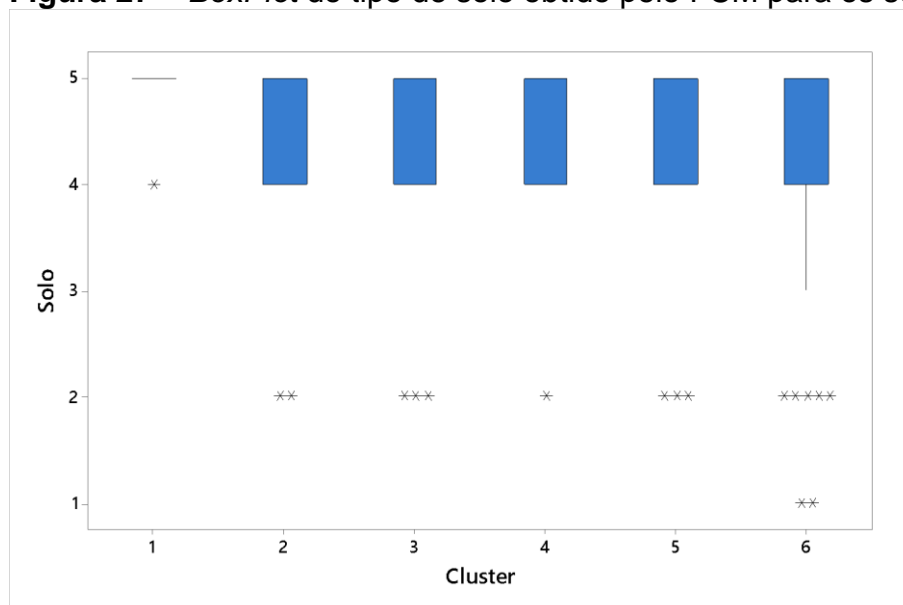
Tabela 12 – Análise estatística da temperatura média obtida pelo FCM para os seis *clusters*.

<i>Cluster</i>	Temperatura centroide	Desvio padrão
1	23,57a	0,517
2	23,69a	0,869
3	23,81ab	0,985
4	23,66a	1,111
5	23,84ab	1,233
6	22,98b	1,653

Fonte: Elaborada pela autora.

Ao analisar a Figura 27, que representa a análise estatística dos resultados obtidos para os seis *clusters* em relação ao tipo de solo, observa-se que, assim como no *K-means*, apesar do FCM ter identificado apenas dois tipos de solo como predominantes (Latosolos e Argissolos), alguns *clusters* são compostos por elementos que apresentam outros tipos de solos menos indicados para a produção de cana de açúcar.

Figura 27 – *BoxPlot* do tipo de solo obtido pelo FCM para os seis *clusters*.



Fonte: Elaborada pela autora.

O *cluster* 1, com maior índice de produtividade, tem apenas um outliers com tipo de solo que não seja do tipo 5, Latossolo, que é o município de Piracicaba que possui predominância do solo tipo 4, Argissolo. O *cluster* 2, 3, 4 e 5, apesar de também serem compostos basicamente pelos solos 4 e 5, possuem alguns *outliers* do

solo numerado como 2, que é o Neossolos. O *cluster* 6 é o com maior número de *outliers* e o único que, além de ter elementos com solo do tipo 2, Neossolos, tem do tipo 1, Espodosolos.

Apesar do algoritmo FCM, assim como o *K-means*, não ter possibilitado uma análise mais aprofundada da relação entre o tipo de solo e a produtividade obtida pelos *clusters*, a análise estatística realizada por meio do gráfico *BoxPlot* permitiu concluir que o tipo de solo possui influência na produtividade.

Na Tabela 13, pode-se verificar o baixo desvio padrão do *cluster* um, por ser formado apenas por elementos com os tipos de solo 1 e 2, enquanto que os outros *clusters* apresentam valores maiores para o desvio padrão.

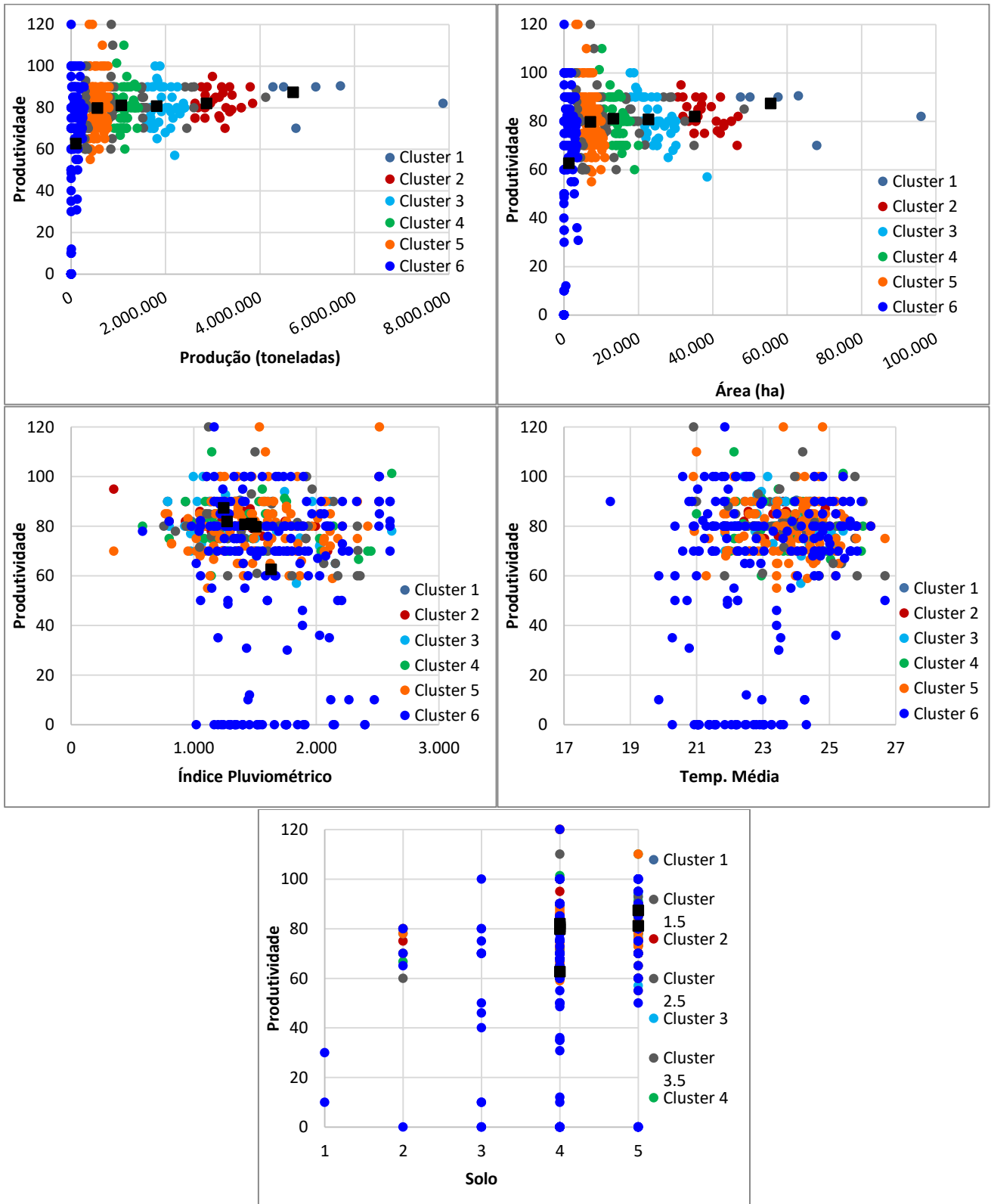
Tabela 13 – Análise estatística do tipo de solo obtido pelo FCM para os seis *clusters*.

<i>Cluster</i>	Tipo de solo centroide	Desvio padrão dos dados
1	4,86ab	0,3780
2	4,45a	0,7942
3	4,50a	0,7752
4	4,53ab	0,5704
5	4,34ab	0,5845
6	4,12b	0,7294

Fonte: Elaborada pela autora.

Na Figura 28, avalia-se a relação de cada atributo discutido neste trabalho com a produtividade. Os pontos coloridos representam os objetos inseridos em cada um dos *cluster*. Neste caso, encontra-se também nos gráficos alguns pontos cinzas, que representa os elementos que estão na fronteira segundo os resultados do FCM. Isso significa que todos os elementos que obtiverem índice menor que 0,55 (em uma escala de 0 até 1) de pertencimento a um mesmo *clusters*, foram definidos como elementos de fronteira.

Figura 28 – Gráficos de dispersão relacionando os atributos à produtividade, obtidos com o FCM.



Fonte: Elaborada pela autora.

No *cluster* 1, dos sete municípios que foram agrupados, um deles foi identificado na fronteira com o *cluster* dois. Batatais apresentou grau de pertinência

de 0,4223 para o *cluster* um e 0,3856 para o *cluster* dois. Apesar de apresentar alta produtividade, característica que contribuiu para o município permanecer teoricamente no *cluster* um, Batatais possui índice pluviométrico acima da média do *cluster*, com 1.378 mm em 2017, aproximando-se mais do *cluster* dois, com centroide de 1.287mm, enquanto que o *cluster* um apresentou um centroide igual a 1.180mm.

O município de Santo Antônio da Alegria está situado na região de fronteira entre o *cluster* cinco e seis. Apesar de apresenta uma média produtiva relativamente alta, de 80 toneladas por hectare, o município registrou em 2017 uma temperatura média muito baixa, de 21,5 °C, o que o aproximou do *cluster* seis. Além disso, o solo predominante do município é do tipo 2, Neossolos, outro fator que o aproxima do padrão identificado no *cluster* seis.

A matriz de pertinência dos municípios em região de fronteira esta apresentada no APENDICE D.

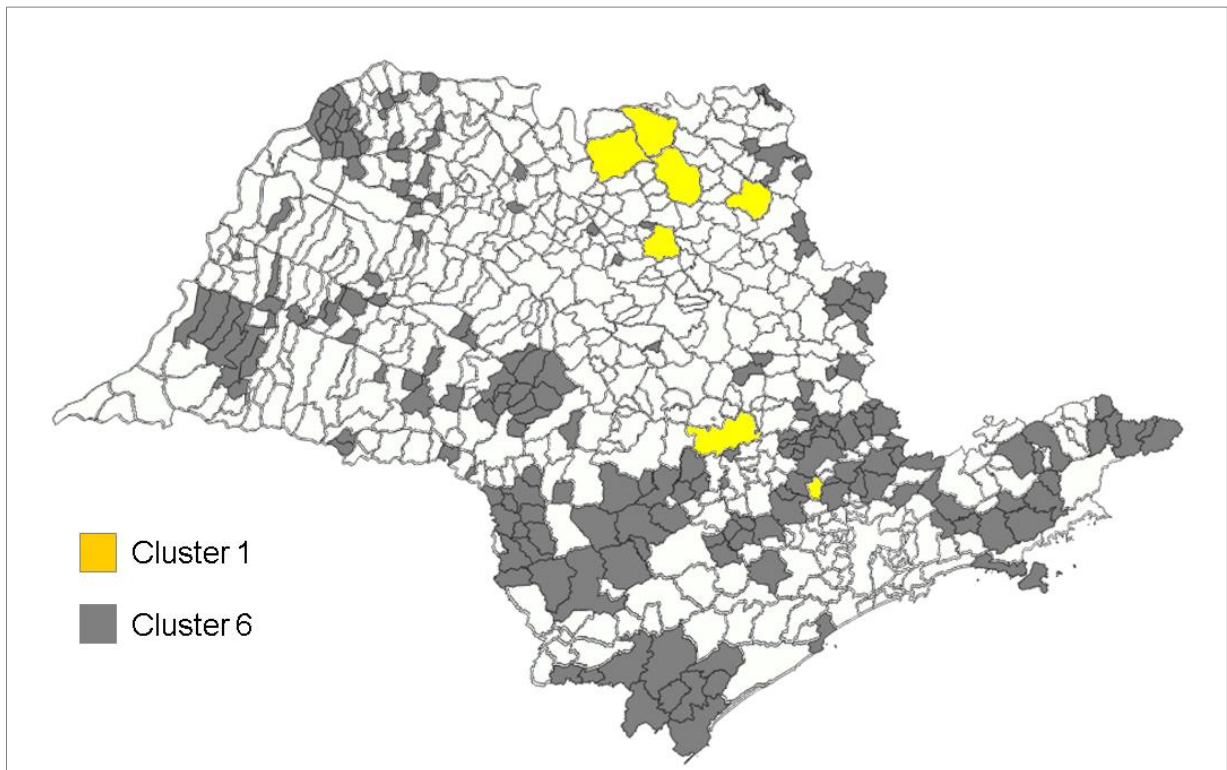
Ao analisar a relação entre a produtividade e a produção de cana de açúcar, assim como a relação da produtividade com a área, nota-se um comportamento muito semelhante ao resultado já obtido pelo *K-means*. A produtividade em pequenas produções varia desde zero até 120 toneladas por hectare, demonstrando a dificuldade de controle e padronização da produtividade em pequenas áreas produtivas. Por outro lado, percebe-se que em grandes áreas possui uma padronização maior da produtividade, com municípios conseguindo manter-se próximos a 80 toneladas por hectare.

Ao relacionar em um gráfico de dispersão o índice pluviométrico e a produtividade, assim como o gráfico que relaciona com a temperatura média, torna-se difícil identificar uma possível influência das variáveis na produtividade.

Na relação entre o tipo de solo e a produtividade acontece a mesma situação, porém, percebe-se que o *cluster* seis é o único que apresenta baixos níveis de produtividade combinados com solos do tipo um, dois e três.

O mapa demonstrado na Figura 29 representa os municípios que foram agrupados no *cluster* 1, ou seja, o *cluster* com maior índice de produtividade.

Figura 29 – Representação dos municípios pertencentes aos *clusters* um e seis do FCM.



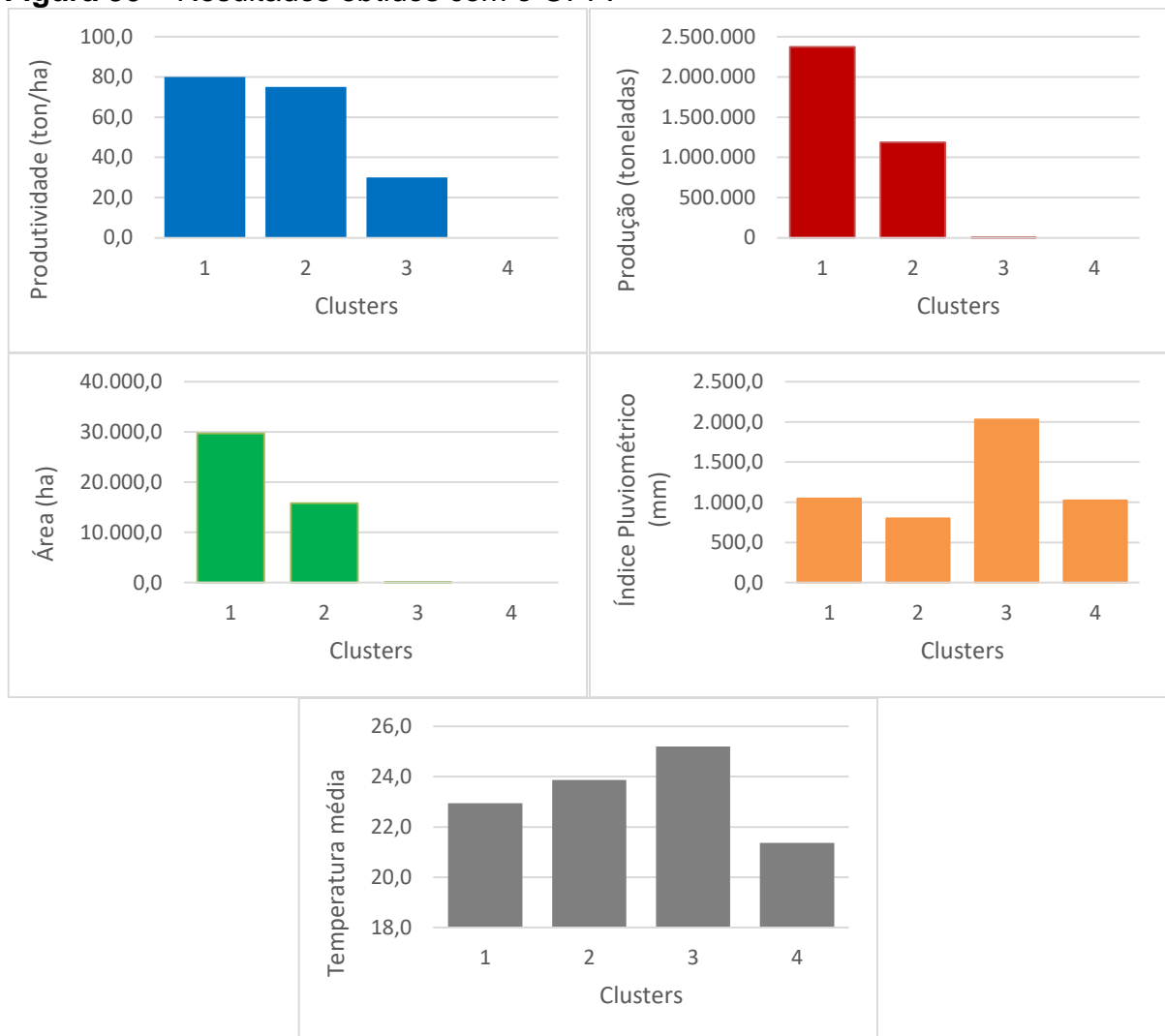
Fonte: Elaborada pela autora.

Assim como ocorreu no mapeamento realizado com os resultados do *K-means*, os municípios estão concentrados no Nordeste, Oeste e no Centro paulista. Os municípios pertencentes ao *cluster* um são os mesmos dos aglomerados pelo *K-means*, com exceção de Guaraci e Valparaíso, que não fazem parte deste *cluster*. Temos então com maior média de produtividade são: Barretos, Batatais, Guáira, Ituverava, Jaboticabal, Morro Agudo e Piracicaba

4.3 Resultados e análises obtidas por meio do OPF

Os resultados obtidos nos quatro *clusters* do OPF estão demonstrados na Figura 30. Nesta Figura, cada coluna de cada gráfico refere-se a um *cluster* encontrado pelo algoritmo. O eixo vertical de cada gráfico mostra o valor do medóide de cada grupo, que são objetos representativos de um *cluster* com um conjunto de dados cuja dissimilaridade média para todos os objetos no *cluster* é mínima, sendo que os medóides são semelhantes em conceito a médias ou centroides, mas os medóides são sempre restritos a serem membros do conjunto de dados.

Figura 30 – Resultados obtidos com o OPF.



Fonte: Elaborado pela autora.

Ao analisar os gráficos de barras gerados pelos resultados dos quatro *clusters* que foram agrupamentos pelo algoritmo OPF, percebe-se que, assim como nos outros algoritmos, os gráficos de área destinada ao plantio de cana de açúcar e de produção da cultura possuem médias com tendências semelhantes para cada *clusters*. Porém, diferentemente do *K-means* e do *Fuzzy C-means*, os rótulos para produção e área do *cluster* 4 foram iguais a zero. Dos 56 municípios que estão inseridos no *cluster* quatro, 32 possuem área e produção igual a zero.

Considerando o gráfico da produtividade, o *cluster* um e dois, que são os com maiores índices de produtividade, são formados por municípios que também possuem grandes áreas e produções, demonstrando mais uma vez a influência da quantidade produzida na obtenção de grandes índices de produtividade.

A relação das variáveis produtividade com o índice pluviométrico não ficou tão clara no OPF como nos outros algoritmos apenas analisando os gráficos de barras, levando em consideração que o índice pluviométrico do *cluster* 1, 1047 mm, e o do *cluster* quatro, 1021 mm, são muito próximos. Em relação a temperatura, observa-se que o *cluster* quatro, assim como nos outros algoritmos, foi aquela com menor temperatura média e produtividade, com uma média inferior a 22° em 2017.

O resultado obtido pela clusterização para o tipo de solo predominante em cada um dos *clusters*, apresentado no Quadro 10, apresentou resultado diferente dos outros métodos. O OPF foi capaz de identificar um *cluster* com rótulo de tipo de solo que não havia sido identificado antes neste trabalho

Quadro 10 – Tipo de solo predominante de cada *cluster* do OPF.

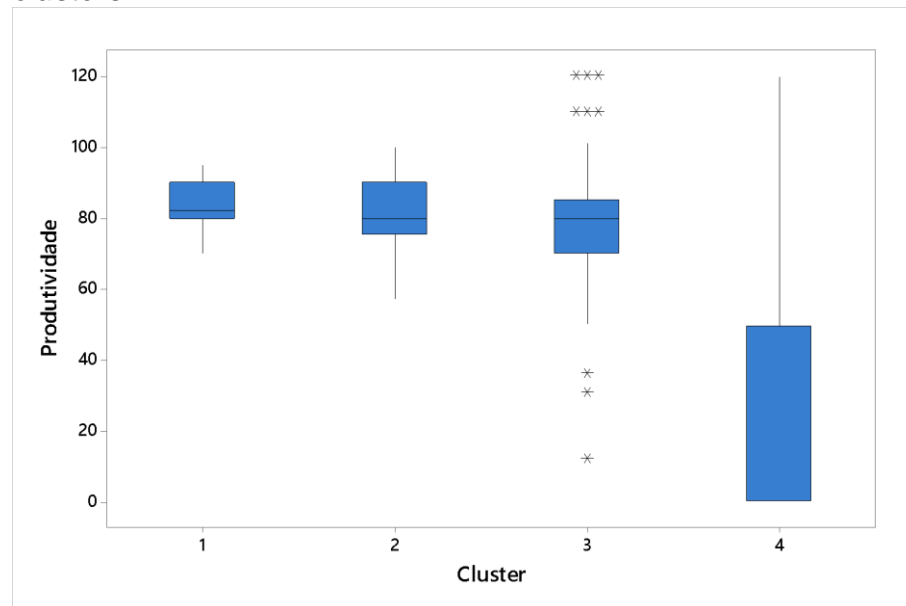
Cluster	Tipo de solo
1	Latossolos
2	Latossolos
3	Argissolos
4	Neossolos

Fonte: Elaborado pela autora.

Aprofundando a discussão a respeito dos resultados obtidos pelo OPF, realizou-se análises estatísticas e elaborou-se gráficos *Boxplot* identificando a mediana de cada *cluster*, a caixa de amplitude interquartílica e os traços que se estendem de ambos os lados da caixa.

Ao analisar a Figura 31, que representa a análise estatística dos resultados obtidos para os quatro *clusters* em relação a produtividade, observa-se medianas equilibradas para os *clusters* um, dois e três, apesar de apresentarem sutil tendência de queda, e uma mediana bem abaixo para o *cluster* quatro. O *cluster* um apresenta pequenos traço que se estende para cima e para baixo da caixa, demonstrando uma pequena amplitude dos dados para baixo, ao contrário do *cluster* quatro com grande amplitude. Assim como nos outros algoritmos, o *cluster* um não apresentou *outliers* que demonstrassem uma maior dispersão dos dados e neste caso, o mesmo aconteceu para o *cluster* dois.

Figura 31 – *BoxPlot* da produtividade obtida pelo OPF para os quatro *clusters*.



Fonte: Elaborada pela autora.

Nos algoritmos *K-means* e FCM, o *cluster* com menor produtividade apresentou mediana abaixo de 80 toneladas por hectare e várias ocorrências de *outliers* com produtividade zero, o que no caso do OPF ocorreu de forma diferente. No OPF o *cluster* quatro não apresentou nenhuma ocorrência de *outliers*, isso porque caixa de amplitude interquartílica dos dados varia de zero a aproximadamente 40 toneladas por hectares, o que quer dizer que 75% dos dados estão dentro deste intervalo.

Com isso, pode-se concluir que o agrupamento dos dados no OPF também foi eficiente no agrupamento dos municípios com produtividade 0, neste caso, para agrupar municípios que baixa produtividade, mais eficiente que os outros métodos.

No caso dos algoritmos *K-means* e FCM, sabe-se que o centro (rótulo) de cada *cluster* está relacionado a média das ocorrências aglomeradas. Porém, no OPF, o rótulo de cada *cluster* é representado pelos dados de um município específico que melhor representa as características de todos os aglomerados naquele *cluster*. Por isso, na Tabela 14 foram apresentadas tanto a mediana dos dados como a média, para assim ser possível comparar os dois valores com os rótulos de OPF e analisar qual mais se aproxima.

Tabela 14 – Análise estatística da produtividade obtida pelo OPF para os quatro *clusters*.

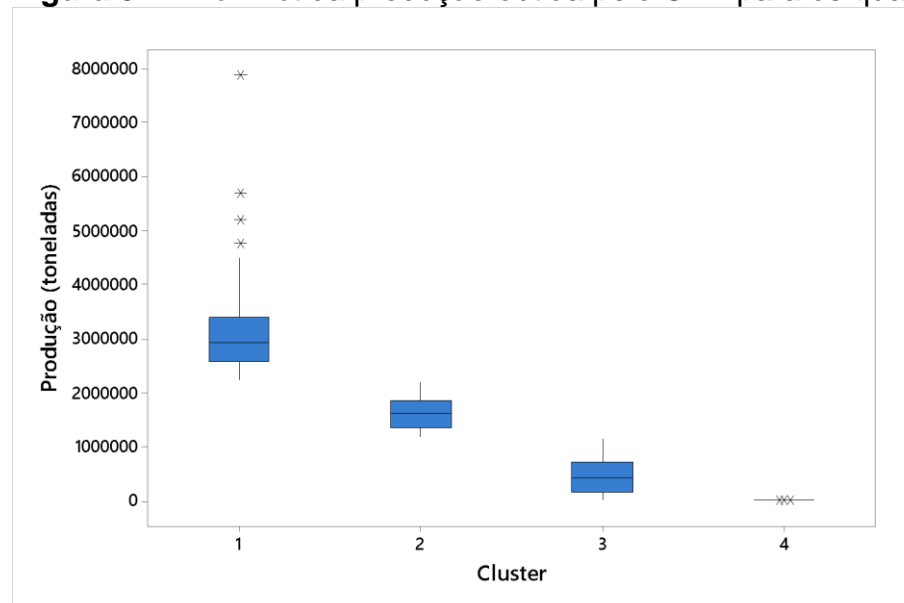
<i>Cluster</i>	Rótulo do OPF	Mediana da produtividade	Produtividade média	Desvio padrão
1	80	82	82,57a	6,04
2	75	80	81,05a	8,09
3	30	80	79,41a	12,31
4	0	0	22,40b	32,17

Fonte: Elaborada pela autora.

A produtividade média dos *clusters* um, dois e três possuem a mesma letra em relação a suas médias, isso significa que, estatisticamente os dados não apresentam diferença significativa, diferenciando-se de maneira significativa apenas o *cluster* quatro.

Figura 32, representa a análise estatística dos resultados obtidos para os quatro *clusters* em relação a produção.

Figura 32 – *BoxPlot* da produção obtida pelo OPF para os quatro *clusters*.



Fonte: Elaborada pela autora.

A mediana de cada *clusters* é bastante distinta, variando de uma mediana que se aproxima de 3 milhões de toneladas até uma mediana igual a zero. Observa-se também que os traços que se estendem de ambos os lados das caixas, representado a amplitude dos dados, são pequenos, o que em um primeiro momento permite concluir que os dados são concisos e com desvio padrão baixo, sendo o *cluster* um o que apresenta maior dispersão dos dados.

Assim como no *K-means* e no FCM, o *cluster* um apresentou um elemento *outlier*, que representa o município de Morro Agudo, que é o maior produtor de cana de açúcar do estado de São Paulo, com uma produção em 2017 de quase 8 milhões de toneladas. O *cluster* quatro não apresentou outliers, assim como no gráfico da produtividade, por ter seus dados concentrados em zero toneladas de produção.

Nos valores médios e medianas pertencentes a cada *cluster*, demonstrados na Tabela 15, nota-se que a mediana identificada na análise estatística mais se aproxima dos rótulos obtidos pelo OPF do que as médias, assim como aconteceu na tabela da produtividade. Em relação ao desvio padrão, confirma-se a concisão dos dados pertencentes a cada *clusters*, demonstrada pelos baixos valores de desvio padrão, com exceção do *cluster* um, que apresentou alguns outliers.

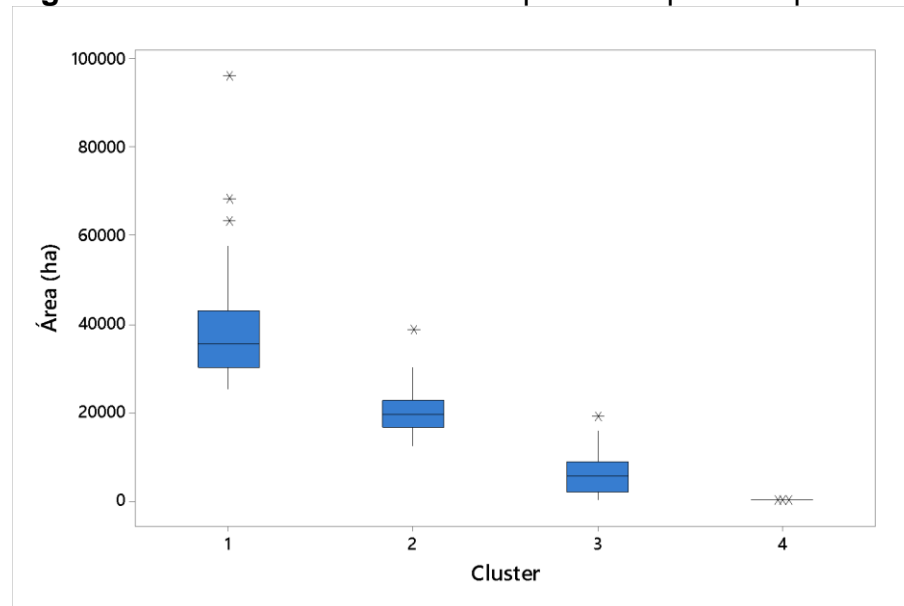
Tabela 15 – Análise estatística da produção obtida pelo OPF para os quatro *clusters*.

<i>Cluster</i>	Rótulo do OPF	Mediana da produção	Produção média	Desvio padrão
1	2.375.120	2.916.000	3.184.789a	1.003.498
2	1.185.000	1.603.073	1.608.149b	292.016
3	2.400	420.000	455.807c	326.918
4	0	0	278d	442

Fonte: Elaborada pela autora.

Ressalta-se também que o agrupamento do método OPF para 4 *clusters*, causada por diferenças no conjunto das variáveis analisadas, teve grande influência da variável produção, uma vez que a análise estatística revelou diferença significativa nas médias de todos *clusters*, pois a produção média dos dados de cada *cluster* que não compartilham uma mesma letra são significativamente diferentes.

A Figura 33, representa a análise estatística dos resultados obtidos para os quatro *clusters* em relação a área, que se comportou de maneira semelhante ao gráfico da produção do método OPF, com medianas variando de maneira significativa e uma amplitude dos dados muito pequeno no *cluster* quatro, que estão inseridos os municípios com baixíssima área e produção de cana de açúcar em 2017.

Figura 33 – BoxPlot da área obtida pelo OPF para os quatro clusters.

Fonte: Elaborada pela autora.

Nos valores médios e medianas pertencentes a cada *cluster*, demonstrados na Tabela 16, nota-se que a mediana identificada na análise estatística mais se aproxima dos rótulos obtidos pelo OPF do que as médias. Em relação ao desvio padrão, confirma-se a concisão dos dados pertencentes a cada *clusters*, demonstrada pelos baixos valores de desvio padrão, com exceção do *cluster* um, que apresentou alguns *outliers*.

Tabela 16 – Análise estatística da área obtida pelo OPF para os quatro clusters.

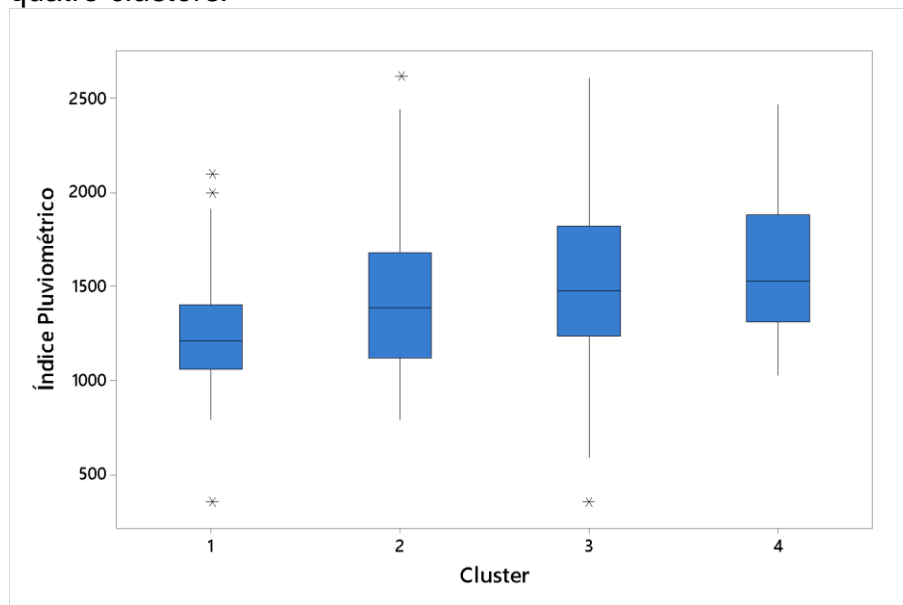
Cluster	Rótulo do OPF	Mediana da área	Área média	Desvio padrão
1	29.689	35.463	38.642,6a	12.024,4
2	15.800	19.293	20.090,6b	4.581,8
3	40	5.500	5.800,4c	4.200,1
4	0	0	5,9d	9,7

Fonte: Elaborada pela autora.

O agrupamento do método OPF para 4 *clusters*, causado por diferenças no conjunto das variáveis analisadas, teve grande influência da variável produção, uma vez que a análise estatística revelou diferença significativa nas médias de todos *clusters*, pois a produção média dos dados de cada *cluster* que não compartilham uma mesma letra são significativamente diferentes.

A Figura 34, que representa a análise estatística em relação ao índice pluviométrico, comportou-se de forma semelhante aos identificados nos métodos *K-means* e *Fuzzy C-means*. O Gráfico *BoxPlot* elaborado para analisar os resultados obtidos pelo OPF demonstrou-se importante, visto que ao analisar apenas os rótulos identificados pelo algoritmo, no início desta seção, não foi possível identificar um padrão em relação ao agrupamento, considerando que o *cluster* dois foi rotulado com menor índice pluviométrico, não atingindo nem mil milímetros em 2017. Esse comportamento não foi semelhante no gráfico BloxPlot, que mostra a menor mediana identificada no *cluster* 1 com tendência de aumento para os demais.

Figura 34 – *BoxPlot* do índice pluviométrico obtido pelo OPF para os quatro *clusters*.



Fonte: Elaborada pela autora.

Diferentemente dos outros algoritmos que apresentaram desvio padrão alto no *cluster* com menor índice de produtividade, o alto desvio padrão neste caso está no *cluster* três, como demonstrado na Tabela 17. Além disso, mais uma vez, a mediana mais se aproximou do rótulo do OPF do que a média.

Tabela 17 – Análise estatística do índice pluviométrico obtido pelo OPF para os quatro *clusters*.

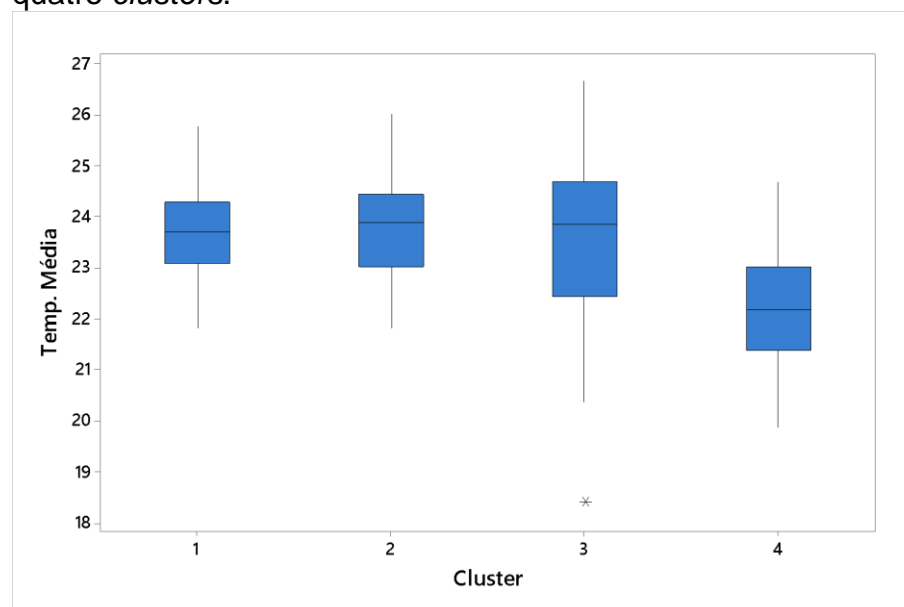
<i>Cluster</i>	Rótulo do OPF	Mediana do índice pluviométrico	Índice Pluviométrico médio	Desvio padrão
1	1.047,5	1.208,3	1.257a	315,4
2	799,2	1.389,1	1.441a	401,8
3	2.027,5	1.474,3	1.554ab	419,5
4	1.021,2	1.526,1	1.618b	373,9

Fonte: Elaborada pela autora.

Pela Tabela 17 podemos identificar que apenas o *cluster* um e dois não apresentaram médias com diferenças significativas para a estatística, com os dois compartilhando da mesma letra. Porém, revela-se que houve diferença significativa em relação aos outros *clusters*, o que significa houve influência do índice pluviométrico no agrupamento dos dados.

A análise estatística para a temperatura média está representada na Figura 35.

Figura 35 – *BoxPlot* da temperatura média obtida pelo OPF para os quatro *clusters*.



Fonte: Elaborada pela autora.

Neste caso, assim como no índice pluviométrico, o gráfico *BoxPlot* elaborado demonstrou-se importante para as análises, visto que ao analisar apenas os rótulos identificados pelo algoritmo, no início desta seção, não foi possível identificar um padrão em relação ao agrupamento, pois enquanto os *clusters* um, dois

e três demonstravam uma tendência de aumento da temperatura média, o *cluster* quatro foi que o apresentou menor média.

No caso da Figura 35, as medianas dos *clusters* um, dois e três são muito similares, enquanto que a mediana do *cluster* quatro apresenta-se muito abaixo. Isso pode ser confirmado na Tabela 18, que mostra que o *cluster* quatro é o único com diferença estatística significativa, com letra diferente das demais médias.

Tabela 18 – Análise estatística da temperatura média obtida pelo OPF para os quatro *clusters*.

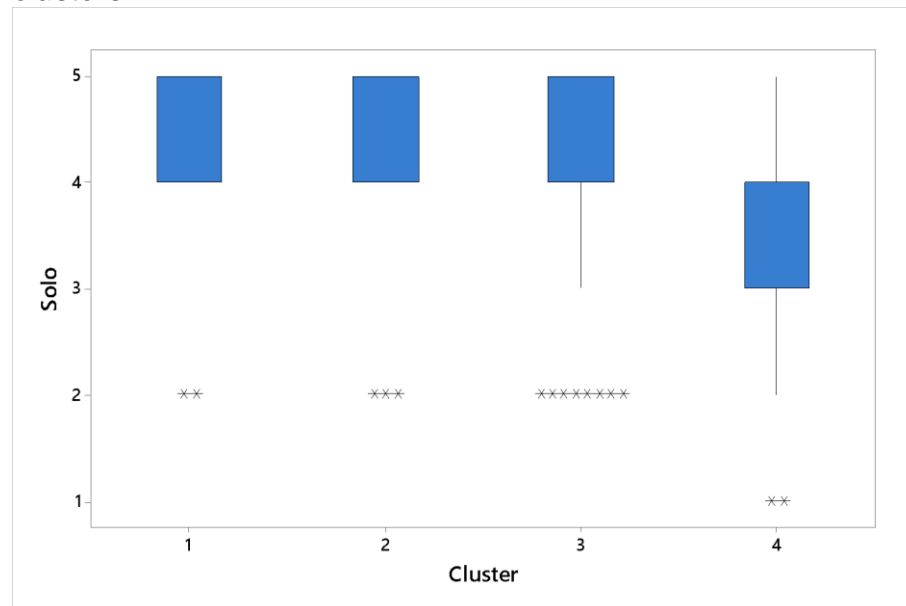
<i>Cluster</i>	Rótulo do OPF	Mediana da temperatura média	Temperatura média	Desvio padrão
1	23,0	23,7	23,70a	0,882
2	23,9	23,8	23,81a	0,992
3	25,2	23,8	23,58a	1,459
4	21,4	22,2	22,16b	1,154

Fonte: Elaborada pela autora.

Ao analisar A Figura 36, que representa a análise estatística dos resultados obtidos para os quatro *clusters* em relação ao tipo de solo, observa-se que, o *cluster* um é composto por basicamente municípios com tipo de solo quatro e cinco, tendo poucos *outliers*. O *cluster* três, apesar de ter 75% dos municípios com solos do tipo quatro e cinco, apresentou uma quantidade considerável de solo tipo dois. O *cluster* quatro é o único que tem inseri municípios com solo tipo um, com maiores ocorrências de solos do tipo três e quatro.

Esse resultado apresentado pelo *cluster* 4 diferencia o resultado do OPF dos demais. Apesar de nos outros algoritmos o último *cluster* apresenta maior quantidade de *outliers* com solos do tipo um e dois, ainda assim eles eram compostos por 75% de municípios com solos quatro e cinco. Esse resultado do *cluster* quatro com 75% dos municípios com solos do três e quatro demonstra que o tipo de solo foi fundamental para o agrupamento deste *cluster*.

Figura 36 – *BoxPlot* do tipo de solo obtido pelo OPF para os quatro *clusters*.



Fonte: Elaborado pela autora.

A importância do tipo de solo no agrupamento do *cluster* quatro pode ser confirmado na Tabela 19, que mostra similaridade entre as médias dos *clusters* um, dois e três, apesar de um leve queda, e o *cluster* quatro sendo o único com diferença estatística significativa, com letra diferente das demais médias.

Tabela 19 – Análise estatística do tipo de solo obtido pelo OPF para os quatro *clusters*.

<i>Cluster</i>	Rótulo do OPF	Mediana do tipo de solo	Tipo de solo médio	Desvio padrão
1	5	5	4,6a	0,6958
2	5	4	4,5a	0,6979
3	4	4	4,3a	0,6122
4	2	4	3,8b	0,8761

Fonte: Elaborada pela autora.

Especificamente neste caso, a média se aproximou mais do rótulo do OPF do que a mediana.

Na Figura 37 nota-se que, apesar de agrupar somente quatro *clusters*, o comportamento das variáveis neste método foi muito parecido com os dois outros métodos. Os dados agrupados no *cluster* um, com pequena produção, possuem uma enorme variabilidade de produtividade, o oposto do que acontece no *cluster* um, em

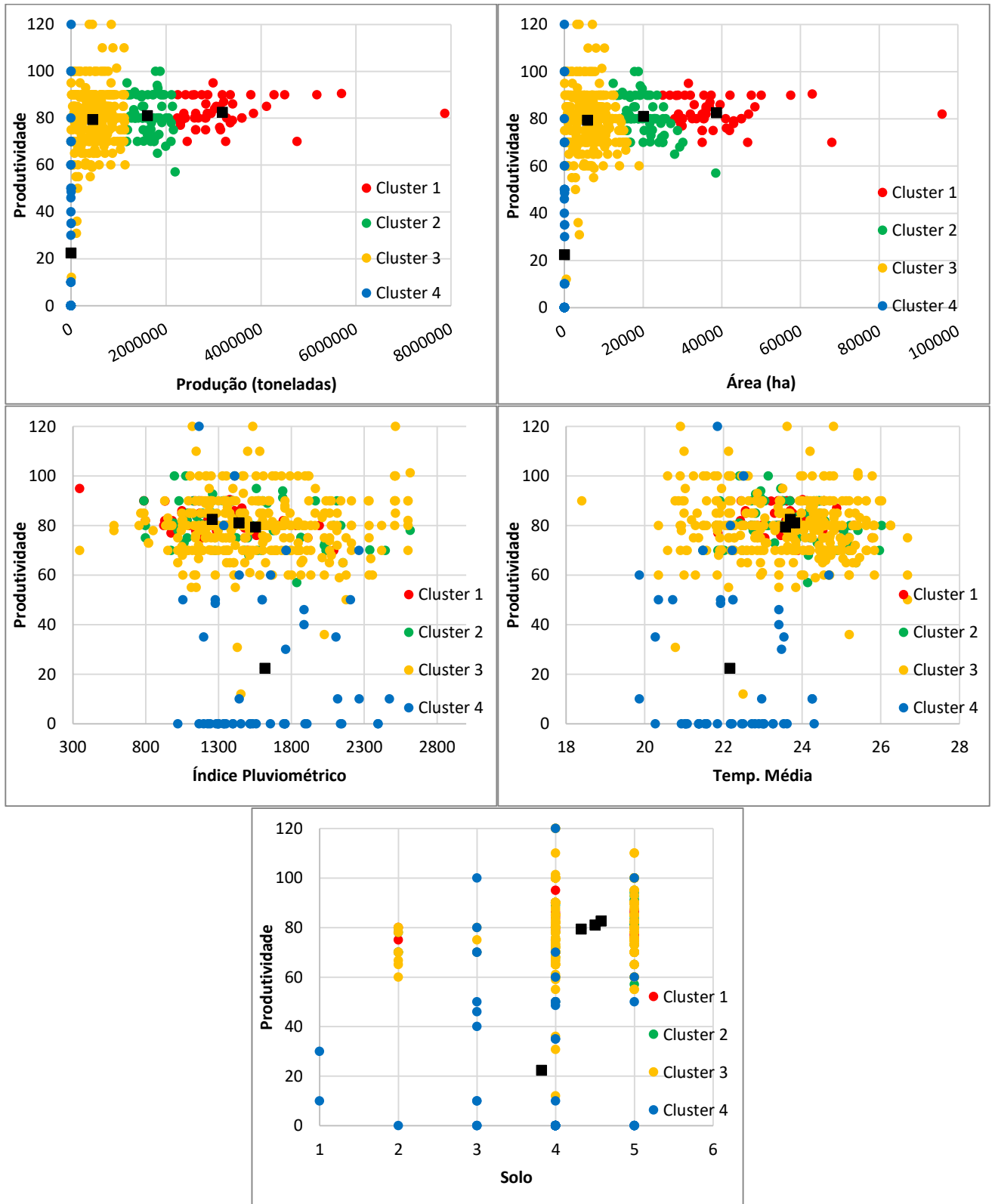
que todos os elementos possuem uma produção acima de 2 milhões de tonelada e uma produtividade se mantendo ente 60 e 100 toneladas por hectare.

A relação entre a área e produtividade possui um comportamento semelhante ao gráfico de dispersão que relacionou a produção com a produtividade.

Ao relacionar em um gráfico de dispersão a temperatura média e a produtividade, nota-se que, assim como no índice pluviométrico, apenas um *cluster*, em azul, aparece de forma mais clara com seus pontos afastados dos outros, assim como seu centro, que possui menos temperatura média e também menor produtividade.

Em relação ao gráfico de dispersão que relaciona a produtividade e o tipo de solo, observa-se que apenas os *clusters* três e quatro apresentaram objetos com produtividade abaixo de 60 toneladas por hectare, comprovando mais uma vez a influência do tipo de solo na obtenção de produtividade.

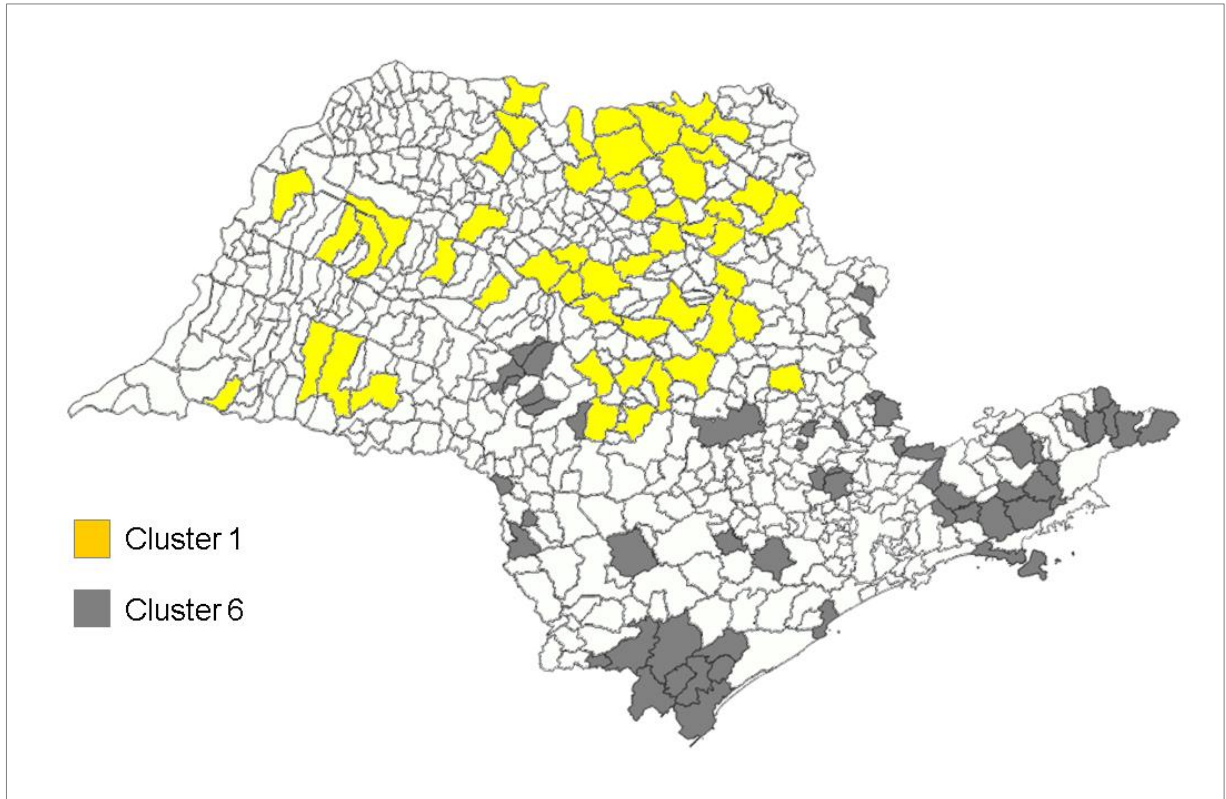
Figura 37 – Gráficos de dispersão relacionando os atributos à produtividade, obtidos com o OPF.



Fonte: Elaborada pela autora.

O mapa demonstrado na Figura 38 representa os municípios que foram agrupados no *cluster 1*, ou seja, o *cluster* com maior índice de produtividade.

Figura 38 – Representação dos municípios pertencentes aos *clusters* um e seis do OPF.



Fonte: Elaborada pela autora.

O mapeamento dos municípios no algoritmo OPF foi diferente do *K-means* e FCM, pois o OPF aglomerou maior quantidade de municípios no *cluster* um. Porém, da mesma forma como ocorreu nos outros algoritmos estudados neste trabalho, os municípios pertencentes ao *cluster* um, com maior média produtiva, também estão concentrados na parte de cima do mapa do estado de São Paulo.

Além de agrupar os municípios já pertencentes ao *cluster* um dos outros algoritmos, Barretos, Batatais, Guáira, Guaraci, Ituverava, Jaboticabal, Morro Agudo, Piracicaba e Valparaíso, o OPF agrupou também: Altinópolis, Andradina, Araçatuba, Araraquara, Araras, Bebedouro, Boa Esperança Do Sul, Borborema, Brotas, Colina, Colômbia, Descalvado, Dois Córregos, Guararapes, Ibitinga, Ipuã, Itajobi, Itápolis, Jardinópolis, Jaú, José Bonifácio, Lençóis Paulista, Lins, Luiz Antônio, Martinópolis, Miguelópolis, Novo Horizonte, Olímpia, Palestina, Paraguaçu Paulista, Paulo De Faria,

Pederneiras, Penápolis, Pitangueiras, Rancharia, Ribeirão Preto, Sandovalina, São Carlos, São Joaquim Da Barra, São Manuel, Sertãozinho, Tanabi, Taquaritinga.

De acordo com Aguiar et al (2009), a região Centro-Norte do estado tem maior tradição no cultivo da cana-de-açúcar e por esse motivo é a região com maior concentração desta cultura agrícola. A Região de Ribeirão Preto possui a maior porcentagem de ocupação do território destinado a produção de cana de açúcar, com 51,0% da sua área ocupada pelo cultivo da cultura. A região de Barretos, Franca e Central possuem mais de 36,0% da sua área ocupada por cana de açúcar. Em contrapartida, no leste do estado, na região de São Paula, São José dos Campos, Santos e Registro são regiões com pouco desenvolvimento da cultura (AGUIAR et al, 2009).

5 CONSIDERAÇÕES FINAIS

O presente trabalho atingiu o objetivo de identificar padrões na produção de cana de açúcar no estado de São Paulo por meio da utilização de algoritmos de agrupamento de dados, extraindo conhecimentos e fornecendo indicativos de fatores que contribuam para um melhor padrão de produção.

Notou-se que os *clusters* com maior produtividade, aglomeraram também municípios com alta produção e área destinada à cana de açúcar e maior temperatura média, porém com menor índice pluviométrico registrado no ano. Conforme a produtividade diminuiu, foi reduzida também a produção, área e temperatura média, enquanto o índice pluviométrico aumentou.

Os *clusters* com menor centroide de produtividade estão concentrados municípios com produções e áreas baixíssimas, índice pluviométrico que ultrapassa 1600mm ao ano e temperatura bem abaixo das registradas nos demais *clusters*. Isso pode significar que a quantidade produzida pode influenciar na obtenção de grandes índices de produtividade. O clima no estado de São Paulo favorece esse cenário, pois, em geral, o estado tem um período de chuvas bem distribuído (de setembro ou outubro até março ou abril), que coincide com período quente (altas temperaturas) e de crescimento dos colmos da cana. Esse comportamento se repetiu nos três algoritmos de reconhecimento de padrões utilizados neste trabalho.

O resultado obtido pela *clusterização* para o tipo de solo foi diferente apenas nos centroides identificados para cada *cluster*. Enquanto o *K-means* e o FCM tiveram como centroide para todos os *clusters* apenas os solos mais indicados para produção de cana de açúcar, que são os Latossolos e Argissolos, mesmo no *cluster* com menor índice de produtividade, o OPF identificou para o *cluster* com menor produtividade solo Neossolos.

Foi possível atingir também o objetivo de avaliar as ferramentas de Inteligência Artificial *K-means*, *Fuzzy C-means* e OPF no reconhecimento de padrões no agronegócio, sendo que os três métodos foram eficazes no agrupamento dos dados, proporcionando resultados que, quando analisados de forma mais profunda, percebe-se que são semelhantes. Comprovou-se a eficácia do algoritmo OPF, que ainda é um método considerado inovador, principalmente no agrupamento de dados relacionados em agronegócio. Além disso, os padrões foram analisados e discutidos estatisticamente.

Apesar das semelhanças apresentadas nos resultados dos métodos, cada um deles teve um papel importante na discussão dos resultados. O *K-means* é um método tradicional e comprovadamente eficiente no agrupamento de dados, muito utilizado para realizar comparações e avaliar resultados de métodos ainda considerados novos. O FCM proporcionou a identificação de sobreposição dos dados, permitindo a identificação do grau de pertinência dos municípios para seus *clusters*. O OPF, que foi utilizado com o intuito de avaliar sua eficácia no agrupamento dos dados, apresentou resultados robustos. Porém, o OPF possui como principal diferença o fato de apresentar medóide como legenda de seus *clusters* e não centroide como no *K-means* e FCM. Percebeu-se que, muitas vezes, os medóides se distanciam consideravelmente da média do grupo.

Forneceram-se mapeamentos e indicativos de representatividade da produção de cana-de-açúcar nos municípios analisados. Nos mapas elaborados com os resultados do agrupamento dos algoritmos foi possível identificar que os municípios analisados que possuem como padrão maior média de produtividade estão concentrados na parte superior do mapa, não sendo identificados municípios com esse padrão no sul do estado de São Paulo.

Sugere-se para os envolvidos no setor do agronegócio, a partir dos resultados obtidos, estudos aprofundados relacionados a cana-de-açúcar e os fatores que influenciam sua produtividade, podendo utilizar outros atributos em análises como a realizada neste trabalho. Quanto a novos estudos relacionadas à Inteligência Artificial e Reconhecimento de Padrões, sugere-se a aplicação dos algoritmos utilizados neste trabalho em outras bases de dados e até mesmo a avaliação de novos algoritmos com a mesma base de dados.

REFERÊNCIAS

AGUIAR, D. A.; SILVA, W. F.; RUDORFF, B. F. T.; SUGAWARA, L. M.; CARVALHO, M. A.; Expansão da cana-de-açúcar no Estado de São Paulo: safras 2003/2004 a 2008/2009. 2009. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 14., 2009, Natal. **Anais eletrônico...** Natal: INPE. 2009. Disponível em: <<http://marte.sid.inpe.br/col/dpi.inpe.br/sbsr@80/2008/11.17.18.21/doc/9-16.pdf>>. Acesso em: 30 abr. 2019.

ALDENDERFER, M. S.; BLASHFIELD, R. K. **Cluster Analysis**. Beverly Hills, CA: Sage Publications, 1984.

ALPAYDIN, E. **Introduction to Machine Learning**. 2. ed. London: The MIT Press, 2010.

ARBEX, M. A. **Avaliação dos efeitos do material particulado proveniente da queima da plantação de cana-de-açúcar sobre a morbidade respiratória na população de Araraquara-SP**, 2001. 188 f. Tese (Doutorado em Medicina) - Universidade de São Paulo, São Paulo, 2001.

ÁVILA, R. C. F. **Os impactos econômicos e ambientais da cultura da cana-de-açúcar no município de Barretos - São Paulo**. 2014. 64 f. Monografia (graduação) - Universidade de Brasília, Instituto de Ciências Humanas, Departamento de Geografia, Universidade Aberta do Brasil, 2014

AY, M.; KISI, O. Modelling of chemical oxygen demand by using ANNs, ANFIS and K-means clustering techniques. **Journal of Hydrology**, v. 511, p. 279-289, 2014.

BAI, C.; DHAVALÉ, D.; SARKIS, J. Complex investment decisions using rough set and fuzzy c-means: an example of investment in green supply chains. **European journal of operational research**, v. 248, n. 2, p. 507-521, 2016.

BASER, F. GOKTEN, S. GOKTEN, P. O. Using fuzzy c-means clustering algorithm in financial health scoring. **The Audit Financiar journal**, v. 15, n. 147, p. 385-385, 2017.

BATALHA, M. O. **Gestão agroindustrial**. 3. ed. São Paulo: Atlas, 2007.

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

BISHOP, C. M. **Pattern Recognition and Machine Learning**. New York: Springer, 2006.

BORGWARDT, S.; BRIEDEN, A.; GRITZMANN, P. An LP-based K-means algorithm for balancing weighted point sets. **European Journal of Operational Research**, v. 263, n. 2, p. 349-355, 2017.

BOVO, A. B. **Um método de tradução de fontes de informação em um formato padrão que viabilize a extração de conhecimento por meio de *link analysis* e teoria dos grafos**. 2004. Dissertação (Mestrado em Engenharia de Produção) - Universidade Federal de Santa Catarina, Florianópolis – SC, ano.

BRYNJOLFSSON, E.; ROCK, D.; SYVERSON, C. Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. **Economics of Artificial Intelligence**. University of Chicago Press, 2017. É livro ou revista?

CAMARA, M. R. G.; CALDARELLI, C. E. Expansão canavieira e o uso da terra no estado de São Paulo. **Estudos Avançados**, v. 30, n. 88, p. 93-116, 2016.

CASTRO, A. A. M.; PRADO, P. P. L. Algoritmos para reconhecimento de padrões. **Revista Ciências Exatas**, v. 8, n. 2002, 2002.

CASTRO, A. M. G; LIMA, S. M. V; CRISTO, C. M. P. N. Cadeia Produtiva: Marco Conceitual para Apoiar a Prospecção Tecnológica. In: SIMPÓSIO DE GESTÃO DA INOVAÇÃO TECNOLÓGICA, 12., 2002, Salvador. **Anais eletrônicos...** Salvador: SGIT, 2002. Disponível em: <http://200.198.192.95/portalmDIC/arquivos/dwnl_1197031881.pdf>. Acesso em: 05 mar. 2018.

CATI. Coordenadoria de Assistência Técnica Integral. **Quem somos**. 2018. Disponível em: < <http://www.cati.sp.gov.br/portal/institucional/quem-somos>>. Acesso em: 06 abr. 2018.

CESNIK, R. Melhoramento da cana-de-açúcar: marco sucro-alcooleiro no Brasil. **Embrapa Meio Ambiente**, 2004. Disponível em: < <https://www.alice.cnptia.embrapa.br/bitstream/doc/15939/1/2007AP008.pdf>>. Acesso em: 05 jan. 2018.

CHEN, S.; SUN, T.; YANG, F.; SUN, H.; GUAN, Y. An improved optimum-path forest clustering algorithm for remote sensing image segmentation. **Computers & Geosciences**, v. 112, p. 38-46, 2018.

CNA. CONFEDERAÇÃO DA AGRICULTURA E PECUÁRIO DO BRASIL.

Comunicado Técnico: Indicadores do PIB. 2018. Disponível em:

<<http://www.cnabrazil.org.br/artigos-tecnicos/comunicado-tecnico-indicadores-do-pib>>. Acesso em: 20 fev 2018.

COLETTA, L. F. S. **Agrupamento de dados fuzzy colaborativo**. São Paulo: USP, 2011. 95 f. Tese (doutorado) - Universidade de São Paulo, São Paulo, 2011.

CONAB. Companhia Nacional de Abastecimento. **Acompanhamento da sara brasileira**. 2018a. Disponível em:

<<file:///C:/Users/acer/Downloads/BoletimZCanaZ4ZLevantamentoZ17-18.pdf>>.

Acesso em: 10 mar. 2018.

CONAB. Companhia Nacional de Abastecimento. **Série histórica das safras**.

2018b. Disponível em: <<https://www.conab.gov.br/index.php/info-agro/safras/serie-historica-das-safras?start=20>>. Acesso em: 10 mar. 2019.

COSTA, A. L. S.; CLEPS, G D G. A produção sucroalcooleira em Morro Agudo (SP) e a migração piauiense. **CAMPO-TERRITÓRIO: revista de geografia agrária**, v. 9, n. 17. 2014.

COSTA, K. A. P.; PEREIRA, L. A. M.; NAKAMURA, R. Y. M.; PEREIRA, C. R.; PAPA, J. P.; FALCÃO, A. X. A nature-inspired approach to speed up optimum-path forest clustering and its application to intrusion detection in computer networks. **Information Sciences**, v. 294, p. 95-108, 2015.

DALFOVO, M. S.; LANA, R. A.; SILVEIRA, A. Métodos quantitativos e qualitativos: um resgate teórico. **Revista Interdisciplinar Científica Aplicada**, Blumenau, v.2, n.4, p.01- 13, sem II. 2008.

DIOLA, V.; SANTOS, F. Fisiologia. In: SANTOS, F.; BORÉM, A.; CALDAS, C. **Cana-de-açúcar: Bioenergia, Açúcar e Etanol. Tecnologias e Perspectivas**. 2. ed. Viçosa: Os Editores, 2011. pag. 25-49.

DUNN, J. C. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. **Journal of Cybernetics**, v. 3, p. 32-57, 1973.

EMBRAPA. Empresa Brasileira de Pesquisa Agropecuária. **Soluções tecnológicas**. 2018. Disponível em: <<https://www.embrapa.br/busca-de-solucoes-tecnologicas/-/produto-servico/49/agritempo>>. Acesso em: 06 abr 2018.

FALCÃO, A. X., STOLFI, J., LOTUFO R. A. The image foresting transform: Theory, algorithms, and applications. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 26.19-29, Jan 2004.

FERNANDES, A. J. **Manual da cana de açúcar**. Livroceres: Piracicaba, 1984.

GAUTHIER, J. **A cladistic analysis of the higher categories of the Diapsida**. Ph.D. thesis, University of California. University Microfilms, International, Ann Arbor, 1984.

GHOSH, S.; DUBEY, S. K. Comparative analysis of K-means and fuzzy c-means algorithms. **International Journal of Advanced Computer Science and Applications**, v. 4, n. 4, p. 35-39, 2013.

GOUVA, M.; KOTROTSIOU, S.; GOURGOULIANNIS, K.; SKENTERIS, N. Perceptions of Roma People Towards Public Health System and a Classification into Homogeneous Groups Using K-means Cluster Analysis. **European Psychiatry**, v. 30, p. 621, 2015.

GUIMARÃES, A. M. **Aplicação de computação evolucionária na mineração de dados físico-químicos da água e do solo**, 2005. 134 f. Tese (Doutorado em Agronomia). Universidade Estadual Paulista "Júlio de Mesquita Filho", Botucatu, 2005.

HESPANHOL, R. M. **Caracterização dos fatores contribuintes em acidentes de pequenas aeronaves da aviação geral brasileira utilizando inteligência artificial**. 2016. 82 f. Dissertação (Mestrado em Transportes) - Faculdade de Tecnologia, Universidade de Brasília – DF, 2016.

HESPANHOL R. M.; PEREIRA, D. R.; J. A. A. S, FORTES. Padrões nos Acidentes na Aviação Geral Brasileira utilizando Floresta de Caminhos Ótimos. In: SIMPÓSIO DE TRANSPORTE AÉREO, 2015. São José dos Campos. **Anais Eletrônicos...** São José dos Campos: SITRAER, 2015. Disponível em: <http://www.researchgate.net/publication/283463252_Padres_nos_Acidentes_na_Aviao_Geral_Brasileira_utilizando_Floresta_de_Caminhos_timos>. Acesso em: 03 mar. 2018.

IAC. Instituto agrônomo. **Mapa pedológico**. 2018. Disponível em: <http://www.iac.sp.gov.br/solosp/pdf/mapa_pedologico_Solos_Estado_de_Sao_Paulo.pdf>. Acesso em: 15 jul 2018.

IEA. Instituto de Economia Agrícola. **Cana de açúcar**. 2018a. Disponível em: <<http://ciagri.iea.sp.gov.br/nia1/cadeia/cadeiaCana.aspx>>. Acesso em: 05 fev. 2019

IEA. Instituto de Economia Agrícola. **Estatísticas da Produção Paulista**. 2018b. Disponível em: <http://ciagri.iea.sp.gov.br/nia1/subjetiva.aspx?cod_sis=1&idioma=1>. Acesso em: 06 fev. 2018.

IEA. Instituto de Economia Agrícola. **Quem somos**. 2019. Disponível em: <<http://www.iea.agricultura.sp.gov.br/out/instituto2.html>>. Acesso em: 26 jan. 2019.

JAIN, A. K.; DUBES, R. C. **Algorithms for clustering data**. NJ, Englewood Cliffs: Prentice-Hall, 1988.

JANK, M. S.; NASSAR, A. M.; TACHINARDI, M. H. Agronegócio e comércio exterior brasileiro. **Revista USP**, n. 64, p. 14-27, 2005.

JAVADI, S.; HASHEMY, S. M.; MOHAMMADI, K.; HOWARD, K. W. F.; NESHAT, A. Classification of aquifer vulnerability using K-means cluster analysis. **Journal of hydrology**, v. 549, p. 27-37, 2017.

KORB, V.; ROSA, C. P.; GUERRA, D.; SANTOS, J. S.; SOUZA, E. L. Produtividade de cultivares de cana-de-açúcar de ciclo precoce na região noroeste do rio grande do sul. **Salão do Conhecimento**, v. 2, n. 2, 2016.

KUANAR, S. K.; RANGA, K. B.; CHOWDHURY, A. S. Multi-view video summarization using bipartite matching constrained optimum-path forest clustering. **IEEE Transactions on Multimedia**, v. 17, n. 8, p. 1166-1173, 2015.

LI, C.; SUN, L.; JIA, J.; CAI, Y.; WANG, X. Risk assessment of water pollution sources based on an integrated K-means clustering and set pair analysis method in the region of Shiyang, China. **Science of the Total Environment**, v. 557, p. 307-316, 2016.

LIBBRECHT, M. W.; NOBLE, W. S. Machine learning applications in genetics and genomics. **Nature Reviews Genetics**, v. 16, n. 6, p. 321, 2015.

LIU, G.; YANG, J.; HAO, Y.; ZHANG, Y. Big data-informed energy efficiency assessment of China industry sectors based on K-means clustering. **Journal of Cleaner Production**, v. 183, p. 304-314, 2018.

LUDKE, M.; ANDRÉ, M. E. D. A. **Pesquisa em Educação**: abordagens qualitativas. São Paulo: E.P.U., 1986.

LUGER, G. F. **Inteligência artificial**. 6. ed. São Paulo: Pearson Education do Brasil, 2014.

MAHELA, O. P.; SHAIK, A. G. Power quality recognition in distribution system with solar energy penetration using S-transform and Fuzzy C-means clustering. **Renewable energy**, v. 106, p. 37-51, 2017.

MARSLAND, S. **Machine learning: an algorithmic perspective**. CRC press, 2015.

MARTINS, G. B.; PEREIRA, D. R.; ALMEIDA, J. G. OPFSumm: on the video summarization using Optimum-Path Forest. **Multimedia Tools and Applications**, p. 1-17, 2018.

MASCARENHAS, S. A. **Metodologia científica**. São Paulo: Pearson Brasil, 2012.

MENDONÇA, M. L. O papel da Agricultura nas Relações Internacionais e a Construção do Conceito de Agronegócio. **Contexto Internacional**, v. 37, n. 2, p. 375, 2015.

MIRANDA, L. L.; VASCONCELOS, A. C. M. D.; LANDEL, M. G. A. **Cana de açúcar**. Campinas: Instituto Agrônômico de Campinas, 2010.

MITCHELL, T. M. **Machine learning**. New York: WCB, 1997.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre Aprendizado de Máquina. In: REZENDE, S. O. **Sistemas Inteligentes** - Fundamentos e Aplicações. Barueri: Manole Ltda, 2003. p. 89-114.

MONFORT, J. La recherche des filières de production. *Economie et Documents*. n. 67. INSEE, França, 93p. 1983.

NALDI, M. C.; CAMPELLO, R. J. G. B. Evolutionary K-means for distributed data sets. **Neurocomputing**, v. 127, p. 30-42, 2014.

NUNES, D. H. F. **Um breve estudo sobre o algoritmo K-means**. Coimbra: UC, 2016. 50 f. Dissertação (mestrado) – Programa de Pós-Graduação em Matemática, Universidade de Coimbra, Coimbra, 2016.

NUNES, T. M.; COELHO, A. L. V.; LIMA, C. A. M.; PAPA, J. P.; ALBUQUERQUE, V. H. C. EEG signal classification for epilepsy diagnosis via optimum path forest–A systematic assessment. **Neurocomputing**, v. 136, p. 103-123, 2014.

PAPA, J. P. **Classificação Supervisionada de Padrões Utilizando Florestas de Caminhos Ótimos**. 2008. 58 f. Tese (Doutorado em Ciência da Computação) – Universidade estadual de Campinas. Campinas - SP.

PAPA, J. P.; FERNANDES, S. E. N.; FALCAO, A. X. Optimum-path forest based on K-connectivity: Theory and applications. **Pattern Recognition Letters**, v. 87, p. 117-126, 2017.

PARASTAR, H.; BAZRAFESHAN, A. Fuzzy C-means clustering for chromatographic fingerprints analysis: A gas chromatography–mass spectrometry case study. **Journal of Chromatography A**, v. 1438, p. 236-243, 2016.

PASSADOR, J. L.; ROSA, A. A.; PASSADOR, C. S. A Comercialização na agroindústria de pequeno porte: a agricultura familiar em evidência: o caso de Londrina. In: CONGRESSO BRASILEIRO DE ECONOMIA E SOCIOLOGIA RURAL, 42., 2004, Cuiabá: **Dinâmicas Setoriais e Desenvolvimento Regional**, 2004. Disponível em: <<http://www.sober.org.br/palestra/12/02O094.pdf>>. Acesso em: 28 abr. 2018.

PASSOS JÚNIOR, L. A.; RAMOS, C. C. O.; RODRIGUES, D.; PEREIRA, D. R.; SOUZA, A. N.; COSTA, K. A P.; PAPA, J. P. Unsupervised non-technical losses identification through optimum-path forest. **Electric Power Systems Research**, v. 140, p. 413-423, 2016.

PEREIRA, T. S. **Uso de inteligência artificial para estimativa da capacidade de suporte de carga do solo**. 2017. 179 f. Tese (Doutorado em Engenharia Agrícola) - Universidade Federal de Santa Maria, Santa Maria, 2017.

PIMENTEL, B. A.; SOUZA, R. M. C. R. A multivariate fuzzy c-means method. **Applied Soft Computing**, v. 13, n. 4, p. 1592-1607, 2013.

PIMENTEL, E. P.; FRANÇA, V. F.; OMAR, N. A identificação de grupos de aprendizes no ensino presencial utilizando técnicas de clusterização. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 14., 2003, Rio de Janeiro. **Anais eletrônicos...** Rio de Janeiro: SBC, 2003. Disponível em: <<http://www.nce.ufrj.br/sbie2003/publicacoes/paper52.pdf>>. Acesso em: 02. fev. 2018.

PISANI, R. J.; Nakamura, R. Y. M.; Riedel, P. S.; Zimback, C. R. L. Falcão, A. X.; PAPA, J. P. Toward Satellite-Based Land Cover Classification Through Optimum-Path Forest. **IEEE Trans. Geoscience and Remote Sensing**, v. 52, n. 10, p. 6075-6085, 2014.

RUSSELL, S.; NORVIG, P. **Inteligência Artificial**. 3. ed., Elsevier Brasil, 2013.

SANTOS, R. C.; GALO, M.; TACHIBANA, V. M. CLASSIFICATION OF LIDAR DATA OVER BUILDING ROOFS USING K-MEANS AND PRINCIPAL COMPONENT ANALYSIS. **Boletim de Ciências Geodésicas**, v. 24, n. 1, p. 69-84, 2018.

SCHWAMMLE, V.; JENSEN, O. N. A simple and fast method to determine the parameters for fuzzy c-means cluster analysis. **Bioinformatics**, v. 26, n. 22, p. 2841-2848, 2010.

SHAFEEQ, B. M. A.; HAREESHA, K. S. Dynamic Clustering of Data with Modified K-means Algorithm. **IPCSIT**, Singapore, v. 27, p. 221-225, 2012.

SILVA, A. J.; MONTEIRO, M. S. L.; LIMA, E. B. Difusão do agronegócio no Brasil: estratégias governamentais. **Universidade Federal do Piauí (UFPI)**, p. 47, 2015.

SILVA, A. B. M.; PORTUGAL, M. S.; CECHIN, A. L. Redes neurais artificiais e Análise de Sensibilidade: uma aplicação à demanda de importações brasileira. **REVECAP**, v. 5 n. 4, 2001. Disponível em: <http://www.ufrgs.br/ppge/pcientifica/2000_11.pdf>. Acesso em: 20 mar. 2018.

SILVA, I. S.; SPRITZER, I. M. P. A.; OLIVEIRA, W. P. A importância da Inteligência Artificial e dos Sistemas Especialistas. In: CONGRESSO BRASILEIRO DE ENSINO DE ENGENHARIA, 32. 2004. Brasília. **Dando forma a uma nova realidade**. Brasília: COBENGE, 2004. Disponível em: <http://www.abenge.org.br/CobengeAnteriores/2004/artigos/09_158.pdf>. Acesso em: 08 fev. 2018.

SOUZA, W. A.; LOTUFO R. A.; RITTNER L. **Análise comportamental da Optimum-Path Forest em diferentes funções métricas**. 2012. 6 f. Monografia (Especialização) – Universidade de Campinas, Campinas - SP.

STETCO, A.; ZENG, X.; KEANE, J. Fuzzy C-means++: fuzzy C-means with effective seeding initialization. **Expert Systems with Applications**, v. 42, n. 21, p. 7541-7548, 2015.

TANG, J. R.; ISA, N. A. M.; CH'NG, E. S. A Fuzzy-C-Means-Clustering Approach: Quantifying Chromatin Pattern of Non-Neoplastic Cervical Squamous Cells. **PloS one**, v. 10, n. 11, p. e0142830, 2015.

UNICAMP. Universidade Estadual de Campinas. **Instituto de Computação**. Disponível em: <<http://www.ic.unicamp.br/>>. Acesso em 03 mar. 2018.

VILELA, R. A. G.; LAAT, E.; LUZ, V. G.; SILVA, A. J. N.; TAKAHASHIM. A. C. Pressão por produção e produção de riscos: a “maratona” perigosa do corte manual da cana-de-açúcar. **Revista Brasileira de Saúde Ocupacional**, v. 40, n. 131, 2015. Disponível em: < <http://www.redalyc.org/html/1005/100541506005/>>. Acesso em: 07 jan. 2018.

WANG, Q. WU, B.; STEIN, A.; ZHU, L.; ZENG, Y. Soil depth spatial prediction by fuzzy soil-landscape model. **Journal of Soils and Sediments**, v. 18, n. 3, p. 1041-1051, 2018.

WITTEN, I. H.; FRANK, E. **Data mining: practical machine learning tools and techniques**. 2. ed.: Morgan Kaufmann, 2005.

WIVES, L. K. **Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos**. 2004.136 f. Tese (Doutorado em Computação) - Universidade Federal do Rio Grande do Sul, Porto Alegre – RS.

XUE, M.; ZHOU, L.; KOJIMA, N.; MUCHANGOS L. S.; MACHIMURA T.; TOKAI, A. Application of fuzzy c-means clustering to PRTR chemicals uncovering their release and toxicity characteristics. **Science of The Total Environment**, v. 622, p. 861-868, 2018.

YIN, J.; SUN, H.; YANG, J.; GUO, Q. Comparison of K-means and fuzzy c-Means algorithm performance for automated determination of the arterial input function. **PloS one**, v. 9, n. 2, p. e85884, 2014.

ZARINBAL, M.; ZARANDI, M. H F.; TURKSEN, I. B. Relative entropy fuzzy c-means clustering. **Information Sciences**, v. 260, p. 74-97, 2014.

ZYLBERSZTAJN, D. **Estrutura de governança e coordenação do agribusiness: uma aplicação da nova economia das instituições**. São Paulo, 1995. Tese (Livro Docência em Administração) - Faculdade de Economia, Administração e Contabilidade. Universidade de São Paulo. 1995.

ZUPERKU, E. J.; PRKIC, I.; STUCKE, A. G.; MILLER, J. R.; HOPP, F. A.; STUTH, E. A. Automatic classification of canine PRG neuronal discharge patterns using K-means clustering. **Respiratory physiology & neurobiology**, v. 207, p. 28-39, 2015.

APÊNDICE A – MUNICÍPIOS COM ESTAÇÕES METEOROLÓGICAS.

Municípios com estações meteorológicas		
Aguai	Guapiaçu	Pereiras
Águas de Santa Bárbara	Guará	Peruíbe
Altinópolis	Guaraçaí	Piacatu
Américo de Campos	Guaraci	Piedade
Amparo	Guaratã	Pindamonhangaba
Andradina	Guararapes	Pindorama
Angatuba	Guararema	Piquerobi
Anhembi	Guaratinguetá	Piracaia
Aparecida D'Oeste	Guareí	Piracicaba
Araçatuba	Guataparã	Pirajuí
Araraquara	Iacanga	Pirangi
Araras	Iacri	Pirapozinho
Arco-Íris	Ibirá	Pirassununga
Arealva	Ibitinga	Pitangueiras
Areias	Igarapava	Planalto
Ariranha	Ilhabela	Platina
Artur Nogueira	Inúbia Paulista	Pompéia
Avanhandava	Iracemópolis	Populina
Avaré	Irapuã	Porto Feliz
Bananal	Itaberá	Potirendaba
Barbosa	Itaí	Pradópolis
Bariri	Itapetininga	Presidente Bernardes
Barra do Turvo	Itapeva	Presidente Epitácio
Barretos	Itápolis	Presidente Prudente
Bastos	Itapuí	Quatá
Batatais	Itapura	Queiroz
Bauru	Itatiba	Rancharia
Bebedouro	Itatinga	Regente Feijó
Bernardino de Campos	Itu	Reginópolis
Boa Esperança do Sul	Ituverava	Registro
Bofete	Jaboticabal	Ribeira
Boituva	Jacareí	Ribeirão Bonito
Borá	Jacupiranga	Ribeirão do Sul
Borborema	Jales	Rifaina
Borebi	Jardinópolis	Rio Claro
Botucatu	Jaú	Riolândia
Bragança Paulista	José Bonifácio	Riversul
Brejo Alegre	Jundiaí	Rosana
Brotas	Junqueirópolis	Rubiácea
Buri	Lagoinha	Sandovalina
Buritama	Lavínia	Santa Albertina
Cabrália Paulista	Limeira	Santa Bárbara D'Oeste
Cachoeira Paulista	Lins	Santa Cruz da Esperança
Caconde	Lucélia	Santa Cruz das Palmeiras
Cafelândia	Luiz Antônio	Santa Cruz do Rio Pardo
Caiabu	Macatuba	Santa Fé do Sul

Caiuá	Macaubal	Santa Maria da Serra
Cajati	Macedônia	Santa Rita do Passa Quatro
Campinas	Magda	Santo Anastácio
Campos Novos Paulista	Marabá Paulista	Santo Antônio da Alegria
Cananéia	Maracaí	Santo Antônio de Posse
Cândido Mota	Marília	Santo Antônio do Aracanguá
Capela do Alto	Mirandópolis	São Carlos
Capivari	Mirante do Paranapanema	São João da Boa Vista
Cardoso	Mococa	São João do Pau D'alho
Casa Branca	Monte Alto	São Joaquim da Barra
Cássia dos Coqueiros	Monte Aprazível	São José do Rio Pardo
Castilho	Morro Agudo	São José do Rio Preto
Catanduva	Motuca	São Luiz do Paraitinga
Catiguá	Nantes	São Manuel
Charqueada	Narandiba	São Pedro
Chavantes	Natividade da Serra	São Pedro do Turvo
Clementina	Nova Granada	São Sebastião
Colina	Novo Horizonte	São Sebastião da Gramma
Colômbia	Nuporanga	São Simão
Corumbataí	Olímpia	Sarutaiá
Cosmorama	Orlândia	Serra Negra
Cravinhos	Oscar Bressane	Sertãozinho
Cruzália	Ourinhos	Socorro
Descalvado	Ouro Verde	Sorocaba
Dobrada	Pacaembu	Sud Mennucci
Dumont	Palestina	Tabatinga
Eldorado	Palmeira D'Oeste	Taguaí
Emilianópolis	Palmital	Tambaú
Espírito Santo do Pinhal	Paraguaçu Paulista	Taquarituba
Estiva Gerbi	Paraibuna	Tatuí
Fernando Prestes	Paranapanema	Teodoro Sampaio
Fernandópolis	Paulínia	Tietê
Franca	Paulistânia	Tupã
Gastão Vidigal	Paulo de Faria	Ubarana
Gavião Peixoto	Pederneiras	Ubirajara
General Salgado	Pedra Bela	Urânia
Getulina	Pedregulho	Urupês
Guaíçara	Penápolis	Valparaíso
Guaíra	Pereira Barreto	Votuporanga

Fonte: Elaborado pela autora

APÊNDICE B – MUNICÍPIOS QUE NÃO POSSUEM ESTAÇÕES METEOROLÓGICAS PRÓPRIAS E OS MUNICÍPIOS LIMÍTROFES A ELAS DOS QUAIS FORAM ADOTADOS OS DADOS PLUVIOMÉTRICOS E DE TEMPERATURA.

Municípios sem Estação Meteorológica	Município limítrofe com Estação Meteorológica
Adamantina	Valparaíso
Adolfo	Ubarana
Águas Da Prata	São João Da Boa Vista
Águas De Lindóia	Socorro
Agudos	Bauru
Alambari	Itapetininga
Altair	Guaraci
Alto Alegre	Penápolis
Álvares Florence	Cardoso
Álvares Machado	Presidente Prudente
Americana	Limeira
Américo Brasiliense	Araraquara
Analândia	Descalvado
Anhumas	Presidente Prudente
Araçoiaba da Serra	Sorocaba
Aramina	Ituverava
Arandu	Avaré
Arapeí	Bananal
Areiópolis	Macatuba
Aspásia	Urânia
Assis	Paraguaçu Paulista
Atibaia	Bragança Paulista
Auriflama	Palmeira D'Oeste
Avaí	Bauru
Bady Bassitt	São José do Rio Preto
Balbinos	Paulo de Faria
Bálsamo	Monte Aprazível
Barra Bonita ²	Jaú
Barrinha	Dumont
Bento de Abreu	Guararapes
Bilac	Clementina
Birigui	Buritama
Bocaina	Bariri
Boracéia	Pederneiras
Braúna	Penápolis
Brodowski	Batatais

² Barra Bonita: Foi utilizada a estação meteorológica do município de Jaú apenas nos anos de 2008, 2009 e 2010, pois no período de 2011 até 2017 existem dados de estação meteorológica no próprio município de Barra Bonita.

Buritizal	Igarapava ³
Cabreúva	Jundiáí
Caçapava	São José dos Campos
Cajobi	Olímpia ⁴
Cajuru	Altinópolis
Campina do Monte Alegre	Paranapanema
Cândido Rodrigues	Fernando Prestes
Canitar	Chavantes
Cedral	São José do Rio Preto
Cerqueira César	Avaré
Cerquilha	Tietê
Cesário Lange	Pereiras
Conchal	Araras
Conchas	Bofete
Cordeirópolis	Araras
Coroados	Clementina
Coronel Macedo	Taquarituba
Cosmópolis	Artur Nogueira
Cristais Paulista	Pedregulho
Dirce Reis	Jales
Divinolândia	São José do Rio Pardo
Dois Córregos	Brotas
Dolcinópolis	Jales
Dourado	Boa Esperança do Sul
Dracena	Junqueirópolis
Duartina	Cabralia Paulista
Echaporã	Marília
Elias Fausto	Capivari
Elisiário	Catanduva
Embaúba	Catanduva ⁵
Engenheiro Coelho	Artur Nogueira
Espírito Santo do Turvo	Paulistânia
Estrela do Norte	Narandiba
Estrela D'Oeste	Fernandópolis
Euclides da Cunha Paulista	Teodoro Sampaio
Fartura	Taguaí
Fernão	Ubirajara
Flora Rica	Junqueirópolis
Floreal	Magda
Flórida Paulista	Pacaembu
Florínea	Cândido Mota
Gabriel Monteiro	Guararapes
Gália	Ubirajara

³ Igarapava: Foi utilizada a estação meteorológica do município de Igarapava apenas nos anos de 2008 até 2013, pois no período de 2014 até 2018 existem dados de estação meteorológica no próprio município de Buritizal.

⁴ Olímpia fica localizado a 19 quilômetros de Cajobi, não sendo um município limítrofe. Olímpia foi escolhido, pois os municípios limítrofes a Cajobi também não possuem estação meteorológica.

⁵ Catanduva fica localizado a 22 quilômetros de Embaúba, não sendo um município limítrofe. Catanduva foi escolhido, pois os municípios limítrofes a Embaúba não possuem estação meteorológica.

Glicério	Brejo Alegre
Guaimbê	Getulina
Guarani D'Oeste	Fernandópolis
Guariba	Jaboticabal
Guzolândia	Palmeira D'Oeste
Herculândia	Tupã
Holambra	Santo Antônio de Posse
Hortolândia	Campinas
Iaras	Avaré
Ibaté	São Carlos
Ibirarema	Campos Novos Paulista
Icém	Guaraci
Iepê	Nantes
Igaraçu do Tietê	São Manuel
Igaratá	Jacareí
Ilha Solteira	Itapura
Indaiatuba	Campinas
Indiana	Caiabu
Indiaporã	Macedônia
Ipaussu	Chavantes
Iperó	Boituva
Ipeúna	São Carlos
Ipiquá	São José do Rio Preto
Iporanga	Barra do Turvo
Ipuã	Guará
Irapuru	Junqueirópolis
Itajobi	Catanduva
Itaju	Ibitinga
Itaóca	Ribeira
Itapira	Espírito Santo do Pinhal
Itaporanga	Riversul
Itirapina	Corumbataí
Itirapuã	Franca ⁶
Itobi	São José do Rio Pardo
Itupeva	Campinas
Jaborandi	Barretos
Jaci	José Bonifácio
Jaguariúna	Campinas
Jarinu	Itatiba
Jeriquara	Pedregulho
Joanópolis	Piracaia
João Ramalho	Rancharia
Júlio Mesquita	Marília
Jumirim	Tietê
Laranjal Paulista	Tietê
Leme	Pirassununga

⁶ Franca fica localizado a 22 quilômetros de Itirapuã, não sendo um município limítrofe. Franca foi escolhido, pois os municípios limítrofes a Itirapuã não possuem estação meteorológica.

Lençóis Paulista	Macatuba
Lindóia	Serra Negra
Lourdes	Buritama
Louveira	Jundiáí
Lucianópolis	Ubirajara
Luiziânia	Getulina
Lutécia	Pompéia
Manduri	Águas de Santa Bárbara
Marapoama	Urupês
Mariápolis	Presidente Prudente
Marinópolis	Palmeira D'Oeste
Martinópolis	Nantes
Matão	Dobrada
Mendonça	Irapuã
Meridiano	Fernandópolis
Mesópolis	Santa Albertina
Miguelópolis	Ituverava
Mineiros do Tietê	Jaú
Mira Estrela	Macedônia
Mirassol	São José do Rio Preto
Mirassolândia	Palestina
Mogi Guaçu	Aguai
Mogi Mirim	Santo Antônio de Posse
Mombuca	Capivari
Monções	Macaubal
Monte Alegre do Sul	Serra Negra
Monte Azul Paulista	Bebedouro
Monte Castelo	São João do Pau D'alho
Monte Mor	Campinas
Murutinga do Sul	Andradina
Nazaré Paulista	Piracaia
Neves Paulista	José Bonifácio
Nhandeara	Monte Aprazível
Nipoã	José Bonifácio
Nova Aliança	Potirendaba
Nova Canaã Paulista	Santa Fé do Sul
Nova Castilho	General Salgado
Nova Europa	Gavião Peixoto
Nova Guataporanga	São João do Pau D'Alho
Nova Independência	Andradina
Nova Luzitânia	Gastão Vidigal
Nova Odessa	Santa Bárbara D'Oeste
Novais	Catiguá
Ocaçu	Campos Novos Paulista
Óleo	Santa Cruz do Rio Pardo
Onda Verde	Nova Granada
Oriente	Marília
Orindiúva	Nova Granada

Oswaldo Cruz	Inúbia Paulista
Ouroeste	Populina
Palmares Paulista	Catanduva
Panorama	Presidente Epitácio
Paraíso	Pirangi
Paranapuã	Populina
Parapuã	Rancharia
Pardinho	Bofete
Parquera-Açu	Jacupiranga
Parisi	Votuporanga
Patrocínio Paulista	Franca
Paulicéia	Castilho
Pedranópolis	Cardoso
Pedreira	Amparo
Pedrinhas Paulista	Cruzália
Piraju	Bernardino de Campos
Piratininga	Bauru
Poloni	Monte Aprazível
Pongaí	Reginópolis
Pontal	Pitangueiras
Pontalinda	Jales
Pontes Gestal	Riolândia
Porangaba	Pirassununga
Pracinha	Inúbia Paulista
Pratânia	Botucatu
Presidente Alves	Pirajuí
Presidente Venceslau	Ouro Verde
Promissão	Ubarana
Quadra	Pereiras
Queluz	Areias
Quintana	Pompéia
Rafard	Porto Feliz
Redenção Da Serra	São Luiz do Paraitinga
Restinga	Franca
Ribeirão Corrente	Ituverava
Ribeirão dos Índios	Santo Anastácio
Ribeirão Preto	Guatapará
Rincão	Motuca
Rinópolis	Piacatu
Rio Das Pedras	Santa Bárbara D'Oeste
Roseira	Lagoinha
Rubinéia	Santa Fé do Sul
Sabino	Lins
Sagres	Inúbia Paulista
Sales	Irapuã
Sales Oliveira	Orlândia
Salmourão	Rubiácea
Saltinho	Piracicaba

Salto	Itu
Salto Grande	Ribeirão do Sul
Santa Adélia	Itápolis
Santa Branca	Jacareí
Santa Clara D'Oeste	Santa Fé do Sul
Santa Cruz da Conceição	Pirassununga
Santa Ernestina	Dobrada
Santa Gertrudes	Rio Claro
Santa Lúcia	Araraquara ⁷
Santa Mercedes	São João do Pau D'Alho
Santa Rita D'Oeste	Santa Albertina
Santa Rosa de Viterbo	São Simão
Santa Salete	Urânia
Santana da Ponte Pensa	Palmeira D'Oeste
Santo Expedito	Presidente Prudente
Santópolis do Aguapeí	Clementina
São Francisco	Urânia
São João das Duas Pontes	Fernandópolis ⁸
São João de Iracema	General Salgado
São José da Bela Vista	Nuporanga
São José do Barreiro	Bananal
Sarapuí	Itapetininga
Sebastianópolis do Sul	Monte Aprazível
Serra Azul	São Simão
Serrana	Cravinhos
Severínia	Olímpia
Silveiras	Cachoeira Paulista
Sumaré	Paulínia
Suzanápolis	Pereira Barreto
Tabapuã	Olímpia
Taciba	Regente Feijó
Taiacu	Monte Alto
Taiuva	Jaboticabal
Tanabi	Palestina
Tapiratiba	São José do Rio Pardo
Taquaral	Pitangueiras
Taquaritinga	Jaboticabal
Tarabaí	Pirapozinho
Tarumã	Maracá
Tejupá	Taquarituba
Terra Roxa	Colina
Timburi	Sarutaiá
Torrinha	Santa Maria da Serra

⁷ Araraquara fica localizado a 15 quilômetros de Santa Lúcia, não sendo um município limítrofe. Araraquara foi escolhido, pois os municípios limítrofes a Santa Lúcia não possuem estação meteorológica.

⁸ Fernandópolis fica localizado a 18 quilômetros de São João das Duas Pontes, não sendo um município limítrofe. Fernandópolis foi escolhido, pois os municípios limítrofes a São João das Duas Pontes não possuem estação meteorológica.

Trabiju	Boa Esperança do Sul
Três Fronteiras	Santa Fé do Sul
Tupi Paulista	Junqueirópolis
Turiúba	Macaubal
Turmalina	Populina
Uchoa	Ibirá
União Paulista	Planalto
Uru	Pirajuí
Valentim Gentil	Votuporanga
Vargem	Piracaia
Vargem Grande do Sul	São Sebastião da Gramma
Viradouro	Pitangueiras
Vista Alegre do Alto	Pirangi
Vitória Brasil	Jales
Zacarias	Planalto

APÊNDICE C – QUESTIONÁRIO: INFLUÊNCIAS DOS FATORES EDAFOCLIMÁTICOS NA PRODUTIVIDADE DA CANA DE AÇÚCAR NO ESTADO DE SÃO PAULO

1. Levando em consideração os tipos de solo do estado de São Paulo, segundo o Instituto Agronômico (IAC): Latossolos, Nitossolos, Cambissolos, Gleissolos, Espodossolos, Organossolos, Neossolos, Argissolo, Planossolo. Quais deles são os mais indicados para o plantio da cana de açúcar e quais os menos indicados?
2. Índices pluviométricos possuem influência na produtividade da cana de açúcar? Para atingir alta produtividade, o adequado são índices pluviométricos mais altos ou regiões menos chuvosas tendem a ser mais produtivas?
3. A temperatura possui influência na produtividade da cana de açúcar? Para atingir alta produtividade, o adequado são temperaturas mais altas ou mais baixas?
4. No ano de 2017, o município Morro Agudo foi o maior produtor atingindo aproximadamente 7.900 mil toneladas, Guaíra ocupou o segundo lugar com 5.700 mil toneladas e Jaboticabal foi o terceiro com 5.179 mil toneladas. Quais são as principais características que fazem com que esses municípios sejam grandes produtores de cana de açúcar?
5. Alguns municípios, segundo o Instituto de Economia Agrícola (IEA), se destacam por atingir produtividade acima de 100 toneladas por hectares, como, por exemplo, Boituva (100 *t/ha*), Bragança Paulista (100 *t/ha*), Guaratinguetá (100 *t/ha*), São José do Rio Preto (100 *t/ha*), Cordeirópolis (110 *t/ha*), Iaras (110 *t/ha*), Jaci (110 *t/ha*), Itupeva (120 *t/ha*), Vargem Grande do Sul (120 *t/ha*). O que leva esses municípios a atingirem níveis de produtividade altos? Os fatores edafoclimáticos possuem influência na produtividade desses municípios?

APÊNDICE D – MATRIZ DE PERTINENCIA DOS MUNICIPIOS QUE ESTAO EM REGIAO DE FRONTEIRA DE ACORDO COM O ALGORITMO FCM.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Municípios
0,4223	0,3856	0,0865	0,0469	0,0331	0,0255	BATATAIS
0,0179	0,2920	0,5473	0,0804	0,0385	0,0239	ITAJOBI
0,0212	0,5000	0,3512	0,0695	0,0354	0,0227	JARDINÓPOLIS
0,0193	0,3445	0,4937	0,0795	0,0388	0,0243	JOSÉ BONIFÁCIO
0,0209	0,5750	0,2892	0,0619	0,0322	0,0208	SÃO MANUEL
0,0209	0,5765	0,2880	0,0618	0,0321	0,0208	TAQUARITINGA
0,0050	0,0283	0,5865	0,2942	0,0591	0,0269	BOTUCATU
0,0048	0,0248	0,3243	0,5373	0,0766	0,0322	CARDOSO
0,0051	0,0271	0,3994	0,4623	0,0741	0,0320	GETULINA
0,0052	0,0276	0,4246	0,4382	0,0728	0,0317	IEPÊ
0,0052	0,0286	0,5261	0,3459	0,0651	0,0292	LAVÍNIA
0,0051	0,0271	0,4022	0,4596	0,0740	0,0320	MATÃO
0,0050	0,0262	0,3670	0,4940	0,0755	0,0322	MOCOCA
0,0050	0,0260	0,3589	0,5022	0,0758	0,0323	MONTE APRAZÍVEL
0,0050	0,0260	0,3583	0,5027	0,0758	0,0323	PATROCÍNIO PAULISTA
0,0050	0,0262	0,3670	0,4940	0,0755	0,0322	SANTA RITA DO PASSA QUATRO
0,0052	0,0276	0,4263	0,4366	0,0727	0,0316	SANTO ANTÔNIO DO ARACANGUÁ
0,0049	0,0258	0,3530	0,5081	0,0760	0,0323	TABAPUÃ
0,0018	0,0066	0,0273	0,3178	0,5838	0,0627	ÁLVARES FLORENCE
0,0018	0,0064	0,0265	0,3005	0,6023	0,0625	AMÉRICO BRASILIENSE
0,0019	0,0069	0,0293	0,3657	0,5338	0,0624	ANHEMBI
0,0020	0,0073	0,0315	0,4476	0,4517	0,0599	ARAMINA
0,0020	0,0073	0,0324	0,5206	0,3819	0,0557	AREALVA
0,0020	0,0073	0,0325	0,5548	0,3502	0,0533	AVARÉ
0,0019	0,0068	0,0284	0,3429	0,5574	0,0627	BARRA BONITA
0,0018	0,0066	0,0275	0,3211	0,5804	0,0627	BIRIGUI
0,0020	0,0073	0,0323	0,5063	0,3953	0,0567	BURITIZAL
0,0020	0,0073	0,0325	0,5343	0,3691	0,0548	DOBRADA
0,0019	0,0071	0,0301	0,3905	0,5085	0,0619	ELIAS FAUSTO
0,0020	0,0071	0,0304	0,4005	0,4984	0,0616	FLORÍNEA
0,0019	0,0072	0,0323	0,5838	0,3238	0,0510	GLICÉRIO
0,0019	0,0070	0,0298	0,3811	0,5180	0,0621	IRAPUÃ
0,0020	0,0072	0,0306	0,4093	0,4896	0,0614	ITAPURA
0,0020	0,0073	0,0325	0,5254	0,3775	0,0554	JACI
0,0020	0,0073	0,0323	0,4978	0,4034	0,0572	MINEIROS DO TIETÊ

0,0018	0,0064	0,0265	0,3005	0,6023	0,0625	NOVA INDEPENDÊNCIA
0,0020	0,0073	0,0323	0,4966	0,4045	0,0573	PAULICÉIA
0,0020	0,0073	0,0319	0,4706	0,4294	0,0587	PIRANGI
0,0019	0,0071	0,0301	0,3905	0,5085	0,0619	PIRAPOZINHO
0,0020	0,0073	0,0324	0,5190	0,3834	0,0558	POPULINA
0,0020	0,0072	0,0309	0,4182	0,4807	0,0611	PRADÓPOLIS
0,0019	0,0071	0,0301	0,3906	0,5084	0,0619	SÃO PEDRO
0,0019	0,0068	0,0288	0,3535	0,5464	0,0626	TATUÍ
0,0020	0,0072	0,0310	0,4248	0,4741	0,0608	VARGEM GRANDE DO SUL
0,0012	0,0035	0,0104	0,0386	0,3571	0,5892	ARTUR NOGUEIRA
0,0012	0,0037	0,0111	0,0445	0,5924	0,3471	BALBINOS
0,0012	0,0036	0,0110	0,0445	0,5979	0,3417	BILAC
0,0012	0,0036	0,0107	0,0402	0,3867	0,5576	BORÁ
0,0012	0,0038	0,0113	0,0445	0,5385	0,4007	BURITAMA
0,0012	0,0037	0,0108	0,0406	0,3960	0,5478	CAIUÁ
0,0012	0,0036	0,0105	0,0394	0,3715	0,5738	CANITAR
0,0012	0,0037	0,0111	0,0423	0,4389	0,5028	CRUZÁLIA
0,0012	0,0036	0,0109	0,0443	0,6122	0,3278	JERQUARA
0,0012	0,0036	0,0107	0,0404	0,3917	0,5523	MANDURI
0,0012	0,0037	0,0111	0,0446	0,5830	0,3564	MIRASSOLÂNDIA
0,0012	0,0037	0,0111	0,0424	0,4406	0,5009	MONÇÊS
0,0012	0,0036	0,0104	0,0388	0,3600	0,5861	MURUTINGA DO SUL
0,0012	0,0036	0,0104	0,0388	0,3600	0,5860	NOVA CASTILHO
0,0012	0,0037	0,0110	0,0421	0,4333	0,5086	ORIENTE
0,0012	0,0036	0,0104	0,0390	0,3636	0,5822	OSVALDO CRUZ
0,0011	0,0034	0,0100	0,0368	0,3271	0,6215	PRESIDENTE VENCESLAU
0,0012	0,0038	0,0113	0,0441	0,5066	0,4330	SAGRES
0,0012	0,0035	0,0102	0,0381	0,3480	0,5990	SANTA CRUZ DA CONCEIÇÃO
0,0012	0,0037	0,0110	0,0421	0,4333	0,5086	SANTO ANTÔNIO DA ALEGRIA
0,0012	0,0036	0,0106	0,0399	0,3816	0,5630	SANTO ANTÔNIO DE POSSE
0,0011	0,0034	0,0098	0,0360	0,3138	0,6360	TURMALINA
0,0012	0,0038	0,0113	0,0439	0,4934	0,4465	VISTA ALEGRE DO ALTO