UNIVERSIDADE DO EXTREMO SUL CATARINENSE - UNESC
CURSO DE CIÊNCIA DA COMPUTAÇÃO

DANIEL NUNES PACHECO

ABORDAGEM DOS ALGORITMOS DE AGRUPAMENTO K-MEANS E FUZZY C-MEANS NA IDENTIFICAÇÃO DE ZONAS PLUVIOMÉTRICAS EM SANTA CATARINA UTILIZANDO O MODELO DE PROCESSO CRISP-DM

CRICIÚMA
2018

DANIEL NUNES PACHECO

ABORDAGEM DOS ALGORITMOS DE AGRUPAMENTO K-MEANS E FUZZY C-MEANS NA IDENTIFICAÇÃO DE ZONAS PLUVIOMÉTRICAS EM SANTA CATARINA UTILIZANDO O MODELO DE PROCESSO CRISP-DM

Trabalho de Conclusão de Curso, apresentado para obtenção de grau de Bacharel no curso de Ciência da Computação da Universidade do Extremo Sul Catarinense, UNESC.

Orientadora: Prof.ª Dra. Merisandra Côrtes de Mattos Garcia

CRICIÚMA

2018

DANIEL NUNES PACHECO

ABORDAGEM DOS ALGORITMOS DE AGRUPAMENTO K-MEANS E FUZZY C-MEANS NA IDENTIFICAÇÃO DE ZONAS PLUVIOMÉTRICAS EM SANTA CATARINA UTILIZANDO O MODELO DE PROCESSO CRISP-DM

Trabalho de Conclusão de Curso aprovado pela Banca Examinadora para obtenção do Grau de Bacharel, no Curso de Ciência da Computação da Universidade do Extremo Sul Catarinense, UNESC, com Linha de Pesquisa em Inteligência Computacional

Criciúma, 30 de novembro de 2018.

BANCA EXAMINADORA

Prof.ª Merisandra Côrtes de Mattos Garcia - Doutora - (UNESC) - Orientadora

Prof. Kristian Madeira - Doutor - (UNESC)

Prof. Cristian Cechinel - Doutor - (UFSC)

RESUMO

A precipitação pluviométrica influencia diretamente na economia do Estado de Santa Catarina, devido ao destaque na agricultura catarinense. Visto que a Organização Meteorológica Mundial recomenda a utilização de trinta anos de dados históricos para encontrar padrões que não tenham interferências com eventos adversos que ocorreram na região se faz necessária a utilização de técnicas computacionais. Existem alguns métodos disponíveis que visam auxiliar na análise desses dados, um deles é conhecido como *CRoss-Industry Standart Process for Data Mining,* possuindo seis etapas em seu processo. Dentre essas etapas, se encontra a de *data mining*, caracterizada pela extração de informações implícitas e potencialmente úteis contidas em um conjunto de dados. O agrupamento, uma das tarefas de *data mining*, tem como objetivo identificar objetos semelhantes entre si por meio da aplicação de técnicas para descrever os dados. A fim de avaliar o agrupamento gerado, são adotadas medidas de qualidade, que consistem em índices estatísticos, responsáveis por verificar os resultados obtidos. Esta pesquisa tem como objetivo comparar por meio de medidas de qualidade os algoritmos de agrupamento *K-means* e *Fuzzy C-means* na identificação de zonas pluviométricas homogêneas no estado de Santa Catarina. Compreendendo as etapas de entendimento e preparação dos dados, *data mining*, avaliação e discussão dos resultados obtidos. Para definir a quantidade de *clusters* foram avaliados os índices de qualidade *Xie and Beni*, Coeficiente de Participação e Coeficiente de Entropia para os resultados do *Fuzzy C-means*, e no algoritmo *K-means* foram utilizados os índices *Sum of Squared Errors* e *Prior Probability*. Ambos os algoritmos apresentaram melhores resultados ao se utilizar três *clusters*, no entanto o *Fuzzy C-means* se demonstrou melhor na espacialização dos objetos dentro dos *clusters* e nas médias pluviométricas (1.448,18mm, 1.364,10mm, 1.665,05mm), apresentando resultados mais homogêneos, enquanto o *K-means* dividiu as estações de forma mais abrupta apresentando uma média de precipitação crescente (1.040,71mm, 1.553,85mm, 1.837,37mm), identificando *clusters* heterogêneos. Mediante este estudo, conclui-se que o algoritmo *Fuzzy C-means* apresenta melhores resultados na identificação de zona pluviometricamente homogêneas no Estado de Santa Catarina.

Palavras-chave: *Big data. Data mining. K-means. Fuzzy C-means.* CRISP-DM.

ABSTRACT

Precipitation directly influences the economy of the State of Santa Catarina, due to the prominence of Santa Catarina agriculture. Since the World Meteorological Organization recommends using thirty years of historical data to find patterns that do not interfere with adverse events that occurred in the region requires the use of computational techniques. There are some methods available that help to analyze this data, one of them is known as CRoss-Industry Standart Process for Data Mining, having six steps in its process. Among these steps is the data mining step, characterized by the extraction of implicit and potentially useful information contained in a data set. The grouping, one of the tasks of data mining, aims to identify objects similar to each other through the application of techniques to describe the data. In order to evaluate the grouping, quality measures are adopted, consisting of statistical indices, responsible for verifying the results obtained. This research aims to compare the K-means and Fuzzy C-means clustering algorithms in the identification of homogeneous rainfall zones in the state of Santa Catarina. Understanding the steps of understanding and preparing the data, data mining, evaluation and discussion of the results obtained. In order to define the number of clusters, we evaluated the Xie and Beni quality indices, the Participation Coefficient and the Entropy Coefficient for the Fuzzy C-means results, and the Sum-Squared Errors and Prior Probability indices were used in the K-means algorithm. Both algorithms presented better results when using three clusters, however, the Fuzzy C-means was better demonstrated in the spatialization of the objects within the clusters and in the pluviometric means (1,448,18mm, 1,364,10mm, 1,665,05mm), presenting more results homogeneous, whereas K-means divided the stations more abruptly, presenting a mean of increasing precipitation (1,040,71mm, 1,553,85mm, 1,837,37mm), identifying heterogeneous clusters. This study concludes that the Fuzzy C-means algorithm presents better results in the identification of pluviometrically homogeneous zones in the State of Santa Catarina.

Key words: Big data. Data mining. K-means. Fuzzy C-means. CRISP-DM.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ADI	*Alternative Dunn Index*
AGNES	*Agglomerative Nesting*
ANA	Agência Nacional das Águas
CPTEC	Centro de Previsão do Tempo e Estudos Climáticos
CURE	*Clustering Using Representatives*
CRISP-DM	*CRoss-Industry Standart Process for Data Mining*
DBSCAN	*Density Based Clustering Method Based on Connected Regions with Sufficiently High Density*
DENCLUE	*Density-based Clustwring*
DI	*Dunn Index*
DIANA	*Divisive Analysis*
DMI	*National Meteorology Works*
EM	*Expectation Maximization*
INPE	Instituto Nacional de Pesquisas Espaciais
KDD	*Knowledge Discovery in Databases*
NASA	*National Aeronautics and Space Administration*
OPTICS	*Ordering Points to Identify the Clustering Structure*
OMM	Organização Meteorológica Mundial
PI	*Partition Index*
SAS	*Statistical Analysis System*
SEE	*Sum of Squared Errors*
SEMMA	*Sample, Explore, Modify, Model, Assess*
STING	*Statistical Information Grid*
WaveCluster	*Clustering Using Wavelet Transformation*
XB	*Xie and Beni Index*
ZCAS	Zona de convergência do Atlântico Sul

SUMÁRIO

1 INTRODUÇÃO

Com o avanço da tecnologia de coleta e armazenamento de dados as organizações acumulam uma vasta quantidade de dados, porém extrair informações destes torna-se uma tarefa desafiadora, pois não é possível utilizar ferramentas e técnicas tradicionais de análise de dados. Isso se deve ao tamanho do conjunto de dados ser muito grande ou a natureza incomum dos elementos que é presente até mesmo em conjuntos relativamente pequenos (TAN; STEINBACH; KUMAR, 2009).

Existem alguns métodos disponíveis que visam auxiliar na análise desses dados, um deles é conhecido como *Knowledge Discovery in Databases* (KDD), tendo como etapa principal o *data mining* que utiliza a aplicação de algoritmos específicos para a extração de padrões nos dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, tradução nossa).

O *data mining* fornece um processo estruturado para descobrir e quantificar padrões escondidos em uma grande quantidade de dados, desempenhando um papel fundamental na tomada de decisão dentro de uma organização (COX, 2005, tradução nossa). Combinando técnicas tradicionais de análise dos dados com algoritmos sofisticados para processar os dados gerados pelas organizações (TAN; STEINBACH; KUMAR, 2009). Dentre as tarefas tradicionais de *data mining* existentes tem-se: caracterização; descrição; análise de associação; classificação e agrupamento (HAN; KAMBER, 2001, tradução nossa).

O agrupamento é uma importante tarefa para o *data mining*, já que ele é capaz de identificar regularidades ou tendências utilizando-se de algoritmos semi-supervisionados ou não supervisionados. O agrupamento consiste em um processo de divisão de conjuntos de objetos em grupos onde cada um representa uma subpopulação (BASU; DAVIDSON; WAGSTAFF, 2008, tradução nossa). Dentre os diferentes tipos de agrupamento tem-se, por exemplo, o particional e o hierárquico que se utilizam da lógica clássica, onde cada objeto só pode assumir um valor, ou seja, ele pertence ou não ao grupo. No entanto, separar os dados de uma forma abrupta dependendo da ocasião, pode ser um problema visto que existem situações onde a transição deve se dar de forma gradual entre os conjuntos (JENSEN; SHEN, 2008, tradução nossa). Nesses casos é utilizado o tipo difuso, que se baseia na lógica *fuzzy*, onde um objeto possui um grau de pertinência a cada grupo, podendo pertencer simultaneamente a mais de um grupo.

Nos projetos de *data mining*, procurando criar um modelo de processo que abrangesse as diferentes áreas de conhecimento, dando a certeza que o projeto possua todas as etapas necessárias e que não seja proprietário, Chapman et al (2000, tradução nossa) criaram em 1996 o *CRoss-Industry Standart Process for Data Mining* (CRISP-DM), que padroniza o ciclo de vida de um projeto de data mining.

No cenário de Aquecimento Global, fenômeno que é caracterizado por mudanças climáticas que afetarão a disponibilidade da água, segurança alimentar, infraestrutura e rendimentos agrícolas, é fundamental o conhecimento do comportamento da precipitação pluviométrica para um planejamento adequado (IPCC AR 4, 2007, 2015, tradução nossa). Segundo Monteiro (2001) o estado de Santa Catarina possui estações bem definidas e suas precipitações são bem distribuídas durante o ano, não havendo uma estação seca.

A precipitação pluviométrica ou chuva é o meio ao qual a água condensada na atmosfera atinge gravitacionalmente o solo. Back, Bonneti e Coan (2014) classificam como um elemento meteorológico que exerce grande influência em quase todas as atividades desenvolvidas em campo, tanto pela ocorrência em excesso ou por sua escassez. Neste sentido a Organização Meteorológica Mundial (OMM, 1989, 2007, tradução nossa) recomenda a utilização de uma base com 30 anos de informações meteorológicas para encontrar padrões que não tenham interferências com eventos adversos que ocorreram na região.

Nesta área, algumas pesquisas já foram realizadas, como por exemplo, os estudos de André et al (2008); Araújo (2013); Boschi, Oliveira e Assad (2011); Boschi, Oliveira e Avila (2009); Dourado (2013) e Dourado, Oliveira e Avila (2012, 2013) que utilizaram a tarefa de agrupamento em dados de precipitações com base no algoritmo *K-means*, a fim de identificar zonas homogêneas dentro de estados brasileiros.

Considerando-se o exposto e que vários dos estudos na área utilizam o *K-means* como método para identificação das zonas pluviométricas, a pesquisa aqui proposta visa a aplicação do *data mining*, segundo o modelo de processo CRISP-DM, por meio da tarefa de agrupamento e dos algoritmos de particionamento (*K-means*) e de lógica *fuzzy* (*Fuzzy C-means*) em dados históricos referentes a um período de 30 anos de precipitações pluviométricas no estado de Santa Catarina. Assim, busca-se identificar por meio de medidas de qualidade em data mining qual destes algoritmos apresenta melhores resultados no que se refere ao domínio de aplicação desta pesquisa aqui proposta.

1.1 OBJETIVO GERAL

Comparar por meio de medidas de qualidade os algoritmos de agrupamento *K-means* e *Fuzzy C-means* na identificação de zonas pluviométricas homogêneas no estado de Santa Catarina.

1.2 OBJETIVO ESPECÍFICO

Os objetivos específicos desta pesquisa são:

a) compreender o conceito de *data mining*, agrupamento, algoritmos *K-means* e *Fuzzy C-means*, medidas de qualidade e CRISP-DM;
b) aplicar o modelo de processo CRISP-DM;
c) aplicar os algoritmos de agrupamento *K-means* e *Fuzzy C-means* em dados de precipitações pluviométricas em Santa Catarina;
d) avaliar por meio de medidas de qualidades os modelos de agrupamento gerados em dados de precipitações pluviométricas;

1.3 JUSTIFICATIVA

Data mining é um processo de descoberta automática de informação em grandes repositórios de dados, empregado nas mais diferentes áreas do conhecimento (TAN; STEINBACH; KUMAR, 2009). Informações estas que podem influenciar a tomada de decisão, antes realizada baseando-se na intuição, por não existirem ferramentas para extrai-las de uma grande base de dados (HAN; KAMBER, 2001, tradução nossa). Segundo Basu, Davidson e Wagstaff (2008, tradução nossa) e Goldschmidt e Passos (2005) o *data mining* pode confirmar ou identificar padrões antes desprezados na área de meteorologia, por exemplo, dentre outras áreas do conhecimento.

O processo de *data mining* possui várias tarefas que podem ser utilizadas, dentre elas a de agrupamento que busca reunir elementos que tenham alguma similaridade entre si formando grupos para serem analisados (TAN; STEINBACH; KUMAR, 2009). Esta tarefa fornece informações úteis sobre os padrões presentes nos dados (CICHOSZ, 2015, tradução nossa), sendo utilizadas amplamente em várias

aplicações, como processamento de imagem, marketing, análise de dados, biologia, entre outros (HAN; KAMBER, 2001, tradução nossa).

A fim de avaliar o agrupamento gerado, são adotadas medidas de qualidade, que consistem em índices estatísticos, responsáveis por verificar os resultados obtidos (JAIN; DUBES, 1988, tradução nossa). Segundo Hamilton e Geng (2007), independente do padrão a ser extraído, essas medidas são importantes, pois podem reduzir tempo e custo do *data mining*.

Com o progresso do *data mining* surgiram alguns modelos de processo que visam orientar a sua implementação, tais como *Sample, Explore, Modify, Model, Assess* (SEMMA), KDD e CRISP-DM (AZEVEDO; SANTOS, 2008, tradução nossa).

O modelo de processo CRISP-DM empregado nesta pesquisa é amplamente adotado em projetos de *data mining*, possuindo uma taxa de utilização de 43%, tanto academicamente quanto na indústria, pois contém uma boa descrição das etapas a serem seguidas (KDnuggets, 2014, tradução nossa).

Desta forma, a aplicação do *data mining* segundo um modelo de processo aparece como uma alternativa para a análise de grandes volumes de dados, como os gerados pelas análises meteorológicas de uma região.

O estado de Santa Catarina é caracterizado por não possuir nenhuma estação seca (clima mesotérmico úmido), graças aos sistemas metrológicos atuantes no estado que são as massas de ar frias, os vórtices ciclônicos[1], os cavados de níveis médios[2], a convecção tropical[3], a Zona de convergência do Atlântico Sul [4](ZCAS) e a circulação marítima[5] junto ao seu relevo (MONTEIRO, 2001), possuindo uma boa distribuição de chuvas dentro de seu território.

O conhecimento dessa importante variável meteorológica dentro do estado pode auxiliar na tomada de decisões de setores que dependem dela, visto que em

---

1 Vórtice ciclônico é um sistema atmosférico de baixa pressão, com ventos associados que giram no sentido horário no hemisfério Sul (Fonte: Centro de Previsão do Tempo e Estudos Climáticos (CPTEC) /Instituto Nacional de Pesquisas Espaciais (INPE), 2018. Disponível em: <https://www.cptec.inpe.br/>. Acesso em: 24/06/218.)

2 Região da atmosfera em que a pressão é baixa. (Fonte: CPTEC/INPE, 2018. Disponível em: <https://www.cptec.inpe.br/>. Acesso em: 24/06/218.)

3 Movimentos internos organizados dentro de uma camada de ar, produzindo o transporte vertical de calor. (Fonte: CPTEC/INPE, 2018. Disponível em: <https://www.cptec.inpe.br/>. Acesso em: 24/06/218.)

4 Faixa de nebulosidade persistente que se estende do Atlântico Sul Central ao sul da Amazônia (Fonte: CPTEC/INPE, 2018. Disponível em: <https://www.cptec.inpe.br/>. Acesso em: 24/06/218)

5 Movimentos de grandes massas de água dentro de um oceano ou mar (Fonte: CPTEC/INPE, 2018. Disponível em: <https://www.cptec.inpe.br/>. Acesso em: 24/06/218.)

alguns casos ela é essencial para o negócio, como na cultura agrícola (DOURADO; OLIVEIRA; AVILA, 2013).

Dentre os projetos que abordaram a constante da precipitação por meio do *data mining* tem-se a pesquisa de André et al (2008) que identificaram as regiões pluviometricamente homogêneas no estado do Rio de Janeiro com dados mensais; Boschi, Oliveira e Assad (2011) que analisaram a precipitação de uma década no Rio Grande do Sul; Dourado, Oliveira e Ávila (2012, 2013) identificaram zonas pluviometricamente homogêneas de precipitação do estado da Bahia; Araújo (2013) identificou zonas homogêneas de precipitações no Rio Grande do Norte; Mello e Leite (2017) realizaram a caracterização da precipitação mensal, sazonal e anual paro o estado do Paraná em períodos secos, normais e chuvosos. Estes estudos citados têm em comum o uso do algoritmo *K-means* para análise dos dados.

Visto que vários estudos abordaram o *K-means*, um algoritmo baseado na abordagem da lógica clássica, na obtenção dos resultados, tem-se uma possibilidade para a análise de outro algoritmo, *Fuzzy C-means*, que poderia ter melhores resultados utilizando os dados das precipitações, pois não é uma abordagem tradicional, sendo mais subjetiva, portanto, geralmente mais parecida com a forma do pensar humano.

1.4 ESTRUTURA DO TRABALHO

Esta pesquisa é composta por 5 capítulos, o Capítulo 1 apresenta o contexto do tema proposto, objetivos e justificativa para realização deste trabalho.

No Capítulo 2 são abordados os conceitos das metodologias de processo para descoberta de conhecimento, *data mining,* tarefas de agrupamento e medidas de qualidade com o propósito de auxiliar no entendimento da pesquisa proposta.

Alguns exemplos de estudos que se utilizam da aplicação dos algoritmos *K-means* e *Fuzzy C-means* são apresentados no Capítulo 2.

No Capítulo 4 é descrito o trabalho desenvolvido, os resultados obtidos com a aplicação das técnicas propostas e a discussão dos resultados.

Por fim, tem-se a conclusão desta pesquisa e algumas sugestões para trabalhos futuros.

2 DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS

Petabytes de dados versam nas redes de computadores e dispositivos de armazenamento de dados todos os dias, dados referentes a quase todos os aspectos possíveis como por exemplo clima, economia e biologia. Resultado da informatização da sociedade e de ferramentas de coleta e armazenamentos de dados em todo o mundo. (HAN; KAMBER; PEI, 2011, tradução nossa).

A existência de dados que envolvem computadores, redes e vidas influenciam agências governamentais, instituições científicas e empresas a dedicam recursos para coletar e armazenar dados. (KANTARDZIK, 2011, tradução nossa).

Para se ter uma ideia do volume dos dados que estamos envolvidos os atuais satélites de observação da Terra da *National Aeronautics and Space Administration* (NASA) geram um terabytes de dados por dia (BRAMMER, 2013, tradução nossa).

Com essa quantidade de dados descobrir informações valiosas e transformá-las em conhecimento organizado virou uma necessidade (HAN; KAMBER; PEI, 2011, tradução nossa).

Extrair conhecimento útil escondido nesses conjuntos dados, complexos e ricos em informações, e atuar sobre esse conhecimento está se tornando cada vez mais comum a quase todos os campos de negócios, ciência e engenharia. (KANTARDZIK, 2011, tradução nossa).

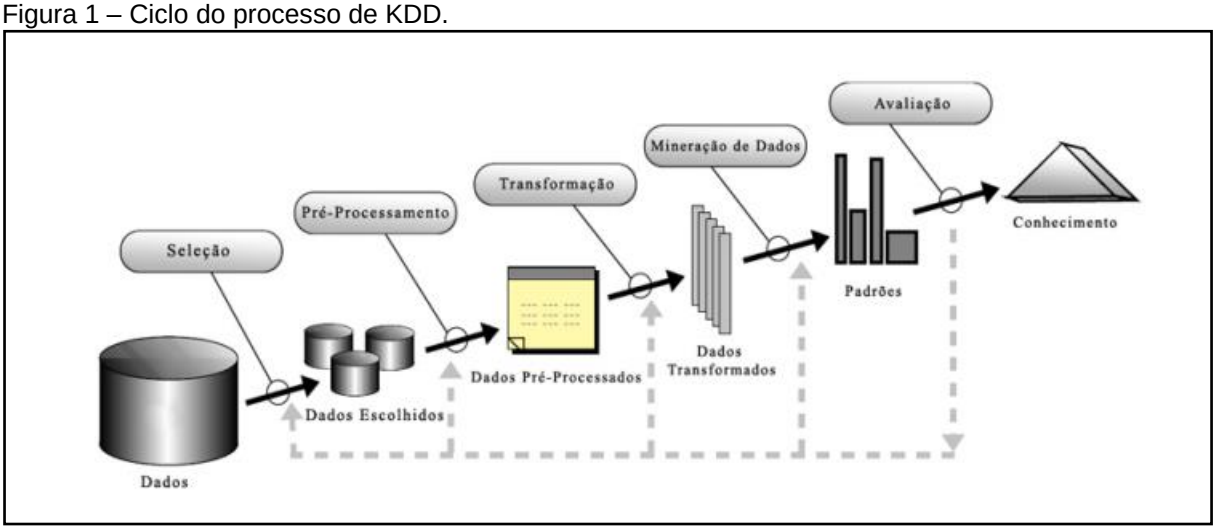
Analisar esses dados manualmente acaba se tornando impraticável em muitos domínios, portanto o trabalho precisa ser automatizado (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, tradução nossa).

O processo de aplicação de uma metodologia para descobrir conhecimento nos dados, por meio de tarefas focadas na análise desses dados levou ao nascimento do *data mining* (FAYYAD; PIATETSKY-SHAPIRO; SMYTH 1996, tradução nossa; HAN; KAMBER; PEI, 2011, tradução nossa; KANTARDZIK, 2011, tradução nossa; YE, 2013, tradução nossa;).

Existem alguns métodos disponíveis que visam auxiliar na análise desses dados, e um deles é conhecido como KDD, possuindo as etapas (figura 1): Seleção, se resume em criar um conjunto de dados de destino no qual a descoberta deve ser realizada; pré-processamento, consiste na limpeza dos dados de destino, com o propósito de obter dados consistentes; transformação, compõe-se pela transformação

dos dados usando redução de dimensionalidade ou métodos de transformação; *data mining*, busca por padrões de interesse da aplicação; Interpretação e avaliação, Interpreta e avalia os padrões encontrados (AZEVEDO; SANTOS, 2008, tradução nossa; FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, tradução nossa).

Figura 1 – Ciclo do processo de KDD.



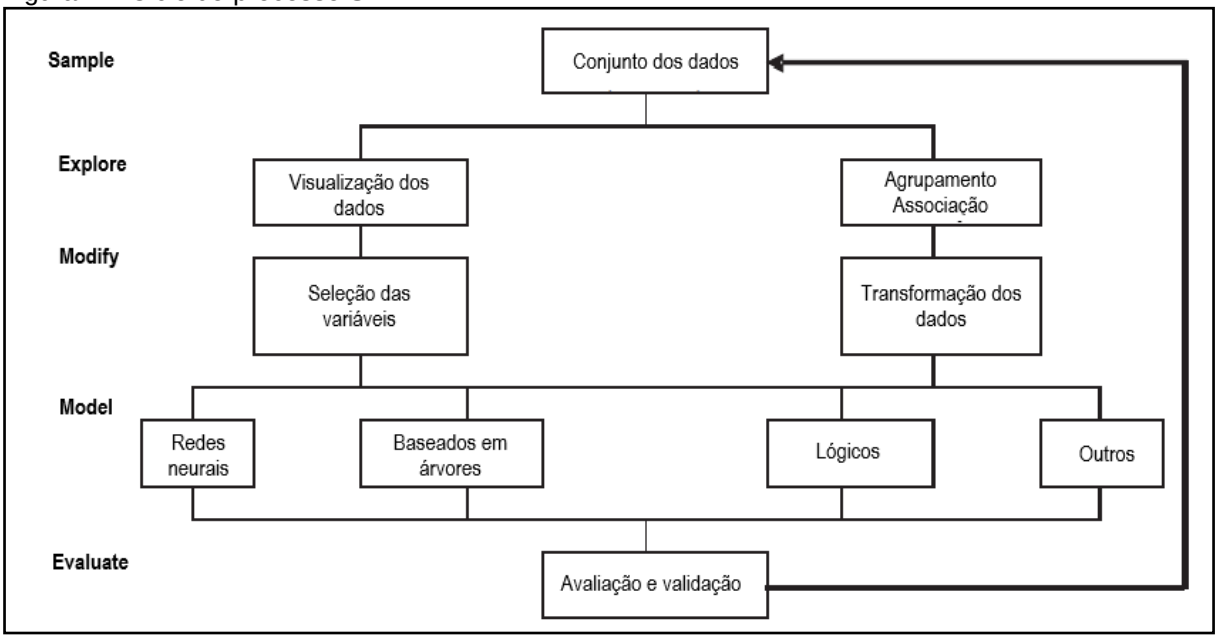
Fonte: Fayyad, Piatetsky-shapiro e Smyth (1996, tradução nossa).

Outra metodologia para efetuar a descoberta de conhecimento é conhecida como SEMMA criado pela *Statistical Analysis System* (SAS) tendo as seguintes 5 etapas.

a) *Sample*: dedica-se na amostragem dos dados, extraindo de um conjunto de dados o suficiente para conter as informações relevantes;
b) *Explore*: compreende a exploração dos dados na busca pelo entendimento dos dados;
c) *Modify*: seleciona e transforma os dados em variáveis para serem utilizados na próxima etapa;
d) *Model:* modela os dados para que o software efetue uma busca automática com uma combinação dos dados procurando encontrar com segurança o resultado desejado;
e) *Assess*: efetua a avaliação dos resultados obtidos verificando a utilidade e confiabilidade.

Na figura 2 é possível visualizar o fluxo da metodologia (AZEVEDO; SANTOS, 2008, tradução nossa, SAS *Institute*, 2017, tradução nossa).

Figura 2 – Ciclo do processo SEMMA.



Fonte: adaptado de SAS Institute (2017, tradução nossa).

Uma das metodologias mais utilizadas segundo KDnuggets (2014, tradução nossa) para descoberta de conhecimento em base de dados, o CRISP-DM, composto por 6 etapas.

Entendimento de negócios, a primeira etapa concentra-se em entender os objetivos e requisitos do projeto, convertendo esse conhecimento em uma definição de problema e um plano preliminar projetado para alcançar o objetivo.

Entendimento de dados, constitui-se na compreensão e coleta dos dados, identificação de problemas com a qualidade dos dados, detecção de subconjuntos interessantes para formar hipóteses para informações ocultas.

Preparação de dados, abrange a construção de um conjunto de dados final com os dados obtidos na etapa inicial.

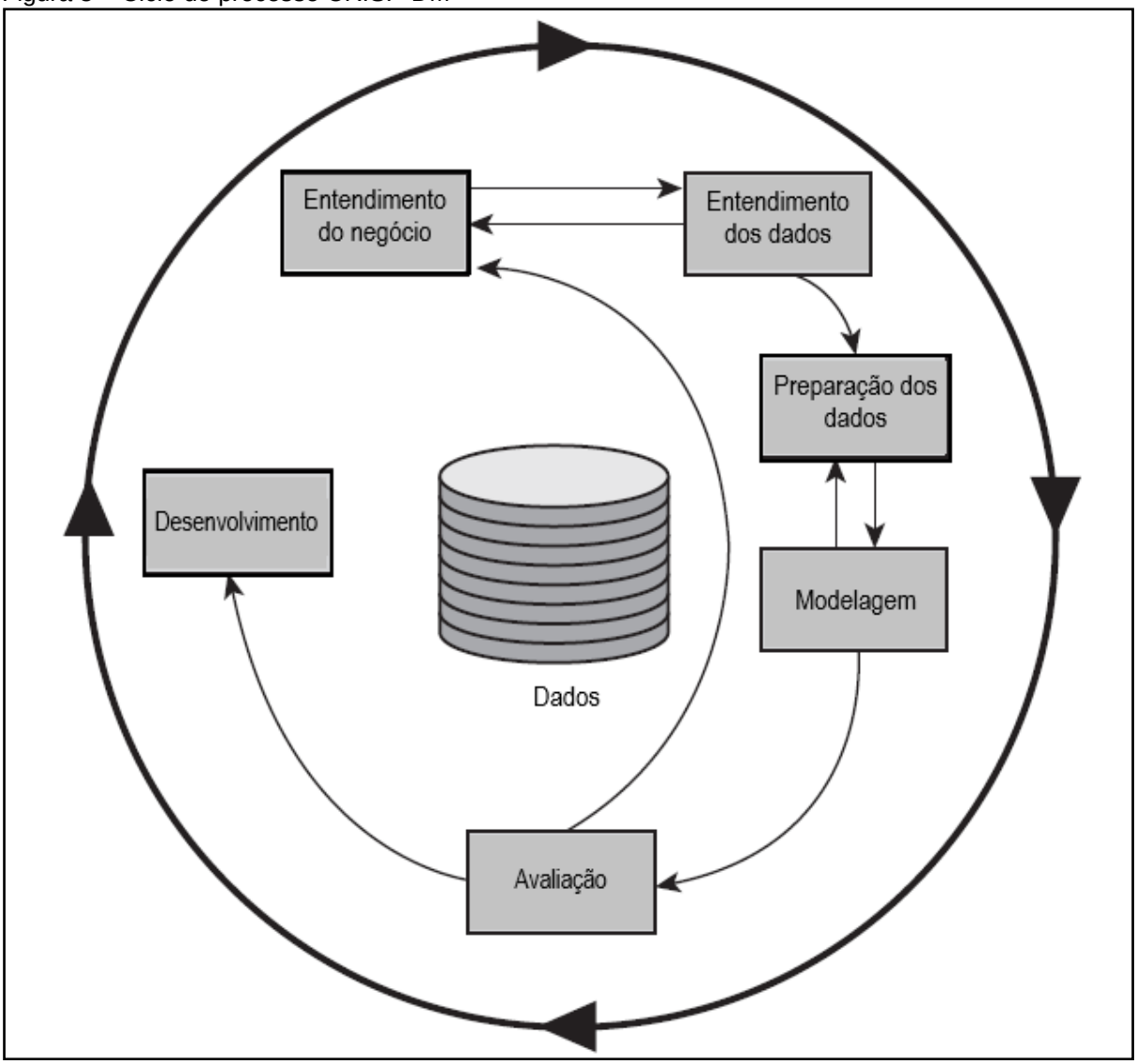
Modelagem, várias técnicas de modelagem são selecionadas e aplicadas e seus parâmetros são calibrados de acordo com a necessidade.

Avaliação, os resultados obtidos são avaliados mais detalhadamente e as etapas executadas para construir o modelo são revisadas para garantir que ele atinja adequadamente os objetivos do negócio.

Implantação/Desenvolvimento, não é o fim do projeto, o conhecimento adquirido deverá ser organizado e apresentado de forma que o usuário possa utilizá-lo, efetuando um relatório identificando os pontos fracos e fortes do projeto e possíveis melhorias a serem efetuadas. Na figura 3 tem-se a representação do ciclo do processo

CRISP-DM. (CHAPMAN et al, 2000, tradução nossa; AZEVEDO; SANTOS, 2008, tradução nossa).

Figura 3 – Ciclo do processo CRISP-DM



Fonte: Chapman et al (2000, tradução nossa).

Segundo Azevedo e Santos (2008, tradução nossa) as etapas do SEMMA e KDD se assemelham, *Sample* pode ser identificada com a Seleção; *Explore* com Pré-processamento; *Modify* com Transformação; *Model* com *Data mining*; *Assess* com Avaliação. Desse modo, o método SEMMA pode ser visto como uma implementação prática dos cinco estágios do método do KDD, uma vez que ele está diretamente ligado ao software SAS *Enterprise Miner*.

Quanto ao CRISP-DM ele também incorpora as etapas do KDD, porém sua comparação não é tão simples como a da metodologia SEMMA, a etapa de

entendimento do negócio pode ser identificada com o desenvolvimento de uma compreensão do domínio de aplicação, o conhecimento prévio relevante e os objetivos do usuário final, etapas iniciais da metodologia do KDD; A fase de Implantação pode ser identificada com a consolidação, incorporando esse conhecimento no sistema. Em relação às etapas restantes, pode-se dizer que: Entendimento de Dados pode ser identificada como a combinação de Seleção e Pré-processamento; Preparação de Dados com Transformação; Modelagem com *Data mining*; Avaliação com Interpretação.

Na tabela 1 é efetuada uma comparação das etapas entre as 3 metodologias de descoberta de conhecimento em base de dados.

Tabela 1 – Comparação das metodologias KDD, SEMMA e CRISP-DM

KDD	SEMMA	CRISP-DM
Pré KDD	-	Entendimento do negócio
Seleção	*Sample*	Entendimento dos dados
Pré-processamento	*Explore*	Preparação dos dados
Transformação	*Modify*	
Data mining	*Model*	Modelagem
Interpretação/Avaliação	*Assess*	Avaliação
Aplicação do KDD	-	Desenvolvimento

Fonte: Santos e Azevedo (2008, tradução nossa).

Portanto, KDD, CRISP-DM e SEMMA têm como um dos objetivos apresentar novos conhecimentos por meio do *data mining* de acordo com a necessidade do negócio a qual é aplicada (AZEVEDO; SANTOS, 2008, tradução nossa).

Por possuir uma documentação detalhada de todas suas etapas, o CRISP-DM se torna mais atrativo na busca de conhecimento em base de dados.

2.1 PROCESSO *CROSS-INDUSTRY STANDART PROCESS FOR DATA MINING* (CRISP-DM)

O CRISP-DM foi desenvolvido por meio dos esforços de um consórcio inicialmente composto pela DaimlerChrysler AG, SPSS Inc., NCR Systems Engineering Copenhagen e OHRA Verzekeringen en Bank Groep B.V.. (CHAPMAN et al, 2000).

Segundo Chapman et al (2000) o ciclo de vida de um projeto de *data mining* consiste em seis etapas, a sequência não é rígida, como é apresentado na figura 3. Suas etapas estão devidamente organizadas, documentadas, estruturadas e definidas, permitindo que um projeto seja facilmente compreendido ou revisado. O resultado de cada etapa determina qual tarefa ou etapa será a próxima a ser realizada. O círculo externo apresentado na figura 3 representa a natureza cíclica do *data mining* já que o processo de descoberta de conhecimento pode não terminar uma vez em que a solução é implantada, pois as lições aprendidas durante o projeto podem desencadear novas questões no domínio da aplicação.

A definição das etapas do CRISP-DM é dada da seguinte forma: compreensão do negócio, esta fase inicial concentra-se na compreensão dos objetivos do projeto e requisitos a partir de uma perspectiva de negócio e logo em seguida converter o conhecimento e aplica-lo em *data mining,* na definição do problema e criação de um plano preliminar destinado a alcançar os objetivos.

Compreensão dos dados, nesta etapa é efetuada a coleta dos dados inicias e após isso são realizadas atividades para se familiarizar com o que foi coletado procurando identificar problemas de qualidade, descobrimentos de *insights* ou detecção de subconjuntos interessantes para formar hipóteses para informação oculta.

Preparação dos dados, o objetivo desta etapa é transformar o conjunto de dados iniciais em um conjunto de dados finais por meio de técnicas de limpeza e formatação dos dados de acordo com a necessidade do projeto.

Modelagem, nesta fase várias técnicas de modelagem são selecionadas e aplicadas, são realizados procedimentos com o intuito de testar a qualidade e validade do modelo utilizado, o sucesso da aplicação também é validado.

Avaliação, os modelos obtidos são avaliados detalhadamente e as etapas executadas para construir o modelo são revisadas com o intuito de verificar se os objetivos foram atingidos, se as tarefas executadas não atenderam o objetivo previsto ela então deve ser executada novamente.

Implantação, a criação do modelo geralmente não é o final do projeto. Mesmo se o propósito do modelo é aumentar o conhecimento dos dados, o conhecimento adquirido deverá ser organizado e apresentado de forma que o usuário possa utilizá-lo, com a elaboração de um relatório final do projeto de forma que sirva como resumo de tudo que foi efetuado e obtido, avaliando o projeto e identificando os

pontos negativos e positivos e o que pode ser melhorado posteriormente. (CHAPMAN et al, 2000, tradução nossa; AZEVEDO; SANTOS, 2008, tradução nossa).

O modelo de processo CRISP-DM é amplamente adotado em projetos de *data mining*, possuindo uma taxa de utilização de 43%, tanto academicamente quanto na indústria, pois contém uma boa descrição das etapas a serem seguidas (KDnuggets, 2014, tradução nossa).

O progresso do *data mining* é definido pela descoberta de informações por meio de métodos automáticos ou manuais. Tornando-se útil em cenários de análise exploratória onde não existem noções pré-determinadas sobre um resultado. Pesquisando informações valiosas e não triviais em grandes volumes de dados constituindo-se no esforço cooperativo de humanos e computadores. Para que os resultados alcançados se tornem satisfatórios é necessário que exista um equilíbrio entre o conhecimento de especialistas humanos na descrição de problemas e objetivos com os recursos computacionais (KANTARDZIK, 2011, tradução nossa).

2.2 *DATA MINING*

A expressão *data mining* foi denominada no final dos anos 80. Desde então *data mining* e descoberta de conhecimento se tornaram assuntos indispensáveis na indústria e nas instituições de ensino, pois fornece informações valiosas em uma grande quantidade de dados (WU, 2012, tradução nossa).

Caracterizado pela extração de informações implícitas e potencialmente úteis contidas em um conjunto de dados com ajuda de especialistas e computadores (WITTEN; FRANK, 2005, tradução nossa).

Segundo Kantardzic (2011, tradução nossa) e Tamilselvi e Kalaiselvi (2013, tradução nossa) o *data mining* pode ser dividido em dois objetivos principais que tendem a ser previsão ou descrição. A previsão envolve a utilização dos conjuntos de dados formados para prever valores futuros que seriam desconhecidos. A descrição concentra-se em encontrar padrões dentro do conjunto de dados que está sendo manipulado. Desse modo é possível classificar o *data mining* em duas categorias: preditiva e descritiva.

A categoria preditiva é composta das seguintes tarefas de *data mining*: regressão, usada para modelar a relação entre uma ou mais variáveis independentes, atributos já conhecidos, na pretensão de predizer possíveis variáveis respostas;

classificação, classifica um item de dados em uma das várias classes predefinidas, emprega um conjunto de exemplos pré-classificados para desenvolver um modelo que possa classificar a população de registros em geral; detecção de mudanças ou desvios, têm como objetivo encontrar as mudanças mais significativas no conjunto de dados na qual é aplicada (ABONYI; FEIL, 2007, tradução nossa; KANTARDZIC, 2011, tradução nossa; TAMILSELVI; KALAISELVI, 2013, tradução nossa).

As tarefas que compõem a categoria descritivas são: sumarização, tarefa descritiva adicional que envolve métodos para encontrar uma descrição compacta para um conjunto (ou subconjunto) de dados; modelagem, encontra um modelo local que descreva dependências significativas entre variáveis ou os valores de um recurso em um conjunto de dados ou em uma parte de um conjunto de dados; agrupamento, tarefa na qual se procura identificar um conjunto finito de categorias ou *cluster* para descrever os dados (KANTARDZIC, 2011, tradução nossa; TAMILSELVI; KALAISELVI, 2013, tradução nossa).

A tarefa de agrupamento pode servir como pré-processamento para outros algoritmos agirem nos clusters gerados. Sendo amplamente utilizada no reconhecimento de padrões (HAN, KAMBER, 2000). A tarefa de agrupamento, será descrita a seguir.

2.2.1 Agrupamento

O agrupamento é uma tarefa comum para análise de dados estatísticos, usada em alguns campos como *business intelligence*, *machine learning*, *data mining*, reconhecimento de padrões, análise de imagens, bioinformática, biologia e segurança (ABONYI; FEIL, 2007, tradução nossa; HAN; KAMBER; PEI, 2011, tradução nossa).

Tendo por objetivo realizar *o* agrupamento de objetos semelhantes entre si e diferentes dos objetos pertencentes a outros *clusters* (BRAMER, 2013, tradução nossa). Por meio da aplicação de técnicas não supervisionadas é possível identificar correlações entre os atributos dos dados, regiões densas e esparsas no espaço de objetos e descobrir o padrão geral de distribuição, as diferenças e semelhanças são avaliadas com base nos valores dos atributos que descrevem os objetos. Para alcançar os objetivos por meio da tarefa de agrupamento existem alguns algoritmos que são aplicados para a formação dos *clusters* (ABONYI; FEIL, 2007, tradução

nossa; HAN; KAMBER; PEI, 2011, tradução nossa; TAMILSELVI; KALAISELVI, 2013, tradução nossa).

Dentro da tarefa de agrupamento é possível classificar os algoritmos de agrupamento nas seguintes categorias: hierárquico; particionamento; densidade; grade e modelos (HAN; KAMBER; PEI, 2011, tradução nossa; TAMILSELVI; KALAISELVI, 2013, tradução nossa).

Existem dois tipos básicos de métodos hierárquicos: aglomerativo, nessa técnica é adotada uma estratégia chamada *bottom-up,* inicialmente cada objetivo representa um grupo. Os agrupamentos mais similares vão se unindo e formando novos agrupamentos até que exista apenas um agrupamento principal. Os algoritmos *Agglomerative Nesting* (AGNES) e *Clustering Using Representatives* (CURE) utilizam esta estratégia; divisivos, adota a estratégia *top-down*, todos os objetos estão no mesmo agrupamento. Os agrupamentos vão sofrendo divisão até que cada objeto represente um agrupamento. O algoritmo *Divisive Analysis* (DIANA) utiliza essa estratégia (BRAMER, 2013, tradução nossa; HAN; KAMBER; PEI, 2011, tradução nossa).

Métodos com base na densidade são utilizados quando os métodos de particionamento e hierárquicos não se aplicam, ou seja, quando a distribuição dos valores dos dados é densa pois acabam não apresentando resultados satisfatórios. Alguns algoritmos que utilizam este método: *O Density Based Clustering Method Based on Connected Regions with Sufficiently High Density (*DBSCAN), *Ordering Points to Identify the Clustering Structure (*OPTICS) e *Density-based Clustwring* (DENCLUE) (BRAMER, 2013, tradução nossa; HAN; KAMBER; PEI, 2011, tradução nossa; KANTARDZIK, 2011, tradução nossa).

Métodos por grade dividem os registros do conjunto de dados em uma estrutura de grade, o tempo de processamento desses métodos costuma ser menor. Os principais algoritmos são o *Statistical Information Grid* (STING) e o *Clustering Using Wavelet Transformation* (WaveCluster) (BRAMER, 2013, tradução nossa; CAMILO; SILVA, 2009; HAN; KAMBER; PEI, 2011, tradução nossa; KANTARDZIK, 2011, tradução nossa).

Quando os dados são gerados por uma série de distribuições os métodos baseados em modelos são utilizados, pois se apoiam na criação de modelos para cada agrupamento e com isso tentam identificar o melhor modelo para cada objeto. Os algoritmos *Expectation Maximization* (EM), uma variação do *K-means*, COBWEB

e CLASSIT se utilizam do método. (BRAMER, 2013, tradução nossa; CAMILO; SILVA, 2009; HAN; KAMBER; PEI, 2011, tradução nossa; KANTARDZIK, 2011, tradução nossa).

Métodos de particionamento têm como objetivo organizar os objetos em um número de *clusters* desejados dentro de um conjunto de dados com n registros e k o número de agrupamentos desejados, tal que k <= n. Os algoritmos mais comuns de agrupamento são*: K-means*, *Fuzzy C-means* e K-*medoids* (BRAMER, 2013, tradução nossa; HAN; KAMBER; PEI, 2011, tradução nossa; KANTARDZIK, 2011, tradução nossa).

2.2.1.1 Algoritmo *K-means*

O K-means é um algoritmo que emprega o critério do erro quadrático, tentando encontrar *clusters* não sobrepostos, baseados na teoria dos conjuntos clássicos e que um objeto pertença ou não a um *cluster* (ABONYI; FEIL, 2007, tradução nossa; MCQUEEN, 1967, tradução nossa; WU, 2012, tradução nossa).

De uma partição aleatória do conjunto inicial de dados começa a reatribuir padrões aos *clusters* pela similaridade entre o padrão de um *cluster* e os centros do *cluster* até que um critério de convergência seja encontrado (ABONYI; FEIL, 2007 tradução nossa; MIRKIN, 2005, tradução nossa).

Figura 4 – Passos do algotimo *K-means*

- **Entrada:**
  1. *k*: Número de *clusters.*
  2. *D*: Conjunto de dados contendo os objetos.
- **Saída:**
  1. Conjunto de *clusters.*
- **Método:**
  1. Escolher aleatoriamente *k* objetos de *D* como centroides iniciais do *cluster;*
  2. Repetir:
     - Atribuir cada objeto ao *cluster* ao qual o objeto é o mais similar, baseado no valor médio dos objetos no *cluster*;
     - Atualizar as médias do *cluster*, ou seja, calcular o valor médio dos objetos para cada *cluster*;
  3. Realizar o procedimento até que não exista nenhuma mudança a ser feita.

Fonte: adaptado Han, Kamber e Pei (2011, tradução nossa).

Uma outra didática de explicação do algoritmo *K-means* segundo Han, Kamber e Pei (2011, tradução nossa), pode-se transformá-lo nos passos apresentados na figura 4.

Segundo Wu (2012, tradução nossa) o progresso do agrupamento *K-means* está subdivido em três aspectos: modelo de generalização; otimização de pesquisa e formato da distância.

O algoritmo tem como objetivo examinar todos os elementos de uma série de dados e associá-los a um centróide, para isso, utiliza-se de uma função capaz de medir a distância entre os centróides. Quanto maior a similaridade, menor é a distância entre os pontos (HAN; KAMBER, 2011).

Para seleção do centróide de cada *cluster* são realizados os cálculos de distância, aplicado ao algoritmo *K-means* tem-se duas formas:

f) Euclidiana: correspondente a distância mínima da soma das raízes quadradas entre os dois pontos considerando todas as coordenadas do espaço de atributos. Sua função matemática é definida por (GOLDSCHMIDT; PASSOS, 2005):

$$distancia(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} \qquad (1)$$

g) Manhattan: soma dos pontos calculados no espaço de atributos. Simples, porém, pode ter pouca precisão. A fórmula é compreendida por (GOLDSCHMIDT; PASSOS, 2005):

$$distancia(p,q) = |p_1 - q_1| + |p_2 - q_2| + \cdots + |p_n - q_n| \qquad (2)$$

Uma dificuldade do *K-means* é não possui tanta eficiência para encontrar *clusters* de tamanhos diferentes, tendo como desvantagens desempenho ruim para *clusters* não-globulares e sensíveis a *outliers* (TAN; KUMAR; STEINBACH, 2006, tradução nossa; WU, 2012, tradução nossa).

2.2.1.2 Algoritmo *Fuzzy C-means*

O agrupamento *fuzzy* permite que os objetos pertençam a vários clusters simultaneamente, com diferentes graus de pertinência (ABONYI; FEIL, 2007, tradução nossa).

Fuzzy C-means é um método de agrupamento difuso que permite que um objeto pertença a dois ou mais *clusters* com grau de associação entre zero e um, sua versão final foi proposta por James C. Bezdek em 1973 (ANDERBERG, 1973, tradução nossa; WU, 2012, tradução nossa).

Amplamente utilizado em aplicações no mundo real em situações que *cluster* não se encontram bem separados normalmente. O procedimento simples e interativo pelo centróide do *Fuzzy C-means* se torna mais atraente quando o volume do conjunto de dados é grande. O armazenamento em *cluster* difuso é mais natural que o *clustering* clássico, pois os objetos nas fronteiras entre várias classes não são obrigados a pertencer totalmente a uma das classes, mas recebem graus de associação entre 0 e 1 indicando suas associações parciais. (WU, 2012, tradução nossa).

O *Fuzzy C-means* pode ser considerado a versão *fuzzy* do tradicional algoritmo *K-means*, portanto, da mesma forma que o *K-means*, este algoritmo utiliza a distância Euclidiana encontrando clusters de forma circular. Assim, o *Fuzzy C-means* pode ser representado pela expressão a seguir (BEZDEK et al, 2005, tradução nossa):

$$J\,(B, U; Z) = \sum_{i=1}^{C} \sum_{j=0}^{N} \left(u_{ij}\right)^{m} \left(d_{ij}^{2}\right) \qquad (3)$$

Onde:

a) C: número total de *clusters*;

b) N: número total de elementos;

c) J: valor a ser minimizado;

d) B: conjunto de *clusters*;

e) U: matriz de pertinências;

f) Z: conjunto de dados;

g) m: parâmetro de *fuzzyficação*;

h) u_{ij}: grau de pertinência do i-ésimo *cluster* e o j-ésimo elemento;

i) d_{ij}: distância entre o i-ésimo *cluster* e o j-ésimo elemento.

Primeiramente é inicializada a matriz de pertinências pelo algoritmo, para então calcular iterativamente as seguintes etapas:

a) determinar os centros dos *clusters*;

b) calcular as distâncias entre elementos e seus grupos;

c) atualizar os graus de pertinência;

d) verificação da condição de parada do algoritmo.

No entanto, o algoritmo possui dificuldades em lidar com dados ruidosos, e algumas limitações na identificação de clusters de diferentes formas (BEZDEK et al, 2005, tradução nossa).

Contudo, o agrupamento é um método não supervisionado tendo resultados dependentes das suposições iniciais e por esse motivo a realização de testes para determinar a qualidade dos resultados obtidos é necessária (GAN; MA; WU, 2007, tradução nossa).

2.3 MEDIDAS DE QUALIDADE

No *data mining* a definição dos índices de validade das medidas de qualidade dos padrões descobertos vêm se tornando uma área de extrema importância para obtenção de resultados satisfatórios na medida que os usuários têm que lidar com numerosas informações que são extraídas e necessitam encontrar quais são as mais interessantes (GENG; HAMILTON, 2007, tradução nossa).

Para identificar se o padrão é interessante ou não existem fatores que dependem somente dos dados brutos na maioria das vezes baseadas em teorias de probabilidade, estatística ou teoria da informação, identificados como objetivos: concisão; generalidade; confiabilidade; peculiaridade; diversidade; e os fatores que dependem dos dados e dos usuários são chamados de fatores subjetivos: novidade; surpresa; utilidade; acionabilidade (GENG; HAMILTON, 2007, tradução nossa).

Segundo Rendón et al (2011, tradução nossa) essas medidas no agrupamento geralmente são definidas pela combinação de dois fatores dos *clusters*: coesão, que que se refere a medida de proximidade dos conjuntos; e separabilidade que indica o quão dois conjuntos são distintos, tal como a distância entre *clusters.*

Índices de validação são empregados para encontrar o número ótimo de *clusters* em uma determinada base de dados. Ajudando a definir a quantidade de *clusters* que encontra partições estáveis, e consequentemente, que melhor definem e explicam a estrutura da base de dados utilizada (KIM et al, 2004, tradução nossa).

Para validação dos clusters gerados pela lógica Fuzzy, Bezdek et al (2005, tradução nossa) propôs os seguintes índices:

$$V_{pc} = \frac{\sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}^2}{N} \quad (4)$$

Onde:

a) V_{pc}: índice do coeficiente de partição;

b) C: quantidade de *clusters*;

c) N: quantidade de elementos;

d) u_{ij}: grau de pertinencia do j-ésimo elemento no i-ésimo *cluster*.

O coeficiente de particição é definido pelo somatório de todos os graus de pertinência, de todos os elementos em relação a todos os clusters, ao quadrado, dividido pelo número de elementos da base. Para validação dos *clusters* procura-se o valor mais aproximado de 1, indicando *clusters* bem definidos, desta maneira deve-se trabalhar os parâmetros na aplicação do algoritmo até que se encontre o valor aceitável (BEZDEK et al, 2005, tradução nossa).

O índice de partição entrópica, também desenvolvido por Bezdek é definido pela expressão (BEZDEK et al, 2005, tradução nossa):

$$V_{pe} = \frac{1}{N}\sum_{j=1}^{n}\sum_{i=1}^{c}[u_{ij}\ \log_{10}(u_{ij})] \quad (5)$$

Onde:

a) V_{pe}: índice do coeficiente de partição entrópica;

b) C: quantidade de *clusters*;

c) N: quantidade de elementos;

d) u_{ij}: grau de pertinência do j-ésimo elemento no i-ésimo *cluster*.

Este índice trabalha de forma contraria ao coeficiente de partição, se trata de uma função minimizadora, quanto mais os valores se aproximarem de zero melhor a identificação dos *clusters*.

Desenvolvido por Xie e Beni, o índice a seguir é chamado de Xie-Beni ou Xie and Beni, devido aos seus criadores. Procurando definir o número ótimo de clusters considerando a separação e compactação dos mesmos. A análise de seus resultados se assemelha ao coeficiente de partição, desta maneira, quando o índice apresentar valores baixos significa que os grupos são bem separados e compactos (KIM et al, 2004, tradução nossa).

A representação matemática de Xie and Beni se dá pela seguinte expressão:

$$V_{xb} = \frac{\left(\sum_{j=1}^{n} \sum_{i=1}^{c} u_{ij}^{m} \, ||x_j - c_i|| \right)}{N \mid (min_{i,k} \, ||c_i - \, c_k||)} \tag{6}$$

Onde:

a) V_{xb}: índice Xie and Beni;

b) C: quantidade de *clusters*;

c) N: quantidade de elementos;

d) u_{ij}: grau de pertinência do j-ésimo elemento no i-ésimo *cluster*;

c) x_j: quantidade de elementos;

c) c_i: quantidade de elementos;

O índice Xie-Beni é definido como o somatório dos graus de pertinência elevados ao elemento fuzzyficador multiplicados pela distância entre elementos e centros dos *clusters* dividido pela quantidade de elementos, sendo que este valor define a compactação dos grupos. O resultado é dividido pela menor distância entre os centros dos *clusters*, este valor define a separação dos grupos (KIM et al, 2004, tradução nossa).

Para validação de *clusters* gerados pelo algoritmo *K-means* Thinsungnoen et al (2015, tradução nossa) recomenda a análise do índice *Sum of Squared Errors* (SSE), sua representação matemática se dá pela expressão:

$$SSE_{(X,\prod)} = \sum_{i=1}^{n} \sum_{x_j \in c_i} ||x_j - m_i||^2 \tag{7}$$

Onde:

a) $SSE_{(x,\prod)}$: *Sum of Squared Errors*;

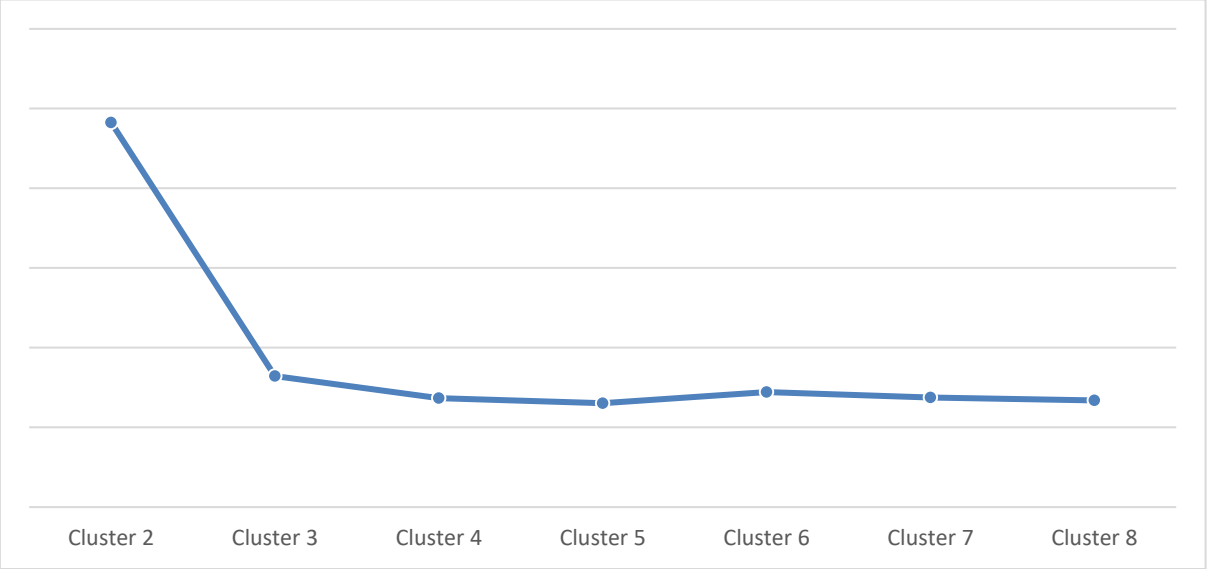
b) X: quantidade de *clusters*;

c) $\prod$: quantidade de elementos;

d) $|| \cdot ||$: distância euclidiana e centróides dos *clusters*

Traduzido literalmente para soma dos erros quadráticos, a análise de seus resultados se busca encontrar qual *cluster* obtive maiores percentuais de mudança em referência ao anterior, também podendo ser analisado por meio de gráficos utilizando o método de Elbow, ou cotovelo.

Figura 5 – Analisando SSE pelo método de Elbow



Fonte: Do autor.

Na figura 5 é possível perceber que no *cluster* 3 os resultados apresentam uma mudança brusca nos seus valores, posteriormente mantendo um padrão em seus resultados. O método do cotovelo analisa de forma visual a diferença entre o *cluster* e o próximo *cluster*, no momento que a diferença deixar de ser significativa entende-se que a quantidade de *clusters* ideal para os dados foi encontrada. De acordo com o método o número de clusters aceitável no exemplo seriam 3.

Outra maneira de se validar os resultados do SSE seria pelo cálculo do percentual de mudança representado na expressão 8, o *cluster* que apresentar o maior percentual de mudança é aquele com o valor mais aceitável para o conjunto de dados que está sendo utilizado:

$$\%\ change = \frac{\left(SSE_{k_{i-1}} - SSE_{k_i}\right) * 100}{SSE_{k_i}} \tag{8}$$

Onde:

a) $\%\ change$: É o percentual de mudança do *cluster*;

b) SSE: É o índice *SSE do cluster*.

Desta forma é de grande importância a aplicação de medidas de qualidade para a obtenção da quantidade de *clusters* possível dentro do conjunto de dados, podendo assim demonstrar resultados melhores no processo de descoberta de

conhecimento. A seguir são apresentados alguns trabalhos que abordam os temas contidos nessa pesquisa.

3 TRABALHOS CORRELATOS

Algoritmos de agrupamento desempenham um papel importante em amplos domínios de aplicação, tais como a área da meteorologia entre outras. A seguir são apresentados alguns casos que se utilizam do algoritmo *K-means* e *Fuzzy C-means* para a obtenção de conhecimento em dados meteorológicos.

3.1 ANÁLISE DE ZONAS HOMOGÊNEAS EM SÉRIES TEMPORAIS DE PRECIPITAÇÃO NO ESTADO DA BAHIA

Em um artigo publicado por Dourado, Oliveira e Avila, em 2013, na Bragantia, utiliza-se a técnica de agrupamento de dados combinado com o algoritmo *K-means* para transformação das séries históricas de precipitação zonas pluviometricamente homogêneas por meio do modelo de processo CRISP-DM.

Foram utilizadas séries históricas de precipitação pluviométrica diárias do sistema de informações hidrológicas Hidroweb da ANA em um período amostral de 30 anos (DOURADO; OLIVEIRA; ÁVILA, 2013).

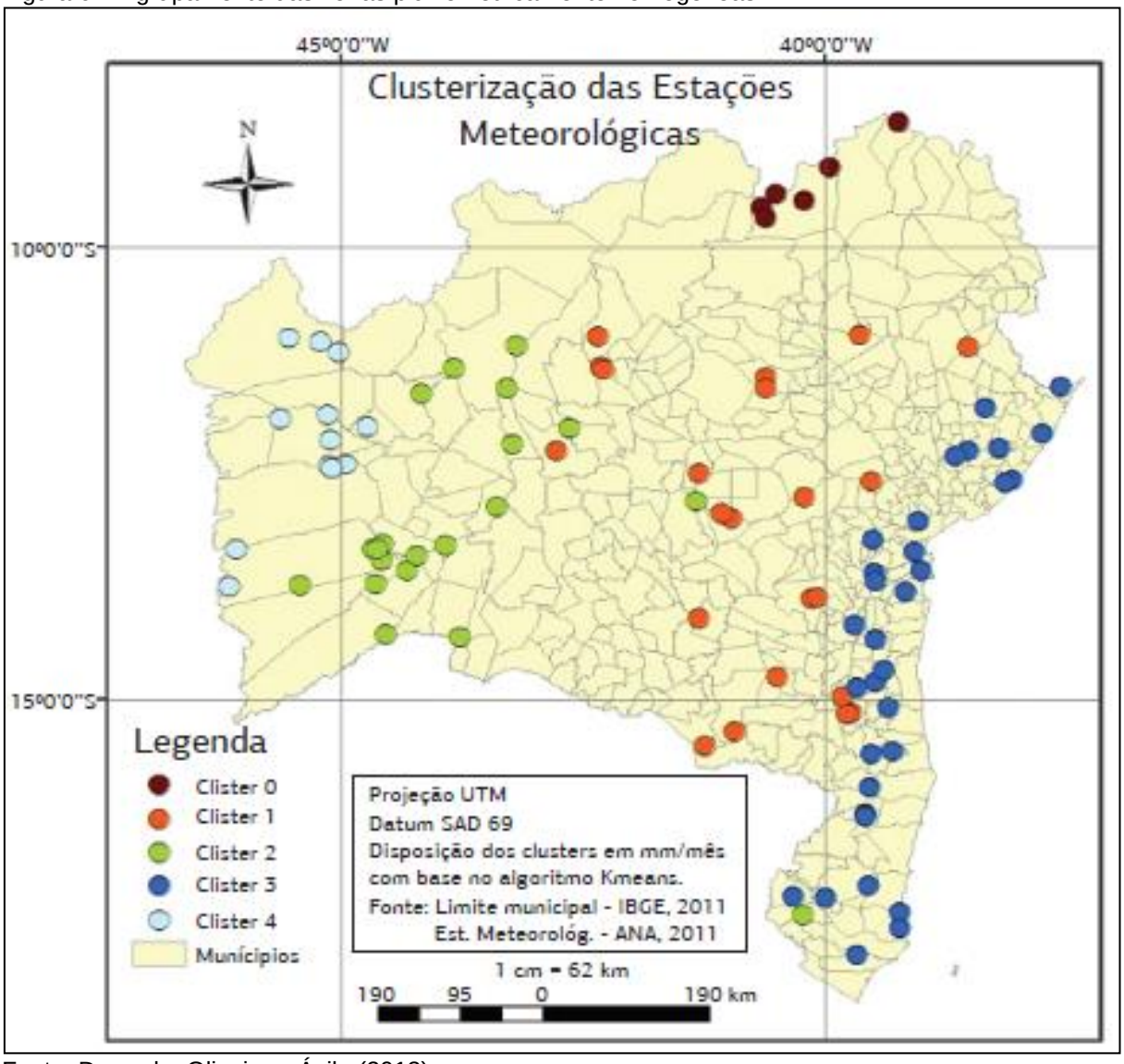
Segundo a padronização da Organização Mundial da Meteorológica as utilizações de um período amostral de 30 anos minimizam as interferências com eventos adversos ocorridos na região (OMM, 1989, 2007, tradução nossa).

Os dados consistidos se resumem a uma planilha composta por 92 linhas (registros), condizentes às estações meteorológicas, e 10.957 colunas (atributos), relacionadas aos valores diários de precipitação (DOURADO; OLIVEIRA; ÁVILA, 2013).

Após a aplicação da técnica de agrupamento por meio do algoritmo *K-means* foi obtido a seguinte representação apresentada na figura 4, os pontos representam as 92 estações meteorológicas do estado (DOURADO; OLIVEIRA; ÁVILA, 2013).

Os valores testados em *K* foram 2, 3, 4 e 5 pois condiziam mais com a realidade dos dados. A escolha desse intervalo deu-se com base em estudos já realizados, e a validação dos *clusters* pela comparação com mapas pluviométricos da região (DOURADO; OLIVEIRA; ÁVILA, 2013).

Figura 6 – Agrupamento das zonas pluviometricamente homogêneas.



Fonte: Dourado, Oliveira e Ávila (2013).

Os resultados mais satisfatórios apontaram para cinco zonas pluviometricamente homogêneas no Estado da Bahia (figura 6), com os dados das precipitações pluviométricas com base no período de 1981 a 2010, as zonas mais secas estão situadas na parte central, de norte a sul do estado.

A alta variabilidade pluviométrica ocorre em quase todas as zonas, principalmente em duas do semiárido com coeficientes de variação iguais a 42 e 28%. Diferencia-se dessa característica a zona pertencente à faixa litorânea, que apresenta regularidade de chuvas durante todo o ano e coeficientes de variação de 15%. As estações chuvosas e secas estão bem definidas. Os valores de precipitação da estação chuvosa representam em torno de 81% dos totais anuais, com destaque para

as zonas situadas no centro-oeste e oeste do estado, com 95 e 96% dos totais anuais. (DOURADO; OLIVEIRA; ÁVILA, 2013).

3.2 CARACTERIZAÇÃO DA PRECIPITAÇÃO MENSAL, SAZONAL E ANUAL PARA O ESTADO DO PARANÁ EM PERÍODOS SECOS, NORMAIS E CHUVOSOS (1977-2006)

Em 2017, Mello e Leite, no XVII Simpósio Brasileiro de Geografia Física Aplicada e I Congresso Nacional de Geografia Física, apresentaram um estudo que aborda o algoritmo *K-means* na caracterização e classificação da precipitação do estado do Paraná.

Os dados utilizados foram retirados da ANA, em um total de 166 estações pluviométricas no período de 1977 a 2006, os dados passaram por uma análise com o intuito de que inconsistências sejam retiradas. Foram definidos 4 grupos para k, cada grupo representa uma região homogênea. De cada grupo foram selecionadas duas estações para que os representassem, a seleção das estações foi estabelecida pelos menores valores de diferença em porcentagem entre a média da estação e a média de todas estações da região homogênea, e também a localização geográfica da estação para a análise dos quantis (MELLO; LEITE, 2017).

Com o intuito de auxiliar na avalição dos resultados foi utilizada a análise de agrupamento hierárquica denominada diagrama de árvore ou dendograma. Utilizou-se o software Statisca 10, o método de ligação "*single linkage*" e a distância euclidiana (MELLO; LEITE, 2017).

Na classificação da precipitação foram utilizados os seguintes quantis:

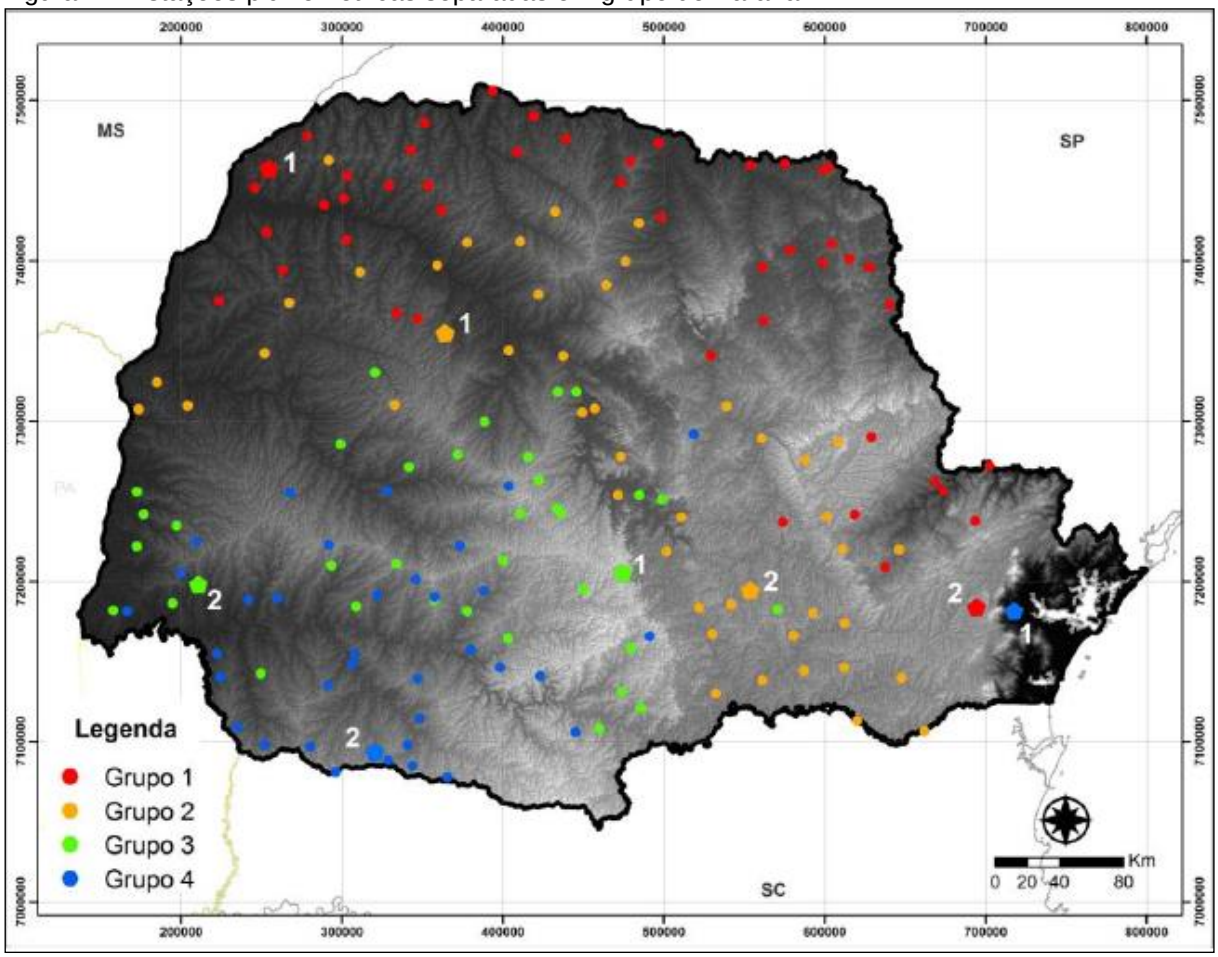
Tabela 2 – Intervalos de classe para classificação da precipitação

Probabilidade	Categoria	Quantificação
$P < Q_{0,15}$	Muito Seco "MS"	15
$Q_{0,15} \leq P < Q_{0,35}$	Seco "S"	**20**
$Q_{0,35} \leq P < Q_{0,65}$	Normal "N"	30
$Q_{0,65} \leq P < Q_{0,85}$	Chuvoso "C"	20
$P > Q_{0,85}$	Muito Chuvoso "MC"	15

Fonte: Mello e Leite (2017).

Após a aplicação das técnicas de agrupamento foi possível identificar o total de estações em cada grupo, no primeiro grupo foram identificadas 47 estações, no segundo 46, no terceiro 36 e no quarto 37. Na figura 7 é possível observar a distribuição das estações e grupos pelo estado (MELLO; LEITE, 2017).

Figura 7 – Estações pluviométricas separadas em grupo do Paraná.



Fonte: Mello e Leite (2017).

Com essas informações foram feitas as classificações de quantis com os dados na forma mensal resultando nas seguintes quantificações: muito seco (13,3%); seco (18,9%); normal (28,6%); chuvoso (18,9%); muito chuvoso (12,8%); misto (7,5%). Para os dados sazonais, as classes apresentaram as seguintes quantificações médias: muito seco (15%); seco (20%); normal (29,2%); chuvoso (18,3%); muito chuvoso (13,3%); misto (4,2%). Para os dados anuais, as classes apresentaram as seguintes quantificações: muito seco (13,3%); seco (16,7%); normal (33,3%); chuvoso (16,7%); muito chuvoso (13,3%); misto (6,7%). Alguns períodos foram classificados

como misto, pois não foi possível definir um padrão. Na figura 8 é possível verificar o resultado da classificação sazonal e anual.

Figura 8 – Classificação pluviométrica sazonal e anual.

	VERÃO	OUTONO	INVERNO	PRIMAVERA	ANUAL
1977	N	S	N	S	S
1978	MS	MS	N	S	MS
1979	S	C	S	C	N
1980	Misto (C/N)	Misto	C	N	N
1981	C	S	MS	S	S
1982	S	MS	MC	MC	C
1983	N	MC	MC	MC	MC
1984	N	N	C	N	N
1985	S	C	MS	MS	MS
1986	C	C	S	S	N
1987	S	MC	N	S	Misto
1988	MS	C	MS	MS	MS
1989	MC	N	C	S	C
1990	N	C	MC	N	MC
1991	S	S	N	MS	S
1992	MS	MC	C	MS	C
1993	C	N	N	N	Misto
1994	C	N	N	N	N
1995	MC	S	S	N	N
1996	MC	N	S	C	C
1997	MC	MS	MC	MC	MC
1998	N	MC	C	C	MC
1999	C	N	N	MS	S
2000	Misto	S	N	C	C
2001	Misto (C/N)	N	C	Misto	N
2002	N	N	S	C	N
2003	N	S	S	N	S
2004	MS	C	N	N	N
2005	MS	N	N	MC	N
2006	S	MS	MS	N	MS

Fonte: Mello e Leite (2017).

A classificação da precipitação é um tanto dependente das análises e interpretações dos pesquisadores, ou seja, com a mesma série histórica o resultado poderia ser diferente (MELLO; LEITE, 2017).

Um único ponto negativo ao utilizar esse método de quantis se dá pelo fato da separação abrupta de estações por exemplo se em janeiro de 2017 chovesse 300 mm, e o $Q_{0,65}$ fosse 299,5, o período seria chuvoso, no entanto, se em 2018 chovesse 299 mm, o período seria considerado normal. Dessa maneira, para uma melhor análise dos resultados é importante analisar os totais pluviométricos para cada período estudado (MELLO; LEITE, 2017).

3.3 CLASSIFICAÇÃO DE SÉRIES DE PRECIPITAÇÃO USANDO O METÓDO *FUZZY* DE CLUSTERIZAÇÃO NA TURQUIA

Dikbas et al em 2011, publicaram na *International Journal of Climatology* um estudo no qual por meio de medidas de qualidade sobre os resultados do *Fuzzy C-means* identificam a quantidade de zonas homogêneas no país Turquia, já que identificar zonas homogêneas pela proximidade das estações ou por regiões podem não apresentar resultados válidos.

A identificação de regiões homogêneas é geralmente o passo mais importante e difícil da análise das precipitações. As bacias são majoritariamente agrupadas e classificadas pelas posições das estações geograficamente próximas no mesmo grupo. No entanto, não é possível dizer que as regiões geradas com essa abordagem sejam homogêneas (DIKBAS et al, 2011, tradução nossa).

Por este motivo, para uma análise mais precisa os métodos de agrupamento tem sido utilizado para identificar regiões homogêneas (DIKBAS et al, 2011, tradução nossa).

Por possuir uma característica em que um vetor pode pertencer a vários grupos com o grau de pertinência especificado com o grau de adesão entre 0 e 1 o algoritmo difuso, *Fuzzy C-means,* pode resultar em mais informações para o conhecimento do usuário no domínio da aplicação (DIKBAS et al, 2011, tradução nossa).

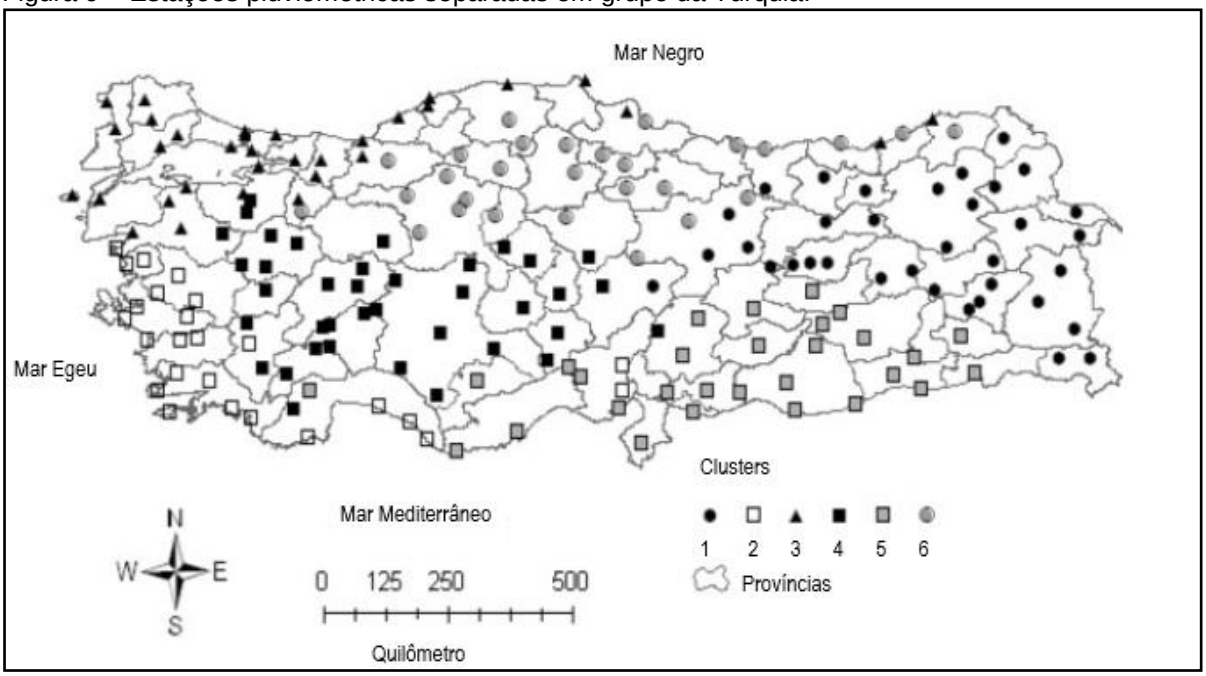
Os dados de precipitação utilizados foram fornecidos *National Meteorology Works* (DMI), a área total das precipitações é de 816.156,70 km², o período dos dados observados é de 1 de outubro de 1967 até 30 de setembro de 1998 (DIKBAS et al, 2011, tradução nossa).

De acordo com os índices de validade utilizados: *Partition Index* (PI) (Bezdek, 1981); *Xie and Beni Index* (XB) (Xie and Beni, 1991); *Dunn Index* (DI); *Alternative Dunn Index* (ADI) e aplicação da técnica *Fuzzy C-means,* foram identificadas seis regiões homogêneas conforme apresentado na figura 7 (DIKBAS et al, 2011, tradução nossa).

Após a obtenção dos resultados com o *Fuzzy C-means* foram realizados o mesmo procedimento com o algoritmo *K-means*, com o intuito de realizar uma comparação e identificar qual apresenta melhor desempenho. Foi possível identificar que o *Fuzzy C-means* foi melhor que o método *K-means* na identificação de zonas

homogêneas, a validação dos resultados se deu pela análise geográfica dos clusters (DIKBAS et al, 2011, tradução nossa).

Figura 9 – Estações pluviométricas separadas em grupo da Turquia.



Fonte: Dikbas et al (2011, tradução nossa).

3.4 APLICAÇÃO DO ALGORITMO *K-MEANS* EM DADOS DA PREVALÊNCIA DE ASMA E RINITE EM ESCOLARES

Martins et al, em 2008, realizaram um estudo utilizando o algoritmo de agrupamento *K-means* na ferramenta *Orion Data Mining Engine* em uma base dados que contém informações referente a prevalência da asma e renite em adolescentes escolares no município de criciúma.

Na realização dos testes foram analisados o tempo de processamento do algoritmo *K-means* sobre os 3010 registros com 5 clusters e variando os atributos os resultados das análises foram satisfatórios (MARTINS et at, 2008).

Foram gerados outros clusters para fins de comparação de resultados utilizando os atributos de saída: sono perturbado e teve asma, foi possível observar que os clusters que não possuem ocorrência de sono perturbado têm uma incidência menor em asma (MARTINS et at, 2008).

Com isso, pode-se afirmar que a utilização do *data mining* na descoberta de conhecimento em dados na área da saúde é de suma importância para a aquisição de conhecimento. Os resultados obtidos foram adequados, porém futuramente torna-

se necessário à sua validação por meio de medidas de qualidade e índices de validade (MARTINS et at, 2008).

3.5 TÉCNICAS DE MINERAÇÃO DE DADOS PARA ANÁLISE DA PRECIPITAÇÃO PLUVIAL DECENAL NO RIO GRANDE DO SUL

Em um artigo publicado na Revista de Engenharia Agrícola, em 2011, Raquel, Stanley e Eduardo apresentam um estudo no qual por meio da técnica de *data mining* utilizando-se da tarefa de agrupamento pelo algoritmo *K-means,* com base na metodologia CRISP-DM analisam a precipitação em um total de seis zonas homogêneas em dois decênios.

Os dados utilizados foram retirados da ANA, o período dos dados foi definido em dois decênios: 1987-1996 e 1997-2008. No total foram analisados os dados consistidos de 79 estações pluviométricas, o único atributo considerado foi a quantidade de chuva mensal (BOSCHI; OLIVEIRA; ASSAD, 2011).

Por meio do programa computacional Weka (WITTEN; FRANK, 2005, tradução nossa) foi utilizado o algoritmo *K-means* (MCQUEEN, 1967, tradução nossa).

A eficiência da técnica de agrupamento na análise do comportamento espaço temporal da precipitação pluviométrica no Estado do Rio Grande do Sul ficou comprovada, a escolhe da quantidade de clusters se deu por pesquisas já realizadas anteriormente (BOSCHI; OLIVEIRA; ASSAD, 2011).

Ao analisar as informações obtidas foi identificado que a precipitação anual apresentou um incremento significativo, entre 20 mm a 240 mm, nos decênios analisados. Os maiores desvios em percentuais ocorreram na zona localizada ao sul do estado enquanto os menores ocorreram na zona norte (BOSCHI; OLIVEIRA; ASSAD, 2011).

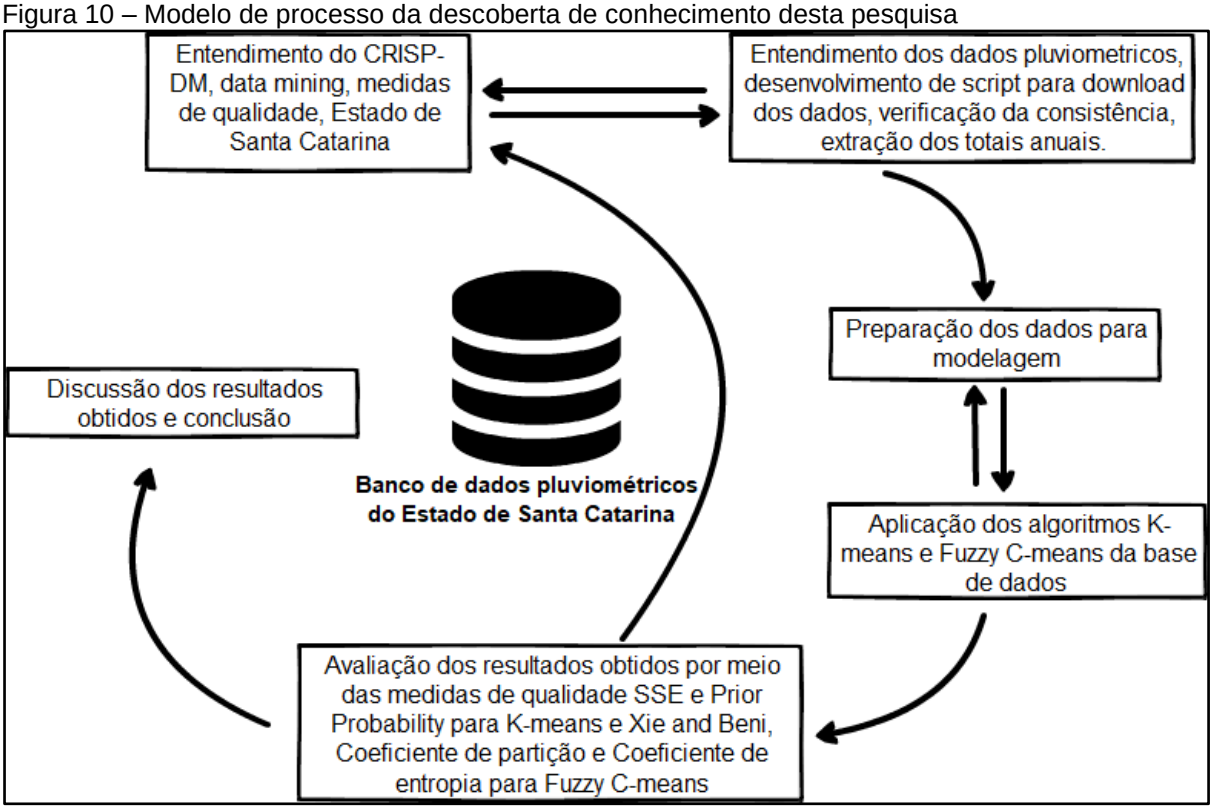
Para trabalhos futuros sugere-se verificar se tais mudanças encontradas sejam analisadas com dados de produção agrícola a fim de identificar os reais impactos na agricultura do estado. Estudar as possíveis causas dessas variações e adicionar outros atributos como umidade e temperatura, verificando assim se existem associações com os fenômenos estudados (BOSCHI; OLIVEIRA; ASSAD, 2011).

4 APLICAÇÃO DOS ALGORITMOS *K-MEANS* E *FUZZY C-MEANS* NA IDENTIFICAÇÃO DE ZONAS PLUVIOMÉTRICAS NO ESTADO DE SANTA CATARINA

A abordagem dos algoritmos se dá pelo fato do algoritmo *K-means* ser amplamente utilizado nas análises de dados pluviométricos em outras regiões, como por exemplo Dikbas et al (2011) no País Turquia, Dourado, Oliveira e Avila (2013) no Estado da Bahia, onde demonstraram resultados satisfatórios de agrupamento ao serem validados por medidas de qualidade.

Esta pesquisa consiste na aplicação de *data mining* por meio da metodologia CRISP-DM para o agrupamento de dados utilizando a técnica tradicional chamada *K-means* e por meio da técnica baseada na lógica *fuzzy* que possui o nome de *Fuzzy C-means.* Posteriormente a aplicação desses algoritmos na base de dados que contempla os totais anuais das precipitações pluviométricas de 42 estações entre 1970 e 2000 do Estado de Santa Catarina, os resultados foram analisados por meio de medidas de qualidade para agrupamento afim de identificar dentre esses algoritmos qual obtém o melhor modelo para identificação de zonas pluviométricas.

Figura 10 – Modelo de processo da descoberta de conhecimento desta pesquisa



Fonte: Do autor

A seguir é apresentado maiores informações sobre área de estudo desta pesquisa, que corresponde ao Estado de Santa Catarina.

4.1 PRECIPITAÇÃO PLUVIOMÉTRICA EM SANTA CATARINA

O Estado de Santa Catarina possui uma área oficial de 95.483 km², com mais 502 km² de águas territoriais, totalizando 95.985 km², correspondente a 1,12 % da área brasileira e 16,61% da Região Sul. Situado entre as latitudes 26º00´S e 30º00´S, e longitudes 48º30´W e 54º00´W (IBGE, 2010).

Tendo grande importância no cenário nacional pois é o primeiro na produção em alho, cebola, maçã, segundo produtor em fumo, terceiro produtor em trigo, quarto produtor em arroz e milho, quinto produtor em batata (PADOLFO et al, 2002).

Segundo a classificação de Köppen-Geiger, o Estado de Santa Catarina é classificado como clima mesotérmico úmido, pois não possui estação seca, incluindo mais dois subtipos: clima subtropical onde a temperatura média no mês mais frio inferior a 18ºC (mesotérmico) e temperatura média no mês mais quente acima de 22ºC, com verões quentes, geadas pouco frequentes e tendência de concentração das chuvas nos meses de verão, contudo sem estação seca definida; clima temperado temperatura média no mês mais frio abaixo de 18ºC (mesotérmico), com verões frescos, temperatura média no mês mais quente abaixo de 22ºC e sem estação seca definida. (PADOLFO et al, 2002).

As principais massas de ar atuantes no Estado são: Tropical Atlântica, Polar Atlântica, Tropical Continental e a Equatorial Continental (BALDO et al, 2000).

Por sua localização geográfica, é um dos Estados que apresenta melhor distribuição de precipitação pluviométrica durante o ano. Os principais sistemas meteorológicos responsáveis pelas chuvas no estado são as frentes frias, os vórtices ciclônicos, os cavados de níveis médios, a convecção tropical, a ZCAS e a circulação marítima. Como possui estas características o estado acaba por ter todos os tipos de precipitações: pluviométrica; neve; granizo; orvalho; e geada (MONTEIRO, 2001).

A precipitação pluviométrica entre elementos meteorológicos que é a que exerce maior influência sobre as condições ambientais e principalmente nas atividades desenvolvidas em campo, dessa forma seu estudo serve de subsídio para o planejamento rural (BALDO et al, 2000; COAN; BACK; BONETTI,2014).

Figura 11 – Mapa de Santa Catarina

Fonte: CEPED-UFSC (2013).

Divido em seis mesorregiões (figura 11): Norte Catarinense, Vale do Itajaí, Grande Florianópolis, Sul Catarinense, Serrana e Oeste Catarinense (IBGE, 2010).

4.2 METODOLOGIA

As etapas metodológicas desta pesquisa são as seguintes: levantamento bibliográfico, por meio do estudo referente a metodologia de processo CRISP-DM, *data mining*, tarefa de agrupamento, series históricas de precipitações pluviométricas, algoritmos *K-means* e *Fuzzy C-means*, medidas de qualidade e pôr fim a análise dos índices a fim de encontrar os melhores resultados para os dados em questão.

Para auxiliar o processo de descoberta do conhecimento foi utilizado a metodologia CRISP-DM, a seguir são apresentadas suas etapas, iniciando-se pela etapa de entendimento dos dados e posteriormente preparação dos dados, modelagem, avaliação, resultados obtidos e discussão.

4.2.1 Entendimento dos dados

Foram utilizadas séries históricas de precipitação pluviométrica anuais, adquiridas no Portal HidroWeb[6] da ANA. De acordo com a padronização da Organização Meteorológica Mundial para caracterização de dados climáticos é necessário ao mínimo 30 anos de dados consistidos, pois dados brutos oriundos das estações podem apresentar problemas como erros de leitura, transcrição, digitação e ausência de dados. A ANA valida a consistência dos dados pela metodologia proposta pela Agência Nacional de Energia Elétrica, baseada no modelo matemático desenvolvido por Holanda e Oliveira (1979).

Na forma que os dados são disponibilizados no HidroWeb só é possível efetuar o download dos dados de uma estação por vez, desta maneira, causaria grande demora na criação de uma base para efetuar a mineração já que existem 815 estações pluviométricas cadastradas no inventário da ANA para Santa Catarina. Em decorrência disso, foi desenvolvido um *script* em python para que simulasse o acesso a aplicação HidroWeb e efetuasse o download dos dados de todas as estações de forma automática.

O funcionamento do script consiste em acessar o banco de dados de inventário da ANA, disponível no HidroWeb, buscando as estações pluviométricas de Santa Catarina. Com essas informações o *script* executa o navegador Firefox informando-o para acessar o HidroWeb, após o acesso o *script* executa as ações necessárias no site como se fosse um usuário realizando a busca, efetuando as ações de cliques, colocando o nome e código da estação e efetuando o *download* de toda série histórica de precipitação disponível para aquela estação. Após efetuar o download o *script* fecha o navegador, busca a próxima estação e repete todo o processo. O processo só é finalizado quando o *script* terminar de realizar o *download* dos dados disponíveis de todas as estações.

Na busca por aumentar a quantidade de estações pluviométricas que atendessem os requisitos mínimos, foi enviada uma solicitação de dados a Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina (EPAGRI) / Centro de Informações de Recursos Ambientais e de Hidrometeorologia de Santa Catarina

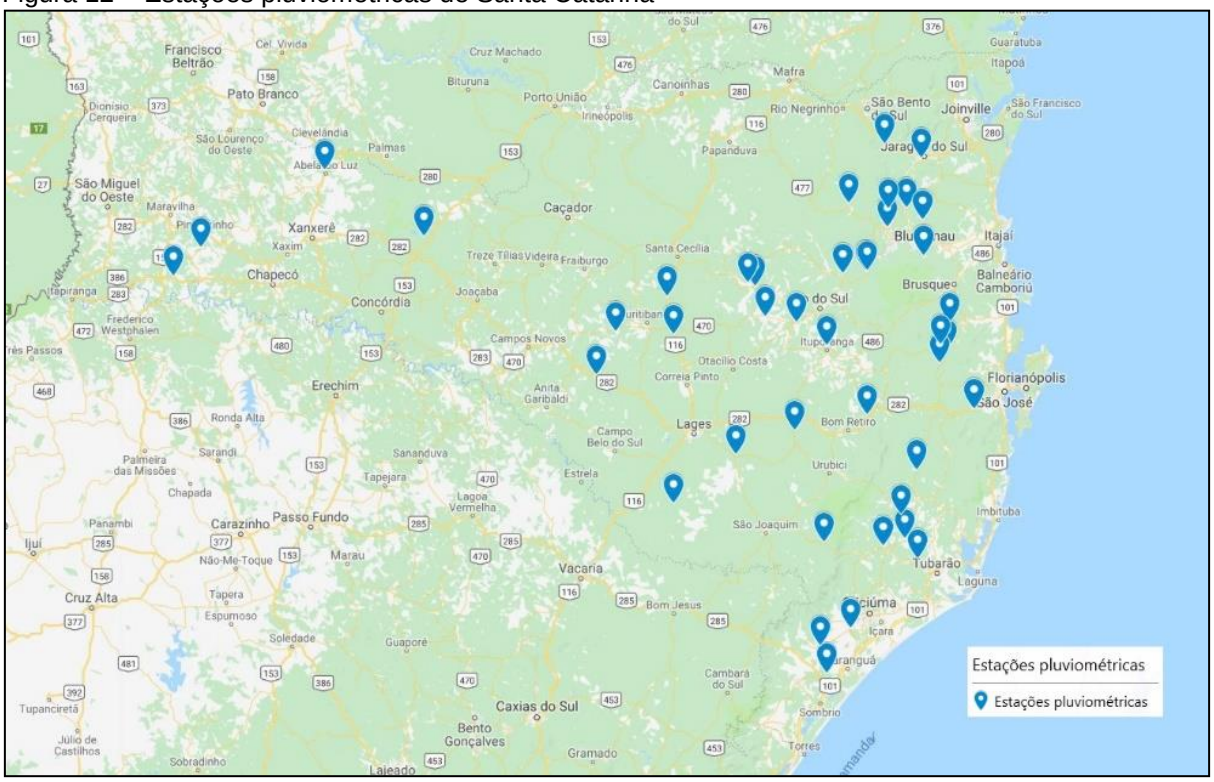
---

[6] Os dados estão disponíveis no Portal Hidro Web em (http://www.snirh.gov.br/hidroweb/)

(CIRAM), onde foram disponibilizados todos os dados de precipitações disponíveis, no entanto, não foi possível acrescentar mais estações visto que nenhuma atendeu o mínimo de 30 anos de dados consistidos.

Desta maneira, o maior período de estações que atenderam os requisitos de 30 anos com dados consistidos é de 1970 a 2000, totalizando 42 estações meteorológicas e 31 anos de dados obtidos da ANA, cobrindo uma boa parte do estado, conforme demonstrado na figura 12.

Figura 12 – Estações pluviométricas de Santa Catarina



Fonte: Do autor.

A próxima etapa, preparação dos dados, consiste na análise desses dados e formatação dos mesmo para a aplicação das técnicas de modelagem.

4.2.2 Preparação dos dados

Os dados obtidos foram importados no software Hidro Sistemas de Informações Hidrológicas 1.3[7], onde foram extraídos para um arquivo CSV contendo todas as informações de precipitações anuais por estação, código da estação, latitude

[7] Hidro Sistemas de Informações Hidrológicas 1.3 disponível em http://www3.ana.gov.br/

e longitude. A utilização desse software se faz necessária pois os dados obtidos são identificados por colunas que exigem um conhecimento que se encontra implícito no banco já que não possuem documentação, utilizando o software ele realiza a leitura e entrega os totais anuais consistidos de cada estação.

Os dados para análise dos *clusters* foram organizados somente com dados numéricos, não foi necessária efetuar a transformação dos dados visto que todos já eram numéricos, totalizando 34 atributos em 42 registros (tabela 3):

Tabela 3 – Descrição da organização dos dados

Atributo	**Descrição**	**Valor**
Estação	Código da estação ANA	Numérico
Latitude	Latitude	Numérico
Longitude	Longitude	Numérico
chuva_1970	Total de precipitação anual	Numérico
chuva_1971	Total de precipitação anual	Numérico
chuva_1972	Total de precipitação anual	Numérico
chuva_1973	Total de precipitação anual	Numérico
chuva_1974	Total de precipitação anual	Numérico
chuva_1975	Total de precipitação anual	Numérico
chuva_1976	Total de precipitação anual	Numérico
chuva_1977	Total de precipitação anual	Numérico
chuva_1978	Total de precipitação anual	Numérico
chuva_1979	Total de precipitação anual	Numérico
chuva_1980	Total de precipitação anual	Numérico
chuva_1981	Total de precipitação anual	Numérico
chuva_1982	Total de precipitação anual	Numérico
chuva_1983	Total de precipitação anual	Numérico
chuva_1984	Total de precipitação anual	Numérico
chuva_1985	Total de precipitação anual	Numérico
chuva_1986	Total de precipitação anual	Numérico
chuva_1987	Total de precipitação anual	Numérico
chuva_1988	Total de precipitação anual	Numérico
chuva_1989	Total de precipitação anual	Numérico
chuva_1990	Total de precipitação anual	Numérico
chuva_1991	Total de precipitação anual	Numérico
chuva_1992	Total de precipitação anual	Numérico
chuva_1993	Total de precipitação anual	Numérico
chuva_1994	Total de precipitação anual	Numérico
chuva_1995	Total de precipitação anual	Numérico
chuva_1996	Total de precipitação anual	Numérico
chuva_1997	Total de precipitação anual	Numérico
chuva_1998	Total de precipitação anual	Numérico
chuva_1999	Total de precipitação anual	Numérico
chuva_2000	Total de precipitação anual	Numérico

Fonte: Do autor.

Existem algumas ferramentas que auxiliam na verificação dos dados em busca de *outliers*, uma delas é conhecida como Weka que é uma ferramenta de *data*

mining frequentemente utilizada em pesquisas da área. Sua distribuição é feita sob a licença *General Public License*[8] e sua manutenção é feita pela Universidade de Wakaito, na Nova Zelândia (WITTEN; FRANK, 2005, tradução nossa).

No Weka[9], na etapa de pré-processamento, foram analisados todos os atributos, identificando que não existia nenhum atributo com valor nulo e todos eram distintos, não sendo necessário executar filtros para a normalização dos dados.

Após o término da preparação iniciou-se a etapa de modelagem, foram aplicados os algoritmos de data mining *K-means* e *Fuzzy C-means* para o agrupamento dos dados.

4.2.3 Modelagem

A etapa da modelagem, ou *data mining*, tem como objetivo a aplicação dos algoritmos *K-means* e *Fuzzy C-means* nos dados obtidos nas etapas anteriores.

Os algoritmos foram executados em dois softwares, o software Weka versão 3.8.2 foi utilizado para a aplicação do algoritmo *K-means* e a aplicação do *Fuzzy C-means* foi efetuada pelo software *Shell Orion Data Mining Engine*.

A *Shell Orion Data Mining Engine* é uma ferramenta de data mining desenvolvida pelo Grupo de Pesquisa em Inteligência Computacional Aplicada do curso de Ciência da Computação da Universidade do Extremo Sul Catarinense (UNESC).

O algoritmo *Fuzzy C-means* implementado na *Shell Orion* por Ademar Crotti Junior em seu trabalho de conclusão de curso intitulado de O Método de Lógica *Fuzzy* pelos algoritmos Robust C-Prototypes e Unsupervised Robust C-Prototypes para a Tarefa de Clusterização na *Shell Orion Data Mining Engine*.

Na aplicação dos algoritmos, utilizou-se um notebook com sistema operacional Windows 10, processador Intel Core i7-4510U 2.00 GHz 2.60GHz e 8GB de memória RAM DDR3, com placa gráfica dedicada AMD Radeon R7 M265 2GB DDR3.

---

8 General Public License. Informações em (http://www.gnu.org)

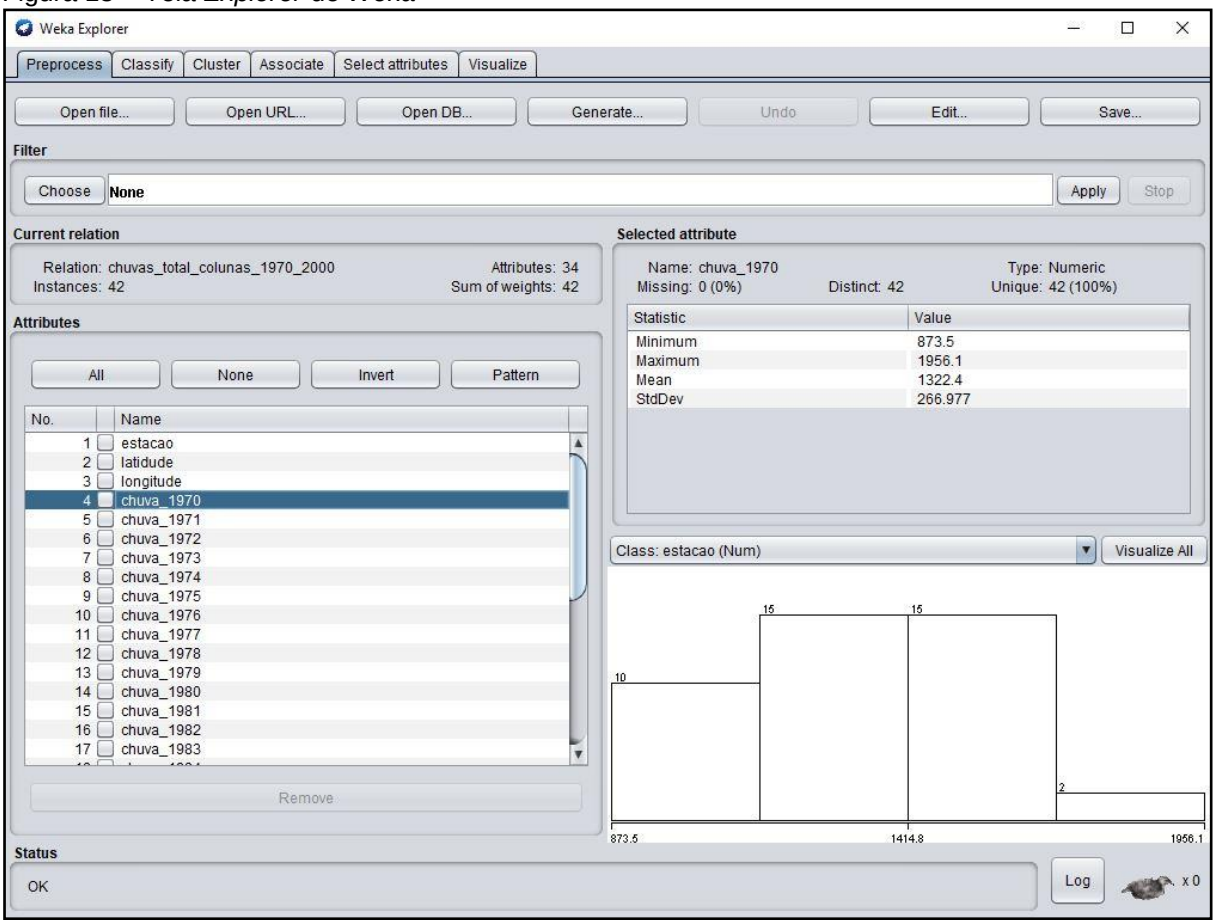
9 *Weka* é disponibilizada gratuitamente em (http://www.cs.waikato.ac.nz/ml/weka/)

4.2.3.1 Aplicação do Algoritmo *K-means*

No software Weka o algoritmo *K-means* é encontrado com o nome de *SimpleKMeans,* para analisar os resultados do algoritmo se faz necessário utilizar a classe chamada de *MakeDensityBasedClusterer* que encapsula o algoritmo *SimpleKMeans* ou qualquer outro algoritmo de agrupamento apresentando no relatório de resultados os índices de validade dos clusters definidos.

Na utilização do Weka não se faz necessário a conexão com banco de dados, pois existe a possibilidade de importar os dados de arquivos CSV. Acessando o Weka encontramos opção *Explorer,* nessa parte do software onde os dados para o agrupamento foram selecionas e posteriormente processados a pelo do *K-means* (figura 13).

Figura 13 – Tela *Explorer* do *Weka*



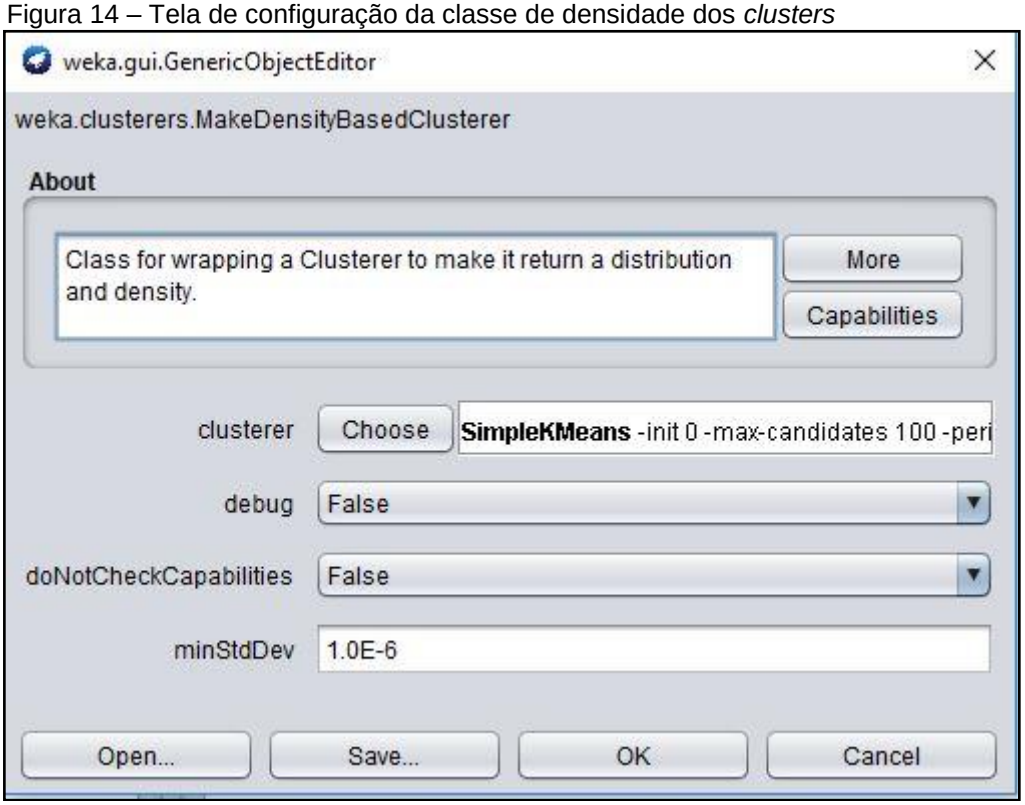
Fonte: Do autor.

Acessando o Weka Explorer é selecionada a base pelo botão *Open file,* após selecionar os dados é possível realizar uma análise verificando se existem

registros nulos por exemplo. Caso a normalização dos dados se faça necessária, existe a possibilidade de aplicar filtros pelo próprio software, nessa pesquisa não foi necessário realizar o procedimento visto que todos os dados estão consistidos.

Após efetuada e seleção dados é aplicada a tarefa de agrupamento pela aba *Cluster*, nela foram ajustados os parâmetros para a aplicação. Após acessar esta etapa é necessário selecionar o método a ser utilizado, em *Choose,* foi utilizado o item *MakeDensityBasedClusterer,* para efetuar a configuração basta clicar sobre o nome que será aberta uma tela para configuração dos parâmetros, como escolher de algoritmo para o agrupamento (figura 14).

Figura 14 – Tela de configuração da classe de densidade dos *clusters*

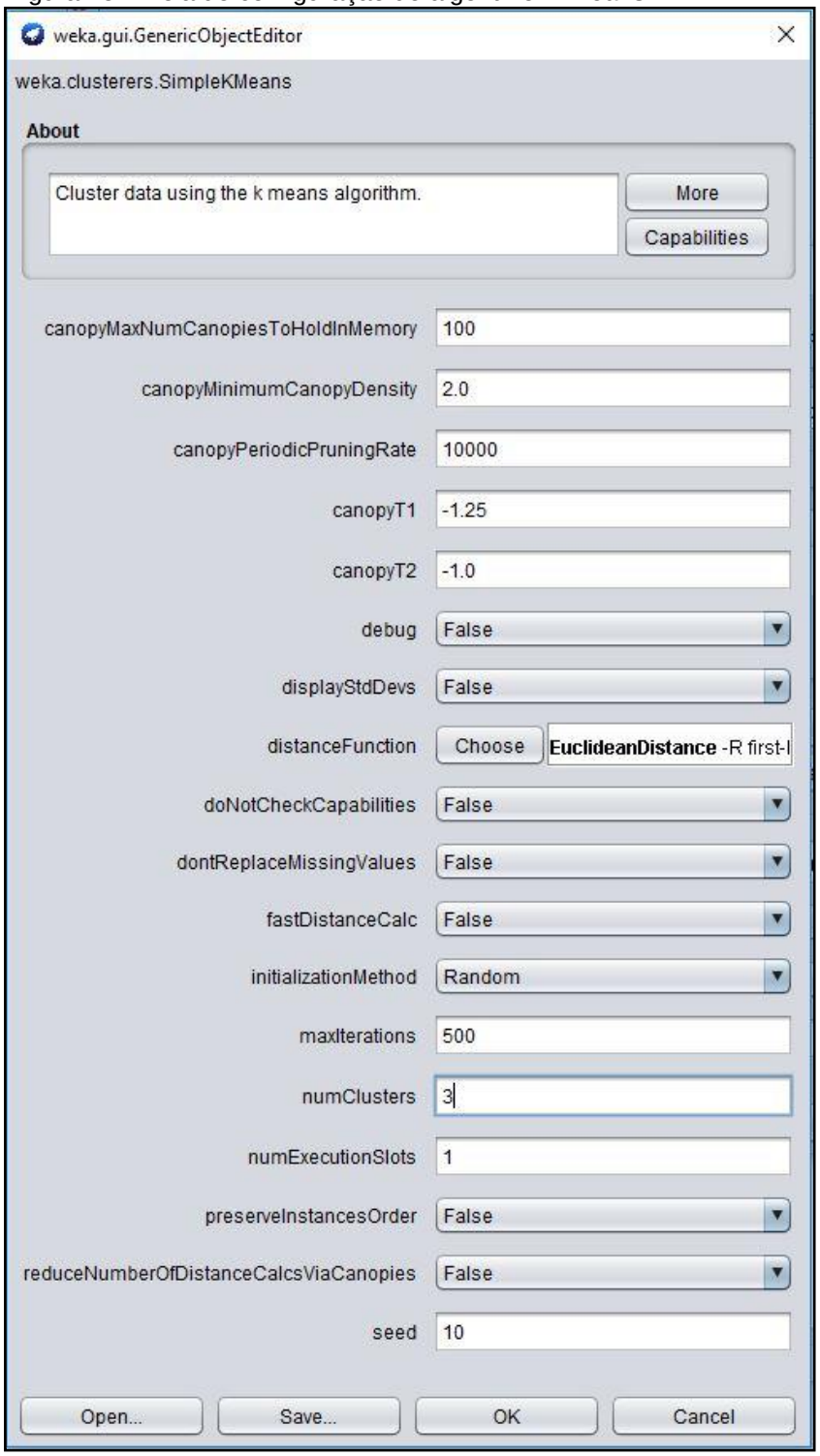


Fonte: Do autor.

Na execução do algoritmo é necessário informar os parâmetros a serem respeitados e atributos serão utilizados. Uma breve explicação dos parâmetros existentes: *clusterer*, algoritmo de agrupamento a ser utilizado; *debug*, definido como true, o algoritmo pode gerar informações adicionais para o console do software; *doNotCheckCapabilities*, os recursos do grupo não serão verificados antes que o agrupamento seja construído; *minStdDev*, definir desvio padrão mínimo permitido.

Após a configuração da classe que envolve o *SimpleKMeans* se faz necessário configurar os parâmetros do algoritmo *K-means* (figura 15).

Figura 15 – Tela de configuração do algoritmo *K-means*



Fonte: Do autor.

Na execução do algoritmo foram configurados os seguintes parâmetros para a execução do algoritmo, sendo que os parâmetros que no nome possuem a nomenclatura *Canopy* não se aplicam ao algoritmo *K-means*: *distanceFunction*: Foram aplicados o cálculo de distância Euclidiana e Manhattan; *numClusters*: 1 até 8; *InitializationMethod*: Foram aplicados testes com *Random*, *K-means++* e *Farthest*

First; maxIterations: Número máximo de iterações foram 500; *numExecutionSlots*: Informado o valor 1, utilizando assim um cpu/núcleo do processador para a execução do algoritmo; *preserveInstancesOrder*: Se marcado como false o algoritmo não irá preservar a ordem das instâncias, ajustando para melhores resultados (tabela 4).

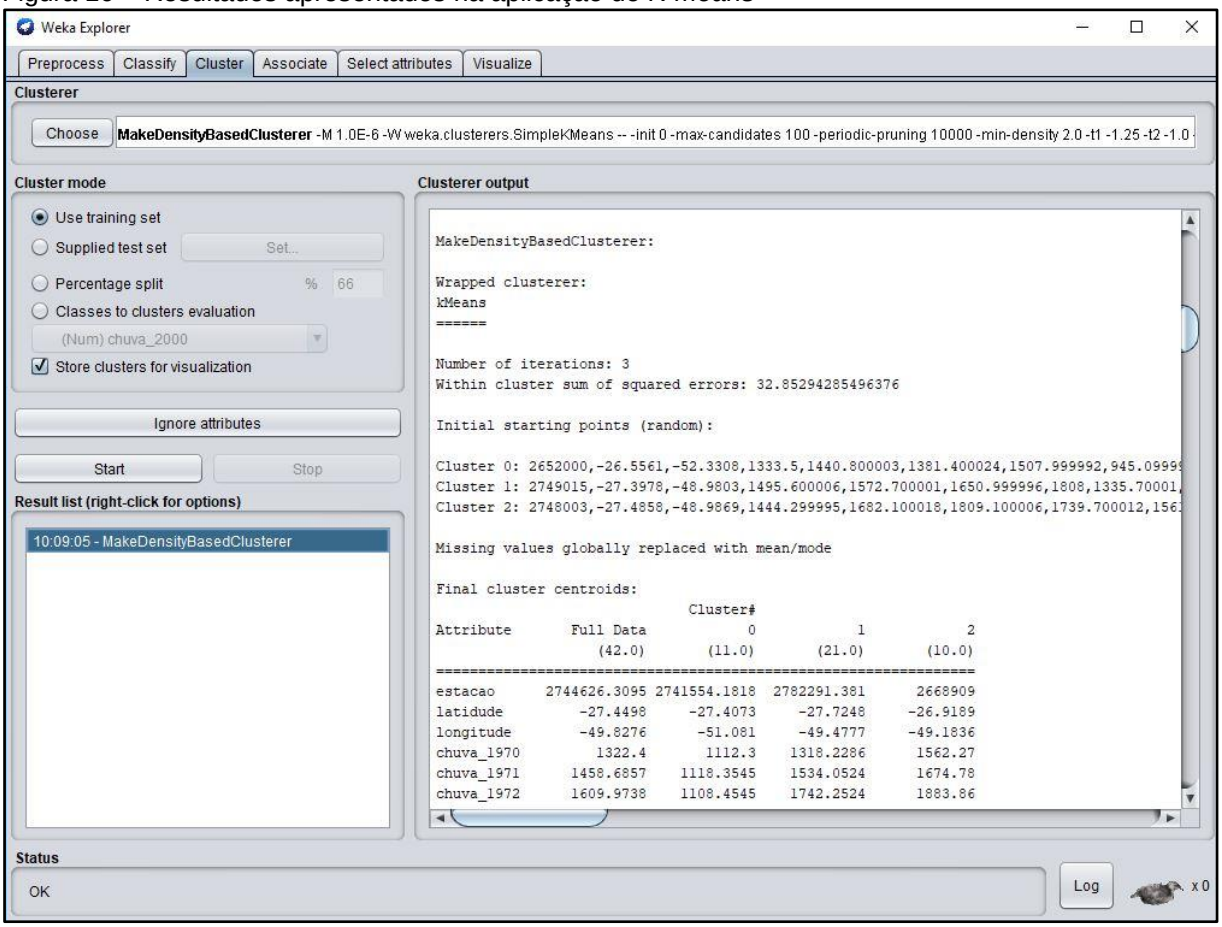
Tabela 4 – Experimentos realizados com o algoritmo *K-means*

Experimento	Clusters	Distância	Inicialização	iterações	Slots
1	1	Euclidiana	*Radom*	500	1
2	1	Euclidiana	*K-means++*	500	1
3	1	Euclidiana	*Farthest First*	500	1
4	1	Manhattan	*Radom*	500	1
5	1	Manhattan	*K-means++*	500	1
6	1	Manhattan	*Farthest First*	500	1
7	2	Euclidiana	*Radom*	500	1
8	2	Euclidiana	*K-means++*	500	1
9	2	Euclidiana	*Farthest First*	500	1
10	2	Manhattan	*Radom*	500	1
11	2	Manhattan	*K-means++*	500	1
12	2	Manhattan	*Farthest First*	500	1
13	3	Euclidiana	*Radom*	500	1
14	3	Euclidiana	*K-means++*	500	1
15	3	Euclidiana	*Farthest First*	500	1
16	3	Manhattan	*Radom*	500	1
17	3	Manhattan	*K-means++*	500	1
18	3	Manhattan	*Farthest First*	500	1
19	4	Euclidiana	*Radom*	500	1
20	4	Euclidiana	*K-means++*	500	1
21	4	Euclidiana	*Farthest First*	500	1
22	4	Manhattan	*Radom*	500	1
23	4	Manhattan	*K-means++*	500	1
24	4	Manhattan	*Farthest First*	500	1
25	5	Euclidiana	*Radom*	500	1
26	5	Euclidiana	*K-means++*	500	1
27	5	Euclidiana	*Farthest First*	500	1
28	5	Manhattan	*Radom*	500	1
29	5	Manhattan	*K-means++*	500	1
30	5	Manhattan	*Farthest First*	500	1
31	6	Euclidiana	*Radom*	500	1
32	6	Euclidiana	*K-means++*	500	1
34	6	Euclidiana	*Farthest First*	500	1
35	6	Manhattan	*Radom*	500	1
36	6	Manhattan	*K-means++*	500	1
37	6	Manhattan	*Farthest First*	500	1
38	7	Euclidiana	*Radom*	500	1
39	7	Euclidiana	*K-means++*	500	1
40	7	Euclidiana	*Farthest First*	500	1
41	7	Manhattan	*Radom*	500	1
42	7	Manhattan	*K-means++*	500	1
43	7	Manhattan	*Farthest First*	500	1
44	8	Euclidiana	*Radom*	500	1
45	8	Euclidiana	*K-means++*	500	1
46	8	Euclidiana	*Farthest First*	500	1
47	8	Manhattan	*Radom*	500	1
48	8	Manhattan	*K-means++*	500	1
49	8	Manhattan	*Farthest First*	500	1

Fonte: Do autor.

Após a execução das 49 possibilidades, que visam descobrir qual combinação apresentará os melhores resultados, foram realizadas as devidas análises conforme os índices de qualidades disponíveis nos resultados apresentados (figura 16). Nesta pesquisa não foram abordadas as distâncias Mahalanobis e Minkowski visto que os trabalhos correlatos aplicaram somente as distâncias Euclidiana e Manhattan.

Figura 16 – Resultados apresentados na aplicação do *K-means*

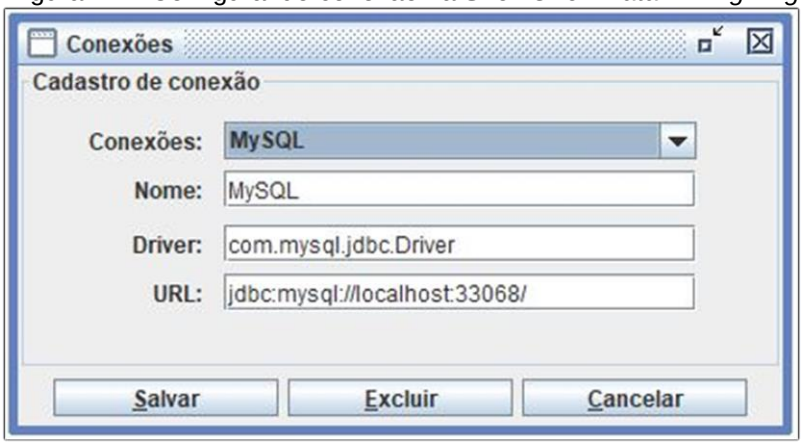


Fonte: Do autor

4.2.3.2 Aplicação do Algoritmo *Fuzzy C-means*

Para aplicação do algoritmo *Fuzzy C-means* foi utilizada a ferramenta *Shell Orion Data Mining Engine*, o primeiro passo a ser executado na ferramenta é a configuração da conexão com o banco de dados utilizado (figura 17).

Figura 17 – Configurando conexão na *Shell Orion Data Mining Engine*



Fonte: Do autor

As configurações que podem ser informadas nesta tela são: conexões: tipo de SGBD que será conectado; nome**:** Nome da conexão; driver: driver a ser utilizado para conexão do SGBD; URL: caminho e porta para o SGDB.

Após a configuração da conexão é necessário efetuar a conexão com a base de dados, informando o nome da base, nome da tabela, usuário e senha do banco de dados (figura 18).

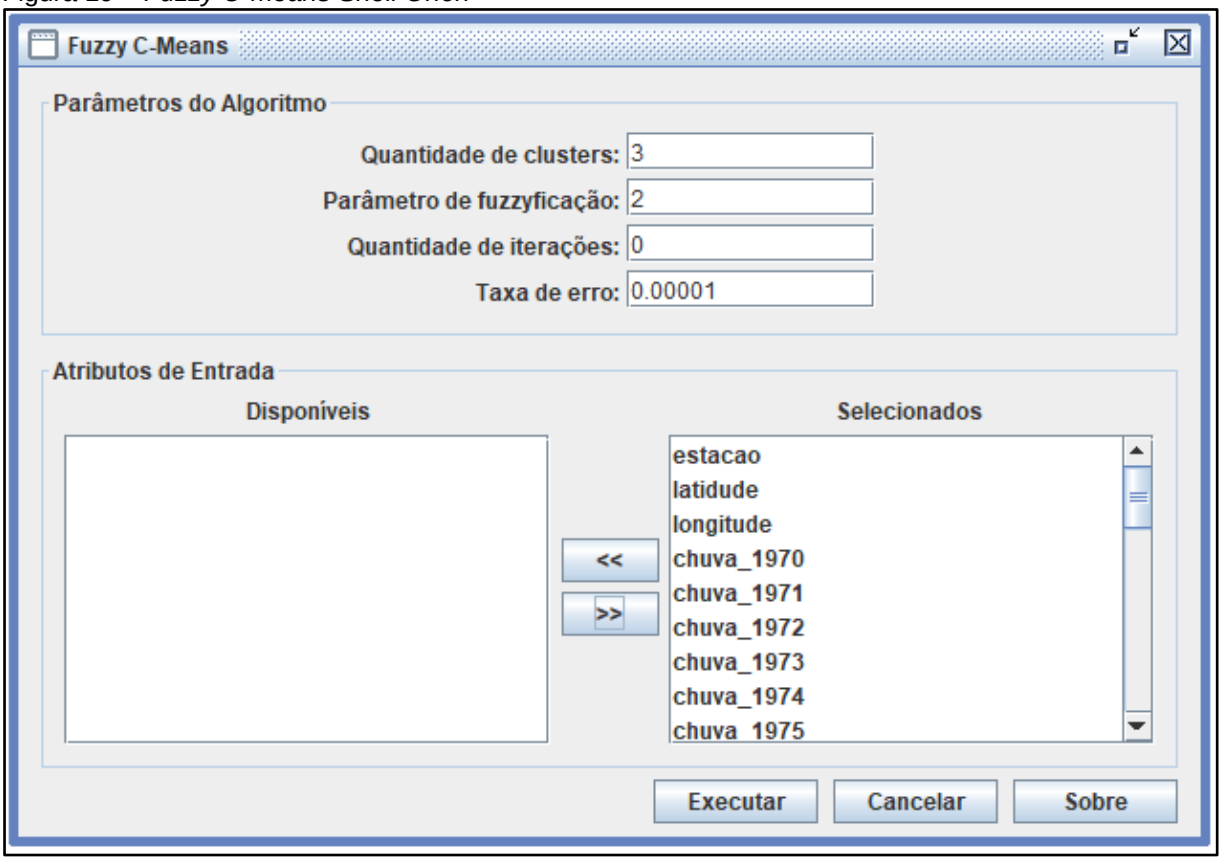
Figura 18 – Conectando a uma base na *Shell Orion*

Fonte: Do autor

As configurações que podem ser informadas nesta tela são: conexão: tipo de SGBD que será conectado; base de dados: base de dados do SGBD, quando for necessário trabalhar com várias tabelas de bases de dados; usuário: usuário de acesso ao SGBD; senha: senha de acesso ao SGBD; tabela: tabela que será utilizada pelo algoritmo.

Após realizadas as etapas anteriores acessou-se o algoritmo *Fuzzy C-means* disponível nas tarefas de agrupamento *fuzzy* na *Shell Orion Data Mining Engine* (figura 19).

Figura 19 – *Fuzzy C-means Shell Orion*



Fonte: Do autor.

Na execução do algoritmo informou-se os parâmetros a serem respeitados e os atributos utilizados (Estação, Latitude, Longitude e chuvas 1970 até chuvas 2000), a quantidade de *clusters* que o algoritmo deverá encontrar, definir o grau de fuzzyficação entre elementos e clusters, a quantidade máxima de ciclos que o algoritmo irá executar e erro aceitável na execução do algoritmo, indicando a parada do algoritmo.

Como o objetivo desta pesquisa é identificar zonas pluviometricamente homogêneas de acordo com as medidas de qualidades encontradas foram realizadas várias execuções do algoritmo com as configurações de quantidade de *clusters* de 2 até 8, parâmetro de fuzzyficação 2, quantidade de iterações ilimitada (0), taxa de erro 0,00001 e todos os atritos de entrada disponíveis (tabela 5).

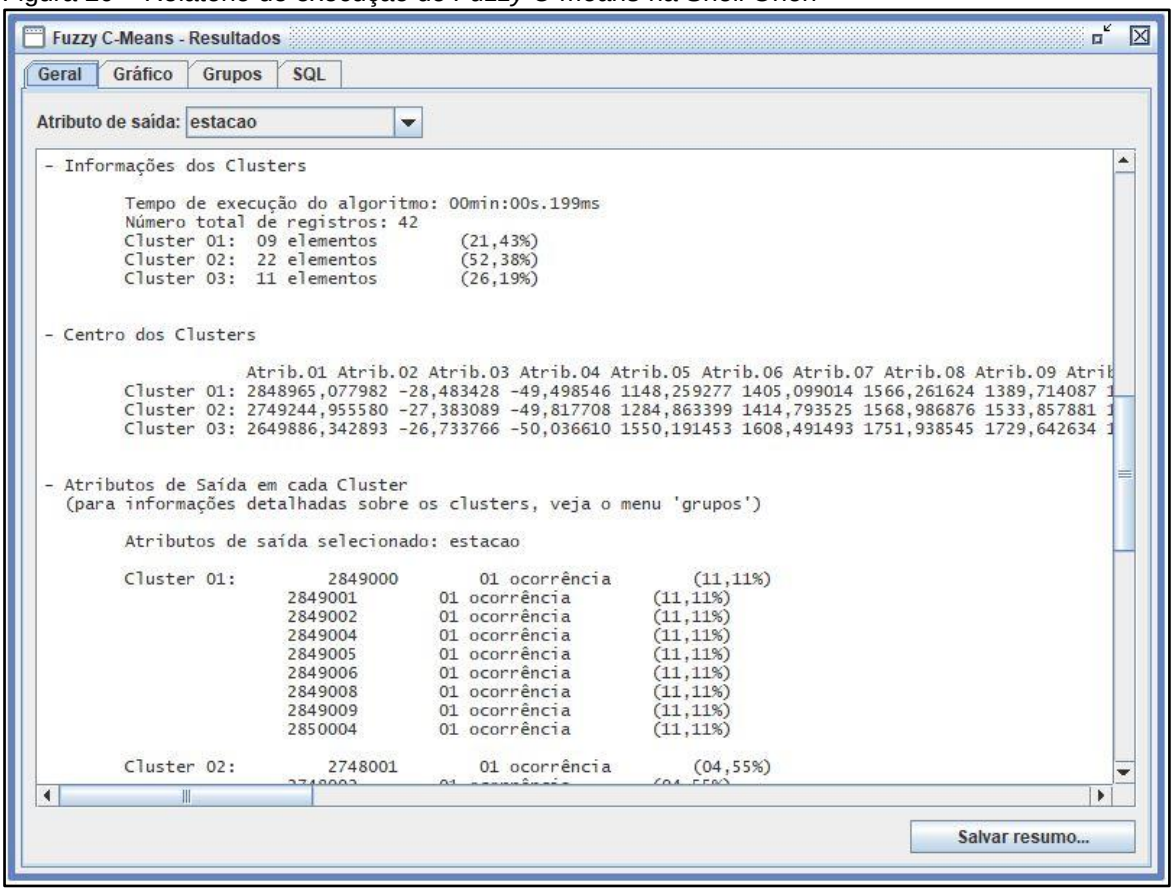
Tabela 5 – Configurações utilizadas na aplicação do *Fuzzy C-means*

Experimento	Quantidade de *clusters*	Fuzzyficação	Iterações	Taxa de erro	Atributos de entrada
1	2	2	0	0,00001	Todos
2	3	2	0	0,00001	Todos
3	4	2	0	0,00001	Todos
4	5	2	0	0,00001	Todos
5	6	2	0	0,00001	Todos
6	7	2	0	0,00001	Todos
7	8	2	0	0,00001	Todos

Fonte: Do autor.

Após a execução a Shell Orion apresenta os resultados em forma de relatório, detalhando os grupos encontrados, centróides e índices de validade (figura 20).

Figura 20 – Relatório de execução do *Fuzzy C-means* na Shell Orion



Fonte: Do autor.

Por meio da distância Euclidiana o algoritmo calcula a separação dos centróides. Após a execução de todas as possibilidades foram realizadas as devidas análises conforme os índices de qualidade disponíveis.

4.3 RESULTADOS OBTIDOS E DISCUSSÃO

Ao avaliar os resultados obtidos de cada algoritmo é visto que ambos apresentaram o mesmo número de grupos de acordo com as medidas aplicadas, porém as disposições das estações ficaram diferentes, no algoritmo *Fuzzy C-means* as estações ficaram visivelmente homogêneas de acordo com suas regiões, isso se deve pelo grau de pertinência que cada objeto possuí, ou seja, mesmo possuindo valores que se adequem a outros centroides o objeto pertencerá ao grupo que possuir mais similaridade em todos os atributos, indo de acordo com a sua lógica difusa.

O algoritmo *K-means* realizou a tarefa de agrupamento de forma mais abrupta nas questões das regiões, o que ocasionou que alguns objetos dos grupos ficassem dentro de outros grupos com uma distância muito aproxima, sendo assim não se demonstrando satisfatório na espacialização dos grupos pelo estado de Santa Catarina.

Desta maneira, pode-se concluir que o algoritmo *Fuzzy C-means* consegue obter resultados mais satisfatórios em questões de espacialização de objetos e média de precipitação, demonstrando capacidade em caracterizar o Estado de Santa Catarina, conforme os dados coletados, em 3 regiões pluviometricamente homogêneas.

4.3.1 Resultados do Algoritmo *K-means*

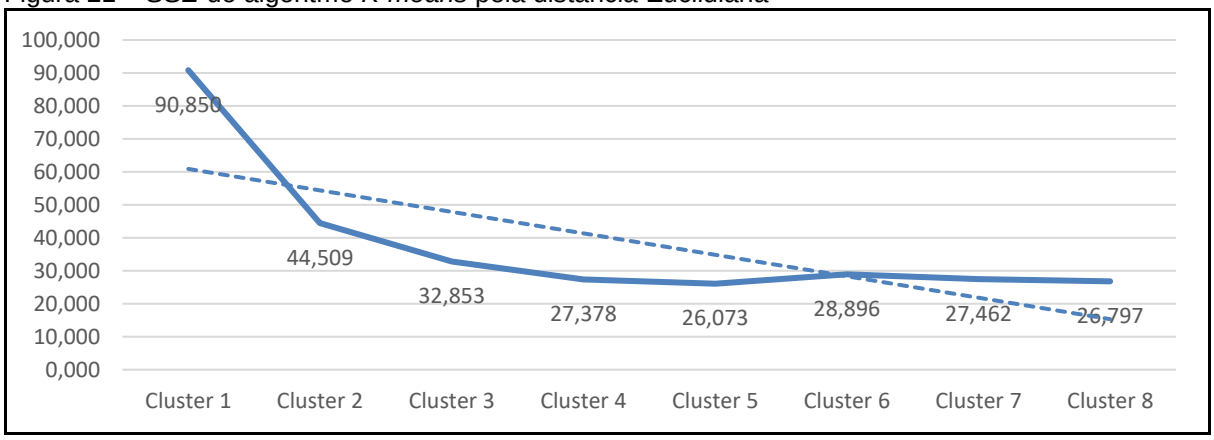
Após a análise dos resultados foi possível identificar que alterar o método de inicialização do algoritmo entre *Random*, *K-means++* e *Farthest First* não modificavam os resultados, os índices de qualidade possuíam o mesmo padrão, desta forma os resultados apresentados foram iniciados com o método de inicialização *Random*. A quantidade de iterações do algoritmo não se mostrou um índice de grande valor pois as execuções retornavam valores abaixo de 10, logo o tempo de execução também ficou sempre abaixo de 1 segundo.

Nos resultados gerados pelo algoritmo utilizando o cálculo de distância Euclidiana é possível identificar pelo índice de qualidade *SSE*, definido como a soma da distância quadrada entre cada membro do *cluster* e seu centróide, qual é o número aceitável de *clusters* para o conjunto de dados. A análise desses resultados pode ser

feita por dois métodos que seriam o método de *Elbow* ou a verificação do percentual de mudança dos *clusters*.

Na figura 21 possuímos os valores do índice SSE de cada aplicação, é possível identificar que o erro diminui à medida que a quantidade de *clusters* aumenta, isso ocorre porque quando o número de *cluster* aumenta, eles ficam menores, então a distorção também é menor.

Figura 21 – *SSE* do algoritmo *K-means* pela distância Euclidiana



Fonte: Do autor.

A análise dos resultados pelo método de *Elbow*, também chamado de método do cotovelo, é feita de forma visual em busca de um ponto onde os resultados se modifiquem em forma de braço ou cotovelo, sendo assim, o melhor número de *clusters* para esses dados seriam 3.

Uma outra maneira analisar o melhor grupo pela *SSE* seria efetuando o cálculo do percentual de mudança entre os grupos, aquele que possuir o maior percentual de mudança é considerado como o melhor, na tabela 6 é confirmado que o agrupamento com 3 centroides possuí o maior percentual de mudança.

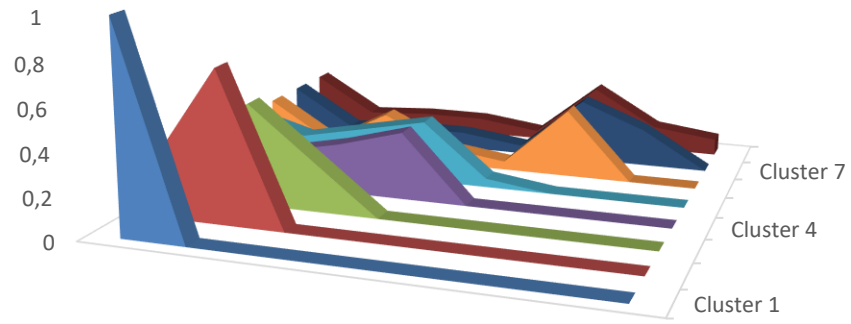
Tabela 6 – Percentual de mudança dos clusters criados pelo algoritmo *K-means* distancia euclidiana

Número de grupos	SSE	Percentual de mudança
2	44,509	-
3	**32,853**	**35,48016%**
4	27,378	19,99615%
5	26,073	5,005825%
6	28,896	-9,769028%
7	27,462	5,220206%
8	26,797	2,483628%

Fonte: Do autor.

Analisando os valores de *Prior Probability*, também é possível identificar valores aceitáveis para o agrupamento com 3 centróides, conforme demonstrado na figura 22. Nesses casos quanto maior a semelhança dos *clusters* melhor o número de agrupamento gerados.

Figura 22 – *Prior probability K-means* pela distância Euclidiana



	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Cluster 1	1							
Cluster 2	0,2727	0,7273						
Cluster 3	0,2667	0,4889	0,2444					
Cluster 4	0,2609	0,1739	0,2391	0,3261				
Cluster 5	0,2553	0,1702	0,2340	0,2979	0,0426			
Cluster 6	0,2500	0,0625	0,2292	0,0833	0,0417	0,3333		
Cluster 7	0,2449	0,0612	0,0816	0,0816	0,0408	0,3061	0,1837	
Cluster 8	0,2400	0,0600	0,0800	0,0800	0,0400	0,3000	0,1200	0,0800

Fonte: Do autor.

Desta forma, pode-se dizer que o número ideal de grupos para os dados pelo algoritmo *K-means* utilizando a distância Euclidiana são 3 grupos, pois a distribuição dos objetos entre os *clusters* possui maior semelhança.

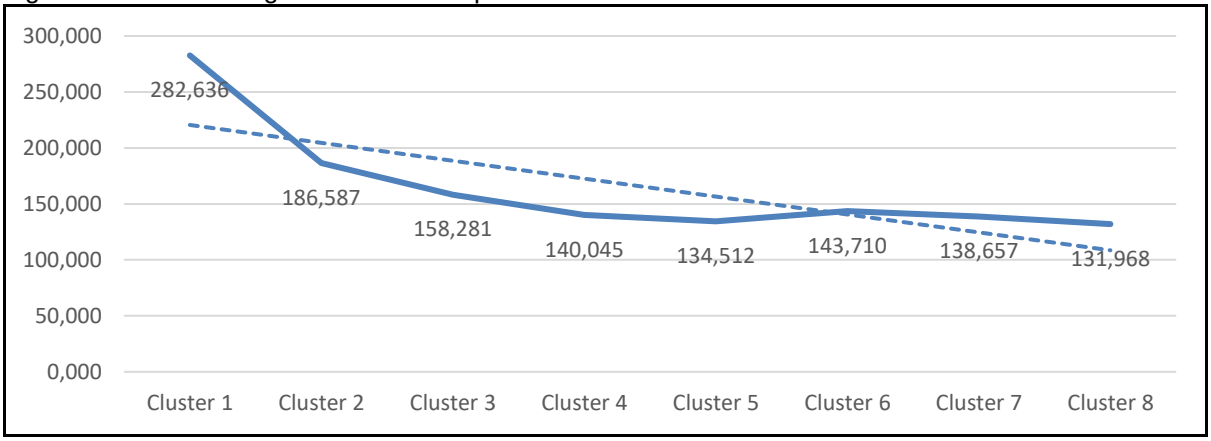
Tabela 7 – Percentual de mudança dos clusters criados pelo algoritmo *K-means* distancia Manhattan

Número de grupos	SSE	Percentual de mudança
2	186,587	-
3	**158,281**	**17,883394%**
4	140,045	13,021755%
5	134,512	4,1135861%
6	143,710	-6,4007317%
7	138,657	3,6444260%
8	131,968	5,0690179%

Fonte: Do autor.

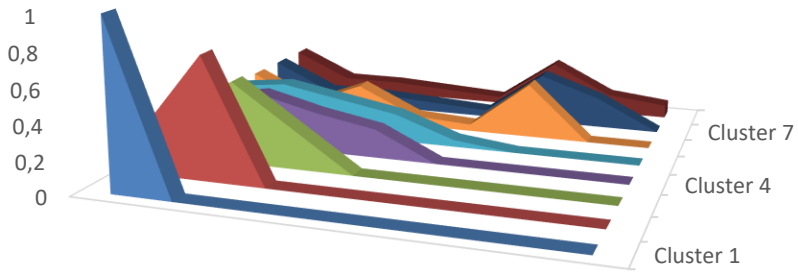
Nos resultados gerados pelo *K-means* utilizando a distância de Manhattan, somente a SSE resultou em valores diferentes, porém apresentou que o número de grupos para os dados são 3, mesmo total apresentado pela distância Euclidiana. As análises foram feitas de acordo com o método de *Elbow*, analisando-se juntamente com o percentual de mudança, e *Prior Probability*[10]*.* Na figura 23 pode-se visualizar os valores da SSE com a distância Manhattan.

Figura 23 – SSE do algoritmo *K-means* pela distância Manhattan



Fonte: Do autor.

Figura 24 – *Prior Probability K-means* pela distância Manhattan



	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Cluster 1	1							
Cluster 2	0,2727	0,7273						
Cluster 3	0,2667	0,4889	0,2444					
Cluster 4	0,2609	0,3261	0,2391	0,1739				
Cluster 5	0,2553	0,2979	0,2340	0,1702	0,0426			
Cluster 6	0,2500	0,0833	0,2292	0,0625	0,0417	0,3333		
Cluster 7	0,2449	0,0816	0,0816	0,0612	0,0408	0,3061	0,1837	
Cluster 8	0,2400	0,0800	0,0800	0,0600	0,0400	0,3000	0,1200	0,0800

Fonte: Do autor.

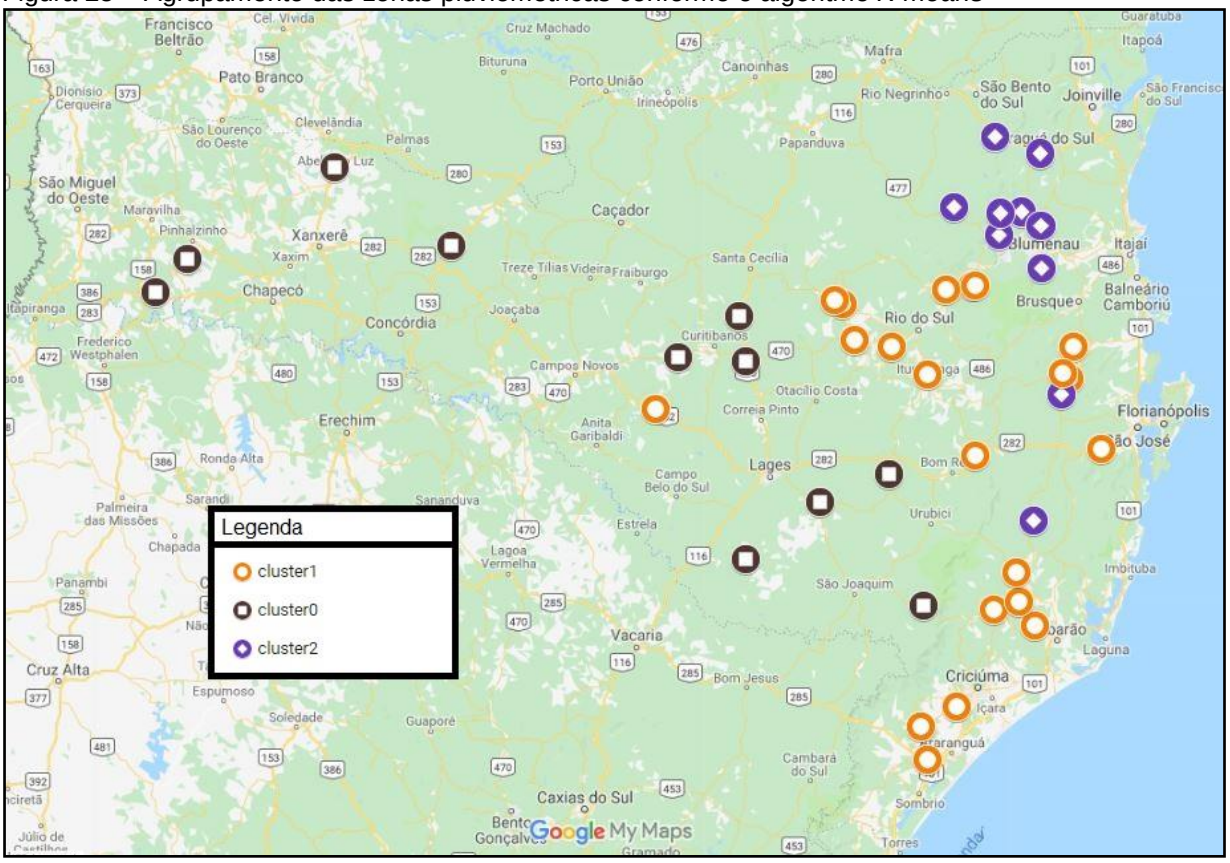
[10] *Prior Probability* distribuição relativa dos valores nos conjuntos apresentados

Os valores de *Prior Probability* (figura 24), ficaram idênticos aos valores dos resultados obtidos pela aplicação da distância Euclidiana.

Visto que os resultados da SSE Euclidiana mostraram valores menores que o da SSE Manhattan e a *Prior Probability* apresentou os mesmos valores em ambas as medidas. Foi escolhida a medida Euclidiana para se analisar os resultados.

Realizando a análise espacial dos *clusters* é possível identificar as zonas pluviometricamente homogêneas do Estado de Santa Catarina de acordo com o algoritmo *K-means* (figura 25).

Figura 25 – Agrupamento das zonas pluviométricas conforme o algoritmo *K-means*



Fonte: Do autor.

O *cluster* 0 abrange as mesorregiões do Oeste Catarinense e Serrana, o *cluster* 1 é contempla as estações da região Serrana, Sul, Vale do Itajaí e Grande Florianópolis, por último, o *cluster* 2 é representado pelas regiões Norte e Grande Florianópolis.

Na tabela 8 é feita a identificação dos agrupamentos encontrados, informando a quantidade de elementos em cada grupo e sua porcentagem de acordo com o total de elementos.

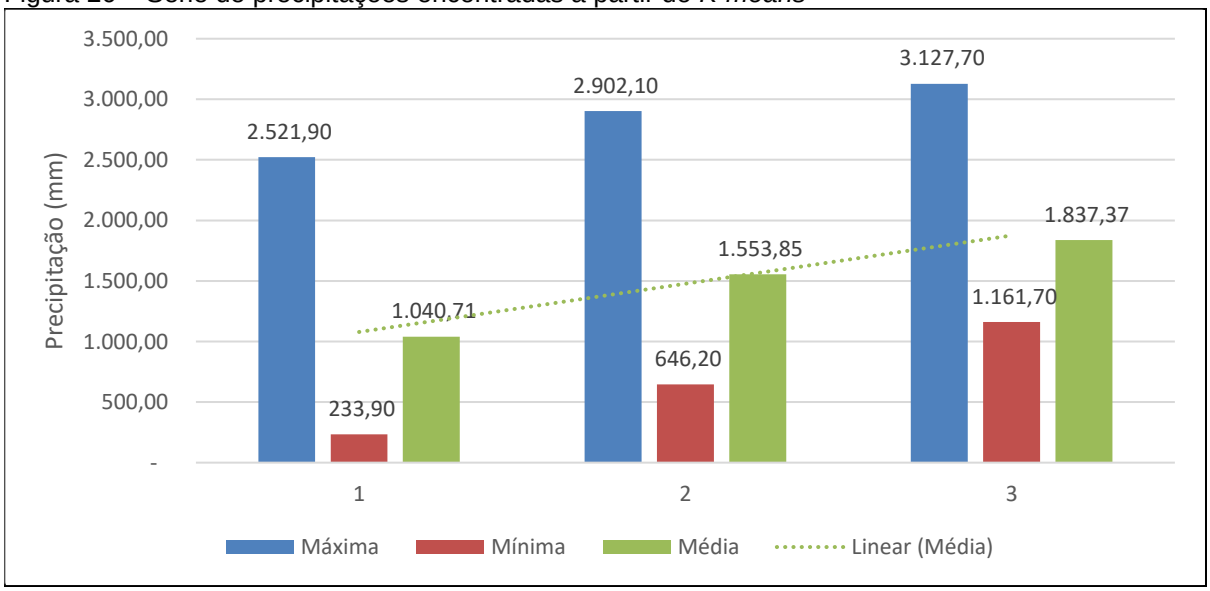
Tabela 8 – Clusters encontrados a partir do algoritmo *K-means*

Grupos	Quantidade de elementos	Porcentagem dos elementos
0	11	26%
1	21	50%
2	10	24%

Fonte: Do autor.

Analisando a série de precipitações dos três clusters encontrados, verificando as máximas, mínimas e a média atingida nesse período, é possível identificar que os valores possuem uma tendência crescente, não apresentando homogeneidade no índice pluviométrico das estações que cada *clusters* representa (figura 26). Na tabela 9, é possível identificar os anos das máximas e mínimas encontradas.

Figura 26 – Série de precipitações encontradas a partir do *K-means*



Fonte: Do autor.

Tabela 9 – Detalhamentos das máximas e mínimas encontradas no *K-means*

Cluster	Máxima(mm)	Ano Máx.	Mínima(mm)	Ano Mín.
0	2.521,90	1983	233,90	1994
1	2.902,10	1983	646,20	1999
2	3.127,70	1983	1.161,70	1988

Fonte: Do autor.

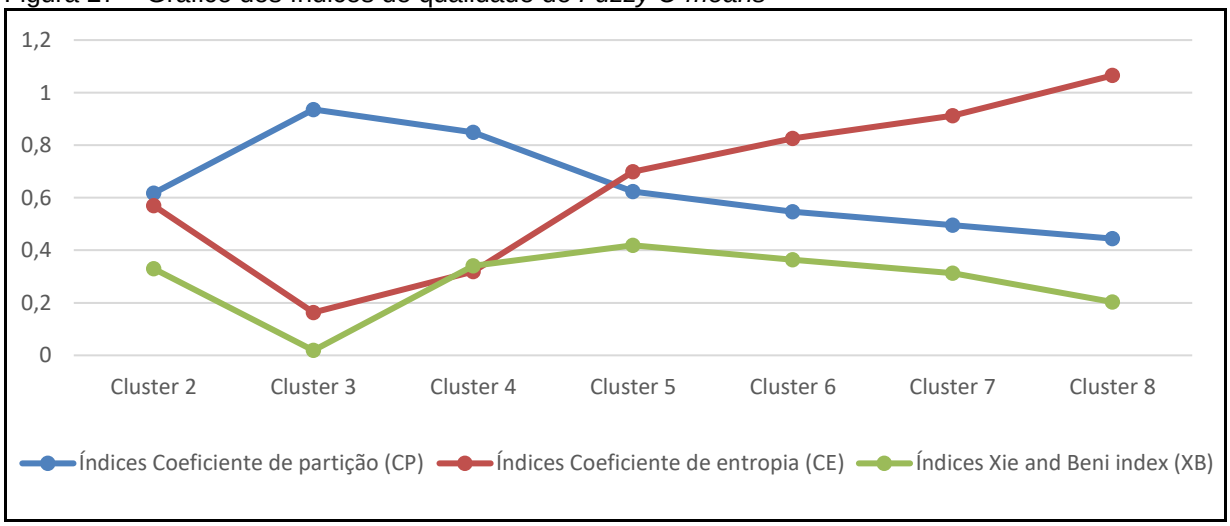
Todos os *clusters* apresentaram o ano de 1983 nas máximas, condizendo com o fato histórico de que em 1983 o estado foi afetado por índices pluviométricos altos, acarretando em várias enchentes (ZIMMERMANN, 2011).

4.3.2 Resultados do Algoritmo *Fuzzy C-means*

Os índices de qualidade analisados no algoritmo *Fuzzy C-means* coeficiente de partição, de entropia e *Xie and Beni* auxiliaram na avaliação dos resultados obtidos. Coeficiente de partição é uma função maximizadora, sendo assim quanto maior o valor mais satisfatório é o resultado, os outros dois índices são minimizadores, quanto menor o valor mais convincente é o agrupamento.

Na figura 27 é possível visualizar os índices obtidos de acordo com os agrupamentos realizados.

Figura 27 – Gráfico dos índices de qualidade do *Fuzzy C-means*



Fonte: Do autor.

De acordo com os resultados, e a função de cada medida, é possível identificar que o melhor número de grupos para os dados desta pesquisa é 3, a partir desse grupo o Coeficiente de participação tende a diminuir seu valor, o coeficiente de entropia e *Xie and Beni* a se elevar, mesmo que o *Xie and Beni* tenha uma tendência após o grupo de diminuir, as outras medidas continuam a se distanciar de valores aceitáveis. Na tabela 10 é possível visualizar detalhadamente os valores dos índices de validade encontrados, onde o máximo coeficiente de partição e os mínimos de entropia e *Xie and Beni* se encontra no grupo 3.

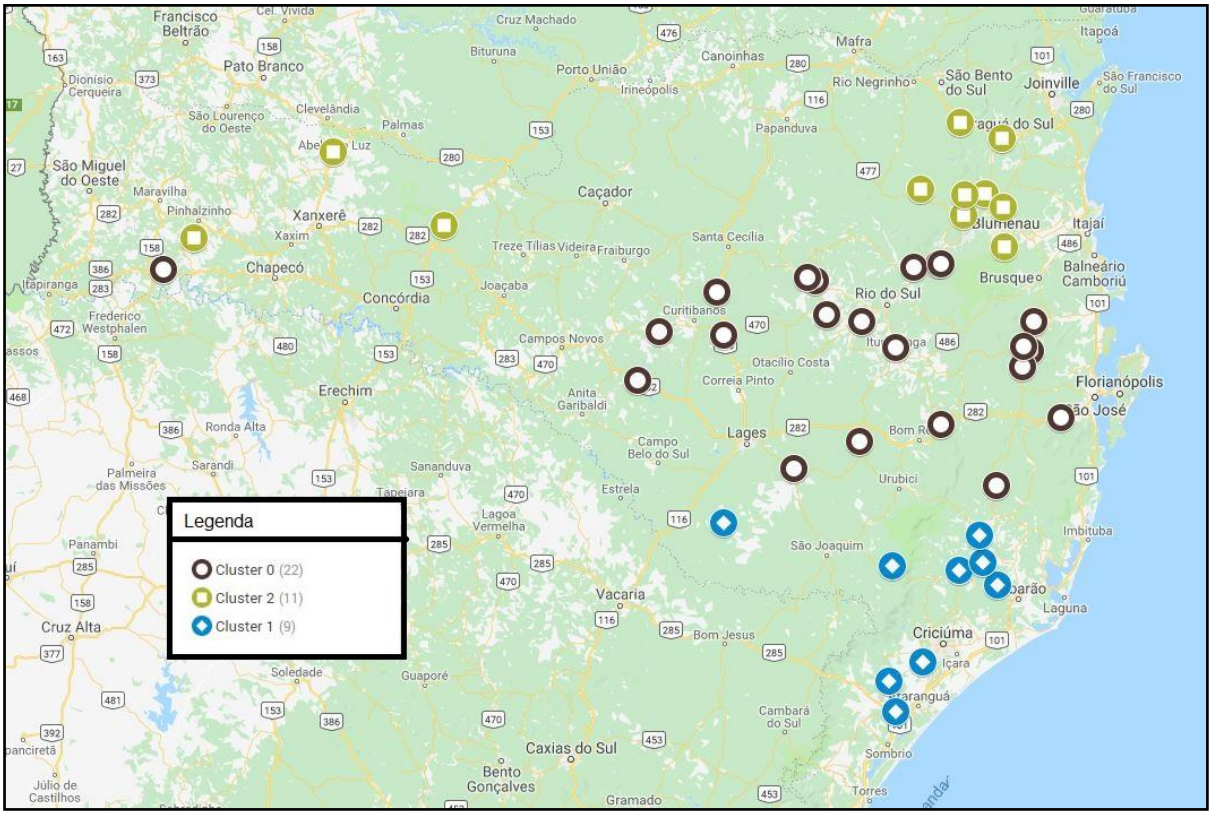
Realizando a análise espacial dos *clusters* foi possível identificar as zonas pluviometricamente homogêneas do Estado de Santa Catarina de acordo com as medidas de qualidade aplicadas (figura 29).

Tabela 10 – índices de validade do algoritmo *Fuzzy C-means*

Grupos	Coeficiente de partição (CP)	Coeficiente de entropia (CE)	*Xie and Beni* (XB)
2	0,617907899	0,569592671	0,330158909
3	**0,935956106**	**0,163320271**	**0,018642285**
4	0,849288702	0,319206485	0,341061615
5	0,624079240	0,698542724	0,418871475
6	0,547111694	0,825904718	0,363527001
7	0,495077763	0,912552506	0,312767318
8	0,444438686	1,065119408	0,202946553

Fonte: Do autor.

Figura 28 – Agrupamento das zonas pluviometricamente homogêneas de acordo com o algoritmo *Fuzzy C-means.*



Fonte: Do autor.

O *cluster* 0 abrange as mesorregiões do Grande Florianópolis, Serrana, Vale do Itajaí e Oeste, já o *cluster* 1 é contempla as estações da região Serrana, Sul, por último, o *cluster* 2 é representado pelas regiões Norte e Oeste.

Na tabela 10 tem-se a identificação dos agrupamentos encontrados, informando a quantidade de elementos em cada grupo e sua porcentagem de acordo com o total de elementos.

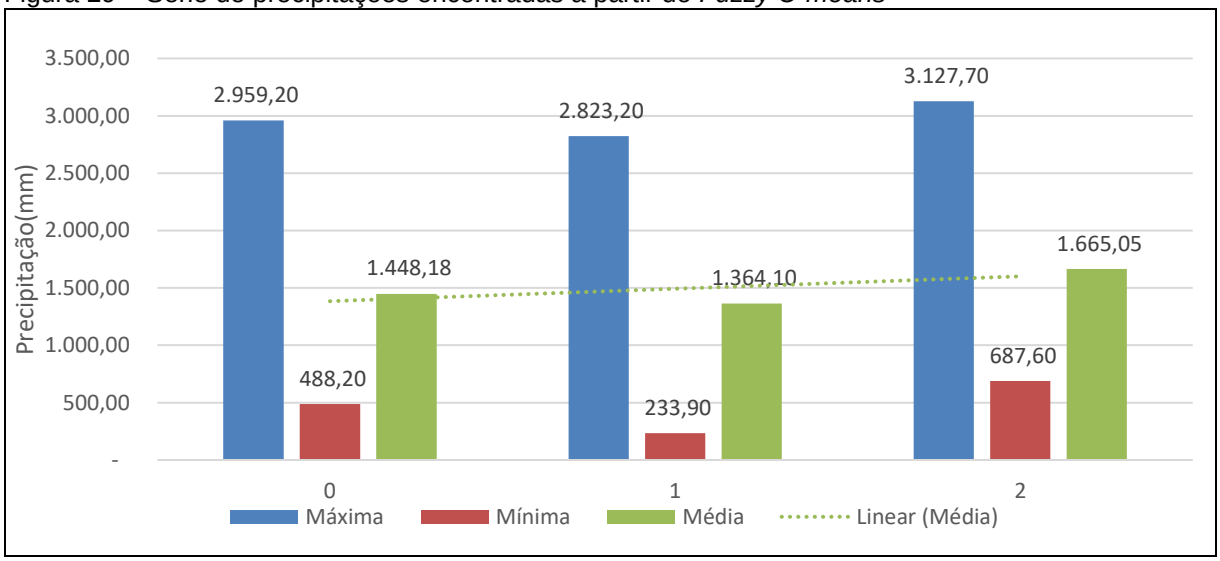
Tabela 11 – Clusters encontrados a partir do algoritmo *Fuzzy C-means*

Grupos	Quantidade de elementos	Porcentagem dos elementos
0	22	52,38%
1	9	21,43%
2	11	26,19%

Fonte: Do autor.

Analisando a série de precipitações pelo *Fuzzy C-means* são identificados três *clusters*, verificando as máximas, mínimas e a média atingida nesse período, é possível identificar que o valor da média possui uma regularidade no índice pluviométrico das estações que cada *clusters* representa (figura 29). Na tabela 12, pode-se identificar os anos das máximas e mínimas encontradas.

Figura 29 – Série de precipitações encontradas a partir do *Fuzzy C-means*



Fonte: Do autor.

Tabela 12 – Detalhamentos das máximas e mínimas encontradas a partir do *Fuzzy C-means*

Cluster	Máxima(mm)	Ano Máx.	Mínima(mm)	Ano Mín.
0	2.959,20	1983	488,20	1985
1	2.823,20	1983	233,90	1994
2	3.127,70	1983	687,60	1988

Fonte: Do autor.

Assim como no *K-means*, na aplicação *Fuzzy C-means* todos os *clusters* apresentaram o ano de 1983 nas máximas, condizendo com o fato histórico de que em 1983 o estado foi afetado por índices pluviométricos altos, acarretando em várias enchentes (ZIMMERMANN, 2011).

4.3.3 Discussão dos Resultados

Avaliando os resultados obtidos de cada algoritmo empregado nesta pesquisa, pode-se verificar que ambos apresentaram os melhores resultados para o mesmo número de grupos de acordo com as medidas aplicadas, porém as disposições das estações e as médias pluviométricas apresentaram diferenças. No algoritmo *Fuzzy C-means* as estações ficaram visivelmente homogêneas de acordo com suas regiões e média de precipitação, isso se deve ao grau de pertinência que cada objeto possui. Portanto, mesmo possuindo valores que se adequem a outros centroides, de acordo com a lógica *fuzzy*, o objeto pertence ao grupo que apresentar maior similaridade em todos os atributos.

O algoritmo *K-means* realizou a tarefa de agrupamento de forma mais abrupta nos dados utilizados, ocasionando que alguns objetos dos grupos ficassem dentro de outros grupos com uma distância muito próxima. Com isso, não se mostrou satisfatório na espacialização dos grupos homogêneos pelo estado de Santa Catarina, no entanto, para identificação de grupos heterógenos o algoritmo se mostrou satisfatório ao serem analisadas as médias pluviométricas de cada *cluster* (1.040,71mm, 1.553,85mm, 1.837,37mm).

Assim, os resultados desta pesquisa apontaram que o algoritmo *Fuzzy C-means* consegue obter resultados mais satisfatórios na questão de obter *clusters* homogêneos, demonstrando capacidade em caracterizar o Estado de Santa Catarina, conforme os dados coletados, em três regiões pluviometricamente homogêneas, visto que as médias de precipitação encontradas nos clusters foram 1.448,18mm para o *cluster* 0, 1.364,10mm para o *cluster* 1 e 1.665,05mm para o *cluster* 2, não apresentando diferenças significativas em seus valores, ao contrário das médias apresentadas pelo *K-means*, que apresentaram heterogeneidade, corroborando com os Dourado, Oliveira e Avila (2012, 2013) onde o algoritmo, aplicado a dados pluviométricos da Bahia, apresentou *clusters* com maiores diferenças nas médias.

Em relação aos trabalhos correlatos esta pesquisa confirma os resultados de André et al (2008); Araújo (2013); Boschi, Oliveira e Assad (2011); Boschi, Oliveira e Avila (2009); Dourado (2013) e Dourado, Oliveira e Avila (2012, 2013) demonstrando algoritmo *K-means* apresenta resultados satisfatórios aplicados a dados pluviométricos.

No entanto, corrobora os resultados de Dikbas et al (2011, tradução nossa), pois mesmo os dois algoritmos mostrando-se satisfatórios na identificação de zonas pluviométricas para determinadas regiões, o algoritmo *Fuzzy C-means* demonstra melhores resultados na busca de grupos homogêneos de regiões pluviométricas, apresentando agrupamentos de forma mais coerente com a espacialização geográfica das estações juntamente com seus atributos pluviométricos. As medidas de qualidade utilizadas por Dikbas et al (2011, tradução nossa) foram *Xie and Beni Index* (1,4), *Dunn Index* (0,05) e *Alternative Dunn Index* (0,002), que resultarem em seis zonas homogêneas.

Outras pesquisas como a de Al-augby et al (2014, tradução nossa) que utilizou dos algoritmos *K-means* e *Fuzzy C-means* na busca de padrões em uma base de dados com informações bancárias para identificar as reações do mercado financeiro referente a notícias falsas e verdadeiras, avaliando por meio do teste estatísticos qui-quadrado da dependência os resultados obtidos, mostrando que o *Fuzzy C-means* obteve melhores resultados.

Contudo, existem pesquisas como a de Yin et al (2014, tradução nossa) que apresentam que o *K-means* obtém melhores resultados sobre o *Fuzzy C-means*, sendo preferível a sua utilização para identificar a melhor função de entrada arterial. Sua validação realizada analisando os resultados de *Time To Peak* (TTP) (*K-means* [0.097s], *Fuzzy C-means* [0.173s]) e *Root Mean Square Error* (RMSE) (*K-means* [0.002], *Fuzzy C-means* [0.004])

Assim, mediante a discussão dos resultados desta pesquisa pode-se observar que a análise do domínio de aplicação e dos dados é de fundamental importância para a identificação do algoritmo que apresenta os melhores resultados, como também para o sucesso do processo de descoberta do conhecimento.

5 CONCLUSÃO

A utilização de uma metodologia para o processo de descoberta de conhecimento é de suma importância para que se alcance com êxito os resultados esperados. O modelo de processo CRISP-DM atendeu todas as necessidades desta pesquisa, auxiliando nos passos a serem executados a cada etapa e na revisão dos processos quando necessário.

Dentre suas principais etapas a de *data mining* contempla a aplicação dos algoritmos sobre os dados que se deseja analisar. A escolha do algoritmo *K-means* foi feita com base na quantidade de pesquisas já realizadas com esse algoritmo, que demonstravam a eficácia da sua utilização nos dados pluviométricos na busca de zonas homogêneas.

Desta forma, a escassez de pesquisas que apresentavam outras linhas de raciocínio fez com que a utilização de outro algoritmo fosse importante. Sendo assim, a escolha do *Fuzzy C-means* aconteceu por sua abordagem baseada na lógica *fuzzy*, podendo apresentar melhores resultados pois não separa seus *clusters* de forma abrupta, mas sim por pertinência entre seus *clusters*.

A dificuldade encontrada, apresentou-se na obtenção dos dados, no *software* HidroWeb só existe a possibilidade de efetuar o *download* dos dados por estação, se tornando um procedimento moroso, desta maneira, foi necessária a implementação de um *script* para que fosse efetuado o *download* de todas as estações pluviométricas da Agência Nacional das águas.

A verificação dos índices de qualidade sobre os resultados obtidos se torna essencial na validação dos *clusters*, com eles pode-se identificar o número de clusters aceitável para o conjunto de dados que foi utilizado. De acordo com as medidas Xie and Beni (0,019), Coeficiente de participação (0,936) e Coeficiente de entropia (0,163) do *Fuzzy C-means* o número de cluster aceitável é três.

Os resultados das medidas aplicadas no *K-means*, *SSE* (32,853) *e Prior Probability* ([0,2667], [0,4889], [0,2444]) também retornaram o número aceitável de três *clusters*.

A aplicação dos dois algoritmos nos dados de precipitações pluviométricas do Estado de Santa Catarina resultou na existência de três zonas pluviométricas, porém o *Fuzzy C-means* apresentou os melhores resultados na espacialização dos

dados e médias pluviométricas (1.448,18mm, 1.364,10mm, 1.665,05mm), demonstrando homogeneidade nos *clusters* obtidos.

Já o algoritmo *K-means* demonstrou-se, mas satisfatório para encontrar clusters heterogêneos, de acordo com a espacialização e médias (1.040,71mm, 1.553,85mm, 1.837,37mm) obtidas.

Desta forma, os resultados obtidos foram satisfatórios, atingindo-se os objetivos propostos que consistiam na compreensão dos conceitos da data mining; agrupamento; algoritmos *K-means* e *Fuzzy C-means*; metodologia CRISP-DM; aplicação dos algoritmos em dados pluviométricos de Santa Catarina e avaliação dos resultados obtidos.

Considerando a pesquisa realizada, tem-se algumas sugestões de trabalhos futuros com o objetivo de aprimorar os conhecimentos na utilização de algoritmos de agrupamentos em dados de precipitações pluviométricas

a) aplicar e analisar os resultados com aplicando outros algoritmos de agrupamento, tais como os Gustafson-Kessel, Gath-Geva, Robust C-Prototypes ou Unsupervised Robust C-Prototypes;
b) analisar os resultados dos algoritmos na identificação de zonas homogêneas na Região Sul do Brasil;
c) realizar a caracterização das zonas pluviométricas do Estado de Santa Catarina, para tal tarefa se faz necessário o conhecimento de um especialista no domínio de aplicação;
d) aplicar a medida de qualidade *silhoutte* para a avaliação dos *clusters* gerados pelo *K-means;*
e) comparar por meio do mapa de isoietas, linhas curvas que representam pontos de igual pluviosidade, a distribuição espacial obtidas por meio do *K-means* e *Fuzzy C-means*.

REFERÊNCIAS

ABONYI, János; FEIL, Balázs. **Cluster Analysis for Data Mining and System Identification .** Berlin: Birkhäuser Verlag AG, 2007.

AL-AUGBY, S.; NERMEND, K.; MAJEWSKI, S.; MAJEWSKA, A.. **A Comparison Of K-Means And Fuzzy C-Means Clustering Methods For A Sample Of Gulf Cooperation Council Stock Markets**. Folia Oeconomica Stetinensia, 2015.

ANDRÉ, R. G. B. et al. Identificação de Regiões Pluviometricamente Homogêneas no Estado do Rio de Janeiro, Utilizando-se Valores Mensais. **Revista Brasileira de Meteorologia**, v. 23, n.4, p. 501-509, 2008. Disponível em: < http://www.scielo.br/pdf/rbmet/v23n4/09.pdf>. Acesso em: 15 out. 2015.

ANDERBERG, M. R., **Cluster analysis for applications**, *Academic Press,* New York, 1973.

ARAÚJO, J. M. S. **Identificação de Áreas com Precipitação Pluvial Homogênea no Estado do Rio Grande do Norte.** 67 f. Dissertação (Mestrado em Engenharia Sanitária) Universidade Federal do Rio Grande do Norte, Natal. 2013. Disponível em:< http://repositorio.ufrn.br/jspui/bitstream/123456789/16006/1/JuremaMSA_DISSERT.pdf >. Acesso em: 15 out. 2015.

AZEVEDO, Ana; SANTOS, Manuel Filipe. KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW. In: EUROPEAN CONFERENCE DATA MINING, 1., 2008, Amsterdam. **KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW.** Amsterdam: Iadis, 2008. v. 1, p. 182 - 185.

BALDO, M. C; ANDRADE, A. R. De; MARTINS, M. L. O. F; NERY, J. T. Análise da precipitação pluvial do Estado de Santa Catarina associada com a anomalia da temperatura da superfície do mar no oceano Pacífico. **Revista Brasileira de Agrometeorologia**, Santa Maria -RS ,v.8,n.2, p.283-293, 2000.

BASU, Sugato; DAVIDSON, Ian; WAGSTAFF, Kiri L. **Constrained Clustering: Advances in Algorithms, Theory, and Applications**. New York: Chapman & Hall/CRC, 2008;

BEZDEK, James et al. **Fuzzy models and algorithms for pattern recognition and image processing.** New York: Springer, 2005.

BOSCHI, R. S., OLIVEIRA, S. R. D. M., ASSAD, E. D. Técnicas de Mineração de Dados para Análise da Precipitação Pluvial Decenal no Rio Grande do Sul. **Revista Engenharia Agrícola**, Jaboticabal, v. 31, n. 6, p. 1189-1201, nov-dez. 2011. Disponível em: < http://www.alice.cnptia.embrapa.br/handle/doc/909223>. Acesso em: 23 set. 2015.

BOSCHI, R. S., OLIVEIRA, S. R. D. M., AVILA, A. M. H. D. Análise da Variabilidade Espaço-Temporal da Precipitação Pluviométrica no Estado do Rio Grande do Sul. In: CONGRESSO BRASILEIRO DE AGROINFORMÁTICA, 7., 2009, Viçosa, MG. **Anais...** Viçosa, MG: UFV, 2009. Disponível em: <http://www.alice.cnptia.embrapa.br/handle/doc/512984>. Acesso em: 23 set. 2015.

BRAMER, Max. **Principles of Data Mining.** 2. ed. London: Springer, 2013.

CHAPMAN, P. et al. **CRISP-DM 1.0: step-by-step data mining guide**. Illinois: SPSS, 78p, 2000. Disponível em: < https://the-modeling-agency.com/crisp-dm.pdf > Acesso em: 13 out. 2015.

COAN, B. De P.; BACK, A. J.; BONETTI, A. V. Precipitação Mensal e Anual Provável no Estado de Santa Catarina. **Revista Brasileira de Climatologia,** v. 15, p. 122-142, jul-dez. 2014. Disponível em: <http://ojs.c3sl.ufpr.br/ojs/index.php/revistaabclima/article/view/38348 >. Acesso em: 14 out. 2015.

COX, Earl. **Fuzzy Modeling and Genetic Algorithms for Data Mining and Explorantion.** San Francisco: Morgan Kaufmann, 2005.

DIKBAS, F. et al. **Classification of precipitation series using fuzzy cluster method.** international journal of climatology, v. 32, p. 1596-1603, 2012.

DOURADO, C. da S. **Mineração de Dados Climáticos para Análise de Eventos Extremos de Precipitação.** 131 f. Dissertação (Mestrado em Engenharia Agrícola) Faculdade de Engenharia Agrícola, Universidade Estadual de Campinas, Campinas. 2013.

DOURADO, C. da S.; OLIVEIRA, S. R. de M.; AVILA, A. M. H. de. Análise de zonas homogêneas em séries temporais de precipitação no Estado da Bahia. **Bragantia**, v. 72, n.2, p. 192-198. 2013. Disponível em: <http://www.scielo.br/pdf/brag/v72n2/v72n2a12.pdf>. Acesso em: 04 set. 2015.

DOURADO, C. da S.; OLIVEIRA, S. R. de M.; AVILA, A. M. H. de. Classificação de anos secos e chuvosos em zonas pluviometricamente homogêneas no Estado da Bahia. In: MOSTRA DE ESTAGIÁRIOS E BOLSISTAS DA EMBRAPA INFORMÁTICA AGROPECUÁRIA, 8., 2012, Campinas. **Resumos...** Brasília, DF: Embrapa, 2012.. Disponível em: <http://www.alice.cnptia.embrapa.br/handle/doc/954506>. Acesso em: 04 set. 2015.

DOURADO, C. da S.; OLIVEIRA, S. R. de M.; AVILA, A. M. H. de. Regionalização da precipitação no estado da Bahia por meio de técnicas de mineração de dados. In: CONGRESSO BRASILEIRO DE METEOROLOGIA, 17.; ENCONTRO DE METEOROLOGIA DOS PAÍSES DO MERCOSUL E ASSOCIADOS, 1.; ENCONTRO SUL AMERICANO DE APLICAÇÕES DO SISTEMA EUMETCast PARA O MONITORAMENTO METEOROLÓGICO E AMBIENTAL, 4.; ENCONTRO DE METEOROLOGIA OPERACIONAL, 2., 2012, Gramado. **Anais**: programa. Gramado: UFRGS, 2012. Disponível em: <http://www.alice.cnptia.embrapa.br/handle/doc/936784>. Acesso em: 23 set. 2015.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. **AI Magazine,** v. 17, n, 3, p. 37-54, 1996. Disponível em: < http://www.aaai.org/ojs/index.php/aimagazine/article/view/1230/1131 >. Acesso em: 12 out. 2015.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. **Data Mining: um guia prático**. Rio de Janeiro: Elsevier, 2005.

GUILLET, Fabrice; HAMILTON, Howard J. **Quality Measures in Data Mining**. Berlin: Springer, 2007.

HAN, Jiawei; KAMBER, Micheline. **Data Mining: Concepts and Techniques**. San Francisco: Morgan Kaufmann , 2001.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data Mining**: **Concepts and Techniques**. 3. ed. Massachusetts: Elsevier, 2012.

HAND, David; MANILLA, Heikki; SMYTH, Padhraic. **Principles of Data Mining.** Massachussetts: Cambridge, 2001.

HOLANDA, C.V.M.; OLIVEIRA, E. **Programa para Homogeneização de Dados – PROHD.** In: SIMPÓSIO DE HIDROLOGIA, 3., 1979, Brasília. Anais... Porto Alegre: Associação Brasileira de Recursos Hídricos, 1979. p.810-845.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Anuário estatístico do Brasil.** Rio de Janeiro: IBGE, 2010 v. 70. Disponível em:< http://biblioteca.ibge.gov.br/visualizacao/monografias/GEBIS%20-%20RJ/AEB/AEB2010.pdf >. Acesso em: 13 nov. 2018.

IPCC – Intergovernmental Panel on Climate Change. **Climate Change 2007: Synthesis Report.** Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Core Writing Team; Pachauri, R.K; Reisinger, A. Geneva, Switzerland, 2008, 104 pp. Disponível em: <https://www.ipcc.ch/pdf/assessment-report/ar4/syr/ar4_syr_full_report.pdf>. Acesso em: 18/11/2015.

IPCC – Intergovernmental Panel on Climate Change. **Climate Change 2014: Synthesis Report.** Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Core Writing Team; Pachauri R.K.; Meyer. L.A. Geneva, Switzerland, 2015, 151 pp. Disponível em: < https://www.ipcc.ch/pdf/assessment-report/ar5/syr/AR5_SYR_FINAL_Front_matters.pdf>. Acesso em: 19/11/2015.

JAIN, Anil K.; DUBES, Richard C. **Algorithms for clustering data.** Englewood Cliffs: Prentice Hall, 1988.

JENSEN, Richard; SHEN, Qiang. **Computational Intelligence and feature selection: Rough and Fuzzy Approaches**. New Jersey: John Wiley & Sons, 2008.

KIM, Y. et al. **A cluster validation index for GK cluster analysis based on relative degree of sharing.** Information Sciences, Vol. 168, p. 255-242, 2004.

KANTARDZIC, Mehmed. **Data Mining: Concepts, Models, Methods, and Algorithms.** John Wiley e Sons. 2003.

KANTARDZIC, Mehmed. **Data Mining: Concepts, Models, Methods, and Algorithms.** 2 ed. John Wiley e Sons. 2011.

CRISP-DM, still the top methodology for analytics, data mining, or data science projects. **KDnuggets**, out. 2014. Disponível em: <http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html>. Acesso em: 20 nov. 2015.

MIRKIN, Boris. **Clustering for Data Mining:** A Data Recovery Approach. Boca Raton: Chapman & Hall/crc, 2005.

MONTEIRO, M. A. Caracterização climática do estado de Santa Catarina uma abordagem dos principais sistemas atmosféricos que atuam durante o ano. **Geosul,** Florianópolis, v. 16, n. 31, p. 69-78, jan-jun. 2001. Disponível em: <https://periodicos.ufsc.br/index.php/geosul/article/view/14052/12896 >. Acesso em: 15 out. 2015.

OMM. **Calculation of monthly and annual 30-year standard normals.** Geneva, 1989. Technical document, n. 341; WCDP, n.10. Disponível em: <http://www.inmet.gov.br/html/clima/OMM_WCDP_N10.pdf>. Acesso em: 23 nov. 2015

OMM. **The Role Of Climatological Normals in a Changing Climate.** Geneva, 2007. Technical document, n. 1377; WCDP, n.61. Disponível em: <https://www.wmo.int/datastat/documents/WCDMPNo61_1.pdf>. Acesso em: 23 nov. 2015

PANDOLFO, C.; BRAGA, H. J.; SILVA JR, V. P. da; MASSIGNAM, A. M., PEREIRA, E. S.;THOMÉ, V. M. R.; VALCI, F.V. **Atlas climatológico do Estado de Santa Catarina**. Florianópolis: Epagri, 2002.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao DATA MINING Mineração de Dados**. Rio de janeiro: Editora Ciência Moderna Ltda. 2009. 900p.

THINSUNGNOEN, Tippaya; KAOUNGKUB, Nuntawut; DURONGDUMRONCHAIB, Pongsakorn; KERDPRASOPB, Kittisak; KERDPRASOPB, Nittaya. **The Clustering Validity with Silhouette and Sum of Squared Errors.** International Conference on Industrial Application Engineering 2015**.** Disponível em: < https://pdfs.semanticscholar.org/8785/b45c92622ebbbffee055aec198190c621b00.pdf >. Acesso em: 13 nov. 2018.

WITTEN, Ian H.; FRANK, Eibe. **Data mining: practical machine learning tools and tecniques.** San Francisco: Morgan Kaufmann, 2005.

WU, Junjie. **Advances in K-means Clustering:** A Data Mining Thinking. Heidelberg: Springer, 2012.

YE, Nong. **Data Mining:** Theories, Algorithms, and Examples. Boca Raton: Crc Press, 2014.

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, 2014. Disponível em: < https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0085884#references> Acesso em: 11 nov. 2018.

ZIMMERMANN, K. Damas. **Revista Santa Catarina em História.** Florianópolis, UFSC, Brasil. ISSN 1984- 3968, v.5, n.2, 2011.

APÊNDICE A – ARTIGO

Algoritmos K-means e Fuzzy C-means Aplicados na identificação de Zonas Pluviométricas em Santa Catarina utilizando o modelo de processo CRISP-DM

Daniel N. Pacheco[1], Merisandra C. de M. Garcia[1]

[1]Universidade do Extremo Sul Catarinense (UNESC) – Criciúma – SC – Brasil

daniel_n_pacheco@hotmail.com, mem@unesc.net

Abstract. *The objective of this research is to identify pluviometrically homogeneous zones in the State of Santa Catarina using the CRISP-DM knowledge discovery process methodology, applying the K-means and Fuzzy C-means clustering algorithms, verifying their results by means of quality measures , in a historical series from 1970 to 2000 of pluviometric data obtained from the National Water Agency database.*

Key words*: Big data, Data mining, K-means, Fuzzy C-means.*

Resumo. *O objetivo desta pesquisa é identificar zonas pluviometricamente homogêneas no Estado de Santa Catarina utilizando a metodologia de processo de descoberta de conhecimento CRISP-DM, aplicando os algoritmos de agrupamento K-means e Fuzzy C-means, verificando seus resultados por meio de medidas de qualidade, em uma série histórica de 1970 a 2000 de dados pluviométricos obtidos da base de dados da Agencia Nacional das Águas.*

Palavras-chave*: Big data, Data mining, K-means, Fuzzy C-means.*

1. Introdução

Com o avanço da tecnologia de coleta e armazenamento de dados as organizações acumulam uma vasta quantidade de dados, porém extrair informações destes torna-se uma tarefa desafiadora, sendo necessária a aplicação de técnicas computacionais de análise de dados devido ao tamanho do conjunto de dados [Tan, Steinbach e Kumar 2009, tradução nossa].

Existem alguns métodos disponíveis que visam auxiliar na análise desses dados, um deles é conhecido como *CRoss-Industry Standart Process for Data Mining* (CRISP-DM), tendo um total de seis etapas, dentre essas etapas, se encontra a de *data mining,* caracterizada pela extração de informações implícitas e potencialmente úteis contidas em um conjunto de dados com a aplicação de tarefas preditivas ou descritivas [Champman et al 2000, tradução nossa].

Um importante índice meteorológico, segundo Back, Bonneti e Coan (2014), que influencia diretamente na economia de Santa Catarina, é a precipitação pluviométrica, devido ao destaque na agricultura catarinense. No entanto, para encontrar

padrões nesses dados a Organização Meteorológica Mundial recomenda a utilização de trinta anos em informações meteorológicas para que eventos adversos não interfiram nos resultados [OMM, 2007, tradução nossa]. Desta maneira é de suma importância a aplicação de técnicas computacionais na obtenção dos resultados.

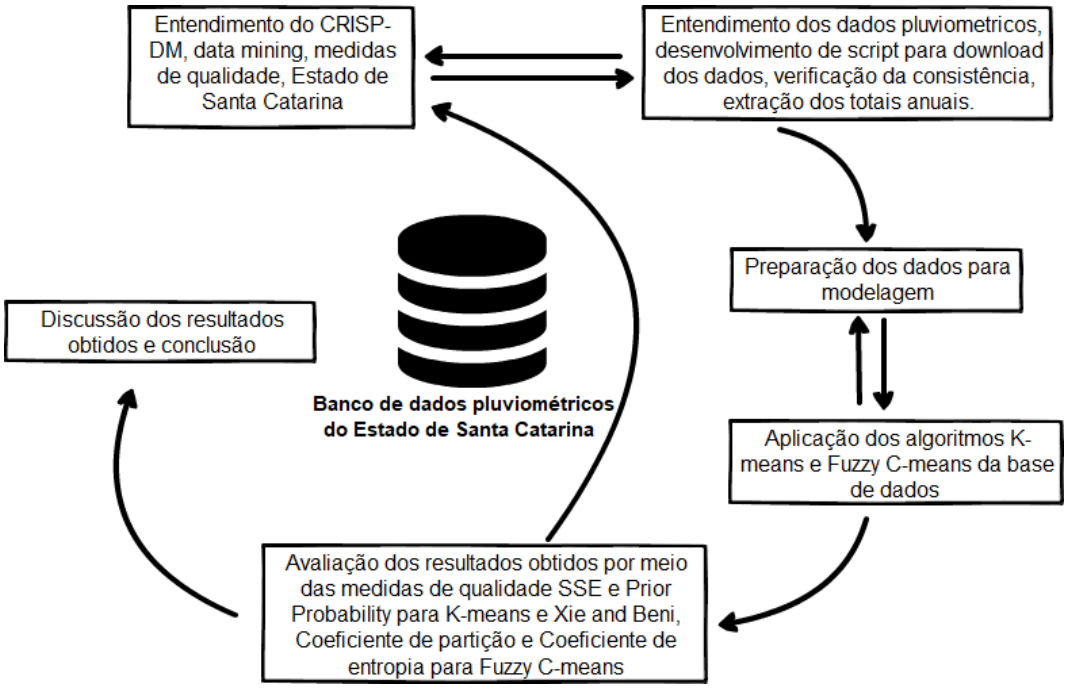
Visto que vários estudos abordaram *o K-means*, como por exemplo, de Dourado, Oliveira e Avila (2013), algoritmo esse baseado na abordagem da lógica clássica, na obtenção dos resultados, tem-se uma possibilidade para a análise de outro algoritmo, *Fuzzy C-means*, que poderia ter melhores resultados utilizando os dados das precipitações, pois não é uma abordagem tradicional, sendo mais subjetiva, portanto, geralmente mais parecida com a forma do pensar humano.

2. Aplicação dos Algoritmos *K-means* e *Fuzzy C-means*

A abordagem dos algoritmos se dá pelo fato do algoritmo *K-means* ser amplamente utilizado nas análises de dados pluviométricos em outras regiões, como por exemplo Dikbas et al (2011) no País Turquia, Dourado, Oliveira e Avila (2013) no Estado da Bahia, onde demonstraram resultados satisfatórios de agrupamento ao serem validados por medidas de qualidade.

Esta pesquisa consiste na aplicação de *data mining* por meio da metodologia CRISP-DM para o agrupamento de dados utilizando a técnica tradicional chamada *K-means* e por meio da técnica baseada na lógica *fuzzy* que possui o nome de *Fuzzy C-means.* Na figura 1 é demonstrado a estrutura do modelo de processo da descoberta de conhecimento aplicado.

Figura 1. Modelo de processo da descoberta de conhecimento desta pesquisa



A área de estudo desta pesquisa corresponde ao Estado de Santa Catarina, que possui uma área oficial de 95.483 km², com mais 502 km² de águas territoriais, totalizando 95.985 km², correspondente a 1,12 % da área brasileira e 16,61% da Região

Sul. Situado entre as latitudes 26º00´S e 30º00´S, e longitudes 48º30´W e 54º00´W [IBGE, 2010].

Tendo grande importância no cenário nacional pois é o primeiro na produção em alho, cebola, maçã, segundo produtor em fumo, terceiro produtor em trigo, quarto produtor em arroz e milho, quinto produtor em batata [PADOLFO et al, 2002].

Por sua localização geográfica, é um dos Estados que apresenta melhor distribuição de precipitação pluviométrica durante o ano. Os principais sistemas meteorológicos responsáveis pelas chuvas no estado são as frentes frias, os vórtices ciclônicos, os cavados de níveis médios, a convecção tropical, a ZCAS e a circulação marítima. Como possui estas características o estado acaba por ter todos os tipos de precipitações: pluviométrica; neve; granizo; orvalho; e geada [MONTEIRO, 2001].

Segundo Coan, Back e Bonetti (2014) a precipitação pluviométrica entre elementos meteorológicos que é a que exerce maior influência sobre as condições ambientais e principalmente nas atividades desenvolvidas em campo, dessa forma seu estudo serve de subsídio para o planejamento rural.

3.1 Entendimento dos Dados

Foram utilizadas séries históricas de precipitação pluviométrica anuais, adquiridas no Portal HidroWeb da ANA. De acordo com a padronização da Organização Meteorológica Mundial para caracterização de dados climáticos é necessário ao mínimo 30 anos de dados consistidos, uma vez que dados brutos oriundos das estações podem apresentar problemas como erros de leitura, transcrição, digitação e ausência de dados. A ANA valida a consistência dos dados pela metodologia proposta pela Agência Nacional de Energia Elétrica, baseada no modelo matemático desenvolvido por Holanda e Oliveira (1979).

Na forma que os dados são disponibilizados no HidroWeb só é possível efetuar o download dos dados de uma estação por vez, desta maneira, causaria grande demora na criação de uma base para efetuar a mineração já que existem 815 estações pluviométricas cadastradas no inventário da ANA para Santa Catarina. Em decorrência disso, foi desenvolvido um script em python para que simulasse o acesso a aplicação HidroWeb e efetuasse o download dos dados de todas as estações de forma automática.

O funcionamento do script consiste em acessar o banco de dados de inventário da ANA, disponível no HidroWeb, buscando as estações pluviométricas de Santa Catarina. Com essas informações o script executa o navegador Firefox informando-o para acessar o HidroWeb, após o acesso o script executa as ações necessárias no site como se fosse um usuário realizando a busca, efetuando as ações de cliques, colocando o nome e código da estação e efetuando o download de toda série histórica de precipitação disponível para aquela estação. Após efetuar o download o script fecha o navegador, busca a próxima estação e repete todo o processo. O processo só é finalizado quando o script terminar de realizar o download dos dados disponíveis de todas as estações.

Desta maneira, o maior período de estações que atenderam os requisitos de 30 anos com dados consistidos é de 1970 a 2000, totalizando 42 estações meteorológicas e 31 anos de dados obtidos da ANA, cobrindo uma boa parte do estado.

3.2 Preparação dos Dados

Os dados obtidos foram importados no software Hidro Sistemas de Informações Hidrológicas 1.3, onde foram extraídos para um arquivo CSV contendo todas as informações de precipitações anuais por estação, código da estação, latitude e longitude. A utilização desse software se faz necessária pois os dados obtidos são identificados por colunas que exigem um conhecimento que se encontra implícito no banco já que não possuem documentação, utilizando o software ele realiza a leitura e entrega os totais anuais consistidos de cada estação.

Os dados para análise dos *clusters* foram organizados somente com dados numéricos, não foi necessária efetuar a transformação dos dados visto que todos já eram numéricos, totalizando 34 atributos (estação, latitude, longitude, chuva de 1970 até 2000) em 42 registros (estações).

Existem algumas ferramentas que auxiliam na verificação dos dados em busca de *outliers*, uma delas é conhecida como Weka que é uma ferramenta de *data mining* frequentemente utilizada em pesquisas da área. Sua distribuição é feita sob a licença *General Public License* e sua manutenção é feita pela Universidade de Wakaito, na Nova Zelândia.

No Weka, na etapa de pré-processamento, foram analisados todos os atributos, identificando que não existia nenhum atributo com valor nulo e todos eram distintos, não sendo necessário executar filtros para a normalização dos dados.

Após o término da preparação iniciou-se a etapa de modelagem, foram aplicados os algoritmos de data mining *K-means* e *Fuzzy C-means* para o agrupamento dos dados.

3.3 Aplicação do Algoritmo *K-means*

No software Weka o algoritmo *K-means* é encontrado com o nome de *SimpleKMeans,* para analisar os resultados do algoritmo se faz necessário utilizar a classe chamada de *MakeDensityBasedClusterer* que encapsula o algoritmo *SimpleKMeans* ou qualquer outro algoritmo de agrupamento apresentando no relatório de resultados os índices de validade dos clusters definidos.

Na utilização do Weka não se faz necessário a conexão com banco de dados, pois existe a possibilidade de importar os dados de arquivos CSV. Acessando o Weka encontramos opção *Explorer,* nessa parte do software onde os dados para o agrupamento foram selecionas e posteriormente processados a pelo do *K-means.*

Na execução do algoritmo foram configurados os seguintes parâmetros para a execução do algoritmo, sendo que os parâmetros que no nome possuem a nomenclatura *Canopy* não se aplicam ao algoritmo *K-means*: *distanceFunction*: Foram aplicados o cálculo de distância Euclidiana e Manhattan; *numClusters*: 1 até 8; *InitializationMethod*: Foram aplicados testes com *Random*, *K-means++* e *Farthest First; maxIterations*: Número máximo de iterações foram 500; *numExecutionSlots*: Informado o valor 1, utilizando assim um cpu/núcleo do processador para a execução do algoritmo; *preserveInstancesOrder*: Se marcado como false o algoritmo não irá preservar a ordem das instâncias, ajustando para melhores resultados.

3.4 Aplicação do Algoritmo *Fuzzy C-means*

Para aplicação do algoritmo *Fuzzy C-means* foi utilizada a ferramenta *Shell Orion Data Mining Engine.* Na execução do algoritmo informou-se os parâmetros a serem respeitados e os atributos utilizados (Estação, Latitude, Longitude e chuvas 1970 até chuvas 2000), a quantidade de *clusters* que o algoritmo deverá encontrar, definir o grau de fuzzyficação entre elementos e clusters, a quantidade máxima de ciclos que o algoritmo irá executar e erro aceitável na execução do algoritmo, indicando a parada do algoritmo.

Como o objetivo desta pesquisa é identificar zonas pluviometricamente homogêneas de acordo com as medidas de qualidades encontradas foram realizadas várias execuções do algoritmo com as configurações de quantidade de *clusters* de 2 até 8, parâmetro de fuzzyficação 2, quantidade de iterações ilimitada (0), taxa de erro 0,00001 e todos os atritos de entrada disponíveis.

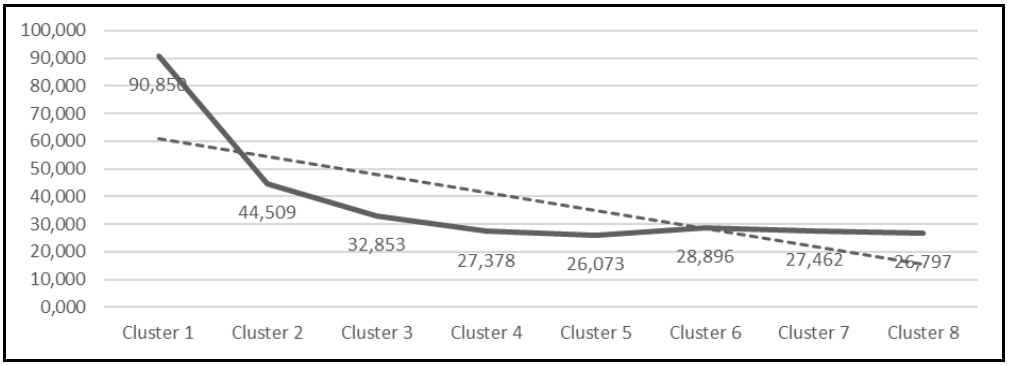
3.5 Resultados Obtidos a partir do *K-means*

Após a análise dos resultados foi possível identificar que alterar o método de inicialização do algoritmo entre *Random*, *K-means++* e *Farthest First* não modificavam os resultados, os índices de qualidade possuíam o mesmo padrão, desta forma os resultados apresentados foram iniciados com o método de inicialização *Random*. A quantidade de iterações do algoritmo não se mostrou um índice de grande valor pois as execuções retornavam valores abaixo de 10, logo o tempo de execução também ficou sempre abaixo de 1 segundo. A alteração das distancias entre Euclidiana e Manhattan não apresentaram diferenças nos resultados finais, sendo assim, foi escolhida a distância Euclidiana para a apresentação.

Nos resultados gerados pelo algoritmo utilizando o cálculo de distância Euclidiana é possível identificar pelo índice de qualidade *SSE*, definido como a soma da distância quadrada entre cada membro do *cluster* e seu centróide, qual é o número aceitável de *clusters* para o conjunto de dados. A análise desses resultados pode ser feita por dois métodos que seriam o método de *Elbow* ou a verificação do percentual de mudança dos *clusters.*

Na figura 2 possuímos os valores do índice SSE de cada aplicação, é possível identificar que o erro diminui à medida que a quantidade de *clusters* aumenta, isso ocorre porque quando o número de *cluster* aumenta, eles ficam menores, então a distorção também é menor.

Figura 2. SSE do algoritmo *K-means* pela distância Euclidiana



A análise dos resultados pelo método de *Elbow*, também chamado de método do cotovelo analisa de forma visual a diferença entre o *cluster* e o próximo *cluster*, no momento que a diferença deixar de ser significativa entende-se que a quantidade de *clusters* ideal para os dados foi encontrada. De acordo com o método o número de clusters aceitável na figura 2 seria 3.

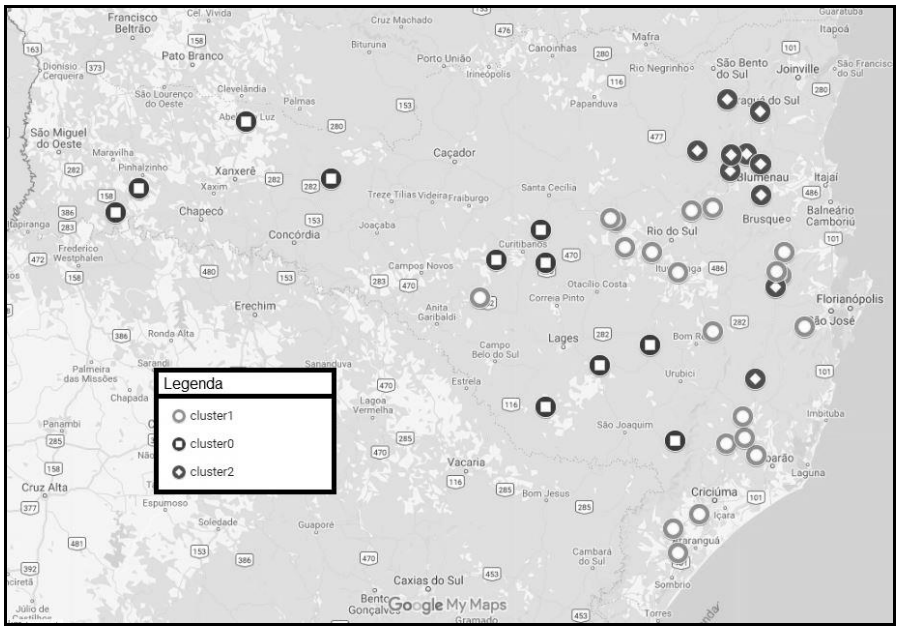
Uma outra maneira analisar o melhor grupo pela *SSE* seria efetuando o cálculo do percentual de mudança entre os grupos, aquele que possuir o maior percentual de mudança é considerado como o melhor, na tabela 1 podemos confirmar que o agrupamento com 3 centroides possuí o maior percentual de mudança.

Tabela 1. Percentual de mudança dos clusters criados pelo algoritmo *K-means*

Número de grupos	SSE	Percentual de mudança
2	44,509	-
3	**32,853**	**35,48016%**
4	27,378	19,99615%
5	26,073	5,005825%
6	28,896	-9,769028%
7	27,462	5,220206%
8	26,797	2,483628%

Realizando a análise espacial dos *clusters* é possível identificar as zonas pluviometricamente homogêneas do Estado de Santa Catarina de acordo com o algoritmo *K-means* (figura 3).

Figura 3. Agrupamento das zonas pluviométricas conforme o algoritmo *K-means*



O *cluster* 0 abrange as mesorregiões do Oeste Catarinense e Serrana, o *cluster* 1 é contempla as estações da região Serrana, Sul, Vale do Itajaí e Grande Florianópolis, por último, o *cluster* 2 é representado pelas regiões Norte e Grande Florianópolis.

Analisando a série de precipitações dos três clusters encontrados, verificando as máximas, mínimas e a média atingida nesse período, é possível identificar que os valores possuem uma tendência crescente, não apresentando homogeneidade no índice pluviométrico das estações que cada *clusters* representa (figura 4). Na tabela 2, podemos identificar os anos das máximas e mínimas encontradas.

Figura 4. Série de precipitações encontradas a partir do *K-means*

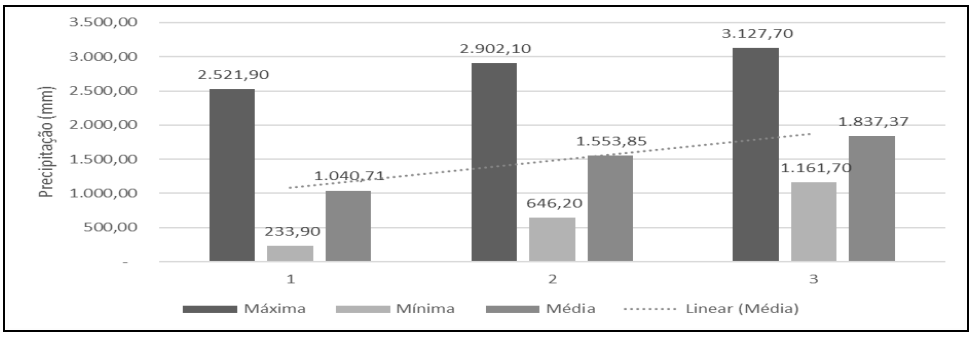


Tabela 2. Detalhamentos das máximas e mínimas encontradas no *K-means*

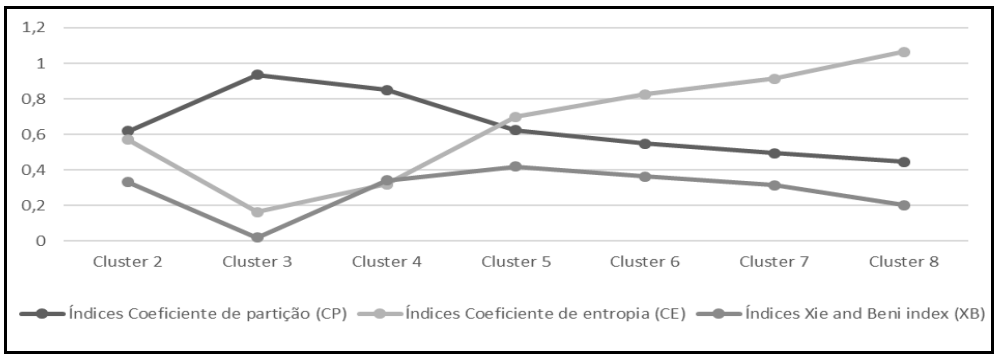
Cluster	Máxima(mm)	Ano Máx.	Mínima(mm)	Ano Mín.
0	2.521,90	1983	233,90	1994
1	2.902,10	1983	646,20	1999
2	3.127,70	1983	1.161,70	1988

Todos os *clusters* apresentaram o ano de 1983 nas máximas, condizendo com o fato histórico de que em 1983 o estado foi afetado por índices pluviométricos altos, acarretando em várias enchentes.

3.6 Resultados Obtidos a partir *Fuzzy C-means*

Os índices de qualidade analisados no algoritmo *Fuzzy C-means* coeficiente de partição, de entropia e *Xie and Beni* auxiliaram na avaliação dos resultados obtidos. Coeficiente de partição é uma função maximizadora, sendo assim quanto maior o valor mais satisfatório é o resultado, os outros dois índices são minimizadores, quanto menor o valor mais convincente é o agrupamento.

Figura 5. Gráfico dos índices de qualidade do Fuzzy C-means

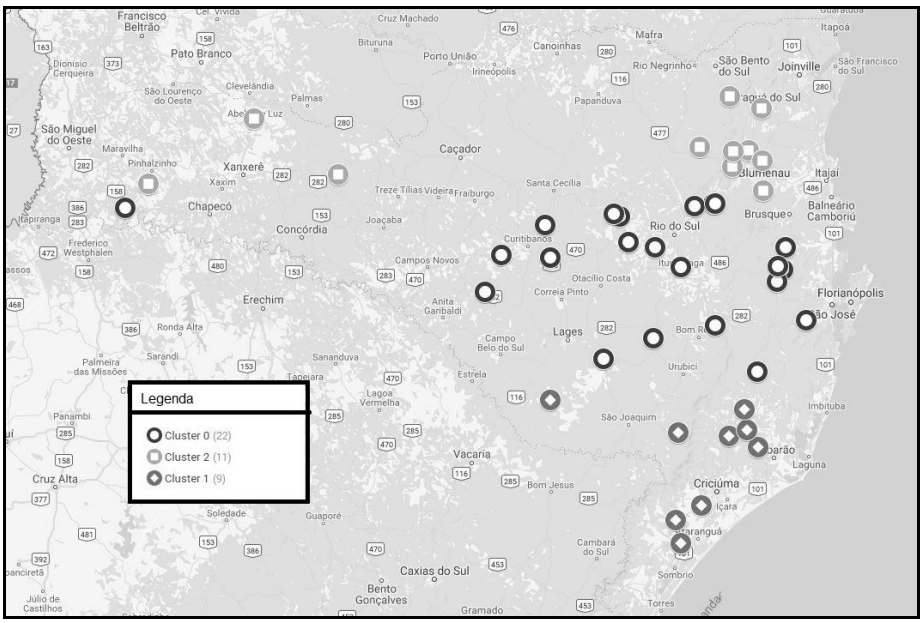


Na figura 5 é possível visualizar os índices obtidos de acordo com os agrupamentos realizados. É possível identificar que o melhor número de grupos para os dados desta pesquisa é 3, a partir desse grupo o Coeficiente de participação tende a

diminuir seu valor, o coeficiente de entropia e *Xie and Beni* a se elevar, mesmo que o *Xie and Beni* tenha uma tendência após o grupo de diminuir, as outras medidas continuam a se distanciar de valores aceitáveis.

Realizando a análise espacial dos *clusters* foi possível identificar as zonas pluviometricamente homogêneas do Estado de Santa Catarina de acordo com as medidas de qualidade aplicadas (figura 6).

Figura 6. Gráfico dos índices de qualidade do Fuzzy C-means

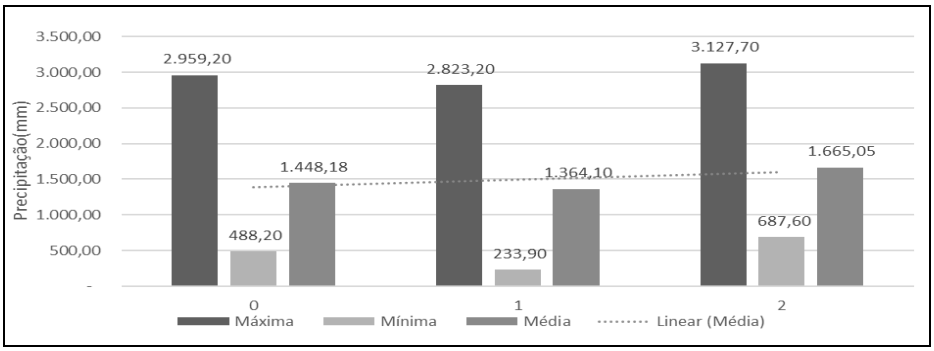


O *cluster* 0 abrange as mesorregiões do Grande Florianópolis, Serrana, Vale do Itajaí e Oeste, já o *cluster* 1 é contempla as estações da região Serrana, Sul, por último, o *cluster* 2 é representado pelas regiões Norte e Oeste.

Tabela 3. Detalhamentos das máximas e mínimas encontradas no K-means

Cluster	Máxima(mm)	Ano Máx.	Mínima(mm)	Ano Mín.
0	2.959,20	1983	488,20	1985
1	2.823,20	1983	233,90	1994
2	3.127,70	1983	687,60	1988

Figura 7. Série de precipitações encontradas a partir do *Fuzzy C-means*



Analisando a série de precipitações dos três clusters encontrados, verificando as máximas, mínimas e a média atingida nesse período, é possível identificar que o valor

da média possui uma regularidade no índice pluviométrico das estações que cada *clusters* representa (figura 7). Na tabela 3, podemos identificar os anos das máximas e mínimas encontradas.

Todos os *clusters* apresentaram o ano de 1983 nas máximas, condizendo com o fato histórico de que em 1983 o estado foi afetado por índices pluviométricos altos, acarretando em várias enchentes.

3.7 Discussão dos Resultados Obtidos

Avaliando os resultados obtidos de cada algoritmo empregado nesta pesquisa, pode-se verificar que ambos apresentaram os melhores resultados para o mesmo número de grupos de acordo com as medidas aplicadas, porém as disposições das estações e as médias pluviométricas apresentaram diferenças. No algoritmo *Fuzzy C-means* as estações ficaram visivelmente homogêneas de acordo com suas regiões e média de precipitação, isso se deve ao grau de pertinência que cada objeto possui. Portanto, mesmo possuindo valores que se adequem a outros centroides, de acordo com a lógica *fuzzy*, o objeto pertence ao grupo que apresentar maior similaridade em todos os atributos.

O algoritmo *K-means* realizou a tarefa de agrupamento de forma mais abrupta nos dados utilizados, ocasionando que alguns objetos dos grupos ficassem dentro de outros grupos com uma distância muito próxima. Com isso, não se mostrou satisfatório na espacialização dos grupos homogêneos pelo estado de Santa Catarina, no entanto, para identificação de grupos heterógenos o algoritmo se mostrou satisfatório ao serem analisadas as médias pluviométricas de cada *cluster* (1.040,71mm, 1.553,85mm, 1.837,37mm).

Assim, os resultados desta pesquisa apontaram que o algoritmo *Fuzzy C-means* consegue obter resultados mais satisfatórios na questão de obter *clusters* homogêneos, demonstrando capacidade em caracterizar o Estado de Santa Catarina, conforme os dados coletados, em três regiões pluviometricamente homogêneas, visto que as médias de precipitação encontradas nos clusters foram 1.448,18mm para o *cluster* 0, 1.364,10mm para o *cluster* 1 e 1.665,05mm para o *cluster* 2.

4. Conclusão

A utilização de uma metodologia para o processo de descoberta de conhecimento é de suma importância para que se alcance com êxito os resultados esperados. O modelo de processo CRISP-DM atendeu todas as necessidades desta pesquisa, auxiliando nos passos a serem executados a cada etapa e na revisão dos processos quando necessário.

A verificação dos índices de qualidade sobre os resultados obtidos se torna essencial na validação dos *clusters*, com eles pode-se identificar o número de clusters aceitável para o conjunto de dados que foi utilizado. De acordo com as medidas Xie and Beni (0,019), Coeficiente de participação (0,936) e Coeficiente de entropia (0,163) do *Fuzzy C-means* o número de cluster aceitável é três. Os resultados das medidas aplicadas no *K-means*, *SSE* (32,853) também retornaram o número aceitável de três *clusters*.

A aplicação dos dois algoritmos nos dados de precipitações pluviométricas do Estado de Santa Catarina resultou na existência de três zonas pluviométricas, porém o *Fuzzy C-means* apresentou os melhores resultados na espacialização dos dados e médias pluviométricas (1.448,18mm, 1.364,10mm, 1.665,05mm), demonstrando homogeneidade nos *clusters* obtidos. O algoritmo *K-means* demonstrou-se, mas satisfatório para encontrar clusters heterogêneos, de acordo com a espacialização e médias (1.040,71mm, 1.553,85mm, 1.837,37mm) obtidas.

Mediante este estudo, conclui-se que o algoritmo *Fuzzy C-means* apresenta melhores resultados na identificação de zona pluviometricamente homogêneas no Estado de Santa Catarina. Enquanto o *K-means* demonstrou-se, mas satisfatório para encontrar clusters heterogêneos.

Referências

CHAPMAN, P. et al. CRISP-DM 1.0: step-by-step data mining guide. Illinois: SPSS, 78p, 2000.

COAN, B. De P.; BACK, A. J.; BONETTI, A. V. Precipitação Mensal e Anual Provável no Estado de Santa Catarina. Revista Brasileira de Climatologia, v. 15, p. 122-142, jul-dez. 2014.

DIKBAS, F. et al. Classification of precipitation series using fuzzy cluster method. international journal of climatology, v. 32, p. 1596-1603, 2012.

DOURADO, C. da S.; OLIVEIRA, S. R. de M.; AVILA, A. M. H. de. Análise de zonas homogêneas em séries temporais de precipitação no Estado da Bahia. Bragantia, v. 72, n.2, p. 192-198. 2013.

GUILLET, Fabrice; HAMILTON, Howard J. Quality Measures in Data Mining. Berlin: Springer, 2007.

HOLANDA, C.V.M.; OLIVEIRA, E. Programa para Homogeneização de Dados – PROHD. In: SIMPÓSIO DE HIDROLOGIA, 3., 1979, Brasília. Anais... Porto Alegre: Associação Brasileira de Recursos Hídricos, 1979. p.810-845.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. Anuário estatístico do Brasil. Rio de Janeiro: IBGE, 2010 v. 70.

MONTEIRO, M. A. Caracterização climática do estado de Santa Catarina uma abordagem dos principais sistemas atmosféricos que atuam durante o ano. Geosul, Florianópolis, v. 16, n. 31, p. 69-78, jan-jun. 2001.

OMM. The Role Of Climatological Normals in a Changing Climate. Geneva, 2007. Technical document, n. 1377; WCDP, n.61.

PANDOLFO, C.; BRAGA, H. J.; SILVA JR, V. P. da; MASSIGNAM, A. M., PEREIRA, E. S.;THOMÉ, V. M. R.; VALCI, F.V. Atlas climatológico do Estado de Santa Catarina. Florianópolis: Epagri, 2002.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. Introdução ao DATA MINING Mineração de Dados. Rio de janeiro: Editora Ciência Moderna Ltda. 2009. 900p.