ARAMIS FERREIRA MARQUES

APLICAÇÃO DE CLUSTERIZAÇÃO DE DADOS NA BASE DE DADOS DO ZONEAMENTO ECOLÓGICO-ECONÔMICO DE MINAS GERAIS

LAVRAS - MG

2014

ARAMIS FERREIRA MARQUES

APLICAÇÃO DE CLUSTERIZAÇÃO DE DADOS NA BASE DE DADOS DO ZONEAMENTO ECOLÓGICO-ECONÔMICO DE MINAS GERAIS

Monografia apresentada ao Colegiado do Curso de Sistemas de Informação, para a obtenção do título de Bacharel em Sistemas de Informação.

Orientador

Prof. Dr. Ahmed Ali Abdalla Esmin

LAVRAS - MG

2014

ARAMIS FERREIRA MARQUES

APLICAÇÃO DE CLUSTERIZAÇÃO DE DADOS NA BASE DE DADOS DO ZONEAMENTO ECOLÓGICO-ECONÔMICO DE MINAS GERAIS

Monografia de graduação apresentada ao Colegiado do Curso de Bacharelado em Sistemas de Informação, para obtenção do título de Bacharel.

APROVADA em 30 de junho de 2014.

Cristina Lelis Leal Calegari

Marluce Rodrigues Pereira

Ahmed Ali Abdalla Esmin (~~Orientador~~)

LAVRAS-MG
2014

Dedico esse trabalho à toda minha família, minha namorada e amigos.

AGRADECIMENTOS

Agradeço primeiramente a Deus, por ter me dado força para concluir esse trabalho. Agradeço meus pais por sempre me apoiarem nesse período tão importante em minha vida. Agradeço também minha namorada por me apoiar sempre, em todos os momentos desta caminhada. Agradeço também ao Prof. Ahmed Ali Abdalla Esmin pela oportunidade e agradeço também a Prof. Cristina Lelis Leal Calegario, sem sua ajuda não seria possível concluir este trabalho. Um agradecimento especial para a Christiane, Luciene e Fernanda que também me ajudaram neste trabalho. Enfim, agradeço todos que me ajudaram direta e indiretamente.

RESUMO

Atualmente o rápido desenvolvimento de tecnologias e ferramentas de coleta e armazenamento de dados possibilitam que grandes volumes de dados sejam armazenados. Em áreas como negócios, medicina, ciência e engenharia, na web através de buscas feitas em motores de buscas, redes sociais, blogs, e-commerce, são gerados diariamente grandes volumes de dados. Ferramentas tradicionais de análises de dados não são capazes de extrair conhecimento de grandes bases de dados, pois o volume de dados armazenados é elevado e a diversidade desses dados dificultam esse processo. Nesse contexto, as técnicas de mineração de dados são utilizadas para agir em grandes bases de dados com o objetivo de descobrir padrões úteis, que poderiam permanecer ignorados. Esse novo conhecimento gerado pode auxiliar tomadores de decisão em diversas questões referentes aos seus negócios. Assim, o principal objetivo desse trabalho é a aplicação de técnicas de mineração de dados na base de dados do Sistema de Informação do Zoneamento Ecológico Econômico de Minas Gerais(ZEE-MG), a fim de identificar características e padrões ocultos. Estes resultados podem auxiliar, por exemplo, como potencializador no planejamento e elaboração das políticas públicas do estado de Minas Gerais.

Palavras-chave: Banco de Dados; Mineração de Dados; ZEE-MG, Agrupamento

ABSTRACT

Nowadays, the fast development of collection and data storage technologies and tools enable large volumes of data to be stored. In areas such as business, medicine, science and engineering, through web searches on search engines, social networks, blogs, e-commerce, are generated large amounts of data every day. Traditional data analysis tools are not able to extract knowledge from large databases, once the data volume is high and the diversity of these makes the discovery knowledge process a tough task. In this context, data mining techniques are used to explore large databases in order to discover useful patterns, which could still be ignored. This new knowledge generated can support decision makers in various situations related to their business. Thus, the main objective of this work is to make an application of data mining techniques in the information system database of the Ecological Economic Zoning of the Minas Gerais state (ZEE-MG), aiming the identification of hidden patterns and characteristics. These results may help, for example, as a potentializer for the planning and preparation of public policies for the state of Minas Gerais.

Keywords: Data Base; Data Mining; ZEE-MG, Clustering

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Nesta seção primeiramente descreve-se a contextualização, bem como a motivação para a realização deste trabalho. Ao final, apresenta-se o objetivo geral deste trabalho, seguido dos objetivos específicos.

1.1 Contextualização e Motivação

Atualmente vive-se na era da informação, onde são trafegados na rede mundial de computadores terabytes ou petabytes de dados diariamente (HAN; KAMBER; PEI, 2011). Avanços rápidos na tecnologia de coleta e armazenamento de dados, possibilitam que dados relacionados a negócios, sociedade, ciência e engenharia, medicina, e de outros aspectos da vida sejam armazenados, resultando em grandes volumes de dados armazenados. Podem ser citadas algumas aplicações atuais em que é gerado um grande volume de dados, como por exemplo:

- Os satélites atuais da NASA(*National Aeronautics and Space Administration*) de observação da Terra geram um terabyte de dados todos os dias (BRAMER, 2007).
- A indústria médica gera grandes quantidades de dados resultantes de registros médicos, monitoramento de pacientes e imagens médicas (HAN; KAMBER; PEI, 2011).
- Muitas empresas constroem grandes armazéns de dados de transações dos clientes, que podem conter mais de uma centena de milhões de transações (BRAMER, 2007).

- Bilhões de pesquisas feitas em motores de busca na web processam dezenas de petabytes de dados diários (HAN; KAMBER; PEI, 2011).
- Comunidades e redes sociais tornaram-se uma fonte de dados muito importante, gerando um grande volume de dados multimídia (HAN; KAMBER; PEI, 2011).

Nesse contexto, a informação útil tem um papel fundamental na sociedade atual. Os avanços na tecnologia de coleta e armazenamento de dados permitem armazenar grandes quantidades de dados a um custo relativamente baixo. A percepção sobre a possibilidade de extrair conhecimento útil de grande bases de dados aumentou, em paralelo com o desenvolvimento da área de mineração de dados. Esse conhecimento, pode ser fundamental para o crescimento de uma empresa, auxiliar em descobertas na ciência, permitir que se faça previsões do clima e de desastres naturais, auxiliar na identificação de causas e curas de doenças terminais (BRAMER, 2007).

A enorme quantidade de dados coletados e armazenados, ultrapassou em muito nossa capacidade humana de compreensão. Como resultado, o tomador de decisão muitas vezes não se baseia nos dados armazenados nos repositórios de dados, mais sim em sua intuição, pois ele muitas vezes, não dispõem de ferramentas que sejam capazes de extrair o conhecimento desses repositórios (HAN; KAMBER; PEI, 2011).

A existência de enormes volumes de dados armazenados, implica muitas vezes, que eles nunca serão examinados ou serão examinados superficialmente. Algumas técnicas de aprendizado de máquina têm o potencial de extrair conhecimento dos vastos repositórios de dados que inundam organizações, governos e indivíduos (BRAMER, 2007).

Mas a extração de informação útil de grandes bases de dados não é uma tarefa trivial, pois fatores como, por exemplo, volume do conjunto de dados e a natureza dos dados, necessitam de ferramentas poderosas e versáteis para que os dados sejam transformados em conhecimento (TAN; STEINBACH; KUMAR, 2009b).

Dessa necessidade, surgiu a mineração de dados. De acordo com Maimom e RoKach (2010) mineração de dados envolve a inferência de algoritmos que exploram os dados, o desenvolvimento de modelos e o descobrimento de padrões ou informações úteis ocultas em grandes bases de dados. Os modelos são usados para compreender os dados, fazer análises e predições.

De acordo com Bramer (2007) Tan, Steinbach e Kumar (2009b) Olson e Delen (2008), existem várias aplicações em que a técnica de mineração de dados podem ser utilizadas, podemos citar:

- Negócios: aplicações de inteligência de negócios como a criação de perfis de clientes, vendas direcionadas, disposição dos produtos em uma loja ou detecção de fraudes.
- Medicina: diagnósticos médicos mais precisos
- Banco: no relacionamento com o cliente
- Telemarketing: informações online, com fácil acesso aos dados
- Gestão de Recursos Humanos: identificar a probabilidade de rotatividade dos funcionários

1.2 Objetivos

O objetivo geral deste trabalho é aplicar a técnica de *clustering* ou agrupamento na base de dados do ZEE-MG(Zoneamento Ecológico-Econômico de Minas Gerais), a fim de identificar características e padrões ocultos. Estes resultados podem auxiliar por exemplo, no planejamento e elaboração das políticas públicas e das ações em meio ambiente no estado de Minas Gerais. Para atingir esse objetivo geral será necessário atingir os seguintes objetivos específicos:

- Adaptar e aplicar a técnica de acordo com as características dos dados existentes na base de dados;
- Analisar, avaliar e relacionar os resultados obtidos de acordo com o contexto do Zoneamento Ecológico Ecológico de Minar Gerais.

A organização deste trabalho está dividida da seguinte forma. O capítulo 2 apresenta o referencial teórico, onde está descrito os principais conceitos relacionados a mineração de dados. O capítulo 3 apresenta a metodologia, que é constituída pela classificação da pesquisa, uma visão geral do objeto de estudo e os procedimentos metodológicos. O capítulo 4 apresenta o desenvolvimento deste trabalho, onde é descrito como o trabalho foi desenvolvido levando em consideração cada etapa do processo *Knowledge Discovery in Database*(KDD). No capítulo 5 temos a conclusão e trabalhos futuros. Por último, o capítulo 6 apresenta o apêndice.

2 REFERENCIAL TEÓRICO

Nesta seção, são apresentados os conceitos básicos relacionados a mineração de dados que são indispensáveis para a realização deste trabalho.

2.1 Atributos e Tipos de Atributos

De forma geral, um conjunto de dados, é formado por objetos de dados. Um objeto de dados pode ser representado por um tupla em um banco de dados, e uma tupla é formada por um conjunto de características ou atributos que representam aquele objeto. Por exemplo, poderíamos ter um conjunto de dados composto por vários objetos clientes, onde cada objeto cliente é caracterizado por uma tupla contendo vários atributos, como por exemplo, o código identificador do cliente, nome e endereço (HAN; KAMBER; PEI, 2011).

Os dados de entrada para o processo de mineração de dados podem ter vários formatos e podem ser armazenados de formas diferentes. Em mineração de dados, o tipo de um atributo é definido pelo conjunto de valores nominal, ordinal, binário e numérico que ele pode assumir (BRAMER, 2007; HAN; KAMBER; PEI, 2011).

2.1.1 Nominal

Atributos nominais são utilizados para categorizar objetos, como por exemplo, cor de cabelo de um objeto pessoa. Uma variável nominal não necessariamente é representada por símbolos ou nomes, ela também pode assumir valores numéricos. Neste caso, os valores numéricos não tem interpretação matemática. Por exemplo, na caso do atributo cor de cabelo, pode-se atribuir valores numéricos, 1 para loiro,

2 para castanho e assim por diante. Neste caso, os valores numéricos atribuídos não seriam utilizados de forma quantitativa (HAN; KAMBER; PEI, 2011).

2.1.2 Binário

Atributos binários são considerados tipos especiais de atributos nominais, pois são representados somente por dois valores, como por exemplo, 1 ou 0, verdadeiro ou falso. No caso dos valores 1 ou 0, 1 normalmente significa ativo ou presente e 0 significa inativo ou ausente. No caso de verdadeiro ou falso, a variável assume o tipo booleano (BRAMER, 2007; HAN; KAMBER; PEI, 2011).

2.1.3 Ordinal

Segundo Bramer (2007) as variáveis ordinais são semelhantes às variáveis nominais. O que difere os dois tipos, é que uma variável ordinal tem valores que possuem uma ordem significativa.

Um exemplo de atributos ordinais seria o nível de satisfação de clientes de uma loja no que diz respeito ao atendimento recebido. A satisfação do cliente poderia ser representada pelas seguinte categorias: 0 - muito insatisfeito, 1 - pouco insatisfeito, 2 - satisfeito, 3 - pouco satisfeito e 4 - muito satisfeito (HAN; KAMBER; PEI, 2011).

2.1.4 Numérico

Segundo Han, Kamber e Pei (2011) atributos numéricos são quantitativos, podem assumir valores inteiro ou reais e são divididos em dois tipos: intervalar ou proporcional.

Escala Intervalar: para esse tipo de atributo, as diferenças entre os valores são significativas. Datas de calendário e temperatura em *Celsius* ou *Fahrenheit* são exemplos de atributos intervalares. No caso da temperatura, podemos dizer que 20°C é cinco graus mais elevado que 15°C. Este raciocínio também é válido para as datas (HAN; KAMBER; PEI, 2011).

Escala Proporcional: para esse tipo de atributo, diferenças e proporções são significativas. Atributos para medir peso, altura, latitude e longitude, quantidades monetárias são exemplos de atributos proporcionais (HAN; KAMBER; PEI, 2011; TAN; STEINBACH; KUMAR, 2009b).

2.2 Mineração de Dados

Data Mining ou Mineração de dados tem como objetivo a descoberta automática ou semi-automática de conhecimento útil, a partir de grandes bases de dados (HAN; KAMBER; PEI, 2011) (MAIMOM; ROKACH, 2010) (WITTHEN; FRANK; HALL, 2011). De acordo com Tan, Steinbach e Kumar (2009b) as técnicas de mineração de dados são utilizadas para agir em grandes bases de dados com o objetivo de descobrir padrões úteis, que poderiam permanecer ignorados. Ainda segundo Tan, Steinbach e Kumar (2009b) a mineração de dados originou-se a partir de várias áreas de pesquisa, como por exemplo:

1. Amostragem, estimativa e teste de hipóteses a partir de estatísticas.

2. Algoritmos de busca, técnicas de modelagem e teorias de aprendizagem da inteligência artificial, reconhecimento de padrões e aprendizagem de máquina.

Segundo Han, Kamber e Pei (2011) a mineração de dados trata-se de um processo interdisciplinar, envolvendo áreas como por exemplo, estatística, aprendizagem de máquina, visualização e ciência da informação.

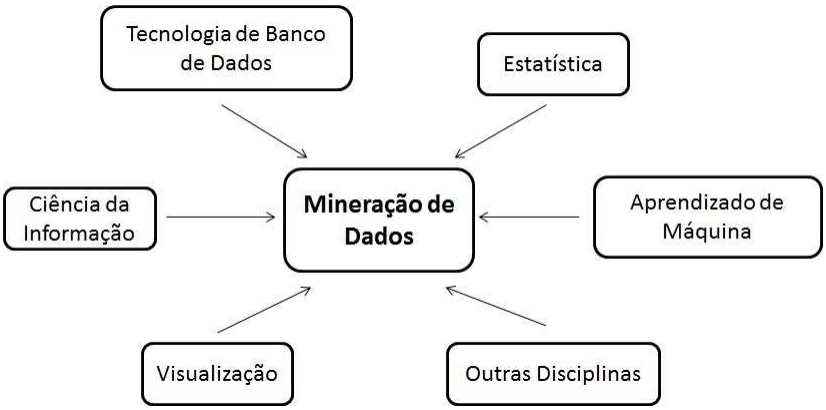

Figura 1: Mineração de Dados como um processo interdiciplinar (HAN; KAMBER; PEI, 2011)

Ainda segundo Han, Kamber e Pei (2011) são utilizadas diferentes técnicas dependendo da abordagem e dos tipos de dados a serem extraídos no processo de mineração de dados. Técnicas como redes neurais, *fuzzy* e/ ou teoria dos conjuntos, representação do conhecimento, programação lógica indutiva podem ser utilizadas dependendo da abordagem da mineração de dados.

Segundo Han, Kamber e Pei (2011) o termo mineração de dados é frequentemente associado ao processo de descoberta de conhecimento como um todo, em meios como, indústria, mídia e no meio de pesquisa. Por esta razão, o autor define o termo Mineração de Dados de uma maneira mais ampla, como um processo de descoberta de padrões e conhecimento a partir de grandes quantidades de dados. A mineração de dados é considerada pela maioria dos autores pesquisados como parte de um processo maior, mais conhecido como KDD (*Knowledge Discovery in Database*), sendo considerado um dos passos mais importantes do

processo (MAIMOM; ROKACH, 2010; TAN; STEINBACH; KUMAR, 2009b; HAN; KAMBER; PEI, 2011; BRAMER, 2007).

2.2.1 Tarefas de Mineração de Dados

De acordo Han, Kamber e Pei (2011), em geral, as tarefas de mineração de dados podem ser classificadas em duas categorias: descritivas e preditivas. As tarefas de mineração de dados descritivas caracteriza as propriedades dos dados disponíveis em um conjunto de dados destino, por sua vez, tarefas de mineração de dados preditivas prevê dados não disponíveis a partir de dados disponíveis, a fim de realizar previsões.

Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996) a fronteira que divide previsão e descrição não é bem definida, pois alguns modelos de previsão podem ser descritivos e vice-versa. Mas mesmo assim essa divisão é útil para fins didáticos.

Segundo Tan, Steinbach e Kumar (2009b), Larose (2004), Han, Kamber e Pei (2011) a mineração de dados pode ser classificada dependendo do tipo de tarefa realizada. A seguir são apresentadas as principais tarefas de mineração de dados.

Associação (*Association*): segundo Larose (2004) a análise de associação é utilizada para descobrir padrões que descrevam a relação entre atributos. Normalmente, usa-se regras de implicação da forma *se atributo A então atributo B*, onde A e B são conjuntos distintos de itens, para representar os padrões descobertos.

A força de uma regra de associação pode ser medida em termos de suporte e confiança. "O suporte determina a frequência na qual a regra é aplicável a um determinado conjunto de dados, enquanto que a confiança determina a frequência

na qual os itens em B aparecem em transações que contenham A"(TAN; STEINBACH; KUMAR, 2009b).

Ainda segundo (TAN; STEINBACH; KUMAR, 2009b) a utilização do suporte e confiança é importante para eliminar regras que podem não ser interessantes. Geralmente, uma regra com suporte baixo tem grande probabilidade de não ser interessante, por exemplo, partindo de uma perspectiva de negócio pode não ser lucrativo promover itens que os clientes raramente compram juntos. Em relação a confiança, quanto maior seu valor, maior a probabilidade de B estar presente em transações que contenham A.

Por exemplo, um supermercado pode descobrir que dos 1000 clientes que compram em uma noite, 200 compraram fraldas e desses 200, 50 compraram cerveja. Assim, a regra de associação seria *Se compra fralda então compra cerveja*, com um suporte de 200/1000 = 20% e uma confiança de 50/200 = 25% (LAROSE, 2004).

Classificação (*Classification*): de acordo com Bramer (2007) a tarefa de classificação é bastante comum em mineração de dados. A classificação é o processo de encontrar um modelo ou função que descreve e distingue classes de dados e conceitos. O modelo é obtido da análise de um conjunto de dados de treino onde seu propósito é ser aplicado a dados não classificados visando categorizá-los em classes (HAN; KAMBER; PEI, 2011).

Basicamente, a classificação consiste em determinar qual classe um determinado objeto pertence, a partir de um modelo de classificação que foi derivado de um conjunto de dados de treinamento onde já é conhecido o valor do rótulo da classe ou do atributo alvo. Por esse motivo a tarefa de classificação é classificada como aprendizado supervisionado (HAN; KAMBER; PEI, 2011).

Segundo Han, Kamber e Pei (2011) uma abordagem geral para resolver problemas de classificação seria fornecer um conjunto de treinamento consistindo de registros onde os rótulos ou atributos alvo sejam conhecidos. O conjunto de treinamento é utilizado para construir um modelo de classificação que é gerado por um algoritmo de aprendizagem. Posteriormente, o modelo é utilizado para classificar um conjunto de dados que contenha rótulos de classes de registros desconhecidos.

Por exemplo, classificar o perfil de um novo colaborador de uma empresa a partir de um modelo gerado através de um conjunto de dados contendo as informações sobre os colaboradores da mesma (BRAMER, 2007).

Segundo Larose (2004) e Han, Kamber e Pei (2011) a tarefa classificação pode ser utilizada em diversas aplicações. Ela pode ser usada, por exemplo, para:

- Determinar se uma transação com cartão de crédito é fraudulenta;
- Identificar se uma pessoa pode ser uma ameaça para a segurança através da análise do seu comportamento pessoal ou financeiro.
- Detecção de mensagens de spam em e-mails baseada no cabeçalho e conteúdo da mensagem
- Classificação de galáxias baseada nos seus formatos

Algumas técnicas como K-vizinhos mais próximo, árvore de decisão e redes neurais são bastante utilizados na classificação em mineração de dados (LAROSE, 2004).

Estimação ou Regressão (*Estimation or Regression*): segundo Han, Kamber e Pei (2011) a classificação e regressão são semelhantes, ambas são uma forma de previsão. A principal diferença entre as duas tarefas é que na classificação

o valor a ser predito é categórico (discreto, sem ordem), já na regressão o objetivo é prever um valor numérico. Assim podemos estimar um valor numérico de uma determinada variável a partir dos valores das demais variáveis (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; BRAMER, 2007).

De acordo com Fayyad, Piatetsky-Shapiro e Smyth (1996), Han, Kamber e Pei (2011) algumas aplicações para a tarefa regressão podem ser:

- Estimar a probabilidade de sobrevivência de um paciente a partir de um conjunto de testes de diagnóstico
- Prever a quantidade de receita que cada item irá gerar durante uma venda, com base nos dados de vendas anteriores

De acordo com Han, Kamber e Pei (2011) o modelo obtido pode ser representado de várias formas, como por regras de classificação, árvores de decisão, formas matemáticas e redes neurais.

Agrupamento (*Clustering*): Segundo Larose (2004) *clustering* tem como objetivo agrupar registros, observações ou casos em classes de objetos similares. Um *cluster* é uma coleção de registros similares entre si, mas são diferentes dos registros que pertencem a outros *clusters* (BRAMER, 2007) (LAROSE, 2004).

De acordo com (HAN; KAMBER; PEI, 2011) as diferenças e semelhanças são avaliadas com base nos valores de atributos que descrevem os objetos e muitas vezes envolvem medidas de distância. Por exemplo, *clustering* pode ser utilizado para identificar subpopulações homogêneas de clientes. A figura abaixo mostra um gráfico 2-D dos clientes no que diz respeito às posições dos clientes em uma cidade. Três *clusters* são evidentes (HAN; KAMBER; PEI, 2011).

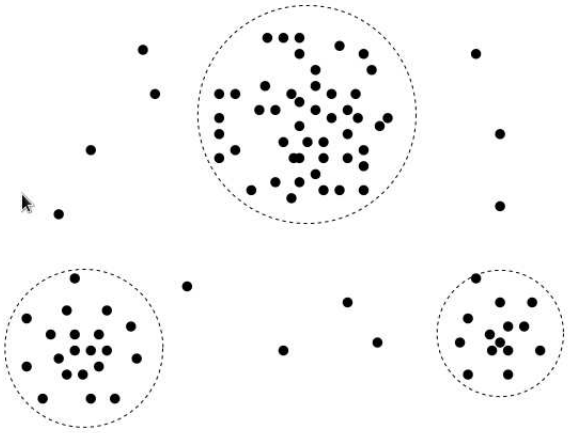

Figura 2: Grupos de clientes homogêneos representados por três *clusters* (HAN; KAMBER; PEI, 2011)

Diferentemente da tarefa de classificação, o *clustering* não necessita de um atributo alvo, por isso seu aprendizado é não-supervizionado (HAN; KAMBER; PEI, 2011).

De acordo com Larose (2004) esta tarefa não tenta classificar, estimar ou prever um valor de um atributo alvo.

Existem várias áreas de aplicação que se beneficiam com o agrupamento de objetos semelhantes, tais como a biologia, segurança, *business intelligence* e pesquisa na web (BRAMER, 2007) (HAN; KAMBER; PEI, 2011) (LAROSE, 2004). Podemos citar como exemplo:

- Na economia pode ser utilizado para encontrar países cujas economias são semelhantes;
- Em uma aplicação de marketing para encontrar grupos de clientes com comportamento de compra similares;

- Em uma aplicação médica onde gostaríamos de encontrar grupos de pacientes com sintomas semelhantes;
- Na classificação de documentos na web;
- Como uma ferramenta de redução de dimensão, quando o conjunto de dados tem centenas de atributos;
- Para agrupar genes que tenham comportamentos semelhantes.

Existem vários métodos ou algoritmos de *clustering*, *K-means* e agrupamento hierárquico são os mais comumente utilizados (BRAMER, 2007).

Detecção de Anomalias (*Outlier Analysis*): Um conjunto de dados pode conter objetos com características ou comportamentos que não estejam em conformidade com os demais. Em algumas aplicações como em detecção de fraudes, características raras podem ser mais interessantes do que as que ocorrem regularmente (HAN; KAMBER; PEI, 2011).

Segundo Tan, Steinbach e Kumar (2009b) a tarefa detecção de anomalias, dado um conjunto de dados, tem como objetivo identificar anomalias ou fatores estranhos cujas características sejam diferentes dos demais. Por exemplo, na figura abaixo, os objetos da região *R* são considerados “anormais”, pois são significativamente diferentes dos demais.

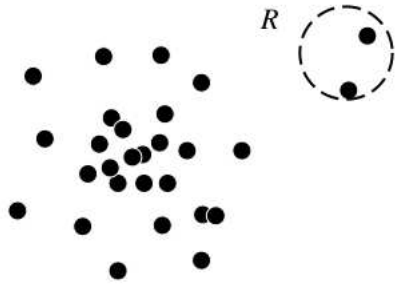

Figura 3: Representação de objetos anormais (HAN; KAMBER; PEI, 2011)

Um exemplo de aplicação seria a detecção de fraudes em cartões de créditos. Normalmente, o número de casos fraudulentos é pequeno comparado ao número de transações legítimas, técnicas de detecção de anomalias podem ser aplicadas nesse contexto para criar um perfil de transações legítimas para o usuário. Quando uma transação é efetuada, ela é comparada com o perfil de transações do usuário, se as características dessa transação forem diferentes do perfil criado, ela é identificada como fraudulenta (TAN; STEINBACH; KUMAR, 2009b).

Segundo Han, Kamber e Pei (2011) detecção de fraudes em cartões de créditos também podem ser identificadas, por exemplo ,pela quantidade de compras efetuadas, pelo local, tipo e frequência de compra.

Detecção de anomalias também está relacionado a detecção de novidades na evolução de um conjunto de dados. Como por exemplo, em mídias sociais a detecção de novidades podem identificar novas tendências. Essas tendências no início podem parecer discrepantes ou anormais. Nesse contexto, a detecção de anomalias e de novidades, compartilham algumas semelhanças no métodos de modelagem de detecção. A diferença entre as duas técnicas é que na detecção de novidades, quando se identifica uma nova tendência, elas são incorporadas no conjunto de da-

dos que possui comportamento normal, não sendo mais tratado como discrepante ou anormal (HAN; KAMBER; PEI, 2011).

2.3 Métodos ou Técnicas de Mineração de Dados

Segundo Maimom e RoKach (2010) uma terminologia bastante usada pela comunidade de aprendizado de máquina refere-se aos métodos de predição como aprendizagem supervisionada e uma parte dos métodos de descrição como aprendizagem não-supervisionada.

Aprendizagem de máquina é uma ampla área da Inteligência Artificial preocupada com o design e desenvolvimento de algoritmos que aprendem padrões presentes em dados fornecidos como entrada. Os padrões aprendidos são utilizados para fazer previsões relativas a novos dados (BAEZA-YATES *et al.*, 2011).

O aprendizado supervisionado requer aprendizagem de uma função a partir de dados de treinamento fornecidos como entrada. Esses dados de treino são então utilizados para aprender uma função de classificação, a qual pode ser utilizada para fazer predições de classes para novos dados cuja classe é desconhecida. Essa abordagem só funciona se a função aprendida puder utilizar dados que não foram vistos anteriormente e prever classes para eles com alta precisão (BAEZA-YATES *et al.*, 2011).

Neste caso, temos como objetivo prever o valor de um atributo para casos que ainda não foram vistos a partir de um conjunto de dados onde esse atributo possui valor, geralmente esse atributo é chamado de atributo alvo ou rótulo.

Na mineração de dados, as tarefas de classificação e regressão são métodos supervisionados (HAN; KAMBER; PEI, 2011).

O aprendizado não-supervisionado é diferente do supervisionado no sentido que não existem dados de treino como entrada (ZHANG *et al.*, 2011).

Segundo Tan, Steinbach e Kumar (2009a), o aprendizado não-supervisionado utiliza grandes quantidades de dados não rotulados para encontrar padrões existentes nestes dados, sendo bastante utilizado onde o conteúdo de grandes bases de dados não é conhecido antecipadamente. O principal interesse deste aprendizado é desvendar a organização dos padrões existentes nos dados através de *clusters* (agrupamentos) consistentes. A tarefa de associação e *clustering*, por exemplo, são classificadas como métodos não supervisionados (BRAMER, 2007).

Existem várias maneiras de classificar os vários métodos de mineração de dados (CIOS *et al.*, 2007; HAN; KAMBER; PEI, 2011; MAIMOM; ROKACH, 2010; WITTHEN; FRANK; HALL, 2011).

A seguir os métodos foram classificados de acordo com cada tarefa de mineração de dados (HAN; KAMBER; PEI, 2011).

2.3.1 Associação

De acordo com Cios *et al.* (2007) assim como *clustering*, a técnica de associação é classificada como método não-supervisionado. Regras de associação são utilizadas para encontrar associações, relações ou dependências entre itens em um conjunto de dados.

Um dos campos mais tradicionais de aplicação da regra de associação é a análise de cesta de compras, que consiste em analisar o comportamento de compra dos clientes para encontrar associação ou relação nos itens que formam sua cesta de compras (HAN; KAMBER; PEI, 2011).

Pode-se descobrir, por exemplo, que um percentual dos clientes compram pão e leite na mesma transação de compra. A descoberta dessas regras podem ser usadas para maximizar os lucros, criar novas estratégias de *marketing* ou adaptar o *layout* das gondolas da loja (CIOS *et al.*, 2007).

Geralmente, a tarefa de associação pode ser vista como um processo dividido em duas etapas: (1) Encontrar todos os conjuntos de itens frequentes cujo valor de suporte é maior do que ou igual ao mínimo apoio e (2) gerar fortes regras de associação a partir dos conjuntos de itens frequentes, essas regras devem satisfazer a um suporte mínimo e uma confiança mínima (HAN; KAMBER; PEI, 2011).

Segundo (HAN; KAMBER; PEI, 2011) *Apriori* é considerado o algoritmo básico para encontrar conjuntos de itens frequentes. O algoritmo foi proposto por (AGRAWAL; SRIKANT, 1994) para descobrir regras de associação entre itens em um grande banco de dados de transação.

2.3.2 Classificação

Segundo Larose (2004) os métodos de classificação, como por exemplo, árvore de decisão, redes neurais e k-vizinhos mais próximos são métodos supervisionados. Isso quer dizer que, primeiro, existe uma variável pré-alvo particular e, segundo, é passado como entrada para um algoritmo, um conjunto de dados em que o valor da variável alvo é conhecido. Assim o algoritmo terá um conjunto de dados em que os valores da variável alvo estará associado com os valores das variáveis de previsão. Abaixo são apresentados os métodos mais comuns utilizados na tarefa de classificação.

K-Vizinhos Mais Próximos: O algoritmo K-vizinhos mais próximos é frequentemente utilizado na tarefa de classificação, sendo utilizado também em estimação e previsão (LAROSE, 2004).

Segundo Witthen, Frank e Hall (2011) no algoritmo k-vizinhos mais próximos cada nova instância é comparada com as instâncias já existentes, utilizando uma função de distância, sendo que a nova instância é classificada de acordo com a instância mais próxima existente. Em alguns casos, mais de um vizinho mais próximo é utilizado para classificar uma nova instância.

Um exemplo deste método seria a classificação de tratamento para determinado paciente. A medição seria indicada de acordo com a classificação de tratamento de um conjunto de pacientes já conhecida, de acordo com os vizinhos mais próximo do novo paciente no conjunto de dados (LAROSE, 2004).

Árvore de Decisão: A árvore de decisão é um método amplamente utilizado na construção de modelos, sendo que sua representação é de fácil interpretação em comparação há outros modelos (BRAMER, 2007).

Uma árvore de decisão é representada por um fluxograma semelhante a uma árvore, sendo que os nós mais internos representam um teste em um atributo, os nós mais externos representam o rótulo da classe e o nó mais alto é chamado de nó raiz. Dependendo do algoritmo, a árvore de decisão gerada pode ser binária ou não binárias. A árvore de decisão pode ser utilizada para classificação, onde teríamos uma tupla(uma linha contendo várias colunas) e um atributo alvo ou rótulo sendo que seu valor não seja conhecido. Os valores da tupla são testados contra cada nó da árvore até o nó folha, formando um caminho. O nó folha detém o valor para o atributo alvo (HAN; KAMBER; PEI, 2011).

Segundo (LAROSE, 2004) o algoritmo C4.5 e o algoritmo CART(Classification and Regression Trees) são alguns dos algoritmos principais para a construção de árvores de decisão.

Redes Neurais: As redes neurais artificiais simulam o aprendizado que ocorre em redes de neurônios naturais. De forma geral, uma RNA é formada por várias unidade de processamento de entrada/saída conectadas, cada conexão tem um peso associado. Esses pesos são ajustados durante a etapa de aprendizagem, de modo que ela seja capaz de classificar corretamento um objeto (HAN; KAMBER; PEI, 2011).

Segundo Larose (2004) uma das vantagens da utilização das redes neurais são sua alta tolerância a dados com ruídos. Isso se deve a capacidade de aprendizado das redes neurais, proporcionando a rede neural, aprender a contornar esse tipo de situação.

Uma desvantagem seria sua difícil interpretabilidade. Por essa razão, inicialmente, as redes neurais não eram muito utilizadas para mineração de dados (HAN; KAMBER; PEI, 2011).

2.3.3 Agrupamento

De acordo com (HAN; KAMBER; PEI, 2011) existem vários algoritmos de agrupamento na literatura, e por essa razão, são difíceis de serem classificados. Sendo assim, os métodos de agrupamento ou *cluster* podem ser divididos em quatro categorias:

Métodos de Particionamento: Dado um conjunto de dados D, com n objetos e k sendo o número de agrupamento escolhido, um algoritmo de particionamento

agrupa os objetos em k grupos, sendo que $k \leq n$. Cada grupo contém pelo menos um objeto e cada objeto pertence a somente um grupo (HAN; KAMBER; PEI, 2011).

Existem vários algoritmos de agrupamento que podem ser encontrados na literatura. Os algoritmos *K-Means* e *K-medoids* são um dos mais conhecidos (VALÊNCIO *et al.*, 2011).

K-Means O *K-Means* utiliza o conceito de centróide. De maneira geral, neste algoritmo os usuários definem o número de *clusters k* que se deseja obter a partir do conjunto de dados inicial. No próximo passo, o algoritmo seleciona de maneira arbitrária k objetos do conjunto de dados como centróides dos *cluster* iniciais. Para cada registro restante, é calculada a distância ou similaridade entre o registro analisado e cada centróide de cada *cluster*. O registro analisado é inserido no grupo onde sua distância é menor, ou seja, onde sua similaridade é maior dentre os registros dos demais grupos. A cada iteração o centro do *cluster* ou centróide é recalculado (BRAMER, 2007; HAN; KAMBER; PEI, 2011).

Segundo (HAN; KAMBER; PEI, 2011) existem algumas variações do método *K-Means*. As variações diferem na escolha do valor de k, no cálculo da dissimilaridade e nas estratégias de cálculo dos centros dos *cluster*. O método *K-modes* é um exemplo de variante do *K-Means*. Algumas diferenças são: agrupamento de dados categóricos e utilização de novas medidas de dissimilaridade para lidar com esses dados.

K-Medoids Segundo (HAN; KAMBER; PEI, 2011) o método *K-Means* é sensível a *outliers*. Em vez de pegar o valor médio dos objetos em um *cluster* como referência, como acontece no *k-Means*, o método *K-Medoids* utiliza objetos

reais para representar os *clusters*. Os objetos restantes são agrupados com o objeto representativo de acordo com a similaridade entre os dois objetos. O particionamento é realizado utilizando o princípio de minimizar a soma das diferenças entre cada objeto e o objeto representativo. Essa estratégia diminui a sensibilidade a *outliers*.

Métodos Hierárquicos: Métodos hierárquicos criam uma decomposição hierárquica que pode ser visualizada por uma representação gráfica dos dados, chamada de dendograma. Um método hierárquico pode ser classificado em dois tipos: aglomerativo ou divisivo (LAROSE, 2004; CIOS *et al.*, 2007).

Aglomerativo Na abordagem aglomerativa ou abordagem *bottom-up* inicialmente cada objeto é considerado um agrupamento. Os grupos mais próximos ou similares são mesclados em *clusters* cada vez maiores, formando um novo grupo ou *cluster* combinado. O processo se repete até que se forme um único cluster ou quando o valor limite pré-definido seja atingido. Podemos citar, como exemplo, o algoritmo AGNES(*AGglomerative NESting*) que utiliza a abordagem aglomerativa (HAN; KAMBER; PEI, 2011).

Divisivo Na abordagem divisiva ou abordagem *top-down* inicialmente todos os objetos são tratados como um único *cluster*. Em cada iteração recursiva, os *clusters* são divididos em grupos menores até que cada objeto represente um *cluster* ou uma condição de parada seja atendida. Podemos citar, como exemplo, o algoritmo DIANA(*DIvisive ANAlysis*) que utiliza a abordagem divisiva (HAN; KAMBER; PEI, 2011).

Métodos Baseados em Densidade: Segundo (HAN; KAMBER; PEI, 2011) os métodos de particionamento e hierárquicos tem dificuldade em encontrar gru-

pos de forma arbitrária, como por exemplo, a figura 4 ilustra *cluster* dispersos em forma de "S"e *cluster* ovais. Tais métodos são projetados para os *cluster* em forma esférica. A ideia geral por trás dos métodos baseado em densidade é que nessa abordagem os *clusters* são considerados regiões de alta densidade separadas por regiões de baixa densidade. Cada ponto deve ter um número mínimo de pontos dentro de sua vizinhança. Existem diversos algoritmos que utilizam essa abordagem, podemos citar, DBSCAN(*Density-Based Spatial Clustering of Applications with Noise*), OPTICS(*Ordering Points to Identify the Clustering Structure*) e DENCLUE (*DENsity-based CLUstEring*).

Figura 4: Representação de clusters de forma arbitrária (HAN; KAMBER; PEI, 2011)

Métodos Baseados em Grade: Os métodos baseados em grade quantifica o espaço objeto em um número finito de células formando uma grade, onde são realizadas todas as operações de agrupamento. Este método é considerado rápido, pois não depende do número de objetos de dados, mas depende do tamanho da grade. Além disso, os métodos baseados em grade podem ser integrados com outros métodos de agrupamento, como por exemplo, os métodos baseado em den-

sidade e hierárquicos. Um exemplo de métodos baseados em grade, é o método STING(*STatistical INformation Grid*) (HAN; KAMBER; PEI, 2011).

A classificação de um método de agrupamento em uma determinada categoria é difícil, pois um algoritmo pode conter princípios de vários métodos de agrupamento e além disso, algumas aplicações pode requerer a integração de várias técnicas de agrupamento (HAN; KAMBER; PEI, 2011).

2.3.4 Detecção de Anomalias

De acordo com (HAN; KAMBER; PEI, 2011) os métodos ou técnicas de detecção de anomalias podem ser dividias em:

Técnicas Baseadas em Modelos: Algumas técnicas de detecção de anomalias constroem um modelo dos dados. Logo, um objeto que não se encaixe bem dentro de um modelo é considerado um objeto anômalo, por exemplo, se o modelo utilizado for um conjunto de *clusters* ou grupos, o objeto anômalo não pertencerá a nenhum grupo ou se o modelo usado for de regressão, um objeto anômalo seria um objeto com o valor relativamente diferente do seu valor previsto (TAN; STEINBACH; KUMAR, 2009b).

Segundo Han, Kamber e Pei (2011) detecção de anomalias pode ser modelado como um problema de classificação. Neste caso, os especialistas podem rotular apenas os objetos normais, assim qualquer objeto que não corresponda ao modelo de objetos normais são considerados anômalos, ou vice-versa.

Em um conjunto de dados, as amostras de objetos anômalos são menores em comparação a amostra de objetos normais, sendo consideradas raras. Neste caso, a raridade das anomalias devem ser levada em consideração na escolha tanto da téc-

nica de classificação quanto das medidas que serão utilizadas para a avaliação. Por isso, a falta de amostras de objetos anômalos podem prejudicar os classificadores construídos para tal finalidade (HAN; KAMBER; PEI, 2011; TAN; STEINBACH; KUMAR, 2009b).

Fatores como distribuição estatísticas dos dados ser desconhecida e nenhum dados de treino estar disponível dificultam a construção de modelos. Nestes casos, as técnicas descritas abaixo podem ser utilizadas (TAN; STEINBACH; KUMAR, 2009b).

Como dito anteriormente, técnicas baseadas em modelos constroem um modelo do dados. Métodos estatísticos criam um modelo para os dados normais, os objetos que não se ajustam ao modelo gerado são considerados anômalos.

Técnicas Baseadas em Proximidades: Nessas técnicas é utilizado o conceito de medida de proximidade entre objetos. De forma geral, objetos anômalos são aqueles que estão distantes dos demais objetos. Os dados podem ser representados como um desenho bi ou tridimensional. Neste caso, os objetos anômalos podem ser identificados por pontos que estejam distantes da maioria (TAN; STEINBACH; KUMAR, 2009b).

2.4 Descoberta de Conhecimento em Banco de Dados - KDD

De acordo com Maimom e RoKach (2010) existe uma grande quantidade de conhecimento oculto esperando para ser descoberto. O volume de dados armazenados em aplicações de negócios, medicina, ciência e engenharia, é cada vez maior. Nesse contexto, extrair conhecimento útil desse grande volume de dados é uma tarefa importante.

O KDD é um processo organizado com várias etapas que visa extrair conhecimento ou padrões úteis que estejam ocultos em grandes bases de dados (MAIMOM; ROKACH, 2010).

De acordo com Tan, Steinbach e Kumar (2009b) o KDD é um processo geral de conversão de dados brutos em informações úteis, sendo assim, a mineração de dados é uma parte integral do processo KDD. O KDD é um processo iterativo e interativo, composto por nove etapas, sendo possível retornar a uma etapa anterior caso necessário. A figura a seguir ilustra o processo KDD em nove etapas.

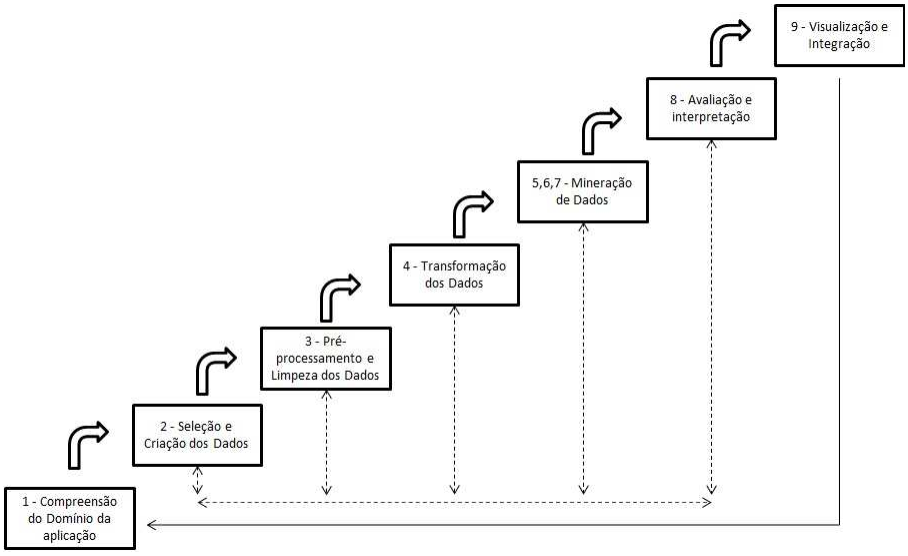

Figura 5: Etapas do Processo KDD (MAIMOM; ROKACH, 2010)

De acordo com (MAIMOM; ROKACH, 2010) as etapas do processo são as seguintes:

1. Compreensão do Domínio da Aplicação: nesta etapa, é feita o levamento dos requisitos necessários. Os responsáveis pelo projeto de KDD precisam

entender e definir os objetivos do usuário final e da aplicação em que o processo irá atuar. A medida que se avança nas etapas posteriores, a revisão deste passo pode ser necessária.

2. Seleção e Criação do Conjunto de Dados sobre o qual a Descoberta será realizada: nesta etapa, é determinado quais dados serão utilizados para a descoberta de conhecimento de acordo com os objetivos definidos na etapa anterior. Isso incluí saber quais os dados que estão disponíveis, obter dados adicionais se for necessário, para que sejam integrados em um único conjunto de dados que contenha os atributos necessários para o processo. A escolha dos atributos é muito importante, pois se faltar algum atributo, todo o processo pode ser comprometido.

3. Pré-processamento e Limpeza de Dados: A qualidade dos dados pode influenciar a eficiência dos algoritmos de mineração. Nesta etapa são realizadas tarefas de limpeza de dados, como a manipulação de valores incompletos(ausência de atributos de interesse, ausência de valores) e remoção de ruídos(erros aleatórios). Também são utilizados métodos de redução ou transformação para diminuir o número de variáveis envolvidas no processo.

4. Transformação de Dados: a transformação de dados antecede a etapa de Mineração de Dados. Nesta etapa, os dados são transformados ou formatados utilizando-se métodos como por exemplo, redução de dimensão, transformação de atributo ou operações de agregação. As quatros etapas seguintes estão relacionadas com a etapa de Mineração de Dados, tendo como foco aspectos algorítmicos de acordo com a peculiaridade do projeto.

5. Escolha da Tarefa Apropriada de Mineração de Dados: nesta etapa é escolhida o tipo de mineração de dados como por exemplo, classificação, re-

gressão ou agrupamento, de acordo com os objetivos definidos na primeira etapa.

6. Escolha do Algoritmo de Mineração de Dados: nesta etapa é escolhido qual será o método específico para ser utilizado na busca de padrões.

7. Utilização do Algoritmo de Mineração de Dados: nesta etapa o algoritmo de mineração de dados é executado de fato. Pode ser que o algoritmo tenha que ser executado várias vezes até que o resultado obtido seja satisfatório.

8. Avaliação: nesta etapa é avaliado os padrões obtidos com relação aos objetivos definidos na primeira etapa do processo. O foco principal da etapa está em compreender o modelo induzido.

9. Utilização do Conhecimento Descoberto: nesta etapa, o conhecimento obtido deverá ser incorporado em outro sistema para ações futuras. O sucesso desta etapa determina a eficácia de todo o processo. Aqui temos agora que lidar com dados dinâmicos, pois em um sistema em produção, a base de dados está sujeita a alterações de caráter estrutural, como por exemplo, adição ou remoção de um atributo em uma tabela, e alterações no dados, como por exemplo, *update* em um atributo passa a ter um valor que não foi considerado anteriormente pelo processo.

2.5 PostgreSQL - Sistema Gerenciador de Banco de Dados Objeto-Relacional

O PostgreSQL é um SGBDOR (Sistema Gerenciador de Banco de Dados Objeto-Relacional de código aberto, derivado do pacote POSTGRE, desenvolvido pelo Departamento de Ciência da Computação da Universidade da Califórnia em *Berkeley*, que já tem mais de 15 anos de desenvolvimento. Nesse tempo, o *Post-*

greSQL tornou-se sinônimo de confiabilidade, integridades de dados e conformidade de padrões entre seus usuários. É suportado por várias plataformas como por exemplo, GNU/Linux, Unix(AIX, BSD, HP-UX, SGI IRIX, Mac OS X, Solaris, Tru64), e MS Windows. Possui suporte completo a chaves estrangeiras, junções, visões, gatilhos e procedimentos armazenados. O PostgresSQL possui uma licença liberal, isso quer dizer que pode ser utilizado, modificado e distribuído por qualquer pessoa para qualquer finalidade, seja particular, comercial ou acadêmica, livre de encargos. O PostgreSQL possui várias extensões, uma delas é o PostGis, que também foi utilizada no desenvolvimento deste trabalho (GROUP, 2014).

2.5.1 PostGis

PostGIS é uma extensão ao SGBDOR PostgreSQL, que permite o armazenamento e manipulação de objetos GIS(Sistemas de Informação Geográfica). Ele inclui suporte para índices espacias *GiST-based R-Tree* e funções para análise e processamento de objetos GIS (WEBGIS, 2014).

A extensão PostGIS possui várias funções de relacionamento de geometria, como por exemplo:

***ST_Intersects(geometria A,geometria B)*:** Retorna *TRUE* se houver cruzamento espacial(compartilhar qualquer região do espaço) entre as geometrias A e B e retorna *FALSE* no caso das geometrias serem disjuntas.

Podemos citar também algumas funções de processamento da geometria, por exemplo:

***ST_Transform(geometria A, integer srid)*:** Retorna uma nova geometria com suas coordenadas transformadas ao SRID referenciado pelo parâmetro inteiro.

***ST_Area(geometria)*:** Retorna a área da geometria, se é um polígono ou multi-polígono.

***ST_Intersection(geometria A,geometria B)*:** Retorna uma geometria que representa a interseção atribuída do ponto das Geometrias e retorna *null* se as geometrias não compartilham qualquer espaço.

As funções acima foram necessárias na etapa de pré-processamento dos dados do processo KDD.

2.6 Zoneamento Ecológico-Econômico do Estado de Minas Gerais

O Zoneamento Ecológico Econômico do Estado de Minas Gerais – ZEE-MG é um macro diagnóstico do estado de Minas Gerais, apresentado como uma ferramenta web de geoprocessamento.

O estado de Minas Gerais possui 853 municípios, com uma superfície de 586.852,35 km^2 e aproximadamente 19.597.330 de habitantes. O estado é dividido por nove regionais administrativas do Conselho de Política Ambiental - COPAM, sendo que sua finalidade é "deliberar sobre diretrizes, políticas, normas regulamentares e técnicas, padrões e outras medidas de caráter operacional, para preservação e conservação do meio ambiente e dos recursos ambientais, bem como sobre a sua aplicação pela Secretaria de Estado de Meio Ambiente e Desenvolvimento Sustentável, pelas entidades a ela vinculadas e pelos demais órgãos locais" (SEMAD, 2014).

Os contornos das regionais do COPAM podem ser visualizadas na figura abaixo.

Figura 6: Contorno das Regionais do COPAM

Segundo Scolforo *et al.* (2008) o "ZEE-MG consiste na elaboração de um diagnóstico dos meios geo-biofísico e sócio-econômico-jurídico-institucional, gerando respectivamente duas cartas principais, que são mapas geográficos ou topográficos, a carta de Vulnerabilidade Ambiental e a Carta de Potencialidade Social,

que sobrepostas irão conceber áreas com características próprias, determinando o Zoneamento Ecológico-Econômico do Estado". O ZEE-MG tem a coordenação da Secretaria de Estado de Meio Ambiente e Desenvolvimento Sustentável, participação de todas as Secretarias de Estado de Minas, de outras entidades e da sociedade civil.

Esses diagnósticos mostram a variação espacial das condições naturais e sociais, fornecendo uma importante ferramenta para o planejamento e gestão territorial. Esta ferramenta está disponível na internet através do endereço `http://http://www.zee.mg.gov.br/`.

2.6.1 Objetivos

A ferramenta ZEE-MG tem como objetivo apoiar a gestão territorial fornecendo subsídios técnicos à definição de áreas prioritárias para a proteção e conservação da biodiversidade e para o desenvolvimento, segundo critérios de sustentabilidade econômica, social, ecológica e ambiental. Sendo assim, o ZEE-MG tem importância no planejamento e elaboração das políticas públicas e das ações em meio ambiente, respeitando as características de cada área de desenvolvimento, possibilitando a melhoria na qualidade de vida e dos serviços prestados para a população do estado (SCOLFORO *et al.*, 2008).

2.6.2 Principais Produtos

O Zoneamento Ecológico-Econômico do Estado de MG é resultado da sobreposição de duas cartas:

- Vulnerabilidade Natural (VN): é a incapacidade do meio ambiente de resistir ou recuperar-se de impactos negativos antrópicos, determinada com base em informações presentes e atuais.

- Vulnerabilidade Social (PS): é o conjunto de condições atuais, medido pelas dimensões produtiva, natural, humana e institucional, que determina o ponto de partida de um município ou de uma microrregião para alcançar o desenvolvimento sustentável.

Através da sobreposição da VN e PS, a ferramenta ZEE-MG é capaz de fornecer informações importantes sobre um município ou uma região, como por exemplo, se um município possui alta PS, significa que ele tem condições favoráveis para alcançar o desenvolvimento sustentável ou se uma região apresenta alta VN, isso significa que o conjunto de suas características físicas e bióticas a torna mais vulnerável à utilização de seus recursos. A alta VN pode ser um fator limitante para o licenciamento de atividades potencialmente impactantes, sendo neste caso necessárias medidas mais severas de mitigação e controle de impactos (SCOLFORO *et al.*, 2008).

Além das cartas de VN e PS, o ZEE-MG apresenta outros produtos, como as cartas de Qualidade Ambiental, Risco Ambiental, Áreas Prioritárias para Conservação, Áreas Prioritárias para Recuperação, Nível de Comprometimento dos Recursos Hídricos Superficiais e Nível de Comprometimento dos Recursos Hídricos Subterrâneos (SCOLFORO *et al.*, 2008).

2.6.3 Aplicações do ZEE-MG

Os produtos disponibilizados pelo ZEE-MG, são capazes de:

- Orientar o direcionamento da ocupação do território para áreas aptas para suportar determinado uso, desestimulando o desenvolvimento em áreas inaptas;
- Indicar áreas aptas que necessitam ser recuperadas antes da sua utilização;
- Apoio à gestão territorial, orientando as decisões do poder público e da sociedade civil;
- Nas análises de processos de regularização e gestão ambiental;
- Em outras finalidades, por se constituir um completo banco de dados sobre a situação sócio-econômica e ambiental do estado de MG.

Todas as informações desta seção podem ser encontradas no link `http://trilhasdosaber.meioambiente.mg.gov.br/Zee/`, disponibilizado pelo governo do estado de Minas Gerais.

3 METODOLOGIA

De acordo com Jung (2009) “a pesquisa e desenvolvimento experimental é utilizado para se obter novos produtos, processos e conhecimentos.” Dessa forma, a pesquisa é utilizada como ferramenta pra descobrir novos conhecimentos, enquanto que o desenvolvimento é a aplicação desses novos conhecimentos para se obter resultados práticos.

3.1 Tipo de Pesquisa

Ainda segundo Jung (2009), a pesquisa pode ser classificada como básica ou aplicada. Este trabalho é classificado como uma pesquisa aplicada, pois é aplicada uma técnica de mineração de dados na base de dados do ZEE-MG, a fim de identificar características e padrões ocultos. Como em uma pesquisa aplicada o resultado a ser obtido é a solução do problema estudado, pretendemos auxiliar no planejamento e desenvolvimento de políticas públicas no governo do estado de Minas Gerais.

3.2 Visão Geral

Os dados utilizados no estudo são originados da carta de potencialidade social do ZEE-MG, que é formada por quatro componentes: produtivo, natural, humano e institucional.

De acordo com Scolforo *et al.* (2008) “a potencialidade social pode ser definida como o conjunto de condições atuais, medido pelas dimensões produtiva,

natural, humana e institucional, que determina o ponto de partida de um município ou de uma microrregião para alcançar o desenvolvimento sustentável".

Um município por exemplo possui, indústrias, agroindústrias, hospitais, instituições de cursos superiores, áreas de preservação ambiental, enquanto que outros não. O somatório de todas essas capacidades essas capacidades indica a potencialidade social de cada município (SCOLFORO *et al.*, 2008).

Cada componente (produtivo, natural, humano e institucional) é formado por um conjunto de fatores condicionantes, sendo que cada fator condicionante é formado por um conjunto de indicadores. Segundo Scolforo *et al.* (2008) cada indicador é formado por um conjunto de variáveis originadas de orgãos oficiais do estado de MG ou do governo federal, bem como de instituições oficiais responsáveis pela divulgação de dados estatísticos, onde aos dados são os mais atualizados possíveis, sobre condições produtivas, humanas, naturais e institucionais de cada município do estado de MG.

Assim tem-se uma estrutura metodológica de potencialidade social para diagnosticar a realidade dos municípios de Minas Gerais. Essa estrutura pode ser visualizada na figura 7.

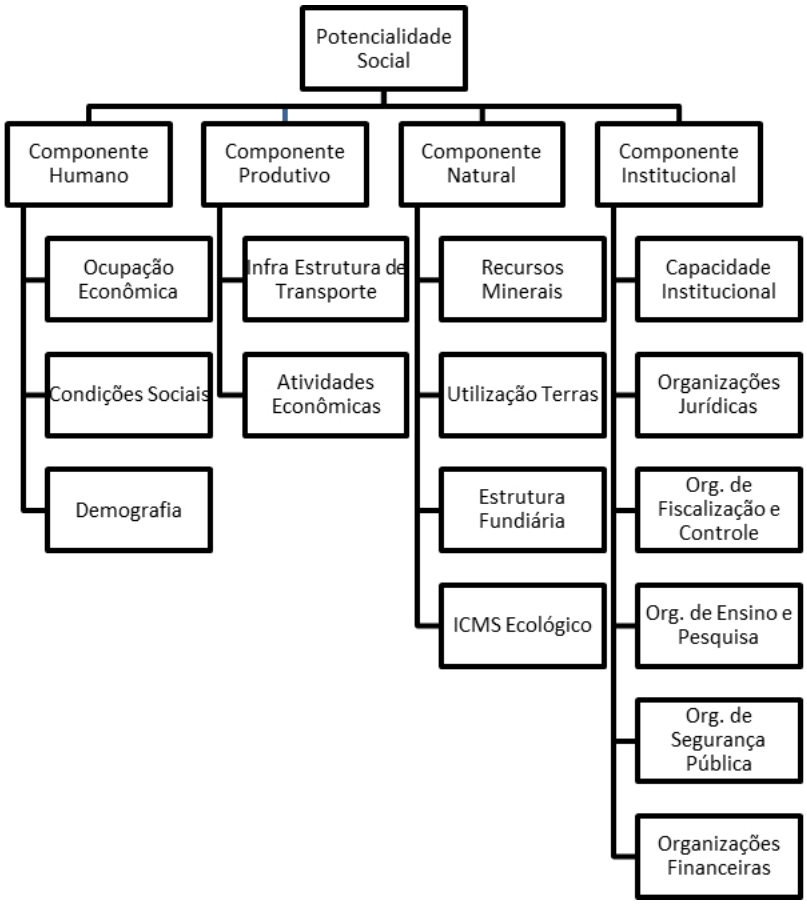

Figura 7: Estrutura metadológica (Componentes e Fatores Condicionantes) da Carta de Potencialidade Social (SCOLFORO *et al.*, 2008)

O componente humano trata dos aspectos ligados à satisfação das necessidades humanas, à melhoria da qualidade de vida e justiça social, direcionados à construção da cidadania. O componente humano, com seus fatores condicionantes e seus respectivos indicadores são apresentados na figura 8 (OLIVEIRA *et al.*, 2008).

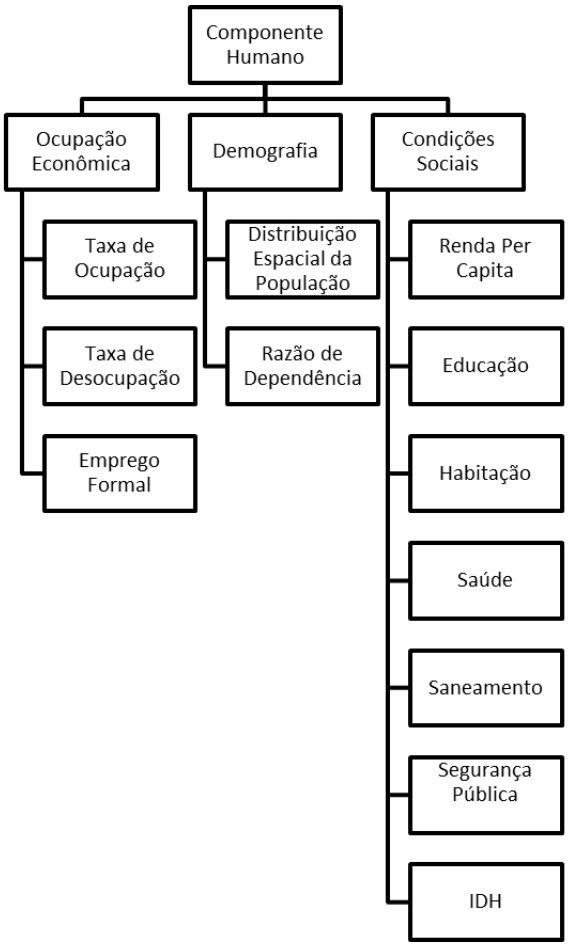

Figura 8: Estrutura metadológica - Componente Humano (OLIVEIRA *et al.*, 2008)

O componente produtivo é determinado pelas condições da infraestrutura de transportes e atividades econômicas. O componente produtivo, com seus fatores condicionantes e seus respectivos indicadores são apresentados na figura 9. (CALEGARIO *et al.*, 2008).

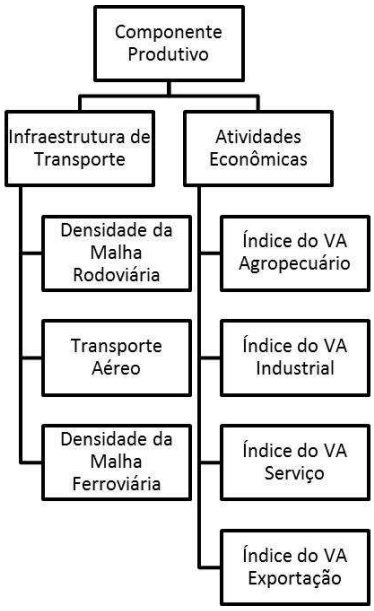

Figura 9: Estrutura metadológica - Componente Produtivo (CALEGARIO *et al.*, 2008)

O componente natural compreende a exploração mineral, a intensidade de uso da terra e a forma de ocupação e convervação do meio ambiente. O componente natural, com seus fatores condicionantes e seus respectivos indicadores são apresentados na figura 10 (AMâNCIO *et al.*, 2008).

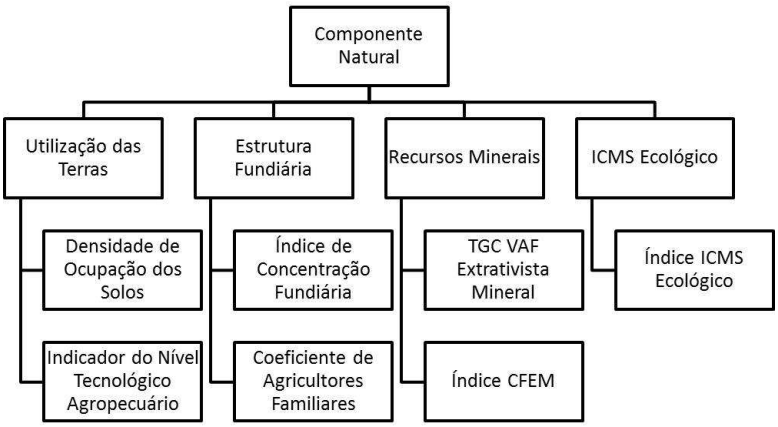

Figura 10: Estrutura metadológica - Componente Natural (AMâNCIO *et al.*, 2008)

O componente institucional se refere à capacidade institucional dos municípios de atender aos cidadãos em suas demandas, sejam de caráter social, ecológico, econômico, político e cultural (SALAZAR *et al.*, 2008).

O componente institucional, com seus fatores condicionantes e seus respectivos indicadores são apresentados na figura 11.

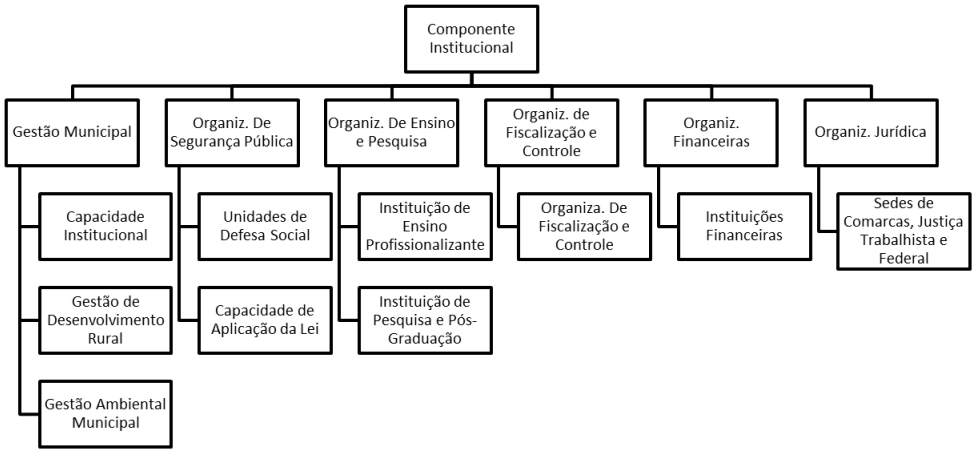

Figura 11: Estrutura metadológica - Componente Institucional (SALAZAR *et al.*, 2008)

Cada indicador é representado por uma entidade no banco de dados do ZEE-MG, onde cada entidade apresenta uma categorização. A tabela 1 exemplifica as cinco categorias utilizadas para classificação dos municípios.

CATEGORIA/PONTOS	TIPO DE POTENCIALIDADE SOCIAL
A = 5	Ponto de Partida em Condições Muito Favoráveis
B = 4	Ponto de Partida em Condições Favoráveis
C = 3	Ponto de Partida em Condições Pouco Favoráveis
D = 2	Ponto de Partida em Condições Precárias
E = 1	Ponto de Partida em Condições Muito Precárias

Tabela 1: Categorização dos Municípios

Segundo Scolforo *et al.* (2008) para categorizar os municípios, comparou-se os dados de um município com os dados dos 853 municípios de Minas Gerais. A categorização é representada por uma valor mínimo e um valor máximo, que por sua vez, é representado pelas letras do alfabeto "A, B, C, D, E". Cada letra representa, respectivamente, um tipo de potencialidade social, "Ponto de Partida em Condições Muito Favoráveis", "Ponto de Partida em Condições Favoráveis",

"Ponto de Partida em Condições Pouco Favoráveis", "Ponto de Partida em Condições Precárias"e "Ponto de Partida em Condições Muito Precárias".

3.3 Procedimentos Metodológicos

O desenvolvimento do trabalho foi baseado nas etapas do processo de descoberta de conhecimento em banco de dados descrito por (MAIMOM; ROKACH, 2010). O processo é composto pela etapa de compreensão do domínio da aplicação, seleção e criação do conjunto de dados sobre o qual a descoberta será realizada, pré-processamento e limpeza de dados, transformação de dados, mineração de dados que envolvem as etapas (escolha da tarefa apropriada de mineração de dados, escolha e execução do algoritmo), avaliação e por último utilização do conhecimento descoberto.

4 DESENVOLVIMENTO

Esta seção aborda as etapas executadas para o desenvolvimento do trabalho. As etapas são baseadas no processo KDD descrito por (MAIMOM; ROKACH, 2010).

4.1 Compreensão do Domínio da Aplicação

O processo de descoberta de conhecimento foi realizado na carta de potencialidade social, formada por quatro componentes: humano, natural, produtivo e institucional. Cada componente é constituído de indicadores que caracterizam as condições atuais (humana, natural, produtiva e institucional) de um município ou microrregião, determinando assim suas condições para alcançar o desenvolvimento sustentável.

Foi necessário criar uma base de dados com extensão espacial para armazenar as informações referentes aos indicadores da potencialidade social, bem como a criação de algumas entidades adicionais para que fosse possível a realização de uma análise quantitativa das informações.

4.2 Seleção e Criação do Conjunto de Dados

Foi criada uma base de dados contendo todos os indicadores que compõem cada componente(produtivo, humano, institucional e natural) da carta de potencialidade social no SGBD PostgreSQL 9.2.4 com extensão espacial PostGis 1.5.8. Cada indicador é representado por uma entidade ou tabela na base de dados criada, totalizando 36 entidades. Os atributos PK e FK, significam respectivamente,

chave primária e chave estrangeira. Podemos visualizar um exemplo da estrutura das entidades criadas na tabela 2.

ATRIBUTO	TIPO	DESCRIÇÂO	PK	FK
gid	integer	Identificador único da entidade	X	
gridcode	integer	Categoria do indicador.		
the_geom	geometry	Coluna georreferenciada		
descricao	text	Descrição da Categoria do indicador.		

Tabela 2: Entidade indicador

Todas as entidades contem cinco atributos, todas com propriedade *not null*, ou seja, tem-se a garantia de que não há valores vazios.

Como dados adicionais, foram criados duas entidades: município e copam. Os atributos da entidade município são descritos a seguir:

ATRIBUTO	TIPO/PRECISÂO	DESCRIÇÂO	PK	FK
gid	integer	Identificador único da entidade.	X	
nm_municipio	character varying(30)	Nome do Município.		
id_copam	integer	Identificador da entidade copam.		X
the_geom	geometry	Coluna georreferenciada.		
cod_ibge	integer	Código IBGE.		

Tabela 3: Entidade município.

Com a entidade município é possível realizar uma junção entre cada registro, através do atributo *the_geom* da entidade município e de cada entidade indicador. Assim, temos a classificação dos indicadores de cada município.

Com a entidade copam, foi possível saber quais municípios pertencem a qual COPAM, tabela 4.

ATRIBUTO	TIPO/PRECISÂO	DESCRIÇÂO	PK	FK
id_copam	integer	Identificador único da entidade.	X	
nm_copam	character varying(30)	Nome da Regional do COPAM.		
the_geom	geometry	Coluna georreferenciada.		

Tabela 4: Entidade copam.

Isso foi possível realizando a junção da entidade copam e município, através do atributo id_copam, sendo que na entidade município esse atributo é chave estrangeira para a entidade copam.

4.3 Pré-processamento, Limpeza e Transformação dos Dados

Para obter a classificação dos indicadores de um município foi construída uma consulta *sql* para fazer a junção entre as entidades município e cada entidade que representa cada indicador. O *select* abaixo foi executado 36 vezes, ou seja, para todos os indicadores, obtendo assim a classificação de cada indicador para todos os 853 municípios de Minas Gerais. Com o *select* abaixo, por exemplo, podo-se obter a classificação do indicador habitação, do componente humano, de cada município do estado de Minas Gerais.

```
1  SELECT *
2  FROM
3    (SELECT m.gid AS gid_mun,
4            ST_Area(ST_Transform((ST_Intersection(m.the_geom,ind.
                 the_geom)), 3310))/10000 AS area,
5            max(ST_Area(ST_Transform((ST_Intersection(m.the_geom,ind
                 .the_geom)), 3310))/10000) over (partition BY m.gid)
                  AS area_max,
6            ind.gridcode,
7            m.nm_municipio
8     FROM zee.ch_zee_habitacao ind,
9          municipio m
10    WHERE ST_Intersects(m.the_geom,ind.the_geom)) x
11 WHERE x.area = x.area_max
12 ORDER BY 1;
```

Existem casos em que um município pode ter mais de uma classificação para um indicador, isso ocorre por que a geometria do indicador não necessariamente se sobrepõe perfeitamente com a geometria do município. Assim a geometria de um município pode ter mais de uma interseção com a geometria do indicador. Para resolver este problema, foi calculada todas as interseções entre a geometria do município com a geometria do indicador (linha 4). Com isso, foi calculada qual era a maior interseção dentre as áreas exibidas, para definir de fato a interseção que realmente representa o valor daquele indicador em relação ao município (linha 5).

Como resultado, foi gerada uma tabela unificada contendo a classificação de todos os indicadores para cada município. Por exemplo, pela tabela 5, podemos visualizar que o município Sapucaí-Mirim possui condições de saneamento muito precárias (valor 1), de habitação pouco favoráveis (3) e condições precárias (2) de distribuição espacial da população.

NM_MUNICÌPIO	CH_SANEAMENTO	CH_HABITAÇÂO	CH_DIST_POPULAÇÂO	...
Sapucaí-Mirim	1	3	2	...
Camanducaia	1	5	3	...
Extrema	1	4	3	...
Toledo	5	3	1	...
Itapeva	1	4	2	...
⋮	⋮	⋮	⋮	

Tabela 5: Exemplo da Tabela Unifica Gerada.

A tabela gerada original contém 853 linhas representando todos os municípios de Minas Gerais, e 36 colunas representando os 36 indicadores.

4.4 Escolha da Tarefa e do algoritmo de Mineração de Dados

A tarefa de mineração de dados escolhida foi a de agrupamento ou *clustering*. Esta tarefa é classificada como descritiva, incluindo aspectos de aprendizado

não-supervisionado. O objetivo de utilizar essa tarefa é descobrir objetos com características semelhantes, no nosso contexto, agrupar municípios que possuem características semelhantes em relação aos vários indicadores.

O método ou algoritmo escolhido foi o *K-Means*. De maneira geral, neste algoritmo os usuários definem o número de *clusters k* que se deseja obter a partir do conjunto de dados inicial. No próximo passo, o algoritmo seleciona de maneira arbitrária k objetos do conjunto de dados como centróides dos *clusters* iniciais. Para cada registro restante, é calculada a distância ou similaridade entre o registro analisado e cada centróide de cada *cluster*. O registro analisado é inserido no grupo onde sua similaridade é maior dentre os demais grupos. A cada iteração o centro do *cluster* ou centróide é recalculado (HAN; KAMBER; PEI, 2011).

Os passos do algoritmo *K-Means* são apresentados abaixo (BRAMER, 2007):

1. Escolher um valor para k
2. Selecionar k objetos de maneira arbitrária. Esses k objetos serão o conjunto inicial de k centróides.
3. Calcular a distância ou similaridade de cada objeto em relação aos centróides.
4. Recalcular os centróides dos k *clusters*.
5. Repetir os passos 3 e 4 até que os centróides não se movam mais.

4.5 Utilização do algoritmo de Mineração de Dados

O algoritmo K-means foi executado por meio da ferramenta Weka. A ferramenta *Waikato Environment for Knowledge Analysis* (Weka) possui uma coleção

de algoritmos de aprendizado de máquina e ferramentas de pré-processamento de dados. Foi originalmente desenvolvida pela Universidade de Waikato (Nova Zelândia), escrito na linguagem Java, de código aberto, permitindo assim que seu código possa ser alterado.

A ferramenta Weka possui um formato de entrada padrão, o formato ARFF. Assim foi necessário transformar as informações da tabela unificada gerada para este formato específico. Pode-se visualizar abaixo um exemplo do arquivo com extensão ARFF gerado. O arquivo original contém a relação de todos os municípios com os 36 indicadores.

```
1  @relation municipio-indices
2
3  @attribute NM_MUNICIPIO string
4  @attribute CP_EXPORTACAO numeric
5  @attribute CH_EDUCACAO numeric
6  @attribute CH_TX_OCUPACAO numeric
7  @attribute CN_CFEM,CN_ICMS numeric
8  @attribute CI_DEFESA_SOCIAL numeric
9  .
10 .
11 .
12
13 @data
14
15 'Sapucai-Mirim',1,3,4,4,5,1
16 'Camanducaia',5,3,3,1,5,1
17 'Extrema',5,5,4,1,5,1
18 'Toledo',1,1,4,1,5,1
19 'Itapeva',4,2,3,1,5,1
20 .
```

```
.
.
```

Em cada execução do algoritmo K-Means na ferramenta Weka, foi passado o valor de k = 5, ou seja, foram gerados cinco *clusters*. O valor inicial de k foi determinado baseado no número de classificações possíveis para cada indicador, ver tabela 1. Em relação a função de distância, foi escolhida a distância euclidiana e foi mantida a ordem em que as instancias foram passadas para o algoritmo.

Depois que os *clusters* são gerados pelo *K-means*, é possível salvar os resultados em um arquivo de saída. As informações contidas nesses arquivos, um exemplo pode ser visualizado logo abaixo, foram importantes nas análises quantitativas.

```
@relation 'municipio-indices-weka.filters.unsupervised.attribute.
    Remove-R2-4,6-12,15-26_clustered'

@attribute Instance_number numeric
@attribute NM_MUNICIPIO string
@attribute CH_EDUCACAO numeric
@attribute CH_TX_OCUPACAO numeric
@attribute CP_EXPORTACAO numeric
@attribute CI_DEFESA_SOCIAL numeric
@attribute Cluster {cluster0,cluster1,cluster2,cluster3,cluster4}

@data
0,Sapucai-Mirim,3,4,1,1,cluster1
1,Camanducaia,3,3,5,1,cluster3
2,Extrema,5,4,5,1,cluster3
3,Toledo,1,4,1,1,cluster0
4,Itapeva,2,3,4,1,cluster3
.
```

18 .

Os indicadores envolvidos em cada execução foram escolhidos de acordo com as regras de associação geradas e classificadas como conhecimento útil e inesperado e de acordo com o conhecimento prévio do especialista. As regras geradas a partir da aplicação da técnica de análise de associação no ZEE-MG podem ser encontradas em (SIDNEY, 2010).

4.6 Avaliação dos Resultados e Discussão

Nesta seção, são apresentados os resultados obtidos através da aplicação da técnica de *clustering* na base de dados do ZEE-MG, considerando as informações relacionadas a potencialidade social. Os indicadores escolhidos para a geração dos clusters são apresentados abaixo.

Componente Humano

- Taxa de Ocupação: Permite tanto o acompanhamento de tendências e de variações no nível de ocupação como, também, subsidiar a formulação de estratégias e políticas de geração de emprego e renda.
- Educação: Está relacionado com as condições econômicas e sociais de um município, estando fortemente ligado ao seu desenvolvimento econômico.

Componente Natural

- CFEM (Compensação Financeira pela Exploração de Recursos Minerais): É devida aos Estados, ao Distrito Federal, aos Municípios, e aos órgãos da administração da União, como contraprestação pela utilização econômica

dos recursos minerais em seus respectivos territórios. Suas receitas deverão ser aplicadas em projetos, que direta ou indiretamente revertam em prol da comunidade local, na forma de melhoria da infra-estrutura, da qualidade ambiental, da saúde e educação.

- ICMS Ecológico: Expressa a existência, no município, de unidades de conservação e a qualidade física dessas áreas, considerando planos de manejo, infra-estrutura, entorno protegido, estrutura de proteção e fiscalização, conforme deliberação normativa do Conselho Estadual de Política Ambiental (COPAM).

Componente Produtivo

- Exportação: As exportações representam as transações comerciais de bens e serviços realizadas com outros países, quanto maior o índice de exportação maior o potencial de aproveitamento de seus recursos econômicos.

Cada indicador mencionado anteriormente pode ser classificado como: muito precário, precário, pouco favorável, favorável ou muito favorável, assim como foi descrito na tabela 1.

Na próxima seção são apresentados os resultados obtidos a partir da execução do algoritmo *k-Means*, implementado na ferramenta Weka.

4.6.1 Indicadores de Exportação, Educação, Taxa de Ocupação e Unidades de Defesa Social

Os indicadores envolvidos na primeira execução do k-Means são Exportação do componente produtivo, Educação e Taxa de Ocupação ambos do componente

humano e o indicador de Unidades de Defesa Social do componente institucional. Os indicadores escolhidos em questão foram originalmente relacionados por uma das regras de associação geradas no trabalho (SIDNEY, 2010).

Segundo Sidney (2010) a regra apresenta a seguinte interpretação "O município que possui condições muito precárias em relação a exportação e muito precárias em relação a saneamento, em 60% dos casos, são considerados muito precários em relação à unidade de defesa social. Esta regra possui um alto grau de confiança, 100% dos casos onde ocorrem condições muito precárias em relação a exportação e muito precárias para saneamento ocorre também defesa social muito precária". Através da análise do especialista, esta regra foi melhorada. Assim foram adicionados mais indicadores na associação, chegando-se assim aos quatro indicadores citados acima.

As características de cada *cluster* gerado são descritas a seguir:

- *Cluster* 0 : foram agrupados 139 municípios no total, representando 16,2954% dos municípios do estado de Minas Gerais, sendo que as condições de educação, taxa de ocupação e exportação são respectivamente precárias, favoráveis e precárias. Já as condições de unidades de defesa social são precárias.

- *Cluster* 1 : foram agrupados 144 municípios no total, representando 16,8815% dos municípios do estado de Minas Gerais, sendo que as condições de educação, taxa de ocupação e exportação são respectivamente favoráveis, muito favoráveis e muito precárias. Já as condições de unidades de defesa social são ambas muito precárias.

- *Cluster* 2 : foram agrupados 207 municípios no total, representando 24,2672% dos municípios do estado de Minas Gerais, sendo que as condições de educa-

ção são precárias e as condições de taxa de ocupação, exportação e unidades de defesa social todas muito precárias.

- *Cluster* 3 : foram agrupados 223 municípios no total, representando 26,1430% dos municípios do estado de Minas Gerais, sendo que as condições de educação e taxa de ocupação são respectivamente favoráveis e pouco favoráveis e as condições de exportação e unidades de defesa social são respectivamente muito favoráveis e precárias.

- *Cluster* 4 : foram agrupados 140 municípios no total, representando 16,4126% dos municípios do estado de Minas Gerais, sendo que as condições de educação e taxa de ocupação são respectivamente favoráveis e precárias e as condições de exportação e unidades de defesa social são ambas muito precárias.

As informações acima podem ser visualizadas resumidamente na tabela 6.

INDICADOR	CLUSTER	DESCRIÇÂO	ESCALA DO INDICADOR	MUNICÍPIOS(%)
Educação, Taxa de Ocupação, Exportação e Defesa Social	0	Emprego favorável independente da exportação.	2,4,2,2	16,2954
	1	Emprego e educação favorável.	4,5,1,1	16,8815
	2	Todos os indicadores precários.	2,1,1,1	24,2672
	3	Educação favorável, seguido de Ocupação e Exportação favoráceis.	4,3,5,2	26,1430
	4	Educação favorável com os outros indicadores precários.	4,2,1,1	16,4126

Tabela 6: Descrição dos *clusters* gerados pela primeira execução do k-Means

A Figura 12 ilustra o mapa do estado de Minas Gerais com a respectiva distribuição de acordo com os agrupamentos gerados.

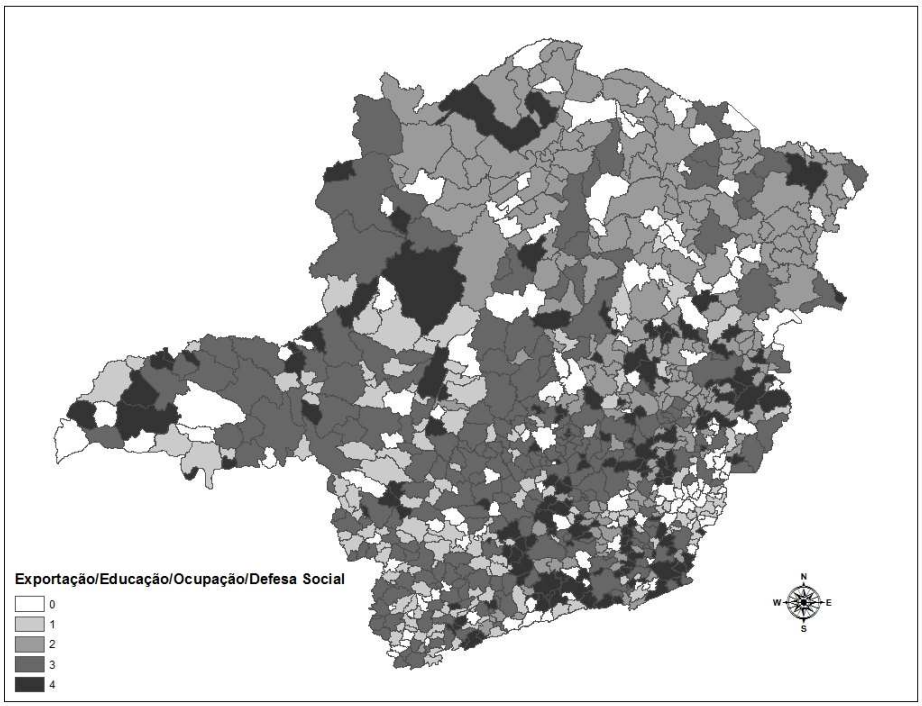

Figura 12: *Clusters* envolvendo os indicadores de exportação, educação, taxa de ocupação e defesa social

De acordo com a figura 12, pode-se analisar que nos *cluster* 2 e 3 o indicador de educação apresenta uma certa influência nos indicadores de taxa de ocupação e exportação, já que ele geralmente está ligado fortemente ao desenvolvimento econômico de um município. Logo, nesses *clusters*, as classificações dos três indicadores citados são similares, ou são todas favoráveis ou são todas precárias. Nos *clusters* 0 e 1, foram agrupados os municípios que não dependem da exportação para gerar empregos, pois geralmente municípios com condições de exportação favoráveis tem melhores condições de taxa de ocupação. Somando-se os dois *clusters*, temos 33,17% dos municípios de Minas Gerais que apresentam essas características. No *cluster* 4, apesar do indicador educação ser um indicador importante

das condições econômicas e sociais e estar ligado a taxa de crescimento de um município, seu valor não influenciou os outros indicadores, ocupação, exportação e unidades de defesa social.

Em relação as 9 regionais do COPAM, a distribuição dos *clusters* pode ser visualizada na figura abaixo.

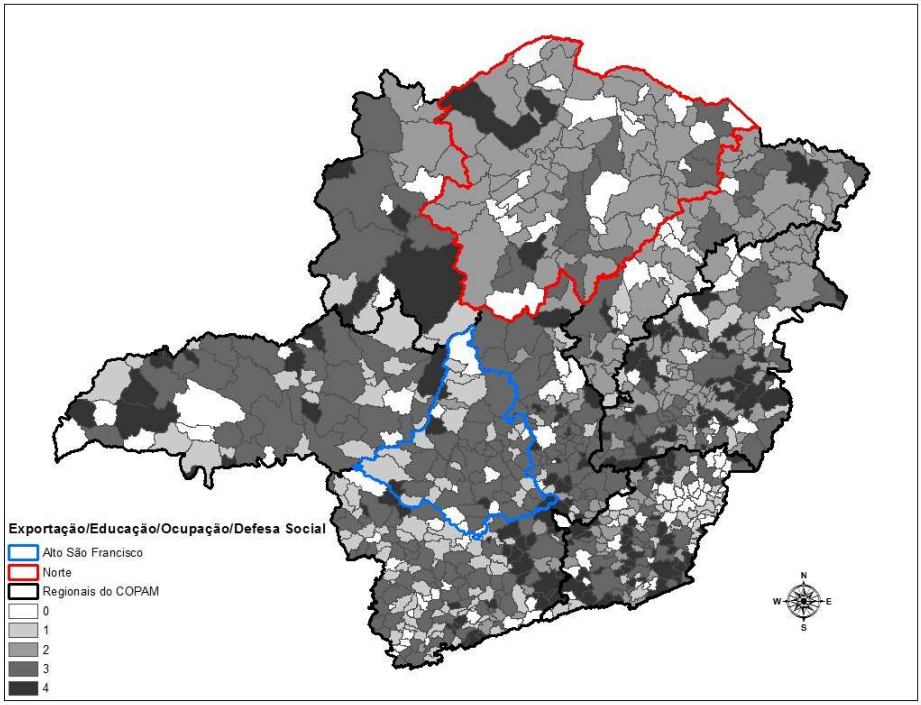

Figura 13: Distribuição dos *clusters* (Exportação, Educação, Taxa de Ocupação e Defesa Social) por regionais do COPAM

Analisando-se cada regional, observamos que em quatro delas (Alto São Francisco, Central, Sul e Triângulo), os municípios pertencentes ao *cluster* 3 são predominantes, ou seja, são municípios onde a educação apresenta uma influência maior, no caso positiva, na taxa de ocupação e exportação desses municípios. No

caso da regional do Alto São Francisco, destacada de azul na imagem, temos o maior percentual de municípios pertencentes ao *cluster 3*. De forma semelhante, temos também que em quatro regionais (Jequitinhonha, Leste Mineiro, Noroeste e Norte), os municípios pertencentes ao *cluster* 2 são predominantes, mas neste caso temos uma influência negativa. No caso da regional Norte, destacada de vermelho na imagem, temos o maior percentual de municípios pertencentes ao *cluster 2*. Na Zona da Mata, os municípios pertencentes ao *clusters* 4 são predominantes, indicando que mesmo que a educação seja forte nesses municípios e ela esteja relacionada com as condições econômicas e sociais e estar ligado a taxa de crescimento de um município, seu valor não influenciou os outros indicadores, ocupação, exportação e unidades de defesa social. A tabela 7 apresenta informações mais detalhadas sobre o que foi apresentado anteriormente.

REGIONAL	**QUANT. DE MUNICÍPIOS**	**CLUSTER PREDOMINANTE**	**QUANT. DE MUNICÍPIOS(CLUSTER)**	**%**
Alto São Francisco	60	3	31	51,66
Central	84	3	37	44,04
Jequitinhonha	56	2	28	50,00
Leste Mineiro	135	2	70	51,85
Noroeste	22	2	6	27,27
Norte	90	2	56	62,22
Sul	177	3	60	33,89
Triângulo	67	3	24	35,82
Zona da Mata	162	4	41	25,30
TOTAL	853			

Tabela 7: Análise dos *clusters* (Exportação, Educação, Taxa de Ocupação e Defesa Social) de acordo com as regionais do COPAM

4.6.2 Indicadores CFEM e ICMS Ecológico

Os indicadores envolvidos na segunda execução do k-Means são CFEM e o Índice ICMS Ecológico ambos do componente natural. Os dois indicadores escolhidos em questão foram originados da seguinte regra de associação:

"O município que possui condições muito precárias a exploração e extração dos recursos minerais, em 40% dos casos, são considerados muito precários ao índice de ICMS Ecológico. Esta regra apresenta uma confiança de 72%" (SIDNEY, 2010).

As características de cada *cluster* gerado são descritas a seguir:

- *Cluster* 0: foram agrupados 157 municípios no total, representando 18,4056% dos municípios do estado de Minas Gerais, sendo que as condições de CFEM e ICMS ecológico são respectivamente, muito favoráveis e favoráveis.
- *Cluster* 1: foram agrupados 141 municípios no total, representando 16,5298% dos municípios do estado de Minas Gerais, sendo que as condições de CFEM e ICMS ecológico são respectivamente, muito favoráveis e muito precárias.
- *Cluster* 2: foram agrupados 359 municípios no total, representando 42,0867% dos municípios do estado de Minas Gerais, sendo que as condições de CFEM e ICMS ecológico são muito precárias.
- *Cluster* 3: foram agrupados 55 municípios no total, representando 6,4478% dos municípios do estado de Minas Gerais, sendo que as condições de CFEM e ICMS ecológico são respectivamente, favoráveis e muito precárias.
- *Cluster* 4: foram agrupados 141 municípios no total, representando 16,5298% dos municípios do estado de Minas Gerais, sendo que as condições do CFEM e ICMS ecológico são respectivamente, muito precárias e muito favoráveis.

As informações acima podem ser visualizadas resumidamente na tabela 8.

A Figura 14 ilustra o mapa do estado de Minas Gerais com a respectiva distribuição de acordo com os agrupamentos gerados.

INDICADOR	CLUSTER	DESCRIÇÃO	ESCALA DO INDICADOR	MUNICÍPIOS(%)
CFEM e ICMS Ecológico	0	Exploradores e preservadores do meio ambiente.	5,4	18,4056
	1	Exploradores e não preservadores do meio ambiente.	5,1	16,5298
	2	Não são nem exploradores e nem preservadores do meio ambiente.	1,1	42,0867
	3	Exploradores e não preservadores do meio ambiente.	4,1	6,4478
	4	Preservadores do meio ambiente.	1,5	16,5298

Tabela 8: Descrição dos *clusters* gerados pela segunda execução do k-Means

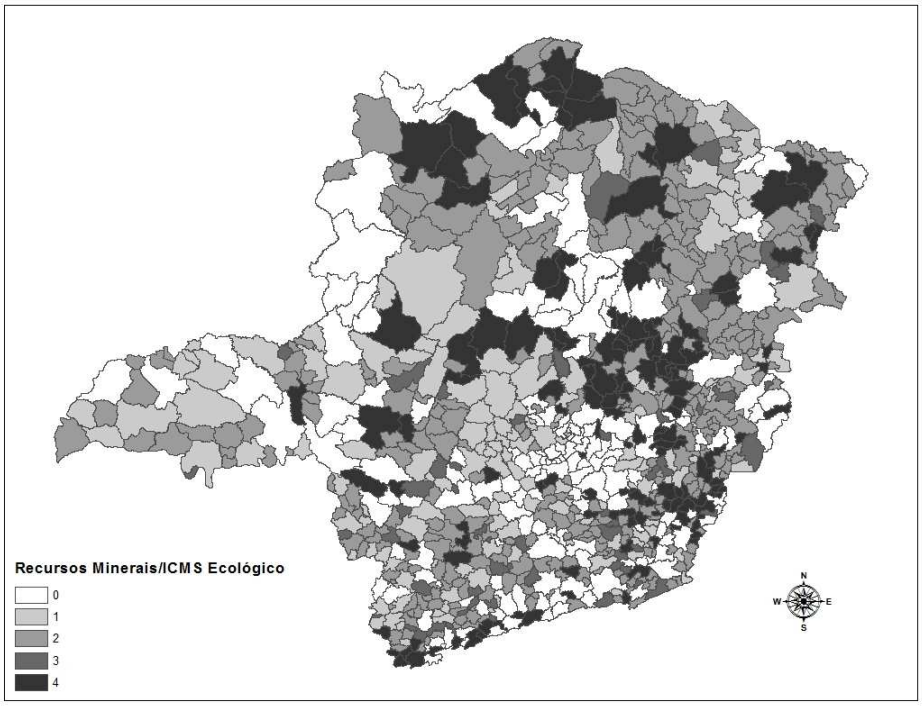

Figura 14: *Clusters* envolvendo os indicadores de CFEM e ICMS Ecológico

Analisando-se a figura 14, pode-se dizer que o *cluster 0* representa os municípios que se destacam, aqueles que tem a consciência de preservar e conservar seu meio ambiente, mesmo que a exploração de seus recursos minerais seja alta, o que gera grandes impactos econômicos e principalmente ambientais. Esses municípios

podem ser vistos como exemplo para os demais, pois eles conseguem manter um bom equilíbrio entre exploração de seu meio ambiente e preservação do mesmo. O *cluster* 1 representa os municípios que são grandes exploradores de recursos minerais, mas que apresentam pouca conscientização de preservação do seu meio ambiente, mesmo que de acordo com a definição do DNPM (Departamento Nacional de Produção Mineral), os recursos da CFEM devem ser aplicados dentre outros aspectos, na melhoria da qualidade ambiental daquele município. Este cenário, pode ser inicialmente interessante para as autoridades competentes no que diz respeito a fiscalização da aplicação dos recursos da CFEM nesses municípios. O *cluster* 2 representa os municípios que não são grandes exploradores de recursos minerais e também não apresentam uma política satisfatória de preservação e conservação do seu meio ambiente. Similar ao *cluster* 1, no *cluster* 3, os municípios exploram, mas não protegem o seu meio ambiente. Somando-se os *clusters* 1 e 3, temos aproximadamente 22,97% dos municípios de Minas Gerais.

Em relação as 9 regionais do COPAM, a distribuição dos *clusters* pode ser visualizada na figura 15.

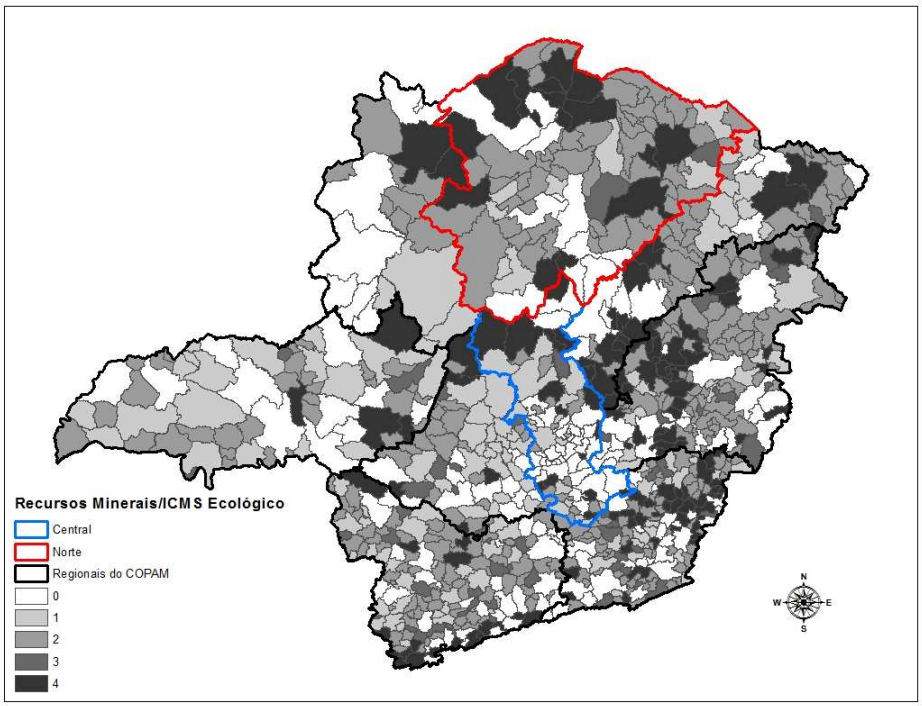

Figura 15: Distribuição dos *clusters* (Recursos Minerais e ICMS Ecológico) por regionais do COPAM

De acordo com a figura 15, pode-se verifica que na maior parte das regionais (Jequitinhonha, Leste Mineiro, Noroeste, Norte, Sul, Triângulo e Zona da Mata) os municípios pertencentes ao *cluster* 2 são predominantes, ou seja, a maioria dos municípios dessas regionais não são exploradores expressivos de recursos minerais e também não preservam seu meio ambiente. Neste caso, a regional Norte, destacada de vermelho, mais de 60% de seus municípios pertencem ao *cluster 2*. Na regional do Alto São Francisco, 38,33 % dos seus municípios pertencem ao *cluster* 1, ou seja, mais de um terço dos seus municípios exploram seus recursos minerais e ao mesmo tempo não aplicam os recursos originados dessa exploração em preservação de seu meio ambiente. A regional Central apresenta um cenário

mais favorável, destaca de azul na imagem, pois 45,23 % de seus municípios, são exploradores de recursos minerais e ao mesmo tempo estão dispostos a preservar seu meio ambiente. Na tabela 9 foi apresentado este cenário mais detalhadamente.

REGIONAL	**QUANT. DE MUNICÍPIOS**	**CLUSTER PREDOMINANTE**	**QUANT. DE MUNICÍPIOS(CLUSTER)**	**%**
Alto São Francisco	60	1	23	38,33
Central	84	0	38	45,23
Jequitinhonha	56	2	27	48,21
Leste Mineiro	135	2	66	48,88
Noroeste	22	2	9	40,90
Norte	90	2	55	61,11
Sul	177	2	74	41,80
Triângulo	67	2	31	46,26
Zona da Mata	162	2	66	40,74
TOTAL	853			

Tabela 9: Análise dos *clusters* (Recursos Minerais e ICMS Ecológico) de acordo com as regionais do COPAM

4.6.3 Indicadores de Educação e CFEM

Os indicadores envolvidos na terceira execução do k-Means são Educação do componente humano e CFEM do componente natural. Os dois indicadores escolhidos em questão foram originados a partir do conhecimento prévio do especialista.

As características de cada *cluster* gerado são descritas a seguir:

- *Cluster* 0: foram agrupados 78 municípios no total, representando 9,1441% dos municípios do estado de Minas Gerais, sendo que as condições de educação e da CFEM são respectivamente, precárias e muito favoráveis.
- *Cluster* 1: foram agrupados 144 municípios no total, representando 16,8815% dos municípios do estado de Minas Gerais, sendo que as condições de educação e da CFEM são respectivamente, favoráveis e muito favoráveis.

- *Cluster* 2: foram agrupados 263 municípios no total, representando 30,8323% dos municípios do estado de Minas Gerais, sendo que as condições de educação e da CFEM são ambas muito precárias.

- *Cluster* 3: foram agrupados 131 municípios no total, representando 15,3575% dos municípios do estado de Minas Gerais, sendo que as condições de educação e da CFEM são ambas muito favoráveis.

- *Cluster* 4: foram agrupados 237 municípios no total, representando 27,7842% dos municípios do estado de Minas Gerais, sendo que as condições de educação e da CFEM são respectivamente, favoráveis e muito precárias.

As informações acima podem ser visualizadas resumidamente na tabela 10.

INDICADOR	**CLUSTER**	**DESCRIÇÃO**	**ESCALA DO INDICADOR**	**MUNICÍPIOS (%)**
Educação e CFEM	0	Recebem CFEM mais tem uma educação precária.	2,5	9,1441
	1	Recursos do CFEM são aplicados na educação.	4,5	16,8815
	2	Recebem pouco recursos do CFEM e apresentam uma educação precária.	1,1	30,8323
	3	Recursos do CFEM são aplicados na educação.	5,5	15,3575
	4	Educação favorável independente do CFEM.	4,1	27,7842

Tabela 10: Descrição dos *clusters* gerados pela terceira execução do k-Means

A Figura 16 ilustra o mapa do estado de Minas Gerais com a respectiva distribuição de acordo com os agrupamentos gerados.

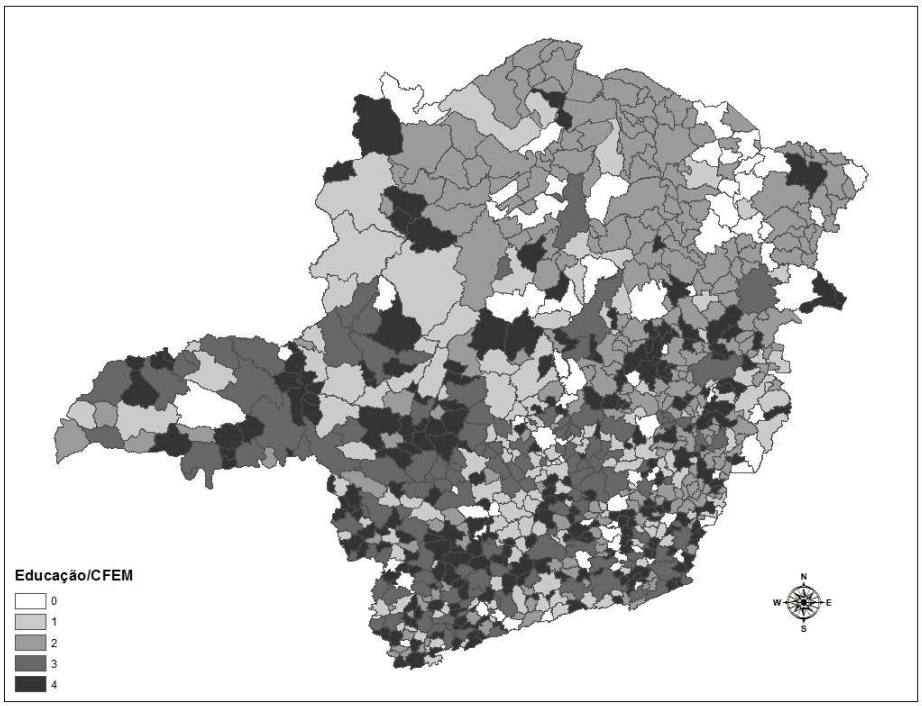

Figura 16: *Clusters* envolvendo os indicadores da CFEM e Educação

Analisando-se a figura 16, pode-se observar que nos *clusters* 1 e 3 as condições da educação e CFEM são favoráveis. Isso nos mostra que, de forma geral, os municípios pertencentes a esses *clusters*, representando aproximadamente 32% dos municípios de MG, aplicam os recursos da CFEM destinados a educação corretamente, já que recebem um valor expressivo da divisão da arrecadação da CFEM, sendo que do total arrecadado, 65% são repassados para o município produtor. O *cluster* 0 é formado pelos municípios que apesar de terem uma alta arrecadação da CFEM não apresentam uma educação satisfatória. Levando em consideração esse contexto, pode ser interessante para as autoridades competentes realizar uma investigação mais detalhada sobre como é feita a aplicação dos

recursos originados da CFEM nesses municípios, afim de identificar uma possível irregularidade.

Em relação às 9 regionais do COPAM, a distribuição dos *clusters* pode ser visualizada na figura abaixo.

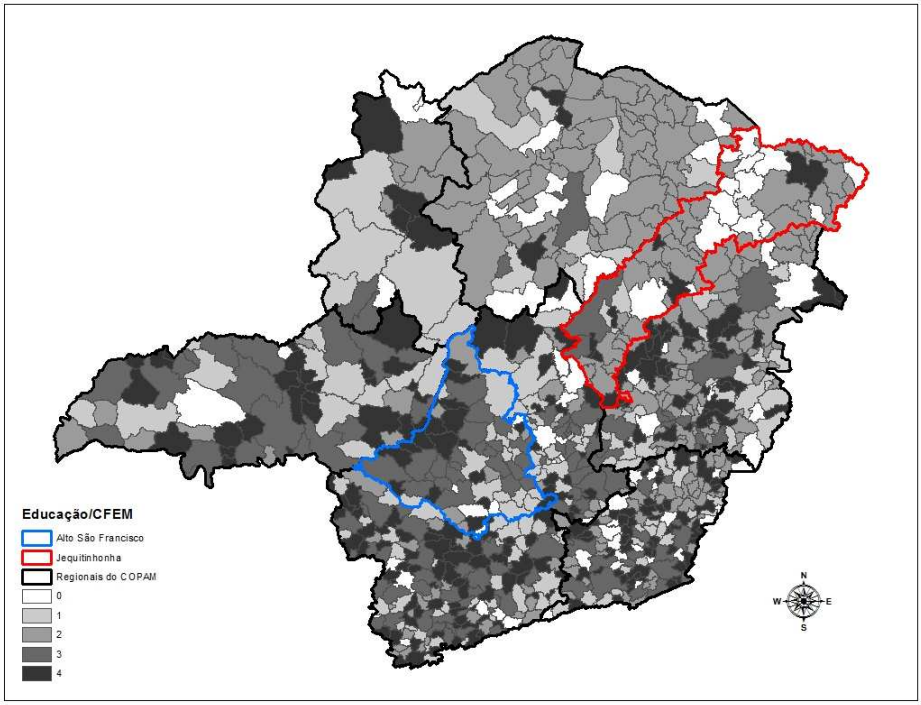

Figura 17: Distribuição dos *clusters* (CFEM e Educação) por regionais do COPAM

Analisando-se a figura 17, observa-se que em quatro regionais (Jequitinhonha, Leste Mineiro, Noroeste e Norte) existe a predominância de municípios pertencentes ao *cluster 2*. Assim nessas regionais, a maior parte de seus municípios apresentam uma baixa arrecadação oriunda do CFEM, ou seja, não são municípios exploradores de recursos minerais, além de possuir condições de educação muito precárias. Neste caso, na regional Jequitinhonha, destacada de vermelho na

imagem, mais de 60% de seus municípios pertencem ao *cluster 2*. Em relação ao *cluster 3*, apenas na regional Alto São Francisco, destacada de azul na imagem, os municípios pertencentes a esse *cluster* são dominantes, ou seja, 38,33 % do total de municípios dessa regional, aparentemente utilizam os recursos originados da arrecadação da CFEM na educação. De forma semelhante, 29,76 % dos municípios pertencentes a regional Central aplicam corretamente seus recursos capitados da arrecadação da CFEM, pois as condições de educação desses municípios são muitos favoráveis. A tabela 11 apresenta de uma maneira mais detalhada e compacta as informações discutidas acima.

REGIONAL	**QUANT. DE MUNICÍPIOS**	**CLUSTER PREDOMINANTE**	**QUANT. DE MUNICÍPIOS(CLUSTER)**	**%**
Alto São Francisco	60	3	23	38,33
Central	84	1	25	29,76
Jequitinhonha	56	2	34	60,71
Leste Mineiro	135	2	65	48,14
Noroeste	22	2	7	31,81
Norte	90	2	66	73,33
Sul	177	4	67	37,85
Triângulo	67	4	27	40,29
Zona da Mata	162	4	56	34,56
TOTAL	853			

Tabela 11: Análise dos *clusters* (Educação e CFEM) de acordo com as regionais do COPAM

5 CONCLUSÂO E TRABALHOS FUTUROS

Esse trabalho apresentou como aplicar a técnica de *clustering* na base de dados do ZEE-MG, levando em consideração as informações referentes a potencialidade social. Para aplicar a técnica, foi necessário passar por todas as etapas do processo de KDD, sendo que as etapas de pré-processamento e transformação dos dados foram importantes para obter o valor de cada indicador para cada município. Com o conhecimento específico do especialista, foram escolhidos indicadores que possuem uma relação em comum, de diferentes componentes. Com os *clusters* obtidos da execução do algoritmo k-means, através da ferramenta Weka, foi possível visualizar a relação de alguns indicadores, como por exemplo, o indicador CFEM e sua relação com os indicadores Educação e ICMS Ecológico. Com a finalização do trabalho foi possível verificar que a base de dados do ZEE-MG tem potencial para a aplicação de mineração de dados. O objetivo de estudar e identificar técnicas de mineração de dados aplicáveis na base de dados do ZEE-MG foi alcançado através da técnica de *clustering*.

Para trabalhos futuros, a técnica de *clustering* poderia ser aplicada periodicamente, de acordo com as atualizações das informações armazenadas na base de dados do ZEE-MG, afim de identificar a evolução de cada município em relação aos *clusters* gerados. Técnicas de mineração de dados poderão ser aplicadas na base de dados do ZEE-MG para extrair padrões úteis referentes a parte de vulnerabilidade natural, que é composta pelos fatores de integridade da flora e fauna, vulnerabilidade dos solos, vulnerabilidade à erosão, índice de unidade e vulnerabilidade dos recursos hídricos.

Referências

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules in large databases. In: *Proceedings of the 20th International Conference on Very Large Data Bases*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994. (VLDB '94), p. 487–499. ISBN 1-55860-153-8. Disponível em: <http://dl.acm.org/citation.cfm?id=645920.672836>.

AMâNCIO, R.; REZENDE, J. B.; BOTELHO, D. de O.; ARRUDA, M. A. *Capítulo 3 - Componente Natural*. [S.l.]: Editora UFLA, 2008.

BAEZA-YATES; RICARDO; RIBEIRO-NETO; BERTHIER. *Modern Information Retrieval: The concepts and technology behind search*. 2. ed. Harlow, England: Pearson Education, 2011. ISBN 978-0-321-41691-9.

BRAMER, M. *Principles of Data Mining*. London, UK: Springer, 2007. (Undergraduate Topics in Computer Science). ISBN 978-1-84628-765-7.

CALEGARIO, C. L. L.; LEITE, E. T.; PEREIRA, N. C.; ARRUDA, M. A. *Capítulo 2 - Componente Produtivo*. [S.l.]: Editora UFLA, 2008.

CIOS, K. J.; PEDRYCZ, W.; SWINIARSKI, R. W.; KURGAN, L. A. *Data Mining A Knowledge Discovery Approach*. 1. ed. [S.l.]: Springer, 2007. ISBN 978-3-540-76916-3.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. Advances in knowledge discovery and data mining. In: FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. (Ed.). Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996. cap. From Data Mining to

Knowledge Discovery: An Overview, p. 1–34. ISBN 0-262-56097-6. Disponível em: <http://dl.acm.org/citation.cfm?id=257938.257942>.

GROUP, T. P. G. D. *About*. Janeiro 2014. Disponível em: <http://www-.postgresql.org/about/>.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. 3rd. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. ISBN 0123814790, 9780123814791.

JUNG, C. F. *Metodologia Aplicada a Projetos de Pesquisa: Sistemas de Informação & Ciência da Computação*. [S.l.]: Taquara, 2009.

LAROSE, D. T. *Discovering Knowledge in Data: An Introduction to Data Mining*. [S.l.]: Wiley-Interscience, 2004. ISBN 0471666572.

MAIMOM, O.; ROKACH, L. *Data Minig and Knowledge Discovery Handbook*. Second. [S.l.]: Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA, 2010.

OLIVEIRA, L. C. F. de S.; LEITE, E. T.; RIBEIRO, L. M. de P.; RESENDE, J. B.; ARRUDA, M. A. *Capítulo 4 - Componente Humano*. [S.l.]: Editora UFLA, 2008.

OLSON, D. L.; DELEN, D. *Advanced Data Mining Techniques*. [S.l.]: Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA, 2008.

SALAZAR, G. T.; OLIVEIRA, E. R. de; SILVA, S. S. da; ARRUDA, M. A.; ROCHA, P. A. M. da; RODRIGUES, L. A. *Capítulo 5 - Componente Institucional*. [S.l.]: Editora UFLA, 2008.

SCOLFORO, J. R.; OLIVEIRA, A. D. de; CARVALHO, L. M. T. de; MARQUES, J. J. G.; LOUZADA, J. N.; MELLO, C. R. de; PEREIRA, J. R.; REZENDE, J. B.; VALE, L. C. C. *Capítulo 1 - Zoneamento Ecológico-Econômico do Estado de Minas Gerais*. [S.l.]: Editora UFLA, 2008.

SEMAD. *COPAM*. Janeiro 2014. Disponível em: <http://www.meioambiente-.mg.gov.br/copam>.

SIDNEY, C. F. *Aplicação de Mineração de Dados no Banco de Dados do Zoneamento Ecológico Econômico de Minas Gerais*. 2010. Disponível em: <http://www.bsi.ufla.br/wp-content/uploads/2013/07/Monografia_Christiane-.pdf>.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introduction to Data Mining - Mineração de Dados*. Rio de Janeiro, RJ, Brasil: Editora Ciência Moderna Ltda, 2009. ISBN 9788573937619.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introdução ao Data Mining*. [S.l.]: Publish, 2009.

VALÊNCIO, C.; MEDEIROS, C. de; ICHIBA, F.; SOUZA, R. de. Spatial clustering applied to health area. In: *Parallel and Distributed Computing, Applications and Technologies (PDCAT), 2011 12th International Conference on*. [S.l.: s.n.], 2011. p. 427–432.

WEBGIS. *Capítulo 6. Referências PostGIS*. Janeiro 2014. Disponível em: <http://www.webgis.com.br/postgis/>.

WITTHEN, I. H.; FRANK, E.; HALL, M. A. *Data Mining Pratical Machine Learning Tools and Techniques*. Third. [S.l.]: Elsevier, 2011.

ZHANG, X.; ZOU, J.; LE, D.; THOMA, G. A structural SVM approach for reference parsing. *BMC Bioinformatics*, v. 12, n. Suppl 3, p. S7, 2011. ISSN 1471-2105. Disponível em: <http://www.biomedcentral.com/1471-2105/12%-/S3/S7>.

6 APÊNDICE

Nesta seção, podemos visualizar uma opção alternativa, com uma variação de cores diferente, das distribuições dos *clusters* de cada execução do algoritmo K-means.

6.1 Indicadores de Exportação, Educação, Taxa de Ocupação e Unidades de Defesa Social

A Figura 18 ilustra o mapa do estado de Minas Gerais com a respectiva distribuição de acordo com os agrupamentos gerados.

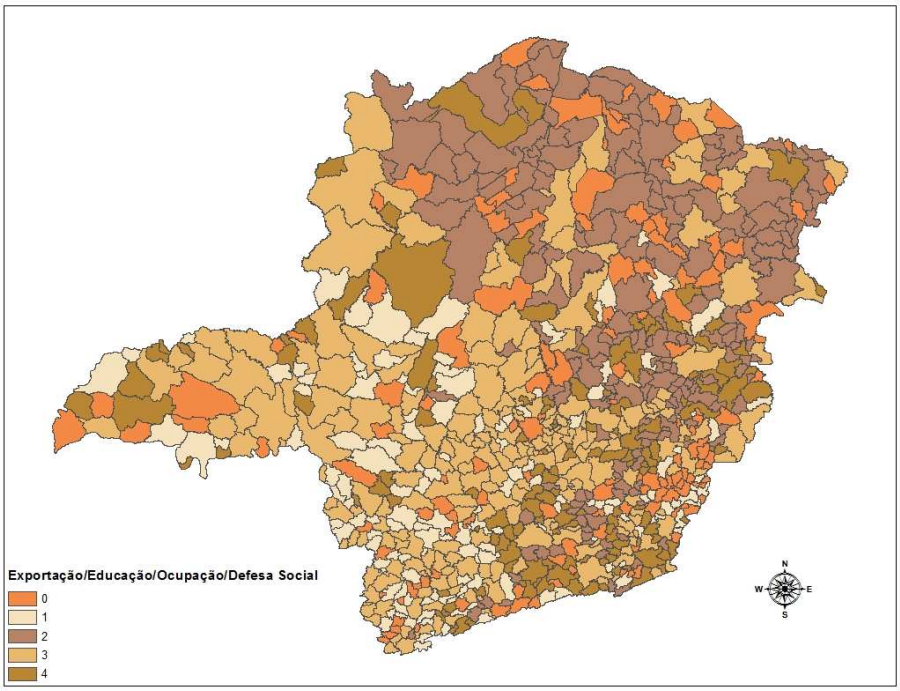

Figura 18: Distribuição dos *clusters* (Exportação, Educação, Taxa de Ocupação e Defesa Social) por regionais do COPAM - Versão Alternativa

Em relação as 9 regionais do COPAM, a distribuição dos *clusters* pode ser visualizada na figura 19.

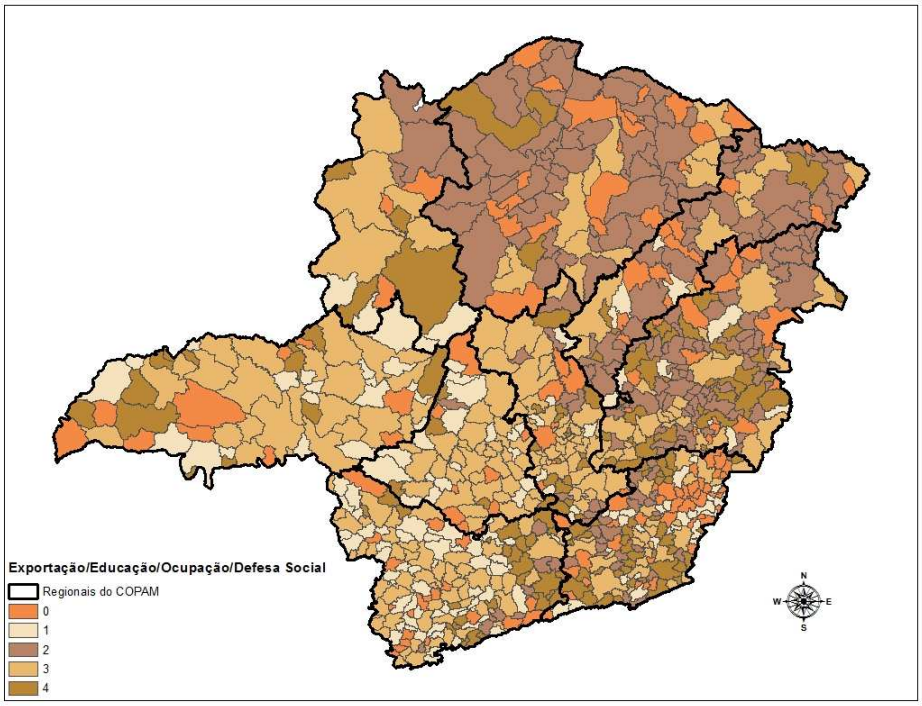

Figura 19: Distribuição dos *clusters* (Exportação, Educação, Taxa de Ocupação e Defesa Social) por regionais do COPAM - Versão Alternativa

6.2 Indicadores CFEM e ICMS Ecológico

A Figura 20 ilustra o mapa do estado de Minas Gerais com a respectiva distribuição de acordo com os agrupamentos gerados.

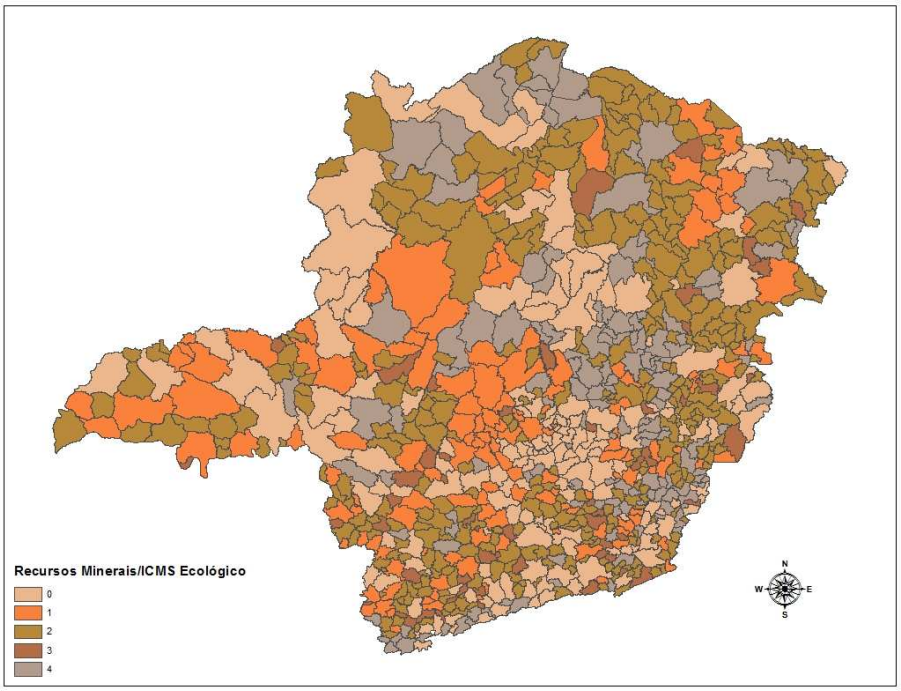

Figura 20: Distribuição dos *clusters* (Recursos Minerais e ICMS Ecológico) por regionais do COPAM - Versão Altenativa

Em relação as 9 regionais do COPAM, a distribuição dos *clusters* pode ser visualizada na figura 21.

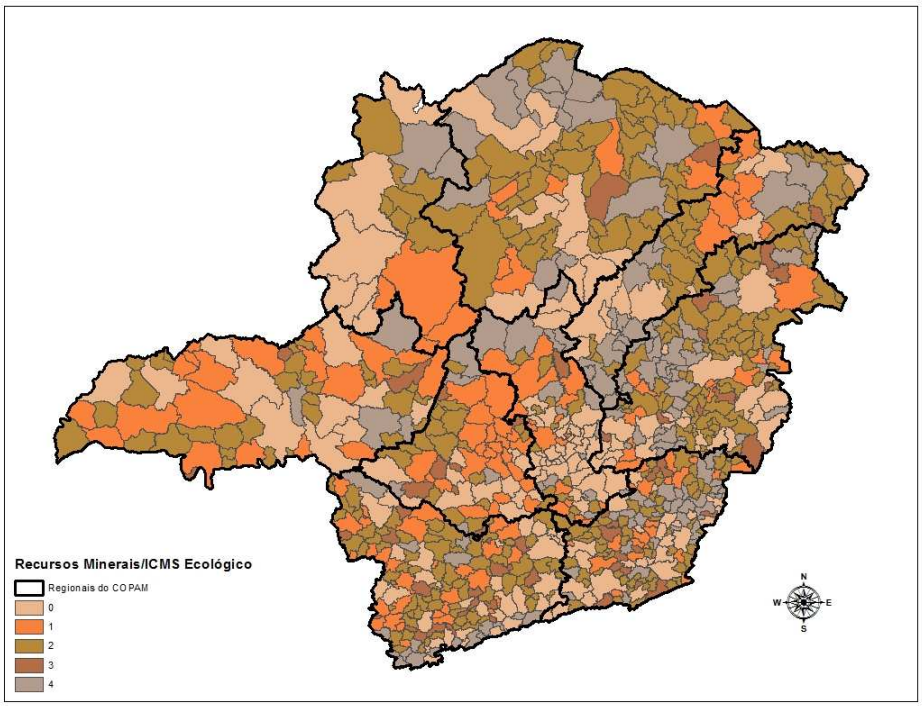

Figura 21: Distribuição dos *clusters* (Recursos Minerais e ICMS Ecológico) por regionais do COPAM - Versão Alternativa

6.3 Indicadores de Educação e CFEM

A Figura 22 ilustra o mapa do estado de Minas Gerais com a respectiva distribuição de acordo com os agrupamentos gerados.

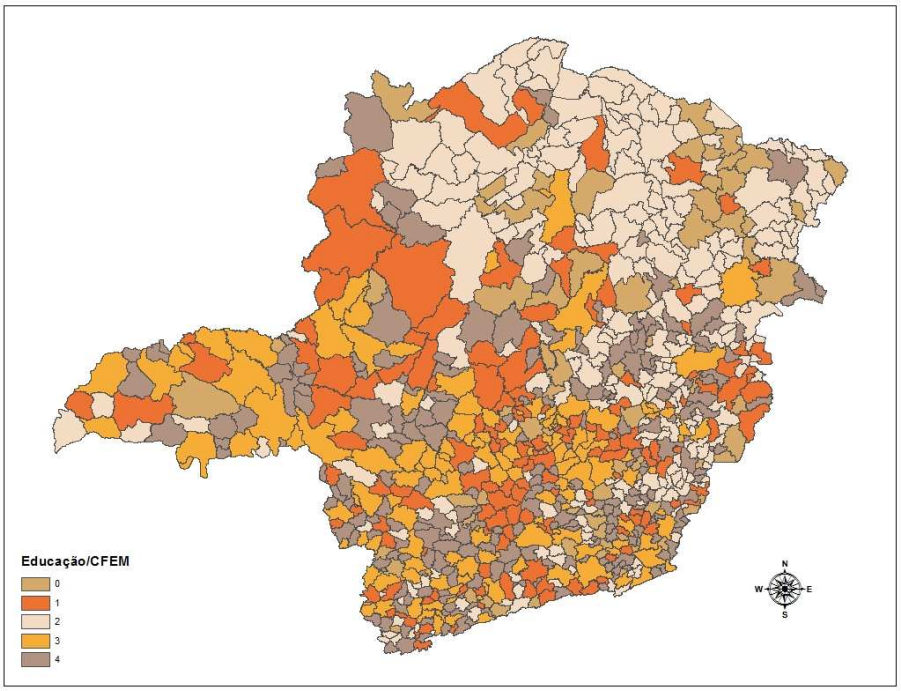

Figura 22: Distribuição dos *clusters* (CFEM e Educação) - Versão Alternativa

Em relação as 9 regionais do COPAM, a distribuição dos *clusters* pode ser visualizada na figura 23.

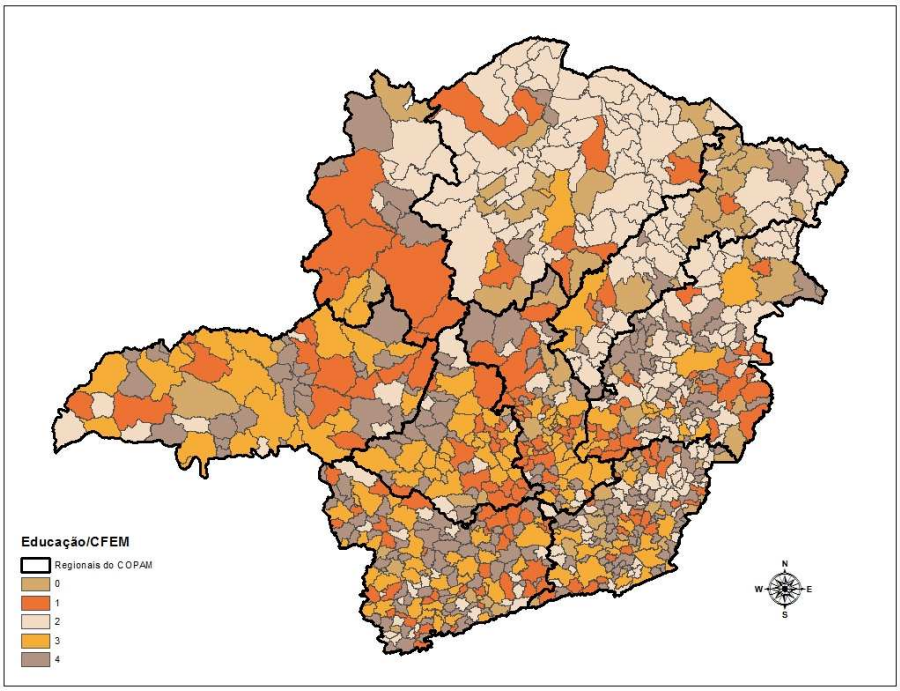

Figura 23: Distribuição dos *clusters* (CFEM e Educação) por regionais do COPAM - Versão Alternativa