UNIVERSIDADE FEDERAL DE SANTA CATARINA

ANÁLISE DOS REGISTROS DE DESASTRES NATURAIS ATRAVÉS DA UTILIZAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS

Mateus Patrício Mello

Florianópolis – SC

2013/1

UNIVERSIDADE FEDERAL DE SANTA CATARINA – UFSC
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA – INE
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

ANÁLISE DOS REGISTROS DE DESASTRES NATURAIS ATRAVÉS DA UTILIZAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS

Mateus Patrício Mello

Trabalho de Conclusão de Curso apresentado como parte dos requisitos para obtenção do grau de Bacharel em Sistemas de Informação

Florianópolis – SC

2013/1

Mateus Patrício Mello

ANÁLISE DOS REGISTROS DE DESASTRES NATURAIS ATRAVÉS DA UTILIZAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS

Trabalho de Conclusão de Curso apresentado como parte dos requisitos para obtenção do grau de Bacharel em Sistemas de Informação.

Orientadora:

_____

Profa. Vania Bogorny

Banca Examinadora:

_____

Prof. Luis Otavio Alvares

_____

Jairo Ernesto Bastos Krüger

AGRADECIMENTOS

Agradeço a professora Vania Bogorny pela orientação na elaboração deste trabalho que muito colaborou para a conclusão do mesmo. Também agradeço ao prof. Luis Otavio Alvares pelas sugestões e conversas, ao Jairo Ernesto Bastos Krüger pela disponibilização dos dados e por ter aceitado participar da banca deste trabalho.

Agradeço também a minha namorada Sofia Beliza Carneiro de Cabral pela compreensão nos momentos em que estive mentalmente e fisicamente distante para a elaboração deste trabalho e também a minha família por sempre ter me apoiado no decorrer da minha faculdade.

Aos meus amigos de faculdade, Matheus Vill, Greici Baretta, Gesiel da Silva, Eduarda Zanette e Guilheme Moser, que desde o começo me acompanharam e ajudaram incontáveis vezes nesses quatro anos e meio de graduação em diversos trabalhos e provas.

RESUMO

Devido ao aumento da ocorrência de desastres naturais e do grau de complexidade, a análise dos registros de desastres tornou-se uma necessidade para que se possa conhecer melhor o histórico brasileiro de desastres. Fazer a análise dos dados históricos de ocorrência de desastres naturais e dos níveis de precipitação de chuvas é um dos pontos fundamentais para que seja possível diminuir o número de vítimas e também para que se possa prever com maior antecedência a ocorrência de um desastre natural.

O que este trabalho pretende é analisar os registros de dados de desastres de todo o Brasil. Para isto foram coletados dados referentes aos desastres naturais, tirando por base o preenchimento, por parte dos municípios, do formulário de avaliação de dados (AVADAN), e também dados referente à ocorrência de do fenômeno *El Niño* e *La Niña*. Além destes dados também foram coletados índices pluviométricos, dados de solo e relevo dos municípios onde estes desastres ocorreram.

O objetivo desta coleta é estudar os desastres através do processo de mineração de dados para a descoberta de padrões entre os desastres. As técnicas de mineração de dados utilizadas neste trabalho foram a clusterização (agrupamento) e classificação, para que se possa obter como resultado os padrões encontrados através da execução de algoritmos das duas técnicas citadas.

Palavras chave: desastres naturais, mineração de dados, análise de dados, classificação e clusterização.

SUMÁRIO

LISTAGEM DE TABELAS

LISTAGEM DE FIGURAS

LISTA DE REDUÇÕES

AVADAN – Formulário de Avaliação de Danos.

ANA – Agência Nacional de Águas.

CEPED – Centro Universitário de Estudos e Pesquisas sobre Desastres.

CPRM – Companhia de Pesquisa de Recursos Minerais (Serviço Geológico do Brasil).

CPTEC – Centro de Previsão de Tempo e Estudos Climáticos.

CRISP-DM – *Cross Industry Standard Process for Data Mining*.

IBGE – Instituto Brasileiro de Geografia e Estatística.

DCBD - Descoberta de Conhecimento em Banco de Dados.

KDD – *Knowledge Discovery Database*

NOAA – *National Oceanic and Atmospheric Administration*

NOPRED – Formulário de Notificação Preliminar de Desastres

SEDEC – Secretaria Nacional de Defesa Civil

SQL – *Structured Query Language*

1. INTRODUÇÃO E MOTIVAÇÃO

A ocorrência cada vez mais frequente de desastres naturais tem sido uma realidade muito comum em diversos países. Cerca de 70% dos desastres ocorridos no mundo são em países em desenvolvimento SORIANO [9]. Segundo os dados levantados pelo Centro Universitário de Estudos e Pesquisa sobre Desastres (CEPED), no Brasil é possível encontrar diversos tipos de desastres naturais, desde inundações e deslizamento, até estiagens e queimadas.

O aumento da complexidade e da quantidade de calamidades, matando e ferindo milhares de brasileiros, fez com que se realizassem cada vez mais estudos no intuito de entender mais sobre desastres naturais e também de diminuir o sofrimento das pessoas. Alguns casos ocorridos são o do município de Ilhota - SC, em 2008, que foi completamente atingido, tendo mais de trinta mortos e milhares de desabrigados e desalojados; os deslizamentos, enxurradas e inundações na região serrana do Rio de Janeiro, em 2011, considerado o pior desastre natural na história brasileira e no ano de 2012, a seca na região nordeste do Brasil, considerada a pior dos últimos 30 anos.

Fazer a análise dos dados históricos de ocorrência de desastres naturais e dos níveis de precipitação de chuvas é um dos pontos fundamentais para que seja possível diminuir o número de vítimas e também para que se possa prever com maior antecedência a ocorrência de um desastre natural.

Uma das maneiras de fazer esta análise é através do processo de mineração de dados, do inglês, *data mining*. O processo de mineração de dados tem como objetivo extrair conhecimento novo, útil e interessante implícito nos dados, e representá-lo de forma acessível para o usuário KUMAR[12]. Atualmente existem diversas técnicas de mineração como classificação, agrupamento, clusterização, regras de associação, entre outros. A técnica de classificação pode ser útil para conceber um modelo de dados que tente predizer a iminência de um desastre natural com base na busca de padrões. No estudo realizado por LIMA [2] foi possível utilizar a técnica de agrupamento ou clusterização para criar grupos que possam ajudar a padronizar o processo de logística humanitária entre os municípios. Ainda segundo LIMA[2] um mesmo desastre pode atingir diferentes localidades ao mesmo tempo e isto sempre exige uma coordenação mútua para as ações de socorro.

A criação de um modelo de dados para previsão e a descoberta de grupos podem ser utilizadas pelas autoridades responsáveis, como a Defesa Civil, na prevenção de desastres naturais. Além disso, pode também ser útil para ajudar os setores responsáveis pela resposta a um desastre.

O que este trabalho pretende é analisar os registros de dados de desastres de todo o Brasil. Estudos sobre desastres naturais no Brasil, em geral, ocorrem sobre uma região geográfica específica. SOUZA [4] e CHAGAS [3] fazem análises em dados da região serrana do estado do Rio de Janeiro e da Serra do Mar paulistana, respectivamente. Já ESTÉBANEZ [3] utiliza uma maior área geográfica, explorando dados de todo o Equador, país que possui uma área de 256.370 km² [7], comparável ao tamanho do estado de São Paulo, com 248.209 km² [6]. Como serão analisados dados de todo o Brasil, logo a área estudada corresponde a todo território brasileiro.

Os dados utilizados neste trabalho foram disponibilizados pelo CEPED, CPRM (Serviço Geológico do Brasil) e NOAA (Administração Oceânica e Atmosférica Nacional). Do CEPED foram utilizados os registros de AVADAN (Formulários de Avaliação de Danos), NOPRED (Formulário de Notificação Preliminar de Desastres), Relatórios de Danos (documento anterior ao AVADAN e NOPRED), decretos e portarias do governo. Do CPRM foi utilizado o índice dos municípios com suscetibilidade a deslizamentos e do NOAA a intensidade anual do *El Niño* e da *La Niña*. A seção 3.1 descreve com maiores detalhes todos os dados que foram obtidos.

1.1 Objetivos

O objetivo geral deste trabalho é um estudo sobre os desastres naturais brasileiros através do processo de mineração de dados para a descoberta de padrões entre os desastres. Os objetivos específicos incluem:

- aplicar algoritmos de classificação e agrupamento para descoberta de conhecimento em registros de desastres naturais cuja origem esteja relacionada com a precipitação de chuvas ocorridas no Brasil**;**
- identificar os possíveis estados em que podem ocorrer desastres naturais mediante um cenário hipotético através das características físicas e climáticas;

- identificar quais os tipos de desastres naturais que mais ocorrem em uma determinada microrregião brasileira;
- relacionar os níveis de chuva com a ocorrência dos desastres naturais;

1.2 Delimitação do Escopo

- O estudo utiliza dados relacionados aos desastres naturais ocorridos em todo o Brasil tirando por base o preenchimento, por parte dos municípios, dos AVADAN, NOPRED, relatórios de danos, decretos e portarias .
- Todo o conjunto de dados analisado corresponde ao período dos anos de 1991 até 2010, tanto das ocorrências dos desastres naturais (AVADAN) quanto dos índices pluviométricos.
- Para a obtenção dos índices pluviométricos foram utilizados dados coletados da ANA (Agência Nacional de Águas).
- Foram utilizados dados relacionados com os registros de desastres naturais de acordo com os AVADAN (preenchidos pelos municípios) e que foram disponibilizados pelo CEPED.

1.3 Organização do Trabalho

O restante do trabalho está divido em quatro capítulos.

O Capítulo 2 apresenta informações relacionadas ao contexto teórico e técnico no qual o trabalho irá se basear. São apresentados os conceitos de mineração de dados, esclarecendo o processo de DCBD (Descoberta de Conhecimento em Bancos de Dados) e os conceitos básicos sobre desastres naturais.

O Capítulo 3 apresenta a metodologia utilizada, além de mostrar como foram obtidos e manipulados todos os dados de modo que fique esclarecido qual foi o conjunto final de registros onde foram aplicados os algoritmos de mineração de dados.

No Capítulo 4 são apresentados os resultados obtidos com a aplicação dos algoritmos utilizados, mostrando os modelos de predição encontrados e a aplicação destes dentro do contexto de desastres naturais.

O quinto e último capítulo apresenta a conclusões obtidas confrontando-as com o objetivo geral do trabalho e os objetivos específicos, além de apresentar trabalhos futuros que podem ser realizados a partir deste estudo.

2. CONCEITOS BÁSICOS

A análise dos dados históricos dos desastres naturais ocorridos no Brasil será realizada através da aplicação de técnicas de mineração de dados. Para isto é necessário entender alguns conceitos de registros dos desastres naturais, processo de mineração, técnicas e os algoritmos que implementam estas técnicas.

2.1 Registros dos Desastres Naturais no Brasil

De acordo com o Atlas Brasileiro de Desastres Naturais CEPED[17], antes de 1990 o documento oficial brasileiro que registrava a ocorrência de um desastre natural era o Relatório de Danos. Posterior a esta data as informações oficiais sobre um desastre começaram a ser registradas através da emissão de um AVADAN e também do Formulário de Notificação Preliminar de Desastre (NOPRED). Para efeito legal, o prefeito deveria oficializar a ocorrência do desastre por meio de um Decreto Municipal . Após a ocorrência de um desastre os formulários deveriam ser encaminhados à Coordenadoria Estadual da Defesa Civil e à Secretaria Nacional de Defesa Civil, onde esta última ou o Ministério da Integração Nacional homologavam o Decreto Municipal através de uma Portaria publicada no Diário Oficial da União. A figura 1 mostra as etapas do processo de oficialização de um desastre natural, deste a incidência do desastre até a publicação da Portaria.

Figura 1 – Esquema do registro de desastres.

Para registrar a ocorrência de um desastre, o município preenchia o AVADAN ou NOPRED em papel. A partir de 2010 um sistema informatizado foi criado e estas informações foram digitalizadas e armazenadas em um banco de dados.

Os principais dados que o AVADAN e o NOPRED disponibilizam são: município de ocorrência, data da ocorrência e tipo do desastre. Outros dados que também são possíveis de obter através deste formulário são estimativas de danos humanos (pessoas desabrigadas, desalojadas, mortas, danos, etc.), materiais e ambientais além de dados referentes aos prejuízos econômicos (produção de indústrias, agricultura e pecuária) e sociais (serviços interrompidos ou prejudicados).

2.2 A Mineração de Dados e o Processo de Descoberta de Conhecimento

A mineração de dados é uma tecnologia que combina métodos tradicionais de análise de dados com algoritmos sofisticados para processar grandes volumes de dados com o intuito de descobrir conhecimento que existe dentro destes grandes volumes TAN [1].

Para que seja possível fazer a mineração de dados, primeiramente é necessário conhecer como funciona o processo de descoberta de conhecimento (KDD - *Knowledge Discovery in Databases*). Segundo TAN [1], a mineração de dados é uma parte integral do KDD e este, por sua vez, faz a transformação de dados e informação em conhecimento.

Dados são números, palavras, figuras, datas ou qualquer sinal desprovido de significado e informação são dados dotados de significado. Um texto escrito em mandarim pode ser um exemplo para diferenciar dados de informação. Para pessoas que sabem ler em mandarim o texto tem um significado (informação) já para pessoas que não sabem o texto contem apenas símbolos (dados). O exemplo da figura 2 mostra esta situação onde é possível visualizar o ideograma chinês que representa a palavra paz.

Figura 2 – Ideograma chinês da palavra paz.

Conhecimento é o conjunto completo de informações, dados, relações que levam à tomada de decisão, realização de tarefas e à criação de novas informações e, conforme já descrito anteriormente, a mineração de dados tem como objetivo descobrir este conhecimento.

O processo de descoberta de conhecimento foi apresentado por FAYYAD [18] e consiste em executar uma série de passos, desde a seleção, pré-processamento e transformação dos dados até a mineração destes dados e a interpretação dos resultados obtidos na mineração. A figura 2 apresenta as etapas do processo de descoberta de conhecimento, mostrando o resultado a cada etapa (dados relevantes, pré-processados, transformados, etc). Na seleção são obtidos os dados relevantes para o problema. No pré-processamento é realizada a limpeza de registros que estão incompletos, redundantes ou que geram incertezas. Na transformação são gerados novos dados a partir dos dados pré-processados. Na etapa de *Data Mining* (modelagem) é feita a busca por padrões nos dados gerando conhecimento e, por último, na interpretação, é onde os padrões serão analisados e compreendidos para que o conhecimento gerado possa ajudar na tomada de decisão.

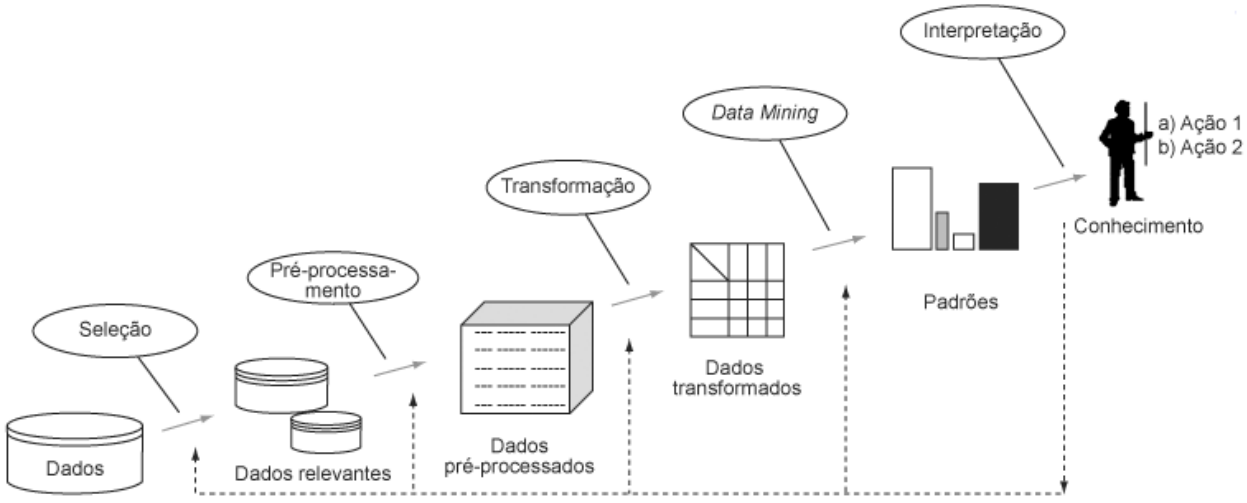


Figura 3 – O processo de KDD

Uma melhor maneira para compreender e elaborar um estudo utilizando mineração de dados é através da metodologia CRISP-DM (Processo Padrão Inter-Indústrias para Mineração de Dados). Esta metodologia, baseada no processo de KDD, foi concebida com o intuito de criar processos que padronizassem o desenvolvimento de projetos de mineração de dados. O guia do CRISP-DM [10] define que o processo de mineração é cíclico e este ciclo está dividido em seis fases: Entendimento do Negócio, Entendimento dos Dados, Preparação dos Dados, Modelagem,

Avaliação e Disponibilização. A figura 3 ilustra a sequência das fases, mostrando através das setas as dependências mais comuns e importantes entre as fases. A seta circular externa simboliza o ciclo natural do processo de mineração de dados.

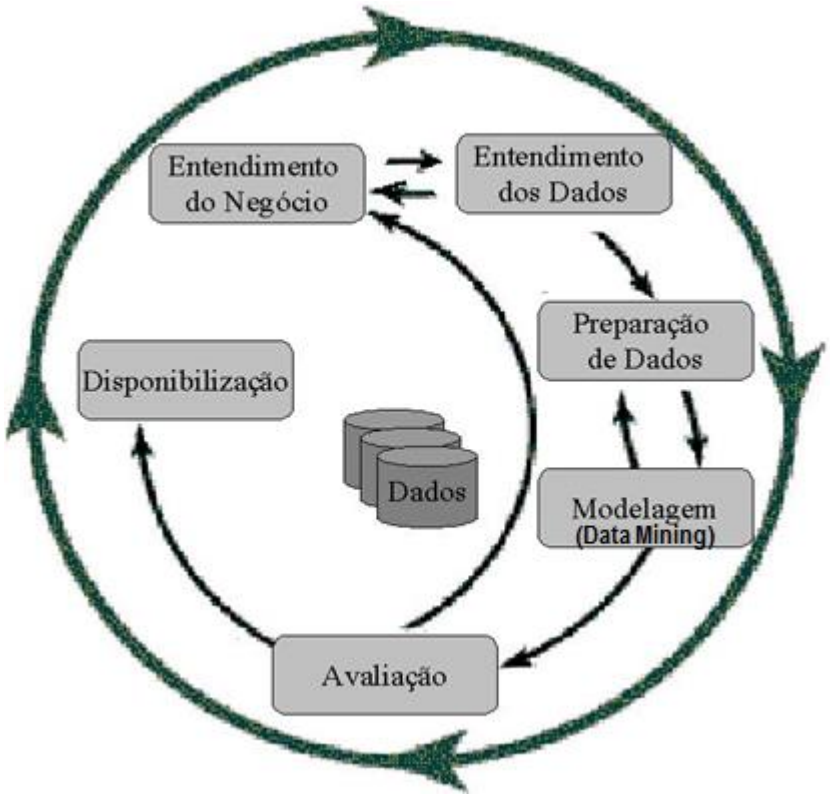


Figura 4 – O ciclo e as fases da metodologia CRISP-DM.

2.2.1 Entendimento do Negócio ou Domínio

O entendimento do domínio é a fase inicial do processo de mineração de dados cujo foco é entender os requisitos do ponto de vista do domínio. Após o entendimento do domínio deve-se definir qual o problema que o projeto irá solucionar [10].

No domínio de desastres naturais existem inúmeras pesquisas feitas analisando dados históricos. Estudar e compreender estes trabalhos são passos importantes para descobrir aspectos que influenciam ou estão relacionados aos desastres.

Também é na etapa de Entendimento do Negócio que é feito o levantamento dos recursos disponíveis para o projeto (dados operacionais, arquivos com informações relevantes, *softwares*

a serem utilizados no projeto) e com base nos recursos levantados listam-se os riscos que podem fazer com que o projeto de mineração de dados fracasse. É relevante criar um glossário com os termos importantes para a compreensão do domínio do problema e determinar os objetivos do projeto [10].

2.2.2 Entendimento dos Dados

Segundo o guia do CRISP-DM [10], a primeira tarefa para entender os dados é a coleta inicial. O objetivo é descrever quais são as fontes de dados, as maneiras como foi possível obtê-los e qualquer tipo de problema encontrado nesta coleta.

Após a coleta inicial ter sido feita é importante descrever as informações técnicas sobre os dados como o formato dos dados, quantidade de registros, identificação dos atributos (colunas), etc. Esta descrição ajuda a fazer a exploração dos dados, gerando algumas estatísticas básicas, fazendo as primeiras descobertas e hipóteses que irão direcionar ao alcance dos objetivos estipulados no processo de mineração de dados. Através da exploração dos dados também é possível verificar a qualidade dos mesmos.

2.2.3 Preparação dos Dados

Segundo ADRIAANS [11], a etapa de preparação dos dados representa cerca de 60% do esforço aplicado em um projeto de mineração. Esta fase visa preparar os dados disponíveis, que geralmente não estão dispostos em formato adequado para a aplicação dos algoritmos de descoberta, análise e a extração de conhecimento ALVARES [12]. Os dados precisam ter qualidade, isto é, estar limpos e compreensíveis, para extrair conhecimento interessante.

ALVARES [12] destaca que é comum os dados estarem representados em formatos diferentes, tais como arquivo-texto, planilhas, bancos de dados e outro tipo de fonte. Por isso é necessária a padronização e a integração dos dados, já que quase todo algoritmo de mineração de dados trabalha com uma única tabela de dados de entrada.

Por ser um dos processos mais importantes na mineração de dados, a preparação dos mesmos está dividida nas seguintes etapas (guia do CRISP-DM [10]):

- **Seleção de dados:** obtenção dos atributos (colunas) mais relevantes seja, de forma manual ou utilizando algoritmos, segmentação dos dados, eliminação direta (excluir itens com um determinado valor em um determinado atributo), agregação de dados (quando se deseja agregar os valores dos atributos de um determinado período), etc. O que motiva fazer esta seleção dos dados é aumentar a simplicidade e a acurácia dos resultados.

- **Limpeza de dados:** basicamente existem três objetivos em realizar a limpeza dos dados. O primeiro é a remoção de dados ausentes ou o preenchimento com valores constantes. O segundo objetivo é a limpeza de inconsistências como, por exemplo, um registro pode ser gravado com latitude ou longitude incorreta do município. O terceiro é a remoção de valores que não estão de acordo com o escopo e o domínio.

- **Transformação de dados:** O objetivo da transformação de dados é obtê-los em um formato mais apropriado para que os algoritmos de mineração consigam chegar a um melhor resultado.

É nesta fase que deve ser feita a normalização dos dados. A normalização é utilizada em atributos numéricos para minimizar os problemas oriundos do uso de unidades e dispersões. Um exemplo deste problema seria o número de habitantes de um município. Existem municípios que possuem milhares de habitantes e outros que possuem milhões e esta diferença pode influenciar no resultado da execução de determinados algoritmos de mineração. Deve-se fazer a transformação os dados para que eles fiquem em uma faixa de intervalo (0 a 1) de modo a evitar a dispersão. A Tabela 1 exemplifica o processo de normalização através do maior valor encontrado no conjunto de dados. O município de São Paulo é o que possui a maior população e com base nisto foi feita a divisão da quantidade de habitantes dos demais municípios (habitantes de *X*/habitantes de São Paulo).

Município	Habitantes	Habitantes Normalizados
São Paulo	11253503	1
Florianópolis	421240	0.04
Porto Alegre	1409351	0.13
Recife	1537704	0.14
Chapecó	183530	0.02

Tabela 1 – Normalização de hab. dos municípios com base no maior valor encontrado.

Caso seja necessário, dependendo do tipo de algoritmo de mineração a ser utilizado, deve-se fazer a transformação de valores numéricos para categóricos e vice-versa (quanto aumentou ou diminuiu a temperatura do Oceano Pacífico – *El Niño/La Niña* ou pode-se criar faixas de temperatura que determinem se foi um ano com *El Niño/La Niña* forte, moderado ou fraco).

- **Construção de dados:** nesta etapa é onde devem ser criados novos atributos que vão melhorar a etapa da mineração. Um exemplo de dado que poderia ser construído seria a densidade demográfica a partir da divisão da população pela área da região.

2.2.4 Mineração dos Dados ou Modelagem

Na metodologia CRISP-DM a fase de mineração dos dados é chamada de modelagem. É na modelagem que ocorrem as execuções dos algoritmos sobre o conjunto de dados. Esta fase é dividida nas seguintes etapas: seleção dos algoritmos, geração do projeto de teste, aplicação dos algoritmos e avaliação do modelo gerado, quantas vezes for necessário para obter o melhor resultado de acordo com o entendimento do negócio.

Cada técnica como classificação, agrupamento, etc, possui diversos algoritmos. Estes algoritmos têm dados de entrada diferentes. Por exemplo, o algoritmo de classificação ID3 QUINLAN [19] utiliza somente valores categóricos, já o algoritmo C4.5 QUINLAN [19] é uma extensão do ID3 que trabalha com valores numéricos. Algumas das principais técnicas de mineração de dados serão descritas na seção 2.3 deste trabalho.

Na etapa de modelagem, todo pré-processamento do KDD já deve ter sido elaborado pelo menos uma vez. A fase de modelagem pode ser executada diversas vezes para ajustar os

conjuntos de parâmetros de cada algoritmo, no intuito de obter resultados mais satisfatórios aos objetivos pré-estabelecidos.

2.2.5 Avaliação

O guia do CRISP-DM [10] divide este processo em três tarefas: avaliação dos resultados, revisão do projeto e determinação dos próximos passos.

Na avaliação dos resultados, deve ser verificado se os resultados atingiram os objetivos do projeto.

Na revisão do projeto deve ser analisado se o resultado é satisfatório para os objetivos definidos. É apropriado fazer uma revisão mais aprofundada da mineração de dados a fim de determinar se existe mais algum fator importante ou tarefa que tenha sido negligenciada.

Dependendo dos resultados da avaliação e revisão do processo, o analista decidirá se o conhecimento descoberto é suficiente para alcançar os objetivos. Caso não seja, deverá ser reiniciado todo o processo de mineração, voltando até a fase de entendimento do negócio e entendimento dos dados, podendo ser necessário realizar novas coletas e repetir toda a etapa de pré-processamento.

2.2.6 Disponibilização dos Padrões

Nesta etapa do processo deve-se produzir um relatório final com os resultados, mostrando os pontos positivos e negativos, os problemas encontrados e trabalhos futuros apresentados na fase de avaliação.

2.3 Principais Técnicas de Mineração de Dados

Existem duas maneiras principais de se fazer descoberta de conhecimento em banco de dados: predição e descrição.

A predição envolve usar valores conhecidos de atributos para predizer o valor desconhecido de uma variável de interesse [13] como, por exemplo: saber se uma determinada região é suscetível a deslizamento a partir de dados geológicos.

Já a descrição se concentra em encontrar padrões que descrevem os dados de forma compreensível para o analista [13]. O agrupamento dos municípios por mês, tendo como base o mês mais chuvoso em cada região pode ser um exemplo de descoberta de conhecimento descritiva.

Na seção 2.3.1 e 2.3.2 são apresentadas duas técnicas de mineração de dados, respectivamente, classificação (técnica preditiva) e agrupamento ou *clustering* (técnica descritiva).

2.3.1 Classificação

TAN [1] define classificação como a tarefa de mineração de dados que organiza objetos em diversas categorias pré-definidas. Classificar um registro é determinar com qual grupo de dados, já classificados anteriormente, este registro apresenta maior semelhança ALVARES[20]. Na classificação cada registro é caracterizado por uma dupla (x,y), onde x é o conjunto de atributos e y o atributo classe. Este último contem os valores no qual os registros devem ser categorizados. É importante que o atributo classe sempre contenha valores qualitativos, pois é isto que diferencia a classificação da regressão.

O resultado da execução de um algoritmo de classificação é a geração de um modelo que pode ser utilizado para atribuir uma classe a diferentes registros ainda não classificados. Supondo o tipo de desastre natural como um atributo classe (y), e características morfoclimáticas, geológicas e precipitações de chuva de uma determinada região como o conjunto de atributos (x), é possível gerar um modelo que possa predizer a relação entre os tipos de desastres e os demais atributos tendo como base um conjunto de ocorrências de desastres naturais. Conforme TAN [1] o modelo de classificação pode ser tratado como uma caixa preta que atribui automaticamente um rótulo de classe quando recebe o conjunto de atributos de um registro desconhecido.

Entre as técnicas de classificação existentes, duas delas são largamente utilizadas: os classificadores *eager* (espertos) e os classificadores *lazy* (preguiçosos).

Nos classificadores *eager*, a partir de uma amostragem inicial (conjunto de treinamento) é construído um modelo de classificação capaz de atribuir uma classe a novos registros. Uma vez

pronto este modelo, o conjunto de treinamento não é mais utilizado. São exemplos de algoritmos: árvores de decisão, redes neurais, redes bayesianas, máquinas de vetores e regras de decisão.

Nos classificadores *lazy* cada novo registro é comparado com todo o conjunto de treinamento e é classificado segundo a categoria do registro que é mais similar. O algoritmo que é o maior exemplo é o Método kNN (*k-nearest-neighbor* – vizinho mais próximo), COVER[21].

2.3.1.1 Árvore de Decisão

A árvore de decisão é uma representação gráfica dos registros de acordo com seus atributos e valores. A Figura 5 ilustra a transformação dos registros de maus pagadores de empréstimos em uma árvore de decisão.

A árvore é gerada a partir do conjunto de registros de treinamento e posterior a isto somente será utilizada para classificar os demais registros. Na árvore, cada nó representa um atributo, os ramos (ligam os nós) correspondem aos valores de um atributo e os nós folha (nós que não possuem sucessores) representam uma classe. A figura destaca o que são atributos, valores e classe. Dois algoritmos que utilizam a árvore de decisão são o ID3 (QUINLAN, 1986) e o C4.5 (QUINLAN, 1993). A diferença entre estes dois algoritmos é que o ID3 só processa atributos descritivos, já o C4.5 também processa atributos numéricos. O C4.5 é uma evolução do ID3.

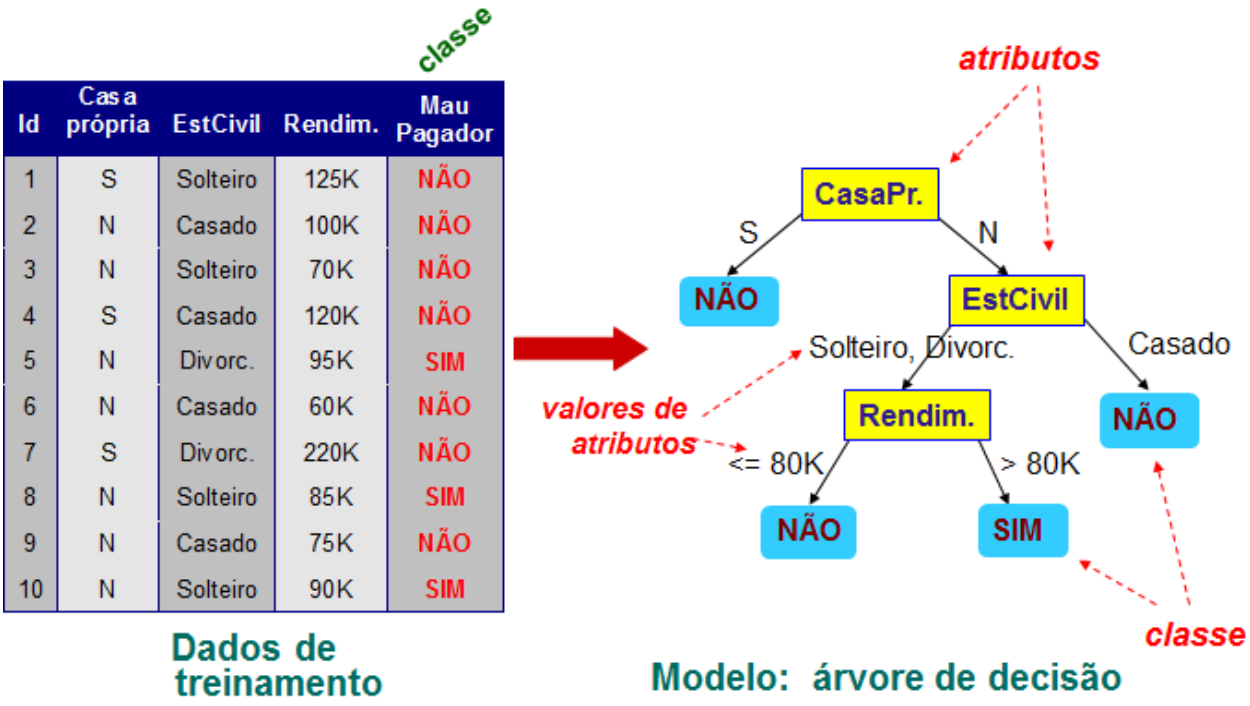


Figura 5 – Os elementos de uma árvore de decisão ALVARES [22].

2.3.2 Agrupamento ou *Clustering*

Segundo TAN [1], a análise de grupo, chamada *clustering*, une registros baseado apenas em informações encontradas nos atributos que descrevem os registros e seus relacionamentos. O ponto principal é entender que quanto maior for a semelhança dentro do grupo (*cluster*) e maior a diferença entre os grupos, melhor será o resultado encontrado, isto é, o agrupamento será mais preciso.

Um dos grandes problemas na utilização de algoritmos que realizam o agrupamento de dados está no próprio conceito do que é um grupo para o domínio estudado, pois um mesmo conjunto de dados pode conter diversos grupos. A Figura 7 retirada do livro de TAN [1] mostra que a divisão dos dados em grupos pode simplesmente ser um artefato do sistema visual humano, pois a partir de um determinado conjunto de registros (A), pode ocorrer a interpretação dos

grupos de até três maneiras diferentes: é possível gerar dois grupos (B), quatro grupos (D) ou seis grupos (C).

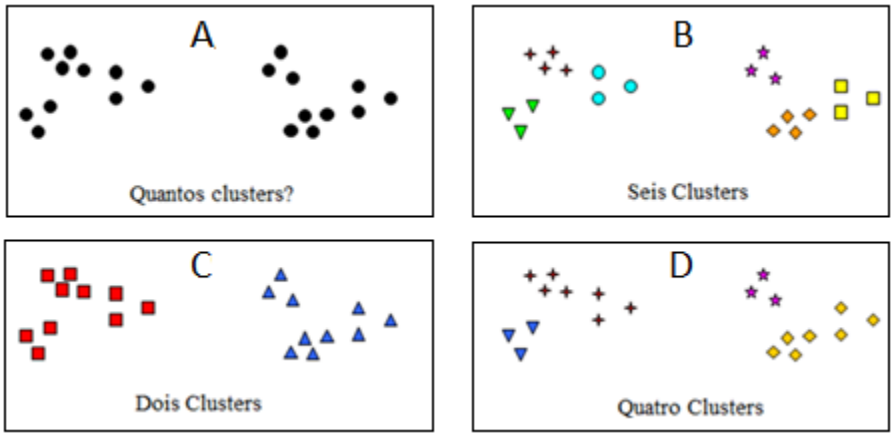


Figura 7 – Formas diferentes de agrupamento do mesmo conjunto de dados ALVARES [22].

Existem diversas técnicas de agrupamento, e a criação dos grupos de registros pode ser abordada das seguintes maneiras: agrupamento particional, hierárquico ou baseado em densidade TAN[1]. No agrupamento particional cada registro pertencerá a um único grupo, não podendo estar inserido dentro de um grupo maior ou menor. Já no agrupamento hierárquico é permitido que um grupo esteja inserido dentro de um grupo maior e que também possua subgrupos (o menor grupo possível sempre será um único registro).

Na Figura 8 é possível observar que, dependendo da interpretação, podemos ter um grupo dentro do outro e estas diferentes interpretações é uma característica de agrupamento hierárquico. Outra abordagem é o agrupamento baseado em densidade. Nesta abordagem um grupo será formado quando existir uma região densa, isto é, uma região com uma grande quantidade de registros.

A diferença desta abordagem para as outras é que ela é bastante tolerante a ruídos (registros que não deveriam pertencer a um determinado grupo) e também que ela consegue gerar grupos em formatos não regulares (círculos ou elipses) TAN[1].

Nas duas seções a seguir são apresentados alguns algoritmos de agrupamento.

2.3.2.1 Agrupamento Particional e o K-means

Segundo TAN [1], o K-means é a uma técnica de agrupamento particional, isto é, cada registro pertencerá a um único grupo. Estes grupos são formados a partir de um conjunto de registros mais próximos ao ponto central que define um grupo, e este ponto recebe o nome de centroide. A quantidade de centroides é um parâmetro que deve ser definido pelo analista de mineração de dados e o valor deste parâmetro pode variar a cada execução da técnica.

A escolha dos centroides iniciais influencia o resultado. O exemplo usado por ALVARES [13], ilustrado na Figura 8, mostra em (B) um particionamento ótimo dos pontos originais (A) e em (C) um particionamento sub-ótimo, considerando três centroides. Em (C), o grupo maior (losangos) acabou sendo dividido em dois pois dos três registros escolhidos como centroides, dois estavam no grupo maior.

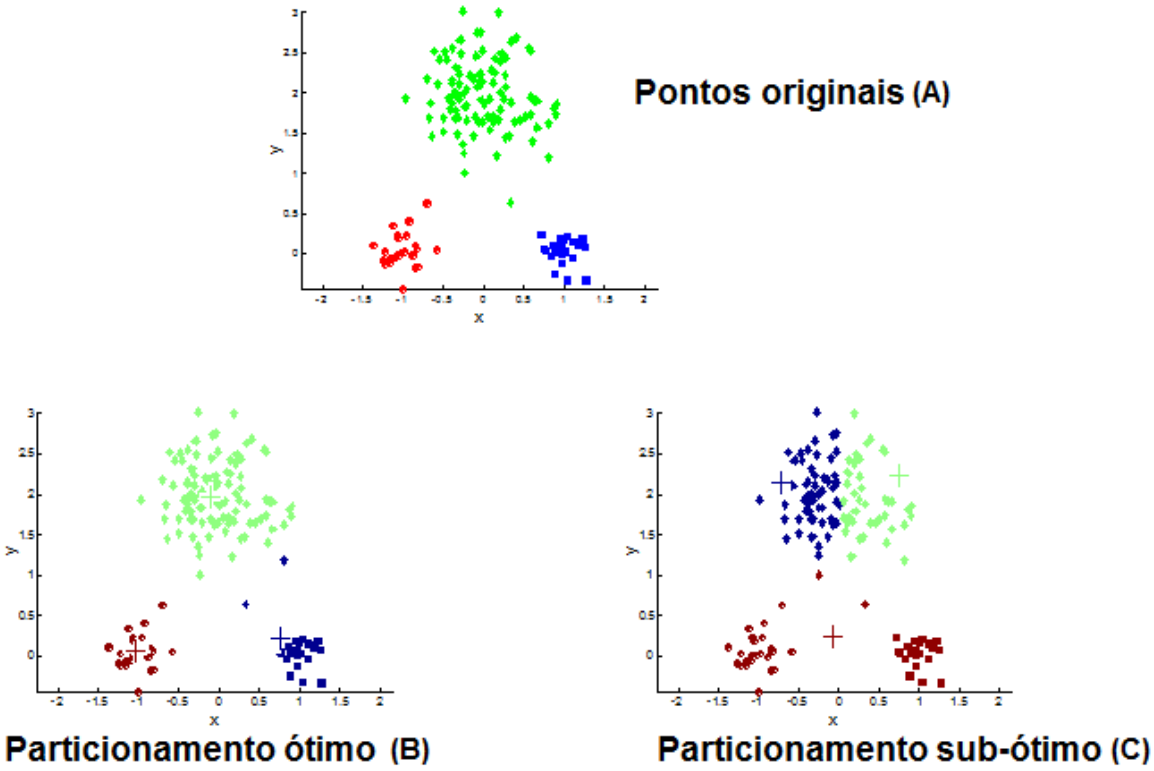


Figura 8 – Gráfico dos objetos mostrando os pontos originais e os grupos gerados.

Algumas limitações encontradas no K-means é que está técnica não se mostra muito eficiente em conjuntos de registros com tamanhos e densidades diferentes e também quando o conjunto de objetos não possui formato esférico. O K-means se baseia da proximidade com um centróide. A figura 9 ALVARES[13] ilustra um caso onde visualmente existe dois grupos distintos, mas que quando aplicado o K-means a divisão ocorre de maneira não-ótima, ficando os grupos gerados de uma maneira que não era esperada.

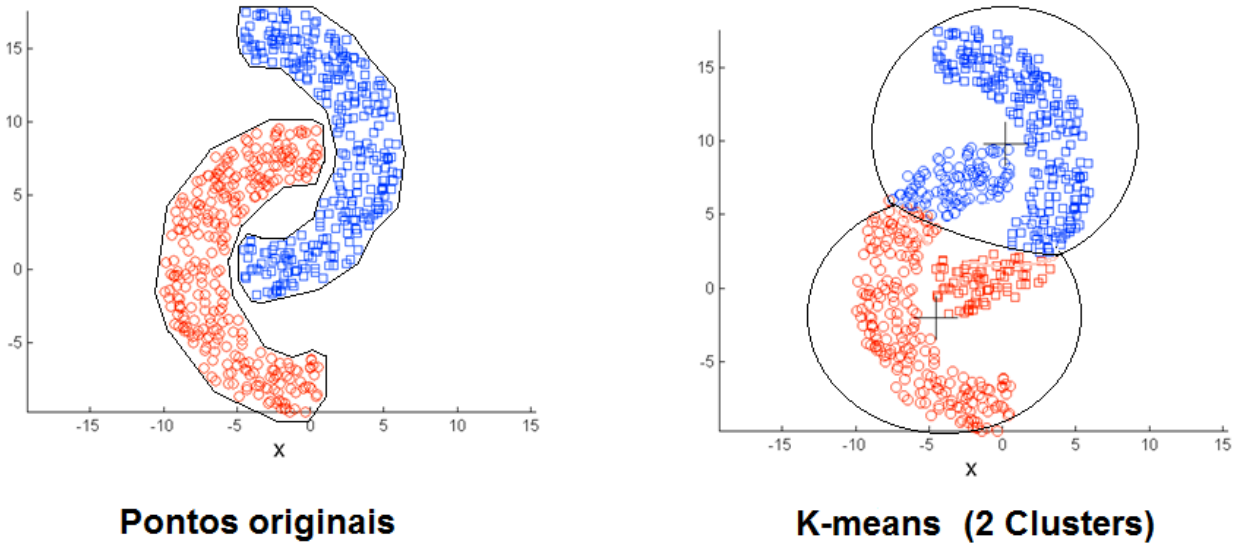


Figura 9 – Limitação do k-means com formatos não esféricos

2.3.2.2 Baseado em Densidade – DBSCAN

O DBSCAN é um algoritmo de agrupamento baseado na densidade dos registros TAN [1]. Uma região densa é uma região de alta densidade de registros. A utilização desta técnica é ideal quando grupos são irregulares ou entrelaçados e também quando existem muitos ruídos externos.

O DBSCAN consegue agrupar os registros através da utilização de dois parâmetros que devem ser definidos pelo analista. O primeiro é a densidade, que é quantidade mínima de registros que devem estar inseridos dentro do raio (*minPts*). O segundo é o raio (*Eps*), que é o espaço que será analisado a partir de um determinado registro. A figura 11 mostra o *minPts = 5* e o *Eps = 1* utilizados pelo DBSCAN e como é feito este agrupamento.

Caso um registro tenha a quantidade mínima de registros dentro do seu raio de tamanho *Eps* ele receberá o nome de *core point* (é o caso do losango na Figura 11, que possui cinco pontos dentro do seu raio). Se o registro não tiver a quantidade mínima de registros dentro do seu raio, mas ele está inserido dentro do raio de um *core point*, ele será um *border point* (é o caso dos círculos na Figura 11). Caso o registro não tenha a quantidade mínima dentro do seu raio e não está dentro do raio de um *core point* (triângulo desenhado na Figura 11)ele será chamado de *noise point*. Os *core points* e *border points* serão considerados os grupos e os *noise points* serão descartados.

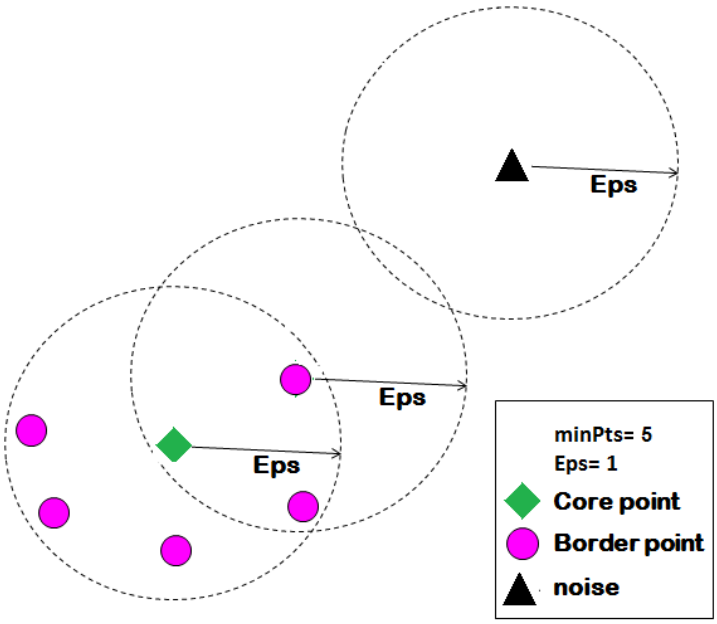


Figura 11 – MinPts, Eps, Core Point Border Point e Noise Point ALVARES[13].

O problema do DBSCAN é quando existem grupos que possuem densidades muito diferentes um do outro e também quando existem dados com alta dimensionalidade, pois a densidade é mais difícil de definir.

3. MATERIAIS E MÉTODOS USADOS PARA O PROCESSO DE MINERAÇÃO DE DADOS

Neste capítulo é apresentada a metodologia adotada com o intuito de mostrar como foram obtidos e manipulados todos os dados de modo que fique esclarecido o conjunto final de registros e quais algoritmos foram utilizados. Todos os resultados apresentados no capítulo 4 estão baseados nas manipulações e restrições descritas neste capítulo.

As subseções deste capítulo foram criadas com base nas etapas do CRISP-DM. A etapa de Entendimento do Negócio ou domínio já foi descrita no capítulo 1 através da introdução ao assunto e objetivos do trabalho. Neste capítulo estão descritas as etapas de Entendimento dos Dados (seção 3.1), seguindo da Preparação dos Dados (seção 3.2) e finalizando com os Algoritmos Utilizados (seção 3.3). As etapas de Avaliação e Disponibilização estão descritas no capítulo 4.

3.1 Entendimento dos Dados de Desastres Naturais

Foram utilizados os dados do CEPED, IBGE, CPRM (Serviço Geológico do Brasil) e NOAA (Agência Americana de Monitoração Climática). Os dados coletados são referentes aos registros de desastres entre 1991 e 2011. Abaixo são descritos os dados que foram utilizados em todo o processo:

- No total foram coletados mais de 41.217 registros oficiais junto à SEDEC (Secretaria Nacional de Defesa Civil) dos mais diversos tipos de desastres naturais ocorridos em todo o Brasil. Este conjunto de dados apresenta como atributos: *nome do desastre natural*, *dia*, *mês*, *ano*, *estado e município de ocorrência e a densidade demográfica do município*. Neste conjunto de dados também existem atributos sobre o próprio desastre: *quantidade de mortos, feridos, desabrigados, desalojados além dos danos materiais e financeiros causados (habitações destruídas e danificadas)*.
- Também foi coletada a temperatura, em grau Celsius, a intensidade do *El Niño* ou *La Niña* no mês e ano em que ocorreu o desastre natural. Segundo pesquisas apresentadas por BARBIERI [14], os fenômenos do *El Niño* e da *La Niña* (aquecimento e resfriamento anômalo

das águas do Oceano Pacífico Sul, respectivamente) são fatores que podem contribuir para a ocorrência de desastres naturais, assim seria importante que também fosse realizado o estudo levando em consideração este dado.

A Agência Americana de Monitoração Climática, com base nos seus estudos, categorizou a intensidade do *El Niño* e da *La Niña* como fraco, moderado e forte. Para que seja considerado que existiu a ocorrência de *El Niño* ou *La Niña* a temperatura do Pacífico deve variar pelo menos 0,5 graus centígrados positivos ou negativos. Com relação aos intervalos de intensidade, para que seja considerado fraco, moderado ou forte a temperatura deve ter alterado entre 0,5 e 0,9 (fraco), 1,0 e 1,4 (moderado) e maior ou igual a 1,5 graus (forte), sendo positivamente para o *El Niño* e negativamente para a *La Niña*.

- Também foi obtido o atributo binário (0,1) se o município possui áreas suscetíveis a deslizamentos. Este atributo foi levantado através de estudos feitos pelo CPRM [16], que é a instituição do governo responsável por organizar e sistematizar o conhecimento geológico do território brasileiro. Caso seja um município que tem áreas suscetíveis o valor será 1, caso contrário 0 (zero).

- Outros dados levantados a partir da ANA (Agência Nacional de Águas) foram os índices pluviométricos mensais dos estados brasileiros, por estação de medição. Para cada mês é quantificado o total de chuvas em milímetro, o número de dias em que choveu, a máxima em milímetros de precipitação em um dia. Estas informações são importantes para identificar a relação entre mês e dia de ocorrência de um desastre natural com os números pluviométricos do mesmo mês.

- Também foram obtidos, do IBGE, dados de relevo e solo predominante de cada município, a partir de arquivos *shapefile*[1] que continham estas informações. Do arquivo *shapefile*[1] sobre relevo foi possível obter características geomorfológicas (formas da superfície terrestre), unidade de relevo (depressão, planalto, etc) e a estrutura de relevo em que se situa o município (Marinhas, Bacia do Paraná, Araucária, Alto do Tocantins, etc). Do arquivo que traz dados sobre o solo brasileiro foi obtido o tipo predominante por município.

Como parte da etapa de entendimento dos dados, foram levantadas algumas informações estatísticas sobre os mesmos com o intuito de direcionar os objetivos do projeto de mineração,

facilitando a escolha dos algoritmos, além de contribuir para o refinamento da descrição dos dados.

A Tabela 2 mostra resumidamente todos os atributos que foram utilizados para o processo de mineração de dados. A coluna da esquerda mostra o nome do atributo e o da direita um exemplo do valor.

Atributo	**Exemplo de Valor**
Tipo do desastre	Deslizamentos
Ocorrência de La Niña/El Niño	*La Niña* Moderada
Mês de ocorrência do desastre	Outubro
Ano de ocorrência do desastre	1999
Estado (UF)	Santa Catarina
Nome da mesorregião	Vale do Itajaí
Nome da microrregião	Blumenau
Nome do município	Blumenau
Munícipio em área de Amazônia	N*(Não)*/S*(Sim)*
Munícipio em área de fronteira	N*(Não)*/S*(Sim)*
Município com área suscetível a deslizamento	0*(Não)*/1*(Sim)*
Domínio morfológico predominante do município	Embasamentos em Estilos Complexos
Subdomínio morfológico predominante do município	Embasamento do Sul/Sudeste
Unidade de relevo predominante do município	Serras
Localização da unidade de relevo	Leste Catarinense
Tipo de solo predominante do município	Solo Podzólicos
Densidade do município (hab/km²)	160,4
Altitude do município (metros)	90
Total de chuvas em um mês (milímetros)	200
Máxima de chuvas em um dia (milímetros)	90
Números de dias com chuva em um mês	20

Tabela 2 – Atributos utilizados no processo de mineração de dados.

Todos os dados foram passados para um sistema de gerenciamento de banco de dados de modo a facilitar a manipulação e o entendimento dos mesmos. Assim, utilizaram-se consultas SQL (Linguagem de Busca Estruturada) e a partir disto foram elaborados alguns gráficos tendo

como base as projeções das consultas. A seguir estão listados estes gráficos que mostram informações referentes aos registros.

Na Figura 12 estão apresentados os tipos de desastres naturais que mais ocorrem no Brasil. O maior número de registros foi referente à estiagem, cerca de 39%; o segundo maior foi enxurrada com aproximadamente 23%; e o terceiro maior foi inundação, com 14,23% dos casos. Desastres naturais como deslizamentos e alagamentos somam aproximadamente 2,5%. Entretanto mesmo sendo um percentual baixo, ainda é importante estudar estes desastres isoladamente devido ao impacto humano que ele gera.

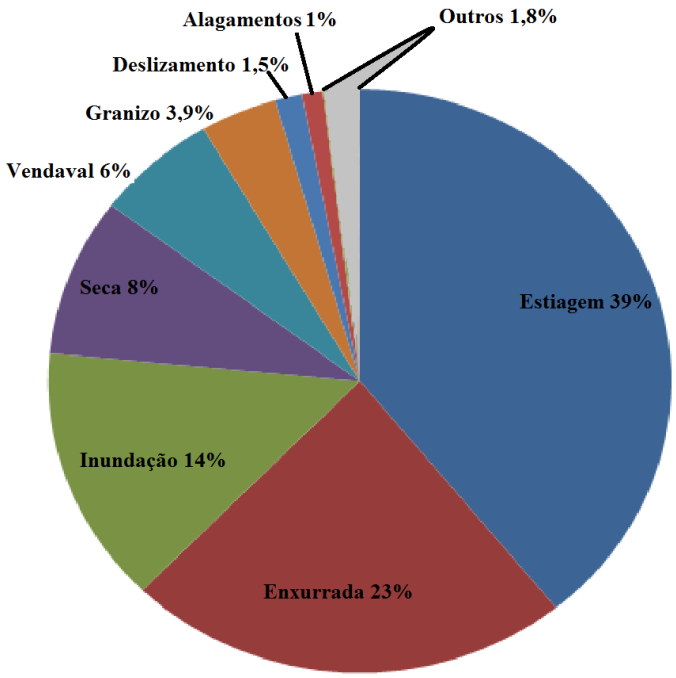


Figura 12 – Dispersão dos relatos de desastres naturais.

Agrupando registros de desastres naturais por microrregião brasileira foi possível encontrar uma boa dispersão da ocorrência dos desastres e isto pode ser um fator importante na aplicação dos algoritmos de mineração, já que diminui o risco do resultado ser tendencioso para um determinado desastre.

De acordo com o IBGE [23], microrregião é um agrupamento de municípios limítrofes com o objetivo de estruturar o espaço geográfico de acordo com a produção agropecuária, industrial, extrativista ou pesca, além de trocas de consumo entre os munícipios e atividades

urbanas e rurais. A Figura 13 mostra o Brasil com todas as 558 microrregiões, sendo que uma microrregião pertence necessariamente a um único estado brasileiro.

Figura 13 – Microrregiões brasileiras.

Um exemplo seria a microrregião de Florianópolis (Figura 14), que abrange os municípios de Antônio Carlos, Biguaçu, Florianópolis, Governador Celso Ramos, Palhoça, Paulo Lopes, Santo Amaro da Imperatriz, São José, São Pedro de Alcântara.

Figura 14 – Microrregião de Florianópolis

Apesar da Figura 15 não mostrar todo o mapa brasileiro (por motivo de falta de espaço), foi feita uma comparação entre todas as microrregiões brasileiras destacando as vinte microrregiões que mais registraram desastres naturais.

A microrregião que obteve a maior quantidade de registros foi a de Chapecó-SC, com 701 registros de desastres naturais (Figura 15). Este valor representa o total de 2% em relação a todos os registros de desastres (41217 ocorrências). Obtendo a relação das vinte microrregiões que mais registraram desastres naturais, Chapecó, representa um total de 10%.

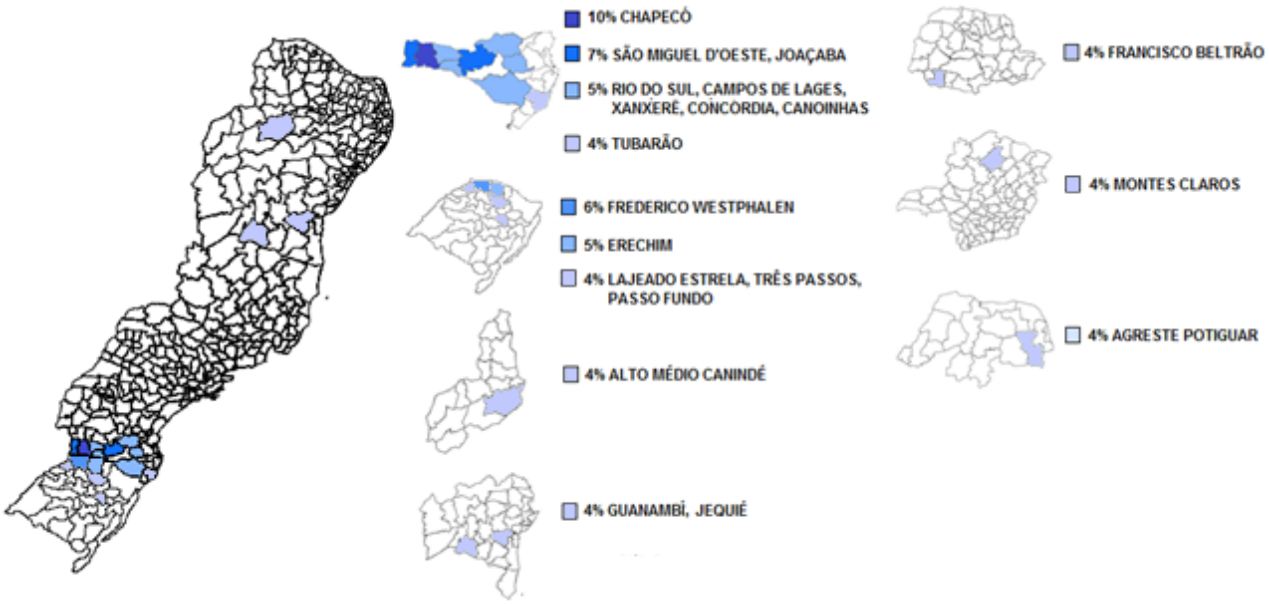


Figura 15 – Dispersão das incidências de desastres naturais por microrregião.

Ainda fazendo uma análise das estatísticas de desastres naturais juntamente com todas as microrregiões, foi possível observar que para as calamidades que ocorrem com maior frequência (estiagem, enxurrada, vendaval, granizo e inundação) a microrregião de Chapecó-SC sempre esteve entre as que mais registraram desastres naturais. As figuras 16, 17, 18, 19, 20 apresentam, respectivamente a dispersão os gráficos dos desastres de estiagem, enxurrada, vendaval, granizo

e inundação. Todas as porcentagens são em relação às microrregiões apresentadas no gráfico e não em relação ao total de registros.

Na Figura 16 é possível verificar que as microrregiões que mais registraram estiagens foram as do oeste catarinense, noroeste gaúcho e da zona de transição do agreste para o sertão baiano.

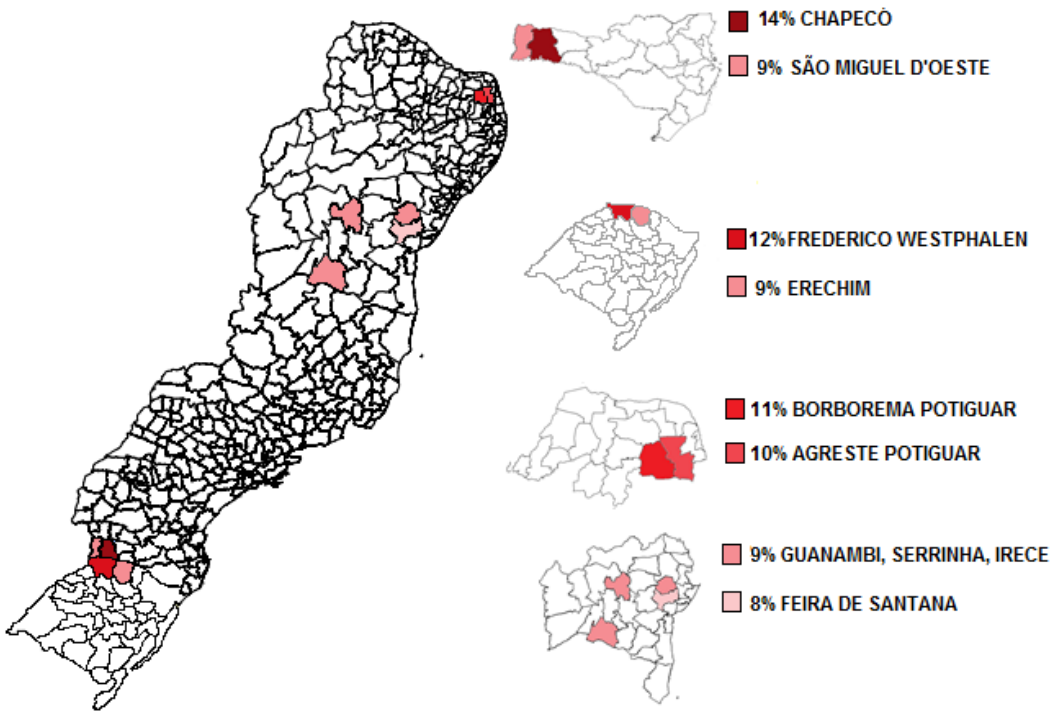


Figura 16 – Dispersão das incidências de estiagens por microrregião.

Já na Figura 17 foram obtidas as 20 microrregiões que mais apresentaram enxurradas. As dez microrregiões que mais apresentaram registros de enxurradas são catarinenses, sendo que outros estados que também apresentaram ocorrências foram Pernambuco, Rio Grande do Sul, Bahia e Espírito Santo.

Sobre Santa Catarina é possível afirmar que o estado é uma unidade federativa que está sempre sob ameaça deste tipo de tragédia. Destaque para a microrregião de Chapecó, que novamente aparece entre as microrregiões que mais registrou ocorrências.

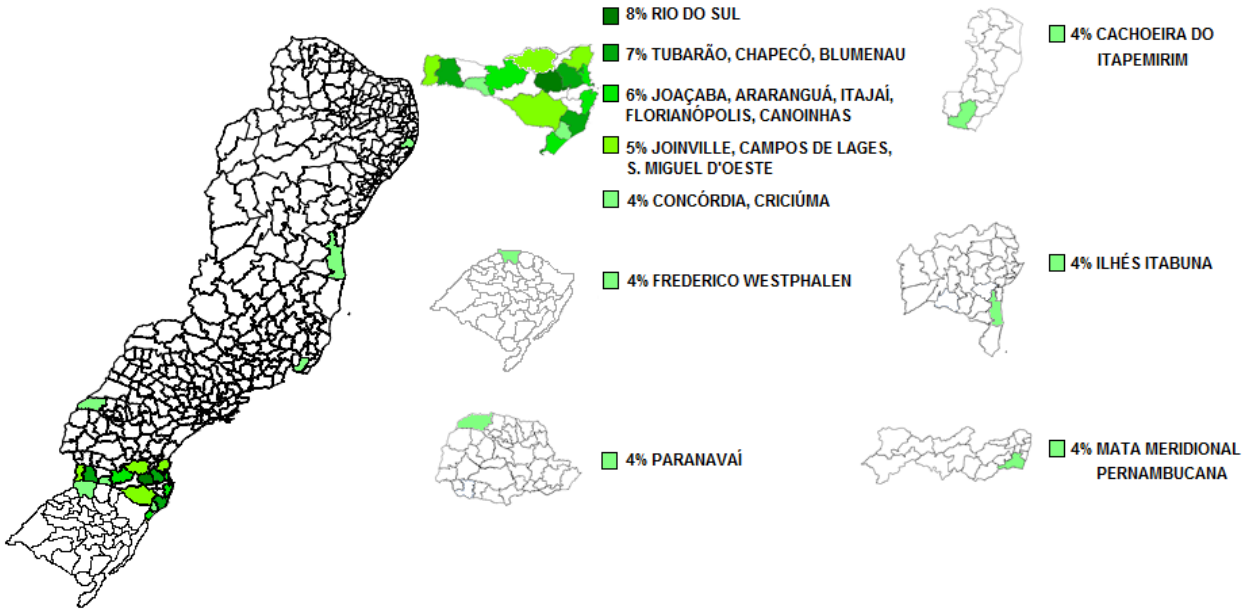


Figura 17 – Dispersão das incidências de enxurradas por microrregião.

Para a ocorrência de vendavais, as microrregiões sulistas foram as que apresentaram a maior quantidade de registros (Gráfico 5), principalmente em microrregiões do interior dos três estados. Este tipo de estatística proporciona realizar um estudo mais focado na região sul.

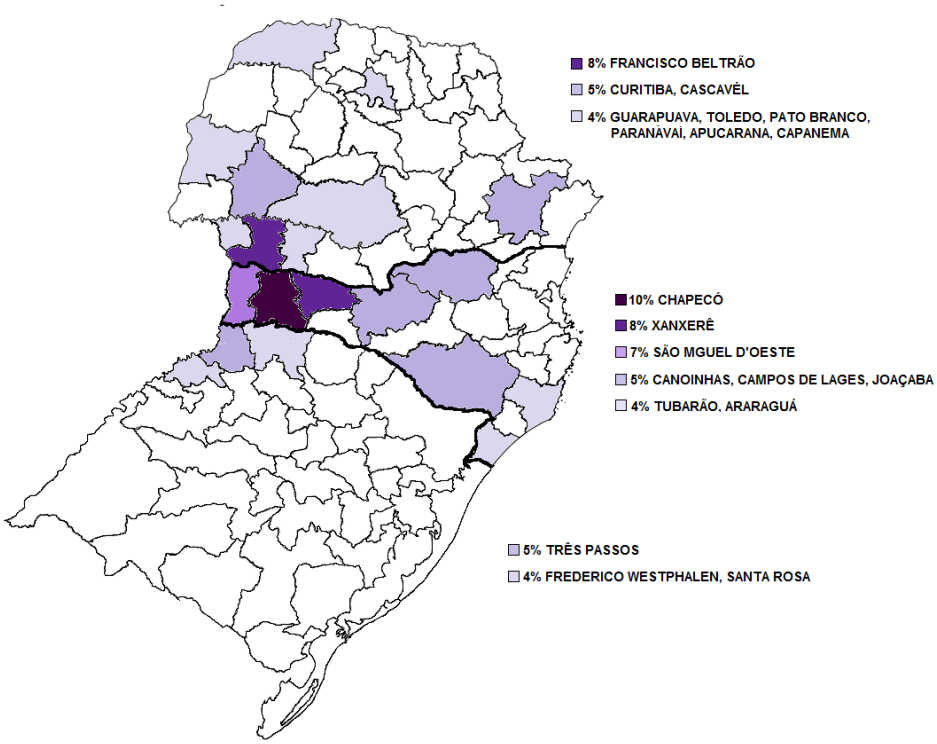


Figura 18 – Dispersão das incidências de vendaval por microrregião.

Novamente a microrregião de Chapecó volta a se destacar quando os desastres analisados foram referentes à ocorrência de granizo, e a região sul também se mostrou ter uma forte influência de características que geram este tipo de desastre.

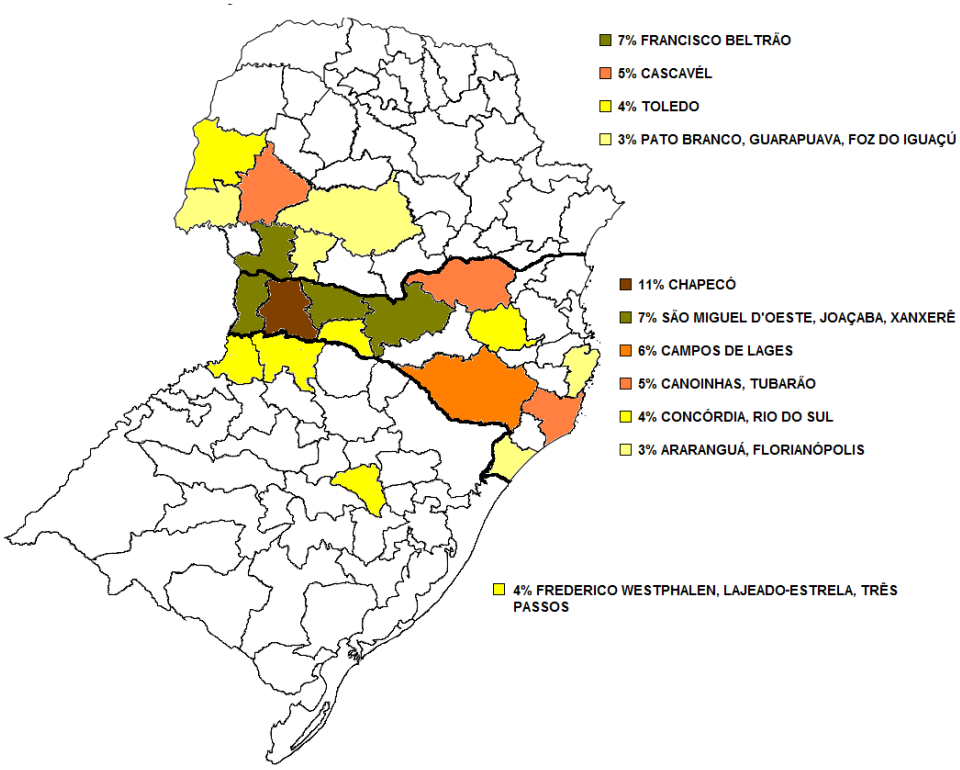


Figura 19 – Dispersão das incidências de granizo por microrregião.

Na análise dos dados referentes ao registro de inundações, além de diversas microrregiões do sul do Brasil, também aparece o Rio de Janeiro, o norte do país (Pará) e o nordeste (Bahia e Rio Grande do Norte). Na Figura 20 estão apresentados os estados brasileiros que possuíam as microrregiões que mais apresentaram registros de inundações (estados da região sul, Minas Gerais, Bahia, Rio Grande do Norte, Rio de Janeiro e Pará).

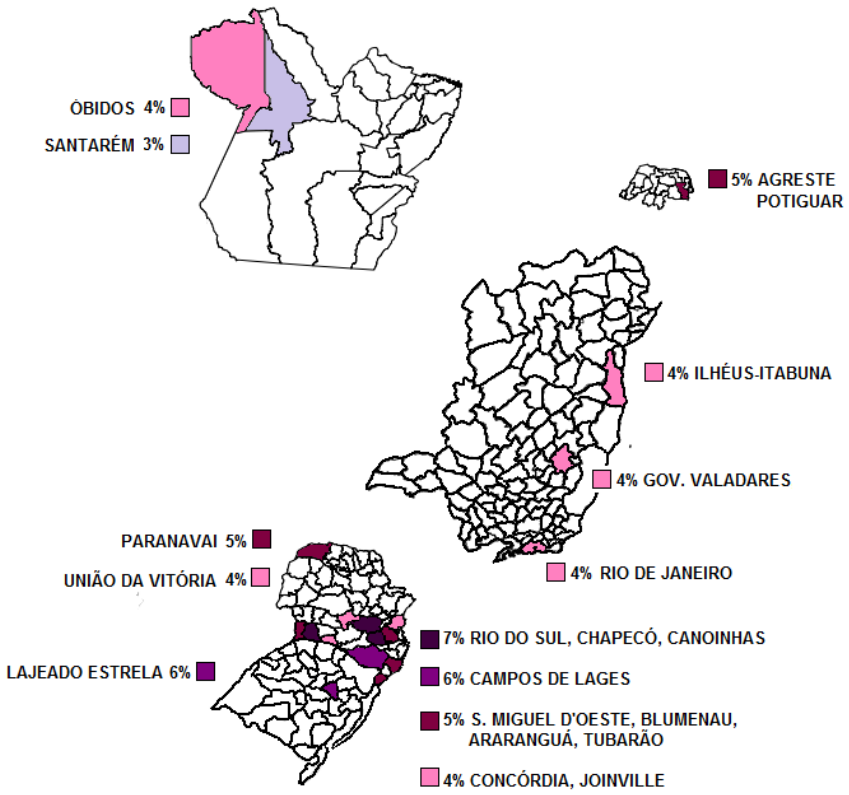


Figura 20 – Dispersão das incidências de inundações por microrregião.

Também foram obtidas estatísticas referentes à ocorrência dos desastres naturais no decorrer dos anos, apresentada na Figura 21. O número em cima de cada barra da figura representa a quantidade de registros que existiram no ano.

Sobre a distribuição dos desastres dos últimos vinte anos observou-se que a média de registros entre 1991 e 2000 era de 927 ocorrências por ano (não foi registrado nenhum ano que ultrapassou dois mil registros), e de 2001 até 2011 ocorreu um salto na média no registro para 2287 ocorrências por ano.

Mesmo com esta informação não é possível afirmar que houve um aumento real na incidência de desastres, visto a fragilidade histórica do Sistema de Defesa Civil em manter atualizados os registros [17]. Com o passar dos anos o sistema de registros tem se fortalecido e a fidelidade dos números só tende a aumentar.

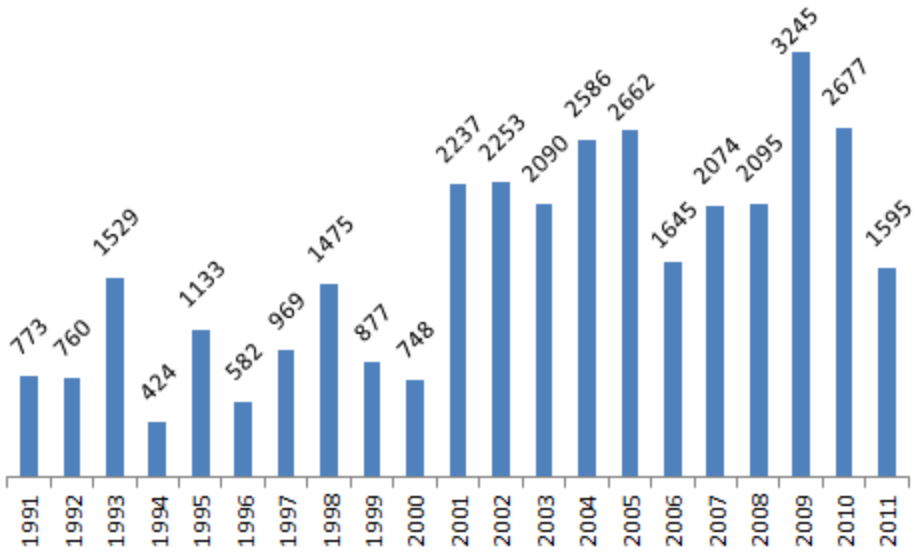


Figura 21 – Distribuição da ocorrência dos desastres naturais no decorrer dos anos.

Com base nestas estatísticas iniciais é possível aprofundar o estudo em algumas regiões para buscar entender porque algumas têm uma maior tendência a possuir mais desastres naturais que outras.

3.2 Preparação dos Dados Referente aos Desastres Naturais

A transformação dos dados foi realizada de três maneiras: manipulação manual das tabelas no MS Office, criação de consultas SQL e a utilização da ferramenta Pentaho Kettle. Optou-se também por utilizar o Kettle, que é uma ferramenta comumente utilizada em *data warehouse,* no processo de extração, transformação e carga de dados, pois o pré-processamento em mineração de dados é relativamente parecido com este processo. Na mineração de dados é necessário unir fontes de dados de tabelas e arquivos distintos, manipular estes dados e condensá-los em um único arquivo. O Kettle mostrou ser uma ferramenta fácil de usar principalmente por permitir uma programação visual de todo o processo de transformação dos dados e também proporcionou a economia de tempo na manipulação dos dados. No Anexo A é possível visualizar a transformação realizada no processo de preparação dos dados. Todas as etapas do processo de preparação dos dados estão descritas nos parágrafos seguintes.

Como nem todos os municípios que possuíam registros de desastres naturais possuíam registros pluviométricos (*máxima de chuva em um dia, total de chuvas no mês e quantidade de dias com chuva no mês*) foi necessário criar duas tabelas, uma com todos os registros de desastres naturais (41.217 registros de desastres) e outra somente com os registros de desastres naturais que ocorrem em municípios que tinham dados pluviométricos (6.864 registros de desastres naturais). Esta divisão ocorreu, pois possuir muitos registros com valores nulos influencia no resultado da execução dos algoritmos de mineração de dados. Com isto, feitas análises considerados os registros com dados pluviométricos e análises não considerando os registros que não possuíam dados pluviométricos.

A primeira tabela gerada, com os 6.864 registros que desastres que possuem os dados pluviométricos, contem os seguintes atributos: *tipo do desastre, mês e ano da ocorrência, unidade federativa, município, densidade demográfica, índice de suscetibilidade de deslizamento, se ocorreu El Niño ou La Niña no mês do desastre e a intensidade, quantidade total de chuvas em milímetro no município no mês do desastre, número de dias em que ocorreram precipitações no município no mês do desastre, quantidade máxima em milímetros de precipitação em um dia também no mês do desastre, tipo de solo predominante do município em que ocorreu o desastre, domínio e subdomínio morfológico do município e unidade do relevo.*

Já a segunda tabela gerada, com todos os 41.217 registros de desastres, contem todos os atributos anteriormente citados exceto os atributos relativos aos índices pluviométricos (*quantidade total de chuvas em milímetro no município, números de dias em que ocorreram precipitações no município* e *quantidade máxima em milímetros de precipitação em um dia*).

3.2.1 Seleção dos Dados

Nos dados obtidos pelo CEPED referente aos registros de desastres naturais foi necessário unir o código do município com outra tabela que possuía a relação entre código e o nome do município. A própria tabela de registros de desastres naturais possui uma coluna com o nome do município, porem não foi utilizada porque existiam problemas de acentuação.

Ainda desta tabela foram excluídos os atributos de danos materiais e financeiros, pois são fatores levantados após o desastre, não sendo considerados relevantes neste trabalho.

Para fazer análises específicas de um determinado desastre natural foram geradas tabelas secundárias baseadas na principal. Foi realizada a segmentação dos registros buscando somente os desastres a serem analisados (vendavais, granizo, deslizamento, alagamentos, etc).

Na tabela específica dos índices pluviométricos as três medidas apresentadas (quantidade total de chuvas no mês, número de dias de chuva, máxima de chuva em um dia) foram definidas por estação pluviométrica.

Como a menor região utilizada foi o município e cada município pode conter mais de uma estação, foram realizadas algumas agregações de valores nos dados. Foi calculada a média da quantidade total de chuvas e obtida a quantidade máxima de chuva em um dia. A Tabela 3 mostra como ficaram as agregações para o município de Blumenau – SC. Este município contem três estações pluviométricas. A quantidade máxima de chuvas no dia foi de 61mm, da estação Blumenau. A quantidade máxima de dias registrada com chuva no mês foi 15 (estação Itupoava Central) e a média do total de chuvas registrados por estação foi de 130,17mm.

Município	**Nome da Estação Pluv.**	**Máxima Dia (mm)**	**Dia da Máxima**	**Numero de Dias Com Chuva**	**Total de Chuva (mm)**
BLUMENAU	BLUMENAU	56	28/09/2005	8	145.60
BLUMENAU	GARCIA DE BLUMENAU	61	26/09/2005	14	127.20
BLUMENAU	ITOUPAVA CENTRAL	42	26/09/2005	15	117.70
		61	**26/09/2005**	**15**	**130.17**

Tabela 3 – Índices pluviométricos do município de Blumenau – SC

Outra tabela que também passou pelo processo de seleção de dados foi a tabela que continha informações referentes ao relevo e solo. Como estas informações foram retiradas de dois arquivos *shapefile* diferentes (um para relevo e outro para solo) e nestes arquivos não continha informações sobre municípios foi necessário realizar a união com outro *shapefile* que continha todos os municípios. Após essa união foi possível saber qual é o tipo de relevo e solo predominante em cada município.

Da tabela de relevo foi excluída a coluna que trazia a informação sobre o tipo de sedimentação do relevo (interiorana ou litorânea) visto que para 60% dos municípios este

registro estava com valor nulo (não definido), o que atrapalharia a aplicação dos algoritmos de mineração de dados.

3.2.2 Limpeza dos Dados

Limpeza dos dados é uma operação básica de remoção de ruídos, atributos incompletos ou erros. O objetivo de realizar a limpeza é filtrar somente os dados relevantes. Conforme já citado no capítulo 2, alguns tipos de operações a serem realizadas são: limpeza de informações ausentes, inconsistências, valores fora do domínio e padronização de dados.

A Tabela 4 mostra o caso das estações do município de Curitibanos – SC, onde existem duas estações pluviométricas em que não foram registrados dados. Para estas estações, e todas as outras onde não existe registro, foi feita a remoção das mesmas do processo de mineração de dados.

Município	Nome da Estação Pluv.	Máxima Dia	Dia da Máxima	Numero de Dias Com Chuva	Total
CURITIBANOS	PONTE ALTA DO NORTE	40.9	8/9/2005	26	92.1
CURITIBANOS	PONTE DO RIO ANTINHAS	33	8/9/2005	26	112.8
CURITIBANOS	BARRAGEM PERY	sr	sr	sr	sr
CURITIBANOS	SALTO PERY	sr	sr	sr	sr
CURITIBANOS	PONTE ALTA DO NORTE – CIFSUL	32	11/9/2005	25	98.2
		40.9	**8/9/2005**	**26**	**101.03**

Tabela 4 – Índices pluviométricos com estações sem registros de Curitibanos – SC

Ao fazer a manipulação de dados também, existiam muitos municípios sem registros pluviométricos, conforme já mencionado na seção anterior. Para que se pudesse realizar uma análise mais consistente da relação entre os índices pluviométricos e os desastres naturais foi necessário remover os desastres cujas cidades não tinham registros de chuvas. Com isto o número de desastres analisados levando em consideração os registros pluviométricos foi de 6.863 (o número total de registros de desastres naturais, depois de toda a limpeza dos dados, foi de 40.936).

Outro problema também apresentado na tabela de registros de desastres naturais foi que quatro registros apresentaram problemas de data. Em todos os casos foram registradas datas que não existiam como: data de ocorrência em 30, 31 de fevereiro e 31 de abril.

Nas demais tabelas não foram encontrados registros que necessitassem passar pelo processo de limpeza dos dados.

3.2.3 Transformação dos Dados

Na tabela disponibilizada pelo CEPED foi feita a transformação dos seguinte valores numéricos para categóricos: o código do desastre natural (COBRADE – Classificação e Codificação Brasileira de Desastres) foi substituído pelo nome do desastre natural com intuito de facilitar a compreensão dos padrões na descoberta de conhecimento; o código do município, microrregião, macrorregião também foram transformados para os valores textuais correspondentes; e conforme já descrito na seção 3.1, a intensidade do *El Niño* e da *La Niña* foi classificado como fraco, moderado e forte. Para que seja atribuída uma das três intensidades a temperatura do *E Niño/La Niña* foi alterada para valores qualitativos onde: entre 0,5 e 0,9 é fraco, 1,0 e 1,4 é moderado e maior ou igual a 1,5 graus é forte, sendo positivamente para o *El Niño* e negativamente para a *La Niña*.

Além disso, foi criada outra coluna que informa se ocorreu El Niño ou La Niña, independente da força do fenômeno, com o intuito de gerar maiores possibilidades de encontrar padrões nos dados. A Tabela 5 mostra alguns exemplos desta transformação onde na coluna mais a esquerda está a temperatura em graus Celsius, na do meio está a temperatura em forma descritiva e na coluna mais a direita o tipo do fenômeno (*La Niña* ou *El Niño*) .

Intensidade do El Niño ou La Niña	Força do Fenômeno	Tipo do Fenômeno Anômalo
0.3	Sem El Niño/La Niña	Sem El Niño/La Nina
0.9	El Niño Fraco	El Niño
-1.9	La Niña Forte	La Niña
1	El Niño Moderado	El Niño

Tabela 5 – Exemplo da transformação de valores para categórico.

Outros quatro atributos que também foram transformados em intervalos foram: *densidade populacional*, *altitude*, *total de chuvas no mês* e *número de dias com chuva*. Para a densidade populacional foi atribuída a escala ilustrada na Tabela 6 (esquerda) a altitude na Tabela 6 (direita).

Densidade (hab/km²)
Menor de 1 hab/km²
Entre 1 e 5 hab/km²
Entre 5 e 10 hab/km²
Entre 10 e 20 hab/km²
Entre 20 e 50 hab/km²
Entre 50 e 100 hab/km²
Mais de 100 hab/km²

Altitude (metros)
Entre 20 e 99,9m
Entre 100 e 199,9m
Entre 200 e 299,9m
…
Entre 900 e 999,9m
Mais de 1000m

Tabela 6 – Escala de densidade populacional (esquerda) e altitude (direita).

Para os índices pluviométricos do munícipios brasileiros (total e máxima de chuvas em um dia) foram criadas duas escalas (Tabela 7). Uma escala que agrupa os registros de 50 em 50 mm (esquerda da Tabela 7) e outra com escala de 100 em 100 mm (direita da Tabela 7). Estes dois intervalos foram definidos a partir da análise visual dos registros de desastres que continham dados pluviométricos.

Total de chuvas e Máxima de chuvas em um dia (50mm)
Menos de 20mm
Entre 20 e 49,9mm
Entre 50 e 99,9mm
Entre 100 e 149,9mm
...
Entre 750 e 799,9mm
Entre 800 e 999,9mm
Mais de 1000mm

Total de chuvas e Máxima de chuvas em um dia (100mm)
Menos de 20mm
Entre 20 e 99,9mm
Entre 100 e 199,9mm
Entre 200 e 299,9mm
...
Entre 700 e 799,9mm
Entre 800 e 999,9mm
Mais de 1000mm

Tabela 7 – Escalas do total de chuvas e máxima de chuvas em um dia em 50mm (esquerda) e 100mm (direita).

Para a quantidade de dias com chuva no mês também foi criada uma escala para transformar o valor numérico em um intervalo. Esta escala está apresentada na Tabela 8 e os intervalos criados foram de 5 em 5 dias com base na análise visual dos registros.

Quantidade de dias com chuva
Menos de 5 dias
Entre 5 e 10 dias
...
Entre 20 e 25 dias
Entre 25 e 31 dias

Tabela 8 – Escala da quantidade de dias com chuva.

Como o índice de suscetibilidade de deslizamento estava com valores 0 (zero) e 1(um) foi feita a transformação do dado para *Não suscetível a deslizamentos* e *Suscetível a deslizamentos*, respectivamente. Isto foi realizado com o intuito de facilitar a interpretação dos dados e para a execução de algoritmos que não aceitam valores numéricos.

O ano de ocorrência do desastre também foi transformado para intervalos de 5 em 5 anos para que fosse possível executar o algoritmo ID3, já que este só aceita valores textuais. A Figura 22 exibe um gráfico mostrando a quantidade de desastres ocorridos (valor sobre cada barra) em cada intervalo de 5 anos.

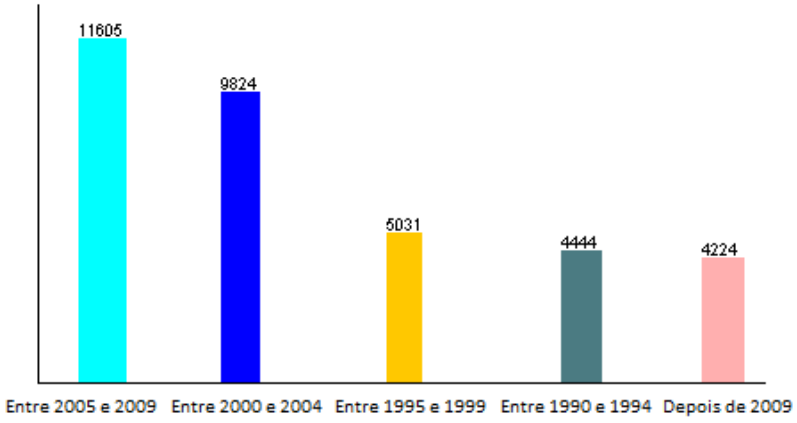


Figura 22 – Distribuição da ocorrência dos desastres naturais nos intervalos de 5 em 5 anos.

3.3 Algoritmos Utilizados

Para alcançar o objetivo do trabalho foram utilizadas duas técnicas de mineração de dados: classificação e agrupamento.

A ferramenta utilizada para aplicar os algoritmos existentes foi o Weka 3.6 FRANK [26]. Esta ferramenta possui diversos algoritmos implementados de praticamente todas as técnicas de mineração. Ela foi desenvolvida pela Universidade de Waikato, Nova Zelândia e é de código aberto e gratuita o que possibilita estudar e alterar o código fonte da ferramenta.

O algoritmo de árvore de decisão utilizado para classificar os registros foi o ID3 (QUINLAN, 1986) e para nível de comparação e análise do algoritmo de classificação com o resultado mais satisfatório, também foi utilizado o C4.5 (QUINLAN, 1993). A princípio foi utilizado como atributo classe o tipo de desastre natural, na tentativa de obter algum resultado interessante e tentar predizer quais atributos tem capacidade de influenciar a ocorrência de um desastre natural.

Já no agrupamento foram utilizados os algoritmos K-means (McQueen, 1967) para gerar grupos que terão sempre registros exclusivos em cada grupo e o DBSCAN (Ester *et al.* 1996) para analisar os registros baseado na densidade.

4. ANÁLISE DOS RESULTADOS

Foram feitas análises com os tipos de desastres que mais apresentaram ocorrências: *enxurrada*, *alagamento* ou *inundação*, *deslizamento*, *seca* ou *estiagem*, *granizo* e *vendaval*. Também realizou-se alguns pré-processamentos como a remoção de determinados atributos e a seleção de um ou mais tipos de desastres naturais para uma análise isolada, com o intuito de obter resultados melhores.

A Tabela 9 apresenta os mesmos atributos considerados no processo de mineração com um exemplo de valor para cada atributo. Todos os valores apresentados nesta coluna levam em consideração o processo de preparação dos dados, isto é, todos os valores que eram números foram passados para intervalos. As linhas destacadas na Tabela 9 mostram os atributos que tiveram seus valores alterados pelo processo de preparação dos dados.

Atributo	**Exemplo de Valor**
Tipo do desastre	Deslizamentos
La Niña/El Niño dividido em intervalos	*El Niño* Fraco; *El Niño* Moderado*; El Niño* Forte*;* Sem *El Niño*/*La Niña*; *La Niña* Fraca*; La Niña* Moderada*; La Niña* Forte
Somente La Niña/El Niño	*La Niña; El Niño;* Sem *El Niño*/*La Niña*;
Mês de ocorrência do desastre	Outubro
Ano de ocorrência do desastre	Entre 1990 e 1994; Entre 1995 e 1999; Entre 2000 e 2004; Entre 2005 e 2009
Estado (UF)	Santa Catarina
Nome da mesorregião	Vale do Itajaí
Nome da microrregião	Blumenau
Nome do município	Blumenau
Munícipio em área de Amazônia	N*(Não);*S*(Sim)*
Munícipio em área de fronteira	N*(Não)*;S*(Sim)*
Município com área suscetível a deslizamento	Suscetível a deslizamentos;Não Suscetível a deslizamentos
Domínio morfológico predominante do município	Embasamentos em Estilos Complexos
Subdomínio morfológico predominante do município	Embasamento do Sul/Sudeste
Unidade de relevo predominante do município	Serras
Localização da unidade de relevo	Leste Catarinense
Tipo de solo predominante do município	Solo Podzólicos
Densidade do município (hab/km²)	Menor que 1; Entre 1 e 5; Entre 5 e 10; Entre 10 e 20; Entre 20 e 50; Entre 50 e 100; Entre 100 e 200; Maior de 200
Altitude do município (metros)	Entre 20 e 99,9m; Entre 100 e 199,9m; Entre 200 e 299,9m;...; Mais de 1000m
Total de chuvas em um mês (50 mm)	Menos de 20mm; Entre 20 e 49,9mm; Entre 50 e 99,9mm; etc...
Máxima de chuvas em um dia (50 mm)	Menos de 20mm; Entre 20 e 49,9mm; Entre 50 e 99,9mm; etc...
Total de chuvas em um mês (100 mm)	Menos de 20mm; Entre 20 e 49,9mm; Entre 50 e 99,9mm; etc...
Máxima de chuvas em um dia (100 mm)	Menos de 20mm; Entre 20 e 49,9mm; Entre 50 e 99,9mm; etc...
Números de dias com chuva em um mês	Menos de 5 dias; De 5 a 10 dias; De 10 a 15 dias; etc

Tabela 9 – Atributos utilizados no processo de mineração de dados com exemplos de valores após o processo de transformação dos dados.

4.1 Execução dos Algoritmos de Agrupamento

As execuções utilizando o algoritmo DBSCAN não resultaram em grupos (*clusters*) que pudessem gerar alguma informação relevante. Foram feitas diversas tentativas alterando os parâmetros do algoritmo (*epsilon* e número mínimo de pontos, *minpoint*), mas os dados estavam todos em um mesmo *cluster* ou em dezenas de *clusters*, dificultando a interpretação, o que acarretava em uma análise imprecisa dos dados. Já com o K-*Means* foi possível encontrar alguns resultados que estão apresentados na sessão 4.1.1.

4.1.1 K-*Means*

Como nem todos os municípios com registros de desastres naturais tem dados pluviométricos (total e máxima de chuvas em um dia e a quantidade de dias em que choveu) foi necessário executar o algoritmo em dois momentos distintos, já que muitos registros com valores nulos para dados pluviométricos influenciaria no resultado da execução do K-*means*.

Primeiramente o algoritmo foi aplicado sobre todos os registros de desastres naturais, não levando em consideração os dados pluviométricos. Posteriormente foram levados em consideração os dados pluviométricos e também foram feitas análises de determinados tipos de desastres naturais.

A seguir serão explicados os resultados tendo como base as microrregiões.

4.1.1.1 Análise por Microrregiões sem Dados Pluviométricos

Para esta primeira análise foram ignorados os atributos *estado(UF)*, *nome da mesorregião* e *nome do município*, já que se pretendia realizar a análise a partir das microrregiões. Com isto, foram considerados os atributos: *tipo do desastre*, *microrregião*, *La Niña/El Niño em intervalos*, *somente La Niña/El Niño*, *município em área de fronteira*, *domínio morfológico*, *unidade do*

relevo (planalto, depressão, planície, etc...), *tipo de solo*, *densidade populacional*, *altitude do município* e *mês de ocorrência do desastre*.

O melhor resultado encontrado foi com a definição de quatro *clusters*. A Figura 23 mostra o padrão encontrado sobre os *clusters* dos desastres de estiagens, por outro lado, na Figura 24, o padrão se refere a desastres de deslizamentos. As duas figuras não possuem relação direta uma com a outra, mas sim dois padrões encontrados em cima dos mesmos quatro *clusters* gerados.

Na Figura 23 é possível visualizar que no *cluster 0*, existe uma relação entre estiagens e o fenômeno *La Niña Fraca*. Segundo o CPTEC (Centro de Previsão de Tempo e Estudos Climáticos), quando se tem este fenômeno a região sul tende a passar por secas e estiagens severas. O valor apresentado, entre parênteses, logo abaixo do número do *cluster* informa o número de registros que foram agrupados no *cluster*.

Na microrregião de Chapecó (*cluster 2*) chegou-se à mesma conclusão dos estudos do CPTEC. Esta tendência maior em agrupar os dados destacando desastres de estiagem ocorre porque a grande maioria dos desastres registrados no Brasil são estiagens ou secas.

Attribute	Cluster 0 (9018)	Cluster 2 (7271)
deslizamento_str	Sucetível a deslizamento	Não sucetível a deslizamento
tipo	Estiagem	Estiagem
nomemicro	Frederico Westphalen	Chapecó
DOM_MORFOE	Bacias Sedimentares Inconsolidadas Pilo-Pleistocenicas	Bacias Sedimentares Inconsolidadas Pilo-Pleistocenicas
UNID_RELEV	Planaltos	Planaltos
SOLOS_TIPO	LATOSOLOS	LATOSOLOS
densidade_intervalo	Entre 20 e 50	Entre 20 e 50
la_nina_el_nino_intervalo	La nina fraca	La nina fraca
la_nina_el_nino_exclusivo	La nina	La nina
alt	Entre 20 e 99,9m	Entre 600 e 699,9m
mes_str	Janeiro	Outubro

Figura 23 – *Clusters* (0 e 2) que relacionam estiagens no sul com a *La Niña*.

Na Figura 24 (gráfico gerado pelo Weka) é possível visualizar os dados agrupados por mês, onde na parte de cima tem-se os tipos de relevo que são suscetíveis a deslizamentos e na parte de baixo os não suscetíveis. Através da mesma figura, verificou-se que regiões de escarpas e reversos tendem a estar em áreas de suscetibilidade a deslizamentos (pontos de cor rosa nos meses de janeiro, fevereiro e março).

Já as regiões de planalto (pontos de cor vermelha no gráfico) podem estar em áreas suscetíveis ou não a deslizamentos, já que aparecem tanto na parte superior do gráfico quanto na

inferior. As depressões (pontos de cor verde) tiveram a maioria dos seus registros feitos em área não suscetíveis a deslizamentos.

Muito embora no eixo *x* esteja apresentado o atributo mês, não foi possível encontrar informações relevantes da predominância de desastres naturais em um determinado mês.

Figura 24 – Relação entre a ocorrência de desastres naturais, relevo e suscetibilidade a deslizamentos.

A Figura 25 mostra que as regiões de Escarpa e Reverso ocupam uma área territorial pequena se comparada ao tamanho do Brasil. Estas áreas aparecerem no extremo norte brasileiro e também na região da Serra do Mar, local onde ocorreu uma das piores tragédias do Brasil, em Petrópolis - Rio de Janeiro. Mesmo ocupando uma área territorial pequena, percebeu-se a relevância em destacá-las devido ao resultado do processo de clusterização.

Figura 25 – Em destaque as regiões de escarpas e reversos no Brasil.

4.1.1.2 Análise por Microrregiões com Dados Pluviométricos

Para que fosse possível realizar a análise considerando os dados pluviométricos, foi necessário remover os desastres naturais que não possuíam tais dados.

Para a análise com os dados pluviométricos foram utilizados os atributos: *tipo do desastre, microrregião, ocorrência de La niña/El niño (fraco, forte, moderado), total de chuvas e máxima de chuvas em intervalo de 50 milímetros (mm), mês do desastre e índice de suscetibilidade a deslizamentos.*

Dos cinco *clusters* criados, os que apresentaram tipos de desastre relacionados com alto índice de chuva também apresentaram relação com o índice de suscetibilidade a deslizamentos (*cluster 2* e *cluster 4*). O *cluster 2* apresentou um total de chuvas no mês entre 250 e 299,9 milímetros e o *cluster 4* entre 300 e 349,9 milímetros.

A Figura 26, que apresenta um gráfico gerado pelo Weka, mostra a dispersão da ocorrência dos desastres dentro dos *clusters* gerados. No eixo *x* tem-se os tipos de desastre e no eixo *y* os *clusters* gerados. Os pontos em vermelho representam desastres que ocorreram em

regiões suscetíveis a deslizamentos e os pontos azuis os desastres que ocorreram em regiões não suscetíveis a deslizamentos.

As microrregiões de Ituporanga (*cluster 4 – destaque na parte superior do gráfico*) e Blumenau (*cluster 2 – destacado no retângulo preto no meio do gráfico*) são as que apresentaram a relação entre o alto índice de chuva e o índice de suscetibilidade a deslizamentos. Já o tipo de desastre que se destacou para os *clusters 2* e *4* foram as enxurradas, onde os círculos em destaque mostram a ligação entre os *clusters* e o tipo de desastre.

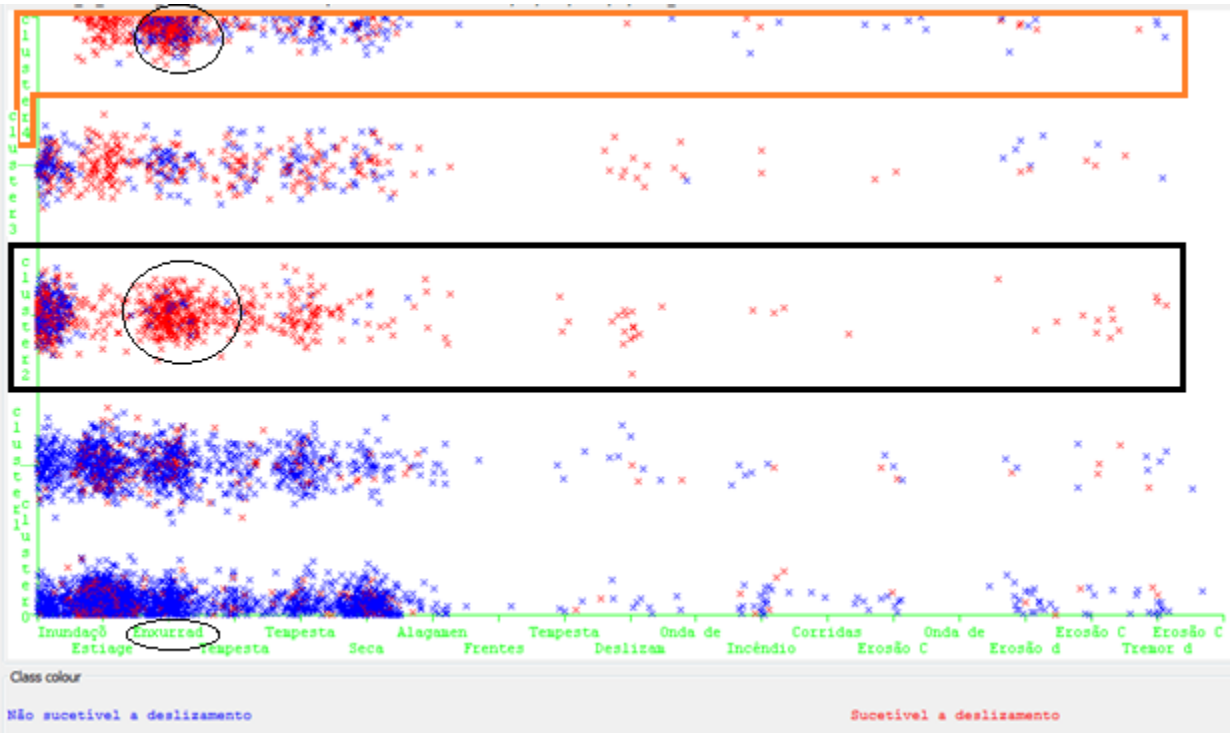


Figura 26 – Relação entre os *clusters (y)*, desastres ocorridos *(x)* e o índice suscetibilidade a deslizamentos *(pontos)*.

Na Figura 27 é possível visualizar a posição geográfica destas duas microrregiões. Ambas se situam no Vale do Itajaí, região conhecida por registrar grandes desastres relacionados com altos índices de chuva.

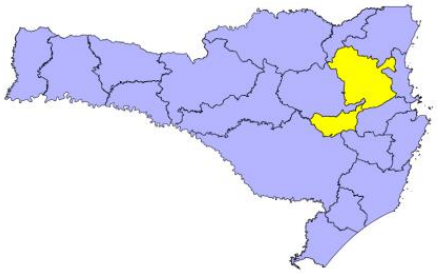


Figura 27 – Em destaque as microrregiões de Ituporanga (mais a baixo) e Blumenau.

4.1.1.3 Análise dos Desastres de Inundações, Enxurradas, Deslizamentos e Alagamentos sem Dados Pluviométricos

A quantidade de registros de desastres naturais exclusivamente dos tipos inundações, enxurradas, deslizamentos e alagamentos; que ocorreram em municípios que não possuíam registros de dados pluviométricos foi grande (apenas 1022 registros, do total de 6.850, possuíam dados pluviométricos). Por isso, primeiramente foi executado do algoritmo K-*means* considerando todos os registros de inundações, enxurradas, deslizamentos e alagamentos, para que se pudesse ter uma análise geral. Com isto, não foram levados em consideração os atributos pluviométricos, já que valores nulos podem influenciar no resultado da mineração.

Nesta análise pode-se chegar a um resultado satisfatório quando se considerou somente os atributos de *densidade populacional, microrregião e ocorrência de La Niña/El Niño (fraco, forte e moderado), altitude e mês de ocorrência*.

Os registros foram divididos em oito *clusters*, e a partir destes atributos, pode-se constatar que em 7 dos 8 *clusters* o atributo *mês* encontra-se na estação de verão. A Figura 28 apresenta os *clusters* gerados a partir dos desastres de inundações, enxurradas, alagamentos e deslizamentos.

A coluna mais à esquerda mostra os nomes dos atributos utilizados. Para o atributo *nomemicro* (nome da microrregião – retângulo em destaque), foi possível observar que houve uma separação dos dados em microrregiões de diversas partes do país (Norte – Baixo Parnaíba Piauiense; Nordeste – Ilhéus-Itabuna; Sul – Araranguá, Joaçaba, Rio do Sul, Paranavai e Curitiba; Sudeste – Rio de Janeiro), mostrando, assim, que o Brasil como um todo tem riscos ter desastres de inundações, deslizamentos, enxurradas ou alagamentos.

Grande maioria do território brasileiro possui uma *densidade populacional* entre 20 e 50 habitantes por quilometro quadrado (hab/km²). O *cluster 3* agrupou os dados pela microrregião do Rio de Janeiro e a *densidade populacional* da microrregião é maior de 150 hab/km² e a área também é suscetível a deslizamento. Todos estes atributos em conjunto fazem um ambiente propicio a ocorrência de desastres relacionados com altos índices pluviométricos. Esta análise é relevante do ponto de vista do estudo de desastres naturais, já que estas características podem acarretar em desastres que desencadeiem elevados danos humanos e materiais.

Attribute	Cluster# 0 (1452)	Cluster# 1 (952)	Cluster# 2 (768)	Cluster# 3 (805)
deslizamento_str	Suscet. a deslizamento	Não suscet. a deslizamento	Suscet. a deslizamento	Suscet. a deslizamento
nomemicro	Rio do Sul	Paranavaí	Curitiba	Rio de Janeiro
densidade_intervalo	Entre 1 e 5	Entre 10 e 20	Entre 20 e 50	Maior de de 150
la_nina_el_nino_intervalo	Sem La nina ou El nino	El nino moderado	El nino fraco	Sem La nina ou El nino
altitude	Entre 200 e 299,9m	Entre 600 e 699,9m	Entre 800 e 999,9m	Entre 20 e 99,9m
mes_str	Fevereiro	Janeiro	Janeiro	Novembro

Attribute	Cluster# 4 (934)	Cluster# 5 (584)	Cluster# 6 (909)	Cluster# 7 (446)
deslizamento_str	Não suscet. a deslizamento	Não suscet. a deslizamento	Não suscet. a deslizamento	Não sucet. a deslizamento
nomemicro	Ilhéus-Itabuna	Araranguá	Joaçaba	Baixo Parnaíba Piauiense
densidade_intervalo	Entre 50 e 100	Entre 100 e 200	Entre 20 e 50	Entre 20 e 50
la_nina_el_nino_intervalo	Sem La nina ou El nino	La nina fraca	Sem La nina ou El nino	La nina fraca
altitude	Entre 400 e 499,9m	Entre 20 e 99,9m	Entre 800 e 999,9m	Entre 20 e 99,9m
mes_str	Dezembro	Março	Maio	Abril

Figura 28 – *Clusters* gerados a partir dos desastres relacionados com altos índices de chuvas.

Com base na análise feita por *microrregião* também foi executado o k-*means* com os mesmos dados, levando em consideração o *estado (UF)* ao invés da *microrregião*. Encontrou-se um padrão entre os estados e os meses do ano, onde a Figura 29 mostra o mapa que relaciona estados e regiões brasileiras e os meses do ano. Este mapa foi criado com o objetivo de facilitar o entendimento dos resultados observados na Figura 30.

No mapa é possível observar que o Norte e Nordeste brasileiro apresenta uma maior tendência a desastres de deslizamentos, inundações, alagamentos e enxurradas nos meses de março, abril e maio. Especificamente o estado da Bahia apresenta esta tendência no mês de dezembro. Já os estados de São Paulo e Paraná tiveram como padrão o mês de janeiro e o Mato

Grosso do Sul apresentou o padrão para o mês de março. A região sul (Santa Catarina, Paraná e Rio Grande do Sul), muito embora não tivesse um mês definido, também apresentou uma grande quantidade de registros (25% de todos os dados agrupados), fazendo com que ganhasse destaque. O parágrafo a seguir mostra de onde foram obtidas estas informações para a criação do mapa.

Figura 29 – Mapa com as distribuições dos desastres de acordo com o mês.

A Figura 30, mostra, para o eixo *x*, os oito *clusters* criados e no eixo *y* os estados brasileiros. Já os pontos do gráfico são os registros de desastres que ocorreram em um determinado mês.

A partir desta figura foi possível concluir que o estado de São Paulo e do Paraná tem uma tendência a ter desastres de inundações, deslizamentos, enxurradas ou alagamentos no mês de janeiro (pontos em cor preta dentro do círculo destacado – *cluster 2*). Na Bahia, a ocorrência se deu em dezembro (pontos em cor cinza dentro do losango destacado – *cluster 4*), no Mato Grosso do Sul ocorreu em março (pontos em vermelho dentro do triângulo destacado – *cluster 5*) e, por fim, na região Norte e Nordeste que além de março (pontos em vermelho dentro do retângulo – *cluster 5*, *6* e *7*) teve ocorrências em abril (pontos em azul dentro do retângulo) e maio (pontos em cor bege dentro do retângulo).

A região sul (Santa Catarina, Paraná e Rio Grande do Sul) não teve um mês específico, entretanto esta é uma região que representa 25% dos clusters, demonstrando a ocorrência de

grande quantidade de registros de desastres relacionados com altos índices de chuva, já que foram utilizados somente os registros de desastres de inundações, deslizamentos, alagamentos e enxurradas. Esta informação é possível observar no *cluster 6* onde existe uma grande quantidade de pontos, mas sem uma cor definida.

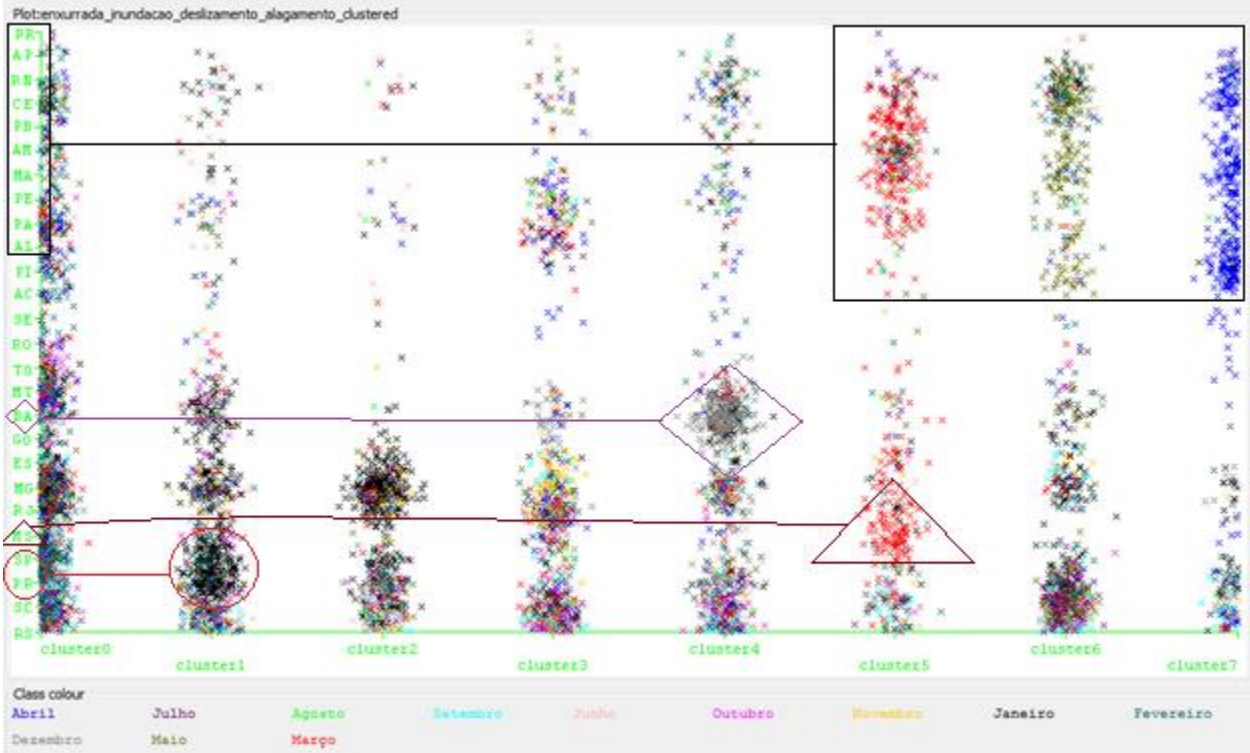


Figura 30 – Relação entre os *clusters (x)*, estados *(x)* e a ocorrência do desastre no mês *(pontos)*.

4.1.1.4 Análise dos Desastres de Inundações, Enxurradas, Deslizamentos e Alagamentos com Dados Pluviométricos.

Na análise dos desastres de inundações, enxurradas, deslizamentos e alagamentos, considerando os dados pluviométricos (*total de chuva no mês*, *máxima de chuva em um dia* e *número de dias em que choveu no mês*) foram utilizados os atributos: *microrregião, ocorrência de La Niña/El Niño (fraco, forte, moderado), índice de suscetibilidade a deslizamentos, mês de ocorrência do desastre*, além dos atributos pluviométricos já citados. Neste caso foi removido o atributo *densidade populacional*, porque sem este atributo foi possível ter *clusters* que gerassem

informações mais relevantes. Os atributos de relevo foram removidos, pois pretendia-se realizar apenas a relação entre as microrregiões e os atributos pluviométricos.

A melhor distribuição dos dados nos *clusters* aconteceu quando se executou o algoritmo K-*Means* considerando cinco *clusters*.

Attribute	Cluster# 0 (353)	Cluster# 1 (192)	Cluster# 2 (218)
nomemicro	Santa Maria	Santarém	Litoral de Camocim e Acaraú
la_nina_el_nino_intervalo	Sem La nina ou El nino	La nina fraca	Sem La nina ou El nino
numdiasdechuvaIntervalo	De 5 a 10 dias	De 25 a 31 dias	De 15 a 20 dias
total_int_50	Entre 300 e 349,9mm	Entre 250 e 299,9mm	Entre 250 e 299,9mm
maxima_int_50	Entre 50 e 99,9mm	Entre 50 e 99,9mm	Entre 50 e 99,9mm
mes_str	Abril	Abril	Maio
deslizamento_str	Não sucetível a deslizamento	Sucetível a deslizamento	Não sucetível a deslizamento

Attribute	Cluster# 3 (110)	Cluster# 4 (149)
nomemicro	Blumenau	Canoinhas
la_nina_el_nino_intervalo	Sem La nina ou El nino	Sem La nina ou El nino
numdiasdechuvaIntervalo	De 20 a 25 dias	De 10 a 15 dias
total_int_50	Entre 200 e 249,9mm	Entre 200 e 249,9mm
maxima_int_50	Entre 100 e 149,9mm	Entre 20 e 49,9mm
mes_str	Maio	Fevereiro
deslizamento_str	Não sucetível a deslizamento	Não sucetível a deslizamento

Figura 31 – *Clusters* gerados considerando atributos pluviométricos.

Analisando os *clusters* apresentados na Figura 31 é possível perceber que os níveis de chuva variaram de 200mm por mês chegando até a 349,9mm. No *cluster 3* a máxima de chuvas registrada em um dia foi entre 100 e 149,9mm, sendo uma quantidade alta se comparada à outros *clusters*. Na maioria dos *clusters* (3/5) a máxima apresentada foi entre 50 e 99,9mm. Já o número de dias com chuva no mês variou de 5 até 25 dias e em quatro dos cinco *clusters* gerados foram em área não suscetível a deslizamento.

Destaque para o *cluster 1*, onde é possível observar que a microrregião de Santarém, pertencente ao estado do Pará, apresentou um número muito elevado de chuvas distribuído durante todo o mês (de 250 e 299,9mm e entre 25 e 31 dias), sendo uma região suscetível a deslizamentos.

Somente para este *cluster* é possível inferir uma relação entre a grande quantidade de chuvas e a incidência de *La Niña fraca*. Segundo o CPTEC quando ocorre o fenômeno da *La Niña*, a região norte do Brasil tende a ter um aumento da precipitação de chuvas e vazão dos rios,

o que pode acarretar em uma maior incidência nos registros de desastres de alagamentos, inundações e enxurradas. Este predomínio acontece no mês de abril.

4.1.1.5 Análise dos Desastres de Secas e Estiagem sem Dados Pluviométricos

Os desastres de secas e estiagem foram os dois tipos de desastre mais registrados no Brasil. Segundo a Defesa Civil, meteorologicamente, a diferença entre seca e estiagem é que a seca é uma estiagem prolongada.

Como a quantidade de registros de secas e estiagem também ocorreram em municípios que não possuíam registros de dados pluviométricos (apenas 3220 registros tinham dados pluviométricos, do total de 19760), assim como para os desastres de enxurradas, inundações, alagamentos e deslizamentos, primeiramente, foi executado o algoritmo considerando todos os registros de secas e estiagens. Com isto, não foram levados em consideração os atributos pluviométricos, já que valores nulos podem influenciar no resultado da mineração.

Quando feita a análise considerando os atributos *estado (UF)*, *densidade*, *La Niña*/*El Niño (fraco, forte, moderado)*, *altitude* e *mês* foram gerados cinco *clusters* e quando feita a análise trocando o atributo *estado (UF)* por *microrregião* foram gerados 10 *clusters*. Os atributos relacionados com relevo foram removidos, pois quando se tentou gerar *clusters* incluindo-os não foi possível chegar a *clusters* que trouxessem informações relevantes.

Para a análise feita com o atributo *estado*, foi possível observar uma relação do fenômeno *La Niña*/*El Niño* com a ocorrência de secas. Segundo o CPTEC quando em tempos de *El Niño*, na região Nordeste costumam ocorrer mais secas ou estiagens. Já na época de *La Niña*, a ocorrência destes mesmos desastres fica mais acentuada em estados sulistas.

De acordo com a Figura 32 é possível perceber que nos *clusters 3* e *4* existe uma maior influência do *El Niño* (pontos verdes no gráfico) nos estados do Nordeste (principalmente Sergipe, Piauí, Alagoas e Bahia) destacados através dos dois círculos interligados, o que vai ao encontro dos estudos anteriormente apresentados pelo CPTEC.

No *cluster 1* (retângulo no gráfico da Figura 32) ocorreu um agrupamento dos desastres que tem uma relação maior com a *La Niña* (pontos vermelhos). Entretanto, não existe um padrão deste fenômeno quando analisado por estado, já que existe uma certa mistura com registros de

desastres que ocorreram sem a influência de *La Niña* ou *El Niño* (pontos azuis no mapa) e *El Niño*, apesar de no Rio Grande do Sul, Santa Catarina e Paraná existirem mais pontos vermelhos (*La Niña*) que os outros.

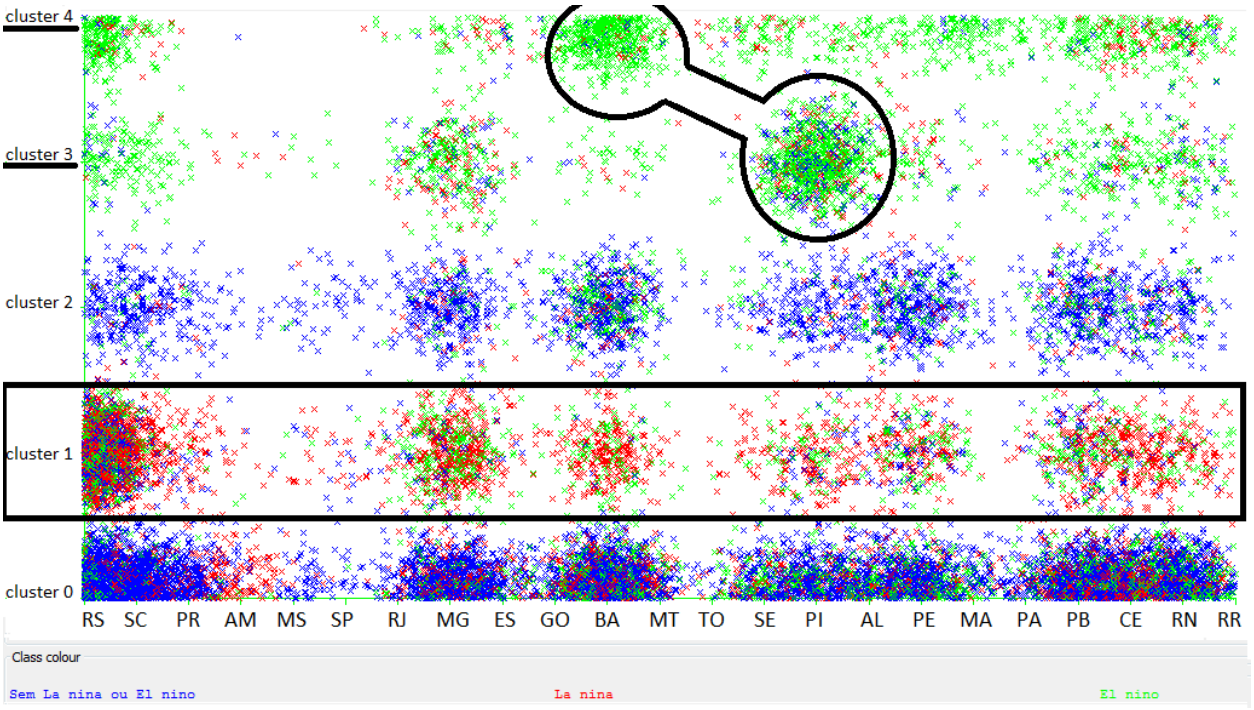


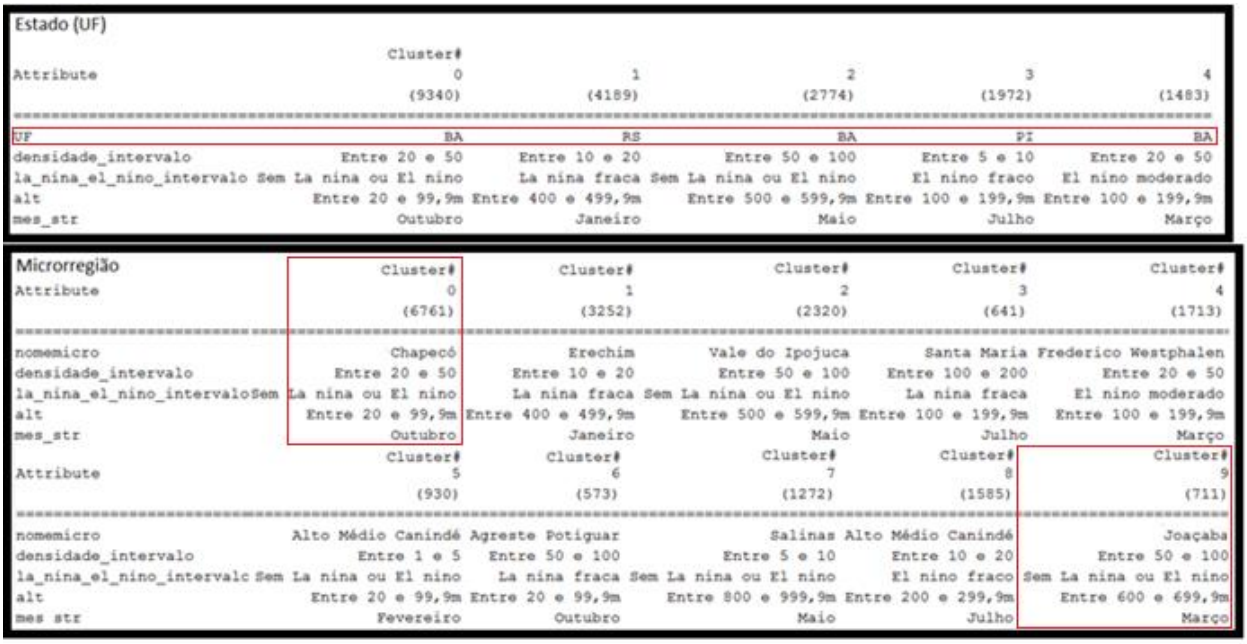
Figura 32 – Relação entre os *clusters (y)*, estados *(x)* e o fenômeno *El Niño*/*La Niña (pontos)*.

A Figura 33 mostra os *clusters* gerados através da utilização do atributo *estado (UF)* (retângulo superior) e do atributo *microrregião* (retângulo de baixo). Para o caso dos *clusters* gerados a partir das microrregiões, ocorreu o aparecimento das microrregiões catarinenses de Chapecó e Joaçaba (*cluster 0* e *cluster 9*, do retângulo de baixo), mesmo o estado sendo só o sexto que mais registra estiagens e secas.

Quando executado o algoritmo por estado, Santa Catarina não aparece nos *clusters* gerados (atributo *UF* do primeiro retângulo).

A partir disto é possível concluir que, de maneira geral, o estado de Santa Catarina não se destaca na incidência de estiagens ou seca, entretanto as microrregiões de Chapecó e Joaçaba são

dois pontos específicos do estado que merecem atenção quanto à incidência de estiagens ou secas.



Estado (UF)

Attribute	Cluster# 0 (9340)	1 (4189)	2 (2774)	3 (1972)	4 (1483)
UF	BA	RS	BA	PI	BA
densidade_intervalo	Entre 20 e 50	Entre 10 e 20	Entre 50 e 100	Entre 5 e 10	Entre 20 e 50
la_nina_el_nino_intervalo	Sem La nina ou El nino	La nina fraca	Sem La nina ou El nino	El nino fraco	El nino moderado
alt	Entre 20 e 99,9m	Entre 400 e 499,9m	Entre 500 e 599,9m	Entre 100 e 199,9m	Entre 100 e 199,9m
mes_str	Outubro	Janeiro	Maio	Julho	Março

Microrregião

Attribute	Cluster# 0 (6761)	Cluster# 1 (3252)	Cluster# 2 (2320)	Cluster# 3 (641)	Cluster# 4 (1713)
nomemicro	Chapecó	Erechim	Vale do Ipojuca	Santa Maria	Frederico Westphalen
densidade_intervalo	Entre 20 e 50	Entre 10 e 20	Entre 50 e 100	Entre 100 e 200	Entre 20 e 50
la_nina_el_nino_intervalo	Sem La nina ou El nino	La nina fraca	Sem La nina ou El nino	La nina fraca	El nino moderado
alt	Entre 20 e 99,9m	Entre 400 e 499,9m	Entre 500 e 599,9m	Entre 100 e 199,9m	Entre 100 e 199,9m
mes_str	Outubro	Janeiro	Maio	Julho	Março

Attribute	Cluster# 5 (930)	Cluster# 6 (573)	Cluster# 7 (1272)	Cluster# 8 (1585)	Cluster# 9 (711)
nomemicro	Alto Médio Canindé	Agreste Potiguar	Salinas	Alto Médio Canindé	Joaçaba
densidade_intervalo	Entre 1 e 5	Entre 50 e 100	Entre 5 e 10	Entre 10 e 20	Entre 50 e 100
la_nina_el_nino_intervalo	Sem La nina ou El nino	La nina fraca	Sem La nina ou El nino	El nino fraco	Sem La nina ou El nino
alt	Entre 20 e 99,9m	Entre 20 e 99,9m	Entre 800 e 999,9m	Entre 200 e 299,9m	Entre 600 e 699,9m
mes_str	Fevereiro	Outubro	Maio	Julho	Março

Figura 33 – *Clusters* gerados a partir dos registros de secas e estiagem.

4.1.1.6 Análise dos Desastres de Secas e Estiagem com Dados Pluviométricos

Na Figura 34 é possível observar que seis dos sete *clusters* registraram menos de 10 dias de chuva e como já era de se esperar, os índices de chuva também foram baixos (seis dos setes registraram menos de 100mm de chuva em um mês – *atributo total_int_50 na Figura 34*). Com relação aos meses, seis dos setes *clusters* tiveram os dados agrupados no primeiro semestre.

O *cluster 3* agrupou os dados pela microrregião de São Miguel do Oeste - SC onde também ocorreu a incidência de *La Niña*. Segundo o CPTEC a região costuma apresentar seca ou estiagem quando ocorre este fenômeno. No *cluster 1*, a maioria dos registros de desastres naturais do *cluster* ocorreram em Montes Claros, Minas Gerais, e este *cluster* apresentou uma predominância de registros de *El Niño*, entretanto segundo o CPTEC a região sudeste não apresenta um padrão característico de mudanças das chuvas para o *El Niño*.

Attribute	Cluster# 0 (602)	Cluster# 1 (411)	Cluster# 2 (443)	Cluster# 3 (232)
nomemicro	Montes Claros	Montes Claros	São Miguel do Oeste	São Miguel do Oeste
densidade_intervalo	Entre 10 e 20	Entre 50 e 100	Entre 20 e 50	Entre 20 e 50
la_nina_el_nino_intervalo	Sem La nina ou El nino	El nino fraco	Sem La nina ou El nino	La nina fraca
numdiasdechuvaIntervalo	De 5 a 10 dias	De 5 a 10 dias	De 5 a 10 dias	De 10 a 15 dias
total_int_50	Entre 50 e 99,9mm	Entre 50 e 99,9mm	Entre 20 e 49,9mm	Entre 200 e 249,9mm
maxima_int_50	Entre 50 e 99,9mm	Entre 20 e 49,9mm	Entre 20 e 49,9mm	Entre 50 e 99,9mm
mes_str	Março	Junho	Fevereiro	Janeiro

Attribute	Cluster# 4 (837)	Cluster# 5 (481)	Cluster# 6 (212)
nomemicro	Montes Claros	Salinas	Almenara
densidade_intervalo	Entre 20 e 50	Entre 20 e 50	Entre 10 e 20
la_nina_el_nino_intervalo	El nino fraco	Sem La nina ou El nino	Sem La nina ou El nino
numdiasdechuvaIntervalo	Menos de 5 dias	Menos de 5 dias	De 5 a 10 dias
total_int_50	Menos de 20mm	Menos de 20mm	Entre 20 e 49,9mm
maxima_int_50	Menos de 20mm	Menos de 20mm	Menos de 20mm
mes_str	Agosto	Maio	Maio

Figura 34 – *Clusters* gerados a partir dos registros de secas e estiagem.

4.1.1.7 Análise dos Desastres de Granizo

A análise dos desastres de granizo foi realizada sem considerar atributos pluviométricos, visto que a quantidade de registros que continham estes atributos era muito pequena para aplicar técnicas de mineração de dados (somente 333 registros de granizo em 20 anos de registros).

A partir dos dados de desastres de granizo foi possível gerar dez *clusters*. Com estes *clusters* foi observado - considerando somente os atributos de *ocorrência de La Niña/El Niño*, *altitude*, *mês de ocorrência do desastre natural* e *microrregião* - que os registros foram agrupados pelos meses de *setembro*, *outubro* e *novembro* (7 dos 10 *clusters*). Outra informação que foi possível extrair é que foram gerados *clusters* somente de microrregiões da região sul, mostrando que esta região está bastante propensa a registros de desastres gerados por granizo.

Fazendo a análise do *mês de ocorrência* e *estado (UF)*, foi possível observar a partir do destaque dado ao *cluster 2*, na Figura 35, que quando ocorre granizo fora da região sul, existe uma preponderância maior de ocorrer no mês de setembro (pontos rosas no gráfico) no estado de Minas Gerais.

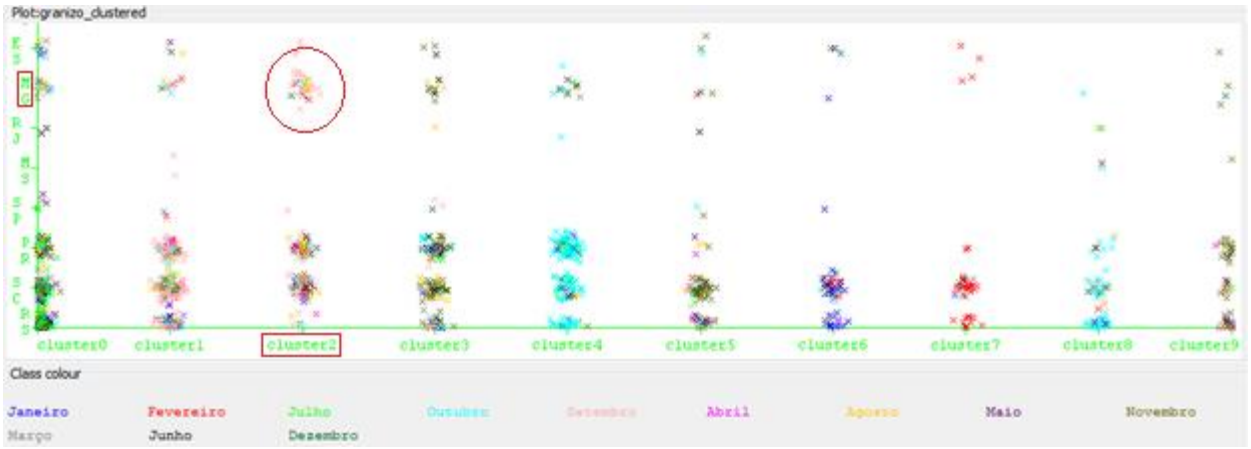


Figura 35 – Relação entre os *clusters (x)*, estados *(y)* e meses *(pontos)*.

4.1.1.8 Análise dos Desastres de Vendaval

Para vendavais também foi realizada a análise não considerando atributos pluviométricos, visto que a quantidade de registros era apenas de 534 para um total de 2.609 desastres de vendavais.

Foram considerados 5 *clusters*, com os atributos *ocorrência de La Niña/El Niño*, *altitude*, *mês de ocorrência do desastre* e a *microrregião*). A maior incidência de desastres desse tipo foi na região sul, muito embora ocorreu um agrupamento com o estado de Minas Gerais, assim como nos registros de desastres de granizo. Quando analisada a Figura 36 é possível observar que os registros de vendavais que tiveram uma maior incidência nos meses de *setembro*, *outubro*, *novembro*, *janeiro* e *fevereiro* (retângulos vermelhos) ocorreram sob influência do fenômeno *El Niño* (pontos azuis) ou *La Niña* (pontos verdes).

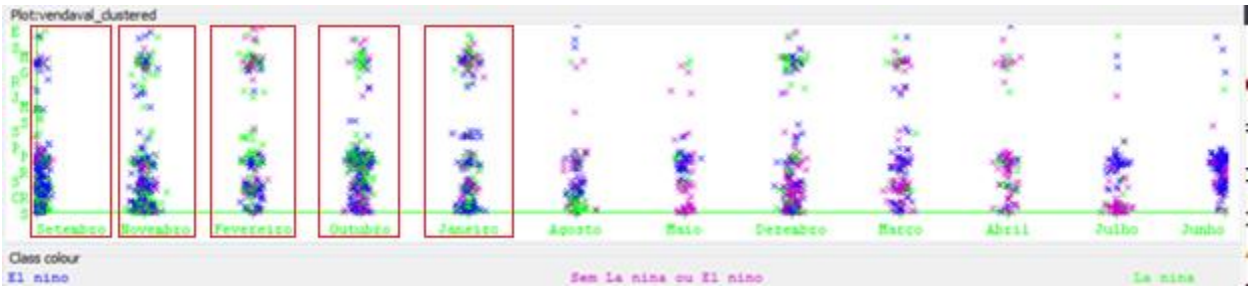


Figura 36 – Relação entre os *meses (x)*, estados *(y)* e ocorrência de *La Niña*/*El Niño (pontos)*.

4.1.2 Resumo dos Resultados de Clusterização

A partir da utilização de algoritmos de clusterização sobre os registros de desastres naturais foi possível confirmar estudos do CPTEC sobre a relação entre o fenômeno *La Niña* e *El Niño* e as regiões brasileiras. Com base nos dados, confirmou-se que desastres de estiagem e secas tem relação com a região sul, quando sob efeito do *La Niña*.

Sobre o relevo existente no Brasil, conclui-se que as regiões de depressão são áreas normalmente não suscetíveis a deslizamentos e que em regiões de escarpas e reversos existe a suscetibilidade a deslizamento.

Outro ponto importante concluído neste trabalho foi que existe uma diferença quando analisados os registros dos estados e das microrregiões. Microrregiões como a de Chapecó, em Santa Catarina, possui uma quantidade grande de registros de estiagem e secas, quando o estado de Santa Catarina é apenas o sexto estado que mais possui registros de secas e estiagem. Isto refletiu diretamente no processo de clusterização, pois foram gerados *clusters* agrupados com microrregiões de Santa Catarina, mas não foram gerados *clusters* com o estado.

Do ponto de vista dos meses do ano verificou-se que o verão é uma estação com uma tendência maior a ocorrer desastres de enxurradas, deslizamentos, alagamentos ou deslizamentos. Também foi possível fazer uma relação de alguns estados ou regiões brasileiras e os meses do ano para estes mesmos tipos de desastres. A Tabela 10 mostra quais foram os padrões concluídos, onde na coluna da esquerda é apresentado o estado ou região e na direita o mês.

Estado ou Região	**Mês Padronizado**
São Paulo e Paraná	Janeiro
Mato Grosso do Sul	Março
Norte e Nordeste	Março, Abril e Maio
Bahia	Dezembro

Tabela 10 – Relação entre os estados ou regiões e os meses para os desastres de enxurradas, deslizamentos, alagamentos ou inundações.

Por último, os desastres de granizos e vendavais tiveram resultados similares. A região Sul é a que mais registra estes desastres, sendo que também é possível encontrar um certo padrão da ocorrência de ambos no estado de Minas Gerais.

4.2 Análise dos Resultados dos Algoritmos de Classificação

Os algoritmos utilizados para fazer a classificação foram o ID3 e o C4.5. Foram analisados registros de enxurradas, inundações, deslizamentos, alagamentos, secas e estiagens, com o intuito de encontrar uma relação entre os índices pluviométricos e estes desastres.

Como se desejou encontrar padrões para desastres naturais que mais afetam vidas humanas, utilizou-se como atributo classe o *tipo de desastre* com os valores enxurradas, inundações, deslizamentos, alagamentos, secas e estiagens.

Também foi criado outro atributo exclusivamente para os algoritmos de classificação (*tem tendência a enxurrada*). O objetivo do atributo é encontrar padrões que caracterizem a ocorrência de enxurrada. Para isto foi feita uma comparação entre os registros de enxurradas e inundação (SIM para registro de enxurrada e NÃO para registros de inundação), visto que são os dois tipos que tiveram mais registros relacionados com altos índices de chuva (1.568 e 944 registros, respectivamente). Para que não houvesse uma influência maior dos desastres de enxurradas, foram considerados somente 944 registros de enxurradas, mesma quantidade de registros de inundações.

Foram efetuados diversos experimentos com os algoritmos de classificação, entretanto em geral, os resultados não atingiram o nível de qualidade que se esperava. O ideal era também ter trabalhado com dados pluviométricos onde não ocorreram registros de desastres naturais para que se pudesse fazer um comparativo entre os valores dos dados em momentos em que ocorreram desastres naturais e momentos em que não ocorreram.

Foi possível observar é que para alguns casos de registros de estiagens e seca os total de chuvas em um mês ultrapassava a quantidade de 150 milímetros. Foi o caso do município de Caruaru, em Pernambuco, onde teve um registro de 142 mm em junho de 2001, mas também apresentou um registro de estiagem para este mês; e também do município de Medina, em Minas Gerais, onde este teve um registro de 336 mm de chuvas no mês de março de 2004 e ao mesmo

tempo apresentou um registro de seca. Isto fez com que os modelos de classificação fossem prejudicados, devido a esta aparente divergência de dados, já que para esta quantidade de chuva não deveriam existir registros de desastres naturais de estiagem ou seca. Uma possível explicação para este problema é que pode ter havido seca ou estiagem em uma parte no município e ter chovido muito em outra parte (local onde está a estação pluviométrica) e como a granularidade de informação é o município, esse tipo de problema pode ocorrer.

Diante desta conclusão, o algoritmo que apresentou resultado mais satisfatório, foi o C4.5. A seção a seguir descreve os resultados obtidos a partir dos atributos chave *tipo de desastre* e *tem tendência a enxurrada*.

4.2.1 C4.5

O algoritmo C4.5 apresentou resultados melhores quando foram utilizados atributos numéricos em vez de descritivos. As duas árvores de classificação geradas a partir das diversas execuções realizadas com o algoritmo C4.5 foram testadas através da validação cruzada, dividindo os dados em 10 partes (*folds*). Para este algoritmo também foi utilizada a poda da árvore gerada, que é uma técnica do algoritmo que permite gerar uma árvore menor e mais compreensível.

Tanto para a execução utilizando com o atributo chave o *tipo de desastre* e o atributo criado *tem_enxurrada*, utilizou-se os atributos: *índice de suscetibilidade a deslizamento*, *unidade do relevo*, *densidade populacional*, *temperatura do aumento das águas do oceano Pacífico* (*El Niño*/*La Niña*), *altitude*, *mês*, *total de chuvas no mês*, *máxima de chuva em um dia* e *quantidade de dias que choveram no mês*.

A utilização do atributo chave *tipo de desastre* gerou um modelo que classificou os registros corretamente em aproximadamente 64% dos casos. Este valor pode ser considerado satisfatório, já que foram analisados quatro tipos de desastres, isto é, a probabilidade de acerto na predição do desastre é de uma em quatro, ou seja, 25%.

A Tabela 11 mostra a matriz de confusão gerada pela execução, mostrando como os registros foram classificados utilizando a árvore gerada. Na primeira coluna é apresentado o tipo

de desastre que se espera ser classificado e na primeira linha qual foi a classificação realmente feita.

Os valores escritos em verde mostram a quantidade de registros que foram corretamente classificados e os escritos em vermelho a quantidade que foi erroneamente classificada. Na primeira linha é possível observar que praticamente metade dos registros de inundações foram registrados como sendo de enxurradas (463 de 944). Já na segunda linha foi possível observar um acerto um pouco melhor para o caso de enxurradas (1.213 de 1.568 registros) e as duas últimas não apresentaram nenhum registro corretamente classificado.

	Inundações	Enxurradas	Alagamentos	Deslizamentos
Inundações	**463**	481	0	0
Enxurradas	355	**1213**	0	0
Alagamentos	18	28	**0**	0
Deslizamentos	7	33	0	**0**

Tabela 11 – Matriz de confusão gerada a partir do atributo chave *tipo de desastre.*

Analisando os resultados da execução do C4.5, algumas informações sobre os desastres que podem ser extraídas é que regiões de escarpa e reversos tem uma probabilidade maior de ocorrer desastres de enxurradas.

Já para o atributo chave criado para verificar se *tem tendência a enxurrada* o modelo gerado classificou corretamente os registros em 74% dos casos. O resultado não foi tão satisfatório, já que o atributo é binário (sim ou não) e a probabilidade de acerto é de 50%. No entanto, já era esperado que a porcentagem fosse um pouco mais baixa, já que foi utilizada a técnica de poda da árvore, o que diminui a acurácia do resultado de quase 80% (antes da poda) para 74% (depois da poda).

A Tabela 12 mostra a matriz de confusão gerada a partir do atributo chave *tem tendência à enxurrada*. Nela é possível observar a porcentagem do resultado da classificação, utilizando a árvore gerada pelo algoritmo C4.5, em números. Os valores escritos em verde e em negrito (645 e 756) mostram a quantidade de registros que foram corretamente classificados, e os escritos em vermelho e em negrito (299 e 189) a quantidade que foi erroneamente classificada.

tem tendência a enxurrada	SIM	NÃO
SIM	**645**	299
NÃO	189	**756**

Tabela 12 – Matriz de confusão gerada a partir do atributo chave *tem tendência à enxurrada.*

Mesmo com este resultado pode-se concluir a partir das análise feitas sobre a execução do algoritmo com o atributo chave *tem tendência a enxurrada*, que regiões de chapadas, planaltos, cristais, colinas e tabuleiros não costumam apresentar enxurradas. Por outro lado, regiões de escarpas e reversos podem apresentar enxurradas se a quantidade total de chuvas em um mês passa de 147 milímetros e regiões de serras podem apresentar desastres de enxurradas quando chove mais de 16 dias em um mês.

5. CONCLUSÃO E TRABALHOS FUTUROS

Devido ao aumento da ocorrência de desastres naturais e do grau de complexidade, a análise dos registros de desastres tornou-se uma necessidade para que se possa conhecer melhor o histórico brasileiro de desastres.

Este trabalho teve como objetivo realizar a coleta de dados e estudar os desastres naturais brasileiros através do processo de mineração de dados para a descoberta de padrões entre os desastres.

A partir da coleta de dados como: registros dos desastres naturais ocorridos no Brasil (AVADAN), intensidade do *El Niño* ou *La Niña*, índice de suscetibilidade a deslizamento de um determinado município, relevo e solo predominante dos municípios e índices pluviométricos; foi feito o entendimento dos dados, levantando informações relevantes que pudessem auxiliar no processo de *data mining*, como a identificação dos tipos de desastres naturais que mais ocorrem em uma determinada microrregião brasileira. Em seguida foi realizada a preparação dos dados para que pudessem ser utilizados os algoritmos de agrupamento (DBSCAN e K-means) e classificação (ID3 e C4.5), esta etapa foi a que mais precisou de tempo e dedicação para a elaboração deste trabalho.

Especificamente a preparação dos dados pluviométricos (máxima de chuvas em um dia, total de chuvas em um mês e quantidade de dias com chuva no mês) foi a que gerou o maior trabalho, pois os dados estavam distribuídos, dentro dos arquivos, de modo que dificultava a união entre os registros de desastre e os dados pluviométricos. Além disto, estes dados também estavam em mais de um arquivo (um para cada estado brasileiro), o que gerou a necessidade de dar uma atenção maior.

Com a seleção, limpeza, e transformação dos dados foi possível executar os algoritmos escolhidos e com o resultado das diversas execuções feitas, pôde-se concluir que existe uma relação entre os registros de desastres naturais e a precipitação de chuvas. Gerou-se grupos (*clusters*) somente com desastres relacionados com altos índices de chuva que variaram de 200mm por mês chegando até a 349,9mm, já a quantidade de chuvas em um dia ficou entre 50mm e 99,9mm.

Por outro lado, desastres relacionados com baixos índices pluviométricos costumam apresentar grupos de no máximo 50mm de chuvas em um mês e em alguns grupos foi possível observar que a máxima em um dia também se aproximou de 50mm.

Além destas conclusões, também se obteve outras que não faziam parte dos objetivos do trabalho, mas que são importantes do ponto de vista da análise dos registros de desastres naturais. São elas:

- Áreas com suscetibilidade a deslizamentos tendem a estar em regiões de escarpas e reversos. Regiões de planalto podem ser tanto suscetíveis quanto não suscetíveis a deslizamentos, e depressões tendem a não ser suscetíveis.

- Existe uma relação de alguns *clusters* criados com a incidência do fenômeno *La Niña* e *El Niño.*

- Desastres relacionados com altos índices de chuva tendem a ocorrer com maior frequência no Norte/Nordeste no mês de abril. Exclusivamente a Bahia costuma apresentar estes tipos de desastre em dezembro. Paraná e São Paulo apresentam desastres de inundações, enxurradas, deslizamentos e alagamentos em janeiro. Este último estado também pode apresentar em fevereiro e o Mato Grosso do Sul em março.

Como nem todos os conjuntos de dados revelaram algum padrão, seria possível obter resultados mais relevantes para uma quantidade maior de registros de desastres naturais, aumentando a precisão dos resultados, bem como mais dados de domínios diferentes (vegetação, clima, agropecuária etc...) referentes aos municípios, microrregiões e estados, aumentando a gama de informações.

Além de trabalhar com outros domínios de dados também se mostrou importante trabalhar com índices, como o índice de suscetibilidade a deslizamentos, criados a partir de estudos realizados anteriormente. Outro ponto a destacar é que existiu muita perda de registros com índices pluviométricos, pois diversos registros ocorreram em municípios que não tiveram dados pluviométricos no ano e mês do desastre.

Para que fosse possível fazer um modelo preditivo mais efetivo para desastres de deslizamentos, inundações e enxurradas, mais informações são necessárias, como por exemplo, a

quantidade de chuva nas últimas vinte e quatro horas, semana e quinzena antes de dos desastres. Além de também ter o número de dias de chuva nestes intervalos (dia, semana e quinzena).

6. REFERÊNCIAS

[1] – TAN, Pang-ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao Data Mining.** Rio de Janeiro: Ciência Moderna, 2009. 900 p.

[2] – LIMA, Fabiana Santos; OLIVEIRA, Daniel de; GONÇALVES, Mirian Buss. **A FORMAÇÃO DE CLUSTERS NA LOGÍSTICA HUMANITÁRIA UTILIZANDO MINERAÇÃO DE DADOS.** Florianópolis, 2011. 12 p.

[3] – ESTÉBANEZ, Katiusca Magdalena Briones. **Identificação de Padrões na Ocorrência de Emergências e Desastres Associados a níveis de Precipitação.** 2012. 88 f. Dissertação (Mestrado) – Curso de Engenharia Civil, Departamento de Coppe, Ufrj, Rio de Janeiro, 2012.

[4] – SOUZA, Fábio Teodoro de. **PREDIÇÃO DE ESCORREGAMENTOS DAS ENCOSTAS DO MUNICÍPIO DO RIO DE JANEIRO ATRAVÉS DE TÉCNICAS DE MINERAÇÃO DE DADOS.** 2004. 115 f. Tese (Doutorado) – Curso de Engenharia Civil, Departamento de Coppe, Ufrj, Rio de Janeiro, 2004.

[5] – CHAGAS, Diego José; CHAN, Chou Sin; CORSI, Alessandra Cristina. **Análise do Banco de Atendimentos da Defesa Civil do Estado de São Paulo.** São Paulo: Xwq, 2010. 9 p.

[6] – WIKIPÉDIA. **São Paulo.** Disponível em: <http://pt.wikipedia.org/wiki/Sao_Paulo>. Acesso em: 22 nov. 2012.

[7] – WIKIPÉDIA. **Equador.** Disponível em: <http://pt.wikipedia.org/wiki/Equador>. Acesso em: 22 nov. 2012.

[8] – GOVERNO Federal. **CONSTITUIÇÃO DA REPÚBLICA FEDERATIVA DO BRASIL DE 1988.** Disponível em: <http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm>. Acesso em: 22 nov. 2012.

[9] – V SEMINÁRIO INTERNACIONAL DE DEFESA CIVIL – DEFENCIL, 2009, São Paulo. **Os desastres naturais, a cultura de segurança e a gestão de desastres no Brasil.** São Paulo: Meio Eletrônico, 2009. 8 p.

[10] – CHAPMAN, Pete et al. **CRISP-DM 1.0:** Step-by-step data mining guide. Eua, Alemanha, Dinamarca, Holanda: Spss, 2000.

[11] – ADRIAANS, P., and D. Zantige. 1996. **Data mining**. Harlow, England

[12] – ALVARES, Luis Otavio. **O processo de KDD.** Disponível em: <http://www.inf.ufsc.br/%7Ealvares/INE5644/processo_DCBD.ppt>. Acesso em: 24 nov. 2012.

[13] – ALVARES, Luis Otavio. **Agrupamento K-means.** Disponível em: <http://www.inf.ufsc.br/%7Ealvares/INE5644/clustering1.ppt>. Acesso em: 26 nov. 2012.

[13] – TAN, Pang-ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao Data Mining.** Rio de Janeiro: Ciência Moderna, 2009. 900 p.

[14] – XIV SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 14., 2009, Natal – Rn. **Relação entre os desastres naturais e as anomalias de precipitação para a região Sul do Brasil.** Santa Maria – Rs: Inpe, 2009. 8 p. Disponível em: <http://www.inpe.br/crs/geodesastres/conteudo/publicacoes/3527_3534_BARBIERI_Relacao_entre_desastres_naturais_2009.pdf>. Acesso em: 01 dez. 2012.

[15] - NULL, Jan. **El Niño and La Niña Years and Intensities.** Disponível em: <http://ggweather.com/enso/oni.htm>. Acesso em: 01 dez. 2012.

[16] – BRASIL, Serviço de Geologia do. **SELEÇÃO DE MUNICÍPIOS CRÍTICOS A DESLIZAMENTOS:** NOTA EXPLICATIVA. Disponível em: <http://www.cprm.gov.br/publique/media/apresentacao_susc.pdf>. Acesso em: 01 dez. 2012.

[17] – **Atlas brasileiro de desastres naturais 1991 a 2010: volume Brasil** / Centro Universitário de Estudos e Pesquisas sobre Desastres. Florianópolis: CEPED UFSC, 2012. 94p.

[18] – FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. The KDD process for extracting useful knowledge from volumes of data. **Communications Of The Acm**, Nova York, n. , p.27-34, 11 nov. 1996.

[19] QUINLAN, J. R.. **Induction of decision trees.** Boston: Induction Of Decision Trees, 1986. 1 v.

[20] ALVARES, Luis Otavio. **Agrupamentos Aglomerativos e DBSCAN.** Disponível em: <http://www.inf.ufsc.br/%7Ealvares/INE5644/aula_5_classificacao_arvores.ppt>. Acesso em: 26 nov. 2012.

[21] – T, Cover; P, Hart. Nearest neighbor pattern classification. **Journals & Magazines**, Boston, n. , p.21-27, 12 jan. 1967.

[22] – ALVARES, Luis Otavio. **Classificação: conceitos básicos e árvores de decisão.** Disponível em: <http://www.inf.ufsc.br/%7Ealvares/INE5644/classificacao_arvores.ppt>. Acesso em: 18 dez. 2012.

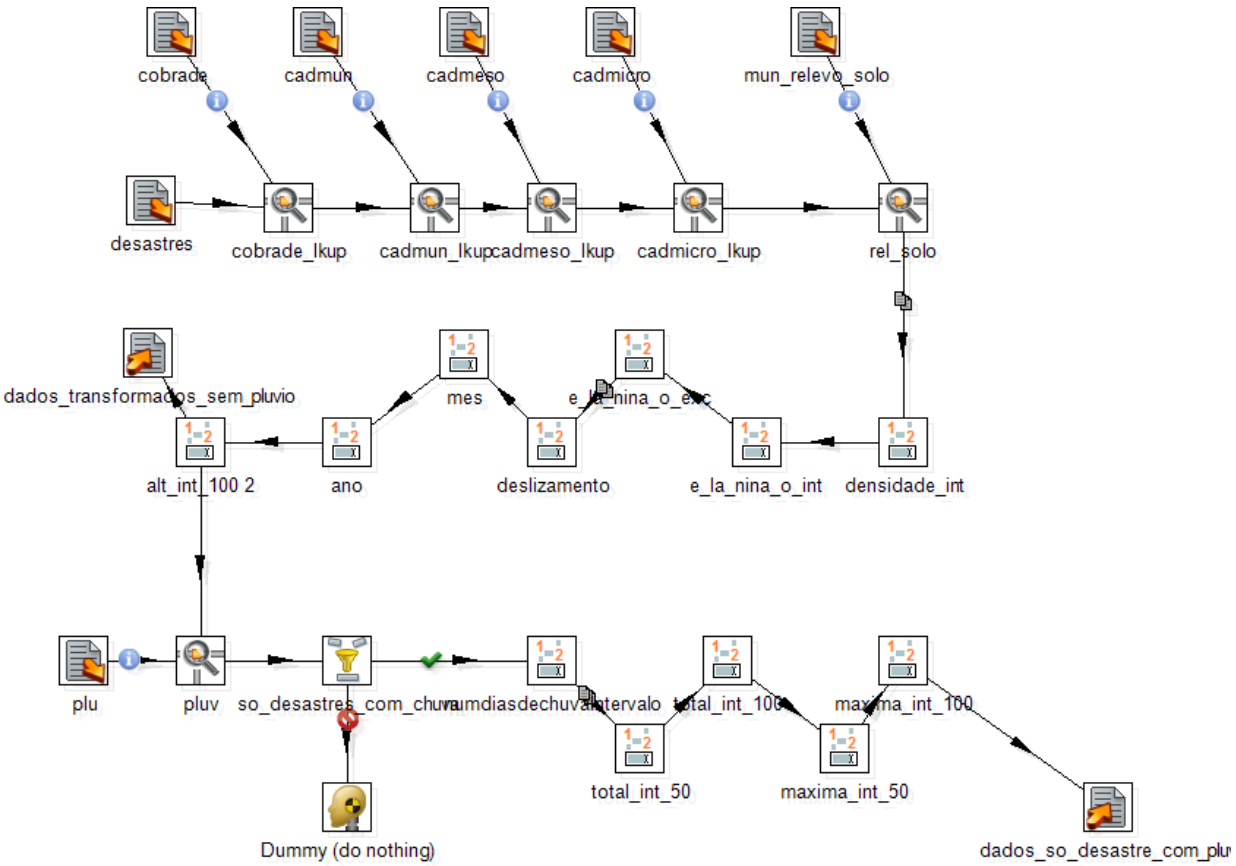
[24] – WIKIPÉDIA. **Microrregião de Florianópolis.** Disponível em: <http://pt.wikipedia.org/wiki/Microrregiao_de_Florianopolis>. Acesso em: 20 dez. 2012.

[25] – ESRI. **ESRI Shapefile Technical Description**. Disponível em: <http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf>. Acesso em: 11 mai. 2013

[26] – WEKA 3.6. **WEKA**. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/>. Acesso em: 11 mai. 2013

7. ANEXOS(S) E APÊNDICE(S)

ANEXO A – *Transformation* dos dados elaborada na etapa de preparação de dados.



APENDICE A – ARTIGO DO TRABALHO DE CONCLUSÃO DE CURSO

Análise Dos Registros De Desastres Naturais Através Da Utilização De Técnicas De Mineração De Dados

Mateus P. Mello[1]

[1] Departamento de Informática e Estatística – Universidade Federal de Santa Catarina (UFSC)
Caixa Postal 476 – 88040-900 – Florianópolis – SC – Brasil

Resumo. Este artigo pretende analisar os registros de dados de desastres de todo o Brasil. Para isto foram coletados dados referentes aos desastres naturais, tirando por base o preenchimento, por parte dos municípios, do formulário de avaliação de dados. Além destes dados também foram coletados dados referentes à ocorrência de do fenômeno *El Niño* e *La Niña*, índices pluviométricos, dados de solo e relevo dos municípios. O objetivo desta coleta é estudar os desastres através do processo de mineração de dados para a descoberta de padrões entre os desastres. As técnicas de mineração de dados utilizadas neste trabalho foram a clusterização (agrupamento) e classificação, para que se possa obter como resultado os padrões encontrados através da execução de algoritmos das duas técnicas citadas.

1. Introdução e Motivação

Segundo os dados levantados pelo Centro Universitário de Estudos e Pesquisa sobre Desastres (CEPED), no Brasil é possível encontrar diversos tipos de desastres naturais, desde inundações e deslizamento, até estiagens e queimadas.

O aumento da complexidade e da quantidade de calamidades, matando e ferindo milhares de brasileiros, fez com que se realizassem cada vez mais estudos no intuito de entender mais sobre desastres naturais e também de diminuir o sofrimento das pessoas.

Fazer a análise dos dados históricos de ocorrência de desastres naturais e dos níveis de precipitação de chuvas é um dos pontos fundamentais para que seja possível diminuir o número de vítimas e também para que se possa prever com maior antecedência a ocorrência de um desastre natural.

Uma das maneiras de fazer esta análise é através do processo de mineração de dados, do inglês, *data mining*. O processo de mineração de dados tem como objetivo extrair conhecimento novo, útil e interessante implícito nos dados, e representá-lo de forma acessível para o usuário KUMAR[12].

Atualmente existem diversas técnicas de mineração como classificação, agrupamento, clusterização, regras de associação, entre outros. A técnica de classificação pode ser útil para conceber um modelo de dados que tente predizer a iminência de um desastre natural com base na busca de padrões. No estudo realizado por LIMA [2] foi possível utilizar a técnica de agrupamento ou clusterização para criar grupos que possam ajudar a padronizar o processo de logística humanitária entre os municípios. Ainda segundo LIMA[2] um mesmo desastre pode atingir diferentes localidades ao mesmo tempo e isto sempre exige uma coordenação mútua para as ações de socorro.

O que este trabalho pretende é analisar os registros de dados de desastres de todo o Brasil. Os dados utilizados neste trabalho foram disponibilizados pelo CEPED, CPRM (Serviço Geológico do Brasil) e NOAA (Administração Oceânica e Atmosférica Nacional). Do CEPED foram utilizados os registros de AVADAN (Formulários de Avaliação de Danos), NOPRED (Formulário de Notificação Preliminar de Desastres), Relatórios de Danos (documento anterior ao AVADAN e NOPRED), decretos e portarias do governo. Do CPRM foi utilizado o índice dos municípios com suscetibilidade a deslizamentos e do NOAA a intensidade anual do *El Niño* e da *La Niña*. A seção 3.1 descreve com maiores detalhes todos os dados que foram obtidos.

2. Conceitos Básicos

2.1. Registros de Desastres

Os dados utilizados neste trabalho foram disponibilizados pelo CEPED, CPRM (Serviço Geológico do Brasil) e NOAA (Administração Oceânica e Atmosférica Nacional). Do CEPED foram utilizados os registros de AVADAN (Formulários de Avaliação de Danos), NOPRED (Formulário de Notificação Preliminar de Desastres), Relatórios de Danos (documento anterior ao AVADAN e NOPRED), decretos e portarias do governo. Do CPRM foi utilizado o índice dos municípios com suscetibilidade a deslizamentos e do NOAA a intensidade anual do *El Niño* e da *La Niña*. A seção 3.1 descreve com maiores detalhes todos os dados que foram obtidos. A figura 1 mostra as etapas do processo de oficialização de um desastre natural, deste a incidência do desastre até a publicação da Portaria.

Figura 1 – Esquema do registro de desastres.

Os principais dados que o AVADAN e o NOPRED disponibilizam são: município de ocorrência, data da ocorrência e tipo do desastre. Outros dados que também são possíveis de obter através deste formulário são estimativas de danos humanos (pessoas desabrigadas, desalojadas, mortas, danos, etc.), materiais e ambientais além de dados referentes aos prejuízos econômicos (produção de indústrias, agricultura e pecuária) e sociais (serviços interrompidos ou prejudicados).

2.2. A Mineração de Dados e o Processo de Descoberta de Conhecimento

A mineração de dados é uma tecnologia que combina métodos tradicionais de análise de dados com algoritmos sofisticados para processar grandes volumes de dados com o intuito de descobrir conhecimento que existe dentro destes grandes volumes TAN [1]. Conhecimento é o conjunto completo de informações, dados, relações que levam à tomada

de decisão, realização de tarefas e à criação de novas informações e, conforme já descrito anteriormente, a mineração de dados tem como objetivo descobrir este conhecimento.

O processo de descoberta de conhecimento foi apresentado por FAYYAD [18] e consiste em executar uma série de passos, desde a seleção, pré-processamento e transformação dos dados até a mineração destes dados e a interpretação dos resultados obtidos na mineração. A figura 2 apresenta as etapas do processo de descoberta de conhecimento, mostrando o resultado a cada etapa (dados relevantes, pré-processados, transformados, etc). Na seleção são obtidos os dados relevantes para o problema. No pré-processamento é realizada a limpeza de registros que estão incompletos, redundantes ou que geram incertezas. Na transformação são gerados novos dados a partir dos dados pré-processados. Na etapa de *Data Mining* (modelagem) é feita a busca por padrões nos dados gerando conhecimento e, por último, na interpretação, é onde os padrões serão analisados e compreendidos para que o conhecimento gerado possa ajudar na tomada de decisão.

Uma melhor maneira para compreender e elaborar um estudo utilizando mineração de dados é através da metodologia CRISP-DM (Processo Padrão Inter-Indústrias para Mineração de Dados). Esta metodologia, baseada no processo de KDD, foi concebida com o intuito de criar processos que padronizassem o desenvolvimento de projetos de mineração de dados.

O guia do CRISP-DM [10] define que o processo de mineração é cíclico e este ciclo está dividido em seis fases.

O entendimento do negócio cujo foco é entender os requisitos do ponto de vista do domínio. Após o entendimento do domínio deve-se definir qual o problema que o projeto irá solucionar [10].

O entendimento dos dados é a etapa onde se deve fazer a coleta e onde devem ser descritas as informações sobre os dados. Nesta etapa é importante fazer a exploração dos dados, gerando algumas estatísticas básicas, fazendo as primeiras descobertas e hipóteses que irão direcionar ao alcance dos objetivos estipulados no processo de mineração de dados.

A preparação dos dados, segundo ADRIAANS [11], representa cerca de 60% do esforço aplicado em um projeto de mineração. Esta fase visa preparar os dados disponíveis, que geralmente não estão dispostos em formato adequado para a aplicação dos algoritmos de descoberta, análise e a extração de conhecimento ALVARES [12] através da seleção, limpeza, transformação e construção de dados. Os dados precisam ter qualidade, isto é, estar limpos e compreensíveis, para extrair conhecimento interessante.

É na modelagem que ocorrem as execuções dos algoritmos sobre o conjunto de dados. Esta fase é dividida nas seguinte etapas: seleção dos algoritmos, geração do projeto de teste, aplicação dos algoritmos e avaliação o modelo gerado, quantas vezes for necessário para obter o melhor resultado de acordo com o entendimento do negócio.,

Avaliação Na avaliação dos resultados, deve ser verificado se os resultados atingiram os objetivos do projeto. Na revisão do projeto deve ser analisado se o resultado é satisfatório para os objetivos definidos. É apropriado fazer uma revisão mais aprofundada da mineração de dados a fim de determinar se existe mais algum fator importante ou tarefa que tenha sido negligenciada.

Disponibilização é a ultima etapa do ciclo, onde nesta etapa do processo deve-se produzir um relatório final com os resultados, mostrando os pontos positivos e negativos, os problemas encontrados e trabalhos futuros apresentados na fase de avaliação.

2.3. Principais Técnicas de Mineração de Dados

2.3.1. Classificação

O resultado da execução de um algoritmo de classificação é a geração de um modelo que pode ser utilizado para atribuir uma classe a diferentes registros ainda não classificados. Supondo o tipo de desastre natural como um atributo classe (y), e características morfoclimáticas, geológicas e precipitações de chuva de uma determinada região como o conjunto de atributos (x), é possível gerar um modelo que possa predizer a relação entre os tipos de desastres e os demais atributos tendo como base um conjunto de ocorrências de desastres naturais. Conforme TAN [1] o modelo de classificação pode ser tratado como uma caixa preta que atribui automaticamente um rótulo de classe quando recebe o conjunto de atributos de um registro desconhecido.

A árvore de decisão é uma representação gráfica dos registros de acordo com seus atributos e valores. A Figura 2 ilustra a transformação dos registros de maus pagadores de empréstimos em uma árvore de decisão.

A árvore é gerada a partir do conjunto de registros de treinamento e posterior a isto somente será utilizada para classificar os demais registros. Na árvore, cada nó representa um atributo, os ramos (ligam os nós) correspondem aos valores de um atributo e os nós folha (nós que não possuem sucessores) representam uma classe. A figura destaca o que são atributos, valores e classe. Dois algoritmos que utilizam a árvore de decisão são o ID3 (QUINLAN, 1986) e o C4.5 (QUINLAN, 1993). A diferença entre estes dois algoritmos é

que o ID3 tem um resultado melhor quando utilizado atributos descritivos, já para o C4.5 o melhor é que seja utilizado atributos numéricos.

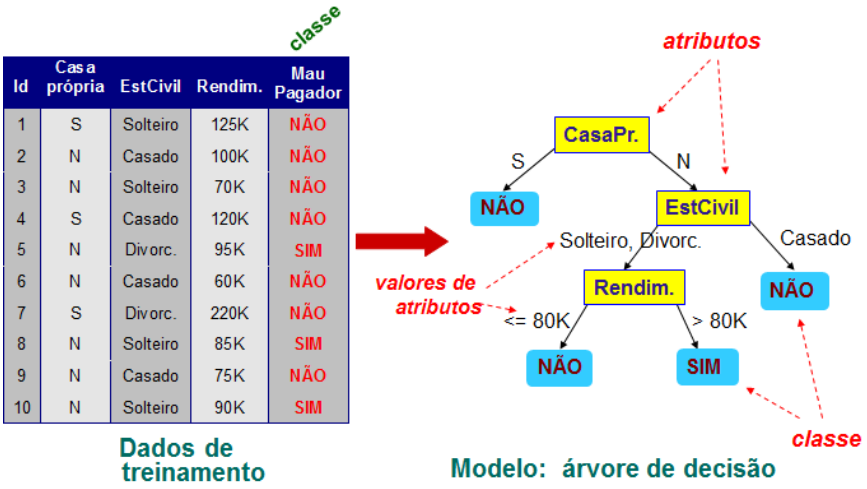


Figura 2. Os elementos de uma árvore de decisão ALVARES [22].

2.3.2. Agrupamento ou Clusterização

Segundo TAN [1], a análise de grupo, chamada *clustering*, une registros baseado apenas em informações encontradas nos atributos que descrevem os registros e seus relacionamentos. O ponto principal é entender que quanto maior for a semelhança dentro do grupo (*cluster*) e maior a diferença entre os grupos, melhor será o resultado encontrado, isto é, o agrupamento será mais preciso.

Existem diversas técnicas de agrupamento, e a criação dos grupos de registros pode ser abordada através do agrupamento particional ou baseado em densidade TAN[1]. No agrupamento particional cada registro pertencerá a um único grupo, não podendo estar inserido dentro de um grupo maior ou menor. Já no densidade um grupo será formado quando existir uma região densa, isto é, uma região com uma grande quantidade de registros.

Segundo TAN [1], o K-means é a uma técnica de agrupamento particional, isto é, cada registro pertencerá a um único grupo. Estes grupos são formados a partir de um conjunto de registros mais próximos ao registro que define um grupo, e este registro recebe o nome de centroide. A quantidade de centroides é um parâmetro que deve ser definido pelo analista de mineração de dados e o valor deste parâmetro pode variar a cada execução da técnica.

Algumas limitações encontradas no K-means é que está técnica não se mostra muito eficiente em conjuntos de registros com tamanhos e densidades diferentes e também quando o conjunto de objetos não possui formato esférico.

Já o DBSCAN é um algoritmo de agrupamento baseado na densidade dos registros TAN [1]. Uma região densa é uma região de alta densidade de registros. A utilização desta técnica é ideal quando grupos são irregulares ou entrelaçados e também quando existem muitos ruídos externos.

O DBSCAN consegue agrupar os registros através da utilização de dois parâmetros que devem ser definidos pelo analista. O primeiro é a densidade, que é quantidade mínima de registros que devem estar inseridos dentro do raio (*minPts*). O segundo é o raio (*Eps*), que é o espaço que será analisado a partir de um determinado registro.

3. Materiais e Métodos Utilizados Para o Processo de Mineração de Dados

As subseções deste capítulo foram criadas com base nas etapas do CRISP-DM. A etapa de Entendimento do Negócio ou domínio já foi descrita no capítulo 1 através da introdução ao assunto e objetivos do trabalho.

3.1. Entendimento dos Dados de Desastres Naturais

No total foram coletados mais de 41.217 registros oficiais junto à SEDEC (Secretaria Nacional de Defesa Civil) dos mais diversos tipos de desastres naturais. Este conjunto de dados apresenta como atributos: *nome do desastre natural*, *dia*, *mês*, *ano*, *estado e município de ocorrência e a densidade demográfica do município*.

Também foi coletada a temperatura, em grau Celsius, a intensidade do *El Niño* ou *La Niña* no mês e ano em que ocorreu o desastre natural.

Outro dado obtido foi obtido o atributo binário (0,1) se o município possui áreas suscetíveis a deslizamentos. Este atributo foi levantado através de estudos feitos pelo CPRM [16], que é a instituição do governo responsável por organizar e sistematizar o conhecimento geológico do território brasileiro. Caso seja um município que tem áreas suscetíveis o valor será 1, caso contrário 0 (zero).

Outros dados levantados a partir da ANA (Agência Nacional de Águas) foram os índices pluviométricos mensais dos estados brasileiros, por estação de medição. Para cada mês é quantificado o total de chuvas em milímetro, o número de dias em que choveu, a máxima em milímetros de precipitação em um dia. Estas informações são importantes para identificar a relação entre mês e dia de ocorrência de um desastre natural com os números pluviométricos do mesmo mês.

Também foram obtidos, do IBGE, dados de relevo e solo predominante de cada município. Foi possível obter características geomorfológicas (formas da superfície terrestre) , unidade de relevo (depressão, planalto, etc) e a estrutura de relevo em que se situa o município (Marinhas, bacia do Paraná, Araucária, Alto do Tocantins, etc).

A Tabela 1 mostra resumidamente todos os atributos que foram utilizados para o processo de mineração de dados. A coluna da esquerda mostra o nome do atributo e o da direita um exemplo do valor.

Tabela 1. Atributos utilizados no processo de mineração de dados

Atributo	**Exemplo de Valor**
Tipo do desastre	Deslizamentos
Ocorrência de La Niña/El Niño	*La Niña* Moderada
Mês de ocorrência do desastre	Outubro
Ano de ocorrência do desastre	1999
Estado (UF)	Santa Catarina
Nome da mesorregião	Vale do Itajaí
Nome da microrregião	Blumenau
Nome do município	Blumenau
Munícipio em área de Amazônia	N*(Não)*/S*(Sim)*
Munícipio em área de fronteira	N*(Não)*/S*(Sim)*
Município com área suscetível a deslizamento	0*(Não)*/S*(Sim)*
Domínio morfológico predominante do município	Embasamentos em Estilos Complexos
Subdomínio morfológico predominante do município	Embasamento do Sul/Sudeste
Unidade de relevo predominante do município	Serras
Localização da unidade de relevo	Leste Catarinense
Tipo de solo predominante do município	Solo Podzólicos
Densidade do município (hab/km²)	160,4
Altitude do município (metros)	90
Total de chuvas em um mês (milímetros)	200
Máxima de chuvas em um dia (milímetros)	90
Números de dias com chuva em um mês	20

O maior número de registros foi referente à estiagem, cerca de 39%; o segundo maior foi enxurrada com aproximadamente 23%; e o terceiro maior foi inundação, com 14,23% dos casos. Desastres naturais como deslizamentos e alagamentos somam aproximadamente 2,5%. 34

Entretanto mesmo sendo um percentual baixo, ainda é importante estudar estes desastres isoladamente devido ao impacto humano que ele gera.

Agrupando registros de desastres naturais por microrregião brasileira foi possível encontrar uma boa dispersão da ocorrência dos desastres e isto pode ser um fator importante na aplicação dos algoritmos de mineração, já que diminui o risco do resultado ser tendencioso para um determinado desastre.

De acordo com o IBGE [23], microrregião é um agrupamento de municípios limítrofes com o objetivo de estruturar o espaço geográfico de acordo com a produção agropecuária, industrial, extrativista ou pesca, além de trocas de consumo entre os munícipios e atividades urbanas e rurais.

Apesar da Figura 3 não mostrar todo o mapa brasileiro (por motivo de falta de espaço), foi feita uma comparação entre todas as microrregiões brasileiras destacando as vinte microrregiões que mais registraram desastres naturais.

A microrregião que obteve a maior quantidade de registros foi a de Chapecó-SC, com 701 registros de desastres naturais (Figura 3). Este valor representa o total de 2% em relação a todos os registros de desastres (41217 ocorrências). Obtendo a relação das vinte microrregiões que mais registraram desastres naturais, Chapecó, representa um total de 10%.

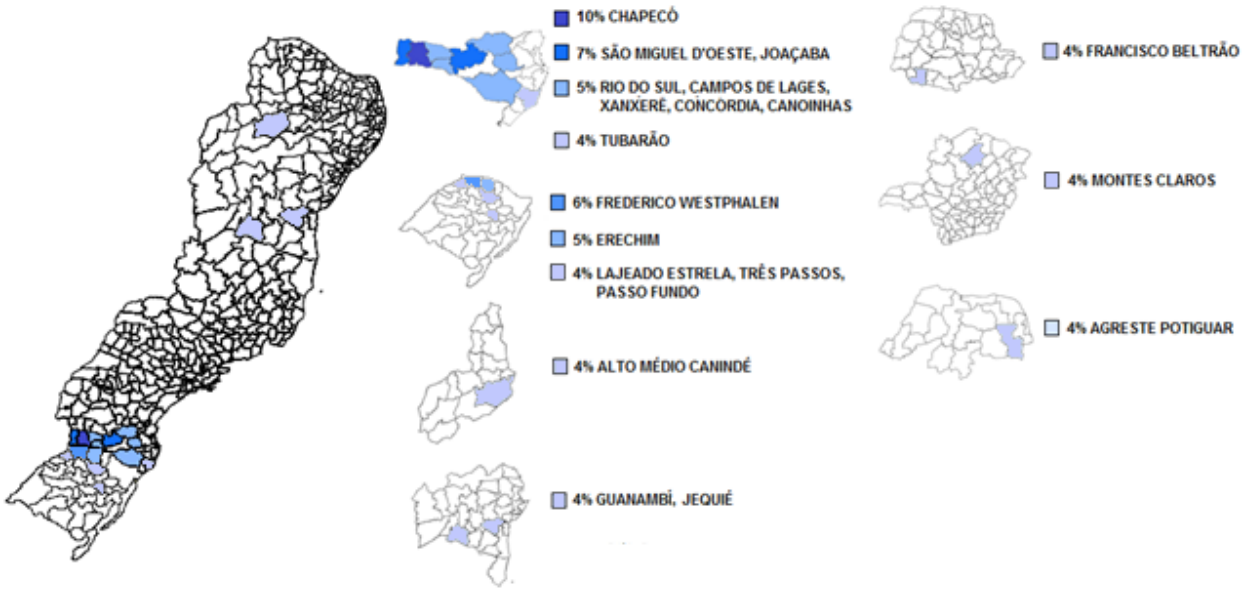


Figura 3 – Dispersão das incidências de desastres naturais por microrregião.

3.2. Preparação dos Dados Referentes aos Desastres Naturais

A transformação dos dados foi realizada de três maneiras: manipulação manual das tabelas no MS Office, criação de consultas SQL e a utilização da ferramenta Pentaho Kettle. Optou-se também por utilizar o Kettle, que é uma ferramenta comumente utilizada em *data*

warehouse, no processo de extração, transformação e carga de dados, pois o pré-processamento em mineração de dados é relativamente parecido com este processo.

Como nem todos os municípios que possuíam registros de desastres naturais possuíam registros pluviométricos (*máxima de chuva em um dia, total de chuvas no mês e quantidade de*

dias com chuva no mês) foi necessário criar duas tabelas, uma com todos os registros de desastres naturais (41.217 registros de desastres) e outra somente com os registros de desastres naturais que ocorrem em municípios que tinham dados pluviométricos (6.864 registros de desastres naturais).

Para fazer análises específicas de um determinado desastre natural foram geradas tabelas secundárias baseadas na principal. Foi realizada a segmentação dos registros buscando somente os desastres a serem analisados (vendavais, granizo, deslizamento, alagamentos, etc).

Na tabela específica dos índices pluviométricos as três medidas apresentadas (quantidade total de chuvas no mês, número de dias de chuva, máxima de chuva em um dia) foram definidas por estação pluviométrica.

A limpeza dos dados é uma operação básica de remoção de ruídos, atributos incompletos ou erros. A Tabela 2 mostra o caso das estações do município de Curitibanos – SC, onde existem duas estações pluviométricas que não foram registrados dados. Para estas estações, e todas as outras onde não existe registro, foi feita a remoção das mesmas do processo de mineração de dados.

Tabela 2 – Índices pluviométricos com estações sem registros de Curitibanos – SC

Município	**Nome da Estação Pluv.**	**Máxima Dia**	**Dia da Máxima**	**Numero de Dias Com Chuva**	**Total**
CURITIBANOS	PONTE ALTA DO NORTE	40.9	8/9/2005	26	92.1
CURITIBANOS	PONTE DO RIO ANTINHAS	33	8/9/2005	26	112.8
CURITIBANOS	BARRAGEM PERY	sr	sr	sr	sr
CURITIBANOS	SALTO PERY	sr	sr	sr	sr
CURITIBANOS	PONTE ALTA DO NORTE – CIFSUL	32	11/9/2005	25	98.2
		40,9	**8/9/2005**	**26**	**101.63**

Na etapa de transformação dos dados a intensidade do *El Niño* e da *La Niña* foi classificado como fraco, moderado e forte. Para que seja atribuída uma das três intensidades a temperatura do *E Niño/La Niña* em graus Celsius foi alterada para valores qualitativos onde: entre 0,5 e 0,9 é fraco, 1,0 e 1,4 é moderado e maior ou igual a 1,5 graus Celsius é forte, sendo positivamente para o *El Niño* e negativamente para a *La Niña*. Além disso, foi criada outra coluna que informa se ocorreu El Niño ou La Niña, independente da força do fenômeno, com o intuito de gerar maiores possibilidades de encontrar padrões nos dados.

Outros quatro atributos que também foram transformados em intervalos foram: *densidade populacional*, *altitude*, *total de chuvas no mês* e *número de dias com chuva* divido os mesmos es escalas que favoreciam o estudo de cada atributo citado.

O ano de ocorrência do desastre também foi transformado para intervalos de 5 em 5 anos para que fosse possível executar o algoritmo ID3, já que este só aceita valores textuais.

3.3. Algoritmos Utilizados

A ferramenta utilizada para aplicar os algoritmos existentes foi o Weka 3.6 FRANK [26]. O algoritmo de árvore de decisão utilizado para classificar os registros foi o ID3 (QUINLAN, 1986) e para nível de comparação e análise do algoritmo de classificação com o resultado mais satisfatório, também foi utilizado o C4.5 (QUINLAN, 1993). A princípio foi utilizado como atributo classe o tipo de desastre natural, na tentativa de obter algum resultado interessante e tentar predizer quais atributos tem capacidade de influenciar a ocorrência de um desastre natural.

Já no agrupamento foram utilizados os algoritmos K-means (McQueen, 1967) para gerar grupos que terão sempre registros exclusivos em cada grupo e o DBSCAN (Ester *et al.* 1996) para analisar os registros baseado na densidade.

4. Análise dos Resultados

Foram feitas análises com os tipos de desastres que mais apresentaram ocorrências: *enxurrada*, *alagamento* ou *inundação*, *deslizamento*, *seca* ou *estiagem*, *granizo* e *vendaval*. Também realizou-se alguns pré-processamentos como a remoção de determinados atributos e a seleção de um ou mais tipos de desastres naturais para uma análise isolada, com o intuito de obter resultados melhores.

A Tabela 3 apresenta os mesmos atributos considerados no processo de mineração com um exemplo de valor para cada atributo. Todos os valores apresentados nesta coluna levam em consideração o processo de preparação dos dados, isto é, todos os valores que eram números foram passados para intervalos.

Tabela 3 – Atributos utilizados no processo de mineração de dados com exemplos de valores após o processo de transformação dos dados.

Atributo	**Exemplo de Valor**
Tipo do desastre	Deslizamentos
La Niña/El Niño dividido em intervalos	*El Niño* Fraco; *El Niño* Moderado*; El Niño* Forte*;* Sem *El Niño*/*La Niña*; *La Niña* Fraca*; La Niña* Moderada*; La Niña* Forte
Somente La Niña/El Niño	*La Niña; El Niño;* Sem *El Niño*/*La Niña*;
Mês de ocorrência do desastre	Outubro
Ano de ocorrência do desastre	Entre 1990 e 1994; Entre 1995 e 1999; Entre 2000 e 2004; Entre 2005 e 2009
Estado (UF)	Santa Catarina
Nome da mesorregião	Vale do Itajaí
Nome da microrregião	Blumenau
Nome do município	Blumenau
Munícipio em área de Amazônia	N*(Não);*S*(Sim)*
Munícipio em área de fronteira	N*(Não)*;S*(Sim)*
Município com área suscetível a deslizamento	Suscetível a deslizamentos;Não Suscetível a deslizamentos
Domínio morfológico predominante do município	Embasamentos em Estilos Complexos
Subdomínio morfológico predominante do município	Embasamento do Sul/Sudeste
Unidade de relevo predominante do município	Serras
Localização da unidade de relevo	Leste Catarinense
Tipo de solo predominante do município	Solo Podzólicos
Densidade do município (hab/km²)	Menor que 1; Entre 1 e 5; Entre 5 e 10; Entre 10 e 20; Entre 20 e 50; Entre 50 e 100; Entre 100 e 200; Maior de 200
Altitude do município (metros)	Entre 20 e 99,9m; Entre 100 e 199,9m; Entre 200 e 299,9m;...; Mais de 1000m
Total de chuvas em um mês (50 mm)	Menos de 20mm; Entre 20 e 49,9mm; Entre 50 e 99,9mm; etc...
Máxima de chuvas em um dia (50 mm)	Menos de 20mm; Entre 20 e 49,9mm; Entre 50 e 99,9mm; etc...
Total de chuvas em um mês (100 mm)	Menos de 20mm; Entre 20 e 49,9mm; Entre 50 e 99,9mm; etc...

Máxima de chuvas em um dia (100 mm)	Menos de 20mm; Entre 20 e 49,9mm; Entre 50 e 99,9mm; etc...
Números de dias com chuva em um mês	Menos de 5 dias; De 5 a 10 dias; De 10 a 15 dias; etc

4.1. Execução dos Algoritmos de Agrupamento

As execuções utilizando o algoritmo DBSCAN não resultaram em grupos (*clusters*) que pudessem gerar alguma informação relevante. Foram feitas diversas tentativas alterando os parâmetros do algoritmo (*epsilon* e número mínimo de pontos, *minpoint*), mas os dados estavam todos em um mesmo *cluster* ou em dezenas de *clusters*, dificultando a interpretação, o que acarretava em uma análise imprecisa dos dados. Já com o K-*Means* foi possível encontrar alguns resultados.

Como nem todos os municípios com registros de desastres naturais tem dados pluviométricos (total e máxima de chuvas em um dia e a quantidade de dias em que choveu) foi necessário executar o algoritmo em dois momentos distintos, já que muitos registros com valores nulos para dados pluviométricos influenciaria no resultado da execução do K-*means*.

A partir da utilização de algoritmos de clusterização sobre os registros de desastres naturais foi possível confirmar estudos do CPTEC sobre a relação entre o fenômeno *La Niña* e *El Niño* e as regiões brasileiras. Com base nos dados, confirmou-se que desastres de estiagem e secas tem relação com a região sul, quando sob efeito do *La Niña*.

Sobre o relevo existente no Brasil, conclui-se que as regiões de depressão são áreas normalmente não suscetíveis a deslizamentos e que em regiões de escarpas e reversos existe a suscetibilidade a deslizamento.

Outro ponto importante concluído neste trabalho foi que existe uma diferença quando analisados os registros dos estados e das microrregiões. Microrregiões como a de Chapecó, em Santa Catarina, possui uma quantidade grande de registros de estiagem e secas, quando o estado de Santa Catarina é apenas o sexto estado que mais possui registros de secas e estiagem. Isto refletiu diretamente no processo de clusterização, pois foram gerados *clusters* agrupados com microrregiões de Santa Catarina, mas não foram gerados *clusters* com o estado.

Do ponto de vista dos meses do ano verificou-se que o verão é uma estação com uma tendência maior a ocorrer desastres de enxurradas, deslizamentos, alagamentos ou deslizamentos. Também foi possível fazer uma relação de alguns estados ou regiões brasileiras e os meses do ano para estes mesmos tipos de desastres. A Tabela 4 mostra quais foram os padrões concluídos, onde na coluna da esquerda é apresentado o estado ou região e na direita o mês.

Tabela 4 – Relação entre os estados ou regiões e os meses para os desastres de enxurradas, deslizamentos, alagamentos ou inundações.

Estado ou Região	Mês Padronizado

São Paulo e Paraná	Janeiro
Mato Grosso do Sul	Março
Norte e Nordeste	Março, Abril e Maio
Bahia	Dezembro

4.2. Análise dos Resultados dos Algoritmos de Classificação

Os algoritmos utilizados para fazer a classificação foram o ID3 e o C4.5. Foram analisados registros de enxurradas, inundações, deslizamentos, alagamentos, secas e estiagens, com o intuito de encontrar uma relação entre os índices pluviométricos e estes desastres.

Como se desejou encontrar padrões para desastres naturais que mais afetam vidas humanas, utilizou-se como atributo classe o *tipo de desastre* com os valores enxurradas, inundações, deslizamentos, alagamentos, secas e estiagens.

Também foi criado outro atributo exclusivamente para os algoritmos de classificação (*tem tendência a enxurrada*). O objetivo do atributo é encontrar padrões que caracterizem a ocorrência de enxurrada. Para isto foi feita uma comparação entre os registros de enxurradas e inundação, visto que são os dois tipos que tiveram mais registros relacionados com altos índices de chuva (1.568 e 944 registros, respectivamente). Para que não houvesse uma influência maior dos desastres de enxurradas, foram considerados somente 944 registros de enxurradas, mesma quantidade de registros de inundações.

Foram efetuados diversos experimentos com os algoritmos de classificação, entretanto em geral, os resultados não atingiram o nível de qualidade que se esperava. O ideal era também ter trabalhado com dados pluviométricos onde não ocorreram registros de desastres naturais para que se pudesse fazer um comparativo entre os valores dos dados em momentos em que ocorreram desastres naturais e momentos em que não ocorreram.

Foi possível observar é que para alguns casos de registros de estiagens e seca os total de chuvas em um mês ultrapassava a quantidade de 150 milímetros. Foi o caso do município de Caruaru, em Pernambuco, onde teve um registro de 142 mm em junho de 2001, mas também apresentou um registro de estiagem para este mês. Isto fez com que os modelos de classificação fossem prejudicados, devido esta aparente uma divergência de dado, já que para esta quantidade de chuva não deveriam existir registros de desastres naturais de estiagem ou seca. Uma possível explicação para este problema é que pode ter havido seca ou estiagem em uma parte no município e ter chovido muito em outra parte (local onde está a estação pluviométrica) e como a granularidade de informação é o município, esse tipo de problema pode ocorrer.

Diante desta conclusão, o algoritmo que apresentou resultado mais satisfatório, foi o C4.5. A seção a seguir descreve os resultados obtidos a partir dos atributos chave *tipo de desastre* e *tem tendência a enxurrada*.

O algoritmo C4.5 apresentou resultados melhores quando foram utilizados atributos numéricos ao invés de descritivos. As duas árvores de classificação geradas a partir das diversas execuções realizadas com o algoritmo C4.5 foram testadas através da validação cruzada, dividindo os dados em 10 partes (*folds*). Para este algoritmo também foi utilizada a poda da árvore gerada, que é uma técnica do algoritmo que permite gerar uma árvore menor e mais compreensível.

Tanto para a execução utilizando com o atributo chave o *tipo de desastre* e o atributo criado *tem_enxurrada*, utilizou-se os atributos: *índice de suscetibilidade a deslizamento*, *unidade do relevo*, *densidade populacional*, *temperatura do aumento das águas do oceano Pacífico* (*El Niño*/*La Niña*), *altitude*, *mês*, *total de chuvas no mês*, *máxima de chuva em um dia* e *quantidade de dias que choveram no mês*.

A utilização do atributo chave *tipo de desastre* gerou um modelo que classificou os registros corretamente em aproximadamente 64% dos casos. Este valor pode ser considerado satisfatório, já que foram analisados quatro tipos de desastres, isto é, a probabilidade de acerto na predição do desastre é de uma em quatro, ou seja, 25%.

A Tabela 5 mostra a matriz de confusão gerada pela execução, mostrando como os registros foram classificados utilizando a árvore gerada. Na primeira coluna é apresentado o tipo de desastre que se espera ser classificado e na primeira linha qual foi a classificação realmente feita.

Tabela 5 – Matriz de confusão gerada a partir do atributo chave *tipo de desastre*

Inundações	**463**	481	0	0
Enxurradas	355	**1213**	0	0
Alagamentos	18	28	**0**	0
Deslizamentos	7	33	0	**0**

Analisando os resultados da execução do C4.5, algumas informações sobre os desastres que podem ser extraídas é que regiões de escarpa e reversos tem uma probabilidade maior de ocorrer desastres de enxurradas.

Já para o atributo chave criado para verificar se *tem tendência a enxurrada* o modelo gerado classificou corretamente os registros em 74% dos casos. O resultado não foi tão

satisfatório, já que o atributo é binário (sim ou não) e a probabilidade de acerto é de 50%. No entanto, já era esperado que a porcentagem fosse um pouco mais baixa, já que foi utilizada a técnica de poda da árvore, o que diminui a acurácia do resultado de quase 80% (antes da poda) para 74% (depois da poda).

A Tabela 6 mostra a matriz de confusão gerada a partir do atributo chave *tem tendência à enxurrada*. Nela é possível observar a porcentagem do resultado da classificação, utilizando a árvore gerada pelo algoritmo C4.5, em números. Os valores escritos em verde e em negrito (645 e 756) mostram a quantidade de registros que foram corretamente classificados, e os escritos em vermelho e em negrito (299 e 189) a quantidade que foi erroneamente classificada.

Tabela 6 – Matriz de confusão gerada a partir do atributo chave *tem tendência à enxurrada.*

tem tendência a enxurrada	SIM	NÃO
SIM	**645**	299
NÃO	189	**756**

Mesmo com este resultado pode-se concluir a partir das análise feitas sobre a execução do algoritmo com o atributo chave *tem tendência a enxurrada*, que regiões de chapadas, planaltos, cristais, colinas e tabuleiros não costumam apresentar enxurradas. Por outro lado, regiões de escarpas e reversos podem apresentar enxurradas se a quantidade total de chuvas em um mês passa de 147 milímetros e regiões de serras podem apresentar desastres de enxurradas quando chove mais de 16 dias em um mês.

5. Conclusão e Trabalhos Futuros

Devido ao aumento da ocorrência de desastres naturais e do grau de complexidade, a análise dos registros de desastres tornou-se uma necessidade para que se possa conhecer melhor o histórico brasileiro de desastres.

Este trabalho teve como objetivo realizar a coleta de dados e estudar os desastres naturais brasileiros através do processo de mineração de dados para a descoberta de padrões entre os desastres.

A partir da coleta de dados como: registros dos desastres naturais ocorridos no Brasil (AVADAN), intensidade do *El Niño* ou *La Niña*, índice de suscetibilidade a deslizamento de um determinado município, relevo e solo predominante dos municípios e índices pluviométricos; foi feito o entendimento dos dados, levantando informações relevantes que pudessem auxiliar no processo de *data mining*, como a identificação dos tipos de desastres naturais que mais ocorrem em uma determinada microrregião brasileira. Em seguida foi realizada a preparação dos dados para que

pudessem ser utilizados os algoritmos de agrupamento (DBSCAN e K-means) e classificação (ID3 e C4.5), esta etapa foi a que mais precisou de tempo e dedicação para a elaboração deste trabalho.

Especificamente a preparação dos dados pluviométricos (máxima de chuvas em um dia, total de chuvas em um mês e quantidade de dias com chuva no mês) foi a que gerou o maior trabalho, pois os dados estavam distribuídos, dentro dos arquivos, de modo que dificultava a união entre os registros de desastre e os dados pluviométricos. Além disto, estes dados também estavam em mais de um arquivo (um para cada estado brasileiro), o que gerou a necessidade de dar uma atenção maior.

Com a seleção, limpeza, e transformação dos dados foi possível executar os algoritmos escolhidos e com o resultado das diversas execuções feitas, pôde-se concluir que existe uma relação dos registros de desastres naturais com a precipitação de chuvas. Gerou-se grupos (*clusters*) levando em consideração somente desastres relacionados com altos índices de chuva que variaram de 200mm por mês chegando até a 349,9mm, já a quantidade de chuvas em um dia ficou entre 50mm e 99,9mm.

Por outro lado, desastres relacionados com baixos índices pluviométricos costumam apresentar grupos de no máximo 50mm de chuvas em um mês e em alguns grupos foi possível observar que a máxima em um dia também se aproximou de 50mm.

Além destas conclusões, também se obteve outras que não faziam parte dos objetivos do trabalho, mas que são importantes do ponto de vista da análise dos registros de desastres naturais. São elas:

- Áreas com suscetibilidade a deslizamentos tendem a estar em regiões de escarpas e reversos. Regiões de planalto podem ser tanto suscetíveis quanto não suscetíveis a deslizamentos, e depressões tendem a não ser suscetíveis.
- Existe uma relação de alguns *clusters* criados com a incidência do fenômeno *La Niña* e *El Niño.*
- Desastres relacionados com altos índices de chuva tendem a ocorrer com maior frequência no Norte/Nordeste no mês de abril. Exclusivamente a Bahia costuma apresentar estes tipos de desastre em dezembro. Paraná e São Paulo apresentam desastres de inundações, enxurradas, deslizamentos e alagamentos em janeiro. Este último estado também pode apresentar em fevereiro e o Mato Grosso do Sul em março.

Como nem todos os conjuntos de dados revelaram algum padrão, seria possível obter resultados mais relevantes para uma quantidade maior de registros de desastres naturais, aumentando a precisão dos resultados, bem como mais dados de domínios diferentes (vegetação, clima, agropecuária etc...) referentes aos municípios, microrregiões e estados, aumentando a gama de informações.

Além de trabalhar com outros domínios de dados também se mostrou importante trabalhar com índices, como o índice de suscetibilidade a deslizamentos, criados a partir de estudos realizados anteriormente. Outro ponto a destacar é que existiu muita perda de registros com índices pluviométricos, pois diversos registros ocorreram em municípios que não tiveram dados pluviométricos no ano e mês do desastre.

6. Referências

TAN, Pang-ning; STEINBACH, Michael; KUMAR, Vipin. Introdução ao Data Mining. Rio de Janeiro: Ciência Moderna, 2009. 900 p.

LIMA, Fabiana Santos; OLIVEIRA, Daniel de; GONÇALVES, Mirian Buss. A FORMAÇÃO DE CLUSTERS NA LOGÍSTICA HUMANITÁRIA UTILIZANDO MINERAÇÃO DE DADOS. Florianópolis, 2011. 12 p.

ESTÉBANEZ, Katiusca Magdalena Briones. Identificação de Padrões na Ocorrência de Emergências e Desastres Associados a níveis de Precipitação. 2012. 88 f. Dissertação (Mestrado) – Curso de Engenharia Civil, Departamento de Coppe, Ufrj, Rio de Janeiro, 2012.

SOUZA, Fábio Teodoro de. PREDIÇÃO DE ESCORREGAMENTOS DAS ENCOSTAS DO MUNICÍPIO DO RIO DE JANEIRO ATRAVÉS DE TÉCNICAS DE MINERAÇÃO DE DADOS. 2004. 115 f. Tese (Doutorado) – Curso de Engenharia Civil, Departamento de Coppe, Ufrj, Rio de Janeiro, 2004.

CHAGAS, Diego José; CHAN, Chou Sin; CORSI, Alessandra Cristina. Análise do Banco de Atendimentos da Defesa Civil do Estado de São Paulo. São Paulo: Xwq, 2010. 9 p.

WIKIPÉDIA. São Paulo. Disponível em: <http://pt.wikipedia.org/wiki/Sao_Paulo>. Acesso em: 22 nov. 2012.

WIKIPÉDIA. Equador. Disponível em: <http://pt.wikipedia.org/wiki/Equador>. Acesso em: 22 nov. 2012.

GOVERNO Federal. CONSTITUIÇÃO DA REPÚBLICA FEDERATIVA DO BRASIL DE 1988. Disponível em: <http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm>. Acesso em: 22 nov. 2012.

V SEMINÁRIO INTERNACIONAL DE DEFESA CIVIL – DEFENCIL, 2009, São Paulo. Os desastres naturais, a cultura de segurança e a gestão de desastres no Brasil. São Paulo: Meio Eletrônico, 2009. 8 p.

CHAPMAN, Pete et al. CRISP-DM 1.0: Step-by-step data mining guide. Eua, Alemanha, Dinamarca, Holanda: Spss, 2000.

ADRIAANS, P., and D. Zantige. 1996. Data mining. Harlow, England

ALVARES, Luis Otavio. O processo de KDD. Disponível em: <http://www.inf.ufsc.br/%7Ealvares/INE5644/processo_DCBD.ppt>. Acesso em: 24 nov. 2012.

ALVARES, Luis Otavio. Agrupamento K-means. Disponível em: <http://www.inf.ufsc.br/%7Ealvares/INE5644/clustering1.ppt>. Acesso em: 26 nov. 2012.

TAN, Pang-ning; STEINBACH, Michael; KUMAR, Vipin. Introdução ao Data Mining. Rio de Janeiro: Ciência Moderna, 2009. 900 p.

XIV SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 14., 2009, Natal – Rn. Relação entre os desastres naturais e as anomalias de precipitação para a região Sul do Brasil. Santa Maria – Rs: Inpe, 2009. 8 p. Disponível em: <http://www.inpe.br/crs/geodesastres/conteudo/publicacoes/3527_3534_BARBIERI_Relacao_entre_desastres_naturais_2009.pdf>. Acesso em: 01 dez. 2012.

NULL, Jan. El Niño and La Niña Years and Intensities. Disponível em: <http://ggweather.com/enso/oni.htm>. Acesso em: 01 dez. 2012.

BRASIL, Serviço de Geologia do. SELEÇÃO DE MUNICÍPIOS CRÍTICOS A DESLIZAMENTOS: NOTA EXPLICATIVA. Disponível em: <http://www.cprm.gov.br/publique/media/apresentacao_susc.pdf>. Acesso em: 01 dez. 2012.

Atlas brasileiro de desastres naturais 1991 a 2010: volume Brasil / Centro Universitário de Estudos e Pesquisas sobre Desastres. Florianópolis: CEPED UFSC, 2012. 94p.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. The KDD process for extracting useful knowledge from volumes of data. Communications Of The Acm, Nova York, n. , p.27-34, 11 nov. 1996.

QUINLAN, J. R.. Induction of decision trees. Boston: Induction Of Decision Trees, 1986. 1 v.

ALVARES, Luis Otavio. Agrupamentos Aglomerativos e DBSCAN. Disponível em: <http://www.inf.ufsc.br/%7Ealvares/INE5644/aula_5_classificacao_arvores.ppt>. Acesso em: 26 nov. 2012.

T, Cover; P, Hart. Nearest neighbor pattern classification. Journals & Magazines, Boston, n. , p.21-27, 12 jan. 1967.

ALVARES, Luis Otavio. Classificação: conceitos básicos e árvores de decisão. Disponível em: <http://www.inf.ufsc.br/%7Ealvares/INE5644/classificacao_arvores.ppt>. Acesso em: 18 dez. 2012.

WIKIPÉDIA. Microrregião de Florianópolis. Disponível em: <http://pt.wikipedia.org/wiki/Microrregiao_de_Florianopolis>. Acesso em: 20 dez. 2012.

ESRI. ESRI Shapefile Technical Description. Disponível em: <http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf>. Acesso em: 11 mai. 2013

WEKA 3.6. WEKA. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/>. Acesso em: 11 mai. 2013