ISSN 1678-6343
DOI: http://dx.doi.org/10.14393/RCG196713

PROPOSTA METODOLÓGICA PARA AVALIAÇÃO DOS REGISTROS SECUNDÁRIOS DE ALAGAMENTOS: UMA ABORDAGEM A PARTIR DE CURITIBA-PARANÁ, BRASIL

Elaiz Aparecida Mensch Buffon
Universidade Federal do Paraná, Departamento de Geografia, Programa de Pós-Graduação em Geografia
Curitiba, PR, Brasil
eambuffon@gmail.com

Mayara Soares de Sousa
Universidade Federal do Paraná, Departamento de Geografia, Programa de Pós-Graduação em Geografia
Curitiba, PR, Brasil
mayara.ssousa93@gmail.com

RESUMO

O primeiro passo da avaliação dos registros de alagamentos é conhecer o tipo de dado, visando identificar passos para a exploração e suas possíveis inter-relações com outros dados. Nesse sentido, o trabalho tem como objetivo propor um encaminhamento metodológico para a avaliação dos registros de alagamentos em Curitiba-PR. Para isso foram utilizados os seguintes dados: registros de alagamentos e precipitação pluviométrica (na escala horária e diária) considerando o intervalo de 2009 a 2012. Os resultados mostraram a existência de registros de casos de alagamentos que não estão associados a eventos de chuva, indicando possíveis erros e/ou inconsistências. Os passos adotados para avaliação dos registros permitiram obter melhores resultados da correlação entre chuva e registros de alagamentos, especialmente, na escala horária. Identificou-se que diferentes valores (mm/hora e mm/24horas) de chuva causam alagamentos, o que precisa ser melhor explorado.

Palavras-chave: Dados secundários. Eventos hidrometeorológicos extremos. Preparação e exploração dos dados. Curitiba.

APPLICATION OF A PROPOSAL FOR DATA MINING OF FLOOD DATA: AN APPROACH FROM CURITIBA-PARANA, BRAZIL

ABSTRACT

The first step of evolution of flood records is to know the type of data, aiming identify steps for exploration and its possible interrelationships with other data. In this sense, the objective of this papper is to propose a methodological approach for the evaluation of flood in Curitiba-PR. For that, we used data from flood and rainfall cases, daily and for hours, considering the interval from 2009 to 2012. The results showed the existence of records of flood events not associated with rain events, indicating possible errors and / or inconsistencies in the data. The steps adopted to evaluate the records allowed better results to be obtained from the correlation between rainfall and flood records, especially on the hourly scale. It was identified that different values (mm/hour e mm/24hours), of rain cause flooding, which needs to be better explored.

Keywords: Second data. Extreme hydrometeorological events. Preparation and exploitation of data. Curitiba.

INTRODUÇÃO

Os registros de alagamentos são importantes fontes de dados para os estudos climáticos relacionados a eventos extremos, ocorridos nas grandes cidades. No entanto, por se tratar de fontes secundárias, a extração de conhecimento a partir destes dados, exige algumas técnicas de tratamento. Neste contexto, algumas etapas do processo de Descoberta de Conhecimento em Banco de Dados (*Knowledge Discovery in Database – KDD*), podem ser utilizadas.

De acordo com Fayyad, Piatetsky-Shapiro e Smyth (1996), KDD refere-se ao processo global de descoberta de conhecimento útil a partir dos dados, sendo o processo não trivial de extração de padrões válidos, novos, potencialmente úteis e compreensíveis a partir de um banco de dados. Os autores descrevem seis etapas do processo de KDD, a saber: 1. Seleção; 2. Pré-Processamento; 3. Transformação; 4. Mineração de Dados; 5. Interpretação / Avaliação; 6. Conhecimento. Recentemente, Maimon e Rocach (2010), adaptaram as etapas do KDD em: 1. Domínio, entendimento e metas do KDD; 2. Seleção e adição de dados; 3. Pré-processamento de dados; 4. Transformação de dados; 5. Escolha da tarefa de mineração de dados apropriada; 6. Escolha do algoritmo de mineração de dados; 7. Emprego do algoritmo de mineração de dados; 8. Evolução e interpretação; 9. Descoberta do conhecimento (visualização e integração).

Conforme Rezende (2005) e Camilo e Silva (2009), existem opiniões divergentes na literatura a respeito dos termos “Mineração de Dados” e “KDD”, alguns autores consideram os termos sinônimos (Fayyad, Piatetsky-Shapiro e Smyth, 1996; Mitchell, 1999; Wei 2003). Enquanto que, outros consideram a Mineração de Dados apenas como um dos passos do processo de KDD, embora seja o passo principal de todo o processo (Mitra, 2002; Sarafis, 2002).

Camilo e Silva (2009) enfatizam que a aplicação da mineração de dados é satisfatória para diversas áreas, inclusive, em tomadas de decisões, filtrando informações relevantes, e fornecendo indicadores de probabilidade. A esse respeito, Cortês *et al*. (2002) inserem a mineração de dados no contexto da Inteligência de Negócios, como uma ferramenta de apoio a tomada de decisão, sendo utilizada principalmente no planejamento estratégico das empresas, dando margem para se pensar também no uso da mineração para políticas públicas, tais como as de planejamento e ordenamento territorial, que necessitam de dados acurados para serem executadas.

Bigolin *et al.* (2003) propõem uma linguagem de consulta que permite automatizar as etapas do processo de descoberta do conhecimento em banco de dados geográficos, atentando para a compreensão dos dados espaciais, que podem ser utilizados no planejamento territorial. Oliveira e Oliveira *et.al.* (2004), também alerta para a necessidade de explorar bases de dados, com vistas a extrair informação/conhecimento para apoio à gestão, tanto de organizações públicas quanto privadas, enfatizando a importância de minerar os dados afim de eliminar as anomalias que podem causar problemas em sua utilização.

No âmbito das Geociências, em especial da Climatologia, o uso de técnicas de Mineração de Dados e KDD, tem se tornado frequente, conforme o exposto nos trabalhos de Boschi et.al. (2011), Pessoa *et. al.* (2012), Ruivo (2013), Bueno (2016), Melanda et. al. (2016), Harwanto et.al. (2017), Talib et. al. (2017), dentre outros. Boschi et. al. (2011), utilizaram as técnicas de mineração de dados para analisar o comportamento espaço-temporal da precipitação pluvial no estado do Rio Grande do Sul para os decênios 1987-1996 e 1997-2006. Pessoa *et. al.* (2012), também fizeram uso da mineração de dados meteorológicos no período de Janeiro e Fevereiro de 2007, para previsão de eventos severos nas regiões do Pantanal Sul Matogrossense, Alto Sorocabana Paulista e parte do Vale do Paranaíba e Litoral Norte. Ruivo (2013), testou os métodos de classificação estatística e árvore de decisão, para analisar as grandes secas do Amazonas ocorridas em 2005 e 2010, e a precipitação extrema ocorrida em Santa Catarina no ano de 2008. Já Bueno (2016), avaliou o potencial de aplicação de técnicas de inteligência artificial (Mineração de Dados e Rede Neural Artificial – RNA) no processo de extração automática de redes de drenagem para a bacia do Rio Mutum-Paraná, no estado de Rondônia. Melanda et. al. (2016), utilizaram a Mineração de Dados, por meio de Árvore de Decisão, para estimar a relação entre a localização de propriedades e o seu valor imobiliário para a cidade de Calgary no Canadá. Harwanto et.al. (2017), utilizaram-se da Mineração de Dados, através da aplicação de técnicas de *Clustering,* Árvore de Decisão e classificação para predição de chuva, voltadas a fins agrícolas na Indonésia. Por fim, Talib et. al. (2017), também fizeram análises de dados meteorológicos a partir da Mineração de Dados, para a cidade de Faisalabad no Paquistão, afim de avaliar possíveis mudanças climáticas.

Em relação aos dados sobre características, fenômenos e processos que acontecem no Brasil, nas escalas federal, estadual e municipal, diversos são os órgãos públicos que possuem a função de disponibilizá-los. Dentre esses órgãos, pode-se citar como exemplos: o Instituto Brasileiro de Geografia e Estatística (IBGE), Agência Nacional das Águas (ANA), Defesa Civil, que disponibilizam de forma gratuita uma gama de dados geográficos constantemente atualizados. Nesse sentido, em função das distintas metodologias de coleta de dados adotadas por cada uma dessas fontes, ressalta-se a importância do tratamento prévio dos dados, que pode ser realizada por meio da aplicação de técnicas de mineração, propiciando assim uma melhor acurácia dos dados para serem utilizados em estudos científicos e relatórios técnicos.

No âmbito dos fenômenos climáticos extremos, os dados referentes aos registros de ocorrências de inundações, alagamentos, deslizamentos, estiagens, erosões, geadas, dentre outros, são oriundos de diversas instituições que utilizam metodologias de trabalho variadas para coleta e armazenamento. Apesar de existir uma integração nacional da Defesa Civil, as instituições estaduais e locais, nem sempre adotam a mesma metodologia de registros da ocorrência de fenômenos climáticos, tais como: inundações, alagamentos, enxurradas.

Neste viés, a presente pesquisa parte do pressuposto de que a execução de algumas etapas de KDD, no tratamento dos registros dos casos de alagamentos disponibilizados pela Defesa Civil, permite eliminar dados com erros e/ou inconsistentes, que podem prejudicar na acurácia de estudos e aplicações. A esse propósito, considera-se dados de alagamento com erros e/ou inconsistentes, aqueles que não estão associados espacialmente e temporalmente com a ocorrência de chuva. Para essa avaliação, construiu-se um caminho metodológico sob a ótica dos passos de KKD apresentado por Fayyad *et al*. (1996). Assim, admite-se como objetivo principal do presente estudo: avaliar os registros secundários de alagamentos da cidade de Curitiba-PR, no período de 2009 a 2012, a partir de algumas etapas de KDD, conforme é apresentado no próximo item deste trabalho. Cabe salientar que neste momento de discussão metodológica, as avaliações não se pautaram em procedimentos automáticos, como pressupõe a Mineração de Dados, mas sim em procedimentos manuais, passíveis de serem automatizados futuramente, através de softwares de mineração de dados, tal como o WEKA, amplamente utilizado na área.

EVENTOS HIDROMETEOROLÓGICOS EXTREMOS: REGISTROS, DEFINIÇÕES E CONCEITOS

No Brasil, a principal fonte de dados das ocorrências de eventos hidrometeorológicos extremos (alagamento, inundação e enxurrada) é a Defesa Civil, que se encontra organizada sob forma de um sistema, e visa integrar ações de governo e da própria comunidade. Sendo assim, a Defesa Civil está estruturada da seguinte forma no território brasileiro: SEDEC - Secretaria Nacional de Defesa Civil (SEDEC); Conselho Nacional de Defesa Civil (CONDEC); Coordenadoria Estadual de Defesa Civil (CEDEC); Coordenadoria Regional de Proteção e Defesa Civil (CORPDEC); e a Coordenadoria Municipal de Defesa Civil (COMDEC). Dentre as mais diversas ocorrências registradas pela Defesa Civil encontra-se os casos de alagamentos, inundações e enxurradas. Mas, essas secretarias e coordenadorias nem sempre estão interligadas, adotando diferentes metodologias de registro e armazenamento dos dados de ocorrências de tais eventos.

No estado do Paraná, a Defesa Civil é constituída de uma coordenadoria estadual, 15 CORPDECs, e mais uma coordenadoria municipal para cada um dos municípios. Sendo assim, os dados de ocorrências de inundações e alagamentos podem ser obtidos em cada uma dessas coordenadorias. Os dados de ocorrências de alagamentos, inundações e enxurradas, no nível de desagregação de cidades e estado, são armazenados em um sistema, que possibilita acesso pela internet (http://www.geo.pr.gov.br/ms4/sisdc/publico/ocorrencias/geo.html). Os dados pontuais (endereço) são fornecidos mediante solicitação nas Coordenadorias Municipais de Defesa Civil.

Entretanto, estes dados nem sempre são diferenciados quanto a sua natureza de ocorrência, ou seja, as ocorrências de inundação, alagamento e enxurrada são agrupadas em uma única categoria denominada de alagamento. Embora estes termos, frequentemente, sejam tratados como sinônimos, o Ministério das Cidades/IPT (2007) conceitua cada um deles de maneira diferente, como segue:

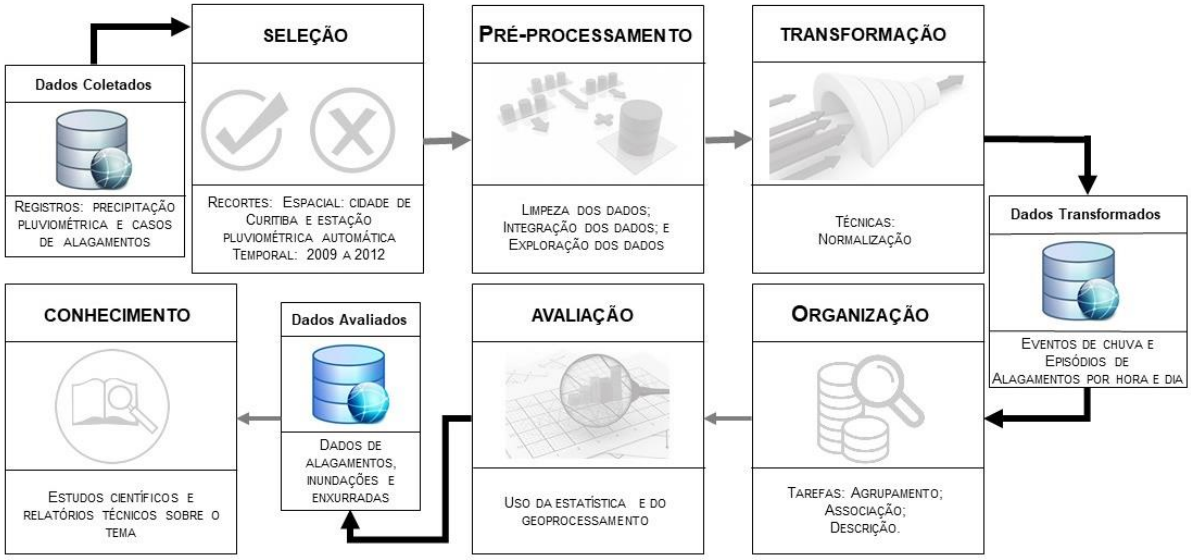
> inundação representa o transbordamento das águas de um curso d'água, atingindo a planície de inundação ou área de várzea. O alagamento é um acúmulo momentâneo de águas em determinados locais por deficiência no sistema de drenagem. A enxurrada é escoamento superficial concentrado e com alta energia de transporte, que pode ou não estar associado a áreas de domínio dos processos fluviais. (MINISTÉRIO DAS CIDADES/IPT, 2007, pgs. 91, 94).

Diversos estudos enfatizam análises em torno das ocorrências de inundações, alagamentos e enxurradas no estado do Paraná, e em razão da ausência da diferenciação desses termos pelos órgãos responsáveis no registro dessas ocorrências, eles apresentam resultados agrupando as ocorrências em um único grupo (LOHMANN, 2011; BUFFON, 2016). Cabe destacar, que essa situação não é restrita ao Paraná, uma vez que Armond (2014) constatou o mesmo problema na ausência da diferenciação nos registros das ocorrências de alagamentos e inundações para o estado do Rio de Janeiro.

METODOLOGIA

A proposta metodológica para avaliação dos registros de alagamentos é embasada no processo de KDD (*Knowledge Discovery in Databases*) que se refere a Descoberta de Conhecimento em Bases de Dados. Esse processo é descrito por Fayyad *et al.* (1996) e envolve vários momentos, sendo esses: seleção, pré-processamento, transformação, mineração e interpretação dos dados. Cada um desses passos, embora não aplicados como um viés de mineração de dados, foram adotados nesta pesquisa. Sendo assim, utilizou-se dos pressupostos teóricos do KDD para elaborar uma proposta metodológica de avaliação dos registros de alagamentos. Essa proposta metodológica é apresenta na figura 1, com uma breve descrição de cada etapa. é adotado como etapa na realização da presente pesquisa, conforme demonstra a figura 1.

Figura 1: Organograma do encaminhamento metodológico adotados no processo de avaliação dos registros de alagamentos.



Elaboração: As autoras (2017).

A realização dos processos apresentados na figura 1 demanda de uma variada gama de dados e de um vasto conjunto de métodos e técnicas, que se apoiam na estatística e no geoprocessamento, afim de se alcançar as respostas para os questionamentos levantados e para atingir os objetivos propostos. Nesse sentido, a seguir são apresentados os dados utilizados nesta pesquisa, e cada um dos passos adotados para cada uma das etapas do encaminhamento metodológico.

OS DADOS

De acordo com Camilo e Silva (2009, p.5) para a execução do processo de avaliação "conhecer o tipo de dados com o qual irá se trabalhar também é fundamental para a escolha do(s) método(s) mais adequado". Assim, na presente pesquisa, utilizaram-se dados de séries de tempos e dados espaciais. Tan, Steinbach e Kumar (2009, p.41) definem os dados de séries de tempos como um tipo especial de dados sequenciais no qual cada registro é uma série de tempo, isto é, uma série de medições feitas no decorrer do tempo. De acordo com os mesmos autores, dados espaciais são aqueles que têm atributos espaciais, como posições ou áreas, como por exemplo os dados climáticos (precipitação, temperatura, pressão) que são coletados para diferentes localizações geográficas. Tan, Steinbach e Kumar (2009, p.42) destacam como aspecto importante dos dados espaciais a auto correlação espacial, isto é, dois pontos próximos tendem a ter valores semelhantes, como por exemplo, para chuva.

COLETA E SELEÇÃO DE DADOS

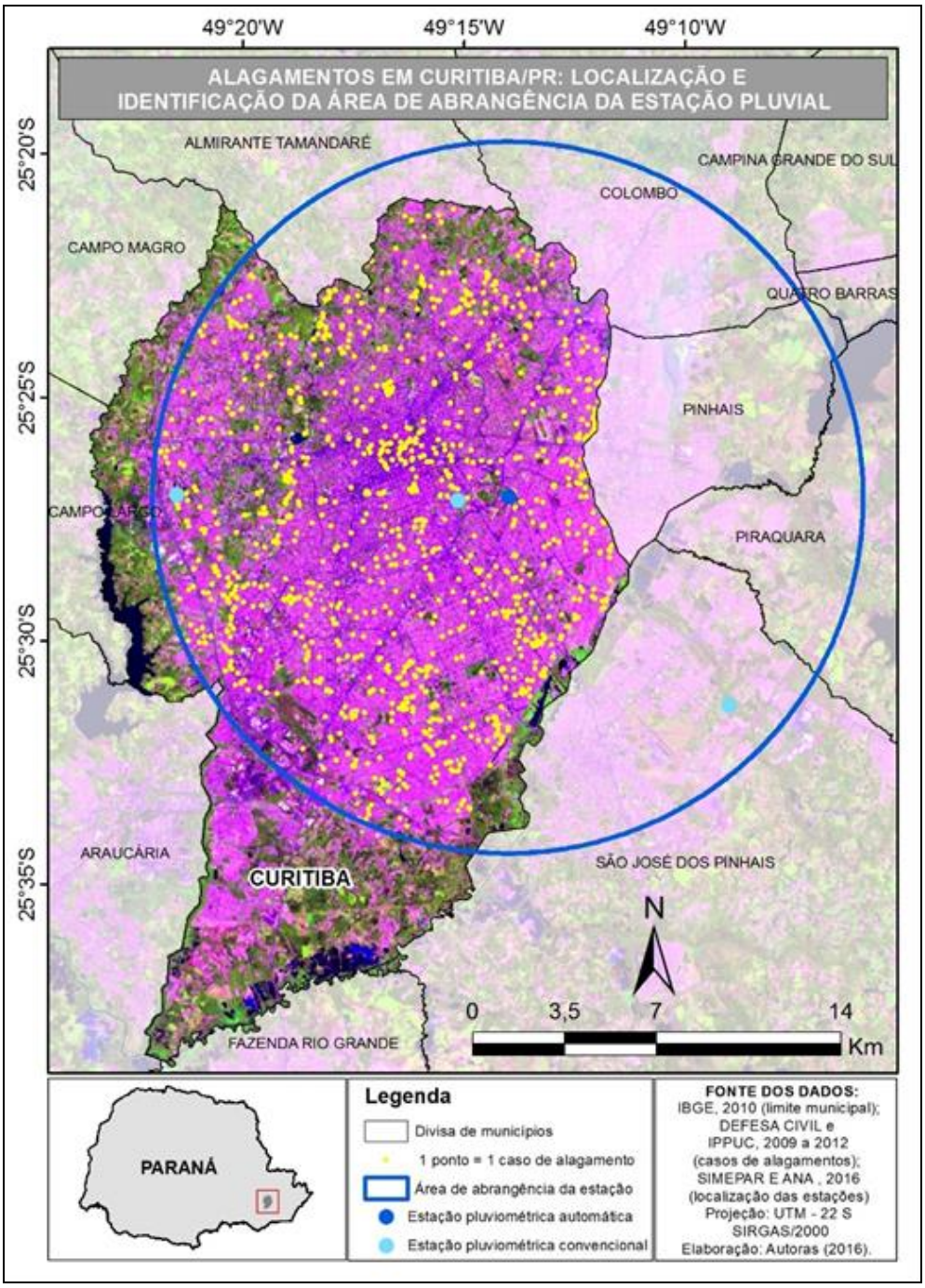
Os dados utilizados nesta pesquisa, registros pontuais dos casos de alagamentos, foram coletados junto à Coordenadoria Municipal de Defesa Civil do município de Curitiba, que foram compilados pelo Instituto de Pesquisa e Planejamento Urbano de Curitiba (IPPUC). De acordo com Lohmann (2011) as informações contidas nos registros são armazenadas por um sistema chamado "SISGESGUARDA" (Sistema de Gerenciamento da Guarda Municipal), sendo esse sistema alimentado por ligações telefônicas recebidas na central de atendimentos e informações da Prefeitura Municipal de Curitiba. As informações armazenadas para cada uma das ocorrências, após o processo de compilação dos registros, são as seguintes: 1) Natureza (somente alagamento); 2) Data de ocorrência (dia/mês/ano/hora); 3) Data de atendimento (dia/mês/ano/hora); 4) Secretaria; 5) Logradouro; 6) Bairro; 7) Regional e, 8) Coordenadas.

Em conjunto com os dados de alagamento utilizou-se dados de precipitação pluviométrica na escala horária, que foram obtidos junto ao Sistema Meteorológico do Paraná (SIMEPAR). Com o objetivo de validar esses dados, buscou-se dados das estações convencionais mais próximas, no que tange a localização da estação automática. Desse modo, por meio da plataforma Sistema de Informações Hidrológicas – *HidroWeb,* da Agência Nacional de Águas, selecionou-se as seguintes estações convencionais: 2549075 – Prado Velho (Curitiba/PR); 2549017 – São José dos Pinhais; 2549126 – Cidade Industrial de Curitiba (Curitiba/PR).

O recorte temporal da pesquisa foi definido com base na disponibilidade de dados, sendo o período selecionado de 2009 a 2012. Embora, existam séries de dados, tanto para os registros de alagamentos e de precipitação, em períodos maiores do que esse selecionado, optou-se por este período em razão da quantidade de dados obtidos na escala horária e dos procedimentos serem manuais. Quanto ao recorte espacial, adotou-se o município de Curitiba/PR, restringindo-se a área de abrangência da principal estação pluviométrica utilizada (SIMEPAR), que permitiu obter dados mais representativos. Essa área é recomendada pela Organização Mundial Meteorológica (WMO, 1994) e equivale a uma área de 575 km² sendo também utilizada em outros estudos no Brasil, tal como o de Correa (2013).

Para caracterizar a localização das estações, tanto a automática (SIMEPAR), quanto as convencionais (ANA/HIDROWEB), elaborou-se um cartograma (Figura 2) com a localização das estações sobrepostas a uma imagem de satélite, a fim de apresentar o contexto urbano em que todas as estações estão inseridas. Além disso, é apresentado um buffer com raio de 13,5 km, que corresponde à área de abrangência da principal estação pluviométrica utilizada. Dentro desse raio são apresentados, em pontos laranjas, todos os registros de casos de alagamentos verificados no período analisado.

Figura 2: Curitiba/PR: Localização dos registros pontuais dos casos de alagamentos (2009 a 2012), das estações pluviométricas automática e convencionais, e da área de abrangência da estação automática.



Elaboração: As autoras (2016).

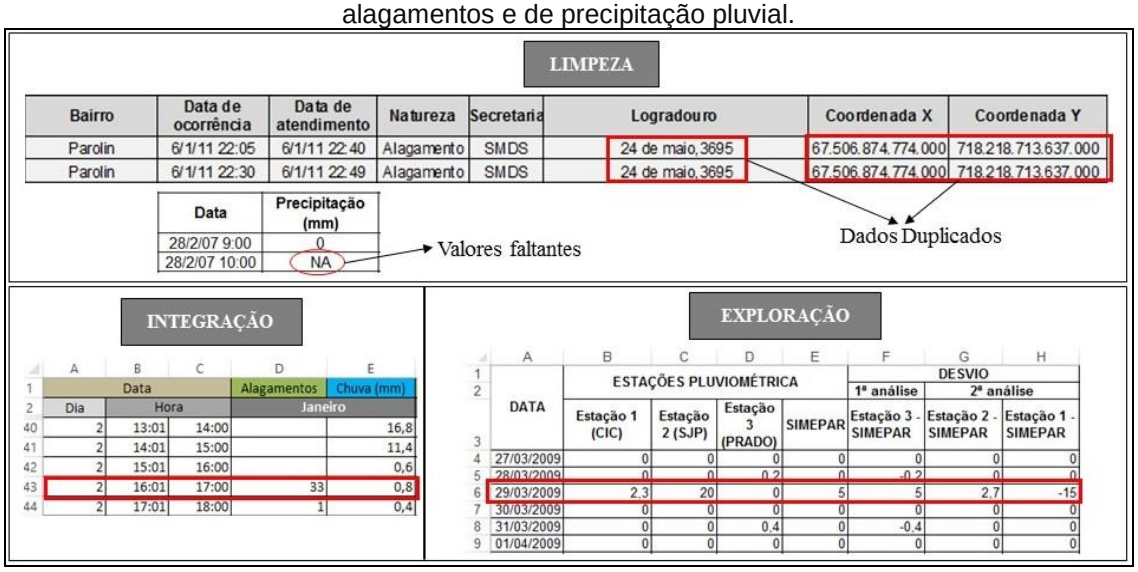
PRÉ-PROCESSAMENTO DOS DADOS

A etapa de pré-processamento dos dados, conforme é apresentada por Mccue (2007) e Olson e Delen (2008), possui fundamental importância, uma vez que é nesse momento que se prepara os dados. Para a visualização integrada dos dados utilizados nesta pesquisa utilizou-se do programa *Microsoft Excel*, por meio de planilhas eletrônicas, que permitiu gerar gráficos e tabelas. A partir disso, buscou-se realizar análises dos dados, tanto para os registros de casos de alagamentos, bem

como, os registros de precipitação pluviométrica, que foram baseadas nas indicações de técnicas de Han e Kamber (2006) e Tan *et al.* (2009) para limpeza, integração, exploração e, quando necessário a transformação dos dados.

A figura 3 permite visualizar exemplos das técnicas adotadas em cada uma das tarefas de pré-processamento dos dados. Em termos gerais, na tarefa de limpeza, buscou-se identificar os atributos duplicados nos registros de alagamento, e valores faltantes dos registros de precipitação. Na tarefa de integração dos dados de alagamento com os de precipitação, construiu-se uma única planilha eletrônica, para conter os valores de precipitação e os casos de alagamentos, por dia e hora de ocorrência.

Figura 3: Exemplos das tarefas de pré-processamento dos dados de registros de casos de alagamentos e de precipitação pluvial.



Elaboração: As autoras (2017).

Ainda, na etapa de pré-processamento conforme demonstra a figura 3, realizou-se a exploração dos dados de registros de precipitação, com o intuito de identificar dados inconsistentes, que podem interferir na análise. Para essa exploração, buscou-se dados de registros de outras três estações convencionais, que estão localizadas a aproximadamente 2.000, 12.700, 11.500 metros da estação pluviométrica automática, principal estação utilizada na pesquisa. Dessa forma, em um primeiro momento fez-se necessário a padronização dos dados. Os dados horários da estação do SIMEPAR são registrados em hora mundial UTC, para isso realizou-se a conversão para UTC−3 ou Horário de Brasília – hora oficial do Brasil, ou seja, é a diferença de fuso horário que subtrai três horas do Tempo Universal Coordenado, e no período do horário de verão adotou-se UTC-2.

Os dados das estações convencionais (ANA/HIDROWEB) são coletados diariamente às sete horas da manhã (Horário de Brasília), sendo assim o valor “k” de precipitação pluvial registrado na data “y”, pertence ao total precipitado entre às sete horas da manhã da data “x” até às sete horas da manhã da data “y”. Nesse sentido, os dados horários da estação automática (SIMEPAR) foram somados, a fim de, de obter o total precipitado diariamente de acordo com a escala de horário das estações convencionais.

Após essa padronização, calculou-se o desvio (mm) dos registros diários de precipitação das duas estações mais próximas, a saber: automática (SIMEPAR) e a convencional 2549075 – Prado Velho (ANA). A fim de, identificar os extremos máximos dos desvios diários foi calculado a média do desvio diário (0,1 mm) e o desvio padrão (3,91 mm). Portanto, adotou-se como valores discrepantes, os que se encontram acima de 3,92 mm, que corresponde à adição da média e do desvio padrão.

Logo, os dias que apresentam valores com desvio acima de 3,92 mm (tanto em valor positivo como negativo), foram avaliados com as estações vizinhas (2549017 – São José dos Pinhais; 2549126 – Cidade Industrial de Curitiba) para verificar qual das duas estações analisadas anteriormente

apresentam menor discrepância dos dados quando comparadas com essas outras duas estações. No total, a estação automática (SIMEPAR) apresentou 142 dias com dados considerados como discrepantes dos valores registrados na estação convencional mais próxima. Dentro desses 142 dias, foram considerados prejudiciais a pesquisa os dados em que uma estação registrou precipitação e a outra registrou 0 mm. Dentro dessas categorias de análise, 13 dias de registros da estação pluviométrica automática foram descartados da análise, em razão de ser discrepante das demais estações.

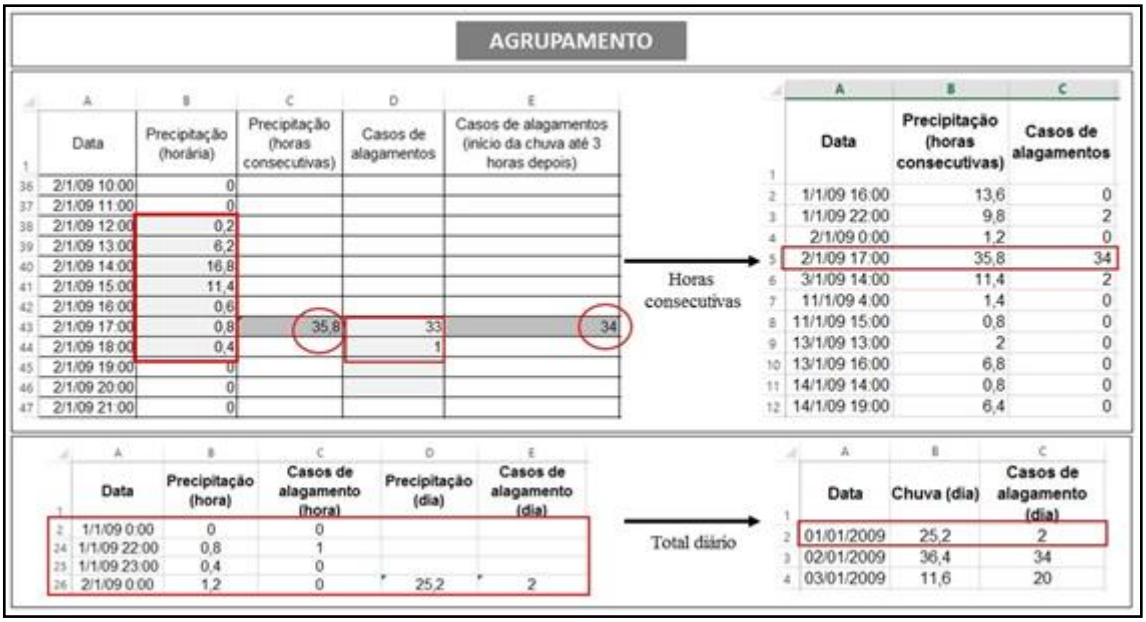
Após essas tarefas, ainda dentro desse momento de pré-processamento dos dados, procedeu-se com a transformação dos dados, que consistiu na aplicação da técnica de suavização. Essa técnica refere-se a exclusão dos valores considerados errados e/ou inconsistentes a partir das tarefas realizadas anteriormente (CAMILO e SILVA, 2009).

TAREFAS DA AVALIAÇÃO DOS REGISTROS

Han e Kamber (2006) definiram, no âmbito da mineração de dados, algumas tarefas importantes, tais como: Agrupamento, Classificação, Associação, Descrição, Estimativa e Predição. Assim, apoiou-se nessas tarefas, para realizar a avaliação dos dados, mesmo que neste trabalho não seja aplicado efetivamente um processo de mineração de dados. Portanto, coloca em evidência as seguintes tarefas para avaliação dos registros de alagamentos:

Tarefa de Agrupamento: refere-se ao agrupamento de registros, observações ou casos em classes de objetos semelhantes. Um agrupamento é um conjunto de registros que são semelhantes entre si, e diferentes de registros em outros agrupamentos (HAN e KAMBER, 2006, p.16). Esse momento é essencial para definição do nível de desagregação dos dados, que pode ser tanto temporal como espacial. Nesta pesquisa, definiram-se dois níveis temporais de desagregação de dados: dia (período de 24 horas) e horário (horas consecutivas com chuva). A figura 4 exemplifica o processo realizado no agrupamento dos valores horários de precipitação e de casos de alagamento, que possibilitaram a transformação dos dados.

Figura 4: Exemplo do agrupamento de dados para obtenção dos dados em dois níveis temporais de desagregação: horas consecutivas de chuva e dia.



Elaboração. As autoras (2017).

Tarefa de Associação: trabalho de encontrar quais atributos estão relacionados, buscando descobrir regras para quantificar a relação entre dois ou mais atributos (HAN e KAMBER, 2006, p.17). Considerando o exposto pelos autores, adotou-se a seguinte regra: se existe registros de casos de

alagamento em determinado dia e hora, então nesse mesmo dia, e em até 3 horas antes do caso de alagamento, deve existir registro de precipitação pluviométrica. Os dados que não se encaixaram nessa regra foram eliminados da análise, por serem considerados inconsistentes.

Tarefa de Descrição: descrevem padrões e tendências dentro dos dados, sugerindo explicações para estes padrões e tendências. Os padrões necessitam ser claros e passíveis de interpretação intuitiva e explicação (HAN e KAMBER, 2006, p.11). Assim, realizou-se o cálculo de porcentagem dos valores de precipitação pluvial verificada em horas consecutivas e seus respectivos casos de alagamentos, dentro de classes determinadas. As classes para os totais de chuva em horas consecutivas aqui adotadas foram: de 0,5 a 5 mm; de 5,1 a 10 mm; de 10,1 a 20 mm; de 20,1 a 30 mm; de 30,1 a 40 mm; de 40,1 a 50 mm; de 50,1 a 70 mm; de 70,1 a 90 mm; e de 100 a 150 mm.

TÉCNICAS DE ANÁLISES DOS REGISTROS DE ALAGAMENTOS PRÉ E PÓS AVALIAÇÃO

Calculou-se a correlação de Pearson que permite verificar a associação entre as variáveis analisadas (dados de chuva e de alagamentos) para avaliar a expressividade dos processos de preparação, exploração e organização dos dados antes e após as exclusões dos dados com erros e/ou inconsistentes. Também, utilizou-se o valor do R^2 (ANDRIOTTI, 2005) que apresenta a parte da variância total de precipitação pluvial e alagamentos, que pode ser explicada pela sua relação linear. Esse valor permite conhecer a proporção da variação total dos casos de alagamentos, explicada pelo ajuste da regressão, permitindo apresentar um valor do coeficiente de determinação.

Tanto para os dados horários, bem como, para os diários, foram calculados dois coeficientes de determinação (r^2), um primeiro com todos os registros de alagamentos (sem avaliação de dados), e um segundo com os registros resultantes após a aplicação da avaliação. Os registros, em ambas as escalas (diária e horária), que foram excluídos das análises foram espacializados, com o intuito de verificar se existe um padrão espacial da ocorrência de registros com erros e/ou inconsistentes.

RESULTADOS E DISCUSSÕES

Considerando como pressuposto a regra fundamental adotada neste estudo, de que os registros de casos de alagamento dependem da ocorrência de chuva, ou seja, registro de precipitação superior a 0,5 mm buscou-se analisar a existência de relação causal *a priori* e *a posteriori* a realização dos processos da avaliação dos registros de alagamento. Outros estudos tais como os de Pradhan *et al.* (2008) e Svoray *et al.* (2011) também trabalharam na perspectiva semelhantes, porém adotando diretamente o conceito de mineração de dados de chuvas, relacionados a outros dados, como solos e topografia, afim de mapear áreas de risco de deslizamento de terras e predição de queda de barrancos, enfatizando o uso de dados de chuva minerados para identificação dessas áreas. Para avaliação dos processos adotados para avaliação dos registros de alagamentos nesta pesquisa, obteve-se como primeiro resultado, através da correlação de Pearson, um aumento na correlação entre os registros de chuva e de casos de alagamento, tanto na escala horária como diária, após a realização do processo de avaliação dos registros (Tabela 1). O aumento mais significativo ocorreu na escala horária, com uma melhora de 24,99% no valor de correlação (Tabela 1).

Tabela 1: Valores da correlação de Pearson para análise da associação dos registros de chuva e alagamento.

Escala temporal	Registros não avaliados e não minerados	Registros avaliados e minerados
Horária	0,1303	0,5214
Diária	0,4756	0,5427

Elaboração: As autoras (2017).

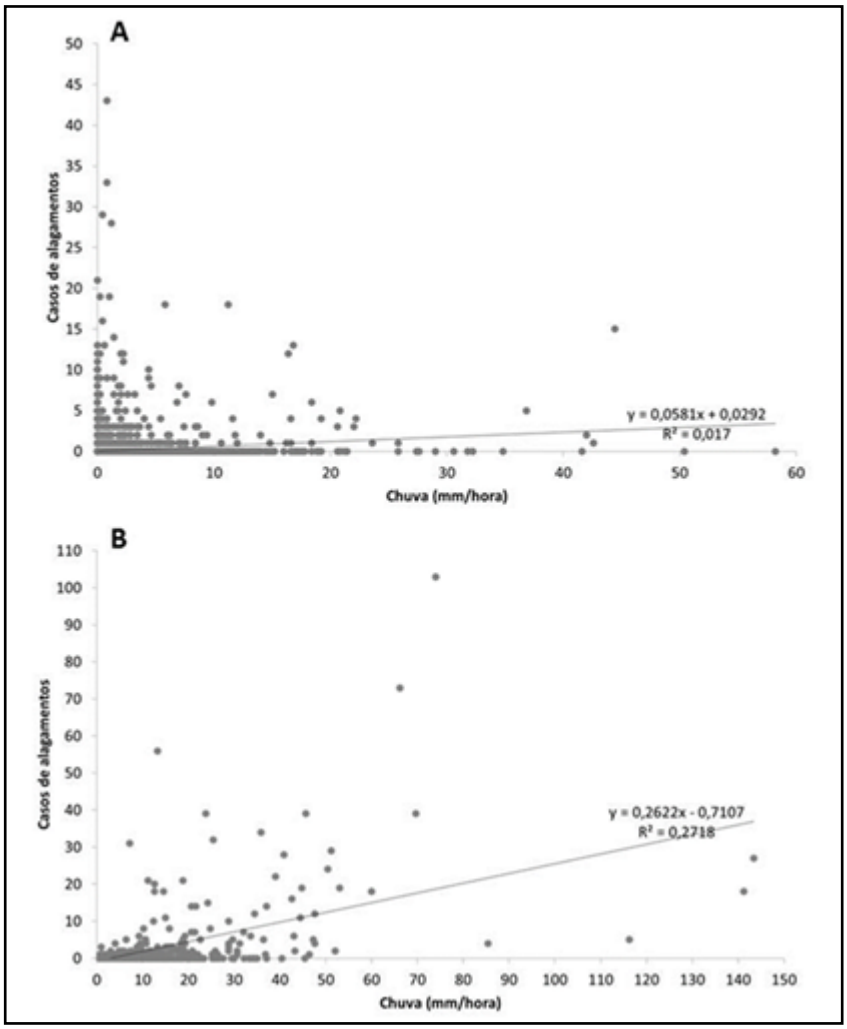
Deste modo, os valores obtidos e apresentados na tabela 1 demonstram a importância da realização dos processos da avaliação de dados, especialmente, nos casos em que a escala adotada visa um maior nível de detalhe na análise, como no caso dos registros de alagamentos na escala horária. Andriotti (2005) apresenta uma quantificação dos valores numéricos de correlação transformando em

valores qualitativos, e de acordo com essa classificação, os dados na escala horária passam da condição de fraca correlação para regular após a execução da avaliação dos registros, com exclusão daqueles com possíveis erros e/ou inconsistentes. Quanto aos dados na escala diária, a correlação permanece dentro da mesma categoria de regular, visto que nesta escala são agrupados os valores registrados por horas em um total de 24h (dia).

Os dados apresentados na tabela 1 referem-se a uma medida da intensidade da relação linear entre duas variáveis, entretanto, não é possível afirmar que quanto maior for o valor de precipitação maior deve ser o número de registros de casos de alagamento. Considerando essa problemática, e a afirmação da necessidade de precaução na interpretação do coeficiente de correlação linear apontada por Andriotti (2005, p. 69), em que essa é "uma interpretação puramente matemática, não implica a existência de causa e efeito", buscou-se obter o valor do coeficiente de determinação (r^2), ou seja, o quantitativo da variância total que pode ser explicada pela sua relação linear.

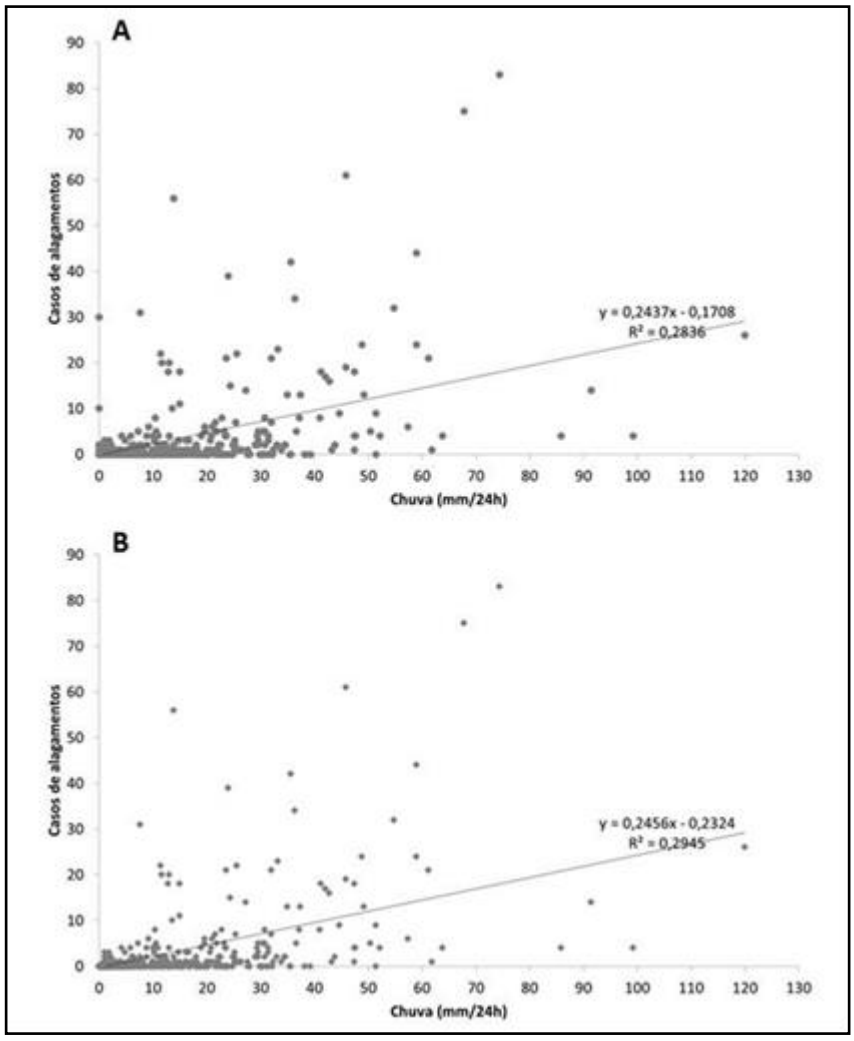
O coeficiente de determinação, por meio da representação da equação linear e seu valor apresentado nas figuras 5 e 6, permite concluir que existe uma relação positiva entre o registro de chuva e casos de alagamentos, entretanto, nem sempre são proporcionais, em razão de dinâmicas ambientais e sociais dos lugares que são afetados por eventos extremos de chuva que causam alagamentos. Dessa forma, as figuras indicam que parte da variação total dos registros permanecem não explicada, entretanto, o valor não explicado pode ser menor quando se aplica a avaliação dos registros.

Figura 5: Relação entre o total de chuva (mm/hora) e os registros de alagamentos (casos/hora). Gráfico A: Registros não avaliados. Gráfico B: Registros avaliados.



Elaboração: Autoras (2017).

Figura 6: Correlação entre o total diário de chuva (mm/24h) e os registros de alagamentos (casos/24h). Gráfico A: Registros não avaliados. Gráfico B: Registros avaliados.

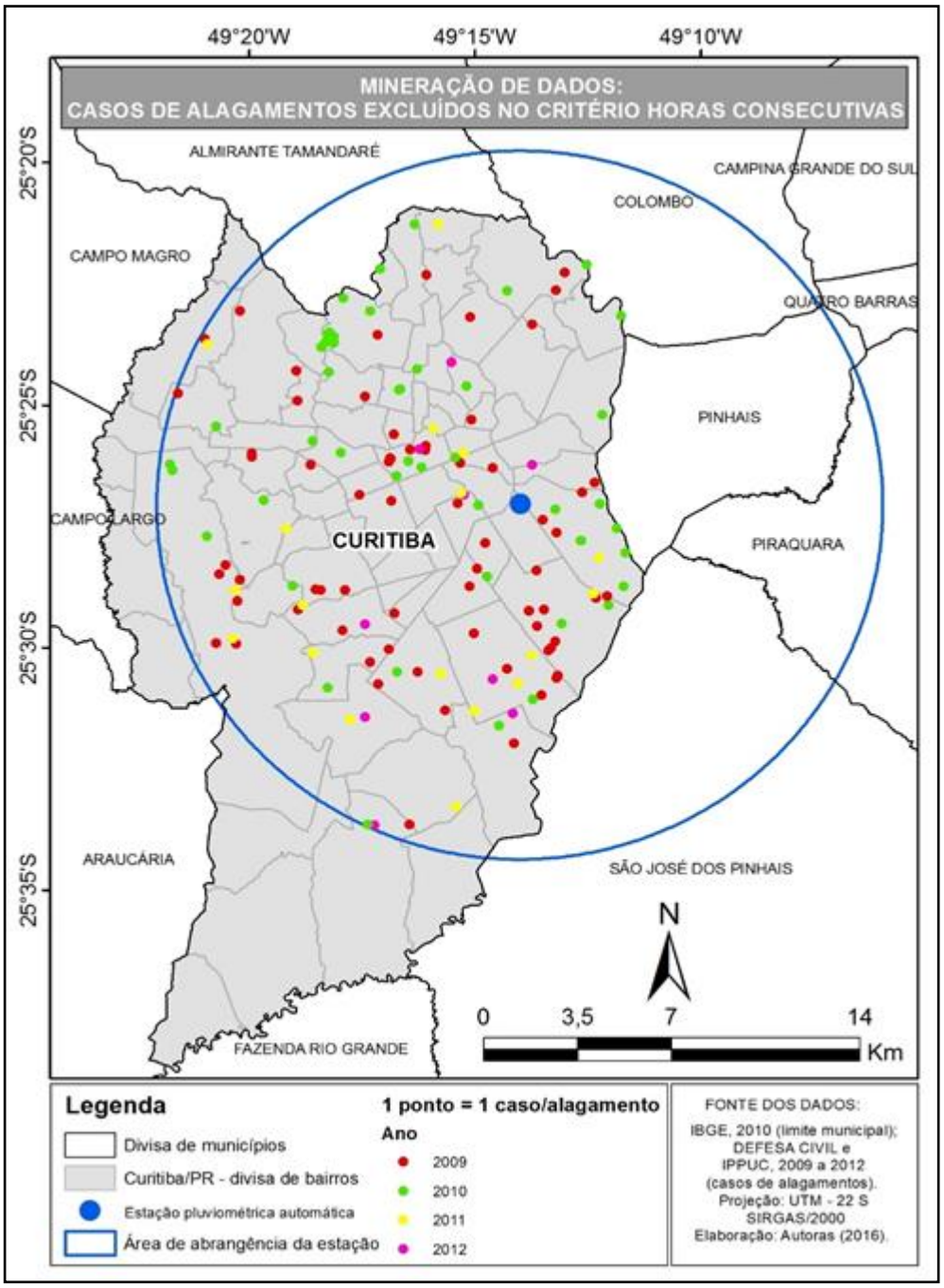


Elaboração: Autoras (2017).

Ainda, dentro do conjunto de avaliação dos registros de alagamentos, observou-se por meio da espacialização dos registros de casos identificados com erros e/ou inconsistentes, que não existe uma concentração em determinada área da cidade (Figuras 7 e 8). Entretanto, ao observar os registros que se inserem dentro da categoria de que não apresentam chuva no dia anterior e, também, no dia da ocorrência, existe uma concentração de casos na porção periférica da cidade. Esses registros concentrados na porção periférica correspondem a 14,08% do total de registros com erros na escala diária.

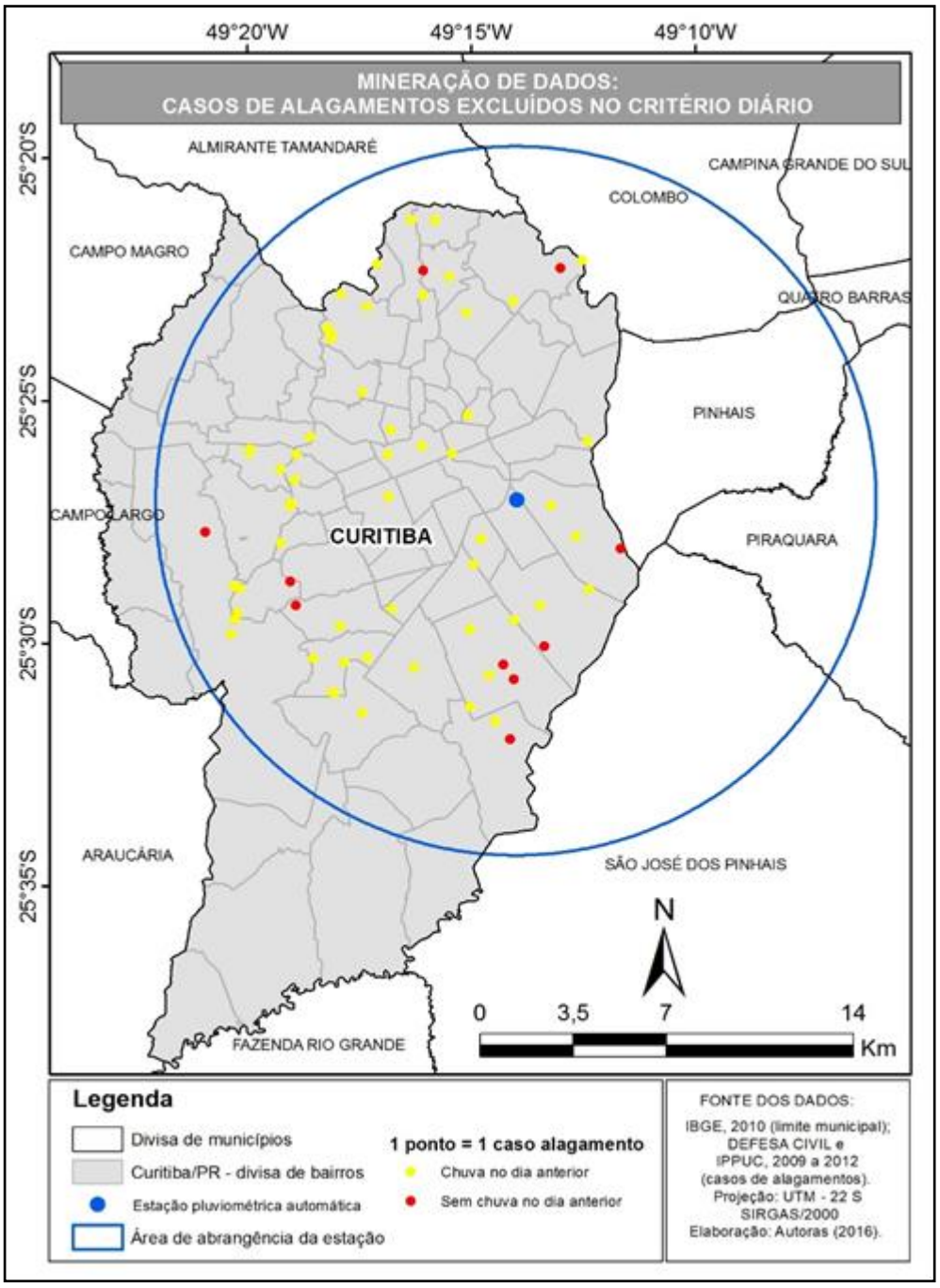
A figura 7 permite afirmar que o número de registros de alagamento com erros e/ou inconsistentes é decrescente ao longo dos anos. Uma análise descritiva dos dados demonstra que 48,99% dos registros com erros são datados de 2009, 32,88% no ano de 2010, 12,08% são de 2011, e 6,04% referem-se ao ano de 2012. Esses valores são base para afirmar que houve uma melhoria nos registros de alagamentos na cidade de Curitiba, dentro do período analisado.

Figura 7: Registros de alagamento na escala horária descartados a partir da aplicação dos processos da avaliação dos registros



Elaboração: As autoras (2016).

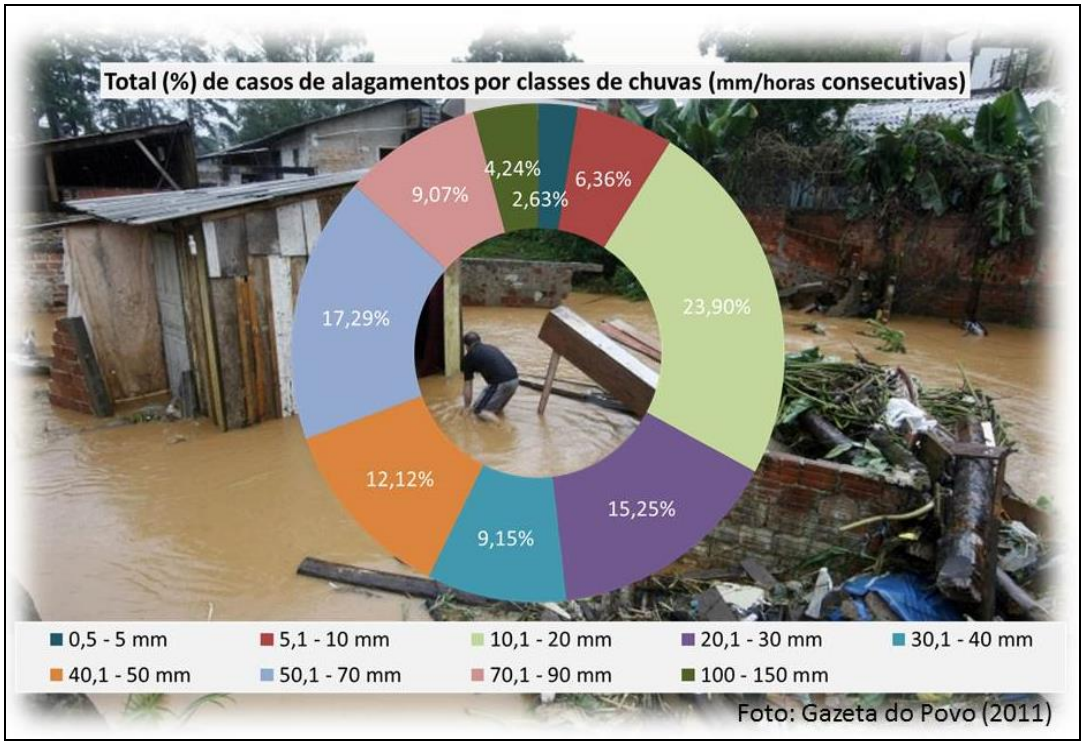
Figura 8: Registros de alagamento na escala diária descartados a partir da aplicação dos processos da avaliação dos registros.



Elaboração: As autoras (2016).

Partindo da análise exploratória dos resultados obtidos com os processos da avaliação dos registros, e considerando que a descrição é uma das tarefas desse processo, construiu-se a figura 9, que tem por finalidade a identificação dos valores totais de chuva, por horas consecutivas, que desencadearam casos de alagamentos. Sendo assim, dentre todas as classes demostradas na figura 9, merecem destaque os eventos pluviais entre 10 e 20 mm que são responsáveis por 23,90% do total de registros de alagamentos. Além disso, verificou-se que houve registros de alagamento durante chuvas com 0,5 a 5 mm em horas consecutivas, e dentre o total de registros de alagamentos, 30,6 % ocorreram em situações com eventos pluviais acima de 50 mm em horas consecutivas (Figura 9).

Figura 9: Descrição dos dados de alagamento e de chuva através da avaliação por classes.



Elaboração: As autoras (2016).

A figura 9, que é parte integrante da discussão dos registros avaliados, se apresenta como relevante em face da importância de diferenciação da natureza dos registros de alagamentos, inundação e enxurrada, discutidos anteriormente. Ao conhecer os padrões de chuva, que são mais frequentes para provocar casos de alagamentos, inundações e enxurradas em determinadas áreas, é possível iniciar um processo de prevenção, especialmente no que se refere ao monitoramento, frente aos impactos decorrentes de tais casos. Nesse sentido, quando os registros não são diferenciados em relação a natureza, a avaliação dos registros, com destaque para a tarefa de descrição, se apresenta como uma eficaz técnica para explorar a possível natureza de tais registros.

Witten e Frank. (2005), Bramer (2007), Olson e Delen (2008) e Camilo e Silva (2009), apresentam outras áreas onde a mineração de dados, ou também, encaminhamentos próprios para avaliação dos dados, são aplicadas de forma satisfatória. A título de exemplo, citam: na área de segurança (detectando atividades criminais e áreas de risco), saúde (com indicação de diagnósticos mais precisos), e na área de gestão de informação dentro das organizações ao servir como ferramenta para tomada de decisão. No contexto do uso de encaminhamentos metodológicos para avaliação dos dados, dados para tomada de decisão, Guo e Mennis (2009) discutem como importantes atividades para a tomada de decisão, uma vez que análises da qualidade dos dados espaciais e temporais são de extrema importância na aplicação do conhecimento.

CONSIDERAÇÕES FINAIS

A proposta metodológica para avaliação dos registros de alagamentos aplicada nesta pesquisa, demandou de alterações em relação ao encaminhamento original proposto por Fayyad *et al.* (1996) no KDD, especialmente, no que concerne ao uso de softwares automáticos para elaboração dos processos. Constatou-se que, neste momento, de discutir propostas metodológicas, essas alterações são essenciais, visto que o objetivo do trabalho foi discutir técnicas e etapas que podem futuramente se tornar automáticas através de softwares para avaliação dos dados secundários. No que concerne ao processo de preparação e exploração dos dados, observou-se que dados secundários de fontes distintas, nem sempre estão integrados, portanto, torna-se necessário realizar a integração de dados em uma única escala espacial e temporal para possibilitar análises comparativas. A análise de um

fenômeno que é dependente de outro, necessita da exploração dos dados que dão sustentação a problemática, no caso desta pesquisa relacionada ao clima, utilizou-se dados de estações pluviométricas localizadas nas proximidades.

Assim, a aplicação da tarefa de associação, com base na regra criada para esse estudo "se existe registros de casos de alagamento em determinado dia e hora, então nesse mesmo dia, e em até 3 horas antes dos casos de alagamento, deve existir registro de precipitação pluviométrica", permitiu identificar dados com erros e/ou inconsistentes. Dessa forma entende-se que a utilização desses dados sem a execução do processo de avaliação dos registros pode prejudicar futuras análises e aplicações, como por exemplo, aquelas associadas a tomada de decisão, ao estabelecimento de correlações espaciais, a criação de projetos, dentre outras.

Os resultados, também, permitem concluir que quando se trabalha com maior nível detalhe nos atributos dos dados, é fundamental realizar procedimentos de agrupamento, visto que esse permite identificar semelhanças e diferenças entre os dados, de forma que esses são transformados. Aqui destaca-se, a dificuldade de caracterizar um tempo médio para ocorrência de eventos hidrometeorológicos extremos após a data e horário de início da precipitação. Essa dificuldade está associada, tanto devido à ausência de estudos desta problemática, bem como a complexidade de tal demanda frente a uma dinâmica urbana de escoamento superficial. Sendo assim, sugere-se novos testes com outros valores de tempo médio de escoamento pluvial superficial, para verificar, por exemplo, se o aumento de tal valor pode resultar em uma correlação maior entre chuva e registros de alagamentos.

Para além da identificação de dados inconsistentes e com erros, a avaliação dos registros possibilitou, a partir da tarefa de descrição, caracterizar padrões e tendências dentro desses dados. A realização do presente estudo demonstrou que os registros de alagamentos em Curitiba não devem ser utilizados, conforme são coletados na íntegra, para retratar em um único conjunto áreas de inundações, visto que, alguns casos estão inseridos dentro das classes de 0,5 a 5 mm, 5,1 a 10 mm. Sendo que, esses valores de precipitação dificilmente podem caracterizar um evento de inundação. Mesmo com essa análise da descrição dos registros, torna-se complexa a diferenciação da natureza do episódio, todavia, entende-se que todos esses registros, não devem ser considerados na análise espacial como eventos impactantes para a sociedade, visto que suas magnitudes são diversas. Como perspectivas de trabalhos futuros, coloca-se a necessidade de investigação com séries temporais mais extensos, e com dados secundários provenientes de diferentes fontes. Além disso, a necessidade de utilização de softwares para criação de procedimentos automáticos no processo de avaliação dos dados.

REFERÊNCIAS

ANDRIOTTI, J. L. S. **Fundamentos de Estatística e Geoestatística**. São Leopordo – RS: Editora da Universidade Do Vale do Rio dos Sinos, 2005.

ARMOND, N. B. **Entre Eventos e Episódios: as excepcionalidades das chuvas e os alagamentos no espaço urbano do Rio de Janeiro**. 239f. Dissertação (Mestrado em Geografia) - Faculdade de Ciências e Tecnologia, Universidade Estadual Paulista Júlio de Mesquita Filho, Presidente Prudente, 2014.

BIGOLIN, N. M.; BOGORNY, V.; ALVARES, L. O. **Uma Linguagem de Consulta para Mineração de Dados em Banco de Dados Geográficos Orientado a Objetos.** In: Conferencia Latino-Americana de Informática. 2003. p. 23-35.

BOSCHI, R. S.; OLIVEIRA, S. R. D. M.; ASSAD, E. D. **Técnicas de mineração de dados para análise da precipitação pluvial decenal no Rio Grande do Sul.** Engenharia Agrícola, 2011.BRAMER, M. **Undergraduate Topics in Computer Science - Principles of Data Mining.** Springer, 2007.

BUENO, L. F. **Inteligência Artificial Aplicada à melhoria da acurácia do mapeamento de redes de drenagem.** 149f. Tese (Doutorado em Geografia) – Setor de Ciências da Terra, Programa de Pós-Graduação em Geografia, Universidade Federal do Paraná, Curitiba, 2016.

BUFFON, E. A. M. **A leptospirose humana no AU-RMC (Aglomerado Urbano da Região Metropolitana de Curitiba/PR) – risco e vulnerabilidade socioambiental**. Dissertação (Mestrado

em Geografia) – Setor de Ciências da Terra, Programa de Pós-Graduação Geografia, Universidade Federal do Paraná, Curitiba, 2016, 171f.

CAMILO, C. O.; SILVA, J. C. **Mineração de dados: Conceitos, tarefas, métodos e ferramentas.** Universidade Federal de Goiás (UFC), p. 1-29, 2009.

CIOS, K. J; PEDRYCZ, W; SWINIARSKI, R. W; KURGAN, L. A. **Data Mining – A Knowledge Discovery Approach.** Springer, 2007.

CORREA, G. G. M. **Distribuição espacial e variabilidade da precipitação pluviométrica na bacia do Rio Piquiri-PR**. Dissertação (Mestrado em Ciências) – Faculdade de Filosofia, Letras e Ciências Humanas, Programa de Pós-Graduação Geografia Física, Universidade de São Paulo, São Paulo, 2013, 104p.

CÔRTES, S. C; PORCARO, R. M.; LIFSCHITZ, S. **Mineração de dados-funcionalidades, técnicas e abordagens.** PUC, 2002.

FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. **From Data Mining to Knowledge Discovery in Databases.** American Association for Artificial Intelligence, 1996.

GUO, D; MENNIS, J. **Spatial data mining and geographic knowledge discovery - An introduction.** Computers, Environment and Urban Systems, n. 33, p. 403-408, 2009.

HAN, J; KAMBER, M. **Data Mining - Concepts and Techniques.** Morgan Kaufmann Publishers, Inc, 2006.

HERWANTO, R.; PURNOMO, R. F. **RAINFALL PREDICTION USING DATA MINING TECHNIQUES**. In: Prosiding International conference on Information Technology and Business (ICITB). 2018. p. 188-193.

LAROSE, D. T. **Discovering Knowledge in Data: An Introduction to Data Mining.** John Wiley and Sons, Inc, 2005.

LOHMANN. M. **Regressão logística e redes neurais aplicadas à previsão probabilística de alagamentos no município de Curitiba, PR**. 230f. Tese (Doutorado em Geografia) – Programa de Pós-Graduação em Geografia, Universidade Federal do Paraná, Curitiba, 2011.

MAIMON, O.; ROKACH, L. **Data mining and knowledge discovery handbook.** Springrer, 2010. https://doi.org/10.1007/978-0-387-09823-4

MCCUE, C. **Data Mining and Predictive Analysis - Intelligence Gathering and Crime Analysis**. Elsevier, 2007.

MELANDA,E.; HUNTER, A.; BARRY,M. **Identification of locational influence on real property values using data mining methods,** Cybergeo : European Journal of Geography [En línea], Sistemas, Modelística, Geoestadísticas, documento 771, Publicado el 04 febrero 2016, consultado el 30 abril 2018. URL: http://journals.openedition.org/cybergeo/27493.

MITCHELL, T. M. **Machine learning and data mining**. Communications of the ACM 42(11), 1999. https://doi.org/10.1145/319382.319388

MINISTÉRIO DAS CIDADES / INSTITUTO DE PESQUISAS TECNOLÓGICAS – IPT – **Mapeamento de riscos em encostas e margens de rios.** Brasília: Ministério das Cidades; Instituto de Pesquisas Tecnológicas – IPT, 2007. 176 p.

MITRA, S. K. Pal, P. M. **Data mining in soft computing framework: A survey.** IEEE Transactions on Neural Networks 13(1), 3–14, 2002. https://doi.org/10.1109/72.977258

OLIVEIRA, P.; RODRIGUES, F.; HENRIQUES, P. **Limpeza de dados: Uma visão geral.** Data Gadgets, p. 39-51, 2004.

OLSON, D. L; DELEN, D. **Advanced Data Mining Techniques.** Springer, 2008.

PESSOA, A. S. A., LIMA, G. R. T., DA SILVA, J. D. S., STEPHANY, S., STRAUSS, C., CAETANO, M., & FERREIRA, N. J. **Mineração de dados meteorológicos para previsão de eventos severos.** Revista Brasileira de Meteorologia, v. 27, n. 1, 2012.

PRADHAN, B; LEE, S. **Landslide risk analysis using artificial neural network model focussing on different training sites**. International Journal of Physical Sciences, v. 4, p.1-15, 2009.

REZENDE, S. O. **Mineração de Dados.** XXV Congresso da Sociedade Brasileira de Computação, 2005.

RUIVO, H. M. **Metodologias de Mineração de Dados em Análise Climática.** Tese (Doutorado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2013. 124p.

SARAFIS, A. M. S. ZALZALA, P. W. T. **A genetic rule-based data-clustering toolkit.** In Congress on Evolutionary Computation (CEC), Honolulu, USA, 2002.

SVORAY, T. et al**. Predicting gully initiation: comparing data mining techniques, analytical hierarchy processes and the topographic threshold**. Earth Surf. Process. Landforms, 2011.

TALIB, M. R., ULLAH, T., SARWAR, M. U., HANIF, M. K., & AYUB, N. **Application of Data Mining Techniques in Weather Data Analysis.** INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND NETWORK SECURITY, 17(6), 22-28, 2017.

TAN, P. N.; STEINBACH, M.; KUMAR, V. **Introdução ao Datamining Mineração de Dados**. Rio de Janeiro: Editora Ciência Moderna Ltda, 2009.

VAREJÃO-SILVA, M.A. **Meteorologia e Climatologia**. Versão Digital, Recife. 2. ed, 2006.

WEI, J. M. **Rough set based approach to selection of node**. International Journal of Computational Cognition 1(2), 25–40, 2003.

WITTEN, I. H; FRANK, E. **Data Mining - Practical Machine Learning Tools and Techniques.** Elsevier, 2005.

WORLD METEOROLOGICAL ORGANIZATION. **Guide to Hydrological Practices, Data Acquisition and Processing, Analysis, forecasting and other Applications**. 5 ed. N 168, Geneva: 1994, p.259.

Recebido em: 28/07/2017
Aceito para publicação em: 10/09/2018