UNIVERSIDADE FEDERAL DO PARANÁ

LUIS FERNANDO BUENO

INTELIGÊNCIA ARTIFICIAL APLICADA À MELHORIA DA ACURÁCIA DO MAPEAMENTO DE REDES DE DRENAGEM

CURITIBA

2016

LUIS FERNANDO BUENO

INTELIGÊNCIA ARTIFICIAL APLICADA À MELHORIA DA ACURÁCIA DO MAPEAMENTO DE REDES DE DRENAGEM

Tese apresentada como requisito parcial à obtenção do grau de Doutor em Geografia, no Curso de Pós-Graduação em Geografia, Setor de ciências da Terra, Universidade Federal do Paraná.

Orientador: Prof. Dr. Tony Vinicius Moreira Sampaio

CURITIBA

2016

B928i Bueno, Luis Fernando
Inteligência artificial aplicada à melhoria da acurácia do mapeamento de redes de drenagem / Luis Fernando Bueno. – Curitiba, 2016.
148 f. : il. color. ; 30 cm.

Tese - Universidade Federal do Paraná, Setor de Ciências da Terra, Programa de Pós-Graduação em Geografia, 2016.

Orientador: Tony Vinicius Moreira Sampaio .
Bibliografia: p. 133-148.

1. Mineração de dados (Computação). 2. Redes Neurais (Computação). 3. Hidrologia. 4. Cartografia. I. Universidade Federal do Paraná. II.Sampaio, Tony Vinicius Moreira. III. Título.

CDD: 006.32

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
Setor CIÊNCIAS DA TERRA
Programa de Pós Graduação em GEOGRAFIA
Código CAPES: 40001016035P1

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em GEOGRAFIA da Universidade Federal do Paraná foram convocados para realizar a arguição da Tese de Doutorado de **LUIS FERNANDO BUENO**, intitulada: "**INTELIGÊNCIA ARTIFICIAL APLICADA À MELHORIA DA ACURÁCIA DO MAPEAMENTO DE REDES DE DRENAGEM.**", após terem inquirido o aluno e realizado a avaliação do trabalho, são de parecer pela sua APROVAÇÃO.

Curtiba, 11 de Agosto de 2016.

Prof TONY VINICIUS MOREIRA SAMPAIO
Presidente da Banca Examinadora (UFPR)

Prof ANGELO EVARISTO SIRTOLI
Avaliador Externo (UFPR)

Prof MARIA CECILIA BONATO BRANDALIZE
Avaliador Externo (UFPR)

Prof SILVANA PHILIPPI CAMBOIM
Avaliador Externo (UFPR)

Prof LUIZ GILBERTO BERTOTTI
Avaliador Externo (UNICENTRO)

AGRADECIMENTOS

Ao Professor Doutor Tony Vinicius Moreira Sampaio pela dedicação e excelência na orientação do trabalho.

Aos amigos de sempre, Itamar Eloi Schlender e Nilce Roos Schlender, pela amizade e incentivo.

À amiga Tatiane Emilio Chechia pelo companheirismo ao longo do curso e pelo apoio na discussão dos conceitos da tese.

À amiga Tânia Maria Azevedo Guimarães Baraúna pelas revisões, contribuições e suporte durante a elaboração da tese.

Aos demais professores e colegas do Programa de Pós-Graduação em Geografia da Universidade Federal do Paraná, de forma especial ao Alex Mota dos Santos.

Como são grandes as riquezas de Deus! Como são profundos o seu conhecimento e a sua sabedoria!
Quem pode explicar as suas decisões?
Quem pode entender os seus planos?
Como dizem as Escrituras Sagradas: "Quem pode conhecer a mente do Senhor?
Quem é capaz de lhe dar conselhos?
Quem já deu alguma coisa a Deus para receber dele algum pagamento?
Pois todas as coisas foram criadas por ele, e tudo existe por meio dele e para ele.
Glória a Deus para sempre! Amém!"
Bíblia Sagrada, Romanos 11:33 a 36.

A sabedoria que vem do céu é antes de tudo pura; e é também pacífica, bondosa e amigável. Ela é cheia de misericórdia, produz uma colheita de boas ações, não trata os outros pela sua aparência e é livre de fingimento.
Bíblia Sagrada, Tiago 3:17.

RESUMO

Mapeamentos das redes de drenagens vêm sendo conduzidos, inicialmente a partir de interpretação visual de imagens, depois com auxílio de algoritmos para extração automática. Em detrimento da melhora na resolução espacial das imagens e na variedade dos algoritmos disponíveis, cada um deles com estratégia diferente para a geração dos canais de drenagem, a acurácia dos mapeamentos ainda é um problema recorrente. Nesta pesquisa avaliou-se o potencial de aplicação de técnicas de inteligência artificial no processo de extração automática de redes de drenagem, visando melhorar a acurácia do mapeamento. Um banco de dados espaciais foi construído, e reuniu dados oriundos do Modelo Digital de Elevação – MDE, parâmetros morfométricos, imagens SAR e SPOT 5, geologia, geomorfologia, hidrogeologia e solo. Uma Rede Neural Artificial – RNA foi criada para classificar amostras nas classes drenagem e não drenagem. A RNA, do tipo *perceptron* multicamadas com algoritmo de retropropagação de erros (*backpropagation*), foi definida com uma camada de entrada com 42 neurônios (quando usadas todas as variáveis possíveis), três camadas escondidas com 119 neurônios e uma camada de saída. A rede foi treinada a partir de quatro conjuntos de dados, e os testes realizados a partir de outros 16 conjuntos distintos de testes contendo amostras diferentes daquelas usadas no treinamento. Percebeu-se que a RNA foi mais eficiente na classificação dos conjuntos de dados com pixel de 2,5 metros, quando foram usadas na camada de entrada da rede todas as variáveis disponíveis e a camada de saída continha apenas duas classes (drenagem e não drenagem). Neste caso, a acurácia total ficou sempre acima de 68%. Foram identificados canais de primeira ordem que não constavam na base cartográfica de referência. A melhoria da acurácia temática e da completude foi observada, atestando que mineração de dados e RNA podem efetivamente contribuir na melhoria dos mapeamentos.

Palavras-chave: Mineração de dados. Redes Neurais Artificiais. Hidrologia. Cartografia. Qualidade de Dados Geoespaciais.

ABSTRACT

Mapping of drainage networks have been performed using visual interpretation of images, at first, then with the assist of automatic extraction algorithms. The limitation of spatial resolution of the available images and the diversity of available algorithms with different approaches in generating drainage channels, the accuracy level of this kind of mapping is still a frequent problem. This research evaluated the potential application of artificial intelligence techniques in auto-extracting process of drainage networks, in order to improve the mapping accuracy. A spatial database was built using data from: the Digital Elevation Model - DEM, morphometric parameters, SAR and SPOT 5 images, geology, geomorphology, hydrogeology and soil. An Artificial Neural Network - ANN was created to classify samples in classes of drainage and non-drainage. The multilayer perceptron ANN, with error backpropagation algorithm, was set with one input layer with 42 neurons (when all possible variables were used), three hidden layers of 119 neurons and an output layer. The network was trained from four datasets, and tests from 16 other distinct sets of tests with different samples from those used in training. The ANN was more efficient in classification of datasets with 2.5 meters pixels when all available variables were used in the network's input layer and the output layer had only two classes (drainage and non-drainage). Following this scenario, the overall accuracy has been always above 68%. First order draining channels were identified where nothing was described in the base map reference. The improvement of thematic accuracy was observed, confirming data mining and RNA as an effective way to contribute to the improvement of this sort of mapping.

Key-words: Data Mining. Artificial Neural Networks. Mapping. Hidrology. Cartography. Geoespatial Data Quality.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

ACP	-	Análise de Componentes Principais
AHP	-	*Analytical Hierarchy Processes*
ARFF	-	Attribute Relation Format File
ASTER	-	*Advanced Spaceborne Thermal Emission and Reflection Radiometer*
AUC	-	*Area Under the Curve*
BHRMP	-	Bacia Hidrográfica do Rio Mutum-Paraná
CONCAR	-	Comissão Nacional de Cartografia
CPRM	-	Serviço Geológico do Brasil
D8	-	*Deterministic Eight-Neighbor*
DEMON	-	*Digital Elevation Model Networks*
D∞	-	*Deterministic Infinity*
DSG	-	Diretoria de Serviço Geográfico
EMBRAPA	-	Empresa Brasileira de Pesquisa Agropecuária
ET-CQDG	-	Especificação Técnica para o Controle de Qualidade de Produtos de Conjuntos de Dados Geoespaciais
ET-PCDG	-	Especificação Técnica para Produtos de Dados Geoespaciais
FD8	-	*Fractional Deterministic Eight-Neighbor*
FN	-	*False Negatives*
FP	-	*False Positives*
GPS	-	*Global Positioning System*
HRV	-	*Visible High-Resolution*
IEC	-	*International Electrotechnical Commission*
INDE	-	Infraestrutura Nacional de Dados Espaciais
INPE	-	Instituto Nacional de Pesquisas Espaciais
ISO	-	*International Organization for Standardization*
KDD	-	Knowledge-Discovery in Databases
MAE	-	*Mean absolute error*
MDE	-	Modelo Digital de Elevação

MDS	-	Modelo Digital de Superfície
MDT	-	Modelo Digital do Terreno
MLP	-	*Multilayer Perceptrons*
MNS	-	Modelo Numérico da Superfície
MNT	-	Modelo Numérico do Terreno
NASA	-	*National Aeronautics and Space Administration*
NDWI	-	*Normalized Difference Water Index*
NIR	-	*Near Infrared*
OGC	-	Open Geoespatial Consortium
PDI	-	Processamento Digital de Imagens
PEC-PCD	-	Padrão de Exatidão Cartográfica dos Produtos Cartográficos Digitais
RAE	-	*Relative absolute error*
Rho8	-	*Random Eight-Neighbor*
RMSE	-	*Root mean squared error*
RNA	-	Rede Neural Artificial
ROC	-	*Receiver Operating Characteristic Curve*
RRSE	-	*Root relative squared error*
SAR	-	*Synthetic Aperture Radar*
SEDAM/RO	-	Secretaria de Estado do Desenvolvimento Ambiental
SGBD	-	Sistemas Gerenciadores de Banco de Dados
SIG	-	Sistemas de Informações Geográficas
SIPAM	-	Sistema de Proteção da Amazônia
SLAR	-	*Side-looking Airbone Radar*
SPOT	-	*Satellite Pour l'Observation de la Terre*
SQL	-	*Structured Query Language*
SRTM	-	*Shuttle Radar Topography Mission*
TauDEM	-	*Terrain Analysis Using Digital Elevation Models*
TN	-	*True Negative*
TP	-	True Positives
UTM	-	*Universal Transversa de Mercator*
WEKA	-	Waikato Environment for Knowledge Analysis
ZSEE/RO	-	Zoneamento Socioeconômico do Estado de Rondônia

SUMÁRIO

1 INTRODUÇÃO

As redes de drenagem, definidas por Christofoletti (1980) como sendo o conjunto de canais de escoamento inter-relacionados que formam a bacia de drenagem, constituem importantes subsídios para estudos geográficos. Na visão de O'Callaghan e Mark (1984), rede de drenagem é um conceito fundamental nas Ciências da Terra, base para a definição das bacias e um componente essencial em modelos hidrológicos e planos de gestão de recursos.

Para Zangh e Guilbert (2012), o sistema de drenagem é o padrão formado por riachos, rios e lagos em uma bacia de drenagem. Os autores consideram o sistema de drenagem como uma parte indivisível da terra, e componentes importantes para as análises de terreno. Couto et al. (2011) afirmaram que os contextos geológico, geomorfológico e os processos estruturais atuantes em determinadas áreas podem ser entendidos a partir do comportamento da rede de drenagem, seus padrões, formas e morfometria.

Akram et al. (2012) ressaltaram a importância das redes de drenagem quando afirmaram que a delimitação das redes de drenagem e a captação são importantes passos para o desenvolvimento de modelos hidrológicos. Silva e Kobiyama (2004) reforçaram a rede de drenagem de uma bacia hidrográfica como relevante variável no entendimento, simulação e previsão de processos hidrológicos e destacaram sua interação com a morfologia local. Silva e Kobiyama (2004) argumentaram, ainda, sobre a importância do mapeamento da rede de drenagem na conservação dos recursos hídricos, sobretudo na conservação das nascentes e o corpo dos rios por meio das matas ciliares.

A extração automática, apoiada por técnicas matemáticas e computacionais, vem sendo amplamente utilizada como forma de obtenção, facilitada e de menor custo, de mapeamentos de redes de drenagem para áreas onde não se dispõe de dados atualizados ou em escalas maiores (AKRAM et al., 2012; FERNÁNDEZ et al., 2012; BANON et al., 2013). Geralmente, por meio de softwares especializados, algoritmos são aplicados sobre Modelo Digital de Elevação – MDE para a extração de diversos parâmetros morfométricos, de direção de fluxo e as linhas de drenagem.

Diversos algoritmos foram desenvolvidos para realizar a extração automática das redes de drenagem, conforme se observa nos trabalhos de O’Callaghan e Mark

(1984), Fairfield e Leymarie (1991), Quinn et al. (1991), Costa-Cabral e Burges (1994) e Tarboton (1997).

Em detrimento dos avanços tecnológicos e conceituais no campo dos algoritmos para extração de redes de drenagem, Sampaio (2008) discutiu aspectos inerentes à subjetividade do processo de mapeamento. Para Sampaio (2008), o mapeamento da rede de drenagem, manual ou automático, continua sujeito a processos subjetivos. O autor exemplifica este problema de subjetividade citando a necessidade de definição de quantidade de *pixel*, que é usado como parâmetro fundamental em vários algoritmos. Para o autor, a definição da quantidade de pixels é subjetiva tanto no que se refere à quantidade a ser utilizada para definir o local aonde se inicia a rede de drenagem, como em relação à escolha da resolução espacial.

Geralmente, o valor da área de captação é confrontado com um limiar que representa a área mínima necessária para a definição de um canal a partir do qual as linhas de drenagem são iniciadas (FERNÁNDEZ et al., 2012). É interessante observar que trabalhos variados comprovaram que tal limiar é dependente das feições geomorfológicas da bacia de drenagem estudada, e que é recomendado adotar limiares diferentes, em cada setor, de acordo com as características do relevo (LOPEZ; CAMARASA, 1999; LIN et al., 2006; FERNÁNDEZ et al., 2012).

Não obstante, é conhecido que as características geomorfométricas das áreas das bacias hidrográficas afetam o funcionamento dos algoritmos de extração de redes de drenagem. Fernandéz et al. (2012) observaram que os resultados derivados da aplicação de um mesmo algoritmo, em áreas de bacias hidrográficas distintas, podem conduzir a mapeamentos com melhor ou pior acurácia. Semelhante conclusão pode ser encontrada em Wilson, Lam e Deng (2007) e em Crombez (2008).

Acerca do termo acurácia e sua utilização no contexto deste trabalho, cabe citar Monico et al. (2009, p. 473), em cujo trabalho defenderam que “o termo acurácia envolve tanto erros sistemáticos como aleatórios, enquanto precisão está unicamente vinculada com erros aleatórios”. Ainda conforme Monico et al. (2009, p. 473), “se acurácia envolve ambos os efeitos (sistemático e aleatório) e precisão somente os aleatórios, o termo acurácia por si só envolve a medida de precisão”.

Acurácia aparece como elemento de qualidade na norma *International Organization for Standardization* – ISO 19157:2013, que tratou da qualidade dos

dados geográficos e definiu elementos da qualidade de dados. De acordo com a citada norma, um elemento de qualidade descreve um determinado aspecto da qualidade do dado geográfico. As categorias definidas na norma ISO 19157 são as seguintes:

- Completude: presença ou falta de feições, seus atributos ou relacionamentos;
- Consistência Lógica: grau de aderência às regras lógicas da estrutura de dados, atribuição e relacionamentos;
- Acurácia Posicional: acurácia da posição das feições num determinado sistema de referência espacial;
- Acurácia Temática: acurácia dos atributos qualitativos, o quão corretos são os atributos não quantitativos e as classificações das feições e seus relacionamentos;
- Qualidade Temporal: qualidade dos atributos temporais e dos relacionamentos temporais entre feições; e,
- Usabilidade: baseada nos requisitos dos usuários e na aderência que têm as informações às necessidades dos usuários.

No Brasil, o normativo que trata da questão da qualidade dos dados geoespaciais é a Especificação Técnica para o Controle de Qualidade de Produtos de Conjuntos de Dados Geoespaciais – ET-CQDG (DSG, 2016), cuja primeira versão foi publicada no início do ano de 2016; tal especificação encontra-se em conformidade com a norma ISO 19157.

Diversos autores relataram também que a acurácia dos mapeamentos automatizados das redes de drenagem tende a ser pior quando os algoritmos são aplicados em áreas de relevo plano (TARBOTON, 1997; FAIRFIELD; LEYMARIE, 1991; PAZ; COLLISCHONN, 2008). Uma possível explicação reside na influência das características do relevo de cada área sobre o cálculo da área de captação, o que tem implicações diretas na escolha dos limiares para extração automática da drenagem (FERNANDÉZ et al., 2012).

Estudos recentes trataram desta questão, com a perspectiva de se conseguir obter redes de drenagens representativas para áreas com diferentes padrões geomorfológicos, como são os casos dos trabalhos de Sampaio (2008),

Banon et al. (2013) e Sampaio e Augustin (2014). Nestes casos, os autores propuseram metodologias auxiliares para o mapeamento da rede de drenagens.

Fernandéz et al. (2012) contribuíram com a discussão do tema e defenderam a ideia de que as diversas variáveis envolvidas no processo de extração automática, a saber: os dados utilizados, os algoritmos de fluxo escolhidos, os parâmetros de operação e as características geomorfométricas das microbacias influenciam diretamente no resultado final das redes extraídas.

A utilização de MDE em Sistemas de Informações Geográficas – SIG e o desenvolvimento de técnicas de Processamento Digital de Imagens – PDI e morfologia matemática possibilitaram automatizar o mapeamento das redes de drenagem. Sampaio (2008) argumentou que tais processos continuam incorporando elementos subjetivos e são dependentes da qualidade das bases cartográficas utilizadas.

Por outro lado, a tecnologia computacional da atualidade, no contexto do tratamento dos dados geoespaciais em estudos geográficos, permite obter dados oriundos de diversas fontes e aplicar técnicas de inferência sobre este conjunto de dados, gerando informações diferenciadas. A combinação de ferramentas tecnológicas de Inteligência Artificial como mineração de dados Espaciais, mineração de dados em Imagens e Redes Neurais Artificiais – RNAs, conjuntamente às já consagradas tecnologias de Banco de Dados e SIG, podem proporcionar a flexibilidade necessária para manipular as diversas variáveis intervenientes e parâmetros relacionados, de modo que se encontre um cenário adequado e que propicie melhor acurácia dos mapeamentos.

Luger e Stubblefild (1988) e Russell e Norvig (2004) são alguns dos autores que mostraram que a Inteligência Artificial faz uso da heurística quando o problema não tem uma solução exata devido a ambiguidades inerentes à declaração do problema ou dados disponíveis; ou quando o problema tem uma solução, mas o custo computacional para encontrá-la torna-se proibitivo. Hou et al. (2011) demonstraram a aplicação deste conceito quando usaram uma busca heurística para tratar das imperfeições do MDE.

Banon et al. (2013) sugeriram a avaliação de outras técnicas de mineração de dados para a extração automática de redes de drenagem e citaram as RNAs como tecnologia promissora.

Neste cenário, é relevante o estudo da acurácia dos mapeamentos da rede de drenagem extraídos por meio de processos automáticos, com vistas a ultrapassar os obstáculos inerentes aos atuais métodos disponíveis. Pode-se, portanto, pensar em novos arranjos tecnológicos que possam contornar as limitações e contribuir para o aumento da acurácia. Assim, surge como questão de pesquisa: investigar se a aplicação de técnicas de Inteligência Artificial pode contribuir para a identificação de padrões nos conjuntos de dados disponíveis, que possam ser úteis para o aumento da acurácia nos mapeamentos de rede de drenagem extraídos automaticamente.

O avanço tecnológico da Computação, quando aplicado à Geografia, pode contribuir significativamente para melhorar o poder de expressividade dos modelos, potencializar o uso de metodologias consagradas e aumentar a capacidade de análise. Como argumentou Ross (2006), a Geografia necessita do uso rotineiro das Tecnologias da Informação para realizar suas análises.

Ao longo do tempo, as bases de dados estão se tornando cada vez mais volumosas. O crescimento de tais bases de dados sugere que, pouco a pouco, ficará difícil manipular manualmente os dados em diversos domínios, incluindo o de Geografia. Estudos vêm sendo conduzidos no sentido da adoção de novas tecnologias que permitam manipular enormes conjuntos de dados, oriundos de fontes diversas e em formatos heterogêneos (GOODCHILD, 2013; GRAHAM; SHELTON, 2013; KITCHIN, 2013; WU et al., 2014).

O uso de ferramentas automatizadas para manipular grandes volumes de dados possibilita a extração de informações oportunas para a geração de novos conhecimentos. Fayyad, Piatetsky-Shapiro e Smyth (1996) defenderam que a necessidade de intensificar as capacidades de análise humana para lidar com o grande número de bytes é de natureza econômica e financeira. Os mesmos autores afirmaram, ainda, que a descoberta de conhecimento em bases de dados é uma tentativa de resolver um problema comum nos dias atuais: a sobrecarga de dados e a grande quantidade de informação.

Complementarmente, considera-se os aspectos abordados por Cunico e Oka-Fiori (2009) que sugeriram, em seu trabalho, a adoção do conceito de totalidade em oposição à perspectiva unitária para análise e avaliação sobre os recursos naturais, no qual os elementos envolvidos não se apresentam de maneira dissociada, e sim interagem de maneira dinâmica e em diferentes escalas. Acredita-

se que o ferramental computacional, sobretudo do ramo da Inteligência Artificial, pode ajudar significativamente este tipo de análise.

1.1 OBJETIVOS

O objetivo geral deste estudo é avaliar o potencial de aplicação das técnicas de inteligência artificial no processo de extração automática de redes de drenagem, visando melhorar a acurácia do mapeamento.

Com o intuito de atingir o objetivo geral, foram estabelecidos os seguintes objetivos específicos:

- Estruturar um banco de dados espaciais que reúna dados vetoriais e matriciais necessários para o estudo da rede de drenagem da Bacia Hidrográfica do Rio Mutum-Paraná;
- Aplicar técnicas de mineração de dados espaciais sobre a base construída, visando gerar subsídios para a extração automática da rede de drenagem;
- Construir um modelo de Redes Neurais Artificiais para apoiar o estudo da rede de drenagem da Bacia Hidrográfica do Rio Mutum-Paraná; e,
- Gerar uma rede de drenagem para a Bacia Hidrográfica do Rio Mutum-Paraná.

1.2 HIPÓTESES

Considerando que as características geomorfométricas da área de estudo interferem no resultado da aplicação dos algoritmos de extração automática de rede de drenagem, e que existem limitações inerentes a tais processos automáticos que podem impactar na acurácia dos mapeamentos derivados, como hipótese básica da pesquisa define-se que a aplicação de técnicas de Inteligência Artificial sobre uma base de dados espaciais poderá, por meio da identificação de padrões diferenciados para cada área de estudo, ser capaz de melhor representar a rede de drenagem e aumentar a acurácia do mapeamento.

De forma complementar, assume-se que uma RNA pode ser configurada para receber parâmetros que expressem as características geomorfométricas da

Bacia Hidrográfica do Rio Mutum-Paraná e seja capaz de identificar seus canais de drenagem, bem como proporcionar a geração de uma nova base cartográfica da rede de drenagem com maior acurácia que as bases cartográficas existentes.

Os princípios teóricos que fundamentaram a formulação da hipótese foram baseados nos trabalhos de Tarboton (1997), Wilson, Lam e Deng (2007), Paz e Collischonn (2008), Sampaio (2008), Fernandéz et al. (2012) e Banon et al. (2013).

1.3 SISTEMATIZAÇÃO DOS CAPÍTULOS

O trabalho está organizado em sete capítulos, incluindo este primeiro que contém a introdução, os objetivos e as hipóteses. Uma caracterização da área de estudo é apresentada no segundo capítulo. O terceiro capítulo apresenta os fundamentos teóricos da pesquisa, por meio da correspondente revisão de literatura, onde são apresentados os principais conceitos usados ao longo do estudo.

No quarto capítulo são descritos os materiais utilizados, assim como os métodos adotados para a estruturação do banco de dados espaciais, para a extração da drenagem e dos parâmetros morfométricos, para a execução das etapas da mineração de dados e para a definição da rede neural artificial.

No quinto capítulo são apresentados os resultados do estudo. Segue-se com a discussão dos resultados no sexto capítulo e, no sétimo capítulo, são expostas as conclusões do trabalho.

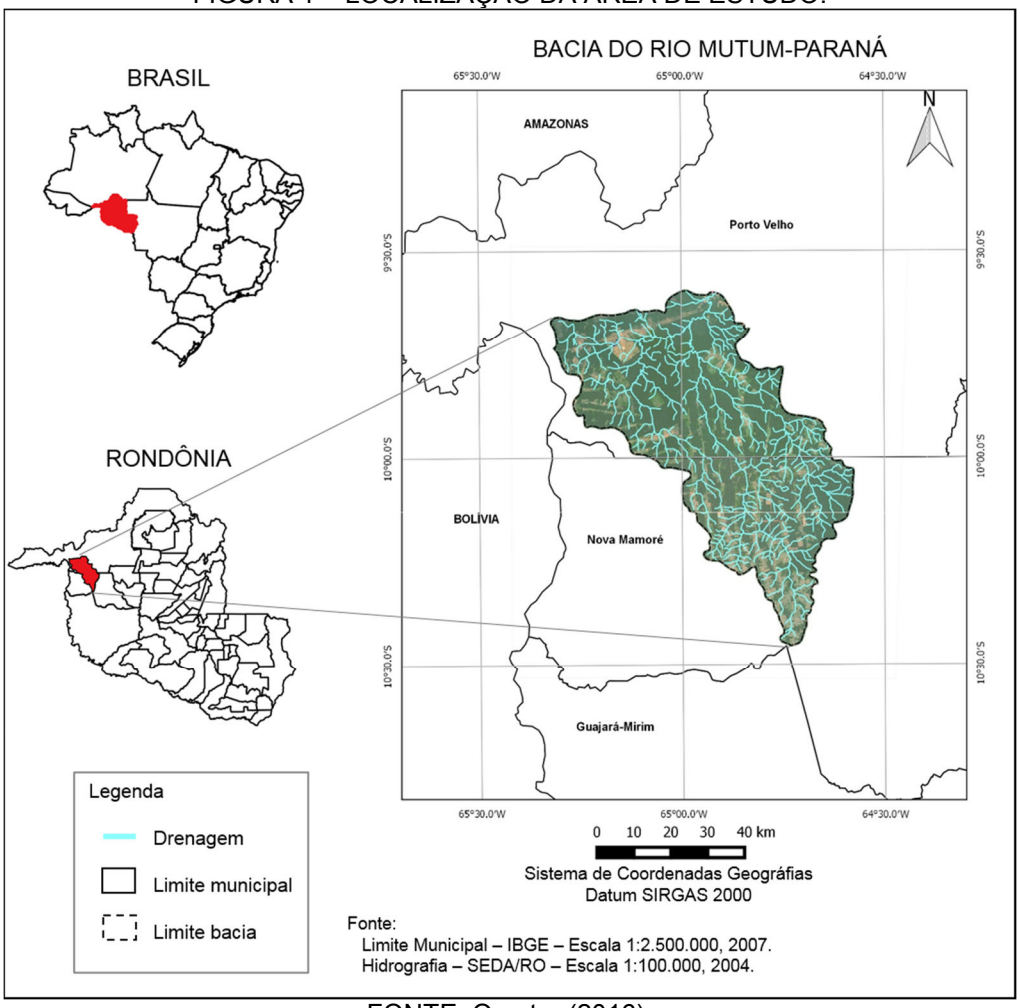
2 ÁREA DE ESTUDO

2.1 CARACTERIZAÇÃO DA BACIA HIDROGRÁFICA DO RIO MUTUM-PARANÁ

A área de estudo corresponde à Bacia Hidrográfica do Rio Mutum-Paraná – BHRMP, localizada no noroeste do Estado de Rondônia, estendendo-se pelos municípios de Porto Velho e Nova Mamoré e delimitada pelas coordenadas 9°34’40’’ e 10°01’49” de latitude Sul e 65°15’34” e 64°57’51” de longitude Oeste, com área de, aproximadamente, 3.560 km^2, conforme FIGURA 1.

A facilidade de acesso por meio de estradas vicinais, que embora não sejam asfaltadas são trafegáveis na maior parte do ano, e a existência de dados geoespaciais foram os principais fatores que justificaram a escolha de área de estudo. Na FIGURA 2 é apresentado um mosaico de imagens de satélite da área.

FIGURA 1 – LOCALIZAÇÃO DA ÁREA DE ESTUDO.



FONTE: O autor (2016).

A Bacia do Mutum-Paraná está incluída na porção sudoeste do Cráton Amazônico, que demonstra uma evolução geológica policíclica iniciada no paleoproterozóico, há aproximadamente 1.750 milhões de anos, sendo reativada por eventos tectono-magmáticos superimpostos até 970 milhões de anos atrás, quando, então, essa parte do Cráton estabilizou-se e, a partir daí, os movimentos tectônicos restringiram-se a reativações de falhas pré-existentes, com maior intensidade durante o período Terciário quando do soerguimento da Cordilheira Andina, tendo os seus reflexos afetado a região sul da Amazônia (RIZZOTO et al., 2005).

Quanto à geologia da área (FIGURA 3), Rizzoto et al. (2005) descreveram as seguintes unidades litoestratigráficas: Complexo Jamari; Formação Mutum-Paraná; Granito Serra da Muralha; Suíte Intrusiva Serra da Providência; Suíte Metamórfica Nova Mamoré; Suíte Laje; Suíte Intrusiva São Lourenço-Caripunas; Suíte Intrusiva Rondônia; Formação Palmeiral; Coberturas Cenozóicas (Formação rio Madeira, Formação Jaci-Paraná, Cobertura Detrito-Laterítica, Sedimentos Aluvionares Argilosos e Arenosos, Sedimentos Aluvionares Indiscriminados).

Adamy e Dantas (2005) descreveram a geomorfologia da Bacia do Mutum-Paraná (FIGURA 4) dividindo-a em duas partes: Bacia do Alto Rio Mutum-Paraná e Bacia do Baixo Rio Mutum-Paraná.

Em relação à Bacia do Alto Rio Mutum-Paraná, afirmaram Adamy e Dantas (2005) que se caracteriza por um relevo colinoso medianamente dissecado, apresentando áreas com uma dissecação variável entre alta a baixa. Ao contrário das bacias dos rios Jaci-Paraná e Candeias, que também são afluentes do rio Madeira entre Porto Velho e Jirau, a bacia do rio Mutum-Paraná, de menor abrangência, não drena as vertentes escarpadas da serra dos Pacaás Novos, mas apenas alguns de seus contrafortes mais rebaixados (ADAMY; DANTAS, 2005). Citando dados do Zoneamento Socioeconômico do Estado de Rondônia – ZSEE/RO (RONDÔNIA, 2002), Adamy e Dantas (2005) apresentaram duas unidades geomorfológicas maiores neste subambiente, representadas pelas Superfícies de Aplanamento e Planícies Aluviais de Rios Secundários.

Segundo os autores, nesta parte da Bacia do Rio Mutum, os terrenos são embasados por rochas do Complexo Jamari, onde predominam solos Podzólicos Vermelho-Amarelos álicos que, de acordo com os autores, foram posteriormente reinterpretados como Argissolos Vermelho-Amarelos Alumínicos e Latossolos Amarelo Alumínicos, sendo caracterizados por solos espessos, argilosos, bem

estruturados e com expressiva variação textural entre os horizontes A e Bt. Um mapa de solos é apresentado na FIGURA 5.

O ambiente foi caracterizado como pouco alterado pela intervenção humana, recoberto por Floresta Tropical aberta, notabilizando-se por sua estabilidade morfodinâmica frente aos processos erosivo-deposicionais e a movimentos de massa (ADAMY; DANTAS, 2005). Para os autores citados, a fraca declividade das vertentes das colinas associada à descontinuidade hidráulica existente no contato dos horizontes A e B dos Argissolos podem desencadear algumas ocorrências erosivas, mas de pouco significado enquanto estes terrenos se mantiverem florestados.

A Bacia do Baixo Rio Mutum-Paraná é constituída pela Bacia do rio Mutum-Paraná e pelas bacias de igarapés menores que drenam diretamente para a margem direita do rio Madeira, tais como os igarapés Jirau e Cirilo, caracterizando-se por um relevo plano, muito pouco dissecado, inserido no Planalto Rebaixado da Amazônia Ocidental. Localmente, pode exibir faixas com um grau de dissecação mais acentuado (ADAMY; DANTAS, 2005).

Novamente, os autores citaram dados do ZSEE/RO para descreverem que a área deste subambiente é caracterizada por uma ampla superfície de aplainamento, de relevo plano a muito suavemente ondulado, entre as cotas de 200 e 300 metros, e que apresentam graus de dissecação variando entre baixo e alto. Os autores afirmaram, ainda, que foram identificados uma baixa ocorrência de relevos residuais, tais como *inselbergs*, *hillocks* e *tors*.

De acordo com Adamy e Dantas (2005), nestes terrenos embasados por rochas do Complexo Jamari predominam Latossolos Vermelho-Amarelos álicos, que se caracterizam por solos muito espessos, argilosos, bem drenados e estruturados. Destacam também os autores que a morfologia quase plana das áreas aplainadas e dos baixos platôs associada a solos e mantos de intemperismo espessos e bem drenados indica uma vulnerabilidade muito baixa com relação aos processos erosivo-deposicionais em terrenos florestados (ADAMY; DANTAS, 2005).

A hidrogeologia da área (FIGURA 6) foi descrita por Melo Júnior (2005), que identificou diferentes unidades hidrogeológicas que serão apresentadas a seguir conforme a caracterização elaborada pelo referido autor.

Os Aquíferos Intergranulares Descontínuos Livres correspondem aos sedimentos terciários da Formação Jaciparaná de composição arenosa, areno-siltosa e areno-argilosa.

Os Aquíferos Intergranulares/Fraturados Contínuos Livres correspondem aos litotipos da Formação Palmeiral, compostos predominantemente por arenitos ortoquartzíticos e paraconglomerados fortemente cimentados. Essa cimentação confere um caráter de rocha cristalina a esta unidade, cuja percolação de água se dá, principalmente, nas fraturas e vênulas geradas pela tectônica imposta a seu arcabouço.

Aquíferos Locais Restritos às Zonas Fraturadas correspondem às rochas vulcânicas ácidas inseridas na Suíte Intrusiva Serra da Providência, bem como aos basaltos de composição vulcânica básica. Segundo Melo Júnior (2005), a permeabilidade desse sistema é variável, comumente baixa; no entanto, os poços que exploram estes aquíferos apresentam produtividade média maior que aqueles que exploram os aquíferos fraturados descontínuos, livres.

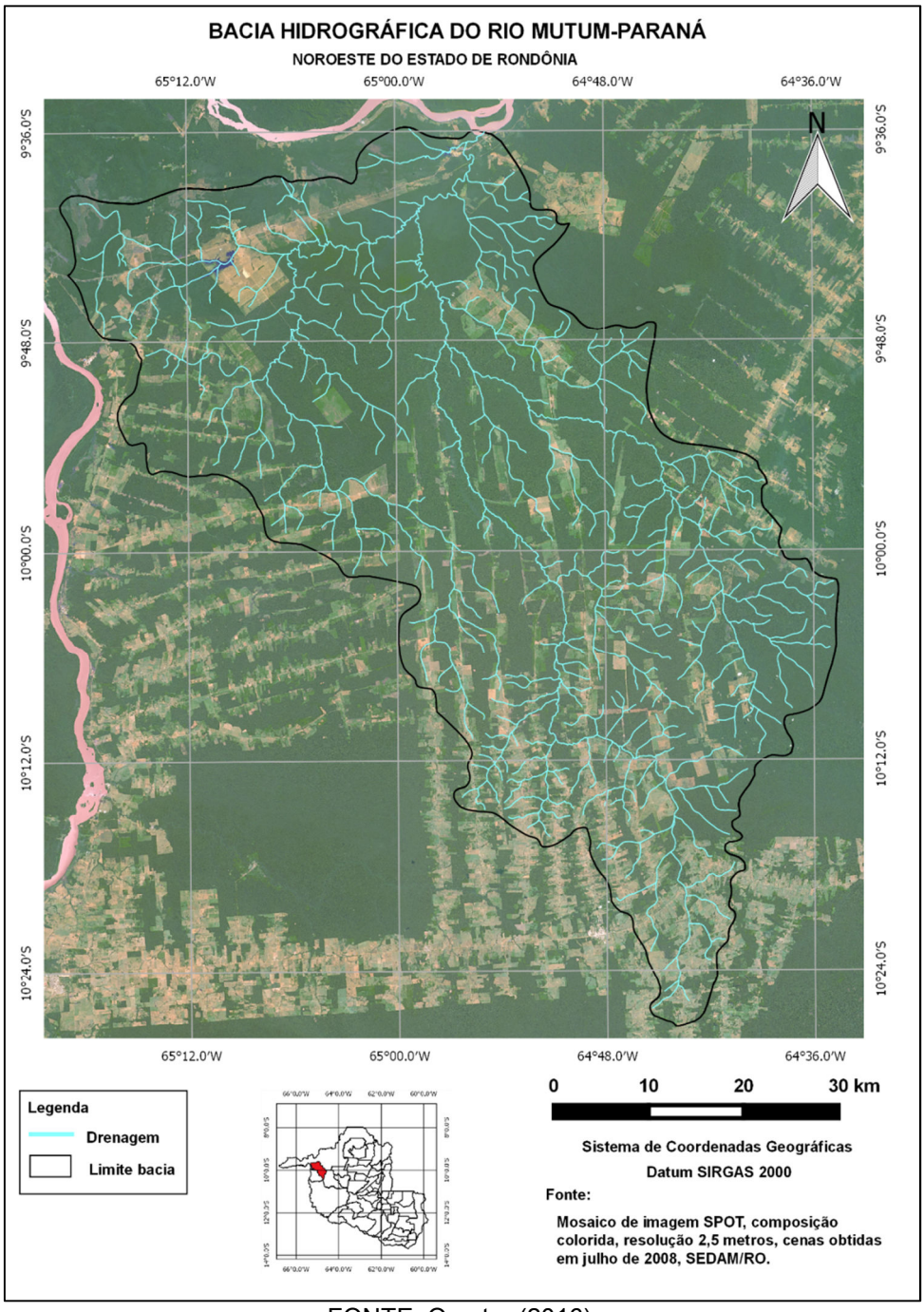
Aquíferos Fraturados Descontínuos Livres correspondem às fraturas abertas existentes nas rochas ortognáissicas de composição granítica do Complexo Jamari. Frequentemente, estes sistemas aquíferos são ampliados pela ocorrência de uma cobertura de sedimentos coluvionares constituída por materiais detrítico argilo-arenosos, com espessura variável. A Suíte Intrusiva Alto Candeias também compõe esse sistema aquífero, sendo composto principalmente por granitos porfiríticos de granulação média a grossa.

Aquífugos são caracterizados por litotipos pouco favoráveis ao armazenamento de água subterrânea, além de apresentarem um relevo bastante acidentado, o que dificulta ainda mais a infiltração. Correspondem às ocorrências da Sequência Metavulcano-Sedimentar e aos sills basálticos, respectivamente das Formações Mutumparaná e Nova Floresta. Enquadra-se, ainda, nesta compartimentação as coberturas detrito-lateríticas e os lateritos maturos da Formação Solimões, os lateritos imaturos mosqueados e concrecionários da Formação Jaciparaná, as rochas das Suítes Intrusivas São Lourenço-Caripunas e Rondônia, as rochas das Suítes Metamórficas Quatro Cachoeiras e Rio Crespo e, finalmente, as rochas do Granito Serra da Muralha.

Melo Júnior (2005) caracterizou, ainda, a vulnerabilidade natural das águas subterrâneas da região da Bacia do Mutum-Paraná, concluindo que os índices de

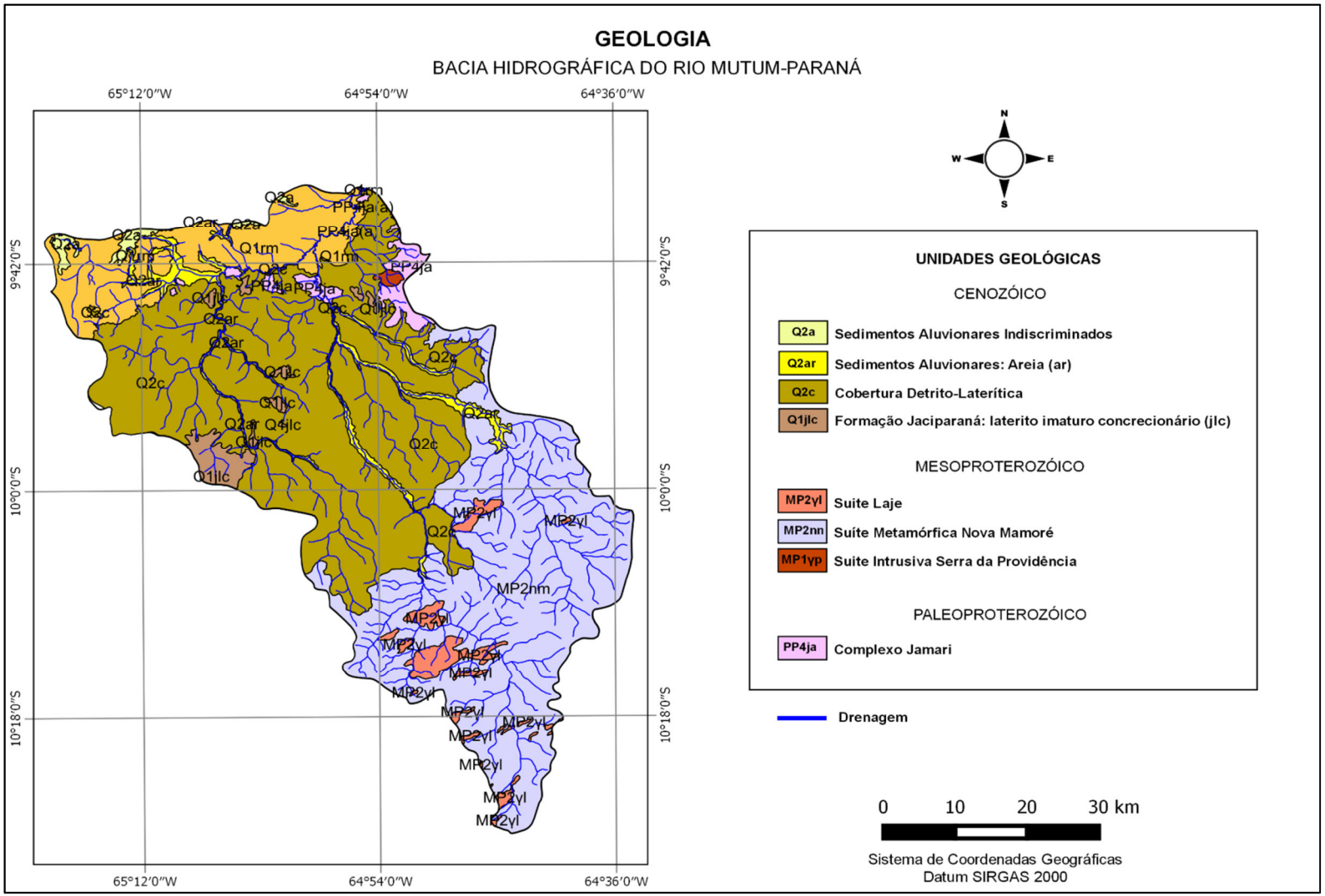
vulnerabilidade variam de altos a extremamente altos nas porções centro-noroeste e centro-sudoeste da bacia.

FIGURA 2 – MOSAICO DE IMAGENS DE SATÉLITE SPOT 5 DA ÁREA DE ESTUDO.



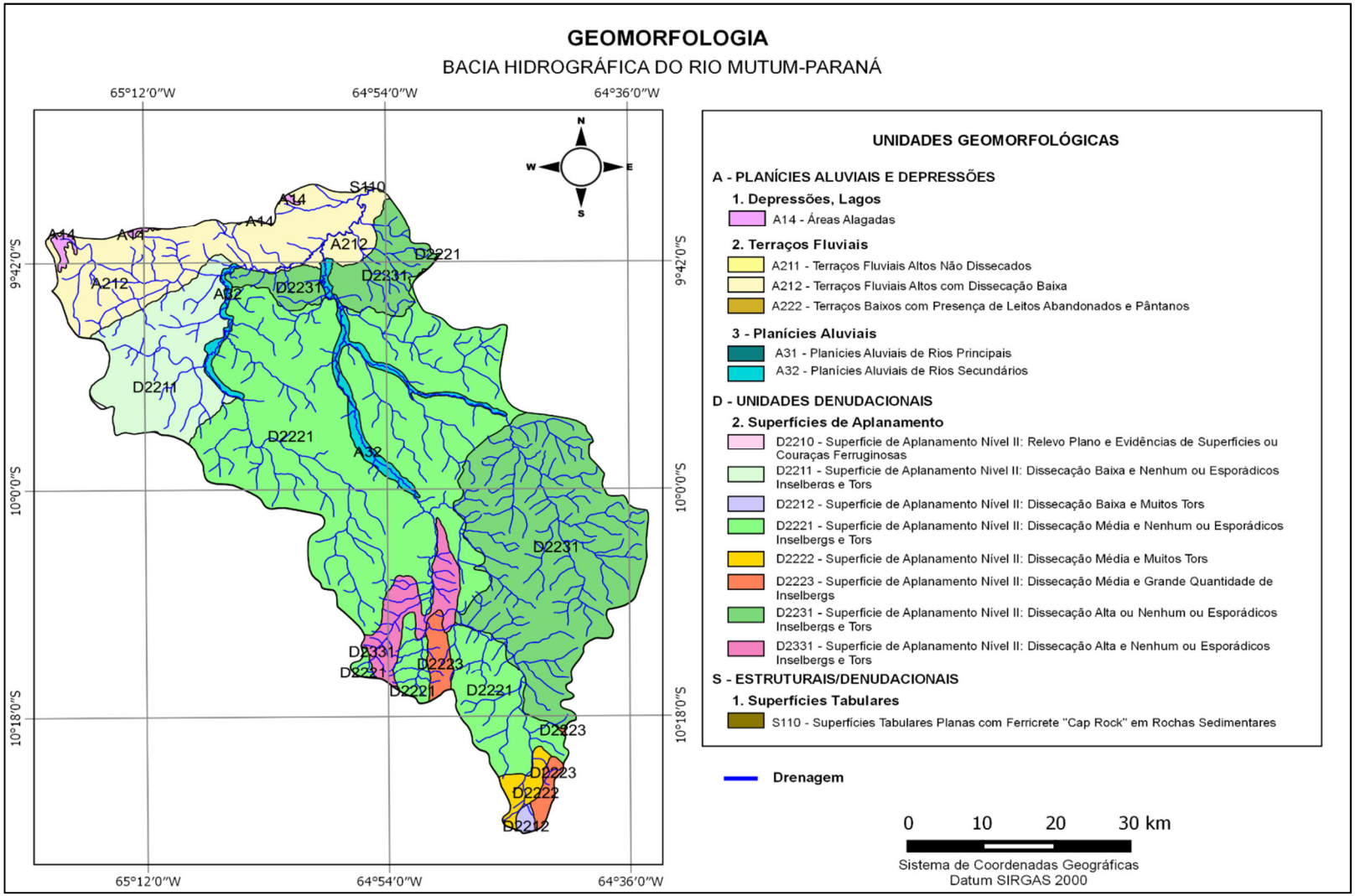
FONTE: O autor (2016).

FIGURA 3 – GEOLOGIA DA ÁREA DE ESTUDO.



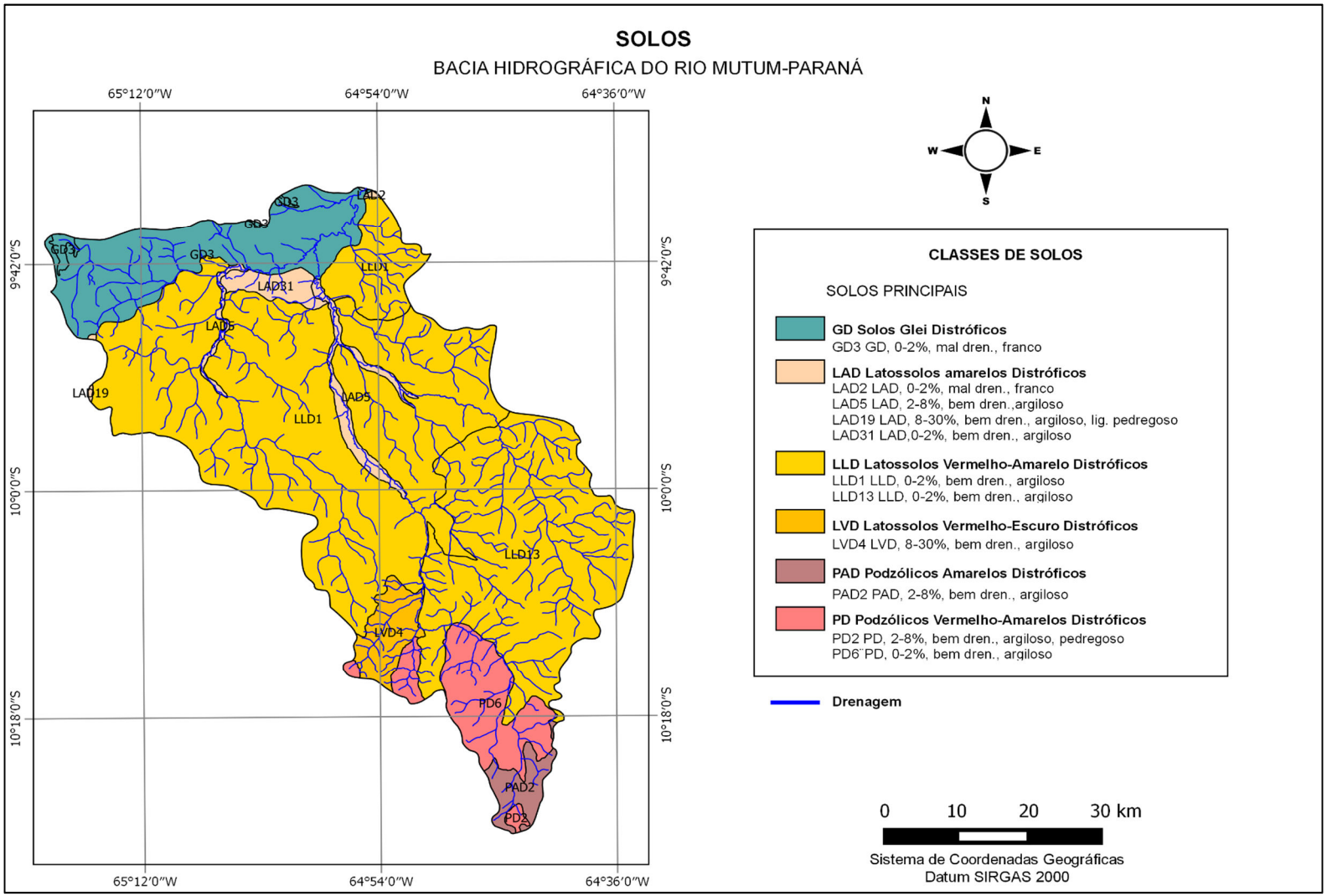
FONTE: Adaptado pelo autor a partir de Rizzoto et al. (2005).

FIGURA 4 – GEOMORFOLOGIA DA ÁREA DE ESTUDO.



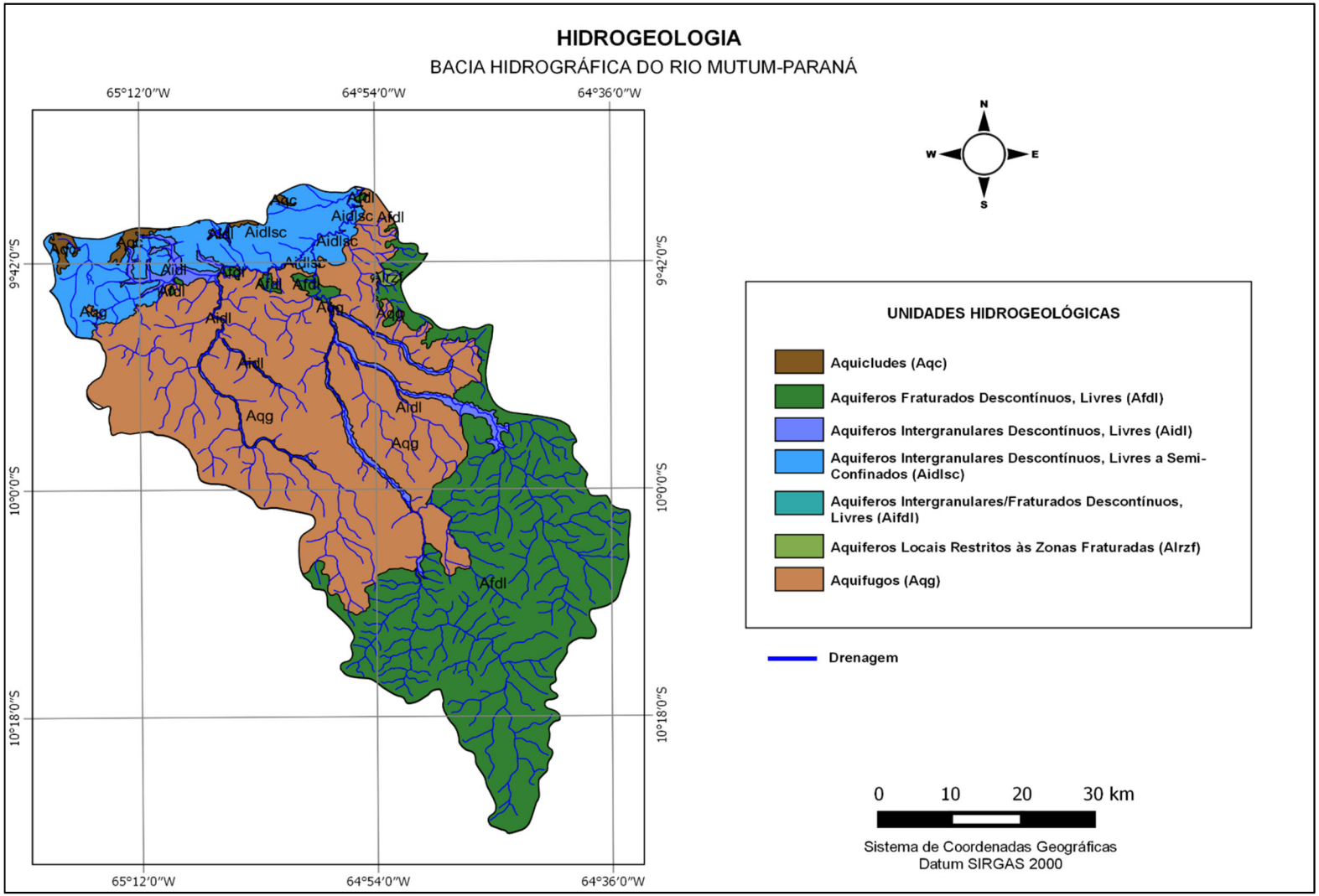
FONTE: Adaptado pelo autor a partir de Adamy e Dantas (2005).

FIGURA 5 – SOLOS DA ÁREA DE ESTUDO.



FONTE: Adaptado pelo autor a partir de Rondônia (2002).

FIGURA 6 – HIDROGEOLOGIA DA ÁREA DE ESTUDO.



FONTE: Adaptado pelo autor a partir de Melo Júnior (2005).

3 FUNDAMENTAÇÃO TEÓRICO-METODOLÓGICA

O presente capítulo apresenta a revisão de literatura realizada durante a pesquisa. Este capítulo pode ser dividido em dois blocos principais, onde o primeiro trata da extração automática de redes de drenagem e padrões morfométricos. Nesta primeira parte da revisão, são expostos o processo de extração automática, os algoritmos de fluxos e parâmetros morfométricos. Já no segundo bloco, o foco recairá sobre a descoberta de conhecimentos em bancos de dados espaciais, com a revisão dos conceitos relacionados aos Bancos de Dados Espaciais, descoberta de conhecimento em bancos de dados, mineração de dados e RNAs.

3.1 EXTRAÇÃO AUTOMÁTICA DE REDES DE DRENAGEM E PADRÕES MORFOMÉTRICOS

Na visão de O'Calleghan e Mark (1984), as redes de drenagem e os canais associados, bem como as bacias de drenagem são conceitos fundamentais em Ciências da Terra. Na definição dos autores mencionados, canais de drenagem referem-se às linhas ao longo das quais processos fluviais atuam para o transporte de água e material mineral de uma região, permitindo que os processos de gravidade em encostas continuem o transporte para paisagens mais baixas. Ainda segundo O'Calleghan e Mark (1984), a topologia e a geometria da rede de drenagem constituem relevantes áreas de estudo dentro da geomorfologia; acrescentando, ainda, que as redes de drenagem são a base para a definição da bacia de drenagem, um componente essencial em modelos hidrológicos e planos de gestão de recursos.

Oliveira, Guasseli e Saldanha (2009) afirmaram que as bacias de drenagem constituem unidades territoriais de planejamento que podem ser tratadas como um sistema onde há entradas, saídas e transformações. Desta forma, os modelos de gerenciamento dos recursos hídricos assumem a bacia hidrográfica como unidade geográfica de referência ou de intervenção, uma vez que nela ocorre boa parte das relações de causa e efeito que envolvem o meio ambiente.

Perspectivas ao estudo das bacias hidrográficas foram citadas por Oliveira, Guasseli e Saldanha (2009), que argumentaram, de um lado, sobre o estudo morfométrico que engloba as análises referentes à hierarquia fluvial, análise areal,

linear e hipsométrica obtidas de mapas, fotografias aéreas e imagens de satélites, indicando as características físicas da bacia; e, por outro lado, sobre o estudo da dinâmica de uma bacia de drenagem, cujos dados são obtidos de coletas e medições realizadas no campo e a partir da elaboração de índices estatísticos e modelos matemáticos referentes à precipitação, infiltração, evaporação e evapotranspiração, escoamento superficial, regime dos cursos d'água, água subterrânea e transporte de sedimentos.

Para Paz e Collischonn (2008), os dados de elevação do terreno provenientes do SRTM e disponibilizados gratuitamente na internet constituem excelentes fontes de informações para a caracterização topográfica de bacias hidrográficas. Os supramencionados autores defenderam a utilização de procedimentos computacionais para a automação da extração da drenagem.

Procedimentos computacionais podem ser facilmente aplicados para extrair, de forma automatizada, a rede de drenagem e diversas outras informações a partir dos Modelos Digitais de Elevação. Esses procedimentos podem ser customizados para elaborar planos de informação específicos para entrada em modelos hidrológicos, agilizando a aplicação destes em bacias de grande porte.

3.1.1 Processo de extração automática

Um Modelo Digital de Elevação – MDE, de acordo com a definição normatizada na Especificação Técnica para Produtos de Conjuntos de Dados Geoespaciais – ET-PCDG (DSG, 2014, p. 5-1), "é um produto cartográfico obtido a partir de um modelo matemático que representa um fenômeno, de forma contínua, a partir de dados adequadamente estruturados e amostrados do mundo real".

De acordo com Brandão e Santos (2009), o MDE é utilizado para calcular os valores que descrevem a altimetria de uma localização geográfica específica, ou dos arredores desta localização, e deve retratar, de maneira precisa, a área em estudo. Para estes autores, o MDE deve ser capaz, também, de representar ou fornecer informações geomorfológicas, ou seja, características especiais do relevo que traduzem formas específicas, tais como: cumeadas, talvegues, etc., bem como as descontinuidades da superfície como falhas geológicas.

As formas de apresentação para MDE também foram descritas na ET-PCDG, que as dividiu em duas representações: solo exposto e solo exposto com os

acidentes naturais e artificiais localizados sobre ele. Dentre as definições da norma, destacam-se as seguintes: Modelo Digital do Terreno – MDT, que é obtido a partir de um modelo matemático que representa o solo exposto, de forma contínua e suavizada, a partir de dados adequadamente estruturados e amostrados da superfície física da Terra; e o Modelo Digital da Superfície – MDS: obtido a partir de um modelo matemático que representa o solo exposto e os acidentes encontrados acima do solo, de forma contínua e suavizada, a partir de dados adequadamente estruturados e amostrados do mundo real.

Para Gong e Xie (2009), a importância do MDE na análise digital de terrenos foi impulsionada pelo desenvolvimento da fotogrametria digital, do sensoriamento remoto e dos Sistemas de Informação Geográfica. Gong e Xie (2009) discutiram a extração de redes de drenagem a partir de MDE com grande volume de dados, apoiada por computação de alto desempenho.

Fernández et al. (2012) descreveram que, de forma geral, o método de extração de redes de drenagem por processos automáticos envolve:

1) O preparo e correção do MDE;
2) O cálculo das direções de fluxo;
3) O cálculo da área de captação; e,
4) O delineamento das linhas de drenagem.

Neste processo, os autores argumentaram a respeito da importância de diversos fatores, como resolução, nível de processamento e características de aquisição dos dados, algoritmo de fluxo utilizado, características geomorfométricas da área a ser analisada e os limiar para a definição.

Akram et al. (2012) defenderam que o sucesso da extração automática das redes de drenagem depende de fatores como a extensão e a disponibilidade dos dados fontes, e, ainda, as características geomorfológicas da área modelada. Banon et al. (2013) defenderam uma metodologia para a extração automática de uma rede de drenagem capaz de representar áreas com diferentes padrões geomorfológicos. Neste trabalho, Banon et al. (2013) basearam a metodologia na extração de atributos do MDE e usaram a mineração de dados para a definição dos atributos mais representativos da rede de drenagem.

Fernández et al. (2012) afirmaram ser necessário considerar as variáveis intervenientes no processo de extração automática de redes de drenagem,

considerando para a obtenção de redes de drenagem com maior acurácia desde o dado utilizado, passando pelos algoritmos de fluxo, os parâmetros de operação, até as características geomorfométricas das microbacias.

Acerca do uso de MDE na extração da rede de drenagem, Hosseinzadeh (2011) defendeu a divisão da bacia em unidades geomorfológicas e o uso de um limiar diferente em cada unidade; tal problema foi destacado, também, por Brandão e Santos (2009).

Brandão e Santos (2009) verificaram que, em áreas planas, ocorreram problemas na determinação da continuidade da rede de drenagem. Entretanto, os autores argumentaram que a utilização dos dados orbitais do SRTM na geração de MDE hidrologicamente consistido viabiliza a extração de variáveis físicas das bacias hidrográficas em SIG em menor intervalo de tempo auxiliando na tomada de decisões relativas à gestão ambiental. As técnicas de sensoriamento remoto, aliadas às técnicas de SIG, mostram-se eficientes para avaliações referentes a dados hidrológicos.

De acordo com Fernández et al. (2012), as características da rede de drenagem extraídas são influenciadas pela geomorfologia da área analisada, sendo necessário o ajuste dos algoritmos de acordo com o tipo de relevo existente. Os autores defenderam a necessidade de identificar limiares específicos para serem aplicados em compartimentos geomorfológicos distintos.

Conforme Pelletier (2013), a maioria dos métodos existentes para extração da rede de drenagem em MDE depende de área de contribuição, ou em limiares pré-estabelecidos pelo usuário que definem a transição de morro para vale. Segundo Pelletier (2013), os primeiros métodos de extração utilizavam apenas a área de contribuição como parâmetro básico, ou uma combinação de área de contribuição, comprimento e inclinação.

Problemas, dificuldades e limitações relacionadas ao uso de MDE, que podem afetar a acurácia das redes extraídas, são comumente discutidos na literatura. O'Calleghan e Mark (1984) demonstraram sua preocupação quanto aos ruídos introduzidos durante a coleta de dados e, deste modo, apresentaram um método para extração de redes de drenagem que objetivou tratar a questão do ruído para delimitar apenas as principais vias de drenagem.

Conforme Fernández et al. (2012), nem sempre é possível obter drenagens fiéis às existentes na paisagem devido à perda de informações que ocorre desde o

levantamento de dados até a extração. Na visão de Fernández et al. (2012), a drenagem existente é parcialmente revelada no relevo, sendo o relevo, por sua vez, simplificado no MDE. Acrescentaram os autores que, muitas vezes, resultados obtidos com programas de modelagem hidrológica expressam uma degeneração da complexidade do traçado existente na paisagem.

No trabalho de Oliveira, Guasseli e Saldanha (2009), os autores preocuparam-se com o procedimento de interpolação para a obtenção de um MDE hidrologicamente corrigido, além de outros aspectos, a saber: a eliminação das depressões artificiais, a introdução de informações sobre as localizações de rede de drenagem e lagos em regiões planas, e as limitações inerentes dos métodos de obtenção da direção de fluxo. A problemática da eficiência dos algoritmos hidrológicos, quando aplicados em áreas de relevo plano, também foi discutida no trabalho de Oliveira, Guasseli e Saldanha (2009).

Tomazoni et al. (2011) mencionaram problemas que podem ocorrer quando do uso de imagens SRTM. Eles afirmaram que tais imagens apresentam dificuldades na localização exata dos rios que possuem matas ciliares, visto que é possível acontecer que a vegetação das árvores cubra o canal dos rios causando a impressão que estes locais sejam mais elevados do que as áreas do entorno. Problema semelhante ao mencionado por Tomazoni et al. (2011) já tinha sido relatado no trabalho de Valeriano e Abdon (2007).

Hou et al. (2011) afirmaram que a questão básica para a extração de redes de drenagem é a determinação da direção do fluxo para cada célula em uma matriz MDE. Para os autores, o MDE fornece uma representação digital contínua da superfície da Terra e argumentaram, ainda, que tais modelos que estão disponíveis para uso são simples de usar e possuem aplicabilidade generalizada para a análise de problemas hidrológicos.

Hou et al. (2011) alertaram para algumas desvantagens quanto ao uso dos MDEs, visto que não são raras as vezes em que descrevem depressões (buracos) e áreas planas, que são consideradas como não possuindo nenhuma drenagem. Tais problemas podem estar relacionados às próprias características dos terrenos que estão sendo representados, como no caso da existência de pedreiras e grutas, podendo surgir erros durante o processo de interpolação dos dados ou mesmo serem introduzidos durante o processo de geração do DEM.

Para Hou et al. (2011), depressões e buracos são desafios para a derivação automática de redes de drenagem totalmente conectadas. Em células com tais ocorrências, o sentido do fluxo não pode ser determinado com referência aos seus vizinhos. Apesar das limitações e dificuldades expostas, inúmeros trabalhos continuam utilizando os MDEs, bem como os algoritmos para determinação de fluxo, inclusive com a discussão dos procedimentos empregados e análise comparativa dos resultados obtidos.

Paz e Collischonn (2008) relataram procedimentos empregados para extrair automaticamente a rede de drenagem a partir do MDE SRTM, para a região geográfica da Bacia do Rio Uruguai. No trabalho dos autores, foram abordadas a geração de direções de fluxo e de áreas acumuladas de drenagem, a delimitação de bacias hidrográficas e a identificação e determinação dos cursos d'água.

Na análise dos resultados obtidos no trabalho de Paz e Collischonn (2008), os autores compararam qualitativamente as redes extraídas automaticamente a partir de variados MDE com a rede vetorial digital pré-existente para a região. Interessantes observações foram tecidas pelos autores que afirmaram que a qualidade da rede de drenagem derivada do MDE decresce com o aumento da resolução e aumenta com o pré-processamento.

Paz e Collischonn (2008) argumentaram que a performance das drenagens extraídas a partir de MDE pode ser explicada, dentre outros fatores, pelo tamanho do *pixel* da imagem *raster* em relação às características do rio, principalmente largura e sinuosidade. Complementarmente, os autores afirmaram que quando a resolução do MDE é inferior à largura do rio, vários *pixels* representam a largura do rio no MDE e possuem valores de elevação praticamente iguais ou com uma diferença mínima não representativa da variação da topografia.

Petsh, Monteiro e Bueno (2012) procederam uma análise comparativa da acurácia de uma rede de drenagem gerada automaticamente em relação a outra extraída diretamente de uma carta topográfica do Município de Ponta Grossa, no Paraná. Interessante observar, neste trabalho de Petsh, Monteiro e Bueno (2012), o relato acerca de problemas da drenagem gerada de forma automatizada que, de acordo com os autores, consistiu na sua generalização em relação aos canais de primeira ordem que acabam influenciando em parâmetros morfométricos, como a densidade de drenagem e a densidade hidrográfica.

Souza, Cruz e Aragão (2011) também desenvolveram um estudo comparativo do desempenho dos MDEs ASTER, TOPODATA e SRTM na obtenção automática de parâmetros físicos de bacias hidrográficas. Neste estudo, os autores consideraram o trabalho com diferentes escalas submetidas a um mesmo algoritmo de fluxo.

Apesar dos resultados satisfatórios na determinação da bacia e das sub-bacias, com suas áreas e perímetros, Souza, Cruz e Aragão (2011) notaram que todos os MDEs foram deficientes na estimativa dos comprimentos de drenagem, na delimitação correta de uma sub-bacia – cujo exutório situava-se em região de cobertura vegetal intensa – e indicaram a necessidade da utilização com cautela de tais produtos, sempre com suporte de outras fontes de informação.

Brubacher et al. (2012) afirmaram que a disponibilização dos dados SRTM aumentou significativamente o volume de estudos que incorporaram a utilização de MDE na extração automática de redes de drenagem, procedimento comum em análises hidrológicas ou ambientais. No estudo de Brubacher et al. (2012), os autores avaliaram e compararam, ainda, a precisão das bases SRTM utilizadas no Brasil (NASA, EMBRAPA e TOPODATA) nos processos de extração de drenagem, bacias, altimetria e no cálculo da extensão dos rios, considerando diferentes padrões morfométricos. Neste sentido, Brubacher et al. (2012), assim como no caso do trabalho de Oliveira, Guasseli e Saldanha (2009), observaram uma tendência de aumento no deslocamento das drenagens nas três bases SRTM à medida que diminui a declividade. Também observaram, relativamente aos erros associados ao cálculo de extensão dos rios, que as maiores discrepâncias ocorreram nas sub-bacias mais planas, com rios sinuosos com diferenças de extensão superiores a 10 km (BRUBACHER et al., 2012).

Fernández et al. (2012) avaliaram a compatibilidade de redes de drenagem extraídas automaticamente, com base no MDE, a partir de algoritmos comumente utilizados na literatura. Fernández et al. (2012) utilizaram quatro bacias no Município de São José dos Campos, São Paulo, como áreas teste para as extrações automáticas das redes de drenagem. Em todas as bacias foram aplicados seis algoritmos de fluxo e conduziram uma análise qualitativa e quantitativa dos resultados, tomando-se como referência a rede de drenagem extraída de carta topográfica na escala 1:50.000.

Dentre os resultados discutidos no estudo de Fernández et al. (2012), é possível destacar o desempenho dos algoritmos, sendo que aqueles que simulam fluxos com múltiplas direções produzem redes mais próximas da realidade. Os autores argumentaram que os limiares influem diretamente nos resultados das redes geradas, produzindo redes mais generalizadas com limiares maiores e mais complexas com limiares menores. Quanto à presença de áreas planas, argumentaram que é condição extremamente desfavorável ao desempenho dos métodos de extração automatizados.

Estudo semelhante foi conduzido por Marques et al. (2011), que analisou a qualidade e a precisão da delimitação automática de bacias hidrográficas, bem como a identificação de segmentos referentes a redes de drenagem utilizando dados SRTM (90m), TOPODATA (30m) e ASTER (20m).

Marques et al. (2011) observaram diferenças nas redes de drenagem extraídas a partir dos distintos MDEs e recomendaram a utilização de alguma fonte complementar de dados para verificar a acurácia dos resultados. Os autores relataram que houve um ganho significativo no detalhamento da rede de drenagem no modelo com *pixel* de 30 metros, sendo que o modelo TOPODATA se mostrou mais acurado.

3.1.2 Algoritmos de fluxo

Tribe (1992) classificou os algoritmos básicos para derivação das redes de drenagem a partir do MDE em três modelos, a saber: aqueles baseados na manipulação de células individuais de elevação; aqueles baseados no conceito de acumulação de fluxo; e, aqueles que combinam os dois primeiros.

Argumentaram Strobl e Forte (2007) que nos algoritmos baseados na manipulação de células individuais de elevação comumente são comparados a curvatura local ou a elevação das células vizinhas, de modo a determinar a célula de fluxo. Os autores comentaram acerca da necessidade de se utilizarem procedimentos de pós-processamento para corrigir a descontinuidade nos segmentos de canais produzidos por estes algoritmos.

Strobl e Forte (2007) acreditam que os algoritmos baseados no conceito de acumulação de fluxo podem produzir redes de drenagem contínuas e explicaram que o princípio envolvido neste tipo de algoritmo baseia-se no fluxo acumulado ao

longo da paisagem. As células de fluxo são definidas como os pontos em que o escoamento superficial é suficientemente concentrado para que os processos fluviais se desenvolvam sobre os declives. Para os supramencionados autores, a principal vantagem desta abordagem é que uma rede de canais contínuos é obtida como resultado, enquanto que a principal desvantagem desta técnica está relacionada à escolha subjetiva de um valor limite que deve ser especificado para definir quais células determinarão o fluxo.

Dentre os algoritmos mais populares relatados na literatura é possível citar: *Deterministic Eight-Neighbor* – D8 de O'Callaghan e Mark (1984); *Fractional Deterministic Eight-Neighbor* – FD8 de Quinn et al. (1991); *Random Eight-Neighbor* – Rho8 de Fairfield e Leymarie (1991); *Digital Elevation Model Networks* – DEMON de Costa-Cabral e Burges (1994); e, *Deterministic Infinity* – D∞ de Tarboton (1997). Em vários trabalhos foram realizadas comparações do desempenho dos algoritmos, tais como as pesquisas de Wilson, Lam e Deng (2007) e Crombez (2008).

De forma geral, o método comumente utilizado por este tipo de algoritmo baseia-se em estimar a área de captação, ou área de contribuição, que consiste na somatória das áreas superficiais das células em que o escoamento contribui para um ponto em questão (FERNÁNDEZ et al., 2012). Fernández et al. (2012) argumentaram que existem diversos algoritmos, conhecidos como algoritmos de fluxo, que realizam o cálculo da área de captação e que estão implantados como funcionalidades dos atuais sistemas SIG.

Wilson, Lam e Deng (2007) compararam o desempenho de algoritmos de determinação de fluxo usados em análises hidrológicas. De acordo com estes autores, cada um destes algoritmos oferece um único método para calcular direção de fluxo e podem resultar em diferentes representações para uma mesma paisagem.

Crombez (2008) também comparou a eficiência computacional e a validade dos resultados obtidos com o uso de algoritmos de fluxo. Neste trabalho, o autor aplicou algoritmos diversos para uma mesma área e analisou os resultados que variaram conforme o algoritmo utilizado. Na visão de Crombez (2008), os algoritmos de direção de fluxo são ferramentas de análise de terreno utilizadas para modelar a transferência de água, sedimentos, contaminantes ou nutrientes em toda a paisagem.

Crombez (2008) explicou que cada algoritmo define como a saída de um determinado ponto ou área será distribuída, e as eventuais disparidades entre os

vários algoritmos dependem, principalmente, da granularidade e dos tipos de escoamento permitidos (simples ou múltiplos). Os algoritmos de fluxo de direção simples permitem o fluxo de descarga para apenas uma célula vizinha com declividade mais baixa, simulando padrões de fluxos convergentes e são eficazes na identificação de redes de fluxo; enquanto que os algoritmos de fluxo de direções múltiplas permitem que a água seja descarregada para todas as células vizinhas na vertente, e são capazes de considerar fluxos convergentes e divergentes.

Crombez (2008) afirmou, ainda, que a capacidade de modelar a dispersão por toda a paisagem é uma característica desejável em muitas análises, porém a eficiência de recursos e requisitos computacionais são, muitas vezes, comprometidas em troca da capacidade de produzir um modelo mais realista. Para o autor, todos os algoritmos apresentam vantagens e desvantagens, e a escolha do algoritmo de roteamento de fluxo em um modelo é etapa importante visto que impacta nos cálculos de curva ascendente da área de contribuição, área específica de captação, índice de umidade, índice de fluxo de energia e vários outros atributos topográficos

3.1.3 Parâmetros morfométricos relacionados

Um dos primeiros trabalhos sobre a extração automática de redes de drenagem, a partir de MDE, foi apresentado por O'Callaghan e Mark (1984). Esta metodologia propõe o uso de um limiar, baseado na área mínima de contribuição, para identificar os pontos aonde a rede de drenagem se origina (nascentes). A área de contribuição de um ponto qualquer representa o número de pontos (ou área) que converge àquele determinado ponto. Métodos baseados na definição de um limiar de área de contribuição são muito simples de implementar e, portanto, muito populares. No entanto, em regiões com diferentes padrões geomorfológicos, a escolha de um único limiar para representar toda a região é extremamente complicada, podendo gerar redes de drenagem com maior ou menor densidade do que a real.

Todavia, outros critérios podem ser utilizados como base para a definição da rede de drenagem. Banon et al. (2013) exemplificaram critérios alternativos, citando a declividade e as curvaturas vertical e horizontal. Argumentaram, os autores, que a declividade e as curvaturas usadas como critério para definir a drenagem podem

indicar áreas de convergência ou divergência dos fluxos de água na superfície, auxiliando na identificação de regiões potencialmente associadas às nascentes.

Estudos detalhados de parâmetros morfométricos, que podem ser extraídos automaticamente, encontram-se nos trabalhos de Collares (2000) e Oliveira, Guasseli e Saldanha (2009).

No estudo de Strobl e Forte (2007), os autores já relatavam que, de forma geral, os procedimentos para a derivação automática das redes de drenagem a partir de MDE fazem uso apenas de variáveis topográficas. Para os autores, em detrimento das resoluções cada vez melhores dos modelos, aliadas com o avanço das ferramentas e técnicas de manipulação de dados espaciais que tornaram viável o uso de vários algoritmos para derivação da rede de drenagem, é notável que as redes geradas nem sempre coincidissem com a rede real.

Sugeriram, portanto, que não apenas as variáveis topográficas influenciam nos padrões de drenagem, e que outras variáveis ambientais não incluídas nestas metodologias precisariam ser consideradas. Semelhantemente, Vogt et al. (2003) também sugeriram o uso de variáveis ambientais diversas para a identificação de redes de drenagem.

Variáveis ambientais consideradas por Strobl e Forte (2007) envolveram temas como topografia, solo e litologia. Adotaram como estudo de caso duas regiões hidrológica e geograficamente distintas e observaram que a rede neural resultante apresentou diferentes grupos de parâmetros para cada uma destas áreas, o que pode sugerir que a rede proposta efetivamente incorporou a capacidade de aprender e responder de forma diversa para cada área de estudo.

Apesar do sucesso declarado para as áreas estudadas, suspeitaram os autores que para outras áreas, com diferentes condições climáticas, topográficas, ambientais e geológicas, poderão ser necessários fatores de entradas distintos para a delimitação da rede de drenagem.

De forma semelhante, recomendaram novas investigações para a incorporação de imagens oriundas de satélite multiespectrais com melhor resolução, visto que usaram imagens do satélite Landsat 7. Na visão de Strobl e Forte (2007), imagens digitais obtidas por sensoriamento remoto são candidatas em potencial para fornecer parâmetros relacionados a diversos fatores ambientais e poderiam ser úteis no processo de extração da rede de drenagens.

3.2 A DESCOBERTA DE CONHECIMENTO EM BANCOS DE DADOS ESPACIAIS

3.2.1 Banco de dados espaciais na era do Big Data

Banco de dados espaciais refere-se a qualquer conjunto de dados que descrevem as propriedades espaciais, semânticas e, possivelmente, temporais de fenômenos do mundo real (BÈDARD, 2005). De forma mais simplificada, banco de dados espaciais pode ser entendido como uma coleção de dados referenciados espacialmente e que atua como um modelo da realidade (REYYA; SUMALLIKA; VASUKI, 2013).

Para Yeung e Hall (2007), os sistemas de banco de dados espaciais tratam-se de tipos singulares de sistemas de banco de dados habilitados especificamente para gerenciar e processar dados espaciais. Tais sistemas suportam tipos de dados espaciais em seus modelos de dados e linguagens de consulta, além de fornecer métodos de indexação espacial e implementar algoritmos eficientes para junções espaciais (GUTTING, 1994).

Manalopoulos, Papadopoulos e Vassilakopoulos (2005) destacaram que o principal objetivo de um sistema de banco de dados espacial é o tratamento eficiente e eficaz dos tipos de dados espaciais em dois, três ou mais espaços dimensionais, e a capacidade para responder a consultas levando em consideração as propriedades dos dados espaciais.

A aplicação de banco de dados espaciais em estudos da Geografia permite a construção de banco de dados capazes de reunir e manipular diversos tipos de dados espaciais. Por exemplo, Ciampalini et al. (2015) relataram a estrutura e o conteúdo de um banco de dados espacial sobre o deslizamento de terras; já o banco de dados proposto por Ciampalini et al. (2015) reuniu dados vetoriais, como geologia, topografia e edificações, além de dados matriciais como MDE e imagens SAR.

Não obstante, a evolução tecnológica dos sistemas de banco de dados espaciais tornou possível o armazenamento de grandes volumes de dados espaciais, ao passo que proliferaram os bancos de dados disponíveis na internet. O compartilhamento dos dados espaciais na rede mundial de computadores foi impulsionado pela criação, disseminação e adoção de padrões que visam a

interoperabilidade de sistemas e facilitam o intercâmbio de dados entre produtores e usuários.

Camboim (2013) abordou a questão da interoperabilidade nas Infraestruturas de Dados Espaciais – IDE e, adicionalmente, discutiu a adoção de uma estratégia de integração de dados geoespaciais da INDE do Brasil e dados abertos por meio de buscas semânticas. O uso de ontologias para a busca de dados espaciais também foi discutido por Camboim e Sluter (2013).

López et al. (2015) propuseram um banco de dados espaciais, com dados ambientais diversos do México, usando tecnologia Big Data para armazenar e recuperar grandes volumes de dados matriciais. De acordo com Chen, Mao e Liu (2014), o termo Big Data é usado para descrever grandiosos conjuntos de dados, uma vez que oportuniza descobrir novos valores, contribui na compreensão de valores ocultos e apresenta novos desafios como organizar e gerir, de forma eficaz, esses conjuntos volumosos de dados. Conceito e características de Big Data foram, também, discutidos por Gandomi e Haider (2015) e Russom (2013).

Ozkose et al. (2015) apresentaram e discutiram definições para Big Data formuladas por variados autores. Destaca-se um conceito abrangente de Big Data que pode ser encontrado em Gantz e Reinsel (2011, p. 6), onde os autores utilizaram o termo para descrever “uma nova geração de tecnologias e arquiteturas, concebidas para economicamente extrair valor de volumes muito grandes de uma ampla variedade de dados, permitindo alta velocidade na captura, descoberta e/ou análise”.

Baseando-se em tal conceito, Chen, Mao e Liu (2014) afirmaram que as características essenciais do Big Data podem ser resumidas em volume (grande volume), variedade (várias modalidades), velocidade (geração rápida) e valor (grande valor, mas baixa densidade). Kitchin (2013) elaborou uma revisão de literatura e condensou o conceito de Big Data nos seguintes termos:

- Quanto ao volume, trata-se de grandes quantidades de dados;
- Quanto à velocidade de obtenção/produção são criados em, ou quase, tempo real;
- Quanto à variedade, podem ser dados estruturados ou não estruturados;
- Quanto ao escopo, geralmente, representam populações ou sistemas inteiros;

- Apresentam melhores resoluções e são fortemente indexáveis, ou seja, potencialmente identificáveis de forma unívoca;
- Relacionais por natureza, contendo campos comuns que permitem a junção de diferentes conjuntos de dados; e,
- Flexíveis, preservando as características de extensionalidade e escalabilidade, isto é, podem ser adicionados novos campos facilmente e podem expandir-se em tamanho rapidamente.

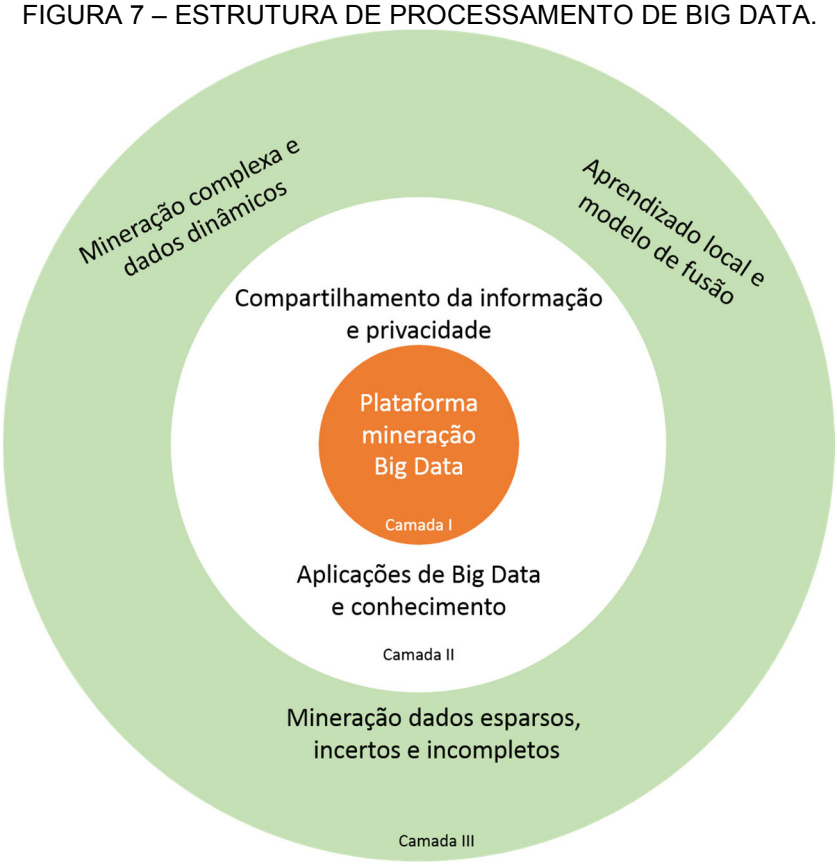
Além do aspecto relacionado ao grande volume de dados inerente ao Big Data, Wu et al. (2014) destacaram outras duas características interessantes que dizem respeito às possibilidades de origem dos dados, notadamente provenientes de fontes autônomas com controle distribuído e descentralizado, e à complexidade dos relacionamentos entre os dados. Para Wu et al. (2014), as fontes de Big Data são autônomas, cada uma delas podendo gerar e coletar informações sem a necessidade de controle centralizado. Os mencionados autores explicaram que no conceito de Big Data o foco passa de apenas encontrar os melhores valores que representam cada observação do universo de discurso para explorar o relacionamento entre os dados, ou seja, explorar as potenciais conexões entre os conjuntos de dados.

É relevante a observação de Chen, Mao e Liu (2014) sobre as consequências do conceito de Big Data no tocante à forma de análise de dados. Os autores defenderam a ideia de que o Big Data desencadeia uma revolução no pensamento, exemplificada pelo fato de: usar todos os dados disponíveis na condução de uma análise; usar grandes quantidades de dados, ao invés de preferir apenas poucos dados acurados; atentar-se para as correlações no lugar de explorar relacionamentos causais; e, usar resultados analíticos em substituição aos especialistas.

Para Kitchin (2014), o Big Data oportuniza novas abordagens para geração e análise de dados, onde a ênfase recai não mais em extrair conhecimento de conjuntos de dados limitados por escopo, temporalidade e tamanho, mas na manipulação de grandes conjuntos de dados, dinâmicos e variados. Na esteira do Big Data, despertou-se a necessidade de novas formas de gestão de dados, técnicas analíticas que dependem de aprendizado de máquina e novos modos de visualização (KITCHIN, 2014).

Wu et al. (2014) propuseram uma visão conceitual da estrutura de processamento do Big Data, representada graficamente conforme a FIGURA 7. Para cada camada, os autores discutiram o desafio da mineração de dados com o Big Data. Na camada I são tratados aspectos relativos ao acesso e à capacidade computacional para lidar com o crescente volume de dados, por vezes armazenados de forma distribuída; na camada II são considerados a privacidade dos dados e os aspectos semânticos relacionados aos domínios para diferentes aplicações; e na camada III os desafios se concentram nos algoritmos de mineração e seu uso considerando as características fundamentais do Big Data.

FIGURA 7 – ESTRUTURA DE PROCESSAMENTO DE BIG DATA.



FONTE: Adaptado pelo autor, a partir de Wu et al. (2014, p. 99).

Graham e Shelton (2013) discutiram o uso do Big Data especificamente na área da Geografia. Cugler et al. (2013) e Evans et al. (2014) adotaram o termo *Spatial Big Data* para discutirem a tecnologia e sua aplicação no trato de dados espaciais. Ainda, encontra-se no trabalho de Vitolo et al. (2015) uma revisão da literatura sobre ferramentas para o processamento de dados ambientais no âmbito de Big Data.

Goodchild (2013) chamou atenção para questões relativas à qualidade dos dados que denominou como *Big Geodata*, como por exemplo a linhagem. Para o autor, em algumas situações, o Big Data pode se caracterizar pela carência de processos normais de controle de qualidade, documentação e amostragem rigorosa. Defendeu, também, a necessidade de investigar a questão da qualidade destes conjuntos de dados.

Bravo e Sluter (2015) demonstraram preocupação semelhante e discutiram a qualidade de dados espaciais no cenário atual de produção de dados espaciais, ou seja, no momento em que a produção de tais dados ocorre não somente por meio de instituições oficiais ou agências governamentais, mas também de forma coletiva, voluntária e colaborativa por meio dos usuários produtores, conectados pela Internet.

3.2.2 Descoberta de conhecimento em bancos de dados

Na visão de Fayyad, Piatetsky-Shapiro e Smyth (1996), a descoberta de conhecimento em bancos de dados (*Knowledge-Discovery in Databases* – KDD) refere-se ao processo global de descoberta de conhecimento útil a partir dos dados. De acordo estes autores, o termo descoberta de conhecimento em bancos de dados enfatiza que o produto final da descoberta de uma abordagem dirigida pelos dados é o conhecimento. Numa definição mais formal, KDD é o processo não trivial de extração de padrões válidos, novos, potencialmente úteis e compreensíveis a partir de uma bancos de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Para Maimon e Rocach (2010), KDD consiste na modelagem e análise exploratória automatizada de grandes volumes de dados.

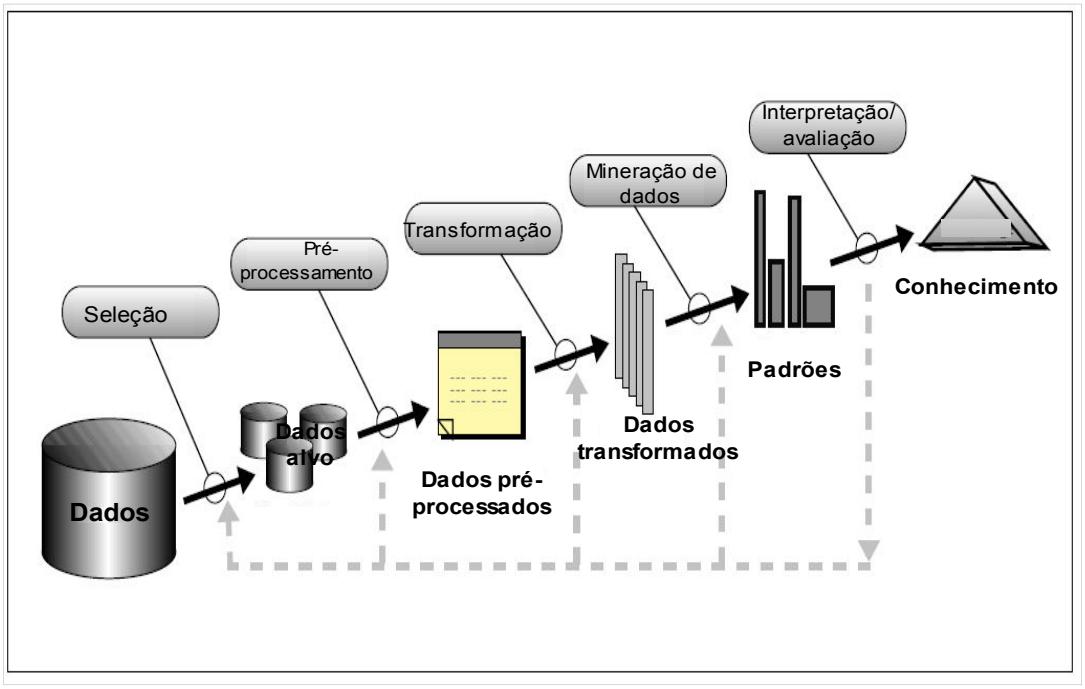
Sobre a natureza interdisciplinar do KDD, Fayyad, Piatetsky-Shapiro e Smyth (1996) argumentaram tratar-se de uma disciplina que surge da interseção de campos de pesquisa, como aprendizagem de máquina, reconhecimento de padrões, bancos de dados, estatística, inteligência artificial, aquisição de conhecimento para sistemas especialistas, visualização de dados e computação de alto desempenho.

Encontram-se relatadas na literatura inúmeras aplicações de KDD em diversas áreas do conhecimento, como na astronomia, geologia, marketing, finanças, detecção de fraudes, manufatura, telecomunicações e internet (FAYYAD;

PIATETSKY-SHAPIRO; SMYTH, 1996; GOEBEL; GRUENWALD, 1999; FRIEDMAN, 2010).

A FIGURA 8, a seguir, apresenta uma visão geral do processo de KDD, conforme o entendimento de Fayyad, Piatetsky-Shapiro e Smyth (1996). Os autores demonstraram, com a ilustração, que a mineração de dados pode ser considerada uma etapa no processo global de KDD.

FIGURA 8 – VISÃO GERAL DOS PASSOS DO PROCESSO DE KDD.



FONTE: Traduzido pelo autor a partir de Fayyad, Piatetsky-Shapiro e Smyth (1996, p. 41).

A natureza iterativa e interativa do processo de KDD foi destacada por Maimon e Rocach (2010), que descreveram nove etapas que devem ser ajustadas para cada tipo de aplicação, argumentando sobre a necessidade de conhecer as possibilidades em cada um dos nove passos do processo. Em outras palavras, mencionaram aspectos “artísticos” do processo de KDD, visto que não é possível apresentar uma fórmula ou fazer uma taxonomia completa para as escolhas certas para cada tipo de passo e aplicação” (MAIMON; ROCACH, 2010). Nos próximos parágrafos apresentam-se os nove passos do processo de KDD, segundo a visão de Maimon e Rocach (2010).

O primeiro passo do processo possui caráter preparatório e prevê a compreensão do domínio da aplicação, quer dizer, o necessário entendimento do universo de discurso que será foco da aplicação do KDD. Este passo inicial requer

que as pessoas envolvidas no processo o entendam e definam os objetivos do usuário final e do ambiente em que o processo do KDD será inserido, incluindo os conhecimentos prévios relevantes ao entendimento do problema. Dada a natureza iterativa e interativa de todo processo, a revisão e afinação deste passo poderá ocorrer *a posteriori*.

Após a compreensão dos objetivos do KDD, segue-se com o pré-processamento dos dados, abrangidos nas segunda, terceira e quarta etapas do processo. O segundo passo do processo prevê a seleção e a criação de um conjunto de dados no qual a descoberta de conhecimento será realizada. Incluem-se nesta etapa a identificação dos dados disponíveis, a obtenção de dados adicionais, a integração dos dados e o estabelecimento dos atributos que serão considerados no processo. Ressalta-se a importância desta etapa, uma vez que as técnicas de mineração de dados aprendem e descobrem a partir do conjunto de dados disponível, que pode ser avaliado como a base de evidência para a construção de modelos. Neste momento, deve-se garantir que os atributos importantes estão presentes no conjunto dos dados, sob pena de comprometer o sucesso do processo.

Se por um lado é aconselhado dispor do maior número possível de atributos na fase de seleção e criação do conjunto de dados, por outro não se pode desprezar o custo da coleta e organização deste repositório de dados. Novamente, deve-se perceber as sutilezas dos aspectos relacionados à iteratividade e interatividade inerentes ao processo, o que permite iniciá-lo com os melhores dados disponíveis e depois acrescentar novos dados ao conjunto inicial, e observar o efeito em termos de descoberta de conhecimento e modelagem do problema (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; MAIMON; ROCACH, 2010; FRIEDMAN, 2010).

O pré-processamento e a limpeza são atividades previstas no terceiro passo do processo, cuja motivação é melhorar a confiabilidade dos dados. Envolve manipular os dados, de forma a eliminar valores nulos, remover ruídos e tratar os dados faltantes ou ausentes que são úteis no contexto da exploração do universo de discurso. Já na fase do pré-processamento, técnicas de mineração de dados podem ser utilizadas visando resolver os problemas e aumentar a confiabilidade do conjunto de dados. Trata-se de uma etapa importante e, por vezes, muito custosa em termos de tempo de trabalho (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; MAIMON; ROCACH, 2010; FRIEDMAN, 2010).

O quarto passo prevê a transformação de dados, procurando gerar dados melhores para a etapa de mineração. Consiste em uma etapa muito específica para cada projeto, porém, de forma geral, utilizam-se métodos como seleção de atributos, re-amostragem e normalização de atributos. Mesmo no caso de não serem conduzidas transformações dos dados no início, as próximas etapas do processo (sobretudo a mineração de dados) podem indicar a necessidade de proceder tais transformações. Ou seja, o processo de KDD poderá resultar numa melhor compreensão das transformações necessárias ao conjunto dos dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; MAIMON; ROCACH, 2010; FRIEDMAN, 2010).

Os quatro passos seguintes, quer dizer as etapas quinta, sexta, sétima e oitava são referentes à mineração de dados e focarão nos algoritmos empregados para cada projeto. No quinto passo, decide-se por qual tipo de mineração de dados será aplicado, quer seja verificação, classificação, regressão ou descrição. A escolha do algoritmo de mineração de dados é o sexto passo do processo, que compreende selecionar o método específico para ser utilizado, tais como Redes Neurais e Árvore de Decisão. Deve-se refletir sobre qual é o algoritmo mais apropriado, considerando as premissas estabelecidas na fase inicial, bem como os parâmetros e táticas de aprendizado de cada algoritmo.

O passo seguinte, sétima etapa do processo, consiste na aplicação do algoritmo de mineração escolhido, podendo ser necessário aplicar o algoritmo repetidas vezes até que um resultado satisfatório seja alcançado. O oitavo passo do processo consiste na etapa de avaliação e interpretação dos padrões resultantes da mineração de dados, relativamente aos objetivos definidos no primeiro passo. Neste momento, avaliam-se as etapas de pré-processamento e seus efeitos nos resultados obtidos com a aplicação do algoritmo de mineração. Esta etapa foca na compreensão e na utilidade do modelo inferido, sendo que o conhecimento descoberto está documentado para uso posterior.

O último passo do processo prevê o uso do conhecimento descoberto na mineração de dados. Nesta etapa, o conhecimento está pronto para ser incorporado em outros sistemas para futuras ações. É possível realizar alterações nos parâmetros escolhidos nas etapas anteriores e medir os efeitos na geração de novos conhecimentos, sendo que o sucesso deste passo determina a eficácia global do processo de KDD.

Segundo Goebel e Gruenwald (1999), é comum uma confusão sobre os termos descoberta de conhecimento em bancos de dados e mineração de dados. Os autores defenderam o uso do termo descoberta de conhecimento em bancos de dados para denotar todo o processo de transformar dados de baixo nível para conhecimento de alto nível, ainda que mineração de dados seja a extração de padrões ou modelos a partir de dados observados. Maimon e Rocach (2010) também situam a mineração de dados como uma etapa do KDD, pois afirmaram ser a mineração de dados o núcleo do processo de KDD, envolvendo a inferência de algoritmos para explorar os dados, desenvolver o modelo e descobrir padrões anteriormente desconhecidos.

Goebel e Gruenwald (1999) afirmaram, ainda, que apesar de ser considerado o núcleo do processo de descoberta de conhecimento, a etapa de mineração de dados corresponde a, aproximadamente, 15 a 20% do esforço global.

Uma taxonomia para os métodos de mineração de dados foi discutida por Maimon e Rocach (2010), que argumentaram sobre sua contribuição no entendimento da variedade, das inter-relações e do agrupamento dos métodos. Dois tipos principais de mineração foram destacados por Maimon e Rocach (2010), a saber: orientados à verificação, onde o sistema verifica as hipóteses do usuário; e, orientados à descoberta, onde o sistema procura novas regras e padrões de forma autônoma.

Sendo assim, métodos de verificação tratam da avaliação de uma hipótese proposta por uma fonte externa, e incluem os tradicionais métodos mais comuns da Estatística (MAIMON; ROCACH, 2010).

Métodos de descoberta são aqueles que identificam automaticamente padrões nos dados, e se subdividem em métodos de previsão e métodos de descrição. Os métodos de descoberta podem ser diferenciados em métodos descritivos, orientados à interpretação dos dados e se concentram na compreensão e na forma como os dados se relacionam, e métodos orientados à previsão, que visam a construção automática de um modelo comportamental que obtém amostras novas e ocultas, além de serem capazes de prever valores de variáveis relacionadas com a amostra (MAIMON; ROCACH, 2010).

Grande parte das técnicas de mineração de dados orientados à descoberta baseiam-se na aprendizagem indutiva, onde um modelo é construído a partir da generalização de um suficiente número de amostras de treinamento, com o

pressuposto de que o modelo gerado possa ser aplicável a amostras desconhecidas. Na visão de Maimon e Rocach (2010), no contexto da mineração de dados interessam, particularmente, os métodos orientados à descoberta, visto que em geral, na maioria dos problemas tratados com mineração de dados, a preocupação recai sobre a descoberta de uma hipótese, ao invés do teste da hipótese.

Gilbert et al. (2010) argumentaram que os principais parâmetros levados em conta para a escolha da técnica de mineração de dados são: o objetivo do problema a ser resolvido e a estrutura dos dados disponíveis. Chikorora (2014) descreveu algoritmos de mineração de dados e fatores a serem considerados na escolha dos algoritmos. A revisão de literatura realizada por Chikorora (2014) adicionou outros fatores aos citados por Gilbert et al. (2010), ou seja, acrescentou os resultados esperados, o tipo de informação que deverá ser usada, a familiaridade com algum algoritmo e, finalmente, os parâmetros de configuração exigidos.

3.2.3 Mineração de dados geoespaciais

Friedman (2010) afirmou que a mineração de dados é um campo vagamente definido, onde as definições são dependentes das experiências e visões dos autores. O autor elencou algumas definições encontradas na literatura, das quais se destacam as seguintes:

- Um processo de extração, a partir de grandes bases de dados, de informações previamente não conhecidas, visando usar tais informações em tomadas de decisão;
- Um conjunto de métodos usados no processo de descoberta de conhecimento para distinguir relacionamentos previamente não conhecidos e padrões entre os dados;
- Um processo de descobrir padrões vantajosos em dados; e,
- Um processo de suporte à decisão se procura em grandes bases de dados por padrões de informação não conhecidos e não esperados.

Para Friedman (2010), mineração de dados (*Data Mining*) é uma disciplina que tem suas origens na Estatística. Entretanto, Hand (1999) defendeu tratar-se de duas disciplinas distintas e que mineração de dados também faz uso de ideias,

ferramentas e métodos de outras áreas – especialmente as áreas de computação, como a tecnologia de banco de dados e a aprendizagem de máquina. Mineração de dados espaciais é o processo de extração de regras de conhecimento, espaciais e não espaciais, a partir de dados espaciais (LI; WANG, 2006).

Na definição formulada por Aldridge (2006), mineração de dados espaciais é um domínio especializado da mineração de dados cujo objetivo é encontrar conhecimento latente ou implícito em dados espaciais. Aldridge (2006) também argumentou sobre os desafios desta técnica, lembrando da natureza intrínseca dos dados espaciais e destacando a grande quantidade de dados espaciais, muitas vezes produzidos por sensoriamento remoto e sistemas de informação geográfica.

Paidi (2012) definiu a mineração de dados espaciais como a extração de conhecimento implícito, relações espaciais ou outros padrões não explicitamente armazenados em bancos de dados. Guo e Mennis (2009) discutiram a mineração de dados espaciais e a descoberta de conhecimento geográfico, argumentando sobre a necessidade de métodos eficientes e eficazes para extrair informações desconhecidas e inesperadas a partir do conjunto de dados de grandes volumes e de alta complexidade.

Na visão dos autores, os avanços da última década permitiram acesso a dados geográficos de alta qualidade para incorporar a informação e a análise espacial em diversos estudos, destacando espacialmente a ampliação das aplicações da tecnologia de GPS (*Global Positioning System*); o compartilhamento de dados espaciais e mapeamentos na *web*; o sensoriamento remoto de alta resolução; e, os serviços baseados em localização. Os autores apresentaram tarefas e métodos de mineração de dados espaciais incluindo classificação, regras de associação, *clustering* e geovisualização multivariada.

Para Shekhar et al. (2004), é mais difícil extrair padrões interessantes e úteis a partir de um conjunto de dados espaciais do que extrair padrões de um conjunto alfanumérico tradicional devido à complexidade dos tipos de dados geográficos, relacionamentos espaciais e auto correlação espacial. Neste sentido, a aplicação da mineração de dados espaciais às Ciências da Terra foi discutida nos trabalhos de Shi e Yang (2012) e de Gotz et al. (2015).

Pradhan et al. (2008) discutiram a aplicação de um modelo de mineração de dados para mapear o risco de deslizamento de terras. Para o estudo, dados topográficos, geológicos e imagens de satélite foram coletados e processados

utilizando SIG e ferramentas de processamento de imagem. Os autores realizaram inferências em bases de dados diversas, tais como topográfica, geológica, de solo e de chuvas, e analisaram os fatores em ambiente de mineração de dados. Os resultados mostraram aproximação entre o mapa gerado e os dados sobre deslizamentos existentes.

Miah (2011) avaliou os métodos de mineração de dados no planejamento de evacuações em situações de emergência. Argumentou o autor que ferramentas eficientes são necessárias para produzir planos que identifiquem as rotas e os horários para evacuar, rápida e eficazmente, populações afetadas por este tipo de situação. O autor, então, sistematizou conhecimentos acerca da aplicação da mineração de dados para o problema das evacuações emergenciais e discutiu desafios e oportunidades desta tecnologia.

No estudo de Svoray et al. (2011) foi demonstrada a possibilidade do uso de mineração de dados na predição de queda de barrancos. Estabeleceram um processo de mineração de dados baseado em árvores de decisão que foi aplicado para identificar áreas de risco de início da queda de barranco. Foi utilizado um banco de dados espacial com dados ambientais, climáticos e antropização. Os autores afirmaram que os resultados demonstram melhor capacidade preditiva da mineração de dados em comparação com a técnica *Analytical Hierarchy Processes* – AHP (processos hierárquicos analíticos) e aquelas tradicionais oriundas da topografia.

No trabalho de Appice, Lanza e Malerba (2007) foi proposta uma arquitetura de um sistema cliente-servidor que integra as tecnologias de Sistemas Gerenciadores de Banco de Dados – SGBD, SIG e mineração de dados espaciais para suportar a interpretação de mapas topográficos.

Date (2004), Elmasri e Navathe (2011) e Silberschatz, Korth e Sudarshan (2012) discutiram, com propriedade, as vantagens do uso de SGBD ao invés do simples armazenamento de arquivos em diretórios de um sistema operacional, tais como a facilidade na gestão de usuários e dos acessos aos dados, o controle da redundância, a independência de dados, a oportunidade de usar restrições de integridade e a melhoria na segurança da informação.

Outra oportunidade que surge é a utilização dos recursos computacionais do equipamento onde o SGBD encontra-se instalado para a realização de operação sobre os dados. Geralmente, sobretudo em ambientes corporativos, o SGBD é instalado em equipamentos servidores com configuração robusta e melhores

recursos do que as estações de trabalho. A extensão PostGIS, por exemplo, proporciona um rico conjunto de funções e operadores para, usando a linguagem de consulta SQL, manipular dados espaciais e executar desde simples consultas, passando pela extração de parâmetros morfométricos a partir de MDE, até complexas operações de análise espacial envolvendo dados vetoriais e matriciais.

De forma semelhante, o trabalho de Xu, Qi e Wang (2008) demonstrou a integração das tecnologias de mineração de dados, SIG e sensoriamento remoto para a construção de um modelo para a mineração de dados espaciais relacionados ao meio ambiente regional. Zhan e Yang (2009) propuseram a utilização da tecnologia de mineração de dados para reconhecer informações geológicas desfavoráveis, considerando o risco geológico. No trabalho, os autores discutiram os bancos de dados geológicos orientados para a mineração de informações geológicas e os métodos de mineração de dados espaciais geológicos.

O estudo de Fuqiang e Gangcai (2009) explorou diferentes níveis socioeconômicos e fatores ambientais que influenciam a adoção de medidas de conservação de solo e água. O objetivo principal do trabalho foi investigar os fatores influentes das medidas de conservação de solo e água, e fornecer informações úteis aos tomadores de decisão. A análise de fatores, com técnicas de mineração de dados, foi utilizada no referido trabalho e os resultados mostraram que a adoção de tais medidas foi influenciada principalmente por quatro fatores, incluindo a eficiência econômica, a fixação humana e uso da terra, o ambiente natural e a qualidade do solo.

Gilbert et al. (2010) discutiram diversas técnicas de mineração de dados aplicados, especificamente, a sistemas ambientais. Neste trabalho, os autores propuseram boas práticas, apresentaram softwares e discutiram desafios da mineração de dados nesta área específica do conhecimento.

3.2.3.1 Produtos de sensoriamento remoto utilizados na mineração de dados

A Mineração de dados também pode ser útil na manipulação de dados oriundos de imagens. Para Zhang, Hsu e Lee (2001), a mineração de dados em imagens denota a sinergia das tecnologias de processamento de imagens e mineração de dados para ajudar na análise e na compreensão de um domínio em imagens. Segundo os autores, é um esforço interdisciplinar que baseia-se em

experiências dos ramos de visão computacional, processamento de imagem, mineração de dados, aprendizado de máquina, banco de dados e inteligência artificial.

A mineração de dados em imagens lida com a extração de conhecimento implícito nos dados da imagem, ou padrões não explícitos em imagens armazenadas, ou entre estas e dados alfanuméricos. Portanto, o objetivo da mineração de dados em imagens é a descoberta de padrões significativos em um determinado conjunto de imagens e o seu relacionamento com dados alfanuméricos (ZHANG; HSU; LEE, 2001). Os autores afirmaram, ainda, que a mineração de dados em imagens visa obter, por meio da extração de conhecimento implícito, relacionamentos entre os dados da imagem ou outros padrões não explicitamente armazenados nas bases de dados de imagens.

Neste contexto, o sensoriamento remoto fornece insumos valiosos para a mineração de dados em imagens, como por exemplo, as imagens de radar. Os sistemas de micro-ondas ativos, conhecidos como radar, são baseados na transmissão de micro-ondas de comprimentos de onda mais longos e na detecção da quantidade de energia retroespalhada na superfície do terreno (JENSEN, 2009).

Segundo Jensen (2009), sensores remotos ativos, como é o caso do radar, não são dependentes da energia eletromagnética do Sol ou das propriedades termais da Terra. Explicou o autor que os sensores remotos ativos geram sua própria energia eletromagnética que: é transmitida do sensor para a superfície do terreno sendo pouco afetada pela atmosfera; interage com o terreno produzindo um retroespalhamento da energia; e, é registrada pelo sensor remoto.

Ainda consoante o autor, as imagens de radar obtidas de aeronaves ou satélites, nos dias atuais, são mapeadas em faixas contínuas do radar aerotransportado de visada lateral (*Side-looking Airbone Radar* – SLAR). A tecnologia de Radar de Abertura Sintética (*Synthetic Aperture Radar* – SAR) é um tipo de SLAR que se difere pela capacidade de imageamento simultâneo com a banda L, nas polarizações HH, VV, VH e HV, e banda X (HH), com resoluções espaciais de 3 m, 6 m e 18 m (BRASIL, 2008). O sensor ativo utilizado permite operações em qualquer hora do dia e em condições meteorológicas adversas.

Castro Filho e Santos (2010) analisaram o potencial de dados de imagens SAR para classificação de uso do solo. Utilizaram técnicas de KDD visando identificar os atributos adequados para discretizar as classes de uso do solo, e

verificaram que as bandas da imagem SAR se mostraram adequadas para serem usadas como atributos na atividade de classificação.

Imagens obtidas por meio de sensores óticos, como as imagens do satélite SPOT 5, são frequentemente consideradas em trabalhos que envolvam a análise de mudanças ao longo do tempo. O satélite SPOT 5 foi lançado em maio de 2002 com bandas no visível, infravermelho próximo e infravermelho de ondas curtas, com 10 metros de resolução espacial e uma banda pancromática de 2,5 metros (JENSEN, 2009).

De acordo com Jensen (2009), os sensores HRV do SPOT operam em dois modos nas porções do visível e do infravermelho refletido no espectro: o primeiro é um modo pancromático correspondendo à observação numa banda espectral ampla e o segundo é um modo multiespectral compreendendo a observação em quatro bandas relativamente estreitas.

Chuanli, Xiaosheng e Qiumin (2013) aplicaram métodos de classificação de imagens multi-temporais, baseados em mineração de dados espaciais, ao monitoramento de áreas úmidas.

O Shuttle Radar Topography Mission – SRTM também é considerado como um bom insumo para mineração de dados em imagens. O SRTM consistiu em um sistema de radar modificado que voou a bordo do ônibus espacial Endeavour durante uma missão no ano 2000 (NASA, 2013). O objetivo da missão era a obtenção de dados de elevação em uma escala quase global para gerar o banco de dados de alta resolução topográfica digital mais completo da Terra.

A pesquisa de Banon et al. (2013) foi direcionada para a definição de critérios a partir da mineração de dados para a extração automática de redes de drenagem. Os autores tomaram como base atributos extraídos do SRTM e propuseram uma metodologia para a extração automática de uma rede de drenagem capaz de representar áreas com diferentes padrões geomorfológicos.

Ainda no trabalho de Banon et al. (2013), a mineração de dados foi usada com o objetivo de definir o conjunto de atributos mais representativo da rede de drenagem. Os autores calcularam atributos morfométricos e baseados na direção de fluxo; além disso, optaram por usar o algoritmo J48, um classificador baseado em Árvore de Decisão, pois afirmaram que na fase de avaliação do pós-processamento este foi o que apresentou um resultado qualitativo superior aos demais classificadores da mesma categoria.

3.2.4 Redes neurais artificiais

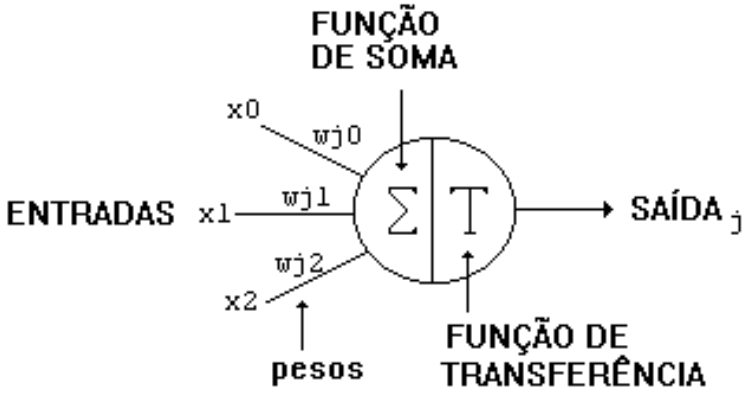
Pesquisadores da área de Inteligência Artificial preocuparam-se em criar RNAs de modo a simular, em ambiente computacional, o funcionamento dos neurônios do cérebro humano e dotar os computadores da habilidade de aprender. McCulloch e Pitts desenvolveram, no ano de 1943, um modelo simplificado de neurônio e, desde aquela época, diversos avanços conduziram para a neurociência computacional (RUSSEL; NORVIG, 2004).

Rede Neural Artificial – RNA consiste de um algoritmo computacional que representa um modelo matemático inspirado na estrutura neural de organismos inteligentes e que procura simular, em computadores, o funcionamento do cérebro humano (SPORL; CASTRO; LUCHIARI, 2011). Para Sporl, Castro e Luchiari (2011), a intenção é que, assim como o cérebro humano, a rede seja capaz de aprender e tomar decisões baseadas em seu próprio aprendizado. Argumentam os autores que, desta maneira, a RNA pode ser interpretada como um esquema de processamento capaz de armazenar conhecimento baseado em aprendizagem e disponibilizar este conhecimento para a aplicação em questão.

Russel e Norvig (2004) explicaram que as redes neurais são compostas por nós conectados por vínculos orientados. Argumentaram os autores que o vínculo do nó j para o nó i é útil para propagar a ativação aj até i; cada vínculo tem um peso numérico Wj, i associado a ele, o que determinará a intensidade e o sinal da conexão.

Ainda de acordo com Russel e Norvig (2004), cada nó i calcula primeiro a soma ponderada de suas entradas e, então, aplica uma função de ativação g para derivar a saída. Tal função de ativação g, que não deve ser linear, precisa ser hábil para permitir que o nó seja ativo quando as entradas corretas forem recebidas (nesse caso indicado com valor próximo de +1) e inativo quando as entradas erradas forem recebidas (nesse caso indicado com valor próximo a 0). A FIGURA 9 apresenta um esquema simplificado de um neurônio artificial.

FIGURA 9 – REPRESENTAÇÃO SIMPLIFICADA DE UM NEURÔNIO ARTIFICIAL.



FONTE: Russel e Norvig (2004, p. 9).

Segundo Strobl e Forte (2007), numa abordagem mais prática, redes neurais consistem de um grupo de equações matemáticas interconectadas que são alimentadas com dados de entrada e que calculam uma saída baseada nessas entradas; sendo que para cada nó da rede os sinais de entrada dos outros nós são totalizados por meio de uma soma ponderada e comparados com um limiar, caso uma função limiar tenha sido definida como função de transferência.

Desta maneira, os autores argumentaram que o algoritmo de aprendizagem de uma rede neural tenta alcançar um ajustamento dos pesos, de modo que ele será capaz de corresponder a um padrão de saída quando for alimentado com o correspondente padrão de entrada. O ajuste dos pesos dos neurônios ocorre de forma iterativa, e acontecerá até se atingir uma situação satisfatória no comparativo do vetor de saída produzido com o vetor desejado.

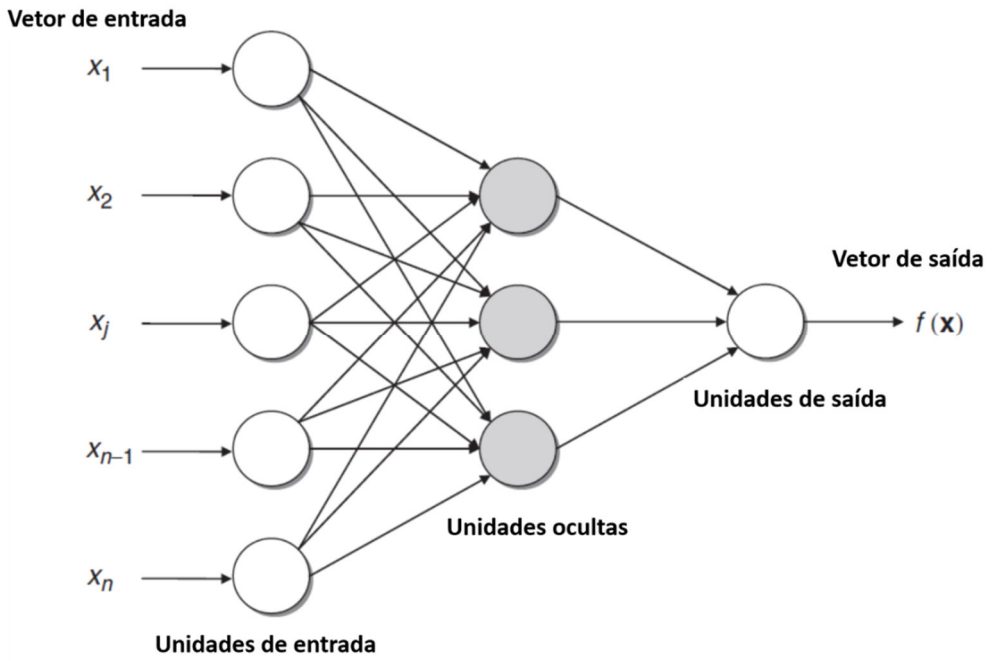
Destacam-se duas categorias de estruturas de redes neurais, a saber: redes de alimentação direta ou acíclicas e redes concorrentes ou cíclicas. Uma rede de alimentação direta representa uma função de sua entrada atual, e não possui nenhum estado interno além dos seus pesos; enquanto que uma rede concorrente utiliza suas saídas para retroalimentar suas próprias entradas, significando que os níveis de ativação da rede formam um sistema dinâmico que pode atingir um estado estável, exibir oscilações ou comportar-se de maneira caótica (RUSSEL; NORVIG, 2004).

Panchal et al. (2011) argumentaram que as redes Multilayer Perceptrons – MLP, com alimentação adiante (*feedforward*) e normalmente treinadas com algoritmo de retropropagação (*backpropagation*), são muito populares. Tratam-se de

redes supervisionadas que aprendem como transformar os dados de entrada em respostas desejadas.

No exemplo simplificado de uma RNA MLP apresentado por Vercellis (2009), ilustrado na FIGURA 10 a seguir, é possível observar como são organizadas as múltiplas camadas e a distribuição de unidades (ou nós) nas camadas. As unidades de entrada recebem os valores de entrada correspondentes aos atributos para cada observação. Unidades ocultas estão conectadas com unidades de entrada, de saída ou mesmo outras unidades ocultas, e transformam os valores de entrada no interior da rede. Unidades de saída recebem conexões de unidades ocultas ou de unidades de entrada e retornam valores de saída que correspondem à predição da variável de resposta.

FIGURA 10 – EXEMPLO DE RNA MLP.



FONTE: Vercellis (2009, p. 261).

Na estruturação de uma RNA, estratégias diversas podem ser usadas para definir a quantidade de camadas e unidades de cada camada (BENGIO; LECUN, 1995; HAN; KAMBER, 2006; VERCELLIS, 2009; GAUTAM; SANDHU; KHULLAR, 2011; PANCHAL et al., 2011; SRIVASTAVA et al., 2014). Russell e Norvig (2004) defenderam que a abordagem habitual consiste em realizar tentativas diversas e optar pela melhor delas, considerando os resultados obtidos com cada uma das configurações testadas.

Para Panchal et al. (2011), a definição e o treinamento de uma RNA MLP envolvem os seguintes passos:

- Selecionar a quantidade de camadas escondidas;
- Decidir a quantidade de neurônios para usar em cada camada escondida;
- Encontrar uma solução ótima global que evite mínimos locais;
- Convergir para uma solução ótima em um razoável período de tempo; e,
- Testar a rede quanto à superadaptação.

Russell e Norvig (2004) discutiram os fundamentos matemáticos inerentes às RNAs, cujas fórmulas serão descritas adiante. As RNAs são compostas por unidades conectadas por vínculos orientados, sendo que um vínculo da unidade *j* para a unidade *i* propaga a ativação a_j desde *j* até *i*. Os vínculos possuem peso numérico $W_{j,i}$ associados, que determinarão o sinal e a intensidade da conexão. A fórmula 1, apresentada abaixo, é utilizada para calcular uma soma ponderada das entradas de uma unidade *i*:

$$in_i = \sum_{j=0}^{n} W_{j,i}\, a_j \tag{1}$$

A fórmula 2 a seguir é aplicada para derivar a saída da unidade *i*, ou seja, corresponde à função de ativação *g* do neurônio artificial:

$$a_{i,j} = g(in_i) = g\left(\sum_{j=0}^{n} W_{j,i}\, a_j\right) \tag{2}$$

A função de ativação implementada no WEKA (software para Mineração de Dados utilizado nesta pesquisa) assume a forma de uma função sigmoide (ou logística). Neste caso, a fórmula 3 apresenta a função sigmoide:

$$f(x) = \frac{1}{(1+ e^{-x})} \tag{3}$$

A regra de atualização de pesos para a camada de saída é definida pela fórmula 4 a seguir:

$$W_{i,j} \leftarrow f(x) = W_{i,j} + \alpha\,.\, a_j\,.\,\Delta_i \tag{4}$$

A regra de propagação para os valores Δ é definida pela fórmula 5 abaixo:

$$\Delta = g'(in_j)\ \sum_i W_{j,i}\Delta_i \tag{5}$$

A regra de atualização de pesos entre unidades de entrada e unidades ocultas é definida pela fórmula 6:

$$W_{k,j} \leftarrow W_{k,j} + \alpha\ .\ a_k\ .\Delta_j \tag{6}$$

onde α é a taxa de aprendizagem.

Russell e Norvig (2004) resumiram o processo de propagação de retorno da seguinte forma:

- Calcular os valores Δ para as unidades de saída, usando o erro observado; e,
- Começando na camada de saída, repetir as etapas a seguir para cada camada na rede, até ser alcançada a camada oculta conectada à camada de entrada:
    - Propagar os valores Δ de volta até a camada anterior; e,
    - Atualizar os pesos entre as duas camadas.

O erro quadrático é definido como mostrado na fórmula 7:

$$E = \frac{1}{2}\sum_i (y_i - a_i)^2 \tag{7}$$

onde a soma se refere aos nós da camada de saída.

Para obtenção do gradiente em relação a um peso específico $W_{j,i}$ na camada de saída, é usada a fórmula 8 derivada a seguir:

$$\frac{\partial E}{\partial W_{i,j}} = -(y_i - a_i)\frac{\partial a_i}{\partial W_{i,j}} = -(y_i - a_i)\frac{\partial g(in_i)}{\partial W_{i,j}} =$$
$$-(y_i - a_i)g'(in_i)\frac{\partial gin_i}{\partial W_{i,j}} - (y_i - a_i)g'(in_i)\frac{\partial}{\partial W_{i,j}}\left(\sum_j W_{j,i}a_j\right) \tag{8}$$

Para obtenção do gradiente em relação aos pesos de $W_{k,j}$ que conectam a camada de entrada à camada oculta, é usada a fórmula 9 derivada a seguir:

$$\begin{aligned}
\frac{\partial E}{\partial W_{k,j}} &= -\sum(y_i - a_i)\frac{\partial a_i}{\partial W_{k,j}} = -\sum(y_i - a_i)\frac{\partial g(in_i)}{\partial W_{k,j}} \\
&= -\sum(y_i - a_i)g'(in_i)\frac{\partial g(in_i)}{\partial W_{k,j}} = -\sum \Delta_i \frac{\partial}{\partial W_{k,j}}\left(\sum_j W_{j,i}a_j\right) \\
&= -\sum \Delta_i W_{j,i}\frac{\partial a_i}{\partial W_{k,j}} = -\sum \Delta_i W_{j,i}\frac{\partial g(in_i)}{\partial W_{k,j}} \\
&= -\sum \Delta_i W_{j,i}g'(in_i)\frac{\partial g(in_i)}{\partial W_{k,j}} \\
&= -\sum \Delta_i W_{j,i}g'(in_i)\frac{\partial}{\partial W_{k,j}}\left(\sum_k W_{k,j}a_k\right) \\
&= -\sum \Delta_i W_{j,i}g'(in_j)a_k = -a_k\Delta_j
\end{aligned} \tag{9}$$

É importante ressaltar que uma das peculiaridades das RNAs é que são consideradas caixas pretas, visto que não se tem o controle absoluto sobre seu aprendizado (TRICHAKIS; NIKOLOS; KARATZAS, 2011; SCHEIDT; BRUNETTO, 2011). Ao longo dos anos, ainda que sem consenso, pesquisas foram conduzidas no sentido de esclarecer o funcionamento interno das redes neurais e reverter esta concepção (BENITEZ; CASTRO; REQUENA, 1997; OLDEN; JACKSON, 2002; HEINERT, 2008).

Russel e Norvig (2004) e Banon (2013) são autores que optaram por outras técnicas que consideraram mais apropriadas e efetivas para conduzir análises de importância de variáveis individualizadas no resultado final de processos de classificação, tais como as árvores de decisão.

Sporl, Castro e Luchiari (2011) discutiram sobre o uso deste tipo de rede na resolução de problemas complexos. Para os autores, a utilização das RNAs na análise ambiental disponibiliza essa nova ferramenta para decisões complexas que envolvem muitos critérios, em que, por vezes, a seleção dos critérios, assim como a definição de seus pesos, são avaliações arbitrárias e subjetivas, dificultando o processo de análise.

A respeito das redes neurais, Zhang, Hsu e Lee (2001) afirmaram que são tolerantes a falhas e são boas para uso em reconhecimento de padrões e previsão de tendências. Os autores afirmaram, ainda, que em se tratando de conhecimento

limitado, os algoritmos de redes neurais são, frequentemente, usados para a construção de um modelo de dados.

A utilização de RNAs na questão da derivação de redes de drenagem foi pesquisada por Strobl e Forte (2007). Os autores discutiram a aplicação deste tipo de rede na exploração de diversos fatores ambientais, juntamente com informações extraídas de imagens de satélite para identificar os principais fatores indicativos da localização de redes de drenagem. Acreditavam os autores que as variáveis identificadas poderiam ser empregadas para obter uma derivação mais acurada das redes de drenagem, em comparação com os métodos tradicionais de extração baseados em MDEs.

Singh e Panda (2011) descreveram o uso de RNAs na modelagem hidrológica. Para este campo de aplicação, de acordo com os autores, é comum o uso do procedimento de treinamento e testes para encontrar a melhor estrutura para a rede neural, porém, argumentaram que se o conjunto de dados for inadequado a rede resultante poderá ser tendenciosa.

Assim, Singh e Panda (2011) propuseram, em sua pesquisa, o uso do procedimento de validação cruzada para estimar a performance da rede, considerando um conjunto de dados reduzido. Os autores desenvolveram modelos de redes neurais para obter a previsão diária de sedimentos, considerando como entradas a precipitação diária, a precipitação dos dias anteriores e as temperaturas diárias mínima e máxima.

O uso das RNAs conjuntamente com sensoriamento remoto também foi discutido por Andrade (2011), que aplicou esta técnica na classificação automática de dados de sensoriamento remoto visando a identificação e o mapeamento do uso e ocupação das terras, dando ênfase na identificação de áreas de cultivo de café nas regiões de Guaxupé, Machado e Três Pontas, em Minas Gerais.

Pradhan e Buchroithner (2010) aplicaram, verificaram e compararam modelos de RNAs para análise da susceptibilidade a deslizamento de terras em três regiões da Malásia. Para suportar o estudo, construíram um banco de dados de deslizamentos de terras que armazenou dados acerca da topografia, dos solos, da geologia e de mapas de cobertura do solo.

Extraíram do banco de dados 11 fatores que influenciam a ocorrência de deslizamentos de terra e computaram pesos de cada um dos fatores. Os autores selecionaram, aleatoriamente, diferentes áreas de estudo para treinar a rede neural

e prepararam nove conjuntos de mapas de susceptibilidade a deslizamentos de terra. Os mapas de susceptibilidade para as áreas estudadas foram obtidos cruzando pesos obtidos dos dados do banco com dados de outras áreas, objetivando verificar a validade do método. Ressaltaram os autores, então, que a acurácia alcançada foi maior que 83% no melhor caso.

Um modelo de rede neural artificial para a simulação de inundações utilizando SIG foi proposto por Kia et al. (2011). O objetivo estabelecido para o trabalho consistiu no desenvolvimento de um modelo de inundação que considerasse diversos fatores que contribuem para inundações, usando técnicas de RNAs e SIG.

Segundo os autores, dados temáticos de chuva, declividade, altitude, acumulação de fluxo, solo, uso da terra e geologia foram gerados usando SIG, sensoriamento remoto e levantamentos de campo. Relativamente à atribuição de pesos, a rede neural foi usada para produzir cotas de inundação e o mapa de inundação foi construído no ambiente do SIG. Para mensurar a performance do modelo, os autores utilizaram quatro critérios, a saber: o coeficiente de determinação; a soma do erro quadrático; o erro quadrático médio; e, a raiz do erro quadrático médio. Concluíram os autores que os resultados demonstraram satisfatoriamente a concordância entre os registros hidrológicos reais e os previstos pelo modelo.

Lin (2011) elaborou um modelo de rede neural e agrupamento geográfico para avaliação de risco do fluxo de detritos em rios e córregos. O estudo teve como objetivo desenvolver um modelo preciso de avaliação de riscos de fluxos de detritos, baseado em rede neural. O autor optou pelo uso de rede do tipo *backpropagation*, devido sua característica de treinamento supervisionado e sua habilidade para resolver complexos problemas de busca de padrões. Casos reais de fluxos de detritos que ocorreram em Taiwan, nos anos de 2007 e 2008, foram usados pelo autor como base de dados, além de considerar, também, dados hidrológicos e geológicos.

Foram selecionados os seguintes fatores influentes como variáveis de entrada para o modelo: gradiente médio; área de captação; área de captação eficaz; precipitação acumulada; intensidade da chuva; e, condições geológicas. Os resultados mostraram que o modelo estabelecido foi bastante adequado para a

avaliação dos riscos de fluxos de detritos, sendo que o erro obtido ficou em 7,04% para os sites geograficamente próximos que foram tratados como grupos.

Mendes e Marengo (2009) compararam RNAs com técnicas de autocorrelação, aplicadas à questão da diminuição da escala temporal (*temporal downscaling*) em cenários sobre o clima na Bacia Amazônica. Os autores verificaram que o modelo de rede neural foi superior e que os resultados comparativos indicaram que as redes neurais são ferramentas potencialmente competitivas para as análises de séries temporais multivariadas.

Silveira (2010) delimitou unidades preliminares de mapeamento de solo a partir de atributos topográficos primários e secundários, integrados com operações de tabulação cruzada e classificação por RNAs. O autor aplicou dois métodos para predição de unidades preliminares de mapeamento de solos: tabulação cruzada e integração por RNAs. Argumentou, ainda, que em detrimento dos bons resultados oferecidos pelos dois métodos, a classificação por meio das RNAs se mostrou mais eficiente na delimitação de unidades preliminares de mapeamento de solos, apresentando inúmeras vantagens em relação à tabulação cruzada.

Complementarmente, Kapageridis (2002) relatou uma série de aplicações de RNA em estudos ambientais e de mineração. Krasnopolskya e Schillerb (2003) discutiram aplicações de redes neurais em medições remotas de geofísica. O uso das RNAs em estudos relacionados aos solos pode ser encontrado em Sirtoli (2008), Arruda, Demattê e Chagas (2013) e Sirtoli et al. (2013).

3.2.5 Análise de componentes principais

Dentre as etapas da mineração de dados, pode-se destacar a seleção de atributos. Para Vercellis (2009), o propósito da seleção de atributos é eliminar do conjunto de dados aqueles que não são considerados relevantes para os objetivos das atividades de mineração de dados. Um dos aspectos mais críticos no processo de aprendizagem é a escolha da combinação de variáveis preditivas mais adequadas para explicar o fenômeno investigado (VERCELLIS, 2009).

A seleção de atributos pode melhorar os modelos resultantes e, também, contribuir para a precisão e capacidade de generalização de tais modelos (VERCELLIS, 2009; KIM; STREET; MENCZER, 2003).

Neste sentido, a Análise de Componentes Principais – ACP é uma técnica utilizada em mineração de dados para a seleção de atributos (VICINI, 2005; VERCELLIS, 2009). Vicini (2005) afirmou que a ACP permite identificar medidas responsáveis pelas maiores variações entre os resultados, sem perdas significativas. Basicamente, a ACP contribui para reduzir o conjunto de dados que será analisado, sobretudo quando os dados são constituídos de variáveis inter-relacionadas.

A ACP pode ser de grande utilidade "quando houver muitas variáveis interagindo concomitantemente no fenômeno ou no processo estudado, e que não se pode postular, com base nos dados disponíveis, uma estrutura particular destas variáveis" (ANDRIOTTI, 1997, p. 30).

Sendo assim, a ACP é uma técnica de análise multidimensional linear cujo objetivo é classificar os elementos de um conjunto em classes de elementos próximos ou similares e de estabelecer o balanço de correlações entre as variáveis originais utilizadas no estudo (ANDRIOTTI, 1997). O primeiro componente principal será a combinação linear com a maior variância; o segundo componente principal será a combinação linear com a maior variância na direção ortogonal do primeiro componente, e assim por diante (RENCHER, 2002).

As fórmulas utilizadas no algoritmo para ACP, nos termos discutidos por Varella (2008), são apresentadas a seguir. Denomina-se matriz *X*, a matriz de dados de ordem *n* x *p*, onde p são características observadas de n indivíduos de uma população π. As características observadas são representadas pelas variáveis x_1, x_2, x_3, ..., x_p.

$$X = \begin{vmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{np} \end{vmatrix} \quad (10)$$

Tomando a matriz de dados X, pretende-se transformar as variáveis x_1, x_2, x_3, ..., x_p em variáveis y_1, y_2, y_3, ..., y_p não correlacionadas e com variâncias ordenadas, para que seja possível comparar os indivíduos usando apenas as variáveis Y_{is} que apresentam maior variância. Para tal, a solução é dada a partir da matriz de covariância S ou da matriz de correlação R.

A partir da matriz *X*, estima-se a matriz de covariância ∑ da população π representada por *S*.

$$S = \begin{vmatrix} \hat{V}ar(x_1) & \hat{C}ov(x_1x_2) & \hat{C}ov(x_1x_3) & \cdots & \hat{C}ov(x_1x_p) \\ \hat{C}ov(x_2x_1) & \hat{V}ar\,(x_2) & \hat{C}ov(x_2x_3) & \cdots & \hat{C}ov(x_2x_p) \\ \hat{C}ov(x_3x_1) & \hat{C}ov(x_3x_2) & \hat{V}ar\,(x_3) & \cdots & \hat{C}ov(x_3x_p) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{C}ov(x_px_1) & \hat{C}ovv(x_px_2) & \hat{C}ov(x_px_3) & \cdots & \hat{V}ar\,(x_p) \end{vmatrix} \quad (11)$$

Denomina-se matriz *Z* a matriz de correlação da matriz de dados X, onde as variáveis z_1, z_2, z_3, ..., z_p correspondem à padronização das variáveis x_1, x_2, x_3, ..., x_p.

$$Z = \begin{vmatrix} z_{11} & z_{12} & z_{13} & \cdots & z_{1p} \\ z_{21} & z_{22} & z_{23} & \cdots & z_{2p} \\ z_{31} & z_{32} & {}_{33} & \cdots & z_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & z_{n3} & \cdots & z_{np} \end{vmatrix} \quad (12)$$

Os componentes principais são normalmente obtidos a partir da matriz de correlação *R*, resolvendo-se a equação característica da matriz, ou seja,

$$det[R - \lambda I] = 0 \qquad \text{ou} \qquad |R - \lambda I| = 0 \quad (13)$$

$$R = \begin{vmatrix} 1 & r(x_1x_2) & r(x_1x_3) & \cdots & r(x_1x_p) \\ r(x_2x_1) & 1 & r(x_2x_3) & \cdots & r(x_2x_p) \\ r(x_3x_1) & r(x_3x_2) & 1 & \cdots & r(x_3x_p) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r(x_px_1) & r(x_px_2) & r(x_px_3) & \cdots & 1 \end{vmatrix} \quad (14)$$

Sejam λ_1, λ_2, λ_3, ..., λ_p as raízes da equação característica da matriz R ou S, então:

$$\lambda_1 > \lambda_2 > \lambda_3 > \cdots, \lambda_p. \quad (15)$$

Para cada autovalor λi existe um autovetor $\tilde{a}_i$:

$$\tilde{a}_i = \begin{bmatrix} a_{i1} \\ a_{i2} \\ \vdots \\ a_{ip} \end{bmatrix} \quad (16)$$

Os autovetores $\tilde{a}_i$ são normalizados, isto é, a soma dos quadrados dos coeficientes é igual a 1 e, ainda, são ortogonais entre si, possuindo as seguintes propriedades:

$$\sum_{j=1}^{p} a_{ij}^2 = 1 \quad (\tilde{a}_i' . \tilde{a}_i = 1) \quad (17)$$

E ainda,

$$\sum_{j=1}^{p} a_{ij} . \ a_{kj} = 1 \quad (\tilde{a}_i' . \tilde{a}_k = 0 \ para \ i \neq k) \quad (18)$$

Sendo $\tilde{a}_i$ o autovetor correspondente ao autovalor λi, então o i-ésimo componente principal é dado por:

$$Y_i = \ a_{i1}X_1 + \ a_{i2}X_2 + \cdots + \ a_{ip}X_p \quad (19)$$

4 MATERIAIS E MÉTODOS

Neste capítulo serão descritos os materiais, os métodos e os procedimentos adotados durante a pesquisa. Na seção 4.1 são listados os materiais utilizados, incluindo os dados, os softwares e o hardware. Na seção 4.2 são detalhados os métodos e os procedimentos adotados para a construção do banco de dados; a extração de parâmetros morfométricos e índices; a construção da RNA; e, as medidas que foram usadas para avaliar o desempenho da RNA.

4.1 MATERIAIS UTILIZADOS

Os dados espaciais utilizados na pesquisa foram obtidos junto a diversas instituições governamentais, tais como: o Sistema de Proteção da Amazônia – SIPAM, o Serviço Geológico do Brasil – CPRM e a Secretaria de Estado do Desenvolvimento Ambiental – SEDAM-RO.

Foi realizada, assim, uma seleção dos dados de modo a definir apenas aqueles que abrangiam a área geográfica relativa à bacia do estudo. Inicialmente, os dados vetoriais selecionados foram reprojetados para o sistema SIRGAS 2000, UTM Zona 20S. Os dados vetoriais considerados neste estudo são apresentados na TABELA 1 a seguir:

TABELA 1 – DADOS VETORIAIS QUE ABRANGEM A ÁREA DA BHRMP UTILIZADOS NA PESQUISA.

TEMA	ESCALA	ANO	FONTE
Geologia	1:250.000	2005	CPRM
Geomorfologia	1:250.000	2005	CPRM
Hidrogeologia	1:250.000	2005	CPRM
Hidrografia	1:100.000	2004	SEDAM/RO
Solo	1:250.000	2002	SEDAM/RO
Vegetação	1:250.000	2002	SEDAM/RO

FONTE: O autor (2016).

Os mosaicos de imagens para a área de Mutum-Paraná foram gerados, reprojetados para o sistema SIRGAS 2000, UTM Zona 20S, e convertidos para o formato GeoTiff a partir dos seguintes conjunto de dados:

- Imagens de radar SAR/SIPAM: segmentos 5 a 9 que se referem a áreas das localidades de Nova Mamoré e Porto Velho no Estado de Rondônia, obtidas durante a Missão MMA/Rondônia, executada em abril de 2008; modo de imageamento Quad L+X; banda L; polarizações VV, HH, HV e VH; 8 bits; resolução espacial de 6 metros; *unsigned integer*; sensor R99 B; direção de visada E/W e W/E; altura do voo 35.000 pés; sistema de coordenadas geográficas (lat/long); *Datum* WGS 84;
- Imagens de satélite SPOT: missão 5; obtidas durante os meses de julho e agosto de 2008; multiespectral (quatro bandas) e pancromática; resolução espacial de 10 metros (multiespectral) e 2,5 metros (pancromática); sistema de coordenadas geográficas (lat/long); *Datum* WGS 84; e,
- MDE SRTM: versão 3; 1 arco por segundo (30 metros), obtido junto ao site da NASA.

Como as imagens SPOT foram obtidas em diferentes datas, foi procedida uma calibração ótica visando reduzir a variação nas medidas de radiância, que são influenciadas, por exemplo, pelas condições atmosféricas e pela posição do sensor (FLOOD et al., 2013; FISHER; DANAHER, 2013). Após este processamento, pixels da imagem passam a armazenar não mais um número digital relativo a tons de cinza, mas sim um valor de reflectância adequado para uso em comparações e cálculo de índices.

O software SIG escolhido para uso no projeto foi o QGIS, versão 2.10 Pisa, que, por meio das opções de processamento, incorpora funcionalidades do SAGA versão 2.1, Grass versão 6.4.3 e Orfeo Toolbook. O QGIS também dispõe das funcionalidades providas pela biblioteca GDAL. Adicionalmente, funcionalidades do pacote TauDEM (*Terrain Analysis Using Digital Elevation Models*), que consiste num conjunto de ferramentas para a extração e análise de informações hidrológicas e topográficas a partir de MDEs, e que implementa diversos algoritmos para modelos hidrológicos, foram configuradas para funcionar no ambiente do QGIS.

Foi utilizado o software WEKA – *Waikato Environment for Knowledge Analysis*, versão 3.6 como ferramenta de mineração de dados. WEKA é um software desenvolvido e mantido pela Universidade de Waikato, que consiste numa coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados.

Para gerenciar o banco de dados do estudo foi usado o SGBD PostgreSQL, versão 9.4.4. PostgreSQL é um sistema para banco de dados que implementa conceitos avançados de administração de bases de dados, robusto e amplamente utilizado.

Visando permitir o armazenamento de dados espaciais, foi necessário usar, também, a extensão PostGIS, versão 2.2. PostGIS é uma extensão que habilita o PostgreSQL manipular dados espaciais em conformidade com padrões estabelecidos pelo Open Geospatial Consortium – OGC, instituição que estabelece os padrões de interoperabilidade para dados geoespaciais.

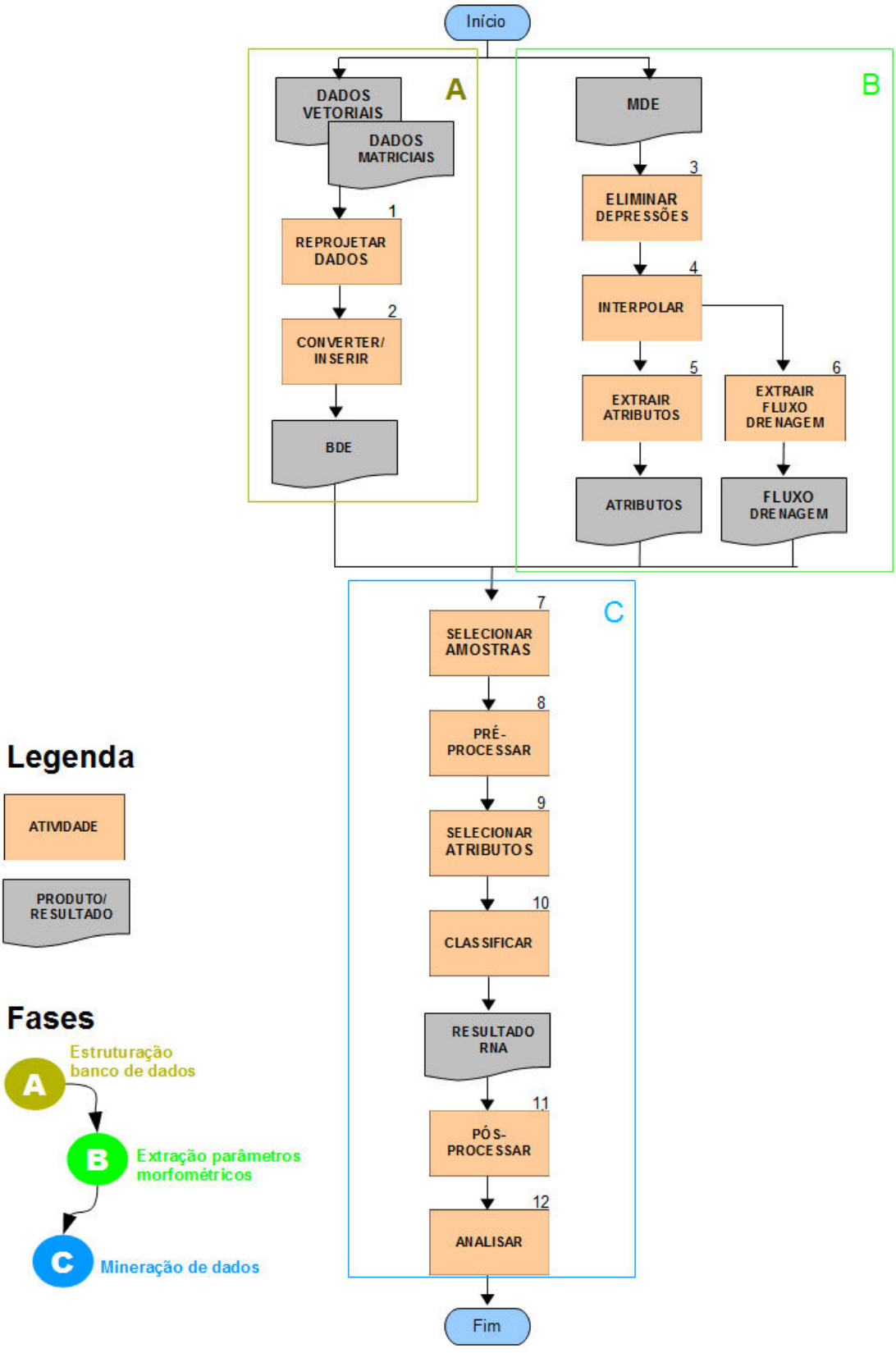
O banco de dados criado no SGBD PostgreSQL/PostGIS foi estruturado em conformidade com a especificação *Simple Feature Access – SFA*, que também pode ser denominada de padrão ISO 19125. Este documento especifica a maneira para armazenar e acessar dados espaciais por meio da linguagem SQL (OGC, 2010). Além disso, a versão 2.2 do PostGIS suporta o padrão *ISO/IEC 13249-3 SQL/MM Spatial*, que estende o padrão SFA e adiciona características como transformação de coordenadas e métodos para validação de geometrias (STOLZE, 2003; OGC, 2010).

Quanto aos recursos computacionais, foi utilizada uma estação de trabalho EliteDesk HP, configurada com 16 GB de memória RAM e processador Intel Core I7. O sistema operacional adotado foi o Ubuntu, versão 14.04 64 bit.

4.2 MÉTODOS E PROCEDIMENTOS

Na FIGURA 11 é apresentado um fluxograma simplificado com as principais etapas desenvolvidas, que serão abordadas nas seções seguintes. A TABELA 2 contém a descrição das atividades do fluxograma.

FIGURA 11 – FLUXOGRAMA SIMPLIFICADO DAS PRINCIPAIS ETAPAS DA PESQUISA.



FONTE: O autor (2016).

TABELA 2 – DESCRIÇÃO DAS ATIVIDADES DO FLUXOGRAMA SIMPLIFICADO DA PESQUISA.

N°	ATIVIDADE	DESCRIÇÃO
1	Reprojetar dados	Reprojetar os dados vetoriais e matriciais para o sistema SIRGAS 2000, UTM Zona 20S
2	Importar/Inserir	Importar dados vetoriais e dados matriciais para formato de tabela do PostgreSQL/PostGIS
3	Eliminar depressões	Eliminar depressões (espúrias) do MDE
4	Interpolar	Interpolar o MDE para obter resoluções semelhantes àquelas das imagens SPOT e SAR
5	Extrair atributos	Extrair os parâmetros morfométricos para a área de estudo
6	Extrair fluxo drenagem	Extrair a direção de fluxo da drenagem da área de estudo
7	Selecionar amostrar	Selecionar amostrar para treinamento e teste da RNA
8	Pré-processar	Converter os dados para o formato ARFF do WEKA
9	Selecionar atributos	Selecionar atributos usando a técnica de Análise de Componentes Principais
10	Classificar	Classificar os dados da área, usando RNA, até obtenção de resultado desejado (processo interativo e iterativo).
11	Pós-processar	Converter arquivo ARFF para o formato vetorial, eliminar linhas de drenagem desconectas
12	Analisar	Analisar os resultados obtidos

FONTE: O autor (2016).

4.2.1 Estruturação do banco de dados

Os dados vetoriais selecionados para a pesquisa foram convertidos e inseridos no banco de dados por intermédio do aplicativo *shp2pgsql*. Este aplicativo converte um arquivo vetorial no formato ESRI Shapefile para comandos em linguagem de consulta *Structured Querie Language* – SQL, que permite o carregamento dos dados no PostgreSQL/PostGIS.

De forma semelhante, as imagens utilizadas neste estudo foram convertidas e inseridas no banco de dados com o auxílio do aplicativo r*aster2pgsql*. A exemplo do *shp2pgsql*, *raster2pgsql* converte arquivo no formato de imagem para comandos SQL que podem ser inseridos no PostgreSQL/PostGIS. Para a modelagem conceitual dos dados foi adotado o modelo OMT-G (BORGES, 1997).

4.2.2 Extração da rede de drenagem, parâmetros morfométricos e NDWI

Esta etapa consiste no pré-processamento do MDE, cuja finalidade é a eliminação de eventuais depressões espúrias no modelo SRTM. É comum a

ocorrência de “buracos” em um MDE, ou seja, a existência de uma célula ou um grupo de células com uma elevação mais baixa do que todas as células adjacentes (CIMMERY, 2010). Corrigir as depressões do MDE é necessário para permitir a correta execução dos algoritmos de direção de fluxo (OLAYA; CONRAD, 2009; CIMMERY, 2010). Para tanto, foi usado o algoritmo proposto por Wang e Liu (2006) que remove as depressões identificadas, ao passo que preserva as inclinações ao longo do trajeto de escoamento.

Foi necessário alterar a resolução espacial do MDE, visto que para o uso nas etapas de mineração de dados deveria ter a mesma resolução das imagens, ou seja, 6 metros para imagens SAR e 2,5 metros para imagens SPOT. As reamostragens foram feitas por meio de interpolações utilizando o método estatístico do vizinho mais próximo (*Nearest-neighbord interpolation*). A opção por este método de interpolação se deu por sua característica de que o valor interpolado seja um dos valores originais, quer dizer, a interpolação não produziu novos valores, afetando somente a resolução da imagem de entrada (NOVO, 1989).

Os parâmetros morfométricos utilizados nesta pesquisa foram calculados de acordo com modelo matemático baseado no método amplamente difundido de obter primeiras derivativas e derivativas parciais de um polinômio bi-quadrático de segundo grau que representa uma superfície. Este polinômio pode ser representado pela equação (EVANS, 1980; WOOD, 1996):

$$z = ax^2 + by^2 + cxy + dx + ey + f \qquad (20)$$

onde *z* é a estimativa de elevação em um ponto (x,y), e a até *f* representam os coeficientes que definem a superfície quadrática.

Os coeficientes da expressão polinomial são estimados tomando como base uma janela de 3x3 células, conforme FIGURA 12 a seguir:

FIGURA 12 – CODIFICAÇÃO DAS CÉLULAS PARA CÁLCULO DOS COEFICIENTES POLINOMIAIS.

Z_1	Z_2	Z_3
Z_4	Z_5	Z_6
Z_7	Z_8	Z_9

FONTE: Wood (1996).

O polinômio pode, ainda, ser escrito na forma geral cônica, a saber (WOOD, 1996; EVANS, 1980):

$$ax^2 + 2hxy + by^2 + 2jx + 2ky + m = 0 \quad (21)$$

onde *h=c/2*, *j=d/2*, *k=e/2*, e *m=f-z*.

De acordo com Wood (1996), a taxa de variação de elevação em ambas as direções x e y pode ser utilizada para identificar a direção e a magnitude da inclinação mais acentuada. Estes dois parâmetros podem ser encontrados tomando as derivadas parciais de primeira ordem em relação a x e y. A declividade pode ser encontrada através da combinação dos dois componentes parciais derivativos (WOOD, 1996):

$$\frac{dz}{dxy} = \sqrt{\left(\frac{\delta z}{\delta x}\right)^2 + \left(\frac{\delta z}{\delta y}\right)^2} \quad (22)$$

As derivativas parciais para x e y são dadas como (WOOD, 1996):

$$\frac{\delta z}{\delta x} = 2ax + cy + d \quad (23)$$

$$\frac{\delta z}{\delta y} = 2by + cx + e \quad (24)$$

Para encontrar a declividade no ponto central da superfície quadrática, mediante adoção de um sistema de coordenadas local com a origem localizada no ponto de interesse, onde x = y = 0, dá-se a partir da seguinte fórmula (WOOD, 1996):

$$\frac{\mathrm{dz}}{\mathrm{dxy}} = \sqrt{d^2 + e^2} \tag{25}$$

A declividade geralmente é expressada em graus, a saber (EVANS, 1980; ZEVENBERGEN; THORNE, 1987; WOOD, 1996):

$$declividade = \text{?rct?n}\left(\sqrt{d^2 + e^2}\right) \tag{26}$$

De forma semelhante, para se obter o aspecto, considera-se o ângulo polar descrito pelas duas derivativas parciais ortogonais (WOOD, 1996):

$$aspecto = \text{?rct?n}\left(\frac{e}{d}\right) \tag{27}$$

Curvatura pode ser separada em dois componentes ortogonais em que os efeitos do processo gravitacional ou são maximizados (perfil de curvatura) ou minimizados (plano de curvatura) (WOOD, 1996). Para os cálculos de curvatura, foram consideradas as fórmulas seguintes (EVANS, 1980; WOOD, 1996):

$$perfil\ de\ curvatura = \ profc = \frac{-200\left(ad^2 + be^2 + cde\right)}{\left(e^2 + d^2\right)\left(1 + d^2 + d^2\right)^{1.5}} \tag{28}$$

$$profc_{MAX} = \text{-?-b} + \sqrt{(a-b)^2 + \ c^2} \tag{29}$$

$$profc_{MIN} = \text{-?-b} - \sqrt{(a-b)^2 + \ c^2} \tag{30}$$

$$plano\ de\ curvatura = \ planc = \frac{200\left(bd^2 + ae^2 - cde\right)}{\left(e^2 + d^2\right)^{1.5}} \tag{31}$$

$$curvatura\ longitudinal = longc = \text{-2}\left(\frac{ad^2 + be^2 + cde}{d^2 + e^2}\right) \tag{32}$$

$$curvatura\ transversal = crosc = \text{-2}\left(\frac{bd^2 + ae^2 - cde}{d^2 + e^2}\right) \tag{33}$$

Para o cálculo do índice de umidade, seguiu-se a aplicação da seguinte fórmula (MOORE; GRAYSON; LADSON, 1991; QUINN et al., 1991; GRUBER; PECKHAM, 2009):

$$TWI = \ln\left[\frac{A}{\tan(\beta)}\right] \quad (34)$$

onde *A* é a área de captação específica e β é o ângulo de declividade local.

O índice de corrente de máximo fluxo também foi calculado a partir da área de captação específica *A* e do ângulo de declividade local β, da seguinte forma (MOORE; GRAYSON; LADSON, 1991; GRUBER; PECKHAM, 2009):

$$TWI = A * \text{t}\square\text{n}(\beta) \quad (35)$$

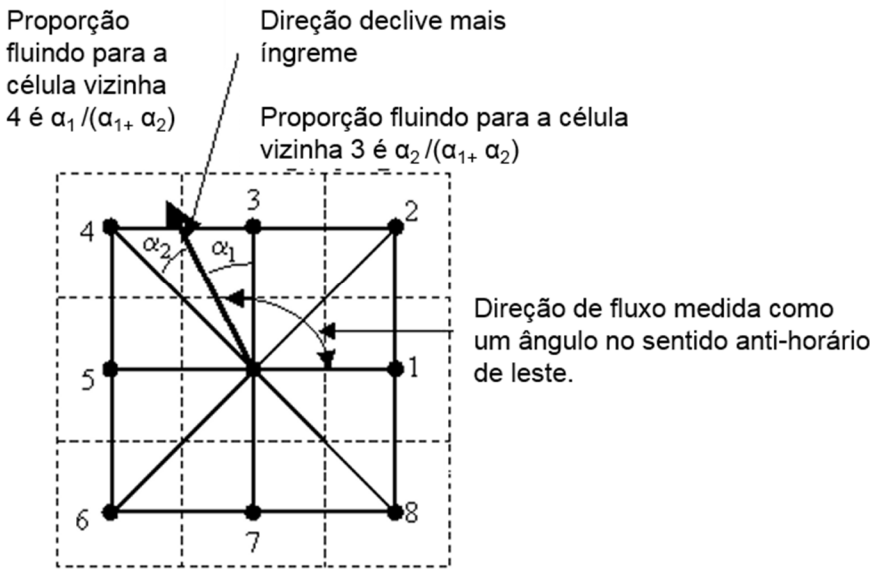
O índice de rugosidade do terreno foi obtido mediante a aplicação da seguinte fórmula (RILEY; DE GLORIA; ELLIOT, 1999):

$$TRI = Y\ \left[\sum\left(x_{ij} - x_{00}\right)^2\right]^{\frac{1}{2}} \quad (36)$$

onde x_{ij} é a elevação de cada célula vizinha à célula (0,0).

O algoritmo para determinação da direção de fluxo escolhido para o estudo foi o *Deterministic infinity* – D∞, desenvolvido por Tarboton (1997). Neste método, a direção do fluxo é definida como o declive mais íngreme em facetas triangulares em cada ponto da grade (janela 3x3). A direção do fluxo é codificada como um ângulo em radianos no sentido anti-horário de leste com valor contínuo entre 0 e 2 pi, conforme demonstrado na FIGURA 13 a seguir. O fluxo resultante é proporcionado entre as duas células vizinhas que definem a faceta triangular com o declive mais acentuado.

FIGURA 13 – DETERMINAÇÃO DA DIREÇÃO DO FLUXO COM O ALGORITMO D∞.



FONTE: Tarboton (1997).

Os cálculos foram realizados por intermédio dos módulos do SAGA, disponíveis no menu Processamento do QGIS. Os principais módulos utilizados, em sequência, são apresentados na TABELA 3 abaixo. Foram adotados valores (*default*) padrão em todos os módulos.

TABELA 3 – PRINCIPAIS MÓDULOS DO SAGA UTILIZADOS PARA EXTRAÇÃO DOS ATRIBUTOS.

SEQ	MÓDULO	ENTRADA	SAÍDA
1	Fill Sinks (Wang e Liu)	MDE original	MDE hidrologicamente consistido
2	Resampling	MDE resolução original	MDE reamostrado
3	Slope, Aspect, Curvature	MDE	Declividade, aspecto, curvatura (geral, perfil, plano, longitudinal, transversal, máxima, mínima)
4	Channel Network and Drainage Basins	MDE	Direção de fluxo, conectividade de fluxo, Ordem de Straher, bacias de drenagem, canais de drenagem
5	Catchment Area	MDE	Área de captação
6	Topographic Wetness Index	Declividade, área de captação	Índice de umidade
7	Terrrain Ruggedness Index	MDE	Índice de rugosidade do terreno
8	Stream Power Index	Declividade, área de captação	Índice de corrente de máximo fluxo

FONTE: O autor (2016).

A partir das imagens SPOT foi calculado o índice de água por diferença normalizada (*Normalized Difference Water Index* – NDWI), conforme definido por McFeeters (1996), cuja fórmula é:

$$NDWI = \frac{\gamma_{green} - \gamma_{NIR}}{\gamma_{green} + \gamma_{NIR}} \quad (37)$$

onde γ é a reflectância no comprimento de ondas verde (*green*) ou no infravermelho próximo (NIR).

4.2.3 Seleção de amostras

O reconhecimento da área de estudo relativa à BHRMP foi feito por meio de visitas *in-loco* nos meses de agosto e setembro de 2013, quando foram levantados dados sobre a existência ou não do fluxo de água ou canal em campo. O registro dos pontos dos locais visitados em campo foi feito com uso de equipamento GPS, com acurácia melhor que 12 metros em 90% dos pontos, equivalente ao PEC-PCD classe B, na escala de 1:50.000.

A representatividade amostral para a coleta de dados foi definida nos termos especificados por Andriotti (2003), nos moldes da fórmula:

$$n = \frac{1}{\mathrm{E}^2} \quad (38)$$

onde *E* representa o erro amostral tolerável.

O conhecimento do número total de nascentes (cartografadas) permitiu corrigir a fórmula em função do tamanho (*n*), a saber:

$$n = \frac{N \,.\, n_o}{N + n_o} \quad (39)$$

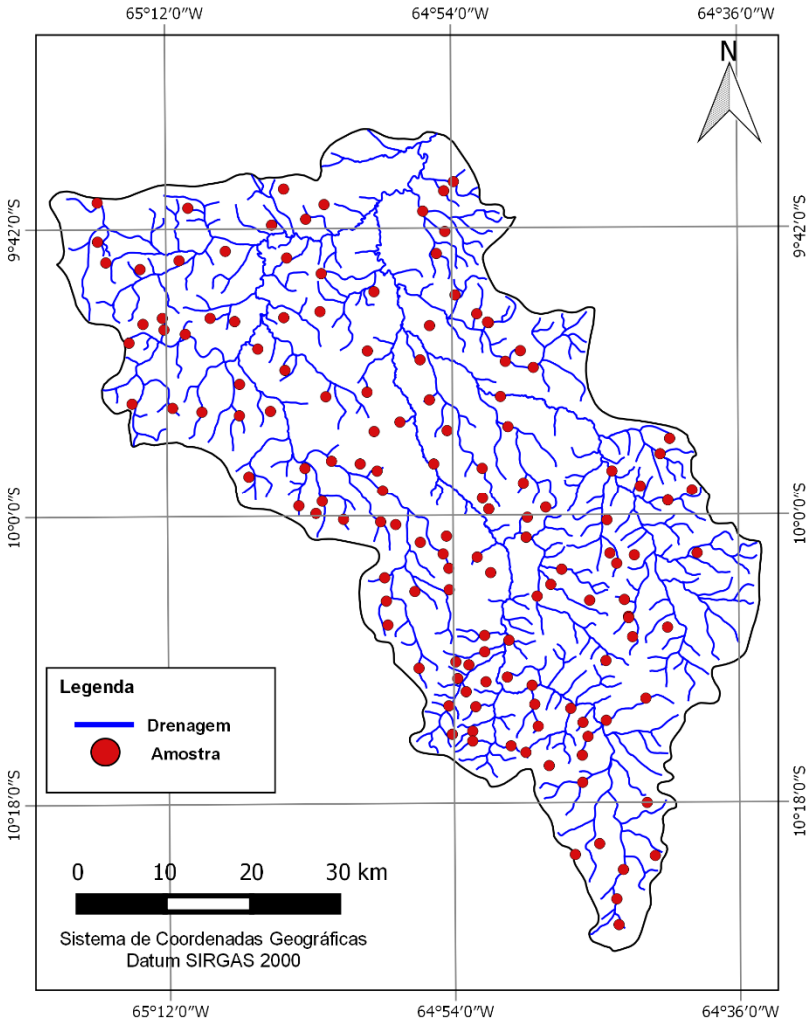
onde n_0 é o tamanho da amostra.

Assim, considerando que o total de nascentes cartografadas é igual a 321 (N=321) e o tamanho da amostra é igual a 236 (n_0=236), resultou em um *n* amostral mínimo de 142 bacias de drenagem, para um erro amostral tolerável de menos de 6,5%:

$$n = \frac{(321*236)}{(321+236)} = 142 \quad (40)$$

Na FIGURA 14, a seguir, são apresentados os pontos amostrais verificados em campo e que foram usados para validação dos resultados da RNA.

FIGURA 14 – PONTOS AMOSTRAIS VERIFICADOS EM CAMPO.



FONTE: O autor (2016).

4.2.4 Atividades de mineração de dados

O software utilizado para a mineração de dados foi o WEKA, que se destaca pela variedade de algoritmos disponíveis, incluindo o algoritmo *multilayerperceptron* para RNA. Para esta pesquisa, os fatores considerados para a escolha de técnicas e algoritmos foram os seguintes:

1. Os objetivos da pesquisa;

2. Os tipos de dados contidos no banco de dados da pesquisa; e,
3. Os algoritmos implementados no WEKA.

Nas seções seguintes são apresentadas as atividades de mineração de dados executadas durante a pesquisa, bem como são descritos os principais conceitos matemáticos envolvidos. O processo de mineração de dados foi sistematizado com embasamento nas tarefas básicas propostas por Fayyad, Piatetsky-Shapiro e Smyth (1996), Goebel e Gruenwald (1999) e Rocha (2005), compreendendo as etapas de pré-processamento, seleção de atributos, classificação e pós-processamento.

4.2.4.1 Pré-processamento

Um requisito para o uso do WEKA consiste em converter os dados para o formato de arquivo *Attribute Relation Format* File – ARFF. Um arquivo ARFF é estruturado em duas partes: o cabeçalho (*header*), que contém o nome da relação e a especificação dos atributos, e a seção de dados (*data*), que contém os valores para cada atributo, na ordem em que foram declarados no cabeçalho. Cada linha da seção de dados representa uma ocorrência (instância).

A FIGURA 15, a seguir, apresenta um exemplo de arquivo no formato ARFF, com destaque para o cabeçalho, que contém a declaração dos atributos, seguido da seção de dados que contém exemplo de instâncias e seus respectivos valores.

FIGURA 15 – EXEMPLO DE ARQUIVO NO FORMATO ARFF, COM A DECLARAÇÃO DE 4 ATRIBUTOS DO TIPO NUMÉRICO E DUAS INSTÂNCIAS DE DADOS REPRESENTADAS.

```
% Cabeçalho (header)
% Title: Pontos amostrais (conjunto treinamento)
@RELATION amostra_treinamento
@ATTRIBUTE image_vpixel numeric
@ATTRIBUTE mde_vpixel numeric
@ATTRIBUTE aspect numeric
@ATTRIBUTE catchment_flow numeric
@ATTRIBUTE class {dre,ndr}

%Dados (data)
@DATA
92,178,0.1671018004,35.9994659420,14426.3583980000
60,145,1.5707963705,215.9967956500,5565995.5000000000
```

FONTE: O autor (2016).

Os arquivos ARFF com os dados da pesquisa continham os dados correspondentes aos pontos que compunham os conjuntos de treinamento e teste. Cada ponto equivaleu a um pixel do arquivo matricial, portanto, foram preparados arquivos para dados com pixel de 2,5 metros e pixel de 6 metros.

A definição do tamanho dos conjuntos de dados para treinamento e teste é um importante fator nas aplicações de RNA (FOODY, 2008; FIGUEROA et al., 2012). De forma semelhante, o mesmo pode ser dito quanto ao tamanho das amostras nas atividades de classificação de dados de Sensoriamento Remoto (LI et al., 2014; CONGALTON, 1991; HASHEMIAN; ABKAR; FATEMI, 2004).

O total de amostras dos conjuntos de dados usados neste trabalho foi definido observando valores apresentados nos trabalhos de: Congalton (1991), que sugeriu de 75 a 100 amostras por classe; Hashemian, Abkar e Fatemi (2004), que adotou de 50 a 70 amostras por classe; e de, Li et al. (2014) que propôs 200 amostras por classe.

Foram compostos quatro conjuntos para uso na fase de treinamento da rede. Outros conjuntos de exemplos independentes, que não participaram da fase de treinamento, foram criados e usados para a fase de teste; tais conjuntos são apresentados na TABELA 4.

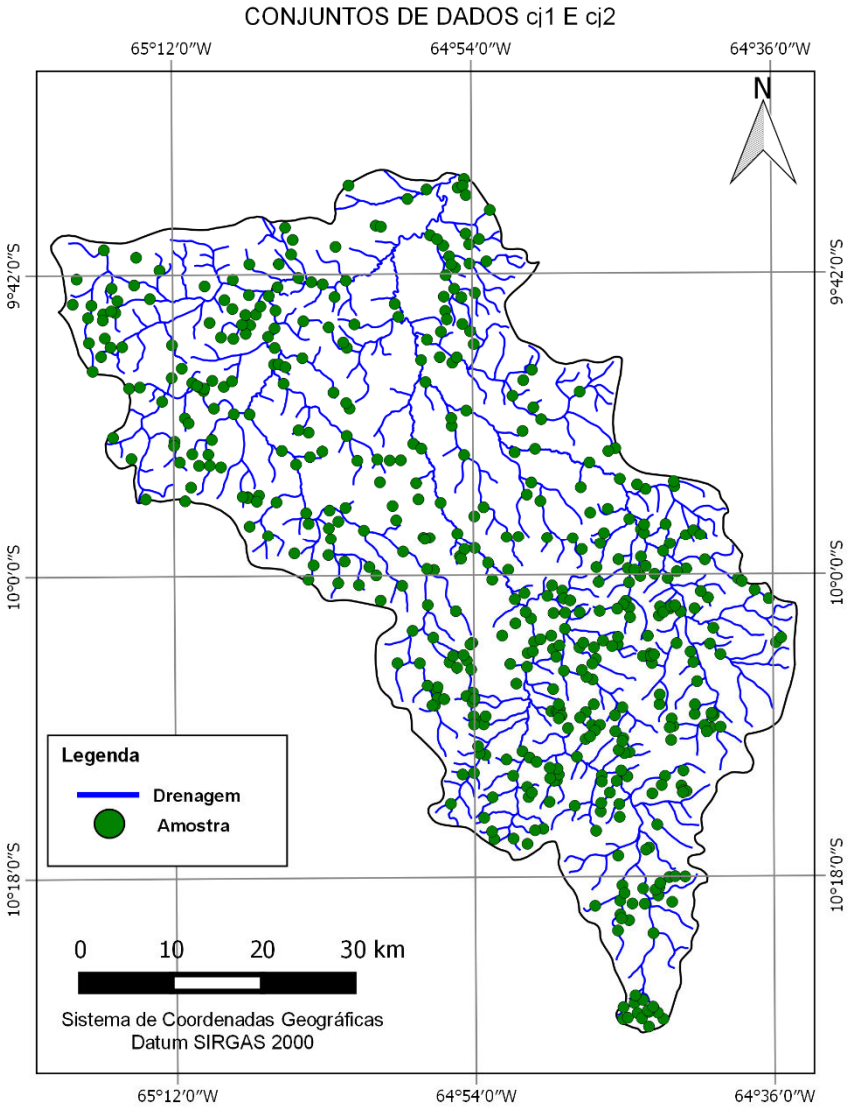
TABELA 4 – Conjuntos de dados usados na Mineração de Dados.

Conjunto	Tipo	Tamanho pixel (metros)	Quantidade amostras	Classes de saída (dre=drenagem; ndr= não drenagem; nas=nascente)
cj1	Treinamento	6	490	dre, ndr
cj2	Treinamento	2,5	490	dre, ndr
cj3	Treinamento	6	939	dre, ndr, nas
cj4	Treinamento	2,5	939	dre, ndr, nas
cj5	Teste	6	300	dre, ndr
cj6	Teste	2,5	300	dre, ndr
cj7	Teste	6	300	dre, ndr, nas
cj8	Teste	2,5	300	dre, ndr, nas
cj9	Teste	6	170	dre, ndr
cj10	Teste	2,5	170	dre, ndr
cj11	Teste	6	170	dre, ndr, nas
cj12	Teste	2,5	170	dre, ndr, nas
cj13	Teste	6	222	dre, ndr
cj14	Teste	2,5	222	dre, ndr
cj15	Teste	6	222	dre, ndr, nas
cj16	Teste	2,5	222	dre, ndr, nas
cj17	Teste	6	522	dre, ndr
cj18	Teste	2,5	522	dre, ndr
cj19	Teste	6	482	dre, ndr
cj20	Teste	2,5	482	dre, ndr

FONTE: O autor (2016).

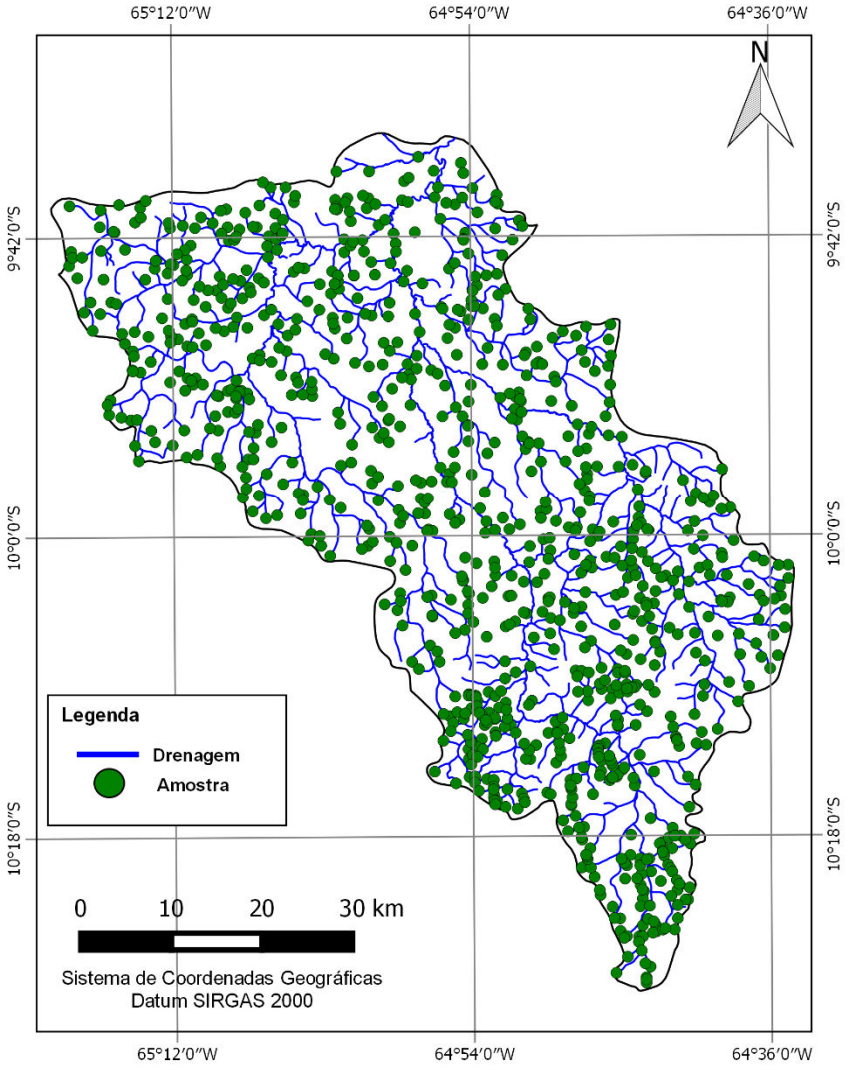
Os conjuntos de treinamento e teste constituíram-se amostras representativas da área de estudo, conforme pode ser observado nas figuras 16 a 21, que ilustram a distribuição na área geográfica da bacia dos pontos utilizados para a composição dos conjuntos.

FIGURA 16 – PONTOS UTILIZADOS NOS CONJUNTOS cj1 E cj2.



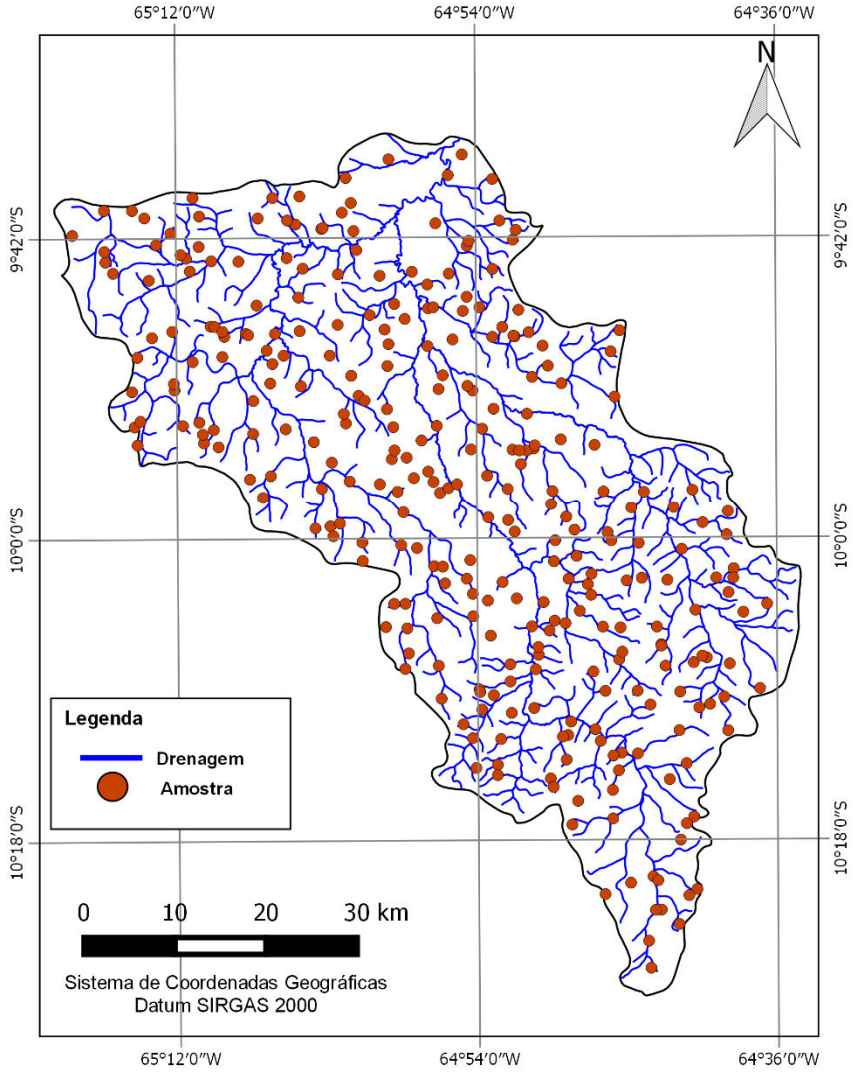
FONTE: O autor (2016).

FIGURA 17 – PONTOS UTILIZADOS NOS CONJUNTOS cj3 E cj4.



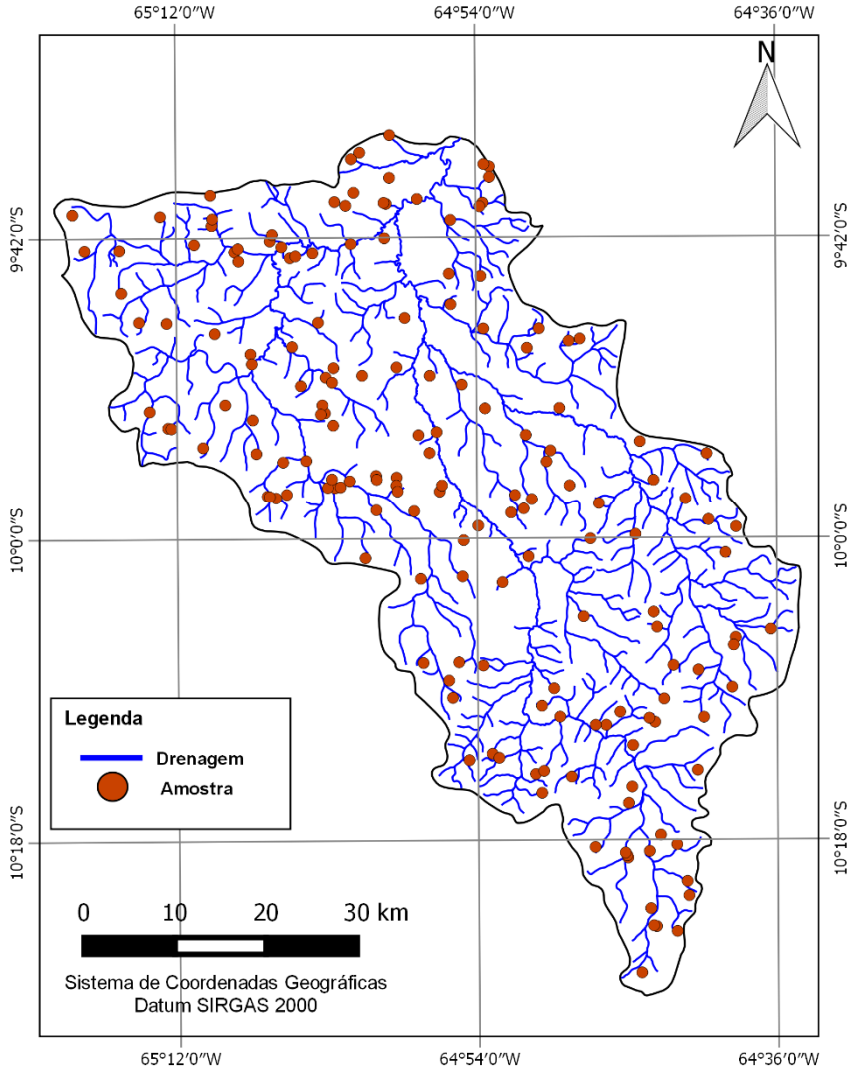
FONTE: O autor (2016).

FIGURA 18 – PONTOS UTILIZADOS NOS CONJUNTOS cj5, cj6, cj7 E cj8.



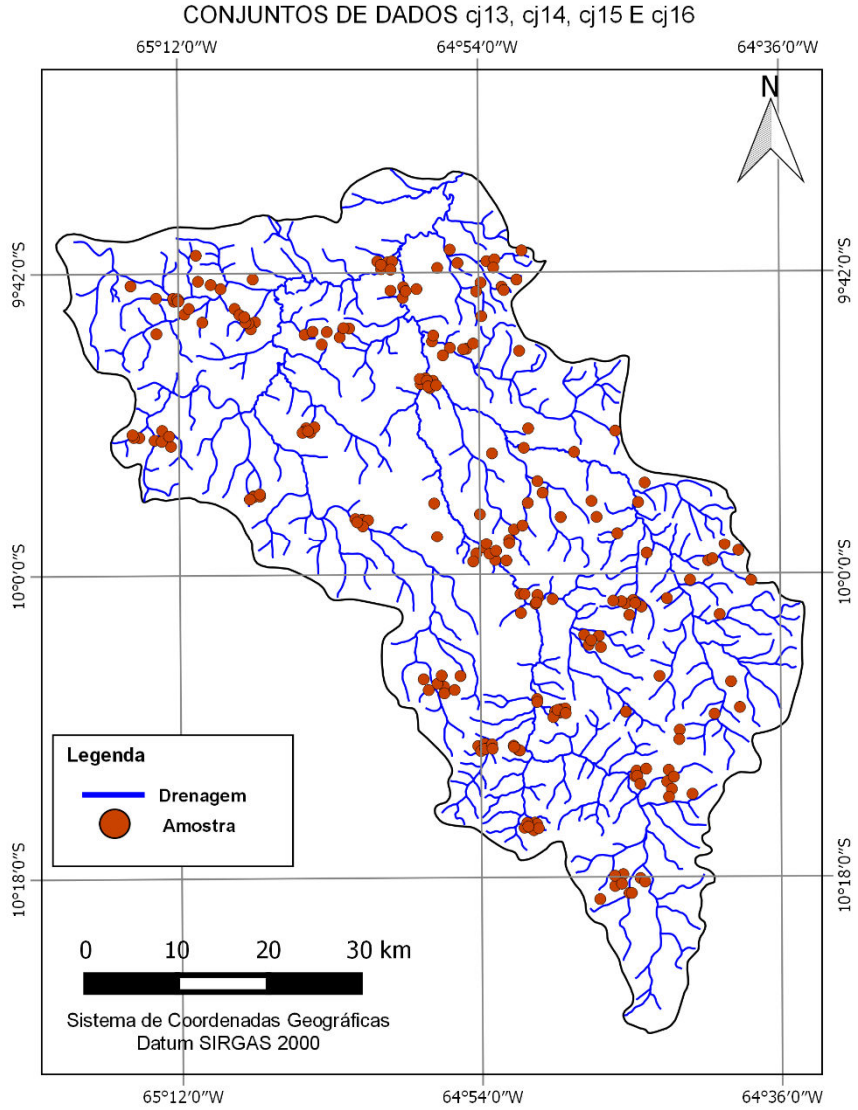
FONTE: O autor (2016).

FIGURA 19 – PONTOS UTILIZADOS NOS CONJUNTOS cj9, cj10, cj11 E cj12.



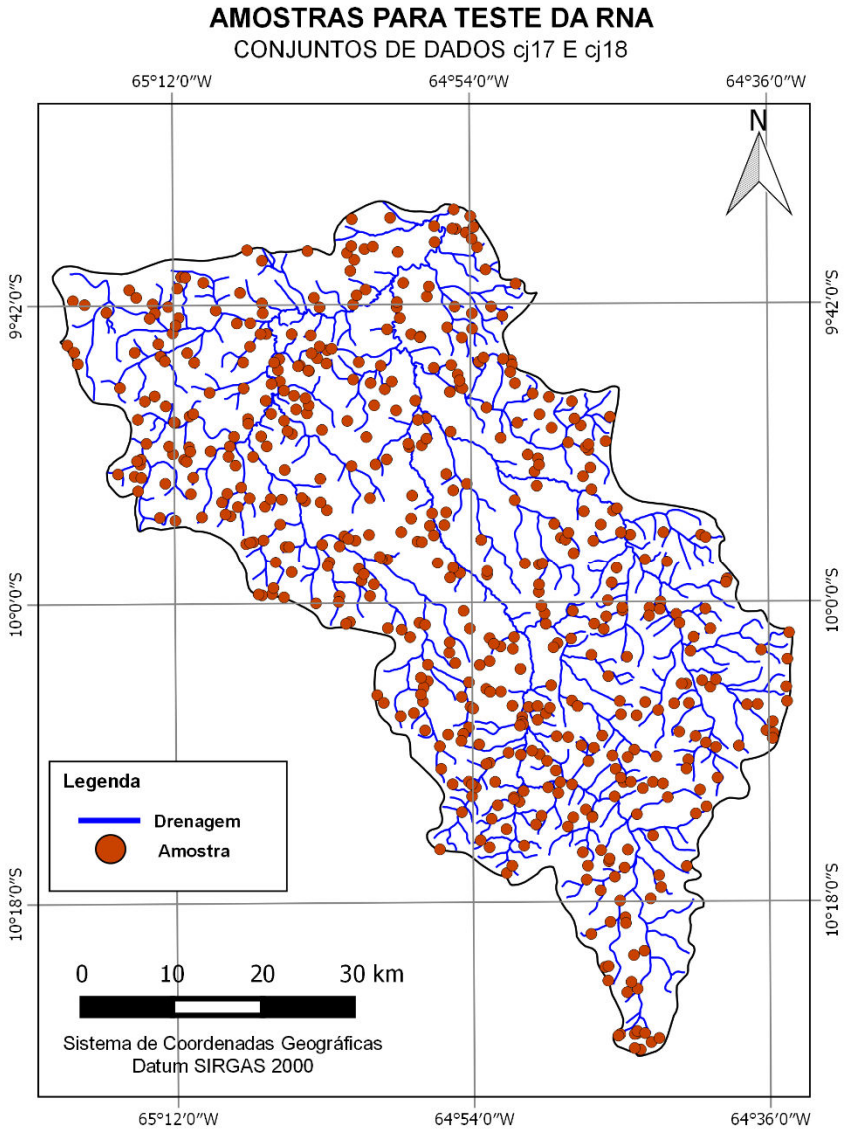
FONTE: O autor (2016).

FIGURA 20 – PONTOS UTILIZADOS NOS CONJUNTOS cj13, cj14 cj15 E cj16.



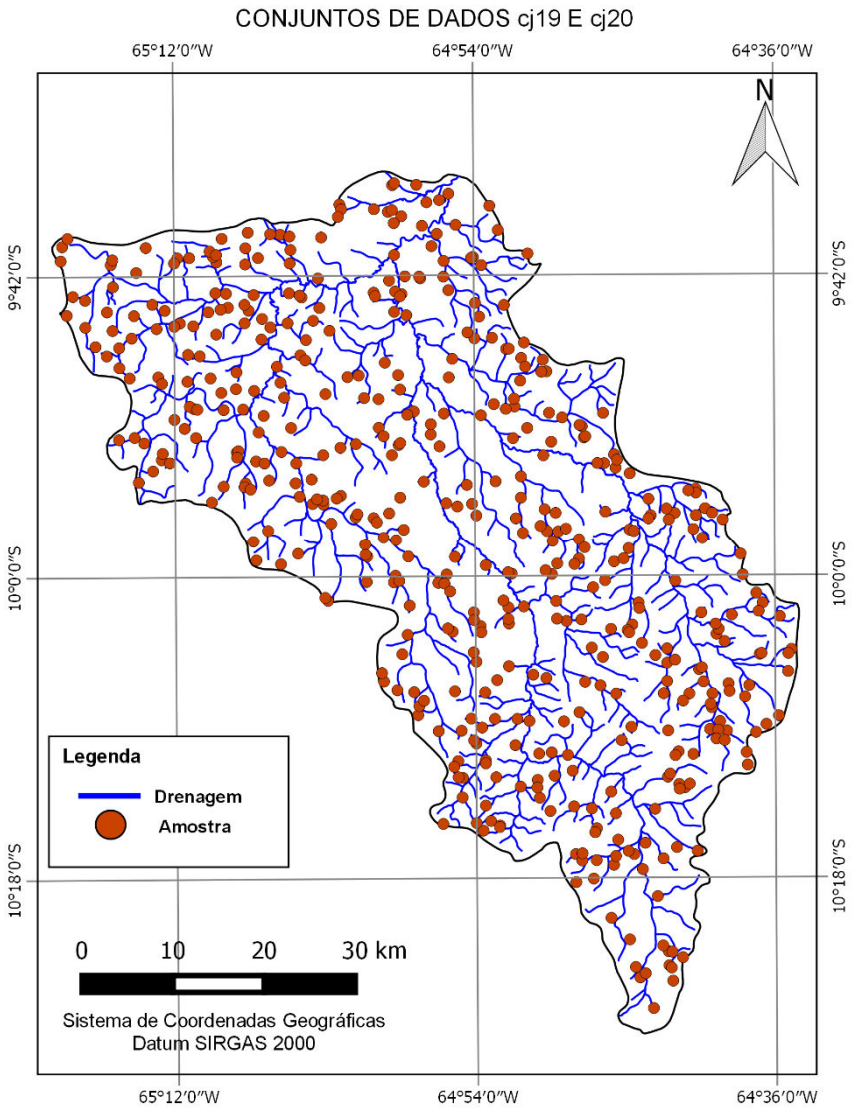
FONTE: O autor (2016).

FIGURA 21 – PONTOS UTILIZADOS NOS CONJUNTOS cj17 e E cj18.



FONTE: O autor (2016).

FIGURA 22 – PONTOS UTILIZADOS NOS CONJUNTOS cj19 E cj20.



FONTE: O autor (2016).

4.2.4.2 Seleção de atributos

Esta etapa objetivou retirar previamente do processamento aqueles dados que não são relevantes para a mineração de dados, contribuindo para a melhoria do modelo e da obtenção de resultados mais objetivos (KIM; STREET; MENCZER, 2003; VERCELLIS, 2009).

A seleção dos atributos para mineração de dados foi feita empregando a Análise de Componentes Principais – ACP, também disponível no WEKA por meio do algoritmo *PrincipalComponents*, observando a fundamentação relatada por Varella (2008) e apresentada na seção 3.2.5.

4.2.4.3 Classificação: definição do modelo da rede neural artificial

O WEKA implementa a rede de alimentação direta com múltiplas camadas (*Multilayer Perceptrons*) usando função de ativação sigmoide e algoritmo de aprendizado por retropropagação de erros (b*ackpropagation*), cuja fundamentação matemática foi discutida por Russell e Norvig (2004) e apresentada na seção 3.2.4.

O modelo de RNA usado neste estudo foi definido com suporte do WEKA. O WEKA adota estratégia baseada na quantidade de atributos de entrada e classes de saída para definir a quantidade de camadas ocultas. Deste modo, as opções disponíveis no WEKA permitem definir as camadas ocultas a partir das fórmulas 41 a 44 seguintes:

$$Quantidade\ camadas\ escondidas = \frac{(quantidade\ atributos + quantidade\ classes)}{2} \quad (41)$$

$$Quantidade\ camadas\ escondidas = Quantidade\ de\ atributos \quad (42)$$

$$Quantidade\ camadas\ escondidas = \ Quantidade\ de\ classes \quad (43)$$

$$Quantidade\ camadas\ escondidas = Quantidade\ de\ atributos\ + Quantidade\ de\ classes \quad (44)$$

Ao final do processamento dos dados de treinamento da RNA, o WEKA fornece um sumário com estatísticas úteis para avaliar o resultado alcançado (*Evaluation on training set* – Summary). As fórmulas 45 a 56 (CARDOSO, 2008; ISRAEL, 2006; LANDIS; KOCH, 1977) serão utilizadas para comparar os resultados e escolher as melhores redes.

O índice ou estatística Kappa (*Kappa statistic*) é uma medida que indica o grau de concordância entre dois classificadores, que considera as probabilidades de que as concordâncias tenham acontecido ao acaso, conforme a fórmula 45, em que *Pr(α)* é a concordância observada e *Pr(ε)* é a concordância esperada ao acaso.

$$K = \frac{\Pr(\alpha) - Pr(\varepsilon)}{1 - \Pr(\varepsilon)} \quad (45)$$

O erro absoluto médio (*Mean absolute error* – MAE) indica a média do afastamento de todos os valores fornecidos pelos classificadores e o seu real valor, calculado pela equação 46, onde *n* é o número de amostras, x_i é o valor fornecido pelo classificador para a i-ésima amostra e $\bar{x}$ é a média dos valores de todas as amostras.

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|x_i - \bar{x}| \tag{46}$$

A raiz do erro médio quadrático (*Root mean squared error* – RMSE) estima a qualidade do classificador e pode ser calculada pela equação 47, onde *n* é o número de amostras, x_i é o valor fornecido pelo classificador para a i-ésima amostra e $\bar{x}$ é a média dos valores de todas as amostras.

$$RMSE = \sqrt{\frac{1}{n}}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{47}$$

O erro relativo médio (*Relative absolute error* – RAE) também é usado para estimar a qualidade de um classificador, e pode ser calculado pela fórmula 48 onde *n* é o número de amostras, x_i é o valor fornecido pelo classificador para a i-ésima amostra e $\bar{x}$ é a média dos valores de todas as amostras, e $\bar{\bar{x}}_i$ é o valor correto que deve ser fornecido pelo classificador.

$$RAE = \frac{\sum_{i=1}^{n}|x_i - \bar{\bar{x}}_i|}{\sum_{i=1}^{n}|x_i - \bar{x}|} \tag{48}$$

A raiz do erro quadrático relativo (*Root relative squared error* – RRSE) é outra medida usada para estimar a qualidade de um classificador, podendo ser calculada pela fórmula 49, onde *n* é o número de amostras, x_i é o valor fornecido pelo classificador para a i-ésima amostra, $\bar{x}$ é a média dos valores de todas as amostras e $\bar{\bar{x}}_i$ é o valor correto que deve ser fornecido pelo classificador.

$$RRSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{\bar{x}}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{49}$$

Positivos verdadeiros (*True positive* – TP) referem-se ao número de instâncias previstas positivas e que, realmente, são positivas. Falsos positivos (*False positive* – FP) referem-se ao número de instâncias previstas positivas, mas que são, na verdade, negativas. Por outro lado, negativos verdadeiros (*True negatives* – TN) referem-se à quantidade de instâncias previstas como negativo e que são, na realidade, negativo. Falsos negativos (*False negative* – FN) dizem respeito ao número de instâncias previstas como negativo, mas que são, na realidade, positivo.

Sensitividade (*Recall*) corresponde à porcentagem de amostras positivas classificadas corretamente sobre o total de amostras positivas, cuja fórmula é definida por:

$$Recall = \frac{true-pos}{pos} \tag{50}$$

A taxa de falsos negativos também é conhecida como especificidade (*Specifity*), dada pela seguinte fórmula:

$$Specificity = \frac{true-neg}{neg} \tag{51}$$

Precisão (*Precision*) equivale à porcentagem de amostras positivas classificadas corretamente sobre o total de amostras classificadas como positivas.

$$Precision = \frac{true-pos}{true-pos+falso-pos} \tag{52}$$

F-*measure* ou F-s*core* é uma média ponderada de precisão e recall, da seguinte forma:

$$F = \frac{2*(Precision*Recall)}{(Precision+Recall)} \tag{53}$$

Curvas ROC (*Receiver Operating Characteristic Curve* – ROC) representam a relação entre sensitividade e a especificidade. Neste caso, considera-se a medida da área abaixo da curva ROC (*Area Under the Curve* – AUC) para comparar a

performance de classificadores, sendo que quanto maior a área AUC melhor a performance global do classificador.

Para a avaliação dos resultados obtidos com o processamento dos conjuntos de teste foi considerada a matriz de erro proposta por Congalton e Green (2009).

FIGURA 23 – MATRIZ DE ERROS.

		Classe atual				
		A	B	C	D	∑
Classe prevista	A	n_{AA}	n_{AB}	n_{AC}	n_{AD}	n_{A+}
	B	N_{BA}	N_{BB}	N_{BC}	N_{BD}	n_{B+}
	C	n_{CA}	n_{CB}	n_{CC}	n_{CD}	n_{C+}
	D	N_{DA}	n_{DB}	n_{DC}	n_{DD}	n_{D+}
	∑	$N+_{A}$	n_{+B}	n_{+C}	n_{+D}	n

FONTE: Congalton e Green (2009).

A acurácia total é dada pelo número de acertos dividido pelo número total de ocorrências. Na matriz de erro, a diagonal principal representa os pontos classificados corretamente, de acordo com os dados de referência. Os erros estão localizados fora da diagonal principal. Na fórmula 54, *q* equivale ao número de classes.

$$Acurácia\ total = \frac{\sum_{k=1}^{q} n_{kk}}{n} X\ 100 \tag{54}$$

Para cada uma das categorias da matriz de erro pode-se encontrar sua acurácia individual. Congalton e Green (2009) apresentaram os conceitos de acurácia do ponto de vista do usuário (ou confiança), que representa os erros de comissão; e acurácia do ponto de vista do produtor, que indica a probabilidade de um *pixel* de referência ser classificado corretamente e representa os erros de omissão.

$$Acurácia\ do\ usuário = \frac{n_{ii}}{n_{i+}} \tag{55}$$

$$Acurácia\ do\ produtor = \frac{n_{ii}}{n_{+i}} \tag{56}$$

A norma ET-CQDG define duas medidas para acurácia da classificação, a saber: exatidão global da classificação (calculada de maneira similar à Fórmula 54) e o índice Kappa (calculado de maneira similar à Fórmula 45).

Quanto à completude, a ET-CQDG define medidas para os elementos excesso e omissão, calculados conforme as fórmulas 57 e 58.

$$Excesso = \frac{itens\ em\ excesso}{tamanho\ da\ referência} \tag{57}$$

$$Omissão = \frac{itens\ em\ falta}{tamanho\ da\ referência} \tag{58}$$

A taxa de acerto é prevista na norma ISO 19157 e referenciada na norma ET-CQDG como uma medida básica de qualidade, nos termos expressos na fórmula 59.

$$Taxa\ de\ acerto = \frac{Quantidade\ de\ acertos}{número\ de\ elementos} \tag{59}$$

4.2.4.4 Pós-processamento

Uma etapa de pós-processamento dos dados foi realizada visando a obtenção da rede de drenagem final. Após o processamento dos dados no Weka, e a consequente classificação pela RNA, os conjuntos de dados foram ser convertidos do formato ARFF (texto) para o formato vetorial (pontos). Esta conversão foi realizada no QGIS, e o arquivo vetorial resultante foi armazenado no banco de dados.

Em seguida, os pontos classificados como drenagem foram convertidos para linha. Nesta etapa do processo, um arquivo vetorial continha as linhas representando os trechos de drenagens previstos pela RNA. Os trechos de drenagem com dimensões pequenas (menores que 8 mm) foram excluídos, visando atender a regra de aquisição da geometria para a escala de 1:100.000, conforme estabelecido na norma ET-ADGV (DSG, 2011). No caso da existência de trechos de drenagens isolados (linhas desconectadas) foi realizada uma edição vetorial para remoção de tais feições.

5 RESULTADOS E DISCUSSÃO

Este capítulo se destina a apresentar os resultados e está dividido em cinco seções que tratarão, respectivamente: da avalição da acurácia da base cartográfica da rede de drenagem de referência; do banco de dados da pesquisa; da definição da RNA; da classificação dos dados por meio da RNA; e, do mapeamento da rede de drenagem da BHRMP realizado com a metodologia da pesquisa.

5.1 AVALIAÇÃO DA ACURÁCIA DO MAPEAMENTO DA REDE DE DRENAGEM DA BACIA HIDROGRÁFICA DO RIO MUTUM-PARANÁ

A escala de representação dos dados 1:100.000 foi considerada por Silva (2001) como nível intermunicipal e cobre, com relativo detalhe, os processos que se desdobram neste nível. De acordo com o autor, nesta escala se considera a dependência dos eventos e processos que existem no espaço geográfico abrangido pelos limites de mais de um município. O autor destacou, ainda, o uso das escalas de 1:50.000 e 1:100.000 nos estudos de pequenas bacias hidrográficas.

A rede de drenagem de referência contém 741 canais de drenagem mapeados, e representa 321 nascentes (316 nascentes se considerar a diminuição de 5 nascentes mapeadas incorretamente). Com base nas imagens SPOT 5 e apoio de levantamentos de campo, foram observadas 186 nascentes não mapeadas que constituem os erros de omissão. Por outro lado, cinco nascentes foram mapeadas de forma equivocada, caracterizando, assim, os erros de excesso. Analisando o conceito de completude, a taxa de acerto ficou em 62,94%, itens em excesso 0,99% e itens omitidos 37,05%.

$$Taxa\ de\ acerto = \frac{316}{502} X\ 100 = 62{,}94\% \quad (60)$$

$$Excesso = \frac{5}{502} X\ 100 = 0{,}99\% \quad (61)$$

$$Omissão = \frac{186}{502} X\ 100 = 37{,}05\% \quad (62)$$

Outra possibilidade de avaliação é dada pelas normas ISO 19157 e ET-CQDG, as quais consideram a acurácia temática que, para Weber et al. (1999), é também conhecida como acurácia de atributos e retrata a fidelidade dos dados descritivos, com uma avaliação geral da identificação de entidades e atribuição de valores de atributo no conjunto de dados. Após verificação de todos os trechos de primeira ordem mapeados, constatou-se que 162 trechos foram mapeados de forma incorreta e não correspondiam à primeira ordem, visto que deveriam ser de segunda ou terceira ordem. Portanto, a taxa de acerto do mapeamento, levando-se em conta a correta identificação da ordem dos canais de primeira ordem mapeados (340 mapeados corretamente), ficou em 67,72%.

$$Taxa\ de\ acerto\ = \frac{340}{502} X\ 100 = 67{,}72\ \% \tag{63}$$

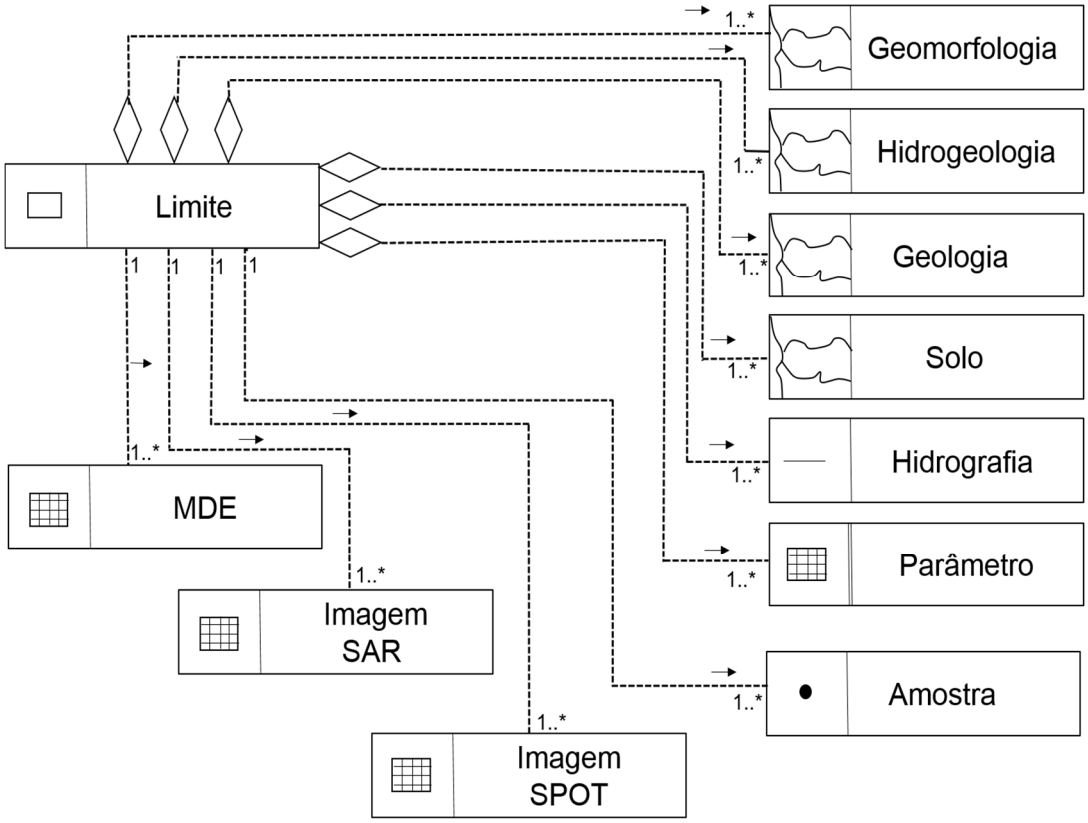
Nos termos discutidos nos trabalhos de Antunes e Lingnau (1997), Selby (1985) e Sampaio (2008), os níveis de acurácia observados estabelecem restrições quanto ao uso deste mapeamento para a realização de estudos hidrológicos, ambientais e de planejamento em nível intermunicipal.

Ainda no que tange à acurácia temática, particularmente se considerar o atributo regime de drenagem, é interessante observar que a classificação dos cursos de água como intermitentes/permanentes foi realizada de forma correta. Durante os trabalhos de campo foi possível perceber que os rios mapeados como intermitentes, de fato, apresentam as características que o permitem classificá-los desta forma. Com efeito, na área de estudo é marcante a existência de rios intermitentes, a exemplo do que ocorre em diversas outras bacias inseridas na Região Amazônica.

5.2 BANCO DE DADOS DA PESQUISA

Na FIGURA 24 é apresentado o diagrama de classes para o banco de dados da pesquisa. A notação simplificada do diagrama segue os conceitos definidos para o modelo conceitual OMT-G (BORGES, 1997). O banco de dados reuniu todos os dados vetoriais e matriciais usados na pesquisa. O dicionário de dados é apresentado na TABELA 5.

FIGURA 24 – DIAGRAMA CONCEITUAL SIMPLIFICADO DO BANCO DE DADOS DA PESQUISA, EM NOTAÇÃO OMT-G.



FONTE: O autor (2016).

TABELA 5 – CONJUNTOS DE DADOS USADOS NA MINERAÇÃO DE DADOS.

continua

Classe	**Descrição**		
Limite	Limites da bacia hidrográfica.		
Atributo	**Tipo**	**Tamanho**	**Descrição**
gid	Inteiro	-	Identificador único
nome	Alfanumérico	20	Nome da bacia hidrográfica
geom	Geometria	-	Representação da geometria do objeto
Classe			**Descrição**
Hidrografia	Canais de drenagem inseridos nos limites da bacia hidrográfica		
Atributo	**Tipo**	**Tamanho**	**Descrição**
gid	Inteiro	-	Identificador único
nome	Alfanumérico	100	Nome atribuído ao trecho de drenagem
regimedrenagem	Alfanumérico	12	Regime de drenagem do trecho
geom	Geometria	-	Representação da geometria do objeto
Classe			**Descrição**
Geomorfologia	Unidades geomorfológicas inseridas nos limites da bacia hidrográfica.		
Atributo	**Tipo**	**Tamanho**	**Descrição**
gid	Inteiro	-	Identificador único
codigogeom	Alfanumérico	5	Código da unidade geomorfológica
padraorelevo	Alfanumérico	50	Padrão de relevo da unidade geomorfológica
unidadegeom	Alfanumérico	120	Nomenclatura da unidade geomorfológica
geom	Geometria	-	Representação da geometria do objeto

continuação

Classe	Descrição		
Hidrogeologia	Unidades hidrogeológicas inseridas nos limites da bacia hidrográfica.		
Atributo	**Tipo**	**Tamanho**	**Descrição**
gid	Inteiro	-	Identificador único
simbolohidro	Alfanumérico	8	Símbolo da unidade hidrogeológica
unidadehidro	Alfanumérico	120	Nomenclatura da unidade hidrogeológica
geom	Geometria	-	Representação da geometria do objeto
Classe	**Descrição**		
Geologia	Geologia da área da bacia hidrográfica.		
Atributo	**Tipo**	**Tamanho**	**Descrição**
gid	Inteiro	-	Identificador único
unidadelitoe	Alfanumérico	120	Nome da unidade litoestratigráfica
litotipo	Alfanumérico	120	Nome dos litotipos associados
classerocha	Alfanumérico	120	Classificação das rochas
geom	Geometria	-	Representação da geometria do objeto
Classe	**Descrição**		
Solo	Classificação dos solos da bacia hidrográfica.		
Atributo	**Tipo**	**Tamanho**	**Descrição**
gid	Inteiro	-	Identificador único
classe	Alfanumérico	120	Classificação do solo
legenda	Alfanumérico	8	Legenda da classe de solo
geom	Geometria	-	Representação da geometria do objeto
Classe	**Descrição**		
MDE	Modelo Digital de Elevação da área da bacia hidrográfica.		
Atributo	**Tipo**	**Tamanho**	**Descrição**
gid	Inteiro	-	Identificador único
geom	Geometria	-	Representação da geometria do objeto
Classe	**Descrição**		
Imagem SAR	Imagem de radar SAR da área da bacia hidrográfica.		
Atributo	**Tipo**	**Tamanho**	**Descrição**
gid	Inteiro	-	Identificador único
geom	Geometria	-	Representação da geometria do objeto
Classe	**Descrição**		
Imagem SPOT	Imagem de satélite SPOT 5 da área da bacia hidrográfica.		
Atributo	**Tipo**	**Tamanho**	**Descrição**
gid	Inteiro	-	Identificador único
geom	Geometria	-	Representação da geometria do objeto
Classe	**Descrição**		
Amostra	Amostras verificadas em campo.		
Atributo	**Tipo**	**Tamanho**	**Descrição**
gid	Inteiro	-	Identificador único
nome	Alfanumérico	20	Nome da bacia hidrográfica
elevacao	Inteiro	-	Elevação do ponto, verificada com GPS
Mapeado	Booleano	-	Indica se o ponto foi considerado no mapeamento de referência
Classe	Alfanumérico	3	Classificação do ponto amostral (drenagem - dre, não drenagem - ndr, nascente - nas)
geom	Geometria	-	Representação da geometria do objeto

continuação/conclusão

Classe	Descrição		
Parâmetro	Parâmetros morfométricos e de direção de fluxo extraídos do MDE para pontos da área da bacia hidrográfica.		
Atributo	**Tipo**	**Tamanho**	**Descrição**
gid	Inteiro	-	Identificador único
ci_converg	Real	-	Índice de convergência
cndb_flowc	Real	-	Conectividade do fluxo
cndb_flowd	Real	-	Direção do fluxo
cndb_strah	Real	-	Ordem de Straher
lsf_lsfact	Real	-	Fator topográfico (LS)
mf_aspect	Real	-	Aspecto
mf_crosscu	Real	-	Curvatura seccional
mf_general	Real	-	Superfície generalizada
mf_longitu	Real	-	Curvatura longitudinal
mf_maximum	Real	-	Curvatura máxima
mf_minimum	Real	-	Curvatura mínima
mf_morfome	Real	-	Feição morfométrica
mf_plancur	Real	-	Plano de curvatura
mf_profile	Real	-	Perfil de curvatura
mf_slope	Real	-	Declividade
srtmsca	Real	-	Áreas de captação da bacia (D∞)
srtmang	Real	-	Angulo (D∞)
srtmslope	Real	-	Declividade (D∞)
slfa_flowac	Real	-	Fluxo acumulado
tri_terrai	Real	-	Índice de rugosidade do terreno
twi_topogr	Real	-	Índice de proteção topográfico
ndwi	Real	-	Índice de índice de água por diferença normalizada
geom	Geometria	-	Representação da geometria do objeto
Classe	**Descrição**		
Limite	Representação dos limites da área de estudo.		
Atributo	**Tipo**	**Tamanho**	**Descrição**
gid	Inteiro	-	Identificador único
nome	Alfanumérico	20	Nome da bacia hidrográfica
geom	Geometria	-	Representação da geometria do objeto

FONTE: O autor (2016).

É possível discutir o banco de dados resultante, bem como a maneira como este interage com os demais componentes de software adotados, nas perspectivas descritas nos trabalhos de Appice, Lanza e Malerba (2007) e Xu, Qi e Wang (2008), que defenderam a integração de SIG, SGBD, Sensoriamento Remoto e mineração de dados de forma semelhante ao que aconteceu nesta pesquisa.

A arquitetura adotada no presente estudo se contrapõe àquelas descritas nos trabalhos de Sousa, Souza Filha e Pereira (2014), Lima e Cunha (2014) e Kumar, Asadi e Vutukuru (2016), quando os autores usaram as tecnologias de SIG, SGBD e Sensoriamento Remoto, mas, na prática, falharam em configurar arquiteturas verdadeiramente integradas. De forma geral, nos trabalhos citados, as

tecnologias foram usadas de forma isolada, por meio de softwares que acessavam dados em arquivos armazenados separadamente ou em formatos diversos.

De forma complementar, assim como foi relatado nos trabalhos de Liu (2012), Berhanu, Melesse e Seleshi (2013), Mohd et al. (2014) e Zhu et al. (2014), ressalta-se a opção pelo armazenamento de dados espaciais com apoio de SGBD, de forma integrada ao invés de manter os dados isolados e independentes para cada projeto e/ou aplicação.

Porém, comparativamente aos trabalhos dos autores mencionados, os componentes de software usados nesta pesquisa possuem como característica marcante o suporte a padrões abertos, extremamente importante no sentido de garantir que o banco de dados possa ser facilmente acessado, independente do software escolhido pelo usuário interessado, justamente como foi proposto no Plano de Ação para Implantação da INDE (CONCAR, 2010).

Assim como aconteceu nos trabalhos de Huang et al. (2010), Houston et al. (2011) e López et al. (2015), da maneira que foi estruturado e da forma como os dados foram armazenados, este banco de dados apresenta potencial para uso em diversas pesquisas, visto que armazena variáveis ambientais de uso generalizado em inúmeros estudos, embora tenha sido elaborado especificamente para os propósitos da presente pesquisa.

Como no cenário proposto para funcionamento da INDE do Brasil, dada à escolha do SGBD PostgreSQL/PostGIS, é viável também pensar na distribuição dos dados armazenados pela Internet, por intermédio de *web services*, assim como sugerido para a INDE do Brasil (CONCAR, 2010), e como discutido nos trabalhos de Camboim (2013) e Paparrizos et al. (2014).

5.3 ESTRUTURAÇÃO DA REDE NEURAL ARTIFICIAL

A estruturação da RNA no WEKA envolveu a definição da quantidade de camadas ocultas, da taxa de aprendizagem, do valor de momentum e do número de épocas. Optou-se por uma abordagem de experimentação, semelhante ao que foi sugerido por Russell e Norvig (2004), baseada em testes simultâneos visando obter a maior quantidade de acertos na classificação e menor erro.

A camada de entrada foi definida em função dos atributos usados, ou seja, cada nó da camada de entrada correspondeu a uma variável daquelas que

compuseram o banco de dados da pesquisa. Por meio da camada de entrada, os padrões (valores dos atributos para cada ponto observado) são apresentados à RNA.

Foram definidas, também, três camadas intermediárias, cuja quantidade de nós considerou, inicialmente, a fórmula 41 apresentada na seção 4.2.4.3. Portanto, a RNA foi estruturada com uma camada de entrada (42 neurônios), três camadas escondidas (119 neurônios) e uma camada de saída (2 neurônios). Nas camadas escondidas ocorre a maior parte do processamento, nas quais as características dos dados são extraídas e, posteriormente, usadas para classificar novos dados.

A camada de saída foi definida em função das classes de saída esperadas: drenagem (dre) e não drenagem (ndr). Nesta camada, o resultado final do processamento pela RNA é apresentado. A taxa de aprendizagem – parâmetro importante que influencia na velocidade e na corretude do aprendizado – foi definida como 0,1. O valor de momentum, que geralmente é usado em conjunto com a taxa de aprendizagem, foi definido em 0,09. A quantidade de ciclos de treinamento, também denominado de épocas, foi estabelecida em 35.000.

Na TABELA 6 são apresentados os valores para os principais parâmetros que definiram a RNA final da pesquisa (primeira linha da tabela, destacado em vermelho), e também valores relativos a outros trabalhos que podem ser utilizados para comparação, visto que todos consideraram RNAs para manipular variáveis ambientais em estudos da área de Geociências.

TABELA 6 – Resumo do comparativo entre arquiteturas de algumas RNAs.

continua

Trabalho	Quant. Camadas	Neurônios por camadas	Taxa aprendizagem	Momentum	Número épocas
RNA da pesquisa	**1 x 3 x 1**	**42 x 39 x 39 x 39 x 2**	**0,1**	**0,09**	**35.000**
Gonçalves (1997)	1 x 1 x 1	27 x 14 x 5	0,25 a 0,75	Sem informação	637 a 623.721
Shrestha, Theobald e Nestmann (2005)	1 x 2 x 1	8 x 4 x 4 x 1	0,5	Sem informação	Sem informação
Strobl e Forte (2007)	1 x 1 x 1	48 x 13.447 x 1	Sem informação	Sem informação	Sem informação
Sirtoli (2008)	1 x 2 x 1	12 x * x 3 * melhores arquiteturas: (100,35); (60,18); (29,9) e (81,20)	Sem informação	Sem informação	47 a 148

continuação/conclusão

Trabalho	Quant. Camadas	Neurônios por camadas	Taxa aprendizagem	Momentum	Número épocas
Agarwal, Rai e Upadhyay (2009)	1 x 1 x 1	Sem informação	0,5	0,5	5.000
Mendes e Marengo (2009)	1 x 1 x 1	11 x * x 1 * variou entre 11 e 189	Sem informação	Sem informação	500
Coneglian, Gomes e Ribeiro (2010)	1 x 1 x 1; 1 x 2 x 1; 1 x 3 x 1	Sem informação	0,1 a 0,7	0,2 a 0, 9	2.500
Pradhan e Lee (2010)	1 x 1 x 1	10 x 22 x 2	0,01	0,01	2.000
Poudyal *et al.* (2010)	1 x 1 x 1	10 x 20 x 1	0,01	0,01	Sem informação
Silveira (2010)	1 x 2 x 1	8 x 138 x 43 x 3	0,1	0,09	15.000
Pradhan e Buchroithner (2010)	1 x1 x 1	11 x 23 x 2	0,01	0,01	2.000
Pradhan (2011)	1 x 1 x 1	9 x 19 x 2	0,01	0,01	2.500
Sing e Pand (2011)	1 x 1 x 1	5 x * x 1 * variou entre 1 e 5	Sem informação	Sem informação	100
Lin (2011)	1 x 1 x 1	6 x 6 x 1	Sem informação	Sem informação	Sem informação
Kia *et al.* (2011)	1 x 2 x 1	7 x 20 x 10 x 1	Sem informação	Sem Informação	2.000
Andrade (2011)	1 x 1 x 1	6 x 14 x 5	0,02	0,5	10.000
Memarian e Balasundram (2012)	1 x 2 x 1	10 X 20 x 10 x 1	0,01	0,7	1.200
Arruda, Dematê e Chagas (2013)	1 X 1 X 1	22 x 11 X 9	0,2	Sem informação	10.000

FONTE: O autor (2016).

Tal qual adotada nesta pesquisa, a estratégia de tentativa e erro para definir a estrutura de uma RNA foi defendida por Lawrence, Giles e Tisoy (1996), Russel e Norvig (2004), Mendes e Marengo (2009) e Turban et al. (2010). Silveira (2010) usou tal estratégia para definir a quantidade de neurônios nas camadas intermediárias.

A RNA definida conteve um maior número de camadas escondidas, em comparação aos trabalhos de Shrestha, Theobald e Nestmann (2005), Strobl e Forte (2007), Sirtoli (2008), Pradhan e Lee (2010), Silveira (2010) e Sing e Panda (2011). Semelhante ao que foi discutido por Hamamoto et al. (1988), Lawrence, Giles e Tisoy (1996), Russel e Norvig (2004) e Hernández et al. (2014), percebeu-se que a maior quantidade de camadas ocultas aumentou o poder de representatividade da rede. Por outro lado, as configurações testadas com mais do que três camadas

intermediárias sempre conduziram à superadaptação da rede e demandaram mais recursos computacionais para o processamento. Este comportamento também já tinha sido notado pelos autores citados.

Analisando a taxa de aprendizagem, percebeu-se que o valor final adotado (0,1) foi semelhante àqueles usados nos trabalhos de Coneglian, Gomes e Ribeiro (2010) e Silveira (2010). Os apontamentos de Turban et al. (2010) e Faceli et al. (2011) foram verificados neste estudo, visto que quando testado em valor menor que 0,1 para a taxa de aprendizagem, exigiu-se mais processamento, e quando testado valor maior que 0,1 os resultados foram piores.

O valor definido para o momentum também ficou idêntico àqueles estabelecidos por Coneglian, Gomes e Ribeiro (2010) e Silveira (2010). O valor é também muito próximo aos definidos nos trabalhos de Gonçalves (1997) e de Memarian e Balasundram (2012).

A comparação da quantidade de ciclos de treinamento (ou épocas) estabelecida na pesquisa, em relação a outros trabalhos, permitiu verificar que se, por um lado, a quantidade de 35.000 épocas foi muito superior àqueles adotados por Agarwal, Rai e Upadhyay (2009), Pradhan e Lee (2010), Memarian e Balsundram (2012) e Arruda, Demattê e Chagas (2013), por outro ficou bem abaixo de experimentos de Gonçalves (1997), que chegou a considerar 623.721 épocas.

O trabalho de Gonçalves (1997) teve como objetivo principal propor uma arquitetura para classificação de imagens multiespectrais de sensoriamento remoto, baseada em RNA. Na pesquisa, o autor investigou a performance do algoritmo *Backpropagation* em termos de tempo de execução e número de épocas. Assim como observado nos diversos experimentos realizados por Gonçalves (1997), o número de épocas influencia diretamente no tempo de execução do *Backpropagation*.

5.4 RESULTADOS DA CLASSIFICAÇÃO POR RNA

5.4.1. Treinamento da RNA

Na fase de treinamento da rede, as amostras contidas nos conjuntos de dados cj1, cj2, cj3 e cj4 foram fornecidas como entrada ao software e processadas pelo algoritmo de classificação *multilayerperceptron*. Os resultados foram, então,

analisados, principalmente em relação ao percentual de instâncias classificadas corretamente.

Os melhores resultados foram obtidos usando todas as variáveis disponíveis na pesquisa (MDE, parâmetros morfométricos e de direção de fluxo, geologia, geomorfologia, hidrogeologia, solos, imagem) e considerando duas classes de saída (drenagem, não drenagem).

Para o conjunto de dados com pixel de 6 metros (cj1), no melhor resultado na fase de treinamento foram classificadas corretamente 98,77% das instâncias. O índice Kappa ficou em 0,97. As demais estatísticas serão apresentadas a seguir.

```
=== Dados com pixel de 6 metros ===
=== MDE, parâmetros morfométicos e de direção de fluxo, geologia, geomorfologia, hidrogeologia,
solo, SAR ===
=== dre, ndr ===
=== Evaluation on training set ==
=== Summary ===
Correctly Classified Instances         484               98.7755 %
Incorrectly Classified Instances         6                1.2245 %
Kappa statistic                          0.9755
Mean absolute error                      0.013
Root mean squared error                  0.1107
Relative absolute error                  2.6022 %
Root relative squared error             22.1325 %
Total Number of Instances              490

=== Detailed Accuracy By Class ===
                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   1         0.024      0.976      1         0.988      0.98      dre
                   0.976     0          1          0.976     0.988      0.979     ndr
Weighted Avg.    0.988     0.012      0.988      0.988     0.988      0.979

=== Confusion Matrix ===
   a     b   <-- classified as
 245     0   |   a = dre
   6   239 |   b = ndr
```

O melhor resultado obtido no treinamento foi conseguido usando o conjunto de dados com pixel de 2,5 metros (cj2), tomando todas as variáveis disponíveis e considerando duas classes de saída. Neste caso, 99,38% das amostras foram classificadas corretamente, e o índice Kappa foi de 0,987.

```
=== Dados com pixel de 2,5 metros ===
=== MDE, parâmetros morfométicos e de direção de fluxo, geologia, geomorfologia, hidrogeologia,
solos, SPOT 5 ===
=== dre, ndr ===
=== Evaluation on training set ===
=== Summary ===
Correctly Classified Instances         487               99.3878 %
```

```
Incorrectly Classified Instances         3               0.6122 %
Kappa statistic                          0.9878
Mean absolute error                      0.0074
Root mean squared error                  0.0783
Relative absolute error                  1.4741 %
Root relative squared error             15.6546 %
Total Number of Instances              490

=== Detailed Accuracy By Class ===
                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.996     0.008     0.992     0.996     0.994     0.988     dre
                   0.992     0.004     0.996     0.992     0.994     0.988     ndr
Weighted Avg.      0.994     0.006     0.994     0.994     0.994     0.988

=== Confusion Matrix ===
   a   b   ← classified as
 244   1 |  a = dre
   2 243 |  b = ndr
```

Quando foram consideradas três classes de saída (drenagem, não drenagem e nascente), o número de instâncias classificadas corretamente durante a fase de treinamento foi menor em 3,46% para dados com pixel de 6 metros e 1,63% para dados com pixel de 2,5 metros.

As estatísticas demonstraram que para o conjunto de dados com pixel de 6 metros (cj3), tomando todas as variáveis disponíveis e considerando três classes de saída, o percentual de acertos ficou em 95,31%, com índice Kappa de 0,92.

```
=== Dados com pixel de 6 metros ===
=== MDE, parâmetros morfométicos e de direção de fluxo, geologia, geomorfologia, hidrogeologia, solo, SAR ===
=== dre, nas, ndr ===
=== Evaluation on training set ===
=== Summary ===
Correctly Classified Instances         895               95.3142 %
Incorrectly Classified Instances        44                4.6858 %
Kappa statistic                          0.9258
Mean absolute error                      0.0325
Root mean squared error                  0.157
Relative absolute error                  7.6819 %
Root relative squared error             34.1564 %
Total Number of Instances              939

=== Detailed Accuracy By Class ===
                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.935     0.017     0.951     0.935     0.943     0.977     dre
                   0.978     0.043     0.955     0.978     0.966     0.981     ndr
                   0.925     0.016     0.953     0.925     0.938     0.972     nas
Weighted Avg.      0.953     0.029     0.953     0.953     0.953     0.978

=== Confusion Matrix ===
   a   b   c   <-- classified as
 232  11   5 |  a = dre
   4 442   6 |  b = ndr
```

```
  8    10   221 |   c = nas
```

Para o conjunto de dados com pixel de 2,5 metros (cj4), no treinamento com três classes de saída, o percentual de acertos ficou em 97,75%, com índice Kappa de 0,96.

```
=== Dados com pixel de 2,5 metros ===
=== MDE, parâmetros morfométicos e de direção de fluxo, geologia, geomorfologia, hidrogeologia, solo, SPOT 5 ===
=== dre, nas, ndr ===
=== Dados com pixel de 2,5 metros ===
=== MDE, parâmetros morfométicos, geologia, geomorfologia, hidrogeologia, solo, SPOT 5 ===
=== Evaluation on training set ===
=== Summary ===
Correctly Classified Instances         914               97.754  %
Incorrectly Classified Instances        21                2.246  %
Kappa statistic                          0.9644
Mean absolute error                      0.0161
Root mean squared error                  0.114
Relative absolute error                  3.8202 %
Root relative squared error             24.8145 %
Total Number of Instances              935

=== Detailed Accuracy By Class ===
                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.963    0.006      0.983      0.963     0.973      0.975     dre
                   0.991    0.029      0.97       0.991     0.98       0.982     ndr
                   0.967    0.004      0.987      0.967     0.977      0.98      nas
Weighted Avg.      0.978    0.017      0.978      0.978     0.978      0.98

=== Confusion Matrix ===
   a     b     c   <-- classified as
 235     8     1 |   a = dre
   2   448     2 |   b = ndr
   2     6   231 |   c = nas
```

Ainda durante a fase de treinamento foram experimentadas variações na quantidade de nós da camada de entrada. Tais variações consistiram em retirar determinados atributos e verificar como se comportou o percentual de acertos das redes treinadas. Os demais parâmetros de configuração da RNA (taxa de aprendizagem, momentum, quantidade de camadas intermediárias, épocas) permaneceram os mesmos apresentados na seção 5.3. Para estas experimentações, foi usado apenas o conjunto de dados com pixel de 2,5 metros e com duas classes de saída (cj2).

A primeira variação treinada consistiu em adotar os atributos extraídos do MDE, geologia, geomorfologia, hidrogeologia e solos como nós da camada de

entrada da RNA. Em outras palavras: não foram consideradas a imagem e o NDWI. O percentual de acertos ficou acima de 97%.

```
=== Dados com pixel de 2,5 metros ===
=== MDE, parâm. morfométicos e de direção de fluxo, geologia, geomorfologia, hidrogeologia, solos =
=== dre, ndr ===
=== Evaluation on training set ===
=== Summary ===
Correctly Classified Instances         476               97.1429 %
Incorrectly Classified Instances        14               2.8571 %
Kappa statistic                          0.9429
Mean absolute error                      0.0476
Root mean squared error                  0.1604
Relative absolute error                  9.5225 %
Root relative squared error             32.0788 %
Total Number of Instances              490

=== Detailed Accuracy By Class ===
                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.984     0.041        0.96     0.984      0.972      0.966    dre
                   0.959     0.016       0.983     0.959      0.971      0.966    ndr
Weighted Avg.      0.971     0.029       0.972     0.971      0.971      0.966

=== Confusion Matrix ===
   a   b   <-- classified as
 241   4 |   a = dre
  10 235 |   b = ndr
```

A próxima variação consistiu na retirada das variáveis correspondentes a geologia, geomorfologia, hidrogeologia e solos, e acréscimo da imagem. O percentual de acertos ficou ligeiramente menor, com pouco mais de 96%.

```
=== Dados com pixel de 2,5 metros ===
=== MDE, parâmetros morfométicos e de direção de fluxo, SPOT 5 ===
=== dre, ndr ===
=== Evaluation on training set ===
=== Summary ===
Correctly Classified Instances         471               96.1224 %
Incorrectly Classified Instances        19               3.8776 %
Kappa statistic                          0.9224
Mean absolute error                      0.0505
Root mean squared error                  0.1962
Relative absolute error                 10.1028 %
Root relative squared error             39.2303 %
Total Number of Instances              490

=== Detailed Accuracy By Class ===
                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.976     0.053       0.948     0.976      0.962      0.938    dre
                   0.947     0.024       0.975     0.947      0.961      0.94     ndr
Weighted Avg.      0.961     0.039       0.962     0.961      0.961      0.939

=== Confusion Matrix ===
   a   b   <-- classified as
```

```
 239      6 |   a = dre
  13    232 |   b = ndr
```

A RNA também foi treinada com a camada de entrada em função apenas dos parâmetros extraídos diretamente do MDE. O percentual de acertos alcançado nesta simulação foi abaixo de 96%.

```
=== Dados com pixel de 2,5 metros ===
=== MDE, parâmetros morfométicos e de direção de fluxo ===
=== dre, ndr ===
=== Evaluation on training set ===
=== Summary ===
Correctly Classified Instances         469               95.7143 %
Incorrectly Classified Instances        21                4.2857 %
Kappa statistic                          0.9143
Mean absolute error                      0.0623
Root mean squared error                  0.1918
Relative absolute error                 12.4615 %
Root relative squared error             38.3604 %
Total Number of Instances              490

=== Detailed Accuracy By Class ===
                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.967     0.053      0.948     0.967     0.958      0.965    dre
                   0.947     0.033      0.967     0.947     0.957      0.965    ndr
Weighted Avg.      0.957     0.043      0.957     0.957     0.957      0.965

=== Confusion Matrix ===
   a     b   <-- classified as
 237     8 |   a = dre
  13   232 |   b = ndr
```

Quando a camada de entrada foi definida em função apenas do valor de elevação contido no MDE, juntamente com os valores contidos nas bandas da imagem SPOT, o percentual de acertos ficou em apenas 64%.

```
=== Dados com pixel de 2,5 metros ===
=== MDE, SPOT 5 ===
=== dre, ndr ===
=== Evaluation on training set ===
=== Summary ===
Correctly Classified Instances         314               64.0816 %
Incorrectly Classified Instances       176               35.9184 %
Kappa statistic                          0.2816
Mean absolute error                      0.3912
Root mean squared error                  0.4439
Relative absolute error                 78.2379 %
Root relative squared error             88.7888 %
Total Number of Instances              490

=== Detailed Accuracy By Class ===
                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
```

	0.939	0.657	0.588	0.939	0.723	0.723	dre
	0.343	0.061	0.848	0.343	0.488	0.723	ndr
Weighted Avg.	0.641	0.359	0.718	0.641	0.606	0.723	

=== Confusion Matrix ===

a	b	<-- classified as
230	15	a = dre
161	84	b = ndr

Porém, substituindo os valores de elevação do MDE pelos parâmetros morfométricos obtidos do MDE, em conjunto com os valores contidos na imagem, o percentual de acertos sobe para 95%.

=== Dados com pixel de 2,5 metros ===
=== parâmetros morfométicos e de direção de fluxo, SPOT 5 ===
=== dre, ndr ===
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances	467	95.3061 %
Incorrectly Classified Instances	23	4.6939 %
Kappa statistic	0.9061	
Mean absolute error	0.0708	
Root mean squared error	0.2134	
Relative absolute error	14.1584 %	
Root relative squared error	42.6866 %	
Total Number of Instances	490	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.947	0.041	0.959	0.947	0.953	0.926	dre
	0.959	0.053	0.948	0.959	0.953	0.926	ndr
Weighted Avg.	0.953	0.047	0.953	0.953	0.953	0.926	

=== Confusion Matrix ===

a	b	<-- classified as
232	13	a = dre
10	235	b = ndr

O treinamento da RNA com a camada de entrada contendo apenas a imagem SPOT 5 resultou num percentual de acertos de 70%.

=== Dados com pixel de 2,5 metros ===
=== SPOT 5 ===
=== dre, ndr ===
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances	343	70 %
Incorrectly Classified Instances	147	30 %
Kappa statistic	0.4	
Mean absolute error	0.3874	
Root mean squared err	0.4396	
Relative absolute error	77.4899 %	
Root relative squared er	87.9233 %	

Total Number of Instances 490

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.482	0.082	0.855	0.482	0.616	0.788	dre
	0.918	0.518	0.639	0.918	0.754	0.788	ndr
Weighted Avg.	0.7	0.3	0.747	0.7	0.685	0.788	

=== Confusion Matrix ===

a	b	<-- classified as
118	127	a = dre
20	225	b = ndr

Na TABELA 7 são relatados os resultados obtidos durante a fase de treinamento da RNA.

TABELA 7 – Resultados obtidos durante a fase de treinamento da RNA.

continua

Conjunto	**Variáveis entrada**	**Classes saída** (dre=drenagem; ndr= não drenagem; nas=nascente)	**Quant. amostras**	**% classif. correta**	**Kappa**	**RMSE**
cj1	MDE, parâmetros morfométricos e de direção de fluxo, geologia, geomorfologia, hidrogeologia, solos, imagem	dre, ndr	490	98,77	0,97	0,11
cj2	MDE, parâmetros morfométricos e de direção de fluxo, geologia, geomorfologia, hidrogeologia, solos, imagem, NDWI	dre, ndr	490	99,38	0,98	0,07
cj3	MDE, parâmetros morfométricos e de direção de fluxo, geologia, geomorfologia, hidrogeologia, solos, imagem	dre, ndr, nas	939	95,31	0,92	0,15
cj4	MDE, parâmetros morfométricos e de direção de fluxo, geologia, geomorfologia, hidrogeologia, solos, imagem, NDWI	dre, ndr, nas	939	97,75	0,96	0,11
cj2(a)	MDE, parâmetros morfométicos e de direção de fluxo, geologia, geomorfologia, hidrogeologia, solos	dre, ndr	490	97,14	0,94	0,16

continuação/conclusão

Conjunto	Variáveis entrada	**Classes saída** (dre=drenagem; ndr= não drenagem; nas=nascente)	Quant. amostras	% classif. correta	Kappa	RMSE
cj2(b)	MDE, parâmetros morfométicos e de direção de fluxo, SPOT 5	dre, ndr	490	96,12	0,92	0,19
cj2(c)	MDE, parâmetros morfométicos e de direção de fluxo	dre, ndr	490	95,71	0,91	0,19
cj2(d)	MDE, SPOT 5	dre, ndr	490	64,08	0,28	0,44
cj2(e)	Parâmetros morfométicos e de direção de fluxo, SPOT 5	dre, ndr	490	95,30	0,90	0,21
cj2(f)	SPOT 5	dre, ndr	490	70	0,4	0,43

FONTE: O autor (2016).

Na fase de treinamento, o percentual de acertos sempre foi melhor quando consideradas apenas duas classes de saída. O acréscimo da classe nascente na camada de saída da RNA tornou o classificador menos capaz de reconhecer o padrão de ocorrência de cada classe. Este resultado conflita com aqueles obtidos por Banon et al. (2013), visto que os autores conseguiram maior percentual de acerto quando usadas as três classes.

Uma situação comum quanto ao uso dos algoritmos para extração automatizada da rede de drenagem diz respeito à correta identificação das nascentes. Bosquilia et al. (2013) observaram uma diferença na quantidade de nascentes identificadas para uma mesma área, com a aplicação de variados algoritmos. Assim como afirmado pelos autores, acredita-se que são múltiplos os fatores antrópicos e naturais que influenciam nesta tarefa.

Os melhores percentuais de acertos foram obtidos no treinamento para os conjuntos cj1 (98,77%) e cj2 (99,38%). Sirtoli (2008), que aplicou RNA no mapeamento de solos, alcançou percentuais que variavam entre 90 e 93,89% nos diversos conjuntos treinados; já Silveira (2010), cujo trabalho investigou o uso de RNA na predição de unidades preliminares de mapeamento de solos, conseguiu percentual de 97,33% nesta mesma fase.

Semelhante aos trabalhos de Vogt et al. (2003) e Strobl e Forte (2007), que ao discutirem algoritmos para extração, defenderam que a derivação das redes de drenagem é influenciada por variáveis ambientais diversas, tais como solos,

vegetação, relevo, etc., nesta pesquisa os processamentos realizados permitiram comparar o desempenho da RNA com a mudança na quantidade de variáveis usadas como atributo de entrada. Os resultados desta pesquisa corroboraram com os argumentos dos citados autores, visto que a supressão de quaisquer variáveis reduziu a capacidade de identificação da drenagem pela RNA.

Entretanto, não obstante aos argumentos de Vogt et al. (2003) e Strobl e Forte (2007), quando considerada apenas a imagem de alta resolução, os resultados não foram superiores a 70% de acertos na classificação (70% para o conjunto cj2(f) e 64,08% para o conjunto cj2(d)). Destaca-se a importância do MDE em combinação com os parâmetros dele extraídos, visto que os melhores resultados só puderam ser obtidos quando tais variáveis foram usadas na camada de entrada.

Como no trabalho de Pradhan e Buchroithner (2010), esta estratégia de variar os atributos da camada de entrada da RNA mostrou-se, em certa medida, útil para avaliar a influência de grupos de variáveis ambientais no resultado final da classificação. A metodologia utilizada mostrou-se flexível, sobretudo no sentido de permitir a agregação de atributos obtidos a partir de fontes de dados diversas, que podem ser usados na camada de entrada da RNA.

Este fator é relevante, levando em conta o ambiente de produção de dados geoespaciais discutidos nos trabalhos de Bravo e Sluter (2015) e CONCAR (2010), visto que as facilidades tecnológicas disponíveis contribuíram para o surgimento das iniciativas de mapeamentos colaborativos, onde a produção e a disponibilização de dados na Internet não estão restritas às instituições oficiais.

Acredita-se que estas iniciativas poderão minimizar a carência de dados para a região Amazônica do Brasil, incluindo aqueles com potencial para uso nos estudos sobre as redes de drenagem. Assim, por meio da RNA poderão ser agregados novos atributos e avaliar sua contribuição para o resultado da classificação.

As simulações confirmaram que o melhor resultado no treinamento foi obtido com o uso de todas as variáveis disponíveis na camada de entrada (conjunto cj2, 99,38% de acertos). O pior resultado foi apresentado pelo modelo que considerou como entrada o MDE e as bandas da imagem SPOT 5 (conjunto cj2(d), 64,08% de acertos). Nos trabalhos de Sirtoli (2008) e Silveira (2010) também foi possível observar esta variação no percentual de acertos em função das diferentes

configurações da camada de entrada da RNA, e, ainda, que os maiores percentuais foram obtidos com o uso de todas as variáveis disponíveis.

Entretanto, destaca-se o percentual de acertos para a variação treinada apenas com o MDE, parâmetros morfométricos e de direção de fluxo que ficou em 95,71%. Este resultado sugere que, para esta área de estudo, e com os dados disponíveis para a pesquisa, os parâmetros extraídos do MDE foram determinantes. Todas as demais variáveis ambientais do estudo, incluindo também as imagens, quando usadas junto com o MDE e parâmetros extraídos, foram suficientes para aumentar o percentual de acertos em apenas 3,67%.

Os diferentes percentuais de acertos obtidos com a variação da configuração da camada de entrada da RNA conduziram para a confirmação do que discutiu Sug (2010), quando observou que as técnicas de mineração de dados são dependentes das características do conjunto de dados disponíveis (SUG, 2010). Embora se tenha alterado a configuração da camada de entrada da RNA, os demais parâmetros permaneceram inalterados.

Assim como sugerido por Ellis e Morgan (1999) e Hernández et al. (2014), acredita-se que o erro pode ser reduzido pelo acréscimo de amostras no conjunto de treinamento. Entretanto, em termos práticos, os próprios autores reconheceram que nem sempre se dispõe de muitas amostras para uso no treinamento, pois quanto maior for o número de amostras para o treinamento, maior será a necessidade de recursos (tempo, esforço, equipamentos, etc.) para adquiri-las e processá-las.

5.4.2. Teste da RNA

Após o treinamento, procedeu-se a generalização para os conjuntos de teste, cujos resultados serão apresentados a seguir. Nesta fase, foram fornecidas para a RNA novas amostras, porém sem informar a que classe pertenciam, e a rede treinada se encarregou de classificar cada uma delas. Para o teste da RNA, definiu-se que a camada de entrada conteria nós correspondentes a todas as variáveis disponíveis.

O conjunto cj5 continha 300 amostras, pixel de 6 metros, considerou todas as variáveis do estudo e duas classes de saída. Neste caso, a acurácia total apurada foi de 68%, como pode ser observada na matriz de confusão.

TABELA 8 – Matriz de confusão para o conjunto de teste cj5, dados com pixel de 6 metros, considerando todas as variáveis do estudo, duas classes de saída. Acurácia total 68%.

	DRE	NDR	Total Linha	Acurácia Usuário %
DRE	98	43	141	70
NDR	52	107	159	67
Total coluna	150	150		
Acurácia produtor	65	71		68

FONTE: O autor (2016).

O conjunto cj6 continha 300 amostras, pixel de 2,5 metros, considerou todas as variáveis do estudo e duas classes de saída. Neste caso, a acurácia total apurada foi de 72%.

TABELA 9 – Matriz de confusão para o conjunto de teste cj6, dados com pixel de 2,5 metros, considerando todas as variáveis do estudo, duas classes de saída. Acurácia total 72%.

	DRE	NDR	Total linha	Acurácia Usuário %
DRE	103	35	138	75
NDR	47	115	162	71
Total coluna	150	150		
Acurácia produtor	69	77		72

FONTE: O autor (2016).

O conjunto cj7 continha 300 amostras, pixel de 6 metros, considerou todas as variáveis do estudo e três classes de saída. Neste caso, a acurácia total apurada foi de 38%.

TABELA 10 – Matriz de confusão para o conjunto de teste cj7, dados com pixel de 6 metros, considerando todas as variáveis do estudo, três classes de saída. Acurácia total 38%.

	DRE	NAS	NDR	Total linha	Acurácia Usuário %
DRE	40	27	35	102	39
NAS	20	38	22	80	48
NDR	90	85	93	268	35
Total Coluna	150	150	150		
Acurácia Produtor %	27	25	62		38

FONTE: O autor (2016).

O conjunto cj8 continha 300 amostras, pixel de 2,5 metros, considerou todas as variáveis do estudo e três classes de saída. Neste caso, a acurácia total apurada foi de 45%.

TABELA 11 – Matriz de confusão para o conjunto de teste I, dados com pixel de 2,5 metros, considerando todas as variáveis do estudo, três classes de saída. Acurácia total 45%.

	DRE	NAS	NDR	Total Linha	Acurácia Usuário %
DRE	60	35	33	128	47
NAS	14	46	19	79	58
NDR	76	69	98	243	40
Total coluna	150	150	150		
Acurácia produtor	40	31	65		45

FONTE: O autor (2016).

O conjunto cj9 continha 170 amostras, pixel de 6 metros, considerou todas as variáveis do estudo e duas classes de saída. Neste caso, a acurácia total apurada foi de 73%.

TABELA 12 – Matriz de confusão para o conjunto de teste cj9I, dados com pixel de 6 metros, considerando todas as variáveis do estudo, duas classes de saída. Acurácia total 73%.

	DRE	NDR	Total linha	Acurácia Usuário %
DRE	59	19	78	76
NDR	26	66	92	72
Total coluna	85	85		
Acurácia produtor	69	78		73

FONTE: O autor (2016).

O conjunto cj10 continha 170 amostras, pixel de 2,5 metros, considerou todas as variáveis do estudo e duas classes de saída. Neste caso, a acurácia total apurada foi de 78%.

TABELA 13 – Matriz de confusão para o conjunto de teste cj10, dados com pixel de 2,5 metros, considerando todas as variáveis do estudo, duas classes de saída. Acurácia total 78%.

	DRE	NDR	Total linha	Acurácia Usuário %
DRE	63	15	78	81
NDR	22	70	92	76
Total coluna	85	85		
Acurácia produtor	74	82		78

FONTE: O autor (2016).

O conjunto cj11 continha 170 amostras, pixel de 6 metros, considerou todas as variáveis do estudo e três classes de saída. Neste caso, a acurácia total apurada foi de 51%.

TABELA 14 – Matriz de confusão para o conjunto de teste cj11, dados com pixel de 6 metros, considerando todas as variáveis do estudo, três classes de saída. Acurácia total 51%. K=0,27.

	DRE	NAS	NDR	Total linha	Acurácia Usuário %
DRE	37	15	22	74	50
NAS	12	44	13	69	64
NDR	36	26	50	112	45
Total coluna	85	85	85		
Acurácia Produtor %	44	52	59		51

FONTE: O autor (2016).

O conjunto cj12 continha 170 amostras, pixel de 2,5 metros, considerou todas as variáveis do estudo e três classes de saída. Neste caso, a acurácia total apurada foi de 57%.

TABELA 15 – Matriz de confusão para o conjunto de teste cj12, dados com pixel de 2,5 metros, considerando todas as variáveis do estudo, três classes de saída. Acurácia total 57%.

	DRE	NAS	NDR	Total linha	Acurácia Usuário %
DRE	39	15	12	66	59
NAS	14	42	8	64	66
NDR	32	28	65	125	52
Total coluna	85	85	85		
Acurácia produtor	46	49	76		57

FONTE: O autor (2016).

O conjunto cj13 continha 222 amostras, pixel de 6 metros, considerou todas as variáveis do estudo e duas classes de saída. Neste caso, a acurácia total apurada foi de 67%.

TABELA 16 – Matriz de confusão para o conjunto de teste cj13, dados com pixel de 6 metros, considerando todas as variáveis do estudo, duas classes de saída. Acurácia total 67%.

	DRE	NDR	Total linha	Acurácia Usuário %
DRE	65	27	92	71
NDR	46	84	130	65
Total coluna	111	111		
Acurácia produtor	59	76		67

FONTE: O autor (2016).

O conjunto cj14 continha 222 amostras, pixel de 2,5 metros, considerou todas as variáveis do estudo e duas classes de saída. Neste caso, a acurácia total apurada foi de 77%.

TABELA 17 – Matriz de confusão para o conjunto de teste cj14, dados com pixel de 2,5 metros, considerando todas as variáveis do estudo, duas classes de saída. Acurácia total 77%.

	DRE	NDR	Total linha	Acurácia Usuário %
DRE	73	13	86	85
NDR	38	98	136	72
Total coluna	111	111		
Acurácia produtor	66	88		77

FONTE: O autor (2016).

O conjunto cj15 continha 222 amostras, pixel de 6 metros, considerou todas as variáveis do estudo e três classes de saída. Neste caso, a acurácia total apurada foi de 55%.

TABELA 18 – Matriz de confusão para o conjunto de teste cj15, dados com pixel de 6 metros, considerando todas as variáveis do estudo, três classes de saída. Acurácia total 55%.

	DRE	NAS	NDR	Total linha	Acurácia Usuário %
DRE	59	15	23	97	61
NAS	14	52	14	80	65
NDR	38	44	74	156	47
Total coluna	111	111	111		
Acurácia Produtor %	53	47	67		55

FONTE: O autor (2016).

O conjunto cj16 continha 222 amostras, pixel de 2,5 metros, considerou todas as variáveis do estudo e três classes de saída. Neste caso, a acurácia total apurada foi de 59%.

TABELA 19 – Matriz de confusão para o conjunto de teste cj16, dados com pixel de 2,5 metros, considerando todas as variáveis do estudo, três classes de saída. Acurácia total 59%.

	DRE	NAS	NDR	Total linha	Acurácia Usuário %
DRE	66	11	26	103	64
NAS	12	54	8	74	73
NDR	33	46	77	156	49
Total coluna	111	111	111		
Acurácia produtor	59	49	69		59

FONTE: O autor (2016).

O conjunto cj17 continha 522 amostras, pixel de 6 metros, considerou todas as variáveis do estudo e duas classes de saída. Neste caso, a acurácia total apurada foi de 70%.

TABELA 20 – Matriz de confusão para o conjunto de teste cj17, dados com pixel de 6 metros, considerando todas as variáveis do estudo, duas classes de saída. Acurácia total 70%.

	DRE	NDR	Total linha	Acurácia Usuário %
DRE	163	56	219	74
NDR	98	205	303	68
Total coluna	261	261		
Acurácia produtor	62	79		70

FONTE: O autor (2016).

O conjunto cj18 continha 522 amostras, pixel de 2,5 metros, considerou todas as variáveis do estudo e duas classes de saída. Neste caso, a acurácia total apurada foi de 74%.

TABELA 21 – Matriz de confusão para o conjunto de teste cj18, dados com pixel de 2,5 metros, considerando todas as variáveis do estudo, duas classes de saída. Acurácia total 74%.

	DRE	NDR	Total linha	Acurácia Usuário %
DRE	178	49	227	78
NDR	83	212	285	72
Total coluna	261	261		
Acurácia produtor	68	81		74

FONTE: O autor (2016).

O conjunto cj19 continha 482 amostras, pixel de 6 metros, considerou todas as variáveis do estudo e duas classes de saída. Neste caso, a acurácia total apurada foi de 72%.

TABELA 22 – Matriz de confusão para o conjunto de teste cj19, dados com pixel de 6 metros, considerando todas as variáveis do estudo, duas classes de saída. Acurácia total 72%.

	DRE	NDR	Total linha	Acurácia Usuário %
DRE	159	49	208	76
NDR	82	192	274	70
Total coluna	241	241		
Acurácia produtor	66	80		72

FONTE: O autor (2016).

O conjunto cj20 continha 482 amostras, pixel de 2,5 metros, considerou todas as variáveis do estudo e duas classes de saída. Neste caso, a acurácia total apurada foi de 77%.

TABELA 23 – Matriz de confusão para o conjunto de teste IV, dados com pixel de 2,5 metros, considerando todas as variáveis do estudo, duas classes de saída. Acurácia total 77%.

	DRE	NDR	Total linha	Acurácia Usuário %
DRE	174	42	216	81
NDR	67	199	266	75
Total coluna	241	241		
Acurácia produtor	72	83		77

FONTE: O autor (2016).

A TABELA 24 apresenta um resumo da fase de testes da RNA, com os resultados obtidos para os diversos conjuntos de teste.

TABELA 24 – Resultados obtidos na fase de testes.

Conjunto	Quant. amostras	Acurácia Produtor (%)			Acurácia Usuário (%)			Acurácia total (%)
		DRE	NAS	NDR	DRE	NAS	NDR	
cj5	300	65	-	71	70	-	67	68
cj6	300	69	-	77	75	-	71	72
cj7	300	27	25	62	39	48	35	38
cj8	300	40	31	65	47	58	40	45
cj9	170	69	-	78	76	-	72	73
cj10	170	74	-	82	81	-	76	78
cj11	170	44	52	59	50	64	45	51
cj12	170	46	49	76	59	66	52	57
cj13	222	59	-	76	71	-	65	67
cj14	222	66	-	88	85	-	72	77
cj15	222	53	47	67	61	65	47	55
cj16	222	59	49	69	64	73	49	59
cj17	522	62	-	79	74	-	68	70
cj18	522	68	-	81	78	-	72	74
cj19	482	66	-	80	76	-	70	72
cj20	482	72	-	83	81	-	75	77

FONTE: O autor (2016).

Na fase de testes, os melhores resultados de acurácia total foram obtidos para os conjuntos cj10 (78%), cj14 (77%) e cj20 (77%). Semelhante ao ocorrido no treinamento, melhores resultados foram conquistados quando se manipularam os dados com pixel de 2,5 metros e se consideraram apenas duas classes de saída, com acurácia total variando entre 67 e 78%.

Estes percentuais de acurácia total ficaram próximos aos encontrados por Poudyal et al. (2010) (78,2%), Lin (2011) (68 a 80%) e Arruda, Dematê e Chagas

(2013) (77%). Por outro lado, nos trabalhos de Pradhan e Lee (2009) (83%), Coneglian, Gomes e Ribeiro (2010) (96,32 a 99,84%), Pradhan e Lee (2010) (83%), Pradhan e Buchroithner (2010) (83,99%) e Pradhan (2011) (82 a 87%), os maiores percentuais de acurácia total superam os alcançados na presente pesquisa.

Vale ressaltar que, embora os citados trabalhos não se refiram explicitamente à extração de rede de drenagem, em todos os casos envolveram a manipulação de variáveis ambientais por RNA, do mesmo modo que ocorreu na presente pesquisa. Não obstante aos objetivos distintos, as características dos dados utilizados, os algoritmos adotados no processamento e as arquiteturas definidas são semelhantes.

É importante considerar que os dados usados em cada um dos trabalhos citados influenciaram os respectivos resultados. Nesta pesquisa, conforme discutido anteriormente, o simples acréscimo ou retirada de variáveis na camada de entrada impactou no resultado. A análise comparativa dos resultados aqui obtidos com aqueles observados em outras pesquisas, sugere que o conjunto de dados disponível constitui fator que impacta no resultado final do processo, assim como atestaram Fernández et al. (2012) e Akram et al. (2012).

Dentre os erros percebidos, alguns deles aconteceram justamente em locais de estradas. Na região geográfica abrangida pela pesquisa, a quase totalidade das estradas possuem revestimento definido como leito natural, ou seja, a superfície de rolamento se apresenta no próprio terreno natural (DSG, 2010) e, em muitos casos, o terreno foi repetidamente escavado, ou simplesmente o traçado foi feito nas áreas mais baixas daquele local.

Este tipo de confusão também pode ocorrer (e frequentemente se observa para esta situação) quando da aplicação de algoritmos de extração automática. Erros semelhantes foram observados com a extração automatizada da drenagem por meio do algoritmo D∞ para a área de estudo, em testes conduzidos nesta pesquisa.

Cabe ressaltar que uma parte da BHRMP é coberta por vegetação que cobrem os canais dos rios, portanto, similar aos apontamentos de Valeriano e Abdon (2007), Grohmann, Riccomini e Steiner (2008) e Tomazoni et al. (2011), alguns erros podem estar relacionados com o uso do SRTM como insumo para a RNA, visto que na aquisição dos dados os sinais de radar são refletidos pelo dossel das árvores em áreas densamente florestadas e não pelo terreno subjacente. Assim como ocorrido

no trabalho de Brubacher et al. (2012), alguns trechos de drenagens podem ter sido ignorados possivelmente pelo efeito dossel em dados SRTM.

Deve-se considerar, ainda, a resolução espacial original do MDE, visto que na área de estudo uma grande quantidade de canais de drenagem possui largura e comprimento inferior a 30 metros. Desta forma, é esperado que alguns destes canais de drenagem não sejam representados pelos dados do MDE, embora sua existência possa ser verificada em campo.

Grande parte dos parâmetros usados como entrada para a RNA foi extraído do MDE, que originalmente possui resolução espacial de 30 metros. Paz e Collischonn (2008) já tinham observado que as drenagens extraídas do MDE sofrem influência do tamanho do pixel, com impactos, sobretudo, na largura e sinuosidade.

A exemplo dos argumentos apresentados por Nourani e Zanardo (2014) e Ribeiro (2015), constatou-se no presente estudo que o aumento do tamanho da célula do MDE conduz a uma suavização do terreno e a consequente redução da declividade da paisagem, o que afeta o funcionamento de algoritmos e técnicas que manipulam estes dados.

A área de estudo possui relevo classificado de plano a muito suavemente ondulado (ADAMY; DANTAS, 2005) e, consequentemente, foi observada a ocorrência dos problemas relatados por Fairfileld e Leymarie (1991), Tarboton (1997), Paz e Collischonn (2008) e Brandão e Santos (2009) quanto ao desempenho dos algoritmos de extração automática.

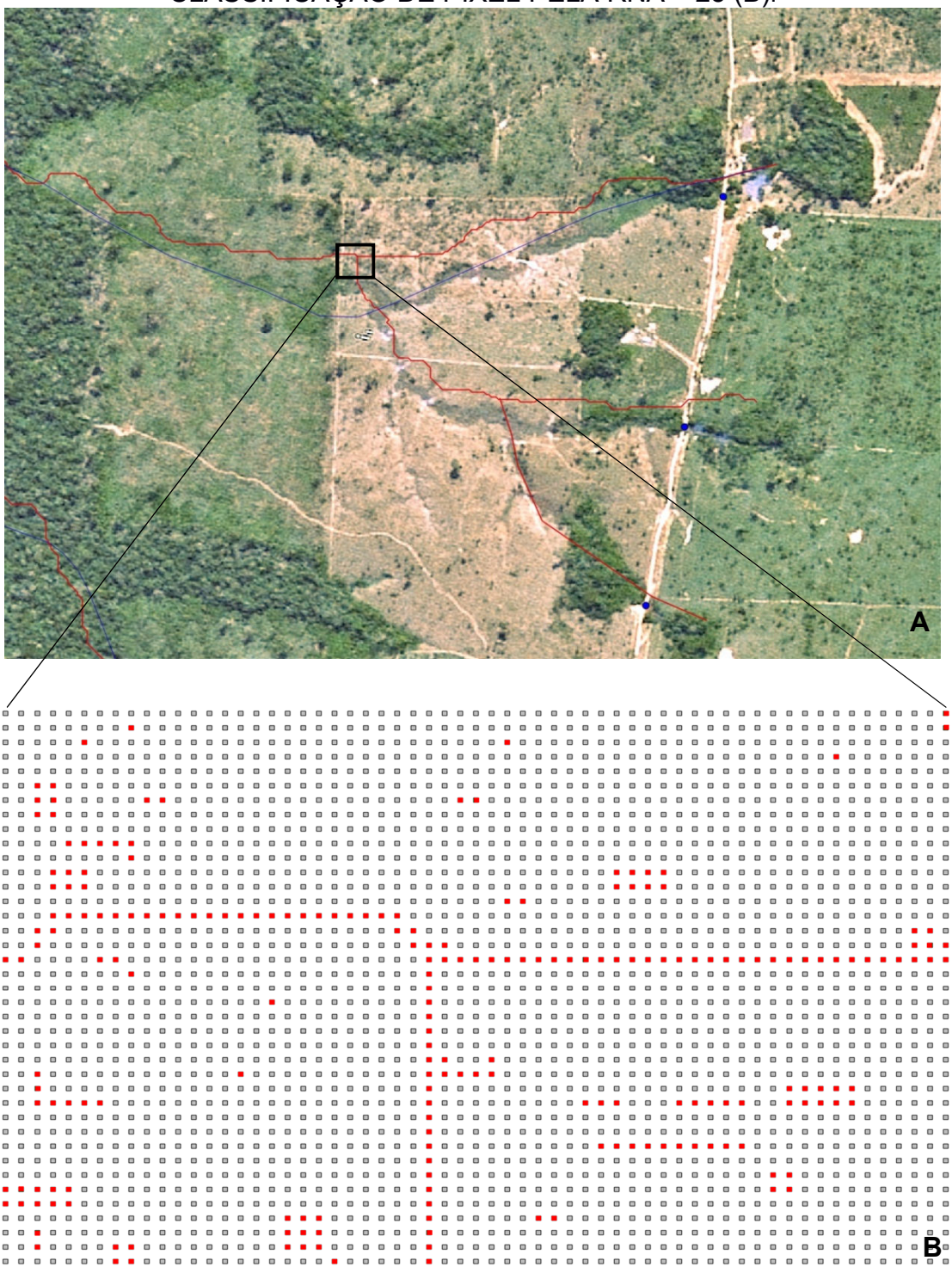
5.5 MAPEAMENTO DA REDE DE DRENAGEM DA BHRMP REALIZADO COM A METODOLOGIA DA PESQUISA

Para obter uma nova rede de drenagem para a BHRMP, foi realizada a generalização para os dados referentes a toda a área de estudo, isto é, todos os pontos da imagem (pixel) foram fornecidos para a RNA, processados e classificados como drenagem ou não drenagem.

A FIGURA 25 apresenta resultados para um recorte da área de estudo. Na FIGURA 25 (A) é demonstrado o mapeamento obtido após o pós-processamento, descrito na seção 4.2.4.4. As linhas em vermelho referem-se à drenagem obtida com a metodologia da pesquisa, e as linhas em azul referem-se à drenagem de referência. Os pontos em azul são amostras de campo. Na FIGURA 25 (B) são

apresentados os pontos da imagem, classificados pela RNA e que não sofreram pós-processamento, para a área destacada na FIGURA 25 (A). Os pontos na cor cinza referem-se a pixel classificado como não drenagem, e os pontos na cor vermelho referem-se a pixel classificado como drenagem.

FIGURA 25 – RESULTADO DO MAPEAMENTO COM PÓS-PROCESSAMENTO PARA UM RECORTE DA ÁREA DE ESTUDO – 25 (A). DETALHE COM A CLASSIFICAÇÃO DE PIXEL PELA RNA – 25 (B).



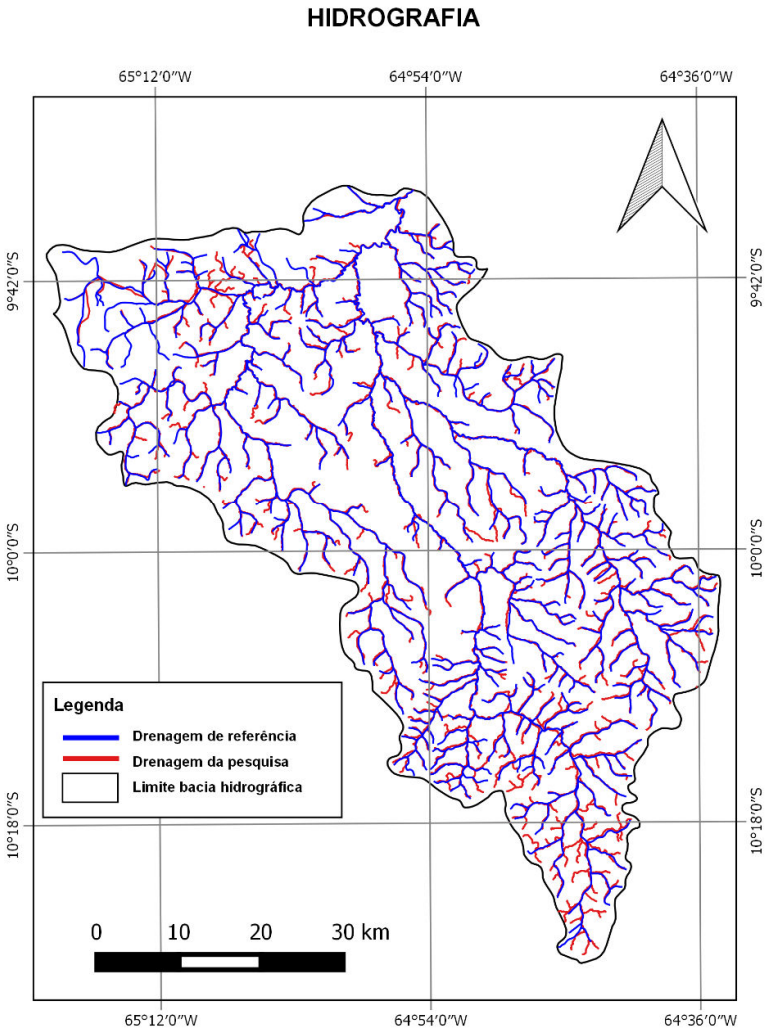
A rede de drenagem gerada com apoio da metodologia proposta contém 893 canais de drenagem e 339 nascentes (327 nascentes mapeadas corretamente se diminuir os erros de excesso). Com base nas imagens SPOT 5 e apoio de levantamentos de campo, foi observado que 163 nascentes não foram mapeadas (erro de omissão) e que 12 nascentes foram mapeadas de forma equivocada (erro de excesso). Quanto à completude, considerando a identificação das nascentes, a taxa de acerto do mapeamento da rede de drenagem ficou em 65,13%. Este resultado é superior àquele observado no mapeamento de referência, que ficou em 62,94%.

$$Taxa\ de\ acerto = \frac{327}{502} X\ 100 = 65{,}13\% \quad (64)$$

$$Excesso = \frac{12}{502} X\ 100 = 2{,}00\% \quad (65)$$

$$Omissão = \frac{163}{502} X\ 100 = 32{,}47\% \quad (66)$$

FIGURA 26 – REDE DE DRENAGEM PARA A BHRMP GERADA COM A METODOLOGIA DA PESQUISA.



FONTE: O autor (2016).

Quanto à acurácia temática, a conferência dos canais de primeira ordem mapeados apontou 110 erros, visto que não se tratavam efetivamente de canais de primeira ordem, mas sim de segunda ou terceira ordem. Portanto, a acurácia do mapeamento quanto a identificação da ordem dos canais (392 canais mapeados corretamente) ficou em 78,08%. Comparativamente a outros mapeamentos realizados, este percentual fica acima do apurado por Sampaio (2008) (64,09%) e abaixo do observado por Souza e Sampaio (2015) (85%).

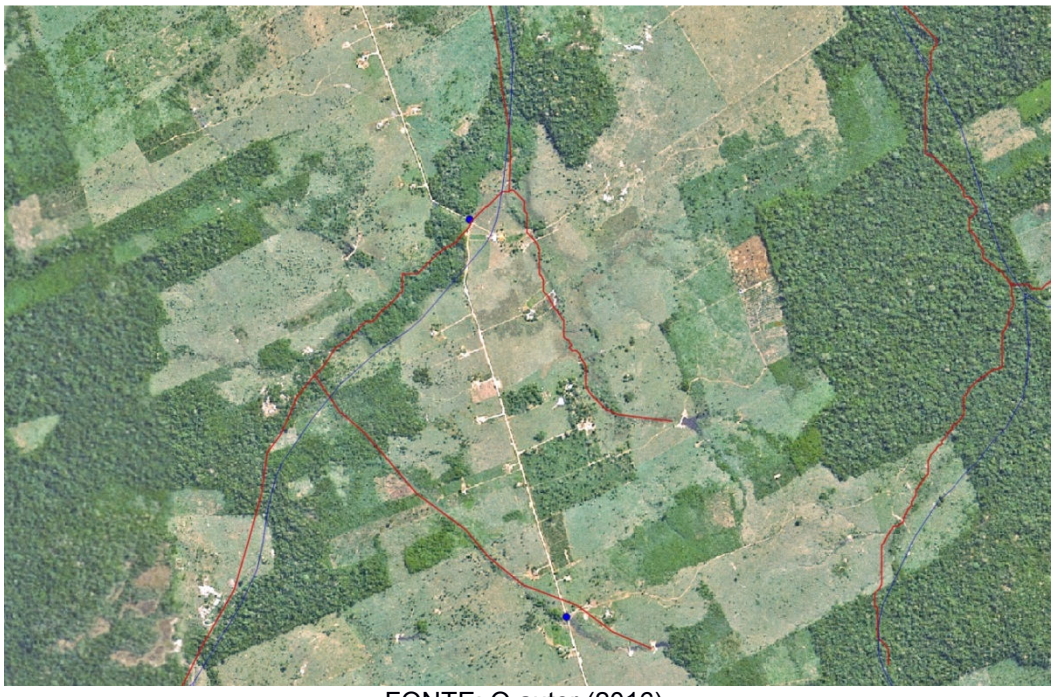
$$Taxa\ de\ acerto\ \frac{392}{502} X\ 100 = 78{,}08 \tag{67}$$

Em comparação aos valores que foram apurados para a base cartográfica da rede de drenagem usada como referência nesta pesquisa, foi possível notar uma melhoria na acurácia. Tal melhoria ocorreu, sobretudo, pela identificação de um maior número canais de primeira ordem. Acredita-se que o uso de parâmetros extraídos da imagem com melhor resolução espacial, e dispostos como atributos de entrada na RNA (bandas da imagem SPOT e índice NDWI), podem ter influenciado positivamente para a melhora.

Este resultado é importante, visto que os métodos de extração automática enfrentam dificuldade justamente na identificação deste tipo de canal, segundo demonstrado na pesquisa de Petsch, Monteiro e Bueno (2012).

Nas figuras 27 e 28 é possível notar exemplos de canais de drenagem de primeira ordem identificados. As linhas na cor azul representam a drenagem de referência e as linhas na cor vermelho representam a drenagem gerada pela pesquisa. Os pontos em azul referem-se a amostras de campo.

FIGURA 27 – DRENAGEM GERADA COM A METODOLOGIA DA PESQUISA (VERMELHO) SOBREPOSTA À IMAGEM SPOT 5.



FONTE: O autor (2016).

FIGURA 28 – DRENAGEM GERADA COM A METODOLOGIA DA PESQUISA (VERMELHO) SOBREPOSTA À IMAGEM SPOT 5.

FONTE: O autor (2016).

Os resultados da pesquisa permitem vislumbrar o uso das saídas da RNA como uma informação valiosa para melhorar os resultados obtidos com os já conhecidos algoritmos para derivação da rede de drenagem, similar às estratégias de uso da heurística defendidas nos trabalhos de Luger e Stubblefild (1988), Russell e Norvig (2004) e Hou et al. (2011).

Estes resultados confirmaram que a metodologia pode ser usada como auxiliar no mapeamento da rede de drenagem, em consonância às proposições de Banon et al. (2013) e de Sampaio e Augustin (2014) que apresentaram alternativas para os métodos tradicionais de extração automática. Em termos práticos, por meio da RNA, o acréscimo dos dados de geologia, geomorfologia, hidrogeologia, solos e imagens com melhor resolução espacial podem melhorar a acurácia do mapeamento de referência.

Não obstante às limitações impostas pelas características dos dados disponíveis, tal qual discutido por Alves Sobrinho et al. (2010), Bosquilia et al. (2013) e Kitchin (2014), deve-se ressaltar que para as regiões onde se detecta a carência de dados (em quantidade e qualidade) é importante a busca por técnicas e métodos que tornem viáveis e potencializem o uso dos dados existentes.

Se considerar as argumentações de Mota, Bueno e Sampaio (2015) quanto ao volume e variedade de dados disponibilizados pela Internet sobre a região

Amazônica, e que tem aumentado cada vez mais, permite-se visualizar a utilização da metodologia proposta inserida no cenário proposto por Chen, Mao e Liu (2014) quanto ao conceito de Big Data.

Mesmo considerando as preocupações de Goodchild (2013) e de Bravo e Sluter (2015) quanto à qualidade dos dados disponíveis, acredita-se que com a automação de algumas etapas, a metodologia empregada nesta pesquisa permitirá adicionar mais variáveis à arquitetura da RNA, podendo aumentar o poder de representatividade da rede e seu potencial para identificar canais de drenagem.

6 CONCLUSÃO

Dados geoespaciais para o estudo da rede de drenagem da BHRMP foram integrados em banco de dados, ficando facilmente acessíveis para manipulação por meio de SGBD, SIG e software de mineração de dados. O banco de dados espaciais foi construído com o propósito específico de contribuir na derivação da rede de drenagem; porém, devido à diversidade dos temas vetoriais e matriciais reunidos é capaz de suportar outros estudos ambientais.

A escolha de um SGBD livre e de código aberto foi essencial para a construção de um banco de dados espaciais aderente a padrões de interoperabilidade, em consonância com os direcionamentos sugeridos pelas iniciativas de criação de infraestruturas para dados espaciais. A arquitetura que sustenta o banco de dados criada para a pesquisa pode facilmente se tornar disponível por meio de *web services*, passando a ficar acessível na Internet ou diretamente por meio de software SIG.

A mineração de dados mostrou-se de muita utilidade quanto à derivação da rede de drenagem, visto que por meio da atividade de classificação foi possível prever a qual classe pertence um pixel da imagem. Esta informação pode ser levada em consideração durante o processo de geração da rede de drenagem, sobretudo naquelas regiões onde os algoritmos de extração automatizada não apresentam bons resultados.

Neste caso, o resultado da classificação, por meio de RNA, poderá ser usado como heurística para orientar o direcionamento do fluxo ou, no mínimo, apontar áreas que definitivamente não correspondem à drenagem. A correta identificação de pontos de drenagem/não drenagem, sobretudo naqueles pontos de indecisão para o algoritmo de derivação, pode contribuir na melhoria da acurácia dos mapeamentos.

A arquitetura da RNA proposta neste trabalho apresentou desempenho satisfatório no processamento das variáveis de entrada, visto que efetivamente conseguiu aprender com base nos padrões oferecidos e foi capaz de generalizar para a área de estudo. Durante o treinamento da RNA foram observados altos percentuais de acertos na identificação das classes das amostras processadas e, posteriormente, nos testes o desempenho da rede também foi aceitável, visto que por vezes foi capaz de classificar novas amostras com acurácia total maior que 67%.

Dentre as diversas configurações testadas, a rede neural artificial que apresentou melhor desempenho reuniu como variáveis de entrada MDE, atributos morfométricos e de direção de fluxo extraídos a partir do MDE, imagem de satélite SPOT 5 e índice de água derivado da imagem SPOT 5. Neste caso, o percentual de acertos sempre ficou acima de 67. Destaca-se, também, a identificação de canais de primeira ordem que não constavam na base cartográfica de referência. A RNA contribuiu, assim, para a identificação de tais trechos, apontando pixels classificados como drenagem.

Uma nova rede de drenagem foi gerada para a área de estudo com diferenças em comparação à base cartográfica de referência, principalmente no que diz respeito aos trechos de primeira ordem. Amostras coletadas em campo possibilitaram a validação dos resultados obtidos em laboratório.

A melhoria da acurácia temática e da completude foi observada, mesmo levando em conta a resolução original do MDE que foi usada para extrair a maior parte dos atributos de entrada da RNA. A área de estudo contém diversos trechos de drenagem com extensão ou largura menor que a resolução espacial do MDE. Os resultados da mineração de dados são dependentes da quantidade e da qualidade dos dados disponíveis.

Os resultados obtidos com o uso da RNA são dependentes, também, dos tamanhos dos conjuntos de dados de treinamento e de teste. A amostragem deve ser suficiente para representar toda a área de estudo, porém na prática nem sempre se dispõe de quantidade suficiente de padrões conhecidos e confirmados para uso na fase de treinamento da rede.

Outra limitação da RNA, observada neste estudo, diz respeito à necessidade de computadores com quantidade apropriada de memória para o processamento dos dados. O hardware exigido para os trabalhos envolveu grande quantidade de memória RAM e processadores de última geração com múltiplos núcleos. A redução do tempo de processamento e a capacidade de suportar processamento de arquivos com grande volume de dados torna a exigência de hardware um fator decisivo para o uso da RNA no apoio à extração de drenagem.

Constatou-se que técnicas de Inteligência Artificial podem ser usadas no processo de extração de redes de drenagem. Particularmente quanto à mineração de dados e RNA, usadas nesta pesquisa, de fato existe potencial para contribuir na melhoria da acurácia do mapeamento.

Cabe destacar o mérito da pesquisa no que se refere à contribuição metodológica no processo de extração da rede de drenagem. Não se trata de substituir os métodos e algoritmos tradicionais de extração automática, mas de acrescentar novas etapas no processo que sejam capazes de contribuir para o entendimento da área e fornecer informações úteis para melhorar o resultado final.

A metodologia proposta permite agregar ampla gama de variáveis, que poderá ser usada como entrada na rede neural artificial e, a partir de então, avaliar sua pertinência em compor o conjunto de atributos que melhor representa determinada região geográfica e que conduzirão melhores resultados na identificação dos canais de drenagem.

A automação de alguns procedimentos poderá facilitar a aplicação da metodologia proposta. Fatores como o tamanho da área geográfica, a quantidade de variáveis e a diversidade dos tipos de dados usados impactam no tamanho dos arquivos que serão manipulados durante o processo. Portanto, as fases de pré e pós processamento são candidatas à automação de procedimentos, que poderá resultar na facilidade de manipulação dos dados por parte dos usuários e na redução do tempo total de trabalho.

REFERÊNCIAS

ADAMY, Amílcar; DANTAS, Marcelo Eduardo. Geomorfologia. In: **Projeto Rio Madeira. Levantamento de informações para subsidiar o estudo de viabilidade do aproveitamento hidrelétrico (AHE) do Rio Madeira. AHE Jirau: relatório final.** Coordenado por Gilmar José Rizzotto e José Guilherme Ferreira de Oliveira, organizado por Marcos Luiz E. S. Quadros, João Marcelo R. de Castro, Antônio Cordeiro, Amílcar Adamy, Homero Reis de Melo Junior e Marcelo Eduardo Dantas. Porto Velho: CPRM – Serviço Geológico do Brasil, 2005.

AGARWAL, Avinash; RAI, R.; UPADHYAY, Alka. Forecasting of runoff and sediment yield using artificial neural networks. **J. Water Resource and Protection**, v. 1, p. 368-375, 2009.

AKRAM, Fatema et al. Automatic delineation of drainage networks and catchments using DEM data and GIS capabilities. In: **Proceedings of the Eighteenth Australasian Fluid Mechanics Conference**, Launceston, Australia, Australasian Fluid Mechanics Society, Hobart, Tasmania, Australia, 2012.

ALDRIDGE, Matthew. Spatial mining techniques and algorithms. In: BERRY, Michael; BROWNE, Murray. **Lecutes notes in Data Mining**. Singapore: World Scientific Publishing, 2006. p. 193-217.

ALVES SOBRINHO, Teodorico et al. Delimitação automática de bacias hidrográficas utilizando dados SRTM. **Engenharia Agrícola**, Jaboticabal, v. 30, n. 1, p. 46-57, 2010.

ANDRADE, Lívia Naiarade. **Redes neurais artificiais aplicadas na identificação automática de áreas cafeeiras em imagens de satélite**. Belo Horizonte, 92 f. Dissertação (Mestrado em Ciências da Computação) – Departamento de Ciências da Computação, Instituto de Ciências Exatas, Universidade Federal de Minas Gerais, Belo Horizonte, 2011.

ANDRIOTTI, José L. S. **Fundamentos de estatística e geoestatística**. São Leopoldo: Ed. Universidade do Vale do Rio dos Sinos, 2003.

_____. Análise de compontentes principais: fundamentos de uma técnica de dados multivariada aplicável a dados geológicos. **Acta Geologia Leopoldensia**, v. XX, n. 44, p. 27-50, 1997.

ANTUNES, A. F.; LINGNAU, C. Uso de índices de acurácia para avaliação de mapas temáticos obtidos por meio de classificação digital. In: Congresso GIS Brasil, 1997, Curitiba. **Anais**... Curitiba: GIS Brasil. 1997. p. 1-15.

APPICE, Annalisa; LANZA, Antonietta; MALERBA, Donato. An integrated platform for spatial Data Mining within a GIS environment. In: **IEEE 23rd International Conference on Data Engineering Workshop**, Istambul, 2007. p. 507-513.

ARRUDA, Gustavo Pais; DEMATTÊ, José Alexandre M.; CHAGAS, César da Silva. Mapeamento digital de solos por Redes Neurais Artificiais com base na relação solo-paisagem. **Rev. Brasileira de Ciências do Solo**, v. 37, p. 27-338, 2013.

BANON, Lise Christine. **Árvores de decisão aplicadas à extração automática de redes de drenagem.** São José dos Campos, 115 f. Dissertação (Mestrado em Computação Aplicada) – Curso de Pós-Graduação em Computação Aplicada, Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2013.

BANON, Lise Christine et al. Definição de critérios a partir da mineração de dados para a extração automática de redes de drenagem. In: **Anais XVI Simpósio Brasileiro de Sensoriamento Remoto – SBSR**, Foz do Iguaçu, Paraná, Brasil, 2013. p. 5753-5760.

BEDARD, Y. Principles of spatial database analysis and design. In: LONGLEY, P. A. et al. **GIS:** principles, techniques, applications & management. 2. ed. Abridged: Wiley, 2005. p. 413-424.

BERHANU, Belete; MELESSE, Assefa M.; SELESHI, Yilma. GIS-based hydrological zones and soil geo-database of Ethiopia. **Catena**, v. 104, p. 21-31, 2013.

BENGIO, Y.; LECUN, Y. Pattern recognition and neural networks. In: ARBIB, M. A. **The handbook of brain theory and neural networks**. Massachusetts: MIT Press, 1995. p. 1-22.

BENÍTEZ, J. M.; CASTRO, J. L.; REQUENA, I. Are artificial neural networks black boxes? **IEEE Transactions on Neural Networks**, v. 8, n. 5, p. 1156-1164, 1997.

BORGES, K. A. V. **Modelagem de dados geográficos –** uma extensão do modelo OMT para aplicações geográficas. Belo Horizonte, 139 f. Dissertação (Mestrado em Administração Pública) – Escola de Governo, Belo Horizonte, 1997.

BOSQUILIA, Raoni Wainer Duarte et al. Comparação entre modelos de mapeamento automático de drenagens utilizando SIG. In: Simpósio Brasileiro de Sensoriamento Remoto, 16, Foz do Iguaçu. **Anais**... Foz do Iguaçu: INPE, 2013. p. 5872-5879.

BRANDÃO, Tayná Freitas; SANTOS, Rosângela Leal. O uso de Imagens SRTM na modelagem de fenômenos hidrológicos (escoamento superficial). In: Simpósio Brasileiro de Sensoriamento Remoto, 14, Natal. **Anais**... Natal: INPE, 2009. p. 4663-4670.

BRASIL. Presidência da República. Casa Civil. Centro Gestor e Operacional do Sistema de Proteção da Amazônia. **Projeto Cartografia da Amazônia –** Documento de Referência. 2008. Disponível em: http://www.sipam.gov.br/dmdocuments/cartografia_versao_final.pdf. Acesso em: 26 fev 2014.

BRAVO, Mezza; SLUTER, Claudia Robbi. O problema da qualidade de dados espaciais na era das informações geográficas voluntárias. **Boletim de Ciências Geodésicas**, v. 21, n. 1, p. 56-73, 2015.

BRUBACHER, João Paulo et al. Avaliação de bases SRTM para extração de variáveis morfométricas e de drenagem. **Geociências**, v. 31, n. 3, p. 381-393, 2012.

CAMBOIM, Silvana Philippi. **Arquitetura para integração de dados interligados aberto à INDE-BR.** Curitiba, 141 f. Tese (Doutorado em Ciências Geodésicas) – Programa de Pós-Graduação em Ciências Geodésicas, Setor de Ciências da Terra, Universidade Federal do Paraná, Curitiba, 2013.

CAMBOIM, Silvana Philippi; SLUTER, Claudia Robbi. Uso de ontologies para busca de dados geoespaciais: uma ferramenta semântica para a infraestrutura nacional de dados espaciais. **Rev. Brasileira de Cartografia**, n. 65/6, p. 1127-1142, 2013.

CARDOSO, Fernando H. B. **Comparativo de classificadores com a plataforma Weka**. Disponível em: www.dsc.ufcg.edu.br/~sampaio/cursos/2008.2/PosGraduacao//dis/FHC.doc. Acesso em: 28 nov 2015.

CASTRO FILHO, Carlos Alberto Pires de; SANTOS, João Roberto dos. Classificação de imagens POLINSAR utilizando técnicas de mineração de dados POLINSAR image classification using data mining technics. **Ambiência Guarapuava (PR)**, v. 6, ed. esp., p. 33-44, 2010.

CHEN, Min; MAO, Shiwen; LIU, Yunhao. Big Data: a survey. **Mobile Netw Appl**, v. 19, p. 171-209, 2014.

CHIKOHORA, Teresa T. A study of the factors considered when choosing an appropriate data mining algorithm. **International Journal of Soft Computing and Engineering (IJSCE)**, v. 4, p. 42-45, 2014.

CHRISTOFOLETTI, Antônio. **Geomorfologia**. São Paulo: Edgard Blucher, 1980.

CHUANLI, Liu; XIAOSHENG, Liu; QIUMIN, Liao. Poyang lake wetland information extraction and change monitoring based on spatial data mining. **Information Technology Journal**, v. 12, p. 6143-6148, 2013.

CIAMPALINI, Andrea et al. Remote sensing as tool for development of landslide databases: the case of the Messina Province (Italy) geodatabase. **Geomorphology**, v. 249, p. 103-118, 2015.

CIMMERY, Vern. **User guide for SAGA (version 2.0.5)**. 2010. Disponível em: http://www.saga-gis.org/en/about/references.html. Acesso em: 24 nov 2015.

COMISSÃO NACIONAL DE CARTOGRAFIA – CONCAR. **Plano de Ação para Implantação da Infraestrutura Nacional de Dados Espaciais**. 2010.

CONEGLIAN, Flavio Marcelo; GOMES, Ingrid Aparecida; RIBEIRO, Selma Regina Aranha. Comparação entre classificadores com rede neural artificial em diferentes áreas de estudo no Paraná. In: III Simpósio Brasileiro de Ciências Geodésicas e Tecnologias da Geoinformação, Recife, 27-30 de julho de 2010. Anais... Recife. 2010.

CONGALTON, Russel G. A review of assessing the accuracy of classifications of remotely sensed data. **Remote Sens. Environ.**, v. 37, p. 35-46, 1991.

CONGALTON, R. G.; GREEN, K. **Assessing the accuracy of remotely sensed data:** principles and practices. New York: Lewis Publishers, 2009.

COLLARES, Eduardo Goulart. **Avaliação de alterações em redes de drenagem de microbacias como subsídio ao zoneamento geoambiental de bacias hidrográficas:** aplicação na bacia hidrográfica do Rio Capivari – SP. São Carlos, 211 f. Tese (Doutorado em Geotecnia) – Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2000.

COSTA-CABRAL, M. C.; BURGES, S. J. Digital Elevation Model Networks (DEMON): a model of flow over hillslopes for computation of contributing and dispersal areas. **Water Resour. Res.**, v. 30, n. 6, p. 1681-1692, 1994.

COUTO, Edivando Vitor do et al. Correlação morfoestrutural da Rede de Drenagem e Lineamentos da Borda Planáltica, Faxinal, Paraná. **Geociências**, v. 30, n. 3, p. 315-326, 2011.

CROMBEZ, K. M. **Comparing flow routing algorithms for digital elevation models.** Digital Terrain Analysis, Project Paper for Michigan State University, p. 1-9, 2008.

CUGLER, Daniel C. et al. Spatial Big Data: platforms, analytics, and science. **Geojournal**, 2013.

CUNICO, Camila; OKA-FIORI, Chisato. Identificação e análise dos condicionantes físicos relevantes à vulnerabilidade ambiental: comparação entre a Serra do Mar e o Primeiro Planalto Paranaense. **Revista Geografar – Resumos do VII Seminário Interno de Pós-Graduação em Geografia**, p. 14-17, 2009.

DATE, C. J. **Introdução a sistemas de bancos de dados**. Rio de Janeiro: Campus, 2004.

Diretoria de Serviço Geográfico – DSG. **Especificações Técnicas para Estruturação de Dados Geoespaciais Vetoriais**. Brasília, 2010.

Diretoria de Serviço Geográfico – DSG. **Especificações Técnicas para Aquisição de Dados Geoespaciais Vetoriais**. Brasília, 2011.

Diretoria de Serviço Geográfico – DSG. **Especificações Técnicas para Produtos de Conjuntos de Geoespaciais**. Brasília, 2014.

Diretoria de Serviço Geográfico – DSG. **Especificações Técnicas para Controle de Qualidade de Dados Geoespaciais. Brasilia, 2016. Produtos de Conjuntos de Geoespaciais**. Brasília, 2016.

ELLIS, Dan; MORGAN, Nelson. Size matters: an empirical study of neural network training for large vocabulary continuous speech recognition. **ICASSP IEEE**, p. 1013-1016, 1999.

ELMASTI, R.; NAVATHE, S. B. **Sistemas de banco de dados**. 6. ed. São Paulo: Pearson, 2011.

EVANS, I. S. An integrated system of terrain analysis and slope mapping. **Zeitschrift fur Geomorphologie**, v. 36, p. 274-295, 1980.

EVANS, Michael et al. Spatial Big Data: case studies on volume, velocity, and variety. In: Big Data KARIMI, Hassan A. **Techniques and Technologies in Geoinformatics**, 2014, p. 2-16.

FACELI, Katti et al. **Inteligência artificial:** uma abordagem de aprendizado de máquina. São Paulo: LTC, 2011.

FAIRFIELD, J.; LEYMARIE, P. Drainage networks from grid digital elevation models. **Water Resources Research**, v. 27, n. 5, p. 709-771, 1991.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to knowledge Discovery in Databases. **AI Magazine**, p. 37-54, 1996.

FERNÁNDEZ, Darcy Carolina Jiménez et al. Extração automática de redes de drenagem a partir modelos digitais de elevação. **Rev. Brasileira de Cartografia**, n. 64/5, p. 619-634, 2012.

FIGUEROA, Rosa L. et al. Predicting sample size required for classification performance. **BMC Medical Informatics and Decision Making**, v. 12, n. 8, p. 1-10, 2012.

FISHER, Adrian; DANAHER, Tim. A Water Index for SPOT 5 HRG Satellite Imagery, New South Wales, Australia, Determined by Linear Discriminant Analysis. **Remote Sens.**, v. 5, p. 5907-5925, 2013.

FLOOD, Neil et al. An operational scheme for deriving standardised surface reflectance from Landsat TM/ETM+ and SPOT HRG Imagery for Eastern Australia. **Remote Sens.**, v. 5, p. 83-109, 2013.

FOODY, Giles M. Sample size determination for image classification accuracy assessment and comparison. In: **Proceeding of the 8th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences**, Shanghai, China, p. 154-162 , 2008.

FRIEDMAN, Jerome H. **Data Mining and Statistics:** what's the connection? Disponível em: http://www-stat.stanford.edu/~jhf/ftp/dm-stat.pdf. Acesso em: 12 out 2010.

FUQIANG, Dai; GANGCAI, Liu. Exploring the determinants of soil and water conservation measures with Data Mining Techniques. In: **International Forum on Computer Science-Technology and Applications**, Chongqing, 2009. p. 380-383.

GANTZ, John; REINSEL, David. Extracting value from chaos. **IDC Iview**, v. 1142, p. 1-12, 2011.

GANDOMI, Amir; HAIDER, Murtaza. Beyond the hype: big data concepts, methods, and analytics. **International Journal of Information Management**, v. 35, p. 137-144, 2015.

GAUTAM, Neha; SANDHU, Parvinder S.; KHULLAR, Sunil. To generate rule for software defect predication on quantitative and qualitative factors using artificial neural networks. In: **Proceedings of International Conference on Intelligent Computational Systems**, ICICS, 2011.

GILBERT, K. et al. Choosing the right data mining technique: classification of methods and intelligent recommendation In: **Proceedings of International Congress on Environmental Modelling and Software Modelling for Environment's Sake**, 2010.

GOEBEL, Michael; GRUENWALD, Le. A survey of data mining and knowledge discovery software tools. **ACM SIGKDD Explorations Newsletter**, v. 1, n. 1, p. 20-33, jun. 1999.

GONÇALVES, Márcio Leandro. **Uma arquitetura neural modular para classificação de imagens multiespectrais de Sensoriamento Remoto**. Campinas, 1997. Dissertação (Mestrado em Engenharia Elétrica) – Curso de Pós-Graduação em Engenharia Elétrica, Universidade Estadual de Campinas, Campinas, 1997.

GONG, J., XIE, J. Extraction of drainage networks from large terrain datasets using high throughout computing. **Computers & Geosciences**, v. 35, n. 2, p. 337-346, 2009.

GOODCHILD, M. F. The quality of big (geo)data. **Dialogs in Human Geography**, v. 3, p. 280-284, 2013.

GOTZ, Markus et al. On scalable data mining techniques for earth science. **Procedia Computer Science**, v. 51, p. 2188-2197, 2015.

GRAHAM, Mark; SHELTON, Taylor. Geography and the future of big data, big data and the future of geography. **Dialogs in Human Geography**, v. 3, p. 255-261, 2013.

GROHMANN, C. H.; RICCOMINI, C.; STEINER, S. S. Aplicações dos modelos de elevação SRTM em geomorfologia. **Rev. Geográfica Acadêmica**, v. 2, n. 2, p. 73-83, 2008.

GRUBER, S.; PECKHAM, S. Land-surface parameters and objects in hydrology. In: HENGL, Tomislav; REUTER, Hannes I. **Geomorphometry:** concepts, software, applications (developments in soil science). v. 33. Elsevier: 2009. p. 171-194.

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

GUTTING, Ralf Hartmut. An introduction to spatil database systems. **VLDB Journal**, v. 3 n. 4, p. 1-33, 1994.

HAMAMOTO, Y. et al. Comparison of pruning algorithms in neural networks. In: HAYASHI, C. **Data Science, Classification, and Related Methods**. Springer, 1988. p. 328-333.

HAN, Jiawei; KAMBER, Micheline. **Data Mining:** concepts and techniques. 2. ed. Morgan Kaufmann Publishers, 2006.

HAND, David J. Statistics and data mining: intersecting disciplines. **ACM SIGKDD Explorations Newsletter**, v. 1, n. 1, p. 16-19, jun. 1999.

HASHEMIAN, M. S.; ABKAR, A. A.; FATEMI, S. B. Study of sampling methods for accuracy assessment of classified remotely sensed data. In: ISPRS Congress, 20., 2004, Istanbul. **Proceedings**... Singapore, ISPRS, 2004.

HEINERT, Michael. Artificial neural networks – how to open the black boxes? In: **AIEG 2008 – First Workshop on Application of Artificial Intelligence in Engineering Geodesy**, p. 1-17, 2008.

HERNÁNDEZ, Luis et al. Artificial neural networks for short-term load forecasting in microgrids environment. **Energy**, v. 75, p. 252-264, 2014.

HOSSEINZADEH, S. R. Drainage network analysis, comparison of digital elevation model (DEM) from ASTER with high resolution satellite image and aerial photographs. **Int. J. Environ. Sci. Dev**, v. 2, p. 194-198, 2011.

HOU, Kun et al. Automatic Extraction of Drainage Networks from DEMs Base on Heuristic Search. **Journal of Software**, v. 6, n. 8, p. 1611-1618, 2011.

HOUSTON, Natalie A. et al. **Geodatabase compilation of hydrogeologic, remote sensing, and water-budget-component data for the high plains aquifer, 2011**. U.S. Geological Survey, 2011.

HUANG, Yue et al. Integrated modeling system for water resources management of Tarim River Basin. **Environmental Engineering Science**, v. 27, n. 3, p. 255-269, 2010.

ISO/TC211. **ISO 19157:2013 Geographic information – Data quality**. 2013.

ISRAEL, Steven. Performance metrics: how and when. **Geocarto International**, v. 21, n. 2, p. 23-32, 2006.

JENSEN, J. R. **Sensoriamento remoto do ambiente:** uma perspectiva em recursos terrestres. 2. ed. São José dos Campos: Parêntese, 2009.

KAPAGERIDIS, I. K. Artificial neural network technology in mining and environmental applications. In: **Proceedings of the 11th International Symposium on Mine Planning and Equipment Selection (MPES)**. VSB Technical University of Ostrava, Prague, 2002.

KIA, Masoud Bakhtyari et al. An artificial neural network model for flood simulation using GIS: Johor River Basin, Malaysia. **Environ Earth Sci**, Springer-Verlag, 2011.

KIM, Y. S.; STREET, W. N.; MENCZER, F. Feature selection in data mining. In: WANG, John. **Data mining:** opportunities and challenges. Hershey: Idea Goup, 2003. p. 80-105.

KITCHIN, R. Big data and human geography: opportunities, challenges and risks. **Dialogues in Human Geography**, v. 3, n. 3, p. 262-267, 2013.

_____. Big Data, new epistemologies and paradigm shifts. **Big Data & Society**, p. 1-12, 2014.

KRASNOPOLSKYA, Vladimir; SCHILLERB, Helmut. Some neural network applications in environmental sciences. Part I: forward and inverse problems in geophysical remote measurements. **Neural Networks**, v. 16, p. 321-334, 2003.

KUMAR, M. Satish; ASADI, S. S.; VUTUKURU, S. S. Integrated Study For Ground Water Quality Analysis Using Remote Sensing And Gis. **International Journal of Applied Chemistry**, v. 12, n. 1, p. 75-86, 2016.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **Biometrics**, v. 33, p. 159-174, 1977.

LAWRENCE, Steve; GILES, C. Lee; TSOI, Chung. **What size neural network gives optimal generalization?** Convergence properties of backpropagation. Maryland: University of Maryland, 1996.

LI, Congcong et al. Comparison of classification algorithms and training sample sizes in urban land classification with Landsat Thematic Mapper Imagery. **Remote Sensing**, v. 6, p. 964-483, 2014.

LI, D. R.; WANG, S. L. **Spatial Data Mining Theories and Applications**. Beijing: Science Press, 2006.

LIMA, Kleber Carvalho; CUNHA, Cenira Maria Lupinacci da. Atualização cartográfica da rede de drenagem para estudo geomorfológico de rios intermitentes e efêmeros do semiárido. **Rev. Brasileira de Cartografia**, n. 66/1, p. 127-136, 2014.

LIN, Jeng-Wen. Neural network model and geographic grouping for risk assessment of debris flow. **International Journal of the Physical Sciences**, v. 6, n. 6, p. 1374-1378, mar. 2011.

LIN, W. et al. Automated suitable drainage network extraction from digital elevation models in Taiwan's upstream watersheds. **Hydrological Processes**, v. 20, p. 289-306, 2006.

LIU, Rui-Juan. A distribute hydrological model integrated with a web-based geographic information system. In: **International Conference on Biomedical Engineering and Biotechnology (iCBEB)**, Macau, May. 28-30, p. 1156-1159, 2012.

LÓPEZ, M. et al. Design scheme for spatial database of climatic and environmental variables in Mexico, integrating Big Data Technology. **Procedia Computer Science**, v. 55, p. 503-513, 2015.

LOPEZ, M. J. G.; CAMARASA, A. M. Use of geomorphological units to improve drainage network extraction from a DEM. Comparison between automated extraction and photointerpretation methods in the Carraixet catchment (Valencia,Spain). **JAGI**, v. 1, n. 3/4, 1999.

LUGER, George F.; STUBBLEFIEL, William. **Artificial intelligence:** structures and strategies for complex problem solving. Berkeley: Addison-Wesley, 1998.

MAIMON, Oded; ROKACH, Lio. **Data mining and knowledge discovery handbook**. Springrer, 2010.

MANOLOPOULOS, Y; PAPADOPOULOS, N. A.; VASSILAKOPOULOS, M. G. **Spatial databases:** technologies, techniques and trends. IGI Global, 2005.

MARQUES, Helder Gustavo et al. Comparação entre os modelos de elevação SRTM, TOPODATA e ASTER na delimitação Automática de rede de drenagem e limite de bacia hidrográfica. In: **Anais XV Simpósio Brasileiro de Sensoriamento Remoto - SBSR**, Curitiba, PR, Brasil, 30 de abril a 05 de maio de 2011, p. 1271 – 1278, 2011.

MCFEETERS, S. K. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. **Remote Sens.**, v. 17, p. 1425-1432, 1996.

MELO JÚNIOR, Homero Reis de. Hidrogeologia. In: **Projeto Rio Madeira. Levantamento de informações para subsidiar o estudo de viabilidade do aproveitamento hidrelétrico (AHE) do Rio Madeira. AHE Jirau: relatório final.** Coordenado por Gilmar José Rizzotto e José Guilherme Ferreira de Oliveira, organizado por Marcos Luiz E. S. Quadros, João Marcelo R. de Castro, Antônio Cordeiro, Amílcar Adamy, Homero Reis de Melo Junior e Marcelo Eduardo Dantas. - Porto Velho: CPRM – Serviço Geológico do Brasil, 2005.

MEMARIAN, Hadi; BALASUNDRAM, Siva Kumar. Comparison between Multi-Layer Perceptron and Radial Basis Function Networks for Sediment Load Estimation in a Tropical Watershed. **Journal of Water Resource and Protection**, v. 4, p. 870-876, 2012.

MENDES, David; MARENGO, José A. Temporal downscaling: a comparison between artificial neural network and autocorrelation techniques over the Amazon Basin in present and future climate change scenarios. **Theor Appl Climatol.**, v. 100, p. 413-421, 2009.

MIAH, Muhammed. Survey of data mining methods in emergency evacuation planning. **Proceeding of Conference for Information Systems Applied Research**, Wilmington North Carolina, v. 4, n. 1815, p. 1-11, 2011.

MOHD, Maina Mamodu et al. Application of web geospatial decision support system for Tanjung Karang rice precision irrigation water management. In: **International Conference on Agricultura, Food and Environmental Engineering (ICAFEE' 2014)**, Kuala Lampur, Malaysia, Jan. 15-16, p. 24-28, 2014.

MONICO, J. F. G. et al. Acurácia e precisão: revendo os conceitos de forma acurada. **Boletim de Ciências Geodésicas**, v. 15 n. 3, p. 469-483, 2009.

MOORE, I. D.; GRAYSON, R. B.; LADSON, A. R. Digital terrain modeling: a review of Hydrological, geomorphological an biological applications. **Hydrological Processes**, v. 5, p. 3-30, 1991.

MOTA, Alex dos Santos; BUENO, Luis Fernando; SAMPAIO, Tony Vinicius Moreira. Dados e informações geoespaciais para análise territorial e ambiental na Amazônia Legal no Brasil. **Rev. Geográfica Venezolana**, v. 56, n. 2, p. 249-267, 2015.

NASA. **Shuttle Radar Topography Mission**. Disponível em: http:// http://www2.jpl.nasa.gov/srtm/. Acesso em: 03 dez 2013.

NOURANI, V.; ZANARDO, S. Wavelet-based regularization of the extracted topographic index from high-resolution topography for hydro-geomorphic applications. **Hydrological Processes**, v. 28, n. 3, p. 1345-1357, 2014.

NOVO, E. M. L. de M. **Sensoriamento remoto:** princípios e aplicações. São José dos Campos: Edgar Blücher Ltda., 1989.

O'CALLAGHAN, J. F.; MARK, D. M. The extraction of drainage networks from digital elevation data. **Computer Vision, Graphics and Image Processing**, v. 28, p. 323-344, 1984.

OGC. **Simple Feature Access - Part 2: SQL option**, 2010. Disponível em: http://www.opengeospatial.org/standards/sfa. Acesso em: 15 nov 2014.

OLAYA, V.; CONRAD, O. Geomorphometry in saga. In: HARTEMINK, A. E.; MCBRATNEY, A. B. **Developments in soil science.** 2009. p. 293-308.

OLDEN, Julian D.; JACKSON, Donald A. Illuminating the “black box”: a randomization approach for understanding variable contributions in artificial neural networks. **Ecological Modelling**, v. 154, p. 135-150, 2002.

OLIVEIRA, G. G.; GUASSELLI, L. A.; SALDANHA, D. L. Avaliação da qualidade da drenagem extraída a partir de dados SRTM. In: **Anais do Simpósio Brasileiro de Recursos Hídricos**, Campo Grande. p. 2745-2751, 2009.

OZKOSE, Hakan et al. Yesterday, today and tomorrow of big data. **Procedia – Social and Behavioral Sciences**, v. 195, p. 1042-1050, 2015.

PANCHAL, Gaurang et al. Behaviour analysis of multilayer perceptrons with multiple hidden neurons and hidden layers. **Proceedings of International Journal of Computer Theory and Engineering**, v. 3, n. 2, p. 332-337, 2011.

PAIDI, A. N. Data mining: future trends and applications. **International Journal of Modern Engineering Research (IJMER)**, v. 2, p. 4657-4663, 2012.

PAPARRIZOS, Spyridon et al. Spatial Data Infrastrucutres (SDIS) in Greece: an assessment of geoportals for carrying out hydrological projects. **International Water Technology Journal**, v. 4, n. 4, p. 222-232, 2014.

PAZ, A. R.; COLLISCHONN, W. Derivação de rede de drenagem a partir de dados do SRTM. **Rev. Geog. Acadêmica**, v. 2, n. 2, p. 84-95, 2008.

PELLETIER, Jon D. A robust, two-parameter method for the extraction of drainage networks from high-resolution digital elevation models (DEMs): evaluation using synthetic and real-world DEMs. **Water Resources Research**, v. 49, p. 1-15, 2013.

PETSCH, Carina; MONTEIRO, Jéssica Barion; BUENO, Marina Brandt. Análise comparativa da acuricidade da rede de drenagem gerada automaticamente e extraída de carta topográfica: estudo de caso no Município de Ponta Grossa – PR. **Revista Geonorte**, v. 2, n. 4, p.1195-1205, 2012.

POUDYAL, Chandra et al. Landslide susceptibility maps comparing frequency ratio and artificial neural networks: a case study from the Nepal Himalaya. **Environ Earth Sci.**, v. 61, p. 1049-1064, 2010.

PRADHAN, Biswajeet; BUCHROITHNER, Manfred F. Comparison and validation of landslide susceptibility maps using an artificial neural network model for three test areas in Malaysia. **Environmental e Engineering Geoscience**, v. XVI, n. 2, p. 107-126, 2010.

PRADHAN, Biswajeet; LEE, Saro. Regional landslide susceptibility analysis using back-propagation neural network model at Cameron Highland, Malaysia. **Landslides**, v. 7, p. 13-30, 2010.

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

PRADHAN, Biswajeet et al. Application of a data mining model fo landslide hazard mapping. **The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences**, v XXXVII, part B8, p. 187-196, 2008.

PRADHAN, Biswajeet. An assessment of the use of an advanced neural network model with five different training strategies for the preparation of landslide susceptibility maps. **Journal of Data Science**, v. 9, p. 65-81, 2011.

QUINN, P. et al. The prediction of hillslope flow paths for distributed hydrological modeling using digital terrain models. **Hydrological Processes**, n. 5, p. 59-80, 1991.

RENCHER, A. C. **Methods of multivariate analysis**. 2. ed. Nova York: John Wiley & Sons, 2002.

REYYA, Suchitra; SUMALLIKA, T.; VASUKI, G. B. M. An approach to store Spatial Big-Data using multi-valued database. **International Journal of Research in Engineering & Technology**, v. 1, p. 15-22, 2013.

RIBEIRO, Hugo José. **Análise da consistência de dados hidrológicos a partir de diferentes Modelos Digitais de Terreno**. Goiânia: UFG, 2015. Dissertação (Mestrado em Engenharia do Meio Ambiente) – Escola de Engenharia Civil, Universidade Federal de Goiás, Goiânia, 2015.

RILEY, S. J.; DE GLORIA, S. D.; ELLIOT, R. A terrain ruggedness that quantifies topographic heterogeneity. **Intermountain Journal of Science**, v. 5, n. 1-4, p. 23-27, 1999.

RIZZOTTO, G. J. **Projeto Rio Madeira. Levantamento de informações para subsidiar o estudo de viabilidade do aproveitamento hidrelétrico (AHE) do Rio Madeira. AHE Jirau:** relatório final. Coordenado por Gilmar José Rizzotto e José Guilherme Ferreira de Oliveira, organizado por Marcos Luiz E. S. Quadros, João Marcelo R. de Castro, Antônio Cordeiro, Amílcar Adamy, Homero Reis de Melo Junior e Marcelo Eduardo Dantas. - Porto Velho: CPRM – Serviço Geológico do Brasil, 2005.

ROCHA, Claudia Lucena. **Análise de fronteiras de reservatório de petróleo através de geoquímica de superfície e mineração de dados**. Rio de Janeiro: UFRJ, 2005. Dissertação (Mestrado em Engenharia Civil) – Programa de Pós-Graduação em Engenharia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2005.

RONDÔNIA. Plano Agroflorestal de Rondônia. **Zoneamento socioeconômico-ecológico do Estado de Rondônia**. Porto Velho, 2002.

ROSS, Jurandyr. **Ecogeografia do Brasil:** subsídios para planejamento ambiental. Sao Paulo: Oficina de Textos, 2006.

RUSSEL, Stuart; NORVIG, Peter. **Inteligência artificial**. 2. ed. Rio de Janeiro: Campus, 2004.

RUSSOM, Philip. **Managing Big Data**. TDI Research, 2003.

SAMPAIO, Tony Vinicius Moreira. **Parâmetros morfométricos para melhoria da acurácia do mapeamento da rede de drenagem uma proposta baseada na análise da Bacia Hidrográfica do Rio Benevente – ES**. Belo Horizonte: UFMG, 2008. Tese (Doutorado em Geografia) – Programa de Pós-Graduação em Geografia, Universidade Federal de Minas Gerais, Belo Horizonte, 2008.

SAMPAIO, T. V. M.; AUGUSTIN, C. H. R. R. Índice de concentração da rugosidade: uma nova proposta metodológica para o mapeamento e quantificação da dissecação do relevo como subsídio a cartografia geomorfológica. **Rev. Brasileira de Geomorfologia**, v. 15, p. 47-60, 2014.

SCHEIDT, Felippe Alex; BRUNETTO, Maria Angelica de Camargo. Modelagem chuva-vazão utilizando redes neurais srtificiais e slgoritmos genéticos. In: Congresso da Sociedade Brasileira de Computaçao, 31, 2011, Natal. **Anais**... Natal: SBC, 2011.

SELBY, M. J. **Earth's changing surface:** an introduction to geomorphology. Oxford: Clarendon Press, 1985.

SHEKHAR, S. et al. Trends in spatial data mining. In: Kargupta, H. et al. (eds.) **Data mining:** next generation challenges and future directions. Menlo Park: AAAI Press, 2004. p. 357-380.

SHI, Guang-ren; YANG, Xin-She. Optimization and data mining for fracture prediction in geosciences. **Procedia Computer Science**, v. 1, p. 1359-1366, 2012.

SHRESTHA, Rajesh Raj; THEOBALD, Stephan; NESTMANN, Franz. Simulation of flood flow in a river system using artificial neural networks. **Hydrology and Earth System Sciences**, v. 9, p. 313-321, 2005.

SILBERSCHATZ, A.; KORTH, H. F.; SUDARSHAN. Sistema de banco de dados. 6. ed. Rio de Janeiro: Elsevier, 2012.

SILVA, J. X. **Geoprocessamento:** para a análise ambiental. Rio de Janeiro: Edição do Autor, 2001.

SILVA, Roberto Valmir da; KOBIYAMA, Masato. Delineamento automático da rede de drenagem em bacias hidrográficas com ênfase em trechos de zero ordem. In: **IAHR AIPH XXI Congresso Latinoamericano de Hidráulica**, São Pedro, São Paulo, Brasil, out. 2004.

SILVEIRA, Claudinei Taborda da. **Análise digital do relevo na predição de unidades preliminares de mapeamento de solos**: integração de atributos topográficos em Sistemas de Informações Geográficas e Redes Neurais Artificiais.

Curitiba: UFPR, 2010. Tese (Doutorado em Geografia) – Programa de Pós-Graduação em Geografia, Universidade Federal do Paraná, Curitiba, 2010.

SINGH, Gurjeet; PANDA, Rabrindra K. Daily sediment yield modeling with artificial neural network using 10-fold cross validation method: a small agricultural watershed, Kapgari, India. **International Journal of Earth Sciences and Engineering**, v. 4, n. 6, p. 443-450, oct. 2011.

SIRTOLI, Angelo Evaristo. **Mapeamento de solos com auxílio da geologia, atributos do terreno e índices espectrais integrados por Redes Neurais Artificiais**. Curitiba: UFPR, 2008. Tese (Doutorado em Geologia) – Pós-Graduação em Geologia – Área de Concentração Geologia Ambiental, Setor de Ciências da Terra, Universidade Federal do Paraná, Curitiba, 2008.

SIRTOLI, Angelo Evaristo et al. Pedometria apoiada em atributos topográficos, índices espectrais e geologia com uso de Redes Neurais Artificiais. **Geociências**, v. 32, n. 3, p. 516-531, 2013.

SOUSA, Benilson Pereira; SOUZA FILHA, Maria Alves de; PEREIRA, Jonas Sousa. Levantamento e quantificação das áreas de preservação permanentes na Área de Proteção Ambiental (APA) das Nascentes de Araguaína a partir de dados de radar interferométrico. **Rev. Tocantinente de Geografia**, ano 3, n. 1, p. 35-47, 2014.

SOUZA, Acácia Maria Barros de; CRUZ, Marcus Aurélio Souza; ARAGÃO, Ricardo. Análise comparativa do uso de modelos digitais de elevação para a caracterização física da bacia do rio Japaratuba. In: **Anais do XIX Simpósio Brasileiro de Recursos Hídricos**, Maceió, p. 1-15, 2011.

SOUZA, Mayara Soares de; SAMPAIO, Tony Vinicius Moreira. Avaliação da acurácia de bases cartográficas: um estudo de caso da rede de drenagem do estado do Paraná na escala 1:50.000 para a carta MI 2818-4. In: Simpósio Brasileiro de Sensoriamento Remoto, 17, 2015, João Pessoa. **Anais**... João Pessoa: INPE, 2015. p. 1713-1719.

SPORL, Christiane; CASTRO, Emiliano; LUCHIARI, Aílton. Aplicação de redes neurais artificiais na construção de modelos de fragilidade ambiental. **Rev. Departamento de Geografia – USP**, v. 21, 2011.

SRIVASTAVA, Nitish et al. Dropout: a simple way to prevent neural networks from overfitting. **Journal of Machine Learning Research**, v. 15, p.1929-1958, 2014.

STOLZE, K. **SQL/MM spatial:** the standard to manage spatial data in relational database systems. Leipzig: BWT 2003, 2003.

STROBL, R. O.; FORTE, F. Artificial neural network exploration of the influential factors in drainage network derivation. **Hydrological Processes**, v. 21, p. 2965-2978, 2007.

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

SUG, Hyontai. The effect of training set size for the performance of neural networks of classification **WSEAS Transactions on Computers**, p. 1297-1306, 2010.

TARBOTON, D. A new method for the determination of flow directions and upslope areas in grid digital elevation models. **Water Resources Research**, v. 33, n. 2, p. 309-319, 1997.

TRIBE, A. Automated recognition of valley lines and drainage networks from grid digital elevation models: a review and a new method. **Jounal Hydrology**, v. 139, p. 263-293, 1992.

TRICHAKIS, I. C.; NIKOLOS, I. K.; KARATZAZ, G. P. Artificial neural network (ANN) based modeling for karstic groundwater level simulation. **Water resources management**, v. 25, p. 1143-1152, 2011.

TOMAZONI, Julio Caetano et al. Uso de modelo digital de elevação gerados a partir do ASTER GDEM e SRTM para caracterização de rede de drenagem. **Rev. Brasileira de Geografia Física**, v. 2, p. 365-376, 2011.

TURBAN, Efrain et al. **Business intelligence:** a managerial approach. 2. ed. Prentice Hall, 2010.

VALERIANO, M. M.; ABDON, M. M. Aplicação de dados SRTM a estudos do Pantanal. RBC. **Rev. Brasileira de Cartografia**, v. 59, p. 63-71, 2007.

VARELLA, Carlos Alberto. **Análise de componentes principais**. Seropédica: UFRRJ, 2008.

VERCELLIS, Carlo. **Business intelligence**. Wiley, 2009.

VICINI, L. **Análise multivariada da teoria à prática**. Santa Maria: 2005. Monografia – Universidade Federal de Santa Maria, 2005.

VITOLO, Claudia et al. Web technologies for environmental Big Data. **Environmental Modelling & Software**, v. 63, p. 185-198, 2015.

VOGT, J. et al. Deriving drainage networks and catchment boundaries: a new methodology combining digital elevation data and environmental characteristics. **Geomorphology**, v. 53, p. 281-289, 2003.

WANG, L.; LIU, H. An efficient method for identifying and filling surface depressions in digital elevation models for hydrologic analysis and modelling. **International Journal of Geographical Information Science**, v. 20, n. 2, p. 193-213, 2006.

WEBER, Eliseu et al. **Qualidade de dados geoespaciais**. Relatório de Pesquisa. Curso de Pós-Graduação em Ciência da Computação, Instituto de Informática, Universidade Federal do Rio Grande do Sul, janeiro de 1999.

WILSON, John P.; LAM, Christine S.; DENG, Yongxin. Comparison of the performance of flow-routing algorithms used in GIS-based hydrologic analysis. **Hidrological Processes**, n. 21, p. 1026-1044, 2007.

XU, Li; QI, Qingwen; WANG, Zhongyuan. Regional ecological environment spatial data mining based on GIS and RS: a case study in Zhangjiajie, China. In: **Fifth International Conference on Fuzzy Systems and Knowledge Discovery**, Jinan Shandong, 2008. p. 266-270.

WOOD, J. **The geomorphological characterisation of digital elevation models**. Phd Thesis, University of Leicester, 1996. Disponível em: http://www.soi.city.ac.uk/~jwo/phd/. Acesso em: 17 dez 2015.

WU, Xindong et al. Data mining with big data. **IEEE Transactions on Knowledge and Data Engineering**, v. 26, n. 1, p. 97-107, 2014.

YEUNG, Albert K. W.; HALL, G. Brent. **Spatial database systems**. Springer, 2007.

ZHAN, Yunjun; YANG, Haiwei. Extracting hazardous geology information based on spatial data mining. In: **Second International Conference on Intelligent Networks and Intelligent Systems**, Tianjian, 2009. p. 514-7.

ZHANG, Ling; GUILBERT, Eric. A study of variables characterizing drainage patterns in river netwoks. In: **International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences**, Melbourne, Australia, v. XXIX-B2, 2012.

ZHANG, Ji; HSU, Wynne; LEE, Mong Li. Image mining: issues, frameworks and techniques. In: **2nd ACM SIGKDD International Workshop on Multimedia Data Mining (MDM/KDD'01)**, San Francisco, CA, USA, 26 aug 2001.

ZHU, Wenju et al. The evaluation system design of GIS-Based oil and gas resources carbon emission database management. In: **35th International Symposium on Remote Sensing of Environment (ISRSE35)**, p. 1-6, 2014.

ZEVENBERGEN, Lyle W.; THORNE, Colin R. Quantitative analysis of land surface topography. **Earth Surface Process and Landforms**, v. 2, p. 47-56, 1987.