Revista Brasileira de Recursos Hídricos
Versão On-line ISSN 2318-0331
RBRH vol. 20 nº.3 Porto Alegre jul./set. 2015 p. 795 - 808

Aplicação de Árvores de Decisão na Modelagem das Concentrações de Nitrato e Fósforo Total: Estudo de Caso no Rio das Velhas

Using Decision Trees to Model Nitrate and Phosphorous Concentrations: A Case-Study in the Velhas River Watershed

Lilia Maria de Oliveira[1] e Philippe Maillard [2]

[1] Departamento de Geografia- IGC, Universidade Federal de Minas Gerais/CEFETMG, MG, Brasil

lilmaria42@gmail.com

[2] Departamento de Geografia - IGC, Universidade Federal de Minas Gerais, MG, Brasil

philippermaillard@yahoo.com.br

Recebido: 27/11/14 - Revisado: 21/01/15 - Aceito: 24/02/15

RESUMO

O uso e ocupação do solo da bacia hidrográfica são responsáveis pela perda ou manutenção da qualidade de suas águas superficiais, especialmente quando tratamos de fontes de poluição difusa. A árvore de decisão é uma ferramenta de aprendizagem de máquina que tem se mostrado promissora para este tipo de estudo. O modelo proposto pela técnica segue uma estrutura hierárquica na busca de descobrir estruturas ocultas em conjuntos de dados complexos e prever o comportamento de uma classe (variável dependente) a partir do conjunto de variáveis independentes. A metodologia aplicada neste trabalho avalia o uso desta técnica no estudo da relação vazão - uso do solo - poluição difusa. A Bacia do Rio das Velhas e os parâmetros de qualidade nitrato e fósforo total foram escolhidos como objeto de investigação. Foram utilizados dados de qualidade de 17 estações de monitoramento e de 8 estações fluviométricas de vazão implantadas na calha principal do Rio das Velhas. O uso do solo, sub-dividido em 8 categorias, foi obtido a partir de um mosaico de 72 imagens RapidEye. As árvores de decisão foram geradas utilizando o algoritmo C4.5 de Quinlan que gerou regras para associar a vazão e as categorias de uso do solo aos parâmetros de qualidade, buscando identificar as categorias de uso do solo responsáveis pela alteração da qualidade das águas superficiais. As árvores de decisão propostas atingiram eficiência de classificação acima de 80% para as duas variáveis avaliadas nitrato e fósforo total. Entretanto, os estudos também apontaram fragilidades na metodologia associadas ao baixo volume de informações existentes para a Bacia do Rio das Velhas

Palavras Chave: *Aprendizagem de máquina. Algoritmo C4.5. Nitrato. Fósforo. Poluição das águas*

ABSTRACT

Land use in a watershed can be responsible for the loss or preservation of the quality of its superficial waters. This especially true for non-point water pollution. Decision trees are machine learning techniques with promising applications in this type of research. A decision tree model follows a hierarchical structure that aims at finding hidden patterns within complex databases and predict the behavior of a class (explained variable) from the independent variables. The methodology used in this article evaluates the applicability of decision trees to study the relationship water discharge – land use – non-point pollution. The Velhas River watershed and the water quality parameters nitrate and phosphorous were chosen as the objects of investigation. Water quality data from 17 stations and discharge data from 8 fluviometric stations, all within the main river channel of the Velhas River were used to carry out our study. Eight categories of land use data were extracted from mosaic of 72 RapidEye images. The decision trees were generated using Quinlan's C4.5 algorithm which created sets of rules associating land use categories to variations of nitrate and phosphorous therefore identifying which ones are the main responsible for loss of water quality. The resulting decision trees were able to correctly classify over 80% of the instances of discrete nitrate and phosphorous classes. However, some shortcomings were also identified and were mainly attributed to the relatively low volume of water quality data.

Keywords: *Machine learning. C4.5 algorithm. Nitrate. Phosphorous. Water pollution*

INTRODUÇÃO

A alteração na qualidade das águas superficiais é denominada poluição e tem como origem fontes poluidoras pontuais e/ou difusas. As fontes pontuais são caracterizadas por despejos que apresentam um único ponto de lançamento no curso de água, por exemplo efluentes domésticos e industriais. As fontes poluidoras difusas são aquelas distribuídas sobre a superfície da bacia, podendo atingir o curso de água em qualquer ponto tendo o escoamento superficial como veículo de transporte de poluentes. A agricultura e a pecuária são atividades que contribuem com a produção de poluição difusa para os cursos de água.

Com o objetivo de avaliar a relação entre o uso do solo e a qualidade das águas superficiais, muitos estudos já foram realizados (CLELAND, 2003; GORSEVSKI et al., 2008; MAILLARD; SANTOS, 2008; PINHEIRO et al., 2014; SHARIFI; HOSSEINI, 2011). Nesses, o uso do solo é normalmente apresentado como o percentual de área ocupada por categoria de uso e é considerado fonte difusa ou uma forma de controlar a poluição difusa para o curso de água.

A seleção da área de influência do uso do solo, normalmente segue duas linhas distintas:

1) a bacia de contribuição ou

2) faixas de largura variada ao longo do curso de água.

A qualidade das águas por sua vez é avaliada pelos parâmetros físicos, químicos e biológicos obtidos através de monitoramento sistemático (GORSEVSKI et al., 2008; IGAM, 2013; MAILLARD; SANTOS, 2008; SHARIFI; HOSSEINI, 2011).

A avaliação da poluição difusa a partir de modelos matemáticos pode ser realizada por meio do uso de diversos softwares disponíveis e aplicações como demonstrado em Huang e Hong (2010), Sharifi e Hosseini (2011) e Zhenyao et al. (2012). A maior dificuldade na utilização de modelos matemáticos se apresenta em função da baixa disponibilidade de dados (SHARIFI; HOSSEINI, 2011).

Uma outra abordagem bastante comum é o uso de regressão múltipla entre qualidade de água, representada pelos parâmetros de monitoramento (ex. turbidez, nitrato, fósforo, DBO, OD, entre outros), e as categorias de uso do solo (HUANG; HONG, 2010; MAILLARD; SANTOS, 2008; SHARIFI; HOSSEINI, 2011).

Recentemente, tem-se utilizado árvores de decisão para obter regras que associem variáveis independentes (ex. uso do solo) a dependentes (ex. parâmetro de qualidade de água). Apesar do seu uso recente em Recursos Hídricos, a mesma vem apresentando resultados promissores na obtenção de regras associando chuva a vazão, nível de água a vazão e uso do solo a qualidade das águas (BHATTACHARYA; SOLOMATINE, 2005; GRUNWALD et al., 2009; LIAO; SUN, 2010; SOLOMATINE; DULAL, 2003).

Árvore de decisão é uma técnica de aprendizado de máquina que, a partir de um conjunto de instâncias (dados de treinamento), compostas por dois ou mais atributos (variáveis independentes e dependente), extrai regras sobre um subconjunto das variáveis independentes, usando, geralmente, uma medida de ganho de informação (BASGALUPP, 2010; CHAU, 2006), para inferir a variável dependente.

A árvore de decisão é construída de forma hierárquica (Figura 1), onde seleciona-se, a partir de um conjunto de instâncias, o atributo (a, b, c) com o maior ganho de informação para dividir esse conjunto de instâncias. A cada divisão seleciona-se novamente um novo atributo para dividi-la até que um critério de poda seja atingido. Ao final deste processo obtém-se regras que associam as variáveis independentes (a, b, c) a variável dependente (M).

Considerando a importância de entender melhor a relação entre o uso do solo e a qualidade das águas, para fontes difusas, este trabalho tem como objetivo avaliar o potencial de árvores de decisão, por meio do uso do algoritmo C4.5 (QUINLAN, 1986), para modelar a relação entre as concentrações de nitrato e fósforo total, vazão e o uso do solo na bacia do rio das Velhas.

Desta forma, avaliamos através de regras, produzidas por árvores de decisão, a associação entre os parâmetros nitrato e fósforo (variáveis dependentes) e uso do solo e vazão (variáveis independentes). As regras geradas dividiram as instâncias, de forma à relacionar uso do solo e vazão a concentração de nitrato e fósforo. A eficiência de classificação, medida pelo número de instâncias classificadas corretamente, foi superior 80%.

Princípios de árvore de decisão

Árvores de decisão utilizam conceitos de inteligência artificial (IA) e de estatística para aprenderem a partir de um conjunto de dados. O conjunto dos dados é constituído de uma série de subconjuntos, definidos por atributos ou características. Os atributos são as variáveis dependentes e independentes, com domínio contínuo, discreto ou categórico. Os atributos são definidos pelos parâmetros de qualidade de água (dependente), de vazão e de uso do solo (ambos independentes). A variável dependente é geralmente denominada de "classe", e tem seu valor determinado a partir das variáveis independentes (BASGALUPP, 2010).

As instâncias contidas nesse conjunto de dados servem de exemplos de treinamento. A partir destes dados realizam-se inferências indutivas que conduzam a condições verdadeiras ou falsas na forma: se (condição) então (ação).

O aprendizado indutivo pode ser dividido em supervisionado e não supervisionado. Quando se conhece o valor da classe, pertencente ao conjunto de exemplos de treinamento, o processo é chamado de indutivo supervisionado. Se o domínio da classe é discreto o problema é de classificação e se é numérico o problema passa a ter solução por regressão ou aproximação de funções.

A obtenção de árvores de decisão por classificação ou regressão, se baseia em particionamento recursivo binário, uma vez que o "nó" pai (ai) é dividido em dois "nós" filhos (bi) (Figura 1) recursivamente, por que o processo pode ser repetido por tratamento de cada "nó" filho como um "nó" pai, até que se atinjam as folhas, que indicam as classes (Mj) (GRUNWALD et al., 2009).

Um mesmo conjunto de atributos pode gerar diversas estruturas de árvore de decisão. O número de árvores de decisão possíveis cresce fatorialmente `a medida que se aumenta o

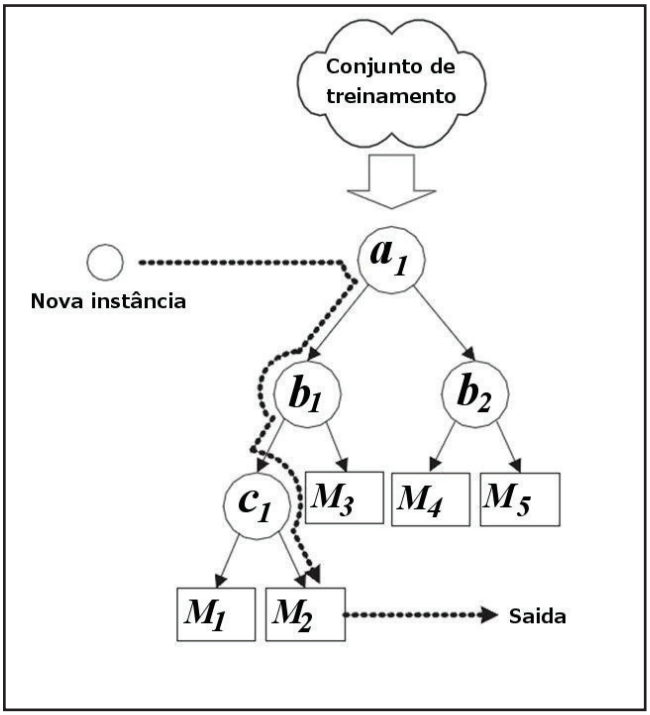


Figura 1 - Esquema Árvore de Decisão
Fonte: Solomatine e Dulal (2003)

número de atributos. Assim normalmente é impraticável definir uma estrutura ótima para um determinado problema, devido ao elevado custo computacional empregado na sua busca.

Nesse sentido, algoritmos baseados em heurística, que é uma técnica utilizada para resolver problemas de elevado nível de complexidade, sem que haja garantia de que a solução obtida seja a ótima, têm sido desenvolvidos para indução de árvores de decisão, apresentando resultados promissores em relação ao tempo e aos recursos computacionais necessários.

De acordo com Basgalupp (2010), o algoritmo Top-down induction of decision tree (TDIDT) é utilizado como base para muitos algoritmos de indução de árvores de decisão, por exemplo, ID3 (QUINLAN, 1986), CART (BREIMAN et al., 1984) e C4.5 (QUINLAN, 1993). De maneira geral, os algoritmos do tipo TDIDT são algoritmos recursivos, ou seja, produzem regras de decisão de forma implícita, em uma árvore de decisão, construída por sucessivas divisões das instâncias de treinamento de acordo com os valores de suas variáveis independentes (X1 e X2 na Figura 2). Assim, busca-se, dentro do conjunto de atributos aquele que melhor divide as instâncias de treinamento em subconjuntos.

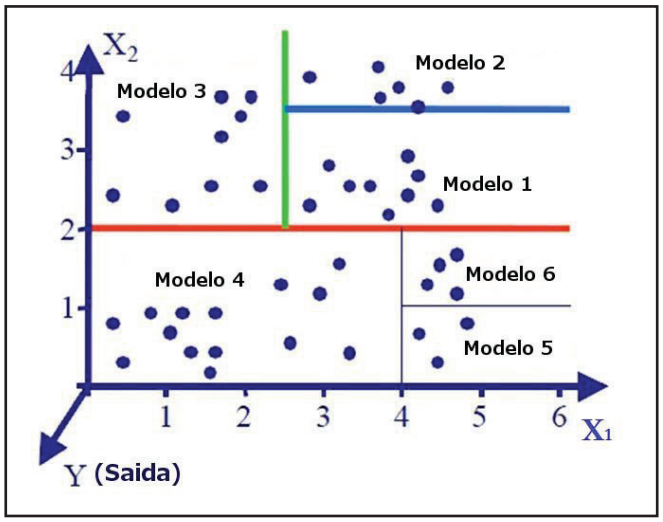


Figura 2 - Processo de divisão de árvore de decisão
Fonte: Solomatine e Dulal (2003)

Este processo se repete até que todos os exemplos de treinamento da classe (variável dependente) estejam classificados, ou até que todos os atributos preditivos tenham sido utilizados.

Os métodos usados para selecionar o melhor atributo preditivo para divisão da árvore de decisão se baseiam em medidas de impureza, de distância e de independência. Assim, a maior parte dos algoritmos de indução busca dividir os exemplos de um "nó-pai", de forma a minimizar o grau de impureza dos "nós-filhos", pois, quanto menor o grau de impureza, mais desbalanceada é a distribuição de classes. Em um determinado "nó", se a impureza é nula, todas as instâncias de treinamento contidas nele irão pertencer a uma mesma classe. Por outro lado, se o grau de impureza é máximo no "nó", cada classe terá o mesmo número de instâncias de treinamento.

Entre as medidas mais comumente utilizadas na divisão das árvores de decisão estão o ganho de informação (Equação 1) e a razão de ganho que é obtida pela divisão do ganho pela entropia (Equação 2 e 3).

$$Ganho = Entropia(pai) - \sum_{i=1}^{n} \frac{N(v_j)}{N} * Entropia(v_j) \quad (1)$$

onde

$$Entropia(nó) = -\sum_{i=1}^{c} p(i/nó) * log_2[p(i/nó)] \quad (2)$$

$$Razão\ Ganho = \frac{ganho}{Entropia(nó)} \quad (3)$$

onde: n é o número de valores do atributo, ou seja, número de nós-filhos, N é o número total de objetos do nó-pai, $N(v_j)$ é o número total de exemplos associados ao nó-filho(v_j), $p(i/nó)$ é a fração de instâncias de treinamento pertencentes à classe i no nó e c é o número de classes.

O ganho de informação utiliza a entropia como medida de impureza, comparando o seu valor para o "nó-pai" (antes da divisão) com a entropia dos "nós-filhos" (após a divisão). O atributo que gerar o maior ganho é escolhido como condição de teste.

A entropia está relacionada com a distribuição dos valores de uma variável, ou seja, um valor elevado quer dizer que há uma distribuição mais uniforme, ao contrário, uma entropia pequena quer dizer que os valores estão concentrados em um ou vários conjuntos ("clusters").

Assim, pode-se dizer que a árvore de decisão consiste em uma sequência de condições que divide o espaço dos dados de acordo com a variável dependente. Isso resulta em um conjunto de ramos que subdividem o espaço dos dados em hiperespaços disjuntos (folhas) da variável a ser prevista. Em uma árvore de decisão, os "nós" se subdividem até que um critério de terminação, ou poda, seja atingido (PAPPENBERGER;

IORGULESCU; BEVEN, 2006).

A poda permite excluir arestas ou sub árvores que representam um aprendizado muito especifico e que não deveria ser aplicado em todas as situações. A poda pode ser classificada em pré-poda ou pós-poda.

A pré-poda é realizada durante o processo de construção da árvore, quando o processo pode ser interrompido, parando de dividir o conjunto de instâncias, transformando, assim, o "nó" corrente em uma folha da árvore. O ganho de informação pode ser utilizado como critério de poda. Caso todas as divisões possíveis utilizando-se um atributo "A" produzam ganhos menores que um limiar predefinido, o "nó" vira folha (BASGALUPP, 2010).

A pós-poda é realizada após a construção da árvore de decisão. Para cada "nó" interno, o algoritmo calcula a taxa de erro caso a árvore seja podada ou não. Se a diferença entre os dois valores for menor que o limiar, a árvore é podada. Este processo se repete progressivamente, gerando uma árvore podada.

O algoritmo C4.5 (J48)

Uma implementação em Java deste algoritmo pode ser encontrada na ferramenta de domínio público WEKA - (Software Waikato Environment for Knowledge Analysis; Witten; Frank, 2000), com o nome de J48.

A tabela 1 lista os parâmetros utilizados para a construção da árvore de decisão do algoritmo J48 (WEKA).

O parâmetro binarySplits (B) serve para determinar se as classes são nominais (verdadeiro) ou numéricas, onde utiliza um operador de comparação do tipo "<" ou ">" (falso).

ConfidenceFactor (C) é o parâmetro de fator de confiança que tem o objetivo de analisar a precisão das regras geradas, pois uma redução no valor padrão (0,25) implica um aumento da poda da árvore.

Os parâmetros minNumObj e numFolds estão associados a recursos de poda da árvore de decisão. Cabe destacar que o numFolds com valor padrão (3) implica a separação de uma parte dos dados para poda e duas partes para construção da árvore. Aplicações deste algoritmo envolvendo recursos hídricos podem ser encontradas em Ali, Qamar e Ali (2013), Bhattacharya e Solomatine (2005), Cheng et al. (2012), Grunwald et al. (2009), Liao e Sun (2010), Schärer, Page e Beven (2006) e Solomatine e Dulal (2003).

MATERIAIS E MÉTODOS

Área de estudo

A Bacia do Rio das Velhas possui uma área de 29.173 km^2 e está localizada na região central do estado de Minas Gerais (Figura 3), onde estão inseridos 51 municípios, que abrigam uma população de, aproximadamente, 4,8 milhões de habitantes (destes, aproximadamente 89% residem em distritos e municípios integralmente inseridos na bacia), sendo que 44 destes estão totalmente inseridos na bacia (IGAM, 2013).

O curso principal do Rio das Velhas percorre uma extensão de 802 km, com nascente no município de Ouro Preto, a uma altitude de, aproximadamente, 1.500 m, e foz na Barra do Guaicuí, município de Várzea da Palma, a uma altitude de 478 m, onde é afluente do Rio São Francisco.

Entre os projetos existentes na Bacia do Rio das Velhas é importante mencionar os projetos Manuelzão e Águas de Minas (Instituto Mineiro de Gestão das Águas - IGAM) ambos iniciados em 1997 (MANUELZÃO, 2014). O biomonitoramento realizado pelo Projeto Manuelzão e pela Universidade Federal de Minas Gerais (UFMG) constatou que os peixes subiam apenas 200 km, a partir da foz do rio das Velhas em 2000. Em 2010, foi verificado que os peixes já avançavam 580 km no rio, chegando bem próximo às áreas consideradas mais degradadas na Região Metropolitana de Belo Horizonte (RMBH). O principal fator responsável pelos resultados obtidos foi o volume de esgoto tratado pela Companhia de Saneamento de Minas Gerais (COPASA) na Bacia do Rio das Velhas, que passou de 41 milhões de m3 em 2003 para 85 milhões de m^3 em 2008, atingindo 127 milhões de m^3 de esgoto tratado em 2010 (Manuelzão, 2014).

O monitoramento da qualidade das águas, por meio da obtenção dos parâmetros físicos, químicos e biológicos na Bacia do Rio das Velhas, é conduzido pelo Projeto Águas de Minas - IGAM desde 1997. Atualmente, existem 82 estações de qualidade das águas na Bacia do Rio das Velhas, distribuídas em diferentes projetos de Avaliação da Qualidade das águas Superficiais: Projeto Pampulha, Projeto Alto Velhas, Meta 2014 (IGAM, 2013). Em 2013, as principais fontes poluidoras, indicadas pelo IGAM como responsáveis pela degradação da qualidade das águas, foram: efluentes sanitários e industriais, pecuária e mineração; esses dois últimos associados à poluição difusa.

Tabela 1 - Parâmetros do algoritmo J48

J48			
Parâmetro	ID	Significado	Padrão
BinarySplits	B	Recurso para utilizar a árvore binária	Falso
ConfidenceFactor	C	Define o fator que será usado para a poda	0,25
minNumObj	M	Define o número mínimo de amostras por folha	2
numFolds	N	Define o número de partições realizada nos dados utilizados para poda com redução de erros.	3
SubtreeRaising	S	Recurso para utilizar subárvore durante a poda	Verdadeiro
Unpruned	U	Recurso para desabilitar o recurso de poda	Falso

Fonte: Adaptado de Banon (2013)

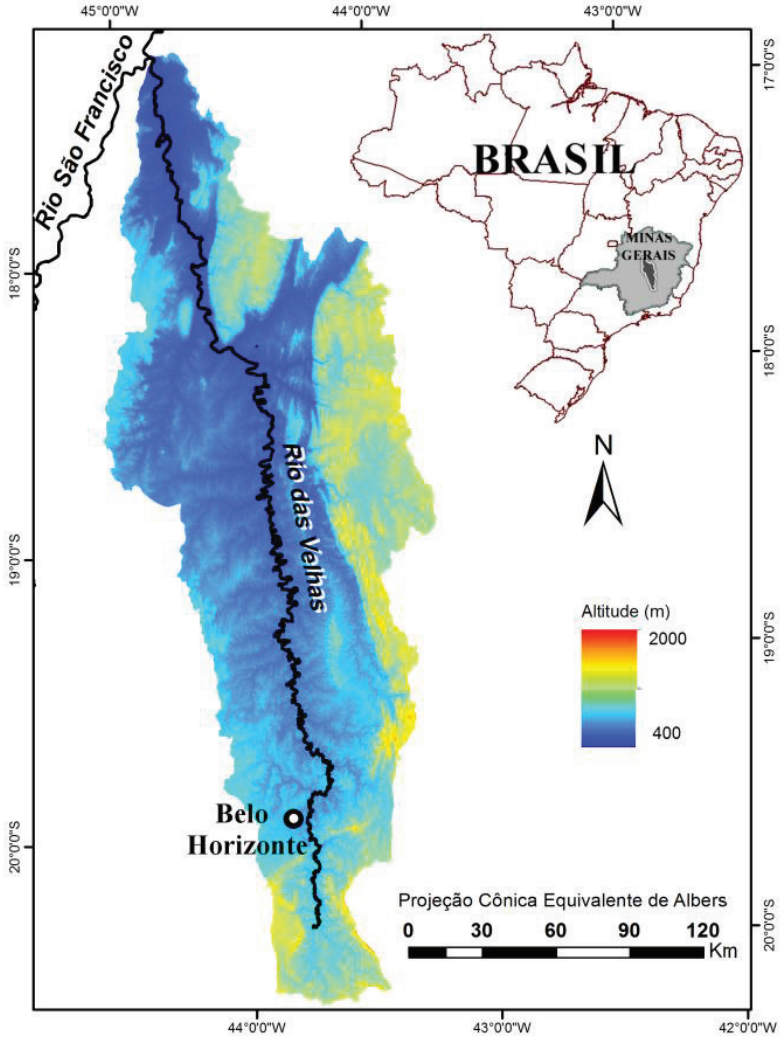


Figura 3 - Localização da Bacia do Rio das Velhas

Dados de qualidade das águas e vazão da bacia do rio das Velhas

Os parâmetros de qualidade Nitrato e Fósforo total foram obtidos mensalmente para o período de 2009 a 2011 em 17 estações monitoradas (Tabela 2) pelo IGAM e localizadas na calha principal do rio das Velhas.

As informações relacionadas às vazões monitoradas nas datas de medição de qualidade foram obtidas para 8 estações, sendo 7 pertencentes à Agência Nacional de Águas (ANA) e 1 à Companhia Energética de Minas Gerais (CEMIG), também localizadas na calha principal do Rio das Velhas.

A tabela 2 apresenta para cada estação de qualidade a correspondente estação de vazão utilizada no estudo.

A espacialização das estações, ao longo da calha do rio das Velhas, é apresentada na figura 4.

Tabela 2 - Estações de monitoramento de qualidade das águas e de vazão utilizadas

Estações de Qualidade	Estações Vazão
BV 013	41152000*
BV 037	
BV 063	41199998
BV 067	
BV 083	
BV 105	41260000
BV 153	
BV 137	41340000
BV 141	41600000
BV 142	41818000
BV 146	
BV 150	
BV 152	
BV 148	41990000
BV 149	
BV 151	
BV 156	41410000

* - Estação CEMIG.

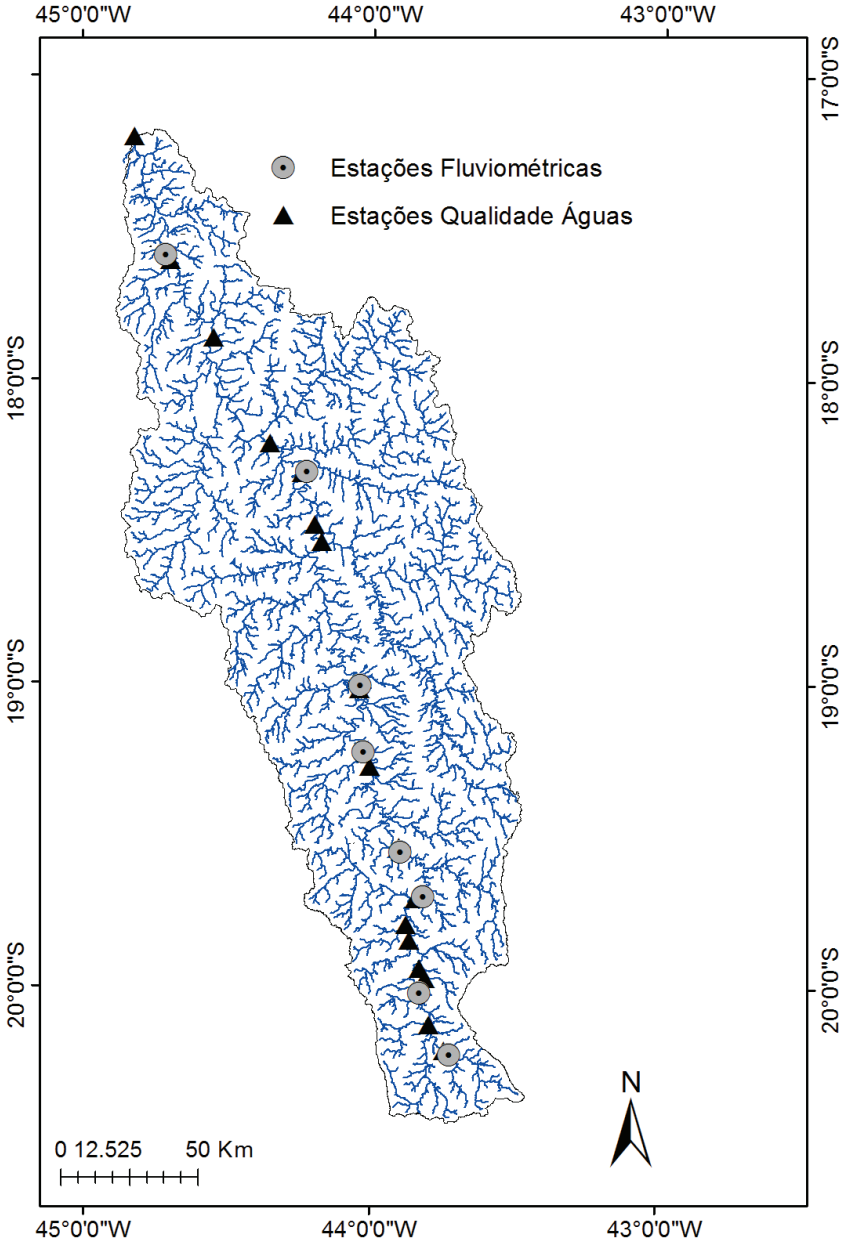


Figura 4 - Bacia do Rio das Velhas com estações fluviométricas e de qualidade de águas

Uso do solo da bacia do rio das Velhas

O mapeamento do uso do solo foi realizado pelo laboratório de Sensoriamento Remoto do Instituto de Geociências da UFMG em parceria com o IGAM. As imagens RapidEye do ano 2010 foram obtidas junto ao Instituto Estadual de Florestas (IEF) (Figura 5).

Este tipo de imagem é adquirida de uma constelação de 5 satélites em 5 bandas espectrais, do azul ao infravermelho próximo, e resolução de 5 metros. A partir de 2011 estas imagens passaram a ser compradas continuamente para todo o território nacional pelo Ministério do Meio Ambiente (MMA) e disponibilizadas para instituições públicas, universidades e institutos tecnológicos (BRASIL, 2014).

O processo de classificação de imagens foi realizado

Tabela 3 - Uso do solo da bacia do rio das Velhas em 2010 (as linhas cinza mostram as categorias agrupadas ou não consideradas)

Categorias	(%)	Categorias finais	(%)
Cerrado (CE)	26,4	Cerrado (CE)	26,4
Campo (CA)	21,5	Campo (CA)	21,5
Pasto (P)	25	Agropastoril (AP)	25,2
Agricultura (A)	0,2		
Mata (M)	16,1	Mata (M)	16,1
Urbano (AU)	2,9	Urbano (AU)	2,9
Mineração (MI)	0,2	Mineral (MI)	4,1
Afloramento(AF)	3,9		
Reflorest. (RE)	3,3	Reflorest. (RE)	3,3
Água (AG)	0,3	Não consideradas	0,5
Sem Info. (SI)	0,2		

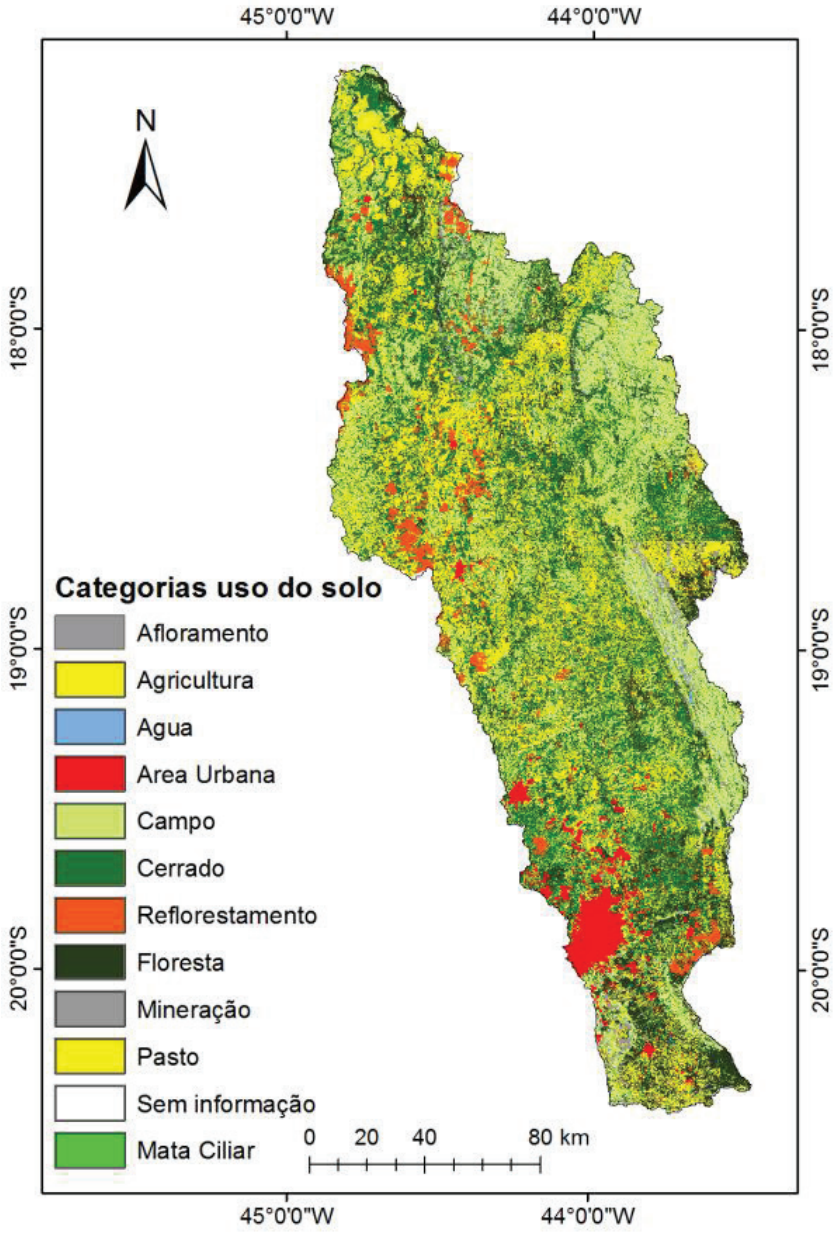


Figura 5 - Tipos de Uso do Solo da Bacia do Rio das Velhas

com o software Trimble® eCognition Developer 8.4 com utilização de regras baseadas, principalmente, na resposta espectral de cada alvo. As categorias mapeadas são apresentadas na tabela 3.

As categorias pasto e agricultura foram agrupadas gerando a categoria Agropastoril. O agrupamento foi realizado devido a dificuldade de diferenciar as duas categorias numa área tão grande como a da Bacia do Rio das Velhas, não sendo possível diferenciar a grande variedade de tipos de culturas, de forma a separa-la da categoria pasto, dificuldade acentuada considerando que as culturas se encontram em várias fases de crescimento e quase sem biomassa verde logo após a safra. Finalmente muitas plantações não são espectralmente diferenciáveis com apenas cinco bandas espectrais dos dados RapidEye.

Adicionalmente, como a categoria mineração ocupava apenas 0,2 % da área, ela foi mesclada com Afloramento dando origem à categoria Mineral. As categorias Água e Sem informação não foram utilizadas nas análises por não terem contribuição previsível como fonte difusa.

Período de análise e preparação da base de dados

As fontes de poluição difusa dependem da existência do escoamento superficial para atingir o curso de água, consequentemente a relação "uso do solo - poluição difusa - vazão" é mais acentuada nos meses chuvosos.

Realizamos as análises considerando um período hidrológico com vazões maiores que a Q50%. Dessa forma, os dois trimestres mais chuvosos foram selecionados para análise:

dezembro-fevereiro e novembro-janeiro dos anos hidrológicos 2009-2010 e 2010-2011. As normais climatológicas de Belo Horizonte (curso médio) e de Pirapora (foz) serviram de base na escolha desses trimestres e foram obtidas do Instituto Nacional de Meteorologia (INMET, 2014). Os anos foram selecionados considerando-se a representatividade do uso e da ocupação do solo obtido para o ano de 2010 (Figura 6).

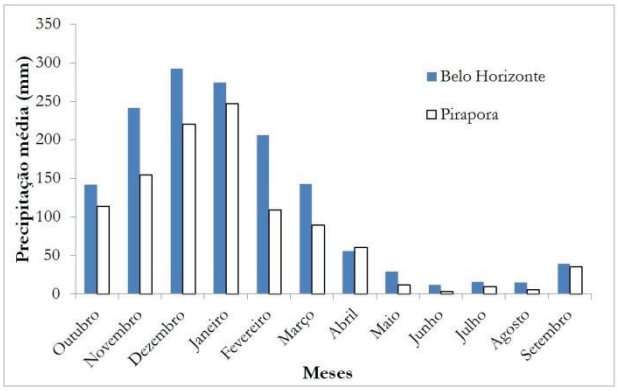


Figura 6 - Normais Climatológicas 1961-1990 para Belo Horizonte e Pirapora
Fonte: INMET (2014)

As 17 estações de qualidade selecionadas possuem área de drenagem variando de 554 km² (BV013) a 28.245 km² (BV149), próximo à nascente e à foz respectivamente. Essas estações apresentam valores de Nitrato e Fósforo Total bastante variáveis em função das fontes poluidoras e da dimensão da sua área de contribuição. O mesmo ocorre com a vazão, onde as estações de monitoramento fluviométrico possuem área de drenagem variando de 490 km² (41152000) a 25.940 km² (41990000).

A variação das áreas de drenagem e da proporção de usos em cada área impossibilita o adequado relacionamento das variáveis vazão, nitrato e fósforo total se forem utilizados seus valores absolutos. Por este motivo, realizamos a normalização destas informações, substituindo as mesmas pela sua permanência no tempo. As curvas de permanência de vazão, nitrato e fósforo total foram traçadas para cada uma das estações avaliadas.

A análise da permanência da vazão no tempo é uma abordagem bastante consolidada em recursos hídricos. Por outro lado, o uso desta mesma abordagem para parâmetros de qualidade é mais recente. A aplicação da permanência de parâmetros de qualidade de águas superficiais podem ser encontradas em Brites (2010), Cleland (2003), Cunha, Calijuri e Mendiondo (2012) e Formigoni et al. (2011).

Como o uso do solo e a permanência da vazão foram utilizados como variáveis independentes e a permanência do nitrato e do fósforo total como variáveis dependentes, selecionamos um período comum para análise destas variáveis. O uso do solo do ano de 2010 foi tido como o centro deste período de análise, com dois anos a mais e a menos, como representativos deste uso do solo.

Os dados de qualidade e vazão de 2008 a 2012 foram selecionados para gerar as curvas de permanência (Figura 7).

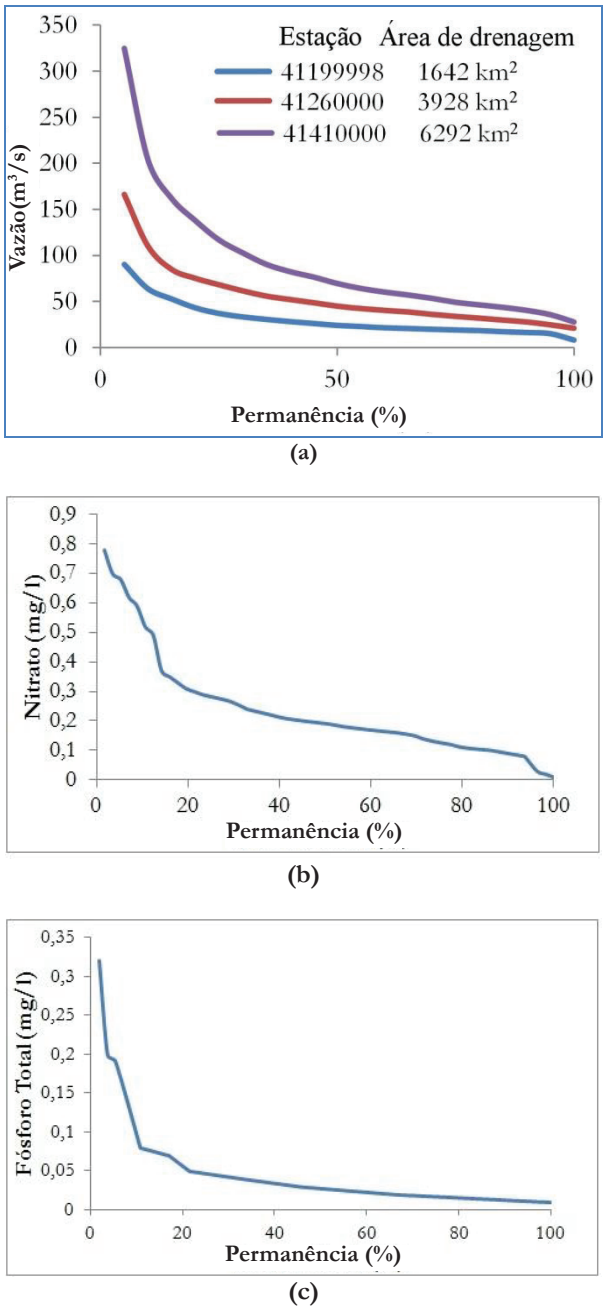


Figura 7 - a) Permanência de vazão das estações 41199998, 41260000 e 41410000, b) Permanência de Nitrato - BV013 e c) Permanência de Fósforo Total - BV013

Considerando que fontes poluidoras difusas tendem a ter o seu comportamento atrelado ao escoamento superficial, ou seja, em bacias com um número elevado de fontes difusas presentes, um aumento no volume escoado (vazão) é acompanhado da perda de qualidade das águas se a poluição difusa for significativa.

Para avaliar este efeito, neste trabalho, é proposto um índice (I_{Q_QA}) que relaciona a permanência da vazão à permanência da qualidade (equação 4):

$$I_{Q\_QA} = \frac{PermQ}{P\_QA} \quad (4)$$

onde: PermQ é a permanência da vazão (%) e $P_{\_QA}$ é a permanência do parâmetro de qualidade (%) e $I_{Q\text{-}QA}$ é a relação obtida pela equação 4.

As etapas de transformação da base de dados são apresentadas na figura 8.

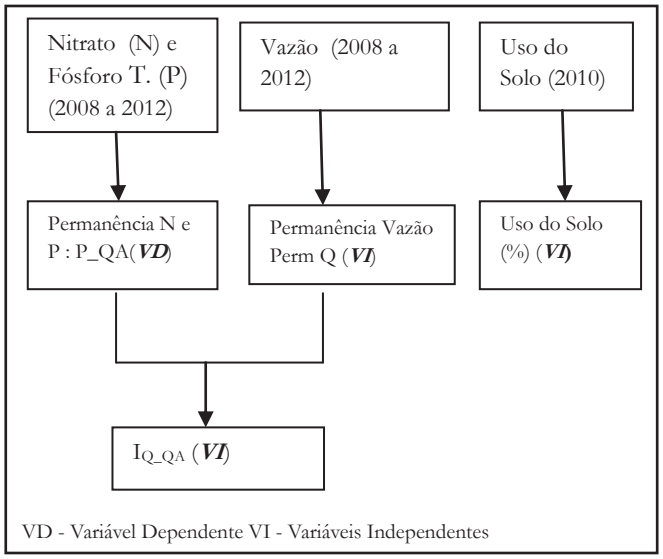


Figura 8 - Etapas de geração das bases de qualidade, vazão e uso do solo

O uso e a ocupação do solo foram avaliados em termos percentuais para as bacias de drenagem dos 17 pontos de qualidade estudados (Figura 8).

Árvore de decisão

As duas variáveis dependentes avaliadas foram permanência do nitrato e do fósforo total, enquanto as variáveis independentes foram: permanência da vazão (PermQ), I_{Q_QA} (Nitrato: $I_{Q\text{-}N}$ e Fósforo Total: I_{Q_P}) e 8 categorias de uso do solo (Tabela 3). O algoritmo C4.5 (J48) trabalha com a variável dependente (classe) discretizada, sendo o número de classes sugerido no WEKA igual a 10, com intervalos iguais. Como a entropia pode variar em função da discretização escolhida adotamos 6 discretizações dos dados em 2, 3, 4, 5, 7 e 10 intervalos, buscando avaliar as diferenças encontradas pelo algoritmo com cada uma delas. As variáveis independentes não necessitam ser discretizadas, sendo que o uso solo foi transformado em valores percentuais para as bacias dos pontos de monitoramento de qualidade das águas. A vazão foi avaliada pela sua permanência no tempo e o I_{Q_QA}, conforme definido na Equação 4. Os parâmetros do algoritmo C4.5 (J48), mostrados na tabela 1, foram utilizados com seus valores padrões.

Validação cruzada

Os algoritmos de indução de Árvores de Decisão constroem geralmente estruturas com mais ramificações que o necessário. O processo de "poda" da árvore busca reduzir os "nós" que trazem pouco ganho de informação.

No C4.5 a poda é realizada após a construção da árvore e determinação da taxa de erro da classificação produzida pela árvore. Se houver dados suficiente, uma parte é utilizada para construção da árvore e outra é reservada para testar e calcular a taxa de erro após a construção da árvore.

Caso os dados sejam escassos, pode-se utilizar um esquema de Validação Cruzada. Neste caso, os dados são divididos em *n* blocos de dimensão semelhante. A aprendizagem faz-se com recurso a *ni* iterações, em que a cada iteração são utilizados *n*-1 blocos para aprendizagem (treinamento) e o bloco que sobrou para teste, sendo este diferente a cada iteração. A técnica de validação cruzada foi adotada em função do número reduzido de instâncias.

Neste trabalho, foram utilizados 95 instâncias para o trimestre Dez-Fev e 104 instâncias para Nov-Jan. A diferença no número de instâncias se deu em função da inexistência de dados em algumas estações em determinadas datas.

RESULTADOS

Análise das classificações

As tabelas 3 e 4 apresentam os resultados das seis discretizações testadas através do percentual de instâncias classificadas corretamente e incorretamente para os trimestres Nov-Jan de 2009 a 2011 (Tabela 4) e Dez-Fev dos mesmos três anos (Tabela 5).

Tabela 4 - Instâncias classificadas corretamente e incorretamente em função da discretização utilizada nos dados de qualidade. Nov.-Jan. 2009-2011

Discr	Nitrato		Fósforo Total	
	Inst. Cor.(%)	Inst. Inc.	Inst. Cor.	Inst. Inc.
2	67	33	69	31
3	83	17	82	18
4	49	51	72	28
5	53	47	67	33
7	38	62	51	49
10	24	76	43	57

Tabela 5 - Instâncias classificadas corretamente e incorretamente em função da discretização utilizada nos dados de qualidade Dez - Fev. 2009-2011

Discr	Nitrato		Fósforo Total	
	Inst. Cor.	Inst. Inc.	Inst. Cor.	Inst. Inc.
2	48	52	65	35
3	59	41	80	20
4	62	38	66	34
5	57	43	57	43
7	47	53	46	54
10	27	73	42	58

Os melhores resultados da classificação do trimestre Nov-Jan foram obtidos para a discretização do nitrato e do fósforo total em 3 classes, obtendo-se 83% e 82% das amostras classificadas corretamente, respectivamente, para o período de Nov-Jan 2009-2011.

Os resultados do trimestre Dez-Fev de 2009-2011 (Tabela 5) apresentaram um número inferior de amostras classificadas corretamente em relação a Nov-Jan (Tabela 4). A melhor árvore de decisão para o nitrato foi obtida com 4 discretizações (Tabela 5), classificando corretamente 62% das instâncias. Para o fósforo total o intervalo de 3 discretizações também obteve melhor resultado com 80% das instâncias classificadas corretamente (Tabela 5).

A variação no número de instâncias classificadas corretamente e incorretamente para as 6 discretizações da variável dependente (permanência de nitrato e fósforo total) é menor para o parâmetro fósforo que para o nitrato (Figuras 9 e 10), considerando os dois trimestres avaliados. As figuras 9 e 10 apresentam os resultados para o período de Nov-Jan.

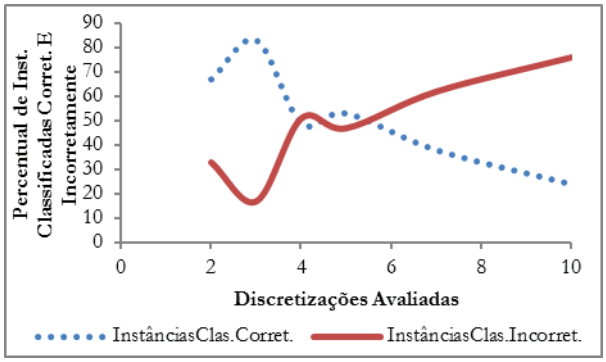


Figura 9 - Percentual de instâncias classificadas correta e incorretamente para Nitrato (Nov- Jan / 2009-2011)

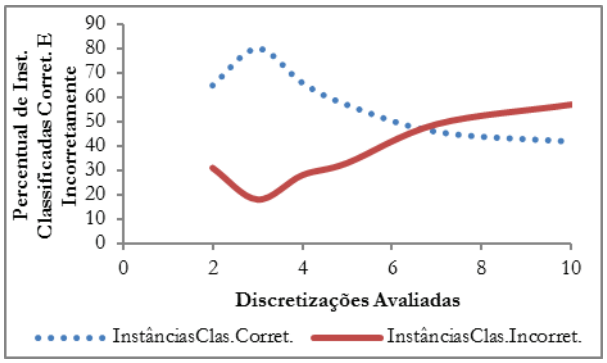


Figura 10 - Percentual de instâncias classificadas correta e incorretamente para Fósforo Total (Nov- Jan / 2009- 2011)

Uma vez que as árvores geradas (Nitrato e Fósforo total) no período de Nov-Jan de 2009-2011 (Figura 11) tiveram melhor resultado de classificação as demais análises apresentadas se referem a este período.

Considerando a figura 11, verifica-se que a separação dos dados ocorreu inicialmente através do uso do índice IQ-QA (equação 4) representado pelos símbolos IQ_N (nitrato) e IQ-P (fósforo total).

O segundo parâmetro utilizado foi o uso do solo, representado pelas categorias: AP, AU, CA e MI para o nitrato e pelas categorias: AU e AP para o fósforo total, figuras 11a e 11b, respectivamente.

No modelo do nitrato (Figura 11a) para IQ_N > 0,72, o algoritmo não conseguiu encontrar categorias de uso do solo e de permanência da vazão (PermQ), que pudessem explicar a permanência do nitrato no intervalo -INF a 40,48%. Os dados desse intervalo apresentam uma média e um desvio padrão do IQ_N de 0,99 e 0,32, respectivamente. Já para IQ_N □ 0,72 a média e o desvio padrão do IQ_N foram de 0,45 e 0,15, respectivamente.

Tem-se ainda que os valores de IQ_N □ 0,72 ocorreram para uma permanência da vazão, em média, de 15%, contra 30% para IQ_N > 0,72. Isso significa que o pressuposto de que a poluição difusa é principalmente observável com vazões elevadas é válido.

Com relação ao uso do solo (urbano e agropastoril), uma elevada dispersão nos dados para IQ_N>0,72 é observada (Tabela 6), o que faz com que a relação entre estas categorias de uso do solo e permanência do nitrato não seja estabelecida.

Tabela 6 - Média e desvio padrão das categorias de uso do solo Agropastoril e Urbano para as faixas de IQ_N

	Agropastoril		Urbano	
I_{Q_N}	Média	Desvio	Média	Desvio
≤ 0,72	18,9	1,7	15,7	1,2
>0,72	23,2	2,6	7,2	5,3

O uso do solo agropastoril (AP), urbano (AU) e o aumento da vazão foram responsáveis pelo aumento da concentração de nitrato representada pelo intervalo - INF a 40,48%.

As bacias de contribuição, das estações de qualidade avaliadas, com menor proporção de área urbana, foram responsáveis pela melhor qualidade da água em relação ao nitrato (40,48 a 70,23%). Tal efeito também foi verificado para a presença de uso do solo das categoriais campo (CA) e mineral (MI), ficando a permanência do nitrato entre 40,48 a INF.

Para o fósforo (Figura 11b), o uso do solo agropastoril (AP) foi associado à maior concentração de fósforo total, - INF a 34,52% para IQ-P > 0,5.

As estações de qualidade com bacias de contribuição com menor percentual de área urbana foram responsáveis por menor concentração de fósforo total, resultando em sua maior permanência no tempo (34,52- 67,26%).

Para IQ-P entre 0.3 e 0,5 não foi possível associar usos do solo à permanência do fósforo total nos três intervalos avaliados e o algoritmo trabalhou somente com as variações da vazão (PermQ), permitindo observar, neste intervalo que o aumento da vazão foi associado ao aumento da concentração de fósforo (Figura 11b).

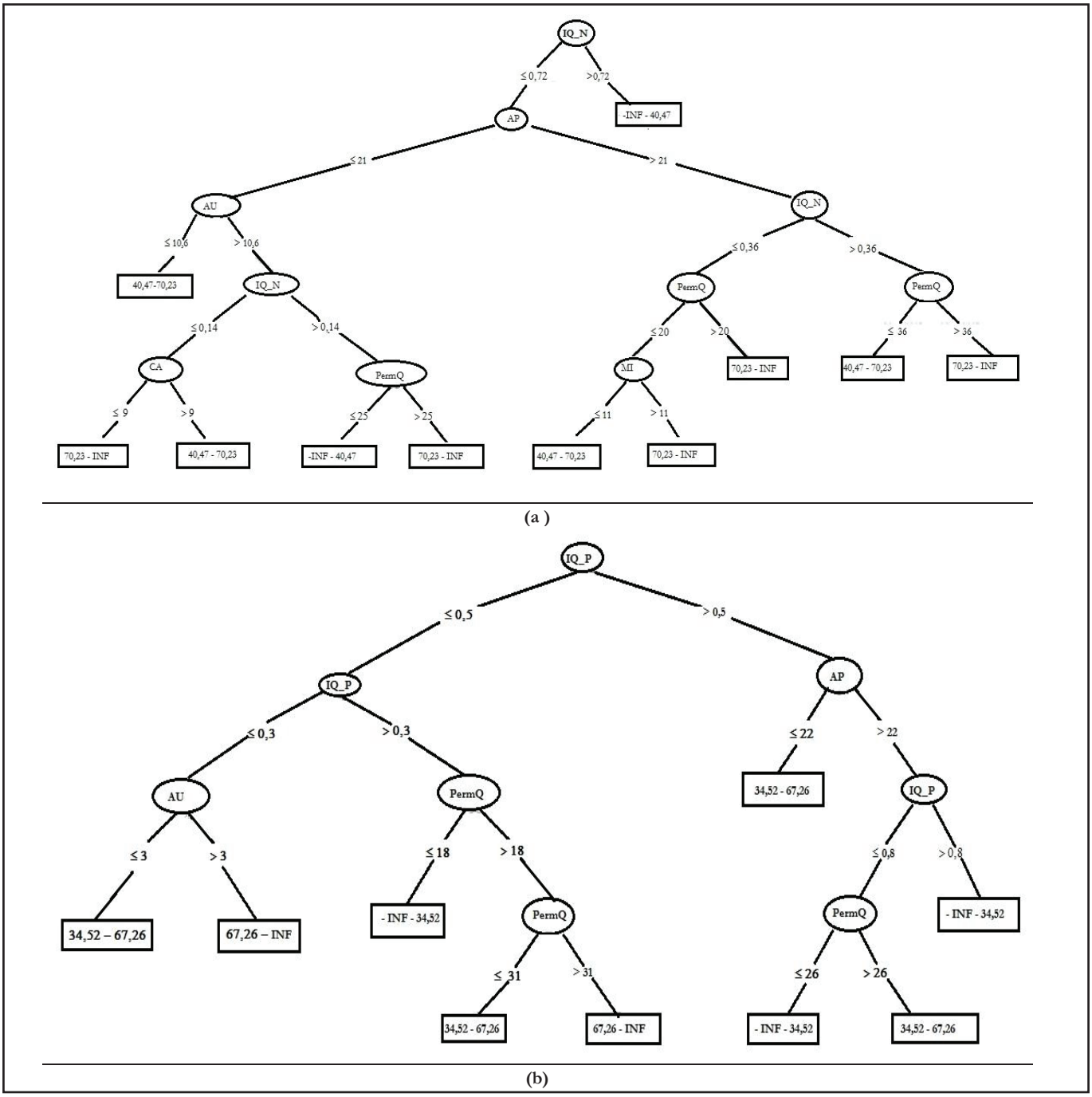


Figura 11 - a) Árvore de decisão para o nitrato Nov-Jan/2009-2011 e b) para o fósforo total Nov-Jan/2009-2011. Legenda: IQ_N - Índice que relaciona a permanência da vazão (PermQ) a permanência do nitrato, AP - Percentual de agropastoril, AU Percentual de área urbana, CA - Percentual de Campo, MI - Percentual de Mineral e IQ_P - Índice que relaciona a permanência da vazão (PermQ) a permanência do fósforo total

Avaliação das regras das àrvores de decisão

Ao todo foram avaliados 104 instâncias (Figura 12) relacionando a permanência de vazão à permanência do nitrato (entre 10.71% e 100%) e do fósforo total (entre 1,78 e 100%.), respectivamente.

A permanência do nitrato e do fósforo total foi discretizada em 3 classes (melhor resultado obtido para classificação) C1, C2 e C3, com diferente número de instâncias (TA) enquadradas em cada classe (Tabela 7).

Tabela 7 - Intervalos de discretização das variáveis dependentes permanência do nitrato e do fósforo total (Nov-Jan)

Parâmetro	C1/(TA)	C2/(TA)	C3/(TA)
Nitrato	-INF-40,5 (18)	40,5-70,2(46)	70,2-INF(40)
Fósforo T.	-INF-34,7 (36)	34,5-67,3(22)	67,3-INF(46)

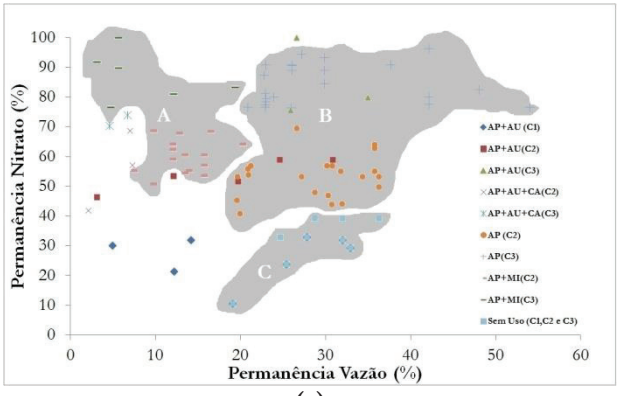
|Legenda: Ci - Intervalo da classe TA- Número instâncias

As classes C1 e C3 representam alta e baixa concentração de nitrato e fósforo total avaliados na estações pesquisadas para o período de Nov-Dez de 2009-2011, respectivamente. A classe C2 possui valores intermediários (médios).

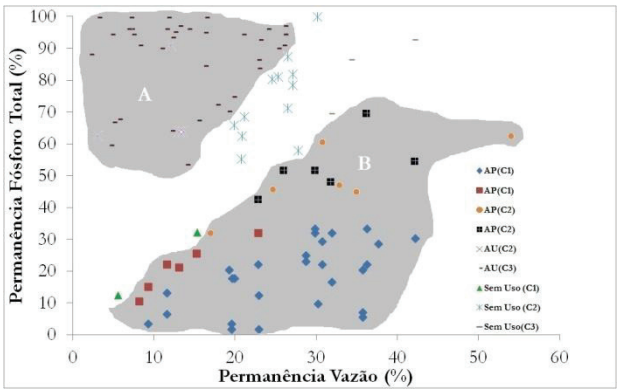
Os agrupamentos das classes para a permanência do nitrato e do fósforo total são apresentados na figura 12. O uso do solo é representado pela cor dos símbolos pontuais. O uso do solo agropastoril (AP) associado ao mineral (MI) foi relacionado à média e baixa concentração de nitrato, agrupamento "A" da figura 12a. A associação do uso urbano (AU) com agropastoril (AP) e campo (CA) apresentou confusão com os demais usos em todas as classes avaliadas (C1, C2 e C3). O agrupamento "B" apresentou os dados classificados com uso do solo agropastoril, também vinculados às classes C2 e C3 do nitrato. A classe C1 do nitrato teve baixa representatividade, em função do número reduzido de amostras nesta classe (TA=18) (Tabela 7).

Para o parâmetro nitrato, o maior problema esteve associado ao desbalanceamento entre as classes da Tabela 7. Ou seja, as classes C2 e C3 contêm a maior parte das instâncias, restando somente 18 instâncias para C1, o que se refletiu na dificuldade de associação da variável uso do solo para esta classe.

Para o fósforo total, o agrupamento "A" da Figura 12b, contém o uso do solo urbano (AU) vinculado às classes de permanência C2 e C3, ou seja, com concentração variando de média a baixa de fósforo total. O uso agropastoril, agrupamento "B", foi relacionado às classes C1 e C2 com concentração variando de média a alta.



(a)



(b)

Figura 12 a)- Agrupamento classes do nitrato b) Agrupamento classes do fósforo Total

DISCUSSÃO

A aplicação do algoritmo C4.5 para o estabelecimento de regras de comportamento entre a permanência da qualidade dos parâmetros nitrato e fósforo total, a permanência da vazão e o uso do solo permitiu obter sucessos de classificação de 83% e 82%, respectivamente.

A eficiência da classificação foi maior para o trimestre chuvoso de Nov-Jan 2009-2011, com vazões com permanência entre 2 e 52%. Para o trimestre Dez-Jan 2009-2011 as vazões tiveram permanência variando entre 3 e 68%. Assim verificamos que os meses com maiores vazões obtiveram melhor resultado na classificação, colaborando com a associação da poluição difusa ao aumento da vazão.

Outro aspecto que reforça esta relação vazão-uso do solo-poluição difusa foi a seleção do índice IQ_QA, como variável inicial para a divisão dos dados, haja visto que no período chuvoso, o IQ-QA deve ser maior para fontes de poluição difusa e menor para fontes pontuais. Em se tratando de fontes pontuais a permanência da vazão diminui e a permanência do parâmetro de qualidade aumenta, indicando uma melhor qualidade das águas, aspecto representado pela permanência do fósforo no cluster A da figura 12b.

Entretanto, foi possível observar que, a discretização utilizada para as variáveis dependentes (permanência de nitrato e fósforo total) influencia a classificação. O aumento no número de intervalos utilizados na discretização promoveu uma queda na eficiência da classificação. Tal fato ocorre, basicamente, em função do reduzido número de instâncias (dados disponíveis).

O monitoramento mensal das variáveis de qualidade não permite a construção de sequências temporais consistentes, para que as variações possam ser acompanhadas. Isto promove um desbalanceamento das classes dificultando a classificação. No caso do nitrato a distribuição de frequência das classes foi desbalanceada, e o algoritmo teve maior dificuldade de prever a classe mais rara (C1 - Tabela 7), confirmando o exposto por Breiman et al. (1984) e Witten e Frank (2000).

Apesar desta dificuldade as árvores de decisão confirmaram que os usos urbano e agropastoril estão associados à alterações de fósforo total e nitrato, verificadas na Bacia do Rio das Velhas, o que corrobora com as fontes poluidoras indicadas por IGAM (2013).

Para o nitrato os clusters A e B (Figura 12a) estão associados ao uso do solo agropastoril, com a permanência do nitrato variando de 40,47 a INF. O modelo definido pela árvore de decisão não conseguiu prever o uso do solo associado a permanência do nitrato de -INF a 40,47 devido a número reduzido de instâncias nesta faixa de permanência (Figura 12a).

A permanência do fósforo total entre 34,5-INF ocorreu para vazões com permanência entre 10 e 25% (cluster A - Figura 12b). O uso do solo urbano foi associado a esta situação, onde o aumento da vazão esta vinculado a uma redução na concentração do fósforo total. O uso do solo agropastoril foi associado a menor permanência do fósforo total (-INF a 34,5) e permanência de vazões entre 8 e 53% (cluster B - Figura 12b). Ou seja, o aumento do escoamento superficial esteve associado ao aumento do fósforo total.

Alguns dados de fósforo total não puderam ser vinculados a nenhuma categoria de uso do solo e são visualizados na figura 12b, através dos pontos não agrupados em clusters.

O uso de árvores de decisão se mostrou viável no processo de identificação das fontes poluidoras na Bacia do Rio das Velhas. Entretanto é necessário o aumento no volume de informações disponíveis para aprimorar a eficiência da ferramenta.

A obtenção de um maior número de observações, através da intensificação das campanhas de monitoramento, permitiria uma melhor exploração das possibilidades do uso de árvores de decisão.

AGRADECIMENTOS

Os autores agradecem ao Instituto Mineiro de Gestão das Águas (IGAM) pela concessão de dados do monitoramento de qualidade de águas da Bacia do Rio das Velhas; à Agência Nacional de Águas (ANA) e à Companhia Energética de Minas Gerais (CEMIG), pela concessão de dados do monitoramento de vazão; ao Instituto Estadual de Florestas (IEF), pela concessão das imagens do sistema RapidEye, ao INMET (Instituto Nacional de Meteorologia), pelo fornecimento das normais climatológicas e ao Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) pela liberação do professor para capacitação.

REFERÊNCIAS

ALI, M., QAMAR, A. M. ALI, B. Data analysis, Discharge classifications, and Predictions of Hydrological Parameters for the Management of Rawal Dam in Pakistan. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING AND APPLICATIONS, 12., 2013, Florida. *Proceedings*... Florida: IEEE, 2013. p. 382-385.

BANON, L. C. *Árvores de decisão aplicadas à extração automática de redes de drenagem*. 2013. 115 f. Tese (Doutorado) - Instituto Nacional de Pesquisas Espaciais - INPE, São José dos Campos, 2013.

BASGALUPP, M. P. LEGAL - *Tree*: um algoritmo genético multiobjetivo lexicográfico para indução de árvores de decisão. 2010. 116 f. Tese (Doutorado) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2010.

BHATTACHARYA, B. E D. P. SOLOMATINE, D. P. Neural networks and m5 model trees in modelling water level–discharge relationship. *Neurocomputing*, v. 63, p. 381-396, Jan. 2005.

BRASIL. Ministério do Meio Ambiente. *Sobre o projeto.* Brasília: MMA, 2014. Disponível em: <http://geocatalogo.ibama.gov.br/sobre.jhtml>. Acesso em: 3 fev. 2015.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. *Classification and regression trees.* Boca Raton: CRC Press, 1984.

BRITES, A. P. Z. *Enquadramento dos corpos de água através de metas progressivas*: probabilidade de ocorrência e custos de despoluição hídrica. 2010. 177 f. Tese (Doutorado) - Escola Politécnica, Universidade de São Paulo, 2010.

CHAU, K.-W. A review on integration of artificial intelligence into water quality modelling. *Marine Pollut. Bull.*, v. 52, n.7, p. 726-733, July 2006.

CHENG, KE-FEI; CHENG, LEI; HUANG, YONG-DONG. Water Quality Evaluation Method Based on J48 Decision Tree Algorithm. *Computer Eng. J.*, v. 38, p. 264-267, 2012.

CLELAND, B. R. Tmdl development from the "bottom up"–part III: duration curves and wet-weather assessments. Proc. Water Environ. *Federation*, n. 4, p. 1740-1766, Sept. 2003.

CUNHA, D. G. F.; CALIJURI, M. C.; MENDIONDO, E. M. Integração entre curvas de permanência de quantidade e qualidade da água como uma ferramenta para a gestão eficiente dos recursos hídricos. *Eng. Sanit. Amb.*, v. 17, n. 4, p. 369-376, out.-dez. 2012.

FORMIGONI, Y.; BRITES, A. P.; FERNANDES C. E.; PORTO, M. Análise crítica da curva de permanência de qualidade de água com base em dados históricos. In: SIMPÓSIO BRASILEIRO DE RECURSOS HÍDRICOS, 19., 2011, Maceió. *Anais*... Porto Alegre: ABRH, 2011.

GORSEVSKI, P. V.; BOLL, J.; GOMEZDELCAMPO, e.; Brooks, E. S. Dynamic riparian buffer widths from potential non-point source pollution areas in forested watersheds. *Forest Ecol. Management*, v. 256, n. 4, p. 664-673, Aug. 2008.

GRUNWALD, S.; DAROUB, S. H.; LANG, T. A.; DIAZ, O. A. Tree-based modeling of complex interactions of phosphorus loadings and environmental factors. *Sci. Total Environ.*, v. 407, n. 12, p. 3772-3783, Jun. 2009.

HUANG, J.; HONG, H. Comparative study of two models to simulate difuse nitrogen and phosphorus pollution in a medium-sized watershed, southeast china. *Estuarine, Coastal Shelf Sc*i., v. 86, n. 3, p. 387-394, Feb. 2010.

IGAM - Instituto Mineiro de Gestão das Águas. *Monitoramento da Qualidade das Águas Superficiais no estado de Minas Gerais* - 10 Trimestre de 2013. Belo Horizonte: Instituto Mineiro de Gestão das Águas, 2013.

INMET- Instituto Nacional de Meteorologia. *Normais Climatológicas de 1961-1990*. [S.l.]: INMET, 2014. Disponível em: <http://www.inmet.gov.br/portal/index.php?r=clima/normaisClimatologicas>. Acesso em: 13 nov. 2014.

LIAO, H.; SUN, W. Foresting and evaluating water quality of chao lake based on an improved decision tree method. *Procedia Environ. Sci.*, v. 2, p. 970-979, 2010.

MAILLARD, P.; SANTOS, N. A. P. A spatial-statistical approach for modeling the effect of non-point source pollution on different water quality parameters in the velhas river watershed–brazil. *J. Environ. Manage.*, v. 86, n. 1, p. 158-170, Jan. 2008.

MANUELZÃO. Sem ações sustentáveis os rios morrem. Edição especial. Belo Horizonte Projeto Manuelzão. *Manuelzão:* saúde, ambiente e cidadania na bacia do rio das Velhas, ano 17, out. 2014. Edição especial.

PAPPENBERGER, F.; IORGULESCU, I.; BEVEN, K. J. Sensitivity analysis based on regional splits and regression trees (sars-rt). *Environ. Modelling Software*, v. 1, n. 7, p. 976-990, July 2006.

PINHEIRO, A. SCHOEN, C., CHULTZ J., HEINZ K. G. H. PINHEIRO, I. G., DESCHAMPS, C. Relação Entre o Uso do Solo e a Qualidade da Água em Bacia Hidrográfica Rural no Bioma Mata Atlântica. *RBRH:* revista brasileira de recursos hídricos, v. 19, n. 3, jul./set. 2014.

QUINLAN, J. R. Induction of decision trees. *Machine Learning*, v. 1, n. 1, p. 81-106, 1986.

QUINLAN, J. R. *C4. 5:* programs for machine learning. San Mateo, CA: Morgan Kaufmann, 1993. (The Morgan Kaufmann series in machine learning.

SCHÄRER, M.; PAGE, T.; BEVEN, K. A fuzzy decision tree to predict phosphorus export at the catchment scale. *J. Hydrol.*, v. 331, n. 3-4, p. 484-494, Dec. 2006.

SHARIFI, S.; HOSSEINI, S. Methodology for identifying the best equations for estimating the time of concentration of watersheds in a particular region. J. *Irrigation Drainage Eng.*, v. 137, n. 11, p. 712-719, Nov. 2011.

SOLOMATINE, D. P.; DULAL, K. N. Model trees as an alternative to neural networks in rainfall runoff modelling. *Hydrol. Sci. J.*, v. 48, n. 3, p. 399-411, 2003.

WITTEN, I. H.; FRANK, E. *Data mining*: practical machine learning tools and techniques with Java Implementations. San Francisco: Morgan Kaufmann, 2000.

ZHENYAO S.; LEI S. QIAN L.; RUIMIN L.; QIAN H.. Impact of spatial rainfall variability on hydrology and nonpoint source pollution modeling. *J. Hydrol.*, v. 472-473, p. 205-215, Nov. 2012.