# Aplicação de Descoberta de Conhecimento em Bases de Dados na Estimativa da Evapotranspiração: um Experimento no Estado do Rio de Janeiro

## Alternate Title: Application of Knowledge Discovery in Databases in Evapotranspiration Estimation: an Experiment in the State of Rio de Janeiro

Fernando Xavier
Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Avenida Pasteur, 458 - Urca
Rio de Janeiro – RJ – Brasil
fernando.xavier@uniriotec.br

Asterio Kiyoshi Tanaka
Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Avenida Pasteur, 458 - Urca
Rio de Janeiro – RJ – Brasil
tanaka@uniriotec.br

Kate Cerqueira Revoredo
Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Avenida Pasteur, 458 - Urca
Rio de Janeiro – RJ – Brasil
katerevoredo@uniriotec.br

## RESUMO

Com o crescimento do volume de dados em diversas áreas, como a Hidrologia, aumenta a necessidade do uso de sistemas de informação para auxílio na manipulação desses dados. Este artigo é um relato de um experimento que usou técnicas de descoberta de conhecimento para a estimativa de um importante componente do ciclo hidrológico: a evapotranspiração. O experimento relatado neste artigo foi realizado com dados meteorológicos e mostrou que alguns algoritmos, como o M5P, apresentam bons resultados quando comparados com os dados históricos da evapotranspiração estimada.

## Palavras-Chave

KDD, Mineração de Dados, Hidrologia, Evapotranspiração, Regressão Linear.

## ABSTRACT

With the growing volume of data in various areas such as Hydrology, there is a need for using information systems to aid in handling such data. This article is a report of an experiment that used knowledge discovery techniques to estimate an important component of the hydrological cycle: evapotranspiration. The experiment reported in this article was done with weather data and showed that some algorithms, such as M5P, present good results when compared to historical data of the estimated evapotranspiration.



## Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Database Applications – *data mining.*

## General Terms

Algorithms, Experimentation, Theory.

## Keywords

KDD, Data Mining, Hydrology, Evapotranspiration, Linear Regression.

## 1. INTRODUÇÃO

A água é um recurso fundamental para o planeta e, por conta disso, existem diversas iniciativas destinadas a estudar os assuntos relacionados a este recurso. Uma dessas iniciativas é o programa Intergovernamental Panel on Climate Change (IPCC), mantido pela Organização das Nações Unidas (ONU) e pela Organização Meteorológica Mundial (WMO), para avaliar as informações científicas, técnicas e socioeconômicas produzidas no mundo, que são importantes para a compreensão das mudanças climáticas [15], que está relacionada a interesses em atividades como: monitoramento de desastres naturais, agricultura, abastecimento de água, geração de energia, entre outros.

Esses interesses são objetos de estudos de ciências como a Hidrologia, que é a ciência que trata da água no planeta, sua ocorrência, circulação e distribuição [3]. As pesquisas relacionadas à Hidrologia usam, por exemplo, dados de séries históricas de chuvas, vazões dos rios, umidade do ar, temperatura, dentre outros, que são importantes para a compreensão dos fenômenos naturais, como o ciclo hidrológico, que tem influência no clima de diversas formas [3].

Dentre os componentes do ciclo hidrológico, destaca-se a evapotranspiração, que é a soma da evaporação da água do solo com a transpiração da vegetação, que retorna para a atmosfera na forma de vapor [2]. Devido ao fato da medida direta da evapotranspiração ser difícil e onerosa, a abordagem corrente é

sua estimativa através de métodos matemáticos [2], como a equação de Penman-Monteith, que é atualmente usada como referência pela Organização das Nações Unidas para Alimentação e Agricultura (FAO) [5]. Uma limitação dessa abordagem é que ela depende da disponibilidade dos valores de todas as variáveis da equação, o que impede que ela seja usada na ausência de valor de uma delas.

Tendo em vista essa limitação, o presente trabalho busca responder à seguinte questão: é possível estimar a evapotranspiração independentemente da disponibilidade de todas as variáveis? Para responder a esta questão, propõe-se uma alternativa para que a estimativa seja feita por modelos gerados através do processo de descoberta de conhecimento em banco de dados (em inglês Knowledge Discovery in Databases - KDD) [6]. Usando *datasets* com séries históricas de dados meteorológicos, espera-se, através de KDD, gerar um modelo que possa estimar o valor da evapotranspiração mesmo que os valores de todas as variáveis do *datasets* não estejam disponíveis.

Para avaliar a solução proposta, os valores calculados no modelo gerado são comparados com os valores históricos, disponíveis nos *datasets* utilizados no experimento executado neste trabalho. Se houver uma boa aproximação entre os valores gerados pelos modelos e os valores das séries históricas, então se pode demonstrar a viabilidade desta abordagem para a estimativa da evapotranspiração.

O restante do artigo está organizado como se segue. Na Seção 2, é apresentada a fundamentação teórica dos conceitos relacionados a este trabalho. Já na Seção 3, é descrita a metodologia utilizada neste trabalho, seguida na Seção 4 pela descrição da execução do experimento. Na Seção 5, é feita uma análise dos resultados encontrados e, na Seção 6, é apresentado um estudo dos trabalhos relacionados indicando as contribuições desde trabalho. Por fim, são feitas as considerações finais a respeito deste trabalho, indicando as contribuições esperadas, limitações e possíveis desdobramentos.

# 2. FUNDAMENTAÇÃO TEÓRICA

## 2.1 Evapotranspiração

A evapotranspiração, medida em milímetros/dia, é referida como a combinação dos processos de evaporação da água do solo e da transpiração da vegetação, que retorna à atmosfera em forma de vapor [2]. Conforme a Figura 1, a evapotranspiração está relacionada a outros componentes do ciclo hidrológico, como as chuvas, capacidade de infiltração do solo, dentre outros.

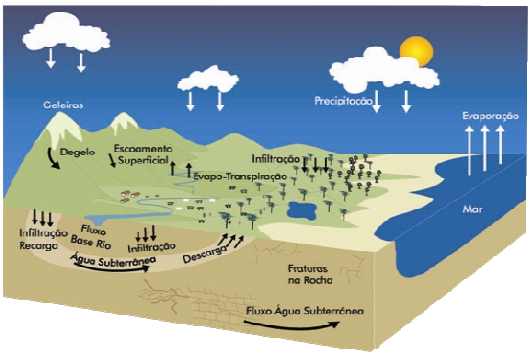


**Figura 1. Ciclo Hidrológico [14]**

Os valores da evapotranspiração são de fundamental importância em áreas como a agricultura, principalmente nas atividades de irrigação do solo. Nessas atividades, a captação e elevação da água são de grande importância, pois, além de consumirem energia e investimentos em equipamentos, podem ter impactos ambientais, como o uso inadequado dos recursos hídricos [18]. Nesse sentido, a disponibilidade do valor da evapotranspiração permite planejar melhor a quantidade de água a ser irrigada e, consequentemente, minimizar o uso dos recursos e os impactos ambientais.

A medida direta da evapotranspiração é difícil e onerosa, ao exigir instalações e equipamentos especiais, justificando seu uso apenas em condições experimentais [2]. Por isso, a abordagem corrente na área da Hidrologia é sua estimativa através de modelos matemáticos, a partir de variáveis meteorológicas, como a equação de Penman-Monteith [5], que é o método recomendado pela FAO para estimativa da evapotranspiração.

A equação de referência da FAO, como outros modelos matemáticos existentes, têm como dependência a disponibilidade dos valores das variáveis e, por isso, a estimativa da evapotranspiração é impossibilitada na ausência do valor de qualquer uma delas. Nesse sentido, o cálculo fica restrito apenas a áreas que dispõem de instrumentos de medição para todas as variáveis [16]. Em pequenas plantações, por exemplo, que dispõem de poucos recursos e têm a necessidade de planejar as atividades de irrigação do solo, a abordagem dos modelos matemáticos para estimativa da evapotranspiração torna-se inviável.

Outra limitação do método de Penman-Monteith é a estimativa da evapotranspiração apenas para áreas pequenas, sem aplicação para áreas maiores [15] . Visando superar essa limitação, existem métodos de estimativa da evapotranspiração que usam dados de sensoriamento remoto, como o método SEBAL [7].

## 2.2 Descoberta de Conhecimento em Banco de Dados

A Descoberta de Conhecimento em Bases de Dados é o processo de análise e interpretação dos dados para identificação de padrões compreensíveis, novos e potencialmente úteis, iterativo e composto pelas seguintes etapas (adaptadas de [6]), ilustradas na Figura 2:

- Pré-Processamento: etapa em que os dados são selecionados das bases de dados e uma sequência de transformações pode ser feita, como a seleção dos atributos que farão parte do conjunto de dados (*dataset*), limpeza dos dados, discretização dos valores, dentre outras.
- Mineração de Dados: Muitas vezes confundida com o processo de KDD em si, a mineração de dados permite aprendizado de modelos que reflitam os dados.
- Pós-Processamento: Avaliação e interpretação dos padrões que podem ser considerados novos conhecimentos [6].

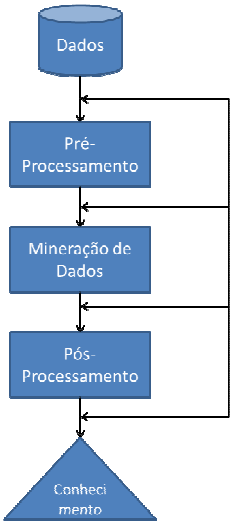


**Figura 2. Etapas do Processo de KDD – Adaptado de [6]**

Na mineração de dados, diversas técnicas podem ser usadas para aprendizado do modelo correspondente aos dados, seja para predição ou descrição:

- Classificação
- Regressão
- Análise de Séries Temporais
- Agrupamento
- Regras de Associação
- Caracterização

Essas técnicas são implementadas por algoritmos, como os baseados em redes neurais, árvores de decisão, regressão linear, dentre outros. O algoritmo M5P, por exemplo, é um algoritmo de classificação baseado em árvores de decisão [23], que aprende modelos para cálculo de valores que podem ser usados de acordo com um ponto de decisão, como o valor de algum atributo do *dataset*.

Esses algoritmos podem ser executados em ferramentas como o Weka [9] que, além da mineração de dados, automatiza as outras tarefas relativas ao processo de descoberta de conhecimento, além de possibilitar que a execução do processo seja repetida ou modificada usando diferentes configurações, como a estratégia de testes, uso de outros algoritmos, escolha da classe a ser analisada (no caso dos algoritmos de classificação), dentre outras.

O uso de KDD, ao automatizar a tarefa de descobrir um modelo que melhor se ajuste aos dados disponíveis, é uma abordagem que pode ser aplicada à estimativa da evapotranspiração, em contextos onde os dados que estão disponíveis podem ser bem diferentes de uma região para outra. Além disso, o uso de KDD é uma alternativa viável para aquelas regiões que não dispõem de todos os dados necessários da equação de referência da FAO para estimativa da evapotranspiração.

## 3. METODOLOGIA CIENTÍFICA

Segundo Recker [20], há a necessidade de se adquirir ao menos três tipos de conhecimento antes de se iniciar uma pesquisa científica:

- Conhecimento sobre o domínio ou tópico de interesse.
- Conhecimento sobre teorias relevantes sobre questões ou fenômenos.
- Conhecimento sobre os métodos de pesquisa que podem ser aplicados para se construir novos conhecimentos, artefatos ou novas questões.

Com base nisso, foram feitas revisões de trabalhos científicos sobre os assuntos relacionados ao tema de pesquisa deste trabalho. A partir dessas revisões, foram identificadas as principais questões relacionadas à estimativa da evapotranspiração, que foram usadas em novas revisões, de modo a identificar nos trabalhos relacionados possíveis problemas de pesquisa.

Com base nessa atividade, foi feita a formulação do problema de pesquisa, seguida da escolha, de acordo com os passos apresentados em [20], de qual estratégia de pesquisa a ser utilizada para se tentar solucionar o problema de pesquisa formulado. Neste trabalho, foi escolhido o método quantitativo, que pode ser definido como um conjunto de técnicas para responder questões de pesquisa com ênfase em dados quantitativos [20]. A razão para essa escolha se deve à natureza do problema de pesquisa, onde se deseja simplificar o cálculo de um componente do ciclo hidrológico.

Com a escolha do método, foi feito o planejamento da pesquisa, adaptando os procedimentos descritos por Recker [20] para esse tipo de método de pesquisa:

- Geração de modelos, teorias e hipóteses
- Desenvolvimento de instrumentos e métodos de medida
- Coleta de dados empíricos
- Modelagem Estatística
- Avaliação dos Resultados

A partir do problema de pesquisa, definiu-se a hipótese de que seria possível estimar a evapotranspiração, independentemente de quais dados estivessem disponíveis. Para teste da hipótese, definiu-se que seria executado o processo de KDD para avaliar os dados meteorológicos disponíveis na base de dados descrita na Seção 4.1 deste artigo.

Como ferramenta de apoio ao processo de KDD, utilizou-se o software Weka, que contém um conjunto de algoritmos de aprendizado de máquina para atividades de pré-processamento, mineração de dados e pós-processamento [9]. Com o auxílio de softwares como o Weka, pode- se executar facilmente para o mesmo *dataset* diversos algoritmos e executar análises como comparação dos resultados, gravação dos modelos gerados, importação de novos algoritmos, pré-processamento das instâncias, entre outras tarefas.

O experimento descrito neste trabalho se deu em três fases, relacionadas aos procedimentos descritos anteriormente e denominadas aqui de: exploração, execução e análise dos resultados. A fase de exploração compreendeu atividades como aprendizado do uso do Weka, análise das fontes de dados existentes, escolha dos atributos que iriam compor o *dataset* do experimento e uso de um *dataset* para análise dos resultados preliminares dos algoritmos de classificação. O objetivo desta fase era aprender sobre os atributos existentes nas fontes de dados e descobrir os algoritmos com melhores desempenhos para execução em outras estações de medição.

Já a fase de execução compreendeu a execução no Weka, do algoritmo escolhido na fase anterior, para mineração dos dados de *datasets* de estações de medição de diferentes regiões. O objetivo dessa fase foi gerar um modelo de estimativa da evapotranspiração para cada região analisada no estudo.

Na fase de análise, os resultados para cada região usada na mineração foram comparados com dados históricos e também com outros trabalhos realizados. O objetivo desta fase foi gerar conhecimento a partir da análise dos resultados deste experimento e discutir possíveis diferenças entre os resultados.

# 4. ESTIMATIVA DA EVAPOTRANSPIRAÇÃO ATRAVÉS DE KDD

## 4.1 Bases de Dados

Os dados foram extraídos do Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP) do Instituto Nacional de Meteorologia (INMET), que é um órgão responsável por prover informações meteorológicas, através do monitoramento, análise e previsão de tempo [13].

A escolha dessa base de dados se justifica porque contém séries históricas de dados meteorológicos para várias localidades do Brasil, coletados através de estações de medição. Além disso, desde 2006, essas séries históricas contêm dados de evapotranspiração, que foram usados como dados de entrada para os algoritmos de aprendizado.

Para o experimento descrito neste trabalho, foram selecionadas todas as estações localizadas no Estado do Rio de Janeiro, descritas na Tabela 1. Conforme mostrado na Tabela 1, uma característica importante dessas estações é que não havia uniformidade na disponibilidade dos valores das variáveis, o que serviu para o objetivo do experimento de estimar a evapotranspiração independentemente de quais variáveis tinham valores nas séries históricas.

Após a seleção da base de dados e definição das estações de medição, foi feita a extração dos dados, selecionando-se na ferramenta disponibilizada pelo INMET todas as variáveis disponíveis, já que essa ferramenta apresenta uma lista de variáveis para as quais se deseja extrair a série histórica de valores.

Essa extração gerou *datasets* com as séries históricas de cada estação que, após a seleção das variáveis na extração, continham os seguintes atributos, com as definições trazidas por Branco [1] e INMET [13]:

- Velocidade Vento Média: Média das velocidades do vento do período.
- Velocidade Vento Máxima Média: Média das velocidades máximas do vento no período.
- Evaporação Piche: Medida da evaporação - em mililitro (ml) ou em milímetros de água evaporada - a partir de uma superfície porosa, mantida permanentemente umedecida por água.
- Evapotranspiração Potencial: Evapotranspiração estimada no mês.
- Insolação Total: Número de horas que a luz solar chegou até a superfície da Terra sem interferência de nuvens.
- Nebulosidade Média: A fração da abóbada celeste que é ocupada por nuvens.
- Precipitação Total: Quantidade de chuvas do período.
- Pressão Média: É a pressão exercida pelo peso da atmosfera em um determinado ponto da superfície da Terra.
- Temperatura Máxima Média: Média das temperaturas máximas do período.
- Temperatura Compensada Média: média entre as temperaturas máximas e mínimas, além de três medidas feitas ao longo do dia (9, 15 e 21 hs).
- Temperatura Mínima Média: Média das temperaturas mínimas do período.
- Umidade Relativa Média: Porcentagem de vapor d'água que há no ar.

**Tabela 1. Estações de medição utilizadas no experimento**

| Estação | *Dataset* |
|---|---|
| Campos | Sem valores de pressão e insolação |
| Cordeiro | Sem valores de velocidade do vento |
| Itaperuna | Poucos valores de pressão |
| Paty do Alferes | Sem valores de pressão |
| Resende | Completo |
| Rio de Janeiro | Sem valores de insolação e poucos valores de pressão |

## 4.2 Avaliação do Algoritmo

Na fase de exploração, foram selecionados para avaliação todos os algoritmos de classificação disponíveis no Weka, para execução em *datasets* que contêm somente atributos numéricos.

As execuções dos algoritmos no Weka foram feitas utilizando o *dataset* da estação de Resende, por conter todas as variáveis, utilizando a estratégia de validação cruzada para particionamento dos conjuntos de treinamento e teste. Como avaliação complementar para seleção do algoritmo, utilizou-se a mesma abordagem usando-se o *dataset* da estação de Itaperuna. Em ambos os casos, o algoritmo de melhor coeficiente de correlação foi o M5P, conforme Tabela 2, que mostra os coeficientes por algoritmo para a estação de Resende:

**Tabela 2. Coeficientes de correção para os principais algoritmos de classificação executados no Weka para o *dataset* da estação de Resende-RJ**

| Algoritmo | Coeficiente de Correlação |
|---|---|
| M5P | 0.9893 |
| MLPRegressor | 0.9853 |
| MultilayerPerceptron | 0.9842 |
| RBFRegressor | 0.9840 |
| Bagging | 0.9812 |

O algoritmo M5P é uma adaptação feita por Wang e Witten [23] para o algoritmo M5, criado por Quinlan [19], que introduziu o conceito de árvores de modelos. Esse tipo de árvore é a combinação de uma árvore de decisão convencional com a possibilidade de ter, nas folhas da árvore, modelos lineares multivariados, como funções de regressão linear [19]. Na adaptação desse algoritmo para o M5P, foram incorporadas formas de tratar casos de instâncias com atributos enumerados ou sem valores, situações comuns em problemas do mundo real [23].

## 4.3 Aprendendo os Modelos Preditivos

Em seguida, após a definição do M5P como algoritmo a ser utilizado, foram feitas execuções do mesmo nos *datasets* das estações escolhidas para este experimento. A Tabela 3 mostra, para cada estação, as equações geradas pelo algoritmo:

**Tabela 3. Modelos gerados pelo algoritmo M5P para cada estação**

| Estação | Decisão | Equação |
|---|---|---|
| Campos | - | EvapoBHPotencial = 1.9879 * NM + 0.0377 * PT + 6.1418 * TMaxM + 2.2551 * TCoM + 5.7273 * TMinM - 0.7818 * URM - 198.7858 |
| Cordeiro | TCoM <= 20.252 | EvapoBHPotencial = 0.1278 * EP + 0.0134 * IT + 0.0093 * PT - 0.6426 * TMaxM + 8.7045 * TCoM - 96.1925 |
| | TCoM > 20.252 | EvapoBHPotencial = 0.312 * EP + 0.0824 * IT + 1.5226 * NM + 0.0242 * PT - 5.0562 * TMaxM + 13.2391 * TCoM + 1.883 * TMinM - 133.0846 |
| Itaperuna | TCoM <= 23.585 | EvP = 0.0381 * EP + 0.0829 * IT + 2.9742 * NM + 0.0362 * PT + 11.7057 * TCoM - 0.6848 * TMinM - 202.7576 |
| | TCoM > 23.585 | EvP = 0.1879 * EP + 0.214 * IT + 7.0959 * NM + 0.0062 * PT + 3.3705 * TCoM + 12.4856 * TMinM - 327.371 |
| Paty do Alferes | TCoM <= 19.635 | EvP = 0.03 * EP + 2.0633 * NM + 0.0115 * PT + 0.5568 * TMaxM + 5.7038 * TCoM - 2.6609 * TMinM - 0.1132 * URM + 24.3532 |
| | TCoM > 19.635 | EvP = 0.1971 * EP + 0.4788 * NM + 0.0362 * PT - 1.4569 * TMaxM + 5.5637 * TMinM - 0.6512 * URm + 102.8731 |
| Resende | TCoM <= 20.276 | EvP = 0.0388 * EP + 0.7079 * NM + 0.0068 * PT - 0.2215 * PM + 8.9358 * TCoM - 0.9358 * TMinM + 107.736 |
| | TCoM > 20.276 | EvP = 4.6224 * VMM + 0.0526 * EP + 2.6542 * NM + 0.0044 * PT - 0.8066 * PM + 11.8915 * TCoM - 0.609 * TMinM + 585.9188 |
| Rio de Janeiro | - | EvP = 0.197 * EP + 4.7078 * NM + 5.9515 * TMaxM + 6.9072 * TCoM + 2.6752 * TMinM - 333.6851 |

Onde as variáveis das equações geradas são descritas na legenda abaixo:

- EP: Evaporação Piche
- EvP: Evapotranspiração Potencial
- IT: Insolação Total
- NM: Nebulosidade Média
- PM: Pressão Média
- PT: Precipitação Total
- TCoM: Temperatura Compensada Média
- TMaxM: Temperatura Máxima Média
- TMinM: Temperatura Mínima Média
- URM: Umidade Relativa Média
- VMM: Velocidade do Vento Média

Nas figuras a seguir, são apresentados gráficos com os valores calculados pelo modelo gerado e os valores históricos dos *datasets* do INMET para cada uma das seis estações analisadas neste experimento. Para cada gráfico, executou-se a regressão linear para medir o grau de precisão entre os valores históricos e previstos, gerando-se o coeficiente $R^2$ e a função que relaciona os dois valores. O coeficiente de determinação, também chamado de $R^2$, é uma medida de ajustamento de um modelo estatístico linear generalizado, como a regressão linear, em relação aos valores observados. O $R^2$ varia entre 0 e 1, indicando, em percentagem, quanto o modelo consegue explicar os valores observados. Quanto maior o $R^2$, mais explicativo é o modelo, melhor ele se ajusta à amostra [26].

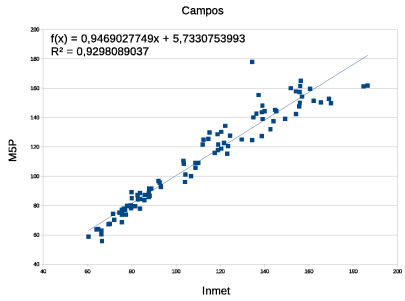


**Figura 3. Evapotranspiração para a estação de Campos-RJ**

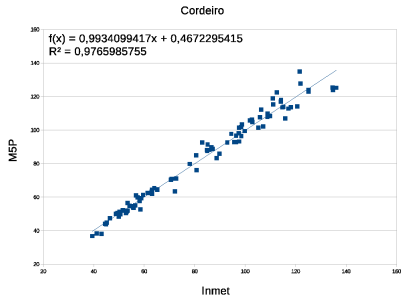


**Figura 4. Evapotranspiração para a estação de Cordeiro-RJ**

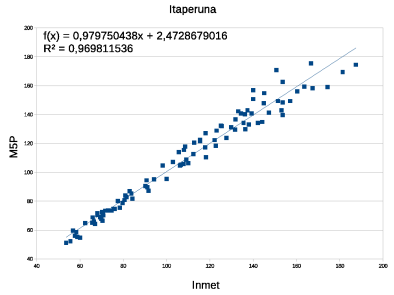


**Figura 5. Evapotranspiração para a estação de Itaperuna-RJ**

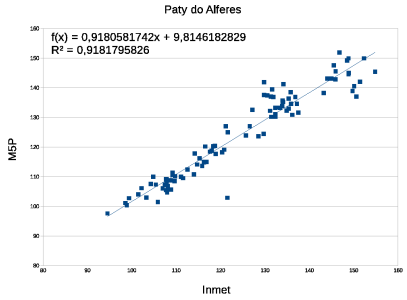


**Figura 6. Evapotranspiração para a estação de Paty do Alferes-RJ**

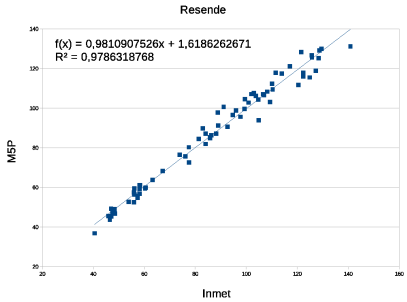


**Figura 7. Evapotranspiração para a estação de Resende-RJ**

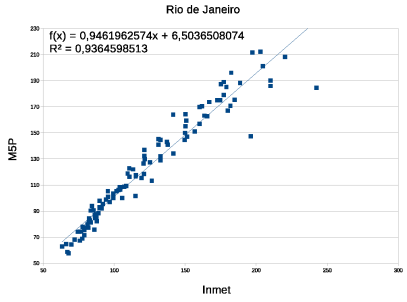


**Figura 8. Evapotranspiração para a estação do Rio de Janeiro-RJ**

## 5. ANÁLISE DOS RESULTADOS

A execução do algoritmo M5P gerou, como resultado, equações para cálculo da evapotranspiração potencial. Nas estações para as quais foram geradas duas equações, nota-se que a temperatura compensada média (TCoM) é o fator de decisão sobre qual equação utilizar. Isso demonstra a importância desse atributo para cálculo da evapotranspiração potencial usando a abordagem proposta neste experimento.

Para duas estações (Campos e Rio de Janeiro), a execução do M5P gerou apenas uma equação. A característica dos *datasets* dessas estações em relação às demais é a ausência de valores em dois atributos: pressão e insolação. O *dataset* da estação do Rio de Janeiro continha apenas algumas instâncias com valores de pressão preenchidos.

Nota-se, na Tabela 3, que nenhuma das equações geradas usou todos os atributos. A explicação para esse fato é que o algoritmo M5P tem, como uma de suas características, a simplificação dos modelos lineares gerados, usando uma estratégia de busca gulosa para identificar e remover todos os atributos que contribuem pouco para o modelo, podendo até substituí-los por uma constante [19].

A única estação que continha todos os atributos preenchidos foi a de Resende, cujo valor de coeficiente de determinação ($R^2$) foi o mais alto, conforme a Figura 7. No entanto, os resultados para estação de Cordeiro apresentaram um desempenho semelhante, conforme a Figura 5. A característica do *dataset* dessa estação é a ausência de valores da velocidade do vento, conforme descrito na Tabela 1.

A semelhança entre os resultados para essas duas estações e o fato de que os valores de velocidade do vento só foram usados em uma única equação gerada entre todas as estações, é uma evidência de essa variável pode não ter relevância no cálculo da evapotranspiração potencial, na abordagem utilizada neste experimento. Na equação de Penman-Monteith, a velocidade do vento é uma das variáveis requeridas, o que invalidaria o uso deste método para a estação de Cordeiro.

## 6. TRABALHOS RELACIONADOS

Em estudo relacionado à agricultura, Haghverdi et al [8] buscaram analisar comparativamente a mineração de dados com outras funções matemáticas conhecidas, para avaliar a produção de grãos baseados em dados de irrigação e evapotranspiração. Os autores identificaram que, nesse campo, os dados de evapotranspiração e salinidade do solo, que são difíceis de medir, podem ser substituídos por outras variáveis como quantidade da água irrigada e salinidade da água irrigada. Mostraram que a capacidade dos métodos de mineração utilizados (árvores de decisão e redes neurais) para introduzir novos preditores de entrada é uma vantagem, em relação aos tradicionalmente utilizados nesse campo.

Com o foco específico na estimativa da evapotranspiração, em [11] e [12] são descritas, entre outras coisas, a importância do uso de dados de evapotranspiração para o planejamento da irrigação de solos. Usaram processos gaussianos para entender a relação entre cada atributo e o valor da evapotranspiração. Entre seus resultados, mostraram que as redes neurais e processos gaussianos trouxeram maior acurácia do que os métodos de regressão linear.

Os trabalhos de Zhou et al [24] e Tang & Li [22] usaram o método Surface Energy Balance Algorithm for Land (SEBAL), que é baseado em dados de imagens de satélite, para comparação com outros métodos de estimativa da evapotranspiração. Já em [10], foi usado o método para estimar a evapotranspiração para uso em um modelo para estimativa da umidade do solo. Segundo Zhu [25], o método SEBAL é adequado para estimativa da evapotranspiração para grandes regiões e, nesse trabalho, o método foi usado em todo o delta do Rio Amarelo, na China.

Alguns trabalhos fizeram uso de redes neurais para estimativa da evapotranspiração, como o feito por El-Shafie el al [5], que modificou uma rede neural e usou apenas dados das temperaturas máximas e mínimas para essa estimativa. Nesse trabalho, a rede neural foi treinada usando dados da evapotranspiração calculados a partir da equação de referência da FAO.

Por fim, o trabalho de Shiri et al [21] indica que um problema dos modelos gerados por mineração de dados sobre dados climáticos é que não podem ter sua performance inferida fora do local de treinamento/teste do modelo. Os autores propuseram um modelo de programação genética para estimar a evapotranspiração, que pode ser usado no local de treinamento ou fora dele, o que pode ser uma vantagem quando os dados não estiverem disponíveis em uma dada estação de medição.

Um aspecto comum a uma parte dos trabalhos relacionados nesta seção é a motivação usada para buscar outros métodos para estimar a evapotranspiração: conforme declarado em [4], o método de Penman-Monteith requer uma quantidade de variáveis que podem não estar disponíveis para o local de uso. Além disso, de acordo com a revisão feita em [15], pode não ser adequado para ser usado em áreas grandes, devido ao fato de não ser prático obter todos os dados necessários para o cálculo.

Na abordagem proposta no presente trabalho, a estimativa da evapotranspiração foi feita independentemente da disponibilidade de uma variável, garantindo flexibilidade no uso das variáveis. Além disso, através do uso de KDD, essa abordagem garante simplificação na estimativa da evapotranspiração, sem o uso de técnicas mais complexas, como as utilizadas pelo método SEBAL.

## 7. CONSIDERAÇÕES FINAIS

Pelo fato da equação de referência da FAO para estimativa da evapotranspiração ser dependente da disponibilidade dos valores de todas as variáveis da equação, buscou-se no presente trabalho uma forma de fazer essa estimativa independentemente de quais valores estejam disponíveis. Para alcançar esse objetivo, foi proposto o uso de descoberta de conhecimento em banco de dados, utilizando-se dados meteorológicos das seis estações de medição do INMET no Estado do Rio de Janeiro.

Para avaliação da solução proposta, foi feita uma comparação dos dados históricos da evapotranspiração potencial, disponíveis nos *datasets* utilizados, com os valores calculados pelos modelos aprendidos na execução do processo de KDD. A medida usada para avaliação para essa comparação foi o coeficiente de determinação ($R^2$), que varia entre 0 e 1, onde 1 indica uma correspondência de 100% entre os valores.

Nos *datasets* de cinco das seis estações usadas neste experimento, pelo menos um atributo estava sem os dados preenchidos. Uma alternativa seria preencher esses dados históricos através de modelos matemáticos tradicionais ou também usar mineração de dados para gerar um modelo de cálculo desses atributos. Optou-se, neste experimento, por não preencher os valores desses atributos vazios para analisar o comportamento do algoritmo nesses cenários.

Como resultado, para todas as estações, a medida de avaliação ($R^2$) esteve acima de 0.9, o que indica uma correspondência de mais de 90% entre o valor calculado e o valor histórico. Com este resultado, foi possível demonstrar que, com o uso de KDD, a evapotranspiração pode ser estimada com boa precisão, mesmo em cenários em que nem todos os valores estão disponíveis.

A abordagem utilizada neste trabalho é a sua principal contribuição, visto que o modelo é gerado a partir dos dados disponíveis e não o inverso. Dessa forma, a abordagem poderia ser utilizada para estimar a evapotranspiração em regiões que não dispõem de instrumentos de medição para todas as variáveis da equação de referência da FAO para estimativa da evapotranspiração.

Como outra contribuição, as equações geradas neste experimento podem ser utilizadas em trabalhos que necessitem de valores estimados da evapotranspiração como atributos de entrada para o cálculo de outra variável. Modelos como os gerados neste experimento também poderiam ser usados em cenários simulados, como os mantidos pelo IPCC.

Outra contribuição deste trabalho é que, para estações de medição com o foco no cálculo da evapotranspiração potencial, não seria necessário dispor de instrumentos de medição para todos os atributos meteorológicos, o que garantiria uma redução de custos na instalação de novas estações de medição. Pela característica do algoritmo de eliminar os atributos que contribuem pouco para o modelo, indicando apenas os que são realmente relevantes, têm-se informação de quais medidas realmente são necessárias para a estimativa da evapotranspiração.

Como trabalho futuro, a abordagem usada neste experimento poderia ser aplicada para estações de outras regiões do País. Uma questão a ser verificada é o desempenho do algoritmo escolhido para as estações do Estado do Rio de Janeiro em estações de outros estados do País, ou mesmo em outros países, onde as condições climáticas podem ser bem diferentes.

Outra possibilidade de trabalho futuro é o agrupamento de estações de acordo com algumas características. Esse agrupamento poderia ser usado para verificar se o modelo gerado para uma estação seria adequado para outra estação do mesmo grupo.

## 8. AGRADECIMENTO

Este trabalho foi parcialmente financiado pela FAPERJ (número do projeto: E-26/110.477/2014)

## 9. REFERÊNCIAS

[1] Branco, P. M. (2014). Elementos que caracterizam o clima. Disponível http://www.cprm.gov.br/publique/cgi/cgilua.exe/sys/start.htm?infoid=1267&sid=129 [acessado em 28-Novembro-2014].

[2] Di Bello, R. C. (2005). Análise do Comportamento da Umidade do Solo no Modelo Chuva-Vazão Smap II–Versão com Suavização Hiperbólica Estudo de Caso: Região de Barreiras na Bacia do Rio Grande-BA (Dissertação de Mestrado, Universidade Federal do Rio de Janeiro).

[3] Eagleson, P. S. (1994). The evolution of modern hydrology (from watershed to continent in 30 years). Advances in water resources, 17(1), 3-18.

[4] El-Shafie, A., Najah, A., Alsulami, H. M., & Jahanbani, H. (2014). Optimized neural network prediction model for potential evapotranspiration utilizing ensemble procedure. Water Resources Management, 28(4), 947-967.

[5] FAO (2014). Organização das Nações Unidas para Alimentação e Agricultura. Chapter 2 - fao penman-monteith equation. [Online; accessado em 10-Janeiro-2015].

[6] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

[7] Ferreira, A., & Meirelles, M. (2011). Implementação preliminar do modelo SEBAL para estimativa da evapotranspiração na Mesorregião do Sul Goiano. Anais XV Simpósio Brasileiro de Sensoriamento Remoto.

[8] Haghverdi, A., Ghahraman, B., Leib, B. G., Pulido-Calvo, I., Kafi, M., Davary, K., & Ashorun, B. (2014). Deriving data mining and regression based water-salinity production functions for spring wheat (Triticum aestivum). Computers and Electronics in Agriculture, 101, 68-75.

[9] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.

[10] Hendrickx, J. M., Pradhan, N. R., Hong, S. H., Ogden, F. L., Byrd, A. R., & Toll, D. (2009, May). Improvement of hydrologic model soil moisture predictions using SEBAL evapotranspiration estimates. In SPIE Defense, Security, and Sensing (pp. 730311-730311). International Society for Optics and Photonics.

[11] Holman, D., Sridharan, M., Gowda, P., Porter, D., Marek, T., Howell, T., & Moorhead, J. (2013, August). Estimating reference evapotranspiration for irrigation management in the texas high plains. In Proceedings of the Twenty-Third international joint conference on Artificial Intelligence (pp. 2819-2825). AAAI Press.

[12] Holman, D., Sridharan, M., Gowda, P., Porter, D., Marek, T., Howell, T., & Moorhead, J. (2014). Gaussian process models for reference ET estimation from alternative meteorological data sources. Journal of Hydrology, 517, 28-35.

[13] INMET (2014). Instrumentos meteorológicos. Disponível em http://www.inmet.gov.br/html/informacoes/sobre meteorologia/instrumentos (acessado em 28-Novembro-2014).

[14] IPCC (2014). Intergovernmental panel on climate change, Disponível em http://www.ipcc.ch (acessado em 28-Novembro-2014).

[15] Liou, Y. A., & Kar, S. K. (2014). Evapotranspiration estimation with remote sensing and various surface energy balance algorithms - A review. Energies, 7(5), 2821-2849.

[16] Manesh, S. S., Ahani, H., & Rezaeian-Zadeh, M. (2014). ANN-based mapping of monthly reference crop evapotranspiration by using altitude, latitude and longitude data in Fars province, Iran. Environment, development and sustainability, 16(1), 103-122.

[17] MMA (2015). Ciclo Hidrológico, Ministério do Meio Ambiente. Disponível em http://www.mma.gov.br/agua/recursos-hidricos/aguas-subterraneas/ciclo-hidrologico. (acessado em 06-março-2015)

[18] Oliveira, M. D., & Carvalho, D. D. (1998). Estimativa da evapotranspiração de referência e da demanda suplementar de irrigação para o milho (Zea mays L.) em Seropédica e Campos, Estado do Rio de Janeiro. Revista Brasileira de Engenharia Agrícola e Ambiental, 2(2), 132-135.

[19] Quinlan, J. R. (1992, November). Learning with continuous classes. In 5th Australian joint conference on artificial intelligence (Vol. 92, pp. 343-348).

[20] Recker, J. (2012). Scientific research in information systems: a beginner's guide. Springer Science & Business Media.

[21] Shiri, J., Sadraddini, A. A., Nazemi, A. H., Kisi, O., Landeras, G., Fard, A. F., & Marti, P. (2014). Generalizability of Gene Expression Programming-based approaches for estimating daily reference evapotranspiration in coastal stations of Iran. Journal of Hydrology, 508, 1-11.

[22] Tang, R., & Li, Z. L. (2015). Evaluation of two end-member-based models for regional land surface evapotranspiration estimation from MODIS data. Agricultural and Forest Meteorology, 202, 69-82.

[23] Wang, Y., & Witten, I. H. (1996). Induction of model trees for predicting continuous classes.

[24] Zhou, X., Bi, S., Yang, Y., Tian, F., & Ren, D. (2014). Comparison of ET estimations by the three-temperature model, SEBAL model and eddy covariance observations. Journal of Hydrology, 519, 769-776.

[25] Zhu, M., Hou, X., Lu, X., & Li, M. (2011, June). Spatial-temporal characters of evapotranspiration in the Yellow River Delta. In Spatial Data Mining and Geographical Knowledge Services (ICSDM), 2011 IEEE International Conference on (pp. 409-412). IEEE.

[26] Wikipedia (2015). Definição de $R^2$. Disponível em http://pt.wikipedia.org/wiki/R%C2%B2 (acessado em 06-março-2015