RBRH – Revista Brasileira de Recursos Hídricos Volume 12 n.3 Jul/Set 2007, 83-92

Aplicação das Técnicas de Mineração de Dados como Complemento às Previsões Estocásticas Univariadas de Vazão Natural: Estudo de Caso para a Bacia do Rio Iguaçu

Marcio Cataldi; Bruno Goulart de Freitas Machado; Simone Borim da Silva; Luiz Guilherme Ferreira Guilhon
Operador Nacional do Sistema Elétrico (ONS)
cataldi@ons.org.br

Carla da C. Lopes Achão
Planning Engenharia e Consultoria

Recebido: 05/12/06 - revisado: 30/04/07 - aceito: 05/07/07

RESUMO

Este trabalho apresenta os resultados obtidos a partir da aplicação de técnicas de mineração de dados e de Redes Neurais com treinamento bayesiano, para o balizamento da escolha da melhor previsão de vazões naturais do sistema de modelos estocásticos PREVIVAZ. Para aplicação desta técnica, foram utilizadas informações de precipitação observada e prevista, além das vazões naturais verificadas nas últimas semanas que antecederam a previsão. O estudo foi realizado para os aproveitamentos hidrelétricos de Foz do Areia e Salto Santiago na bacia do rio Iguaçu. Os resultados obtidos indicam que a utilização desta ferramenta pode ser uma solução simples e eficaz para a diminuição dos erros de previsão em horizonte semanal de vazão natural nesta bacia.

Palavras-chave: *Data Mining; Redes Bayesianas; Modelos Estocásticos; Previsão de vazões.*

HISTÓRICO

A partir do I ENCONTRO TÉCNICO SOBRE PREVISÕES DE SÉRIES TEMPORAIS – SCEN, em maio de 1993, algumas empresas do setor elétrico iniciaram o processo de contratação do desenvolvimento do modelo PREVIVAZ junto ao CEPEL. Naquela ocasião, as empresas participantes do GCOI (Grupo Coordenador para a Operação Interligada) financiaram seu desenvolvimento, que ocorreu entre o segundo semestre de 1994, sendo entregue a versão 1.1 em dezembro de 1996. A partir desse momento, o PREVIVAZ foi utilizado e foi se aperfeiçoando até culminarem, em março de 1998, os testes para sua validação com a posterior elaboração do relatório intitulado "Modelo PREVIVAZ - Testes Finais de Validação – Agosto/1998". Esse relatório foi aprovado pelo GCOI – Resolução RS-G-2946/98 em 06/10/98. Em janeiro de 1999, o modelo foi implantado para todas as bacias do Sistema Interligado Nacional pelo ONS e passou a ser utilizado no Programa Mensal de Operação Eletroenergética (PMO) a partir de fevereiro de 1999. Em 11/04/2000, o ONS apresentou uma comparação entre o modelo PREVAZ (modelo estocástico em base mensal utilizado até então pela ELETROBRÀS) durante todo o ano de 1999, e o modelo PREVIVAZ (Horizonte Semanal), tornando-se um consenso a decisão de se substituir a previsão semanal, até então obtida pela desagregação da vazão mensal do modelo PREVAZ, pela previsão de vazões semanais calculadas pelo modelo PREVIVAZ, a partir do PMO de maio de 2000.

O PREVIVAZ é um modelo de previsão de vazões médias semanais constituído de um conjunto de alternativas de metodologias para previsão de vazões para um horizonte de até seis semanas, em base estocástica, utilizando combinações de modelos estatísticos estacionários ou periódicos, com diferentes métodos de estimação de parâmetros e diferentes tipos de transformações.

APRESENTAÇÃO DO PROBLEMA

As metodologias estocásticas contidas no modelo PREVIVAZ [CEPEL, 2004] contemplam os modelos autoregressivos e de médias móveis, com estrutura estacionária ou periódica, ou seja, os modelos AR(p) e PAR(p), com "p" de até ordem 4, e

PARMA(p,q) e ARMA (p,q), com "p" de até ordem 3 e q de ordem 1. As transformações podem ser logarítmica, Box & Cox ou sem transformação [Guilhon, 2003]. Os métodos de estimação de parâmetros se baseiam no método da máxima verossimilhança e são o método dos momentos, o método de regressão simples e o de regressão em relação à origem das previsões.

O PREVIVAZ divide o histórico em duas metades, estimando, para cada semana, os parâmetros de todos os modelos para a primeira metade, e verificando o erro médio quadrático da previsão com os dados da segunda metade do histórico, conforme a equação (1) a seguir.

$$\sqrt{\frac{\sum_{i=1}^{N} (X^i_{prev} - X^i_{obs})^2}{N}} \qquad (1)$$

onde,

X^i_{prev} – Vazão prevista no instante i.

X^i_{obs} – Vazão observada no instante i.

N – número total de semanas da metade do histórico considerada.

Em seguida, de forma análoga, o PREVIVAZ estima os parâmetros de todos os modelos para cada semana da segunda metade da série e verifica o erro médio quadrático da previsão com os valores da primeira metade do histórico. Calcula-se então a média dos erros médios quadráticos de cada modelo para as duas metades do histórico, e é feita uma ordenação de modo a escolher aquela metodologia que apresente o menor valor médio do erro médio quadrático.

Após a escolha do melhor modelo, o PREVIVAZ estima novamente os parâmetros, considerando agora todas as semanas do histórico e passa a utilizar este modelo com os novos parâmetros calculados para cada semana específica.

Esta modelagem, no entanto, não incorpora informações de precipitação na bacia, sejam estas de precipitação observada ou prevista, as quais são fundamentais na composição da vazão natural afluente.

Neste artigo será mostrado que é possível melhorar o desempenho do sistema PREVIVAZ, mantendo sua característica univariada, a partir da utilização de um critério de seleção capaz de balizar a escolha do melhor modelo deste sistema. Esta escolha ocorre dentro de uma faixa de vazões previstas por um sistema auxiliar, que incorpora outras informações, tais como os dados previstos e observados de precipitação na bacia.

Desta forma, buscou-se por meio da utilização das técnicas de mineração de dados (*Data Mining*), uma classificação para as previsões de vazões naturais de dois aproveitamentos localizados na bacia do rio Iguaçu, a saber: UHE Foz do Areia e Salto Santiago.

Para tal, utilizou-se o aplicativo *WEKA*[1] *Data Mining – Waikato Environment for Knowledge Analysis*, *software* de domínio público e grande portabilidade, implementado em linguagem Java, que possibilita a aplicação de um grande número de tecnologias distintas para o estudo de classificação.

Dentre as técnicas empregadas pelo *software* WEKA para a previsão das faixas de vazão natural, foram testadas aquelas baseadas em Redes Neurais (RN) e Inteligência Artificial (IA), tais como árvores de decisão ID3 e J48; RN do tipo *Multi Layer Perceptron* e *Lazy*, classificadores baseados em regras de associação, além de RN com treinamento bayesiano e algoritmo de automatização de procura baseado na técnica *Hill-Climbing* (subida da encosta ou gradiente). Esta última técnica foi a que apresentou, em todos os experimentos, os melhores resultados. Uma breve descrição do classificador bayesiano disponível no *software WEKA*, utilizado neste estudo, será apresentada a seguir.

TEORIA BAYESIANA

Esse teorema tem como premissa que os itens e relações de interesse são a manifestação de leis de distribuição de probabilidade. É, portanto, uma abordagem essencialmente quantitativa, olhando para o problema como a escolha da melhor hipótese de um espaço de hipóteses, ou seja, aquela que é mais coerente com os dados do problema. [Friedman et al. 1997]

O teorema de Bayes, sendo um dos resultados mais importantes da teoria das probabilidades, é também o princípio fundamental da aprendizagem bayesiana.

Esse teorema pode ser resumido como:

Teorema: Se {A1, A2, ..., Am} é uma partição do espaço de resultados e B um acontecimento qualquer, com P(B)>0, e para cada i P(Ai)>0, então:

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{\sum_{i=1}^{m} P(A_i)P(B \mid A_i)} \qquad (2)$$

[1] http://www.cs.waikato.ac.nz/ml/weka/

$i \in \{1,...,m)$

onde:
P(A) é a probabilidade de ocorrência do acontecimento A
P(A | B) é a probabilidade de A condicionada por B, definida por P (A∩B)/P(B)

De uma maneira geral podemos entender a P(A | B) como sendo a probabilidade do evento A ocorrer, tendo em vista que o evento B já ocorreu.

Uma conseqüência imediata deste teorema pode ser aplicada para dois acontecimentos A e B, tais P(A)>0 e P(B)>0. Neste caso pode-se assumir então que:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \qquad (3)$$

Em um treinamento de redes bayesianas, supomos que o nosso conjunto de dados (instâncias de treino) é designado por D, então pelo teorema anterior, temos uma forma de calcular a probabilidade de ocorrência de uma hipótese h, tendo por base os dados do treinamento, através da seguinte relação:

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)} \qquad (4)$$

Nesse caso a probabilidade P(h | D) é a denominada "probabilidade à posteriori" de ocorrência de um evento em h, dado que tenha ocorrido um determinado evento dentro do conjunto de dados D. O termo P(h) é a "probabilidade à priori" da hipótese h. A probabilidade "à priori" é a probabilidade de ocorrência não condicionada ao treinamento, e somente ao conjunto de dados que compõe o treinamento. Ela é calculada avaliando-se a probabilidade inicial de ocorrência de cada classe dentro do conjunto de treinamento.

Considerando que temos um espaço de hipóteses possíveis H, então em uma rede bayesiana, pretende-se determinar qual a melhor hipótese "à posteriori", levando-se em consideração o conjunto de dados observados D. Se interpretarmos a melhor hipótese como sendo a mais provável, atendendo ao conjunto de dados observados D, isto é, a hipótese com melhor valor de "probabilidade à posteriori", então o valor procurado, deve ser:

$$h_{MAP} = \arg\max P(h \mid D) \qquad (5)$$

onde h ∈ H.

Na equação 5, então, podemos obter o valor de maior probabilidade de ocorrência para um evento "à posteriori" (h_{MAP}), levando-se em consideração o treinamento bayesiano de probabilidades.

O classificador bayesiano irá criar um conjunto de tabelas de probabilidade organizadas em formato árvore, unidas através de diversos nós que formam um conjunto acíclico de busca, conforme apresentado no exemplo da Figura 1. Nesta figura, os nós representam as variáveis de domínio e os arcos as relações de dependência probabilística direta entre as variáveis que as conectam.

A probabilidade de cada tabela seguirá a configuração da rede bayesiana, de modo que o nosso conjunto de variáveis D seja formado por k variáveis, onde D = {xi, ..., xk}, com k>1. A probabilidade bayesiana (P_{BS}) de ocorrência de cada tabela será então [Bouckaert 2004]:

$$P_{BS} = \{P(d \mid pa(d)) \mid d \in D\} \qquad (6)$$

onde pa(d) é a probabilidade de cada subconjunto de d que compõe a tabela.

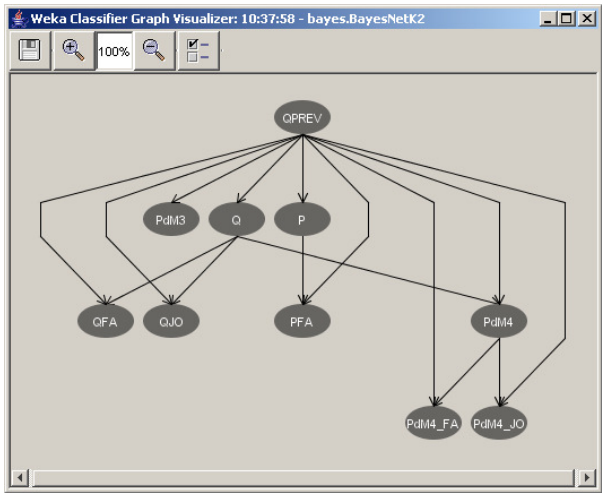

Figura 1 - Exemplo de "árvore de probabilidades bayesianas" para o trecho da bacia do rio Iguaçu a montante da UHE Salto Santiago.

A maior probabilidade bayesiana, ou seja, a classe mais provável de ocorrer, será encontrada através da busca acíclica realizada em todas as tabelas das k variáveis. Essa busca é feita utilizando-se o algoritmo "Hill Climbing". Maiores detalhes podem ser encontrados em [Buntine 1996].

Vale ressaltar que para a solução da equação 6, o classificador bayesiano necessita de três "hiperparâmetros", sendo que somente um deles, o hiperparâmetro α, é ajustável na versão de classificador bayesiano disponível no Weka. De acordo com o valor atribuído a esse parâmetro, será determinado o peso que cada tabela de probabilidades terá na escolha da classe de maior probabilidade de ocorrência. Maiores detalhes sobre o classificador bayesiano podem ser encontrados em Witten e Frank (2000).

DESCRIÇÃO E JUSTIFICATIVA DOS DADOS

Com o objetivo de obter um desempenho satisfatório nos testes com o *software* WEKA, foram testadas inúmeras configurações, a partir da combinação das variáveis de vazão (observada e prevista) e precipitação (observada e prevista). Cabe notar que as variáveis de precipitação previstas incluídas neste estudo correspondem àquelas geradas pelo modelo ETA [Black, 1994] do Centro de Previsão do Tempo e Estudos Climáticos – CPTEC, referentes aos aproveitamentos da bacia do rio Iguaçu (Foz do Areia e Salto Santiago).

Os dados utilizados compreenderam o período entre 1994 e 2003. A fim de validar a metodologia proposta foram testados dois anos: 2002 e 2003. O período de treinamento correspondente ao teste de 2002 foi de 1994 a 2001. Para se testar o ano de 2003, utilizou-se o período de treinamento compreendido entre 1994 e 2002.

Para compor a precipitação observada para o trecho a montante de Foz do Areia, foram utilizados dados de doze postos pluviométricos, ao passo que para o trecho a montante de Salto Santiago, foram utilizados dados de apenas dois postos.

A distribuição dos postos pluviométricos, dos pontos de grade do modelo ETA e das principais usinas da bacia do rio Iguaçu pode ser observados na Figura 2.

É importante notar que para cada aproveitamento obteve-se o melhor desempenho no software de *Data Mining* utilizando uma determinada configuração, isto é, a partir do uso de diferentes conjuntos de variáveis, observando a dependência de cada uma delas em relação às demais variáveis estudadas. A definição da configuração ótima dos aproveitamentos – que possibilitou o melhor desempenho nos testes com *Data Mining* – resultou da aplicação de um conjunto de testes, quando foram avaliadas várias combinações dentre as variáveis disponíveis.

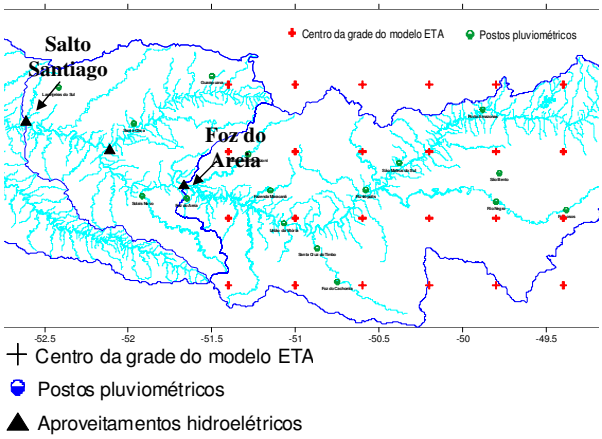

Figura 2 – Topografia da bacia do rio Iguaçu desde a cabeceira até a UHE Salto Santiago.

A escolha das variáveis foi feita através de análises estatísticas padrões nos estudos de *Data Mining*, a saber: matriz correlação, dendogramas e Análise de Componentes Principais (ACP). As faixas de classe destas variáveis foram escolhidas a partir da análise da curva de permanência de cada uma delas, com o objetivo de se obter classes que pudessem caracterizar principalmente períodos de cheia e de recessão, períodos estes onde os erros do modelo PREVIVAZ são maiores, devido principalmente ao fato do modelo não incorporar as informações de precipitação.

Serão apresentadas a seguir algumas figuras utilizadas nas análises realizadas para o trecho da bacia do rio Iguaçu a montante da UHE Foz do Areia. Cabe ressaltar que análises semelhantes foram realizadas para que se pudesse obter a composição final das variáveis apresentadas nas figuras 4 a 7.

A partir dessas análises e da realização de inúmeros testes com o conjunto de treinamento, obteve-se a configuração que possibilitou o melhor desempenho para Foz do Areia. As variáveis descritas na tabela 1.

A precipitação média foi obtida pelo método de Kriging. A metodologia de cálculo, bem como a análise e discussão de seu uso foram abordadas por Cataldi e Machado (2004).

Tabela1 - Variáveis utilizadas para a UHE Foz do Areia.

Sigla	Significado
Q_1	Vazão natural média observada na semana anterior à semana da previsão (m^3/s)
Q	Vazão natural média observada na semana da previsão (m^3/s)
QUV	Vazão natural média observada no posto de União da Vitória na semana da previsão (m^3/s)
QPUV	Vazão natural média prevista para o posto de União da Vitória na semana seguinte à semana da previsão (m^3/s)
Q_PREV ou QM1	Vazão natural média prevista para a semana seguinte à semana da previsão (m^3/s)
P**	Precipitação diária acumulada em 7 dias observada na semana da previsão (mm).
PdM4*	Previsão de Precipitação acumulada para os próximos 4 dias a partir da data da previsão (mm)
PdM3*	Previsão de Precipitação acumulada do 4º ao 7º dia a partir da data da previsão (mm)

*Nos testes realizados considerando a previsão "perfeita" de precipitação essas variáveis foram compostas pelos valores de precipitação observada na semana a ser prevista.

** Nos testes que utilizaram a previsão de precipitação do modelo ETA, os últimos 3 dias dessa variável foram compostos com a previsão de precipitação, visando completar a semana operativa.

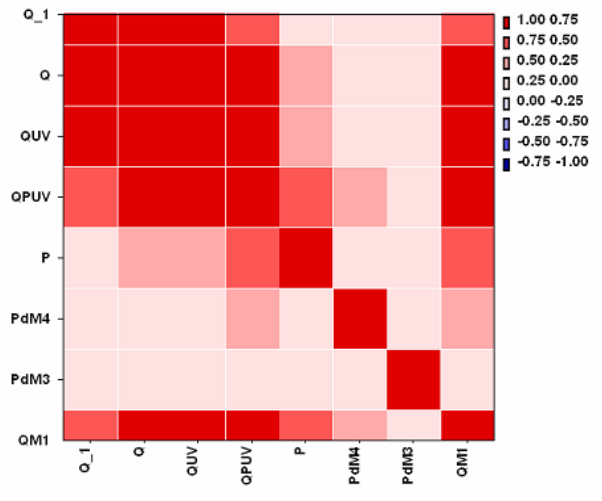

Figura 3 – Matriz correlação.

Na Figura 3 pode-se observar a correlação entre as variáveis usadas no estudo. O dendograma, que representa os agrupamentos obtidos, é mostrado na Figura 4. As análises de componentes principais por variável e por período do histórico são apresentadas nas figuras 5 e 6.

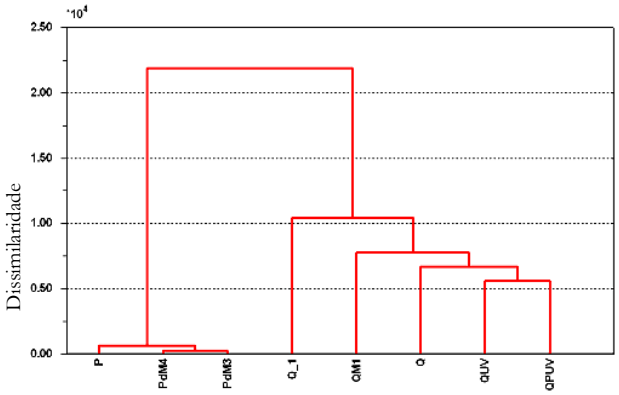

Figura 4 – Dendograma.

A análise das figuras 5 e 6 nos faz perceber que o conjunto de componentes principais formado somente pelas vazões, em geral, só é capaz de representar todo o conjunto de dados nos casos onde ocorrem pequenas variações entre as vazões naturais semanais.

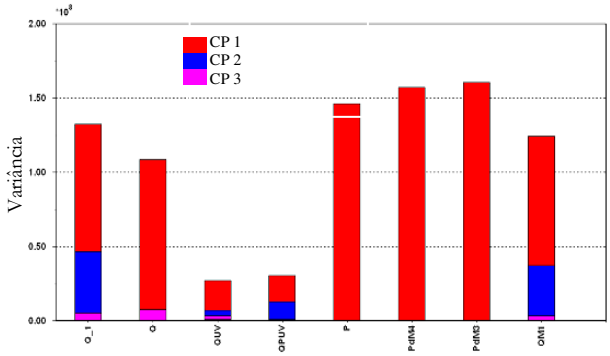

Figura 5 – Análise de Componentes Principais por variável.

As faixas que possibilitaram o melhor desempenho do classificador bayesiano foram obtidas a partir de curvas de distribuição de probabilidades (Figura 7), e estão apresentadas na figura 8.

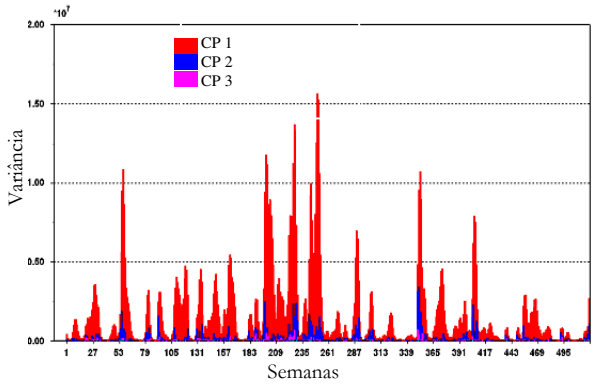

Figura 6 – Distribuição das Componentes Principais ao longo de toda a série histórica.

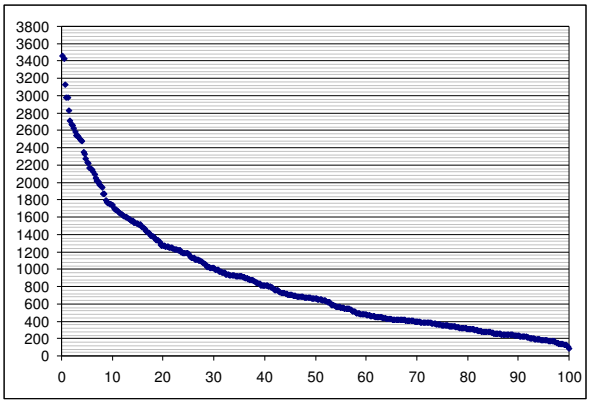

Figura 7 – Curva de permanência em porcentagem das vazões naturais totais a UHE Foz do Areia (m^3/s)

Q (m^3/s)	Pob_4d (mm)	Pob_3d (mm)	Q_1 (m^3/s)
Q ≤ 160	PdM4 ≤ 5	PdM3 ≤ 5	Q_1 ≤ 160
160 < Q ≤ 300	5 < PdM4 ≤ 20	5 < PdM3 ≤ 15	160 < Q_1 ≤ 300
300 < Q ≤ 470	20 < PdM4 ≤ 40	15 < PdM3 ≤ 30	300 < Q_1 ≤ 470
470 < Q ≤ 650	40 < PdM4 ≤ 60	30 < PdM3 ≤ 50	470 < Q_1 ≤ 650
650 < Q ≤ 950	PdM4 > 60	PdM3 > 50	650 < Q_1 ≤ 950
950 < Q ≤ 1220			950 < Q_1 ≤ 1220
Q > 1220			Q_1 > 1220
P (mm)	**PdM4 (mm)**	**PdM3 (mm)**	**QPREV (m^3/s)**
P ≤ 5	PdM4 ≤ 5	PdM3 ≤ 5	QPREV ≤ 160
5 < P ≤ 20	5 < PdM4 ≤ 20	5 < PdM3 ≤ 15	160 < QPREV ≤ 300
20 < P ≤ 40	20 < PdM4 ≤ 40	15 < PdM3 ≤ 30	300 < QPREV ≤ 470
40 < P ≤ 60	40 < PdM4 ≤ 60	30 < PdM3 ≤ 50	470 < QPREV ≤ 650
P > 60	PdM4 > 60	PdM3 > 50	650 < QPREV ≤ 950
			950 < QPREV ≤ 1220
			QPREV > 1220

Figura 8 – Faixas de vazões e precipitação utilizadas para a UHE Foz do Areia

Observando a Figura 3 podemos perceber que a previsão de precipitação dos últimos 3 dias da semana a ser prevista (PDM3) foi a variável que apresentou menor correlação com a vazão da semana a ser prevista (QM1), porém, ela contém informações importantes nos casos onde a variação nos valores de vazão entre as semanas observada e prevista é grande, como pode ser observado na ACP (Figura 5). Observa-se na Figura 4, como era de se esperar, dois grandes grupos formados pelas variáveis analisadas: um formado pelas variáveis de precipitação e outro pelas variáveis de vazão natural.

A ACP apresentada na Figura 5 indica que com os 3 Componentes Principais (CP) encontrados, cerca de 98 % da série poderia ser explicada. A ACP é uma técnica estatística que pode ser utilizada para redução do número de variáveis e para fornecer uma visão estatisticamente privilegiada do conjunto de dados. A ACP consiste em reescrever as variáveis originais em novas variáveis denominadas Componentes Principais - CP, através de uma transformação de coordenadas. Os CP são as novas variáveis geradas através de uma transformação matemática especial realizada sobre as variáveis originais. Cada CP é uma combinação linear de todas as variáveis originais. Nestas combinações, cada variável terá uma importância ou peso diferente.

As variáveis podem guardar entre si correlações que são suprimidas nos CP, ou seja, os componentes principais são ortogonais entre si. Deste modo, cada componente principal traz uma informação estatística diferente das outras. As variáveis originais têm a mesma importância estatística, enquanto que os componentes principais têm importância estatística decrescente, ou seja, os primeiros componentes principais são tão mais importantes que podemos em alguns casos até desprezar os demais.

Vale ressaltar na análise da Figura 5 que o CP (1) é formado pela transformação linear de parte de todas as variáveis do subconjunto de dados, e sozinho é capaz de explicar cerca de 87% dos eventos. Os CP (2) e (3) são formados apenas pela transformação linear dos dados de vazão natural (com defasagem temporal). Podemos observar na Figura 6 que o CP (1) é capaz de explicar os eventos de grande variação entre as semanas (observada e prevista). Já os componentes (2) e (3) juntos não são capazes de identificar essas grandes variações entre as vazões naturais semanais. Essa análise é um indício de que nessas situações as informações de precipitação, tanto observadas quanto previstas, são de fundamental importância para o conhecimento das vazões futuras em complemento ao conhecimento das vazões passadas.

As outras variáveis e configurações analisadas foram: vazões naturas semanais observadas com defasagem de 2 e 3 semanas, previsão de precipitação para 7 dias agrupados e precipitação observada

acumulada nos últimos 7 dias dividida em dois conjuntos, com 4 e 3 dias. Essas variáveis/configurações foram retiradas do estudo por não apresentarem relevância nas análises estatísticas e/ou por comprometerem o desempenho do classificador bayesiano, sendo que em muitas vezes, o próprio modelo de classificação excluía algumas dessas variáveis/configurações. Esta exclusão se deu pelo fato dessas variáveis/configurações não apresentarem uma relação de causa e efeito significativa, do ponto de vista probabilístico, com a variável a ser prevista.

Um resumo do processo e das tecnologias envolvidas para a criação deste tipo de modelo, que foi batizado como Modelo de Previsão de Classes de Vazão (MPCV), pode ser vislumbrado no fluxograma apresentado na figura 9.

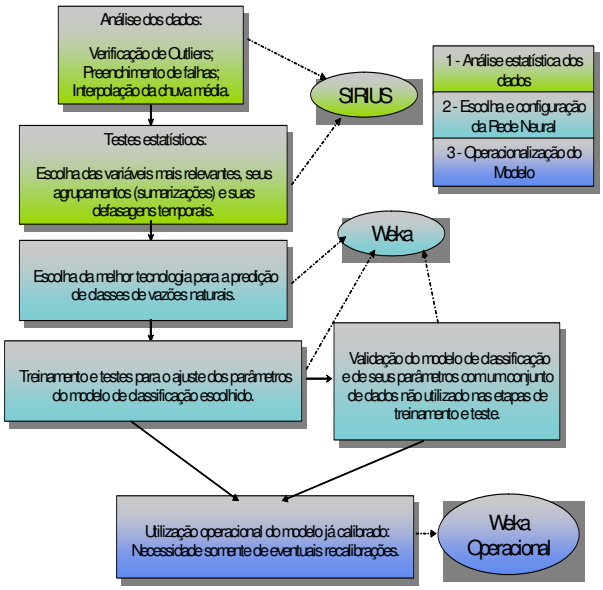

Figura 9 – Fluxograma com as etapas de análise estatística, escolha e configuração da tecnologia de Rede Neural e operacionalização do MPCV.

RESULTADOS OBTIDOS NOS TESTES DO MODELO

Os resultados apresentados a seguir foram obtidos no desenvolvimento do MPCV.

Assim, partindo de informações de precipitação observada e prevista, além das vazões verificadas nas últimas semanas que antecederam a previsão, foram estabelecidas faixas para as variáveis inseridas no classificador bayesiano, de forma a se ter uma classificação associada a cada previsão.

A partir das faixas de vazões semanais previstas pelo classificador bayesiano, interferiu-se na escolha do melhor modelo do PREVIVAZ em todas as semanas em que a sua previsão (aquela realizada pelo modelo melhor classificado pelo sistema PREVIVAZ) se apresentou fora da faixa sugerida. Nestes casos, buscou-se a previsão do modelo melhor posicionado dentro do ranking dos modelos utilizados pelo sistema PREVIVAZ, constantes em seu relatório de resultados, e que estivesse dentro da faixa de vazão determinada pelo modelo de balizamento desenvolvido no *software WEKA*.

Cabe ressaltar que, para algumas semanas, onde o melhor modelo escolhido pelo PREVIVAZ estava fora da faixa determinada pelo *WEKA*, o critério de busca descrito acima não obteve sucesso, tendo em vista o fato de as previsões dos modelos do PREVIVAZ estarem fora da faixa determinada. A alternativa encontrada para contornar este problema foi buscar, dentro do ranking dos modelos, aquele cuja previsão mais se aproximava da faixa prevista pelo classificador bayesiano. Esta metodologia foi testada para dois aproveitamentos da bacia do rio Iguaçu, a saber: Foz do Areia e Salto Santiago, considerando-se dados de previsão perfeita e real de precipitação. Conforme elucidado anteriormente, os dados de previsão real de precipitação foram gerados pelo modelo ETA do CPTEC.

Desta forma, para cada aproveitamento e para cada ano de teste (2002 e 2003), foram obtidos dois conjuntos de resultados: um considerando previsão perfeita de precipitação (obtida através da interpolação dos dados observados nos postos pluviométricos) e o outro considerando a previsão real de precipitação, como será apresentado a seguir.

Tabela 3 - Resumo dos erros médios quadráticos das previsões de vazão natural média semanal relativos aos anos de 2002 e 2003.

UHE	Ano	Previsão de Precipitação	Previvaz (%)	Previvaz com Data Mining (%)
Foz do Areia	2002	Perfeita	28,5	22,8
		Real		26,2
	2003	Perfeita	50,0	35,4
		Real		36,9
Salto Santiago	2002	Perfeita	33,7	24,5
		Real		28,1
	2003	Perfeita	35,1	27,6
		Real		29,4

Nas tabelas 3 e 4 estão disponíveis os principais resultados desse trabalho. Na tabela 3 são apresentadas as comparações dos erros médios quadráticos relativos das previsões de vazão natural média semanal do sistema Previvaz para os anos de 2002 e 2003, com e sem a utilização da metodologia desenvolvida nesse trabalho. Na tabela 4 são apresentadas as comparações somente para as semanas onde foi possível utilizar o critério de seleção proposto. Essas previsões são realizadas pelo Operador Nacional do Sistema Elétrico – ONS uma vez por semana e com uma antecedência de 3 a 4 dias em relação ao início da semana a ser prevista.

Tabela 4 - Resumo dos erros médios quadráticos das previsões de vazão natural média semanal, relativos às semanas em que a aplicação da metodologia interferiu no resultado.

UHE	Ano	Previsão de Precipitação	Previvaz (%)	Previvaz com Data Mining (%)
Foz do Areia	2002	Perfeita	34,7	26,6
		Real	35,7	31,1
	2003	Perfeita	57,3	29,2
		Real	63,3	38,1
Salto Santiago	2002	Perfeita	43,0	26,5
		Real	40,1	30,5
	2003	Perfeita	39,4	25,0
		Real	37,0	24,1

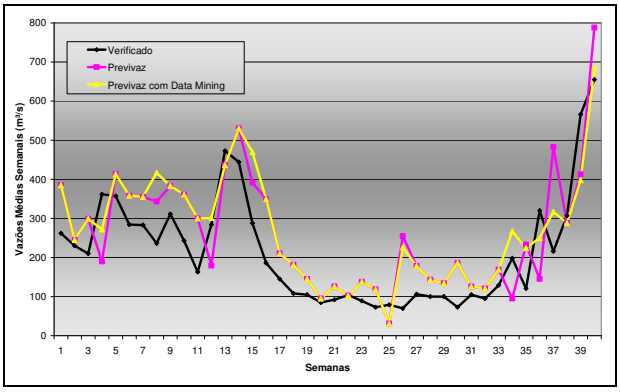

Figura 10 – Acompanhamento da previsão semanal de vazões naturais para a UHE Foz do Areia (m³/s).

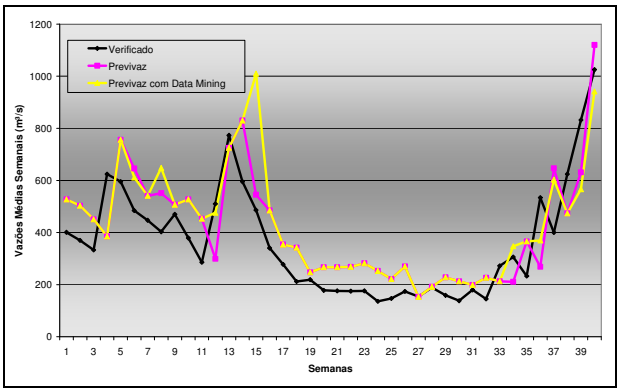

Figura 11 – Acompanhamento da previsão semanal de vazões naturais para a UHE Salto Santiago (m³/s).

MODELO OPERACIONAL

A aplicação do MPCV de modo operacional na rotina de previsão semanal de vazões naturais do ONS, para a bacia do rio Iguaçu, se deu a partir do ano de 2006 com a autorização da Agência Nacional de Energia Elétrica - ANEEL. Desde então, foram avaliados os valores iniciais previstos pelo modelo PREVIVAZ com o resultado do balizamento sugerido pela técnica de mineração de dados. Ambos os valores previstos foram armazenados e comparados com os valores observados na bacia. Nas figuras 10 e 11, pode-se analisar a evolução da previsão semanal de vazões naturais realizadas com o PREVIVAZ e com o MPCV para as UHE Foz do Areia e Salto Santiago, respectivamente. Essas previsões foram comparadas com os dados de vazão natural total observada nesses trechos até o fechamento da primeira semana de outubro de 2006.

De modo geral, verifica-se que para vazões baixas, como as que ocorreram em meados de setembro de 2006 na bacia do rio Iguaçu, o uso do MPCV, na maioria dos casos, não modifica os valores previstos pelo PREVIVAZ.

Em relação ao erro médio obtido entre os valores previstos e observados, nota-se que para o aproveitamento Foz do Areia, no período de 40 semanas, houve 11 ocasiões onde o MPCV indicou uma mudança da faixa de vazões semanais previstas. Deste total, em 8 ocasiões, a alteração resultou numa melhora da previsão e, por conseguinte, na diminuição do erro absoluto entre o valor esperado e o verificado. Em 3 ocasiões a indicação da nova faixa pelo MPCV implicou num afastamento maior da previsão em relação ao valor observado na bacia. Na figura 12 é apresentado à diferença entre os erros médios quadráticos calculados nas ocasiões onde houve alteração do patamar inicial previsto pelo PREVIVAZ. Destaca-se que na 35º semana, foi obtida

uma melhora significativa de cerca de 80% da previsão quando utilizada a técnica de mineração de dados.

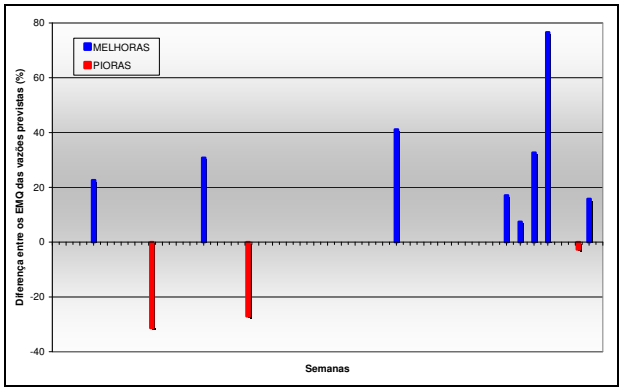

Figura 12 – Diferença entre o erro Médio Quadrático das vazões previstas pelos modelos Previvaz e MPCV para a UHE Foz do Areia (%)

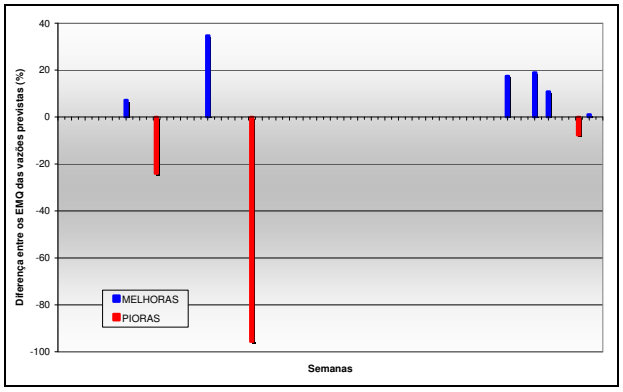

Figura 13 – Diferença entre o erro Médio Quadrático das vazões previstas pelos modelos Previvaz e MPCV para a UHE Salto Santiago (%)

Em relação ao aproveitamento Salto Santiago, como pode ser observado na figura 13, houve um total de 9 mudanças de faixa de vazão prevista pelo MPCV na mesma amostra de 40 semanas do ano de 2006. Deste total, em 6 ocasiões houve aproximação do valor previsto com o verificado e, em 3 ocorrências, houve maior dispersão entre os mesmos. Destaca-se um erro acentuado da previsão na 15º semana, após a alteração da previsão pelo MPCV. Este tipo de erro na previsão de vazões do MPCV geralmente está associado a grandes desvios entre a precipitação prevista pelo modelo ETA e aquela ocorrida na bacia. Pequenas variações entre os totais de precipitação observada e prevista geralmente não implicam em grandes erros nas previsões de vazão do MPCV, já que a previsão de precipitação é inserida no modelo através de faixas de valores, tal como ilustrado na figura 8.

CONCLUSÕES

Este estudo demonstrou que a aplicação das técnicas de *Data Mining* pode se apresentar como uma importante ferramenta para análise de variáveis de interações não lineares, como aquelas que compõem a estrutura dos fenômenos hidrológicos. Dentre as técnicas estudadas, os classificadores bayesianos foram os que apresentaram melhor destreza na predição das classes de vazões naturais, para a maioria dos casos analisados.

Nos anos escolhidos para a validação da metodologia, os resultados obtidos com a interferência do classificador bayesiano melhoraram o índice de acerto das previsões do modelo PREVIVAZ em todas as situações, inclusive naquelas onde foi utilizada a previsão de precipitação do modelo ETA. Destaca-se a previsão para o ano de 2003 no aproveitamento de Foz do Areia, onde os erros foram reduzidos pela metade nas semanas onde o classificador interferiu diretamente no resultado. Cabe ressaltar que a alternativa apresentada nesse estudo é de simples aplicação e possui um custo computacional extremamente baixo.

Em relação à utilização operacional do MPCV, foram observados melhores resultados nas previsões de vazão para UHE Foz do Areia. Esse comportamento pode estar associado a melhor distribuição e cobertura pluviométrica nessa região.

A modelagem estocástica univariada contida no modelo PREVIVAZ, por muitas vezes, dificulta a previsão de mudanças no comportamento das vazões entre uma semana e outra, o que ocasiona um efeito que é conhecido como "efeito sombra". A inserção das variáveis de precipitação como complementação às previsões do PREVIVAZ, se mostrou, ao longo deste estudo, como uma alternativa relativamente eficiente na minimização deste tipo de erro sistemático. Entretanto em muitos casos essa correção não poder ser realizada de forma mais efetiva, devido ao fato de que em algumas semanas, nenhuma das previsões realizadas pelos modelos do sistema PREVIVAZ estar dentro da faixa prevista pelo MPCV.

Por outro lado, nos casos onde a incerteza da previsão de precipitação induziu a previsão do MPCV à faixas de vazões muito distintas dos valores

verificados, os resultados do PREVIVAZ, cuja tendência é se aproximar da média de longo termo, devido a sua natureza estocástica, fizeram com que, especificamente nesses casos, os erros não aumentassem de forma significativa, minimizando o erro associado à inclusão de previsões de precipitação equivocadas no processo de previsão de vazões.

Para dar continuidade a este trabalho esta metodologia está sendo replicada para a bacia do rio Uruguai.

REFERÊNCIAS

Black T.L., 1994: NMC Notes: The New NMC mesoscale Eta model: Description and forecast examples. Weather and Forecasting, 9, 256-278.

Bouckaert, R. B., "Bayesian Network Classifiers in WEKA", *Internal Notes*, 2004

Buntine, W.L. "A guide to the literature on learning probabilistic networks from data", *IEEE Transactions on Knowledge and Data Engineering*, 8:195-210, 1996.

Cataldi, M., Machado, C.O., "Avaliação da previsão de precipitação utilizando a técnica de Downscale do modelo ETA e suas aplicações no setor elétrico", *XIII Congresso de Meteorologia*, 2004.

CEPEL, "Modelo de Previsão de Vazões Semanais Aplicado ao Sistema Hidroelétrico Brasileiro – Modelo Previvaz", Manual de Referência, 2004.

Friedman, N., Geiger, D., Goldszmidt, M., "Bayesian network classifiers". *Machine learning*, 29:131-163, 1997.

Guilhon, L.G.F. "Modelo Heurístico de Previsão de Vazões Naturais Médias Semanais Aplicado à Usina de Foz do Areia", - Dissertação de Mestrado, UFRJ, 2003

Witten, I.H., Frank, E., "Data Mining: Practical machine learning tools and techniques with Java implementations", *Morgan Kaufmann Publishers*, 2000.

Use of Data Mining Techniques to Complement Univariate Stochastic Forecasts of Natural Flow Studied Case by Case in the Iguaçu River Basin

ABSTRACT

This paper presents the results obtained by using software in the public domain that, through Data Mining and Neural Networks with Bayesian training, can lay the foundation to select the most appropriate natural inflow forecast used in the PREVIVAZ stochastic modeling system. This technique utilizes forecast and observed information on precipitation, as well as natural inflow data recorded over the weeks that precede the actual forecast target for the Foz do Areia and Salto Santiago hydroelectric plants located in the Iguaçu River Basin. The results obtained indicate that the using these tools can provide a simple and efficient solution to reduce natural inflow forecast errors on a weekly forecast basis for the Iguaçu River Basin

Keywords: Data Mining; Bayesian Networks; Stochastic Models; Inflow Forecasts.