Previsão da Temperatura da Água no Reservatório de Itaipu Utilizando o Método Não-Linear *k-Nearest Neighbor*

Carlos Andres Ferrero[1,2,4], **Maria Carolina Monard**[2,4],
Huei Diana Lee[1,4], **Simone Frederigi Benassi**[3,4], **Wu Feng Chung**[1,4]

[1]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo
Laboratório de Inteligência Computacional – LABIC
Caixa Postal 668, 13560-970 – São Carlos, SP, Brasil

[2]Centro de Engenharias e Ciências Exatas – Universidade Estadual do Oeste do Paraná
Laboratório de Bioinformática – LABI
Caixa Postal 39, 85856-970 – Foz do Iguaçu, PR, Brasil

[3]Superintendência de Meio Ambiente, Divisão de Reservatórios – Itaipu Binacional
Caixa Postal 39, 85866-900 – Foz do Iguaçu, PR, Brasil

[4]Centro de Estudos Avançados em Segurança de Barragens – CEASB
Parque Tecnológico Itaipu – PTI

{anfer,mcmonard}@icmc.usp.br,huei@pti.org.br

Abstract. *The goal of time series forecasting is extrapolating past behavior into the future. To this end, several methods have been proposed, among them nonlinear forecasting methods. Some nonlinear forecasting methods are based on the k-Nearest Neighbor algorithm, and have been used to model environmental behaviors. In this work, we applied a k-Nearest Neighbor forecasting methodology to predict future Itaipu lake's water temperature. Our methodology considers two prediction functions as well as different numbers of nearest neighbors. Experimental results show that the proposed methodology is suitable to implement decision support systems.*

Resumo. *A previsão de dados temporais é uma tarefa de interesse para diversas áreas, inclusive para o monitoramento ambiental. Métodos para realizar essa tarefa têm sido propostos e aplicados. Especificamente os que permitem prever comportamentos não lineares têm apresentado-se apropriados para a previsão de comportamentos ambientais. Neste trabalho, é apresentada uma metodologia para a previsão de dados temporais utilizando o algoritmo k-Nearest Neighbor para a previsão de dados ambientais. Foram utilizadas duas funções para o cálculo do valor futuro e avaliadas para diferentes valores do parâmetro k do algoritmo. Os resultados mostraram que a metodologia e as funções de predição são promissoras para a construção de sistemas de suporte à previsão de dados.*

1. Introdução

Sistemas computacionais para gerenciamento de dados permitem, cada vez mais, o armazenamento de informações de diversas áreas do conhecimento. Esse acúmulo de

informações faz com que haja a necessidade de utilizar métodos computacionais que permitam organizá-los e analisá-los, com o objetivo de extrair informações adicionais que permitam, por exemplo, auxiliar especialistas em processos de tomada de decisão. Na área de Seguranca de Barragens, essa análise é de grande importância devido ao fato de que os dados coletados periodicamente por sensores contêm informações a respeito do estado da barragem.

Como esses dados coletados por sensores constituem observações realizadas seqüencialmente ao longo do tempo, é fundamental considerar a ordem dos dados. Essa restrição não permite a aplicação direta do processo de mineração de dados. Desse modo, é necessário o desenvolvimento e a aplicação de métodos de pré-processamento, extração de padrões e interpretação de bases de dados temporais. Assim, os dados temporais referentes à Segurança de Barragens necessitam ser pré-processados para, posteriormente, serem analisados por meio do processo de mineração de dados [Witten and Frank 2005]. Nesse sentido, está sendo desenvolvido o projeto Análise Inteligente de Dados de Séries Temporais para Segurança de Barragens, em uma parceria entre o Laboratório de Inteligência Computacional — LABIC — da Universidade de São Paulo — USP / São Carlos —, o Laboratório de Bioinformática — LABI — da Universidade Estadual do Oeste do Paraná — UNIOESTE / Foz do Iguaçu — e o Centro de Estudos Avançados em Segurança de Barragens — CEASB — do Parque Tecnológico Itaipu — PTI.

O projeto contempla três etapas, ilustradas na Figura 1. A primeira consiste no pré-processamento dos dados temporais. A idéia é utilizar diversos métodos de limpeza e transformação de dados, bem como métodos de extração e seleção de características, com o objetivo de obter uma descrição estruturada dos dados a serem analisados, além de construir modelos matemáticos que representem o comportamento das Séries Temporais — ST. Posteriormente, na Etapa 2, com base nessa descrição estruturada dos dados, pode ser aplicado o processo de mineração de dados. Os padrões extraídos nesse processo, conjuntamente com os modelos matemáticos construídos, podem ser considerados para a realização de diversas tarefas de interesse, tais como a análise comportamental de fenômenos, a detecção de anomalias, a predição de eventos e a detecção de padrões, entre outras. Na terceira etapa, avaliação de riscos, o conhecimento extraído e os padrões encontrados na etapa anterior, podem ser utilizados para realizar as atividades de identificação, estimação, avaliação e controle de riscos.

O objetivo deste trabalho em andamento, o qual está inserido dentro da etapa de descoberta de padrões, é apresentar uma metodologia que utiliza o algoritmo k-*Nearest Neighbor* — kNN — para a previsão de dados temporais relacionados ao tema Segurança de Barragens. A metodologia foi aplicada a uma série de dados limnológicos[1], utilizando duas funções para o cálculo do valor futuro e diferentes valores do parâmetro k do algoritmo. A partir dos resultados foi possível observar que a metodologia e as funções de predição são promissoras para a construção de sistemas de suporte à previsão de dados.

O restante deste trabalho está organizado do seguinte modo: na Seção 2 é descrita a metodologia proposta; na Seção 3 é apresentado o estudo de caso em dados ambientais; e, na Seção 4, são apresentados a conclusão e os trabalhos futuros.

[1] A limnologia é a ciência que estuda as águas continentais.

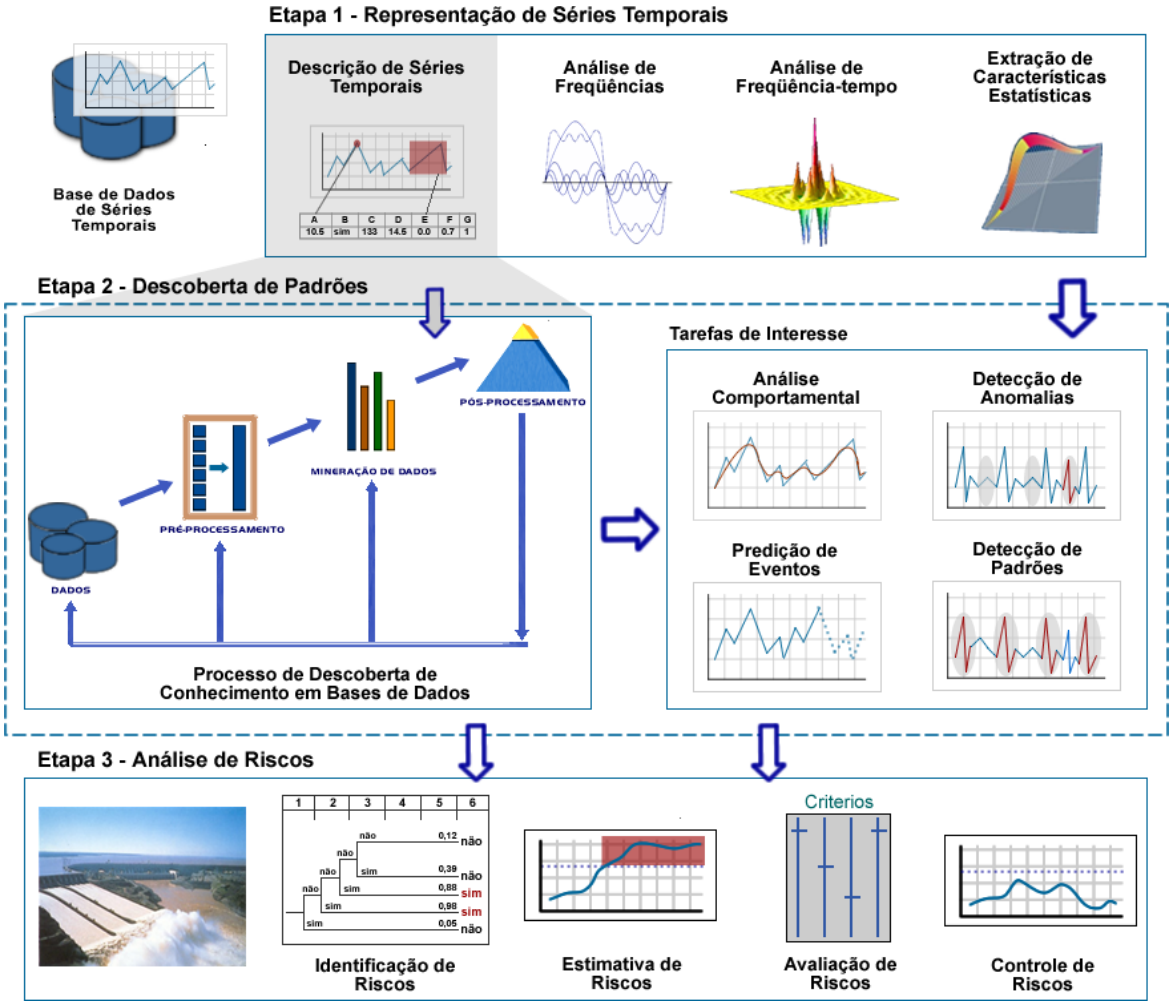

Figura 1. Projeto de Análise Inteligente de Dados de Séries Temporais para Segurança de Barragens.

2. Metodologia de Previsão de Dados Temporais

Uma das tarefas de maior interesse para qualquer área de conhecimento que esteja interessada em analisar fenômenos do ponto de vista temporal, consiste na previsão de valores futuros a partir do histórico da série em questão. Dependendo do objetivo da tarefa, as previsões podem apresentar características específicas. Por exemplo, em problemas de monitoramento, é desejável que as previsões sejam realizadas a curto prazo.

O problema de previsão em séries temporais consiste em predizer o valor de x_{t+1} de uma série temporal $X = (x_1, x_2, \ldots, x_n)$, utilizando os valores anteriores a $t+1$, isto é, $x_t, x_{t-1}, x_{t-2}, \ldots, x_{t-m+1}$, onde m corresponde ao número de valores prévios da série X utilizados para realizar a previsão. De acordo com [Sorjamaa et al. 2007], o cálculo do próximo valor de uma série temporal pode ser definido conforme a Equação 1.

$$x'_{t+1} = f_1(x_t, x_{t-1}, x_{t-2}, \ldots, x_{t-m+1}) \tag{1}$$

As técnicas de previsão utilizam, de modo geral, duas abordagens. A primeira corresponde à utilização de métodos lineares, que consistem em ajustar aos dados modelos Auto-regressivos (AR); Médias Móveis (MA — *Moving Average*); Auto-regressivos

de Médias Móveis (ARMA — *Auto-regressive Moving Average*) e Auto-regressivos de Médias Móveis Integrados (ARIMA — *Auto-regressive Integrated Moving Average*). A segunda abordagem corresponde à utilização de modelos não lineares, isto é, métodos que permitem prever comportamentos e encontrar padrões não lineares [Yankov et al. 2006]. Exemplos desse tipo de métodos são os baseados no algoritmo k-*Nearest Neighbor* e em redes neurais artificiais [Karunasinghe and Liong 2006]. Os baseados no método kNN podem ser apropriados para a modelagem de comportamentos da natureza e constituem o foco deste trabalho.

O método k-*Nearest Neighbor* é um método de aprendizado que permite realizar previsões de valores futuros baseando-se nos valores registrados do passado. A idéia consiste em, considerando os últimos N_u registros ocorridos, encontrar as M_s seqüências de tamanho N_u que apresentam comportamentos similares no passado. Com base nas informações dessas M_s seqüencias, é realizado o cálculo do valor futuro x'_{t+1}.

A metodologia proposta neste trabalho para a aplicação do kNN consiste de quatro fases: (1) Preparação do conjunto de treinamento, (2) Obtenção dos k vizinhos mais próximos, (3) Cálculo do valor futuro e (4) Avaliação do método de previsão. Essas fases são descritas a seguir.

2.1. Preparação do Conjunto de Treinamento

Para que o método kNN possa ser aplicado, é necessário construir o conjunto de séries de treinamento, i.e, todos os exemplos de séries que devem ser lembrados pelo algoritmo kNN para encontrar as seqüências similares. Considerando a ST $X = (x_1, x_2, \ldots, x_n)$ e $U = (x_{n-N_u}, x_{n-N_u+1}, \ldots, x_n)$ os últimos N_u valores de X, cada elemento do conjunto de séries de treinamento $S = \{s_1, s_2, s_{n-N_u}\}$ é definido pela Equação 2.

$$\begin{aligned} s_1 &= (x_1, x_2, \ldots, x_{N_u}, x_{N_u+1}) \\ s_2 &= (x_2, x_3, \ldots, x_{N_u+1}, x_{N_u+2}) \\ s_3 &= (x_3, x_4, \ldots, x_{N_u+2}, x_{N_u+3}) \\ &\vdots \\ s_{n-N_u} &= (x_{n-N_u}, x_{n-N_u+1}, \ldots, x_{n-1}, x_n) \end{aligned} \tag{2}$$

2.2. Obtenção dos *k* Vizinhos mais Próximos

Para a aplicação do kNN em séries temporais, três características importantes devem ser consideradas: (a) o número de registros que devem ser lembrados (tamanho do conjunto de treinamento); (b) a medida a ser utilizada para quantificar a similaridade entre a seqüencia procurada e cada série do conjunto de treinamento; e (c) o número de vizinhos mais próximos utilizado para o cálculo do valor futuro. Essas características, as quais influenciam fortemente o desempenho dos métodos de previsão, são descritas a seguir:

Tamanho da série de treinamento: a complexidade de kNN em tempo e espaço depende do tamanho do conjunto de treinamento [Alpaydin 2004]. Assim, se todos os registros são memorizados, a procura pelas seqüências similares pode tornar-se um processo lento. Para contornar esse problema podem ser consideradas apenas as séries mais representativas do conjunto de treinamento, resumindo a informação mais importante em um conjunto menor de dados. Outra abordagem é considerar apenas as séries mais recentes.

Medida de similaridade: a noção de similaridade é um aspecto decisivo no contexto de análise de dados. Entretanto, o critério de decisão a respeito do que deve ser considerado similar é bastante subjetivo, pois depende de diversos fatores como o domínio de aplicação e do método para o cálculo dessa similaridade. No contexto de séries temporais são usadas distintas medidas para calcular a similaridade entre duas séries temporais, as quais variam de acordo com as características da série e com o modo como essa série é representada [Fink 2004, Vlachos and Gunopulos 2004]. Dentre essas medidas, a distância Euclidiana é a mais utilizada para a comparação de séries temporais [Keogh and Kasetty 2002], determinando a distância no espaço $\Re^m$ entre dois pontos, onde m corresponde ao tamanho da seqüência considerada.

Cardinalidade do conjunto de vizinhos mais próximos: o cálculo do valor futuro depende do número de vizinhos mais próximos (k) que é considerado pelo algoritmo. Os k vizinhos mais próximos constituem o conjunto $S' = \{s'_1, s'_2, \ldots, s'_k\} \subset S$ de cardinalidade $|S'| = k$.

2.3. Cálculo do Valor Futuro

A partir do conjunto S', que contém k séries temporais de comportamento mais próximo, uma função $f(s'_1, s'_2, \ldots, s'_k)$ é utilizada para o cálculo do valor futuro. Diversas funções foram propostas na literatura e, neste trabalho, são utilizadas duas. A primeira, média local, consiste no cálculo da média dos valores que sucedem a cada seqüência [McNames 1999], neste trabalho denominada Média de Valores Absolutos — MVA. A segunda função consiste em, a partir da diferença ocorrida entre o último valor de cada seqüência e o valor que a sucede, calcular o valor futuro pela soma do último valor amostrado e a média das diferenças encontradas. Essa última função é denominada neste trabalho Média de Valores Relativos — MVR.

2.4. Avaliação do Método de Previsão

Nesta etapa, o método de previsão construído é avaliado e validado, utilizando duas abordagens. A primeira consiste na utilização de medidas objetivas e, a segunda, no parecer dos especialistas do domínio. Medidas para avaliar a qualidade de métodos de previsão permitem auxiliar na análise do desempenho desses métodos de modo mais objetivo. Essa análise pode ser realizada por meio de diversas métricas amplamente divulgadas na literatura [Daliakopoulosa et al. 2004, Karunasinghe and Liong 2006]. Uma análise a respeito das características dessas medidas de avaliação é apresentada em [Hyndman and Koehler 2006]. Quatro métricas são utilizadas neste trabalho, três delas, Erro Médio Absoluto — EMA —, Desvio-Padrão Absoluto — DPA — e Coeficiênte de Correlação — R^2 —, são definidas, respectivamente, pelas Equações 3, 4 e 5, onde $Z(t) = (z_1, z_2, \ldots, z_{N_o})$ é a série temporal de valores observados; $\hat{Z}(t) = (\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_{N_o})$ é a série temporal correspondente calculada; N_o corresponde ao número de valores observados; e $e(t) = \hat{Z}(t) - Z(t)$ é a diferença entre os valores preditos e os observados.

$$EMA = \text{média}(|e|) \tag{3}$$

$$DPA = \text{desvio-padrão}(|e|) \tag{4}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N_o}(z_i - \hat{z}_i)^2}{\sum_{i=1}^{N_o} z_i^2 - \frac{\sum_{i=1}^{N_o} \hat{z}_i^2}{N_o}} \tag{5}$$

A terceira medida de avaliação utilizada neste trabalho, denominada Erro de Direção — ED —, consiste na percentagem de valores previstos que cresceram ou decresceram conforme esperado. Para a construção dessa medida são considerados o último valor observado, o valor de previsão observado e o valor previsto. Se o valor de previsão observado e o valor previsto são ambos maiores (ou menores) do que o último valor observado, então a direção do valor de previsão observado e do valor previsto é a mesma, caso contrário, ocorre um ED devido à previsão em direção contraria à esperada.

3. Aplicação da Metodologia para a Previsão de Dados Ambientais

A avaliação de riscos em uma barragem deve ser capaz de identificar problemas e recomendar soluções para esses problemas, tais como estratégias corretivas e operacionais [Pan and He 2000]. Para auxiliar nesse processo de avaliação de riscos, é necessário que sejam realizadas coletas de dados através de monitoramentos freqüentes, com o objetivo de manter a integridade de todas as áreas relacionadas à barragem. Como os dados são temporais, uma solução adequada consiste em representar o problema por meio de séries temporais e analisá-las utilizando técnicas desse tipo de dados.

O monitoramento ambiental é definido como o conjunto de dados físicos, químicos e biológicos de um ecossistema em estudo, que permite obter informações a respeito da qualidade das águas, um tema de importância para a segurança de barragens. O monitoramento consiste de repetidas observações, medidas, registros ambientais e parâmetros operacionais em um período de tempo. A Itaipu Binacional possui uma equipe de especialistas e uma rede de monitoramento de qualidade de água que está distribuída entre o reservatório e os seus afluentes. Os dados coletados pela Itaipu Binacional envolvem a medição de variáveis físicas e químicas climatológicas e da água. As amostras de água são coletadas na superfície e em diferentes profundidades. A partir das amostras coletadas são medidas as seguintes variáveis: alcalinidade total, clorofila a, condutividade, Demanda Bioquímica de Oxigênio (5 dias) — DBO_5 —, Demanda Química de Oxigênio — DQO —, fósforo total, nitrato, nitrito, nitrogênio amoniacal, nitrogênio Kjedahl, oxigênio dissolvido, pH, saturação de oxigênio, sólidos suspensos, temperatura, turbidez e transparência da água e temperatura do ar.

Dentre esses dados, neste trabalho foram considerados registros trimestrais de temperatura da água, coletados em superfície, na estação E5, localizada a 15 km a montante da barragem, no período compreendido entre 1994 e 2004. Nesse período foram coletadas 44 observações, as quais constituem a série temporal de temperaturas $T(t) = \{x_1, x_2, \ldots, x_{44}\}$ a ser analisada. A metodologia apresentada na Seção 3 para a previsão de dados utilizando o algoritmo *k-Nearest Neighbor* foi aplicada a essa série temporal.

Assim, a partir de T foram criados dez conjuntos de experimentação, definidos por $E = \{e_1, e_2, \ldots, e_{10}\}$, em que cada elemento de E é definido por $e_i = \{X_i, x_j\}$, onde X_i corresponde a uma subsérie de X e x_j ao valor a ser previsto. Esses conjuntos foram construídos conforme a Equação 6.

$$
\begin{aligned}
e_1 &= \{(x_1, x_2, \ldots, x_{34}), x_{35}\} \\
e_2 &= \{(x_1, x_2, \ldots, x_{35}), x_{36}\} \\
&\vdots \\
e_{10} &= \{(x_1, x_2, \ldots, x_{43}), x_{44}\}
\end{aligned} \tag{6}
$$

Para cada um destes experimentos foi realizada a previsão de x_j utilizando a metodologia apresentada na Seção 3. Na Etapa (1) da metodologia, preparação do conjunto de treinamento, foi utilizada a série de treinamento contida em cada experimento para construir o conjunto de séries de treinamento S_i e, conseqüêntemente, a série dos últimos valores registrados U_i. Neste trabalho, a cardinalidade de U_i foi definida como $N_u = 4$, que representa o período de um ano de coleta.

Na Etapa (2), obtenção dos k vizinhos mais próximos, foram procuradas as subseqüências de S_i mais próximas de U_i. Na questão (a), sobre o número de subseqüências em S_i a serem consideradas, foi utilizado o conjunto completo, pois o tamanho da amostra não é suficientemente grande para influenciar no desempenho do algoritmo kNN. Em relação à medida de similaridade — questão (b) — foi utilizada a distância Euclidiana e, para obter uma medida orientada à morfologia da seqüência, foi realizada uma tranformação de translação, a qual consiste na subtração do valor médio da seqüência a cada valor da seqüência, antes do cálculo da distância. Para a questão (c), de escolha do número k de vizinhos mais próximos, foram utilizados os valores de $k = 1, 2, 3, 4$ e 5, no intuito de encontrar o número de vizinhos mais próximos que poderia proporcionar melhores valores de previsão.

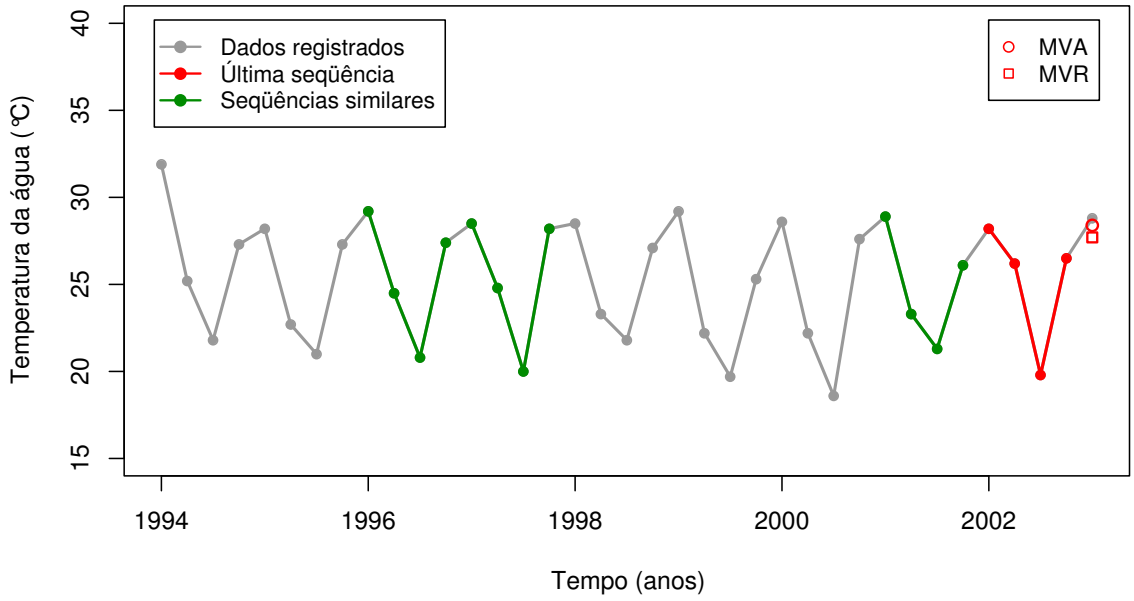

Figura 2. Exemplo do resultado da aplicação das funções MVA e MVR para o cálculo do valor de futuro em dados de temperatura da água.

Na etapa (3), cálculo do valor futuro, foram utilizadas as duas funções apresentadas na Seção 2.3, média dos valores absolutos (MVA) e média dos valores relativos

(MVR). Na Figura 2 é apresentado um exemplo da aplicação do método *kNN* utilizando $k = 3$. No gráfico, a linha cinza representa os valores registrados; a linha vermelha, a seqüência dos 4 últimos valores ocorridos, isto é, os registros de um período de um ano; e, a linha verde, as seqüências similares ao período em questão, encontradas pelo algoritmo. Utilizando os valores sucessores de cada uma das seqüências mais similares foi realizada a previsão do valor do quarto trimestre de 2002. Nessa figura, é ilustrado o resultado da aplicação de duas funções diferentes de previsão (representadas por círculo e quadrado vermelhos, respectivamente).

Os valores de temperatura resultantes da aplicação das funções MVA e MVR para cada k e os correspondentes coeficientes de correlação R^2, são apresentados na Tabela 1.

Tabela 1. Valores de previsão resultantes da aplicação das funções MVA e MVR para cada k.

	Valor	**Valores de previsão (^oC)**									
	real	$k=1$		$k=2$		$k=3$		$k=4$		$k=5$	
Registro	(^oC)	MVA	MVR	MVA	MVR	MVA	MVR	MVA	MVR	MVA	MVR
x_{35}	$19,8$	$20,0$	$21,4$	$20,4$	$21,9$	$19,8$	$22,2$	$20,1$	$22,8$	$20,4$	$23,1$
x_{36}	$26,5$	$28,2$	$28,0$	$27,8$	$27,2$	$27,7$	$27,7$	$27,3$	$26,9$	$27,3$	$26,8$
x_{37}	$28,8$	$28,5$	$26,8$	$28,5$	$27,2$	$28,4$	$27,7$	$28,4$	$27,6$	$28,5$	$27,8$
x_{38}	$23,3$	$24,8$	$25,1$	$24,1$	$24,4$	$23,6$	$24,0$	$24,2$	$24,7$	$23,8$	$24,2$
x_{39}	$20,2$	$20,8$	$19,6$	$19,7$	$19,7$	$19,8$	$19,3$	$20,3$	$19,9$	$20,5$	$20,2$
x_{40}	$25,9$	$27,4$	$26,8$	$27,5$	$28,0$	$27,0$	$27,0$	$26,6$	$26,7$	$26,7$	$26,7$
x_{41}	$28,9$	$28,2$	$28,0$	$28,4$	$27,5$	$28,4$	$28,1$	$28,4$	$27,8$	$28,5$	$27,8$
x_{42}	$20,8$	$22,2$	$22,5$	$24,2$	$24,7$	$23,5$	$23,8$	$23,8$	$24,1$	$24,0$	$24,1$
x_{43}	$18,1$	$19,7$	$18,3$	$19,1$	$17,8$	$19,8$	$18,2$	$19,9$	$18,1$	$20,1$	$17,9$
x_{44}	$24,7$	$25,3$	$23,7$	$26,5$	$25,4$	$26,3$	$24,9$	$26,2$	$24,4$	$26,5$	$24,4$
Coeficiente R^2		$0,95$	$0,87$	$0,90$	$0,80$	$0,92$	$0,87$	$0,93$	$0,84$	$0,93$	$0,84$

Como pode ser observado, a função MVA apresentou sempre valores de correlação maiores para as previsões, considerando cada k, em relação à função MVR. Em relação a MVA, o menor valor de correlação foi $0,90$ correspondepente a $k = 2$, enquanto o maior valor foi de $0,95$ correspondente a $k = 1$. No caso de MVR, o menor R^2 foi de $0,80$ referente a $k = 2$ e, o maior, de $0,87$, referente a $k = 1$ e 3.

Posteriormente, foi realizada uma análise sobre os erros de previsão correspondentes à aplicação de cada função de previsão. Para uma análise mais clara desses erros foi conservado o sinal positivo $(+)$ ou negativo $(-)$, conforme o valor previsto estivesse acima $(+)$ ou abaixo $(-)$ do valor real de temperatura. Na Tabela 2, são apresentados os erros dos valores de previsão de temperatura utilizando as funções MVA e MVR para cada k e os respectivos erro médio absoluto, desvio-padrão absoluto e erro de direção.

Com base nessas informações pode ser evidenciado que, em relação ao erro médio absoluto, a função MVA apresentou o menor valor ($0,99$ com DPA de $0,83$) para $k = 3$ e o maior valor ($1,18$ com DPA $0,93$), para $k = 2$. A função MVR, apresentou menor valor de EMA ($1,12$ com DPA de $1,21$), correspondente a $k = 5$ e maior valor ($1,44$ com DPA de $1,07$) para $k = 2$. O teste estatístico ANOVA para dados emparelhados foi aplicado para verificar a existência de diferença estatisticamente significativa entre os erros de previsão considerando os diferentes valores de k. Os $p - valores$ para as funções MVA e MVR foram $0,8399$ e $0,6205$, respectivamente, não sendo possível afirmar que para

Tabela 2. Erros de previsão utilizando as funções MVA e MVR para cada k.

	Valor	Erros de previsão (^oC)									
	real	$k=1$		$k=2$		$k=3$		$k=4$		$k=5$	
Registro	(^oC)	MVA	MVR	MVA	MVR	MVA	MVR	MVA	MVR	MVA	MVR
x_{35}	$19,8$	$+0,2$	$+1,6$	$+0,6$	$+2,1$	$+0,0$	$+2,4$	$+0,3$	$+3,0$	$+0,6$	$+3,3$
x_{36}	$26,5$	$+1,7$	$+1,5$	$+1,3$	$+0,7$	$+1,2$	$+1,2$	$+0,8$	$+0,4$	$+0,8$	$+0,3$
x_{37}	$28,8$	$-0,3$	$-2,0$	$-0,3$	$-1,6$	$-0,4$	$-1,1$	$-0,4$	$-1,2$	$-0,3$	$-1,0$
x_{38}	$23,3$	$+1,5$	$+1,8$	$+0,8$	$+1,1$	$+0,3$	$+0,7$	$+0,9$	$+1,4$	$+0,5$	$+0,9$
x_{39}	$20,2$	$+0,6$	$-0,6$	$-0,5$	$-0,5$	$-0,4$	$-0,9$	$+0,1$	$-0,3$	$+0,3$	$+0,0$
x_{40}	$25,9$	$+1,5$	$+0,9$	$+1,6$	$+2,1$	$+1,1$	$+1,1$	$+0,7$	$+0,8$	$+0,8$	$+0,8$
x_{41}	$28,9$	$-0,7$	$-0,9$	$-0,5$	$-1,4$	$-0,5$	$-0,8$	$-0,5$	$-1,1$	$-0,4$	$-1,1$
x_{42}	$20,8$	$+1,4$	$+1,7$	$+3,4$	$+3,9$	$+2,7$	$+3,0$	$+3,0$	$+3,3$	$+3,2$	$+3,3$
x_{43}	$18,1$	$+1,6$	$+0,2$	$+1,0$	$-0,3$	$+1,7$	$+0,1$	$+1,8$	$+0,0$	$+2,0$	$-0,2$
x_{44}	$24,7$	$+0,6$	$-1,0$	$+1,8$	$+0,7$	$+1,6$	$+0,2$	$+1,5$	$-0,3$	$+1,8$	$-0,3$
	EMA	$1,01$	$1,22$	$1,18$	$1,44$	$0,99$	$1,15$	$1,00$	$1,18$	$1,07$	$1,12$
	DPA	$0,58$	$0,58$	$0,93$	$1,07$	$0,83$	$0,91$	$0,88$	$1,13$	$0,96$	$1,21$
	ED	$0,00$	$0,00$	$0,00$	$0,00$	$0,00$	$0,00$	$0,00$	$0,00$	$0,00$	$0,00$

algum valor de k os erros de previsão sejam menores em relação aos erros de previsão dos outros valores de k. Por outro lado, em relação ao erro de direção, tanto a função MVA quanto a função MVR não apresentaram ED.

4. Conclusão

O desenvolvimento de sistemas computacionais que permitem o armazenamento e a organização de informação tem promovido um aumento das bases de dados em todas as áreas do conhecimento. Nesse sentido, diversas áreas de pesquisa têm se interessado em compreender fenômenos que transcendem no tempo. Neste trabalho foi apresentada uma metodologia para a previsão de dados temporais utilizando o método k-*Nearest Neighbor*. A metodologia foi aplicada a uma série de dados ambientais de temperatura da água, utilizando as funções de cálculo de valor de previsão, média de valores absolutos e média de valores relativos. O parâmetro k do método *Nearest Neighbor*, o qual especifica a cardinalidade do conjunto de vizinhos mais próximos utilizado para o cálculo do valor futuro, foi avaliado com os valores $k = 1, 2, 3, 4$ e 5, no intuito de encontrar o melhor valor de k para a série de dados utilizada.

Os resultados evidenciam que a função de média de valores absolutos teve melhor desempenho que a função de média de valores relativos, para o conjunto de dados considerado neste trabalho, em relação à medida de coeficiente de correlação. Os melhores valores de k em para o coeficiente de correlação R^2 foram, $k = 1$ para a função MVA e $k = 1$ e 3 para MVR. Por outro lado, em relação ao erro médio absoluto, o melhor valor de k foi $k = 3$ para MVA e $k = 5$ para MVR, em que para cada função não foi possível encontrar diferença estatisticamente significativa em relação à variação do parâmetro k de 1 até 5. Isso indicou que, tanto para a função MVA quanto para a função MVR, nenhum valor k apresentou erro de previsão significativamente menor. Considerando o erro de direção, ambas funções de previsão apresentaram erro de previsão igual a 0, independentemente do valor de k. Desse modo, não foi possível determinar qual função. juntamente com um valor de k, permite obter previsões significativamente mais precisas.

Além disso, os especialistas da área de domínio consideraram os resultados

promissores para a construção de sistemas computacionais de suporte à previsão. Trabalhos futuros incluem a aplicação da metodologia para a previsão de outras variáveis coletadas, como alcalinidade total, clorofila a, entre outras, a utilização de períodos maiores de coleta para aumentar o tamanho do conjunto de séries de treinamento, a utilização de outras funções de cálculo de valor futuro e a aplicação de outros métodos de previsão de comportamentos lineares e não lineares.

5. Agradecimentos

Ao Programa de Desenvolvimento Tecnológico Avançado — PDTA-FPTI/BR — e ao Centro de Estudo Avançados em Segurança de Barragens — CEASB —, pelo auxílio por meio da linha de financiamento de bolsas. Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico — CNPq — pelo apoio financeiro.

Referências

Alpaydin, E. (2004). *Introduction to Machine Learning*. MIT Press, Cambridge — MA, England.

Daliakopoulosa, I. N., Coulibalya, P., and Tsanis, I. K. (2004). Groundwater level forecasting using artificial neural networks. *Journal of Hydrology*, 309(1–4):229–240.

Fink, E. (2004). *Data Mining in Time Series Databases*, volume 57 of *Machine perception and artificial intelligence*, chapter Indexing of Compressed Time Series, pages 43–65. World Scientific, Singapore.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688.

Karunasinghe, D. S. K. and Liong, S.-Y. (2006). Chaotic time series prediction with a global model: Artificial neural network. *Journal of Hydrology*, 323(1-4):92–105.

Keogh, E. and Kasetty, S. (2002). On the need for time series data mining benchmarks: a survey and empirical demonstration. In *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining*, pages 102–110, New York, USA.

McNames, J. (1999). *Innovations in Local Modeling for Time Series Prediction*. Tese de doutorado, Stanford. Disponível em: `www.ece.pdx.edu/~mcnames/Publications/Dissertation.pdf`.

Pan, J. and He, J. (2000). *Large Dams in China: a fifty–year review*. China WaterPower Press, Beijing, China.

Sorjamaa, A., Hao, J., Reyhani, N., Ji, Y., and Lendasse, A. (2007). Methodology for long-term prediction of time series. *Neurocomputing*, 70(16-18):2861–2869.

Vlachos, M. and Gunopulos, D. (2004). *Data Mining in Time Series Databases*, chapter Indexing Time-Series Under Conditions of Noise, pages 67–100. Machine perception and artificial intelligence. World Scientific.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, San Francisco — CA, USA, 2 edition.

Yankov, D., DeCoste, D., and Keogh, E. J. (2006). Ensembles of nearest neighbor forecasts. In *European Conference on Machine Learning*, pages 545–556.