Revista Engenho, vol.11 –Dezembro de 2015

MODELO DE REGRESSÃO POR MÍNIMOS QUADRADOS PARCIAIS PARA DADOS DE MONITORAMENTO DE BARRAGENS

Suellen Ribeiro Pardo Garcia
Universidade Tecnológica Federal do Paraná, Toledo, PR, Brasil

Anselmo Chaves Neto
Universidade Federal do Paraná, Curitiba, PR, Brasil

Sheila Regina Oro
Universidade Tecnológica Federal do Paraná, Francisco Beltrão, PR, Brasil

Tereza Rachel Mafioleti
Universidade Tecnológica Federal do Paraná, Francisco Beltrão, PR, Brasil

Claudio Neumann Junior
Itaipu Binacional, Foz do Iguaçu, PR, Brasil

RESUMO

Movimentos em estruturas de barragens de concreto ocorrem devido às variações do nível do reservatório, temperatura e eventuais deformações permanentes. É de interesse investigar o relacionamento entre essas variáveis ambientais e a resposta da barragem, uma vez que o monitoramento é atividade permanente de engenheiros e técnicos responsáveis pela segurança da obra. Neste trabalho, apresenta-se uma metodologia para a construção de um modelo de regressão por mínimos quadrados parciais, que descreve o relacionamento entre dados de sensores de um pêndulo direto e variáveis ambientais. O modelo condensa informação dos dados em um número pequeno de novas variáveis, o que indica um bom potencial para auxiliar no monitoramento de barragens.

Palavras-chave: Regressão Multivariada, Mínimos Quadrados Parciais, Barragem e Deslocamento.

ABSTRACT

Movements in concrete dam structures occur due to variations in the reservoir level, temperature and eventual permanent deformation. It is of interest to investigate the relationship between these environmental variables and the dam response, since the monitoring is permanent activity of engineers and technicians responsible for the safety of the structure. This paper presents a method for constructing a model of partial least squares regression that describes the relationship between sensor data and a direct pendulum and environmental variables. The model condenses data information in a small number of new variables indicating a good potential to assist in the monitoring of dams.

Keywords: Multivariate Regression, Partial Least Squares, Dam and Displacement

INTRODUÇÃO

O monitoramento da estrutura de uma barragem é realizado por meio de inspeções visuais e instrumentação. A instrumentação instalada gera uma enorme massa de dados, que, se analisada devidamente, fornece informações sobre o comportamento da estrutura mediante efeitos externos, como as variações do nível do reservatório e da temperatura ambiente.

O objetivo da análise dos dados da instrumentação é propiciar informação que possa ser usada em uma interpretação física das deformações e, na previsão, seja do comportamento futuro da própria estrutura ou para estudo do comportamento de uma estrutura semelhante (DENG; WANG; SZOSTAK, 2008, p.1). Modelos estatísticos são propostos na literatura para tal objetivo, como por exemplo, os trabalhos de Ahmadi-Nedushan (2002), Chouinard e Roy (2006), De Sortis e Paoliani (2007), Léger e Lecler (2007), Deng, Wang e Szostak (2008), Yu et al. (2010), Mata (2011), Xi et al. (2011) e Li, Wang e Liu (2013). Esses são baseados em correlações existentes entre fatores, tais como o nível de água do reservatório, temperatura ambiente, desgaste devido ao tempo e a resposta da barragem a alguns tipos de ações, como tensões, deformações e deslocamentos (AHMADI-NEDUSHAN, 2002, p.25).

Dois grandes desafios encontrados ao propor tais modelos para dados de monitoramento de barragens é que, primeiro, as variáveis independentes, ou seja, as variações do nível do reservatório e de temperatura, podem gerar dados multicolineares, de modo que não seja possível utilizar algumas técnicas estatísticas clássicas. Segundo, o número de variáveis dependentes é geralmente alto, pois são muitos os sensores de instrumentação de uma barragem de concreto.

A multicolinearidade cria uma variância compartilhada entre as variáveis, diminuindo, assim, a capacidade de prever a variável dependente, bem como de examinar a importância relativa de cada variável independente (HAIR, *et al.*, 2009, p. 165).

O método de regressão por mínimos quadrados parciais não parte da hipótese de que as variáveis sejam não correlacionadas, ~~e~~ nem requer que os resíduos sigam uma distribuição normal, como ocorre na regressão por mínimos quadrados ordinários. O método de mínimos quadrados parciais utiliza as componentes obtidas, a fim de maximizar a covariância entre as variáveis independentes e as variáveis dependentes

(GARCIA e FILZMOSER, 2011, p.25).

O método generaliza e combina características de Regressão Multivariada, Análise de Correlação Canônica e Análise de Componentes Principais sem impor suas restrições (DENG; WANG; SZOSTAK, 2008, p.2).

O método de mínimos quadrados parciais – ou PLS (*Partial Least Squares*), como é conhecido na literatura – foi desenvolvido na década de 1960, por Herman Wold, como uma técnica econométrica, mas seus maiores defensores são engenheiros químicos. PLS é aplicada em calibração de espectrometria, no monitoramento e controle de processos industriais, em que um processo pode ter centenas de variáveis controláveis e dezenas de saídas (TOBIAS, 1995). Outras aplicações foram em medicina, psicologia e agropecuária, entre outras áreas.

Entre as aplicações de regressão por PLS, destaca-se o trabalho de Deng *et al*. (2008), que apresenta uma análise de deformação tridimensional para um único ponto sobre a barragem. A análise consiste na construção de um modelo, previsão de deformação e análise da contribuição de fatores individuais. A metodologia foi empregada em uma barragem de terra localizada na região central da China. A conclusão, no artigo, foi que o modelo de regressão por mínimos quadrados parciais é mais confiável e tem melhor integridade do que o modelo de regressão múltipla, que, segundo os autores, foi amplamente empregado no monitoramento de barragens.

A proposta do presente artigo é desenvolver um modelo estatístico de regressão multivariada por mínimos quadrados parciais, no qual as leituras dos sensores de um pêndulo direto, instalados em um bloco de concreto do tipo gravidade aliviada, componham a matriz de variáveis dependentes e as variáveis independentes (preditoras) são as leituras da variação do nível do reservatório e juntamente com as leituras dos termômetros de superfície instalados no concreto do bloco. O aspecto relevante desse modelo é sua característica multivariada, ou seja, será proposto um modelo para diversas variáveis de resposta simultaneamente, o que pode auxiliar no monitoramento de barragens de concreto.

MODELOS ESTATÍSTICOS E MONITORAMENTO DE BARRAGENS

Modelos estatísticos, utilizados para analisar e interpretar os dados da instrumentação, são baseados em correlações existentes entre fatores, como o nível de

água do reservatório e a temperatura ambiente, entre outros, e os efeitos causados na barragem, como tensões, deformações e deslocamentos.

Modelos de regressão linear múltipla para dados de monitoramento de barragem são construídos com o objetivo de prever a resposta da estrutura em função das cargas que nela atuam. Esses modelos são baseados em dois pressupostos. O primeiro é que os efeitos são analisados em um período em que a configuração da barragem continua a mesma; o segundo é que a resposta da barragem é separada em efeitos reversíveis (devido à variação do nível do reservatório e temperatura do ar) e irreversíveis (devido ao adensamento, à decantação, degradação ou fluência). A resposta de um instrumento (por exemplo, deslocamento) pode ser modelada da seguinte forma (Ahmadi-Nedushan, 2002, p.9)

$$D_i(t) = F_i(t) + G_i(H) + H_i(T) + \varepsilon_i \quad (1)$$

Onde $F(t)$ é a função que descreve o efeito irreversível, $G(H)$ é a função do nível do reservatório (carga hidrostática), $H(T)$ é a função da temperatura e ε é o erro. Na literatura, são encontradas várias funções propostas para modelar os diferentes componentes de resposta, principalmente, quando se trata de modelar $F(t)$ e $H(T)$. Algumas dessas versões serão comentadas aqui.

No período operacional normal de uma barragem de concreto, o efeito térmico é diretamente relacionado às variações de temperatura, e a inércia térmica cria um atraso na resposta entre a variação de temperatura e as leituras dos instrumentos. Existem duas abordagens para descrever esse efeito térmico: o modelo HST (*hydrostatic, seasonal, time*) e modelos que consideram a temperatura do concreto.

O modelo HST foi proposto inicialmente, em 1958, por Ferry, Will e Beaujoint (CHOUINARD e ROY, 2006, p. 201). Algumas versões são encontradas na literatura para melhor ajuste do modelo ao estudo de caso, cita-se Ahmadi-Nedushan (2002), De Sortis e Paoliani (2007), Xi *et al.* (2011), Mata (2011) e Li, Wang e Liu (2013).

No modelo HST, o efeito do nível do reservatório é modelado por um polinômio de quarto grau; o efeito da temperatura, por uma soma de funções trigonométricas, e os efeitos irreversíveis, por uma função polinomial do tempo (AHMADI-NEDUSHAN, 2002), da seguinte forma:

$$D(t) = H(z) + S(\theta) + T(t) = a_1 + a_2 z + a_3 z^2 + a_4 z^3 + a_5 z^4 + a_6 sen(\theta) + a_7 cos(\theta) + \\ + a_8 sen(\theta) cos(\theta) + a_9 sen^2(\theta) + c_1 t + c_2 t^2 + c_3 t^3 \quad (2)$$

onde $D(t)$ é a variável resposta (por exemplo, deslocamentos), $H(z)$, $S(\theta)$, $T(t)$ são, respectivamente, função do nível do reservatório, função da temperatura e efeito irreversível, onde t é o número de dias desde que se iniciou a análise. As variáveis z e θ são definidas como $z = \frac{h - h_{mín}}{h_{máx} - h_{mín}}$, h nível do reservatório e $\theta = \frac{2\pi j}{365}$, $j = 1, ..., 365$

.

Várias funções são propostas na literatura para modelar a função do efeito irreversível. Por exemplo, De Sortis e Paoliani (2007) utilizam $T(t) = c_0 + c_1 t$; Xi *et al.* (2011) modelam com $T(t) = c_1 \theta + c_2 ln(\theta)$; Mata (2011) utiliza a função $T(t) = c_1 t + c_2 e^{-t}$, e por fim, Li *et a*l. (2013) utiliza $T(t) = c_1 \theta + c_2 ln(\theta + 1)$. As variáveis θ e t são dadas em número de dias ou em ano desde que começou a análise, dependendo da aplicação.

Os coeficientes desconhecidos a_k e c_l são calculados por uma minimização da diferença entre as medidas reais e as medidas obtidas pelo modelo em (2), usando o método dos mínimos quadrados. Nota-se que o modelo HST é construído por meio de funções não lineares, mas, como os valores das variáveis de entrada são conhecidos pelo pesquisador (tempo e nível do reservatório), o modelo se torna linear ao passo que essas variáveis são substituídas nas funções.

Segundo Léger e Leclerc (2007), uma abordagem para modelar os efeitos térmicos seria utilizar os dados dos termômetros embutidos na barragem, que monitoram a evolução transitória de temperaturas do concreto. Substituindo a função da temperatura $S(\theta)$ de HST por

$$S(T) = \sum_{i=1}^{k} b_i T_i \quad (3)$$

onde b_i são os coeficientes e T_i são os dados dos termômetros $1, 2, ..., k$. Esse modelo denomina-se HT_dT (*hydrostatic, direct temperature, time*). Assim,

$$D(t) = H(z) + S(T) + T(t) \quad (4)$$

Neste trabalho, apresenta-se um modelo multivariado, ou seja, consideram-se diversas variáveis de resposta (dependentes), o que difere dos modelos encontrados na literatura. Como, para essa aplicação, estão disponíveis dados dos termômetros embutidos no concreto do bloco, opta-se pela abordagem do modelo HT_dT, descrita por Léger e Leclerc (2007). Para a modelagem do efeito irreversível, ajusta-se a função proposta por Xi *et al*. (2011), $T(t) = c_1 t + c_2 \ln(t)$ onde *t* é dado em anos.

REGRESSÃO POR MÍNIMOS QUADRADOS PARCIAIS

O método de regressão por PLS é uma técnica de estimação do modelo de regressão linear, baseada na decomposição das matrizes de variáveis respostas e de variáveis preditoras. O algoritmo usado examina ambas as matrizes e extrai componentes, que são diretamente relevantes a ambos os conjuntos de variáveis (AHMADI-NEDUSHAN, 2002, p.32).

Segundo Morellato (2010), o método de regressão PLS apresenta as vantagens de:

- modelar regressões com múltiplas variáveis respostas;
- aceitar multicolinearidade
- os fatores produzidos têm alto poder de predição, devido às altas covariâncias com as variáveis resposta.

As desvantagens do método são:

- dificuldade na interpretação das cargas dos fatores;
- os testes de significância dos estimadores dos coeficientes de regressão são realizados via métodos de reamostragem, pois suas distribuições não são conhecidas
- falta de estatísticas de teste para o modelo.

O método de regressão por PLS é, preferencialmente, uma técnica de predição, e não de interpretação, apesar de existirem trabalhos que fazem interpretação dos fatores extraídos via PLS (MORELLATO, 2010, p. 4). Devido ao caráter de predição do método, sua aplicação no monitoramento de barragens se torna interessante, pois a comparação de valores preditos pelo modelo a valores observados traz informação útil sobre o comportamento da barragem.

DESCRIÇÃO DO MODELO

Segue a descrição do método de regressão PLS, baseada no trabalho de (WOLD, SJÖSTRÖM e ERIKSSON, 2001).

Como, em regressão linear multivariada, o objetivo da regressão por PLS é construir um modelo linear, $Y = X\beta + \varepsilon$, Y é uma matriz *nxm* de variáveis de resposta, X é uma matriz *nxr* de variáveis preditoras, β é uma matriz *rxm* dos coeficientes de regressão, e ε é a matriz dos resíduos *nxm*. Essa é uma abordagem livre de distribuição. Dessa forma, os resíduos possuem vetor de médias nulo e matriz de covariâncias igual a $\sigma^2 I$, onde I é a matriz identidade de ordem *nxn*, mas sem distribuição definida.

O modelo encontra poucas "novas" variáveis chamadas de escores de X, ou componentes, ou fatores, que são denotadas por $\underline{t}_a \ (a = 1,2,...,A)$. Essas componentes são preditoras de Y e também modelam X (equações (8) e (6)), ou seja, ambas variáveis, e são assumidas, ao menos parcialmente, e modeladas pelas mesmas variáveis latentes. O objetivo de extrair componentes que consigam capturar as variâncias das variáveis preditoras, e também de obter correlações com as variáveis de resposta é alcançado maximizando a covariância entre X, $\underline{t}_a$ e Y (MORELLATO, 2010, p.6).

O número de componentes A é menor do que o número de variáveis preditoras ($A<r$), e estes componentes são ortogonais, obtidos por combinações lineares das variáveis originais $\underline{x}_r$, com os pesos $\underline{w}_a$ ($a=1,2,...,A$), da seguinte forma:

$$T = XW \tag{5}$$

onde $T = (\underline{t}_1, \underline{t}_2, ..., \underline{t}_A)$ é a matriz *nxa* de componentes (escores ou fatores) e $W = (\underline{w}_1, \underline{w}_2, ..., \underline{w}_A)$ é a matriz *rxa* de pesos.

As matrizes X e Y são decompostas na forma:

$$X = TP' + F \tag{6}$$

$$Y = UC' + G \tag{7}$$

onde T e U são matrizes *nxA* de componentes (escores ou fatores) de X e Y respectivamente, P' e C' são matrizes *Axr* e *Axm* de pesos de X e Y,

respectivamente e *F* e *G* são as matrizes dos resíduos.

As componentes de *X* são boas preditoras de *Y*, ou seja,

$$Y = TC' + E \tag{8}$$

onde *C'* é obtido por mínimos quadrados, dado por

$$\hat{C}' = \left(T'T\right)^{-1} T'Y \tag{9}$$

Para conseguir os coeficientes da regressão por PLS, referentes às variáveis originais, substitui-se (5) em (8), e obtém-se

$$Y = TC' + E = XWC' + E = XB + E \Rightarrow B = WC'$$

ou seja,

$$\hat{B} = W\hat{C}' \tag{10}$$

A j-ésima coluna da matriz $\hat{B}$ corresponde aos coeficientes estimados para o modelo de regressão por PLS da variável de resposta $\underline{y}_j$, $j = 1,...,m$.

Diferentes algoritmos podem ser utilizados na extração das componentes da regressão por PLS. O algoritmo mais popular é o Non-Iterative Partial Least Squares (NIPALS), desenvolvido por Wold, em 1966. Outros algoritmos podem ser encontrados na literatura, como o algoritmo SIMPLS, publicado por De Jong, em 1993, ou o algoritmo Kernel, descrito por Lindgren et al., em 1993, e Rannar et al., em 1994 (MEVIK e WEHRENS, 2007).

Na regressão por PLS, o número de componentes A determina a complexidade do modelo. Com inúmeras e correlacionadas variáveis preditoras, existe risco de sobreajuste, ou seja, obter-se um modelo bem ajustado, porém com pouco ou nenhum poder de predição. Portanto há a necessidade de se verificar o poder de cada componente adicionada no modelo, e parar o processo, quando a inclusão de componentes for não significativa (WOLD, SJÖSTRÖM e ERIKSSON, 2001). A validação cruzada é, normalmente, utilizada para determinar esse número ótimo de componentes (MEVIK e WEHRENS, 2007).

VALIDAÇÃO CRUZADA

O processo consiste em ajustar um modelo com uma observação retirada da amostra. Com o modelo estimado, calcula-se a previsão para essa observação retirada. Esse processo é repetido até que todas as observações da amostra sejam retiradas, suas previsões calculadas e obtém-se com isso uma estatística de erro. Assim, tem-se as estatísticas de erro para o número de componentes $a=1,2,....,A$ e pode-se avaliar qual o número de componentes ideal, ou seja, o número de componentes que minimiza essa estatística.

A estatística de erro é a soma dos quadrados das diferenças entre os valores observados e os previstos. Esses valores constituem a soma dos quadrados dos resíduos da predição (PRESS - *predictive residual sum of squares*), que estima a capacidade preditiva do modelo. A razão $PRESS_a / SS_{a-1}$ é calculada após cada componente, e um componente é considerado significativo quando comparado a um valor crítico fixado. Aqui, SS_{a-1} denota a soma dos quadrados dos resíduos antes da atual componente fixada. Os cálculos continuam até que um componente seja não significativo (WOLD, SJÖSTRÖM e ERIKSSON, 2001, p. 116). O valor crítico de $Q^2 = 1 - PRESS_a / SS_{a-1}$ é igual a 0,0975 com 95% de nível de confiança.

MATERIAIS E MÉTODOS

A Usina Hidrelétrica de Itaipu é uma empresa binacional localizada no Rio Paraná, na fronteira entre Brasil e Paraguai (Figura 1). A barragem foi construída, no período de 1975 a 1982, por ambos os países. Itaipu foi a maior produtora mundial de energia por dois anos consecutivos, 2012 e 2013, porém, em 2014, a Barragem das Três Gargantas, na China, produziu 98,8 milhões de megawatts hora (MWh), contra 98,5 milhões de MWh de Itaipu.

Itaipu conta com, aproximadamente, 2.400 instrumentos (1.358 no concreto, 881 nas fundações e 161 para geodesia), sendo 270 automatizados. Conta também com 5.295 drenos (949 no concreto e 4.346 nas fundações) para acompanhar o desempenho das estruturas de concreto e fundações (ITAIPU BINACIONAL, 2015).

Os dados obtidos pela instrumentação permitem, aos engenheiros, analisar o comportamento da estrutura, sendo o deslocamento um dos parâmetros mais significativos em monitoramento de segurança de barragens. O pêndulo direto foi o

instrumento considerado nesta análise. Esse instrumento capta os movimentos relativos da estrutura, e esses movimentos relativos acontecem devido a fatores externos, como variação do nível do reservatório e da temperatura. Logo, as leituras do nível do reservatório e as leituras dos termômetros de superfície também são consideradas neste estudo.

A Barragem Principal (trecho F) é composta de blocos de concreto do tipo gravidade aliviada, com 16 tomadas de água. Esse é o trecho com maior influência da variação do nível do reservatório, por apresentar os blocos com maior altura.

O trecho F possui 18 blocos duplos de concreto do tipo gravidade aliviada que vão do F1/2 ao F35/36, sendo quatro deles intensamente instrumentados, denominados blocos-chave. Recebem maior quantidade de instrumentos, devido às características do solo em que se localizam e às características da construção. Os blocos-chave no trecho F são os blocos F5/6, F13/14, F19/20 e F35/36. Considerou-se o bloco F19/20 para análise, por ser um dos blocos com maior altura de coluna d'água.

Figura 1: Vista aérea da Usina Hidrelétrica de Itaipu. Em destaque, o Trecho F. Fonte: BUZZI, DYMINSKI e CHAVES NETO (2007).

Considera-se, neste trabalho, as leituras mensais dos instrumentos no período compreendido entre 2000 e 2013, pois, a partir de 2000, as leituras realizadas pelos técnicos apresentaram, em geral, essa periodicidade. No início das leituras da instrumentação, em 1982, a frequência era maior. Com exceção do nível do reservatório, que sempre teve leituras diárias, selecionou-se o nível no dia em que a leitura do pêndulo foi realizada. Outra razão para a escolha do período de 2000 a 2013 é por que foram detectados poucos dados perdidos, não precisando de um

método mais complexo de imputação. Para esses poucos casos, a média da leitura anterior e posterior foi utilizada.

Os dados das leituras são de 5 sensores do pêndulo direto (COF18X, COF19X, COF20X, COF21X, COF22X) e 6 termômetros de superfície (TSF11, TSF12, TSF13, TSF14, TSF15, TSF16). As outras variáveis são a função que modela variação do nível do reservatório (z, z2, z3, z4) e a função que modela o efeito irreversível (t, ln t). A eis consideradas para o modelo.

Tabela 1 apresenta as 17 variáveis consideradas para o modelo.

Tabela 1: Variáveis consideradas no modelo.

Variáveis Dependentes											
COF18X			COF19X		COF20X		COF21X			COF22X	
Variáveis Independentes											
z	z2	z3	z4	t	ln t	TSF11	TSF12	TSF13	TSF14	TSF15	TSF16

O pêndulo direto fornece medidas dos deslocamentos na direção X, no sentido do fluxo (direção montante-jusante), e na direção Y, perpendicular ao fluxo (margem direita-esquerda) medido em milímetros. O presente trabalho limita-se a modelar os deslocamentos no sentido do fluxo (direção X). O nível do reservatório é dado em metros, e a unidade de medida para os termômetros de superfície é em graus Celsius. Na Figura 2 segue a localização dos sensores (CO-F-17, CO-F-18, CO-F-19, CO-F-20, CO-F-21 e CO-F-22) do pêndulo direto no bloco F19/20, a localização dos termômetros de superfície (TS-F-11, TS-F-12, TS-F-13, TS-F-14, TS-F-15 e TS-F-16). As cotas estão identificadas na Figura 2 pela sigla El. (elevação). O sensor CO-F-17 foi excluído da análise, pois o modelo não conseguiu resultado satisfatório para essa variável. Pode-se notar na Figura 3 que as leituras dos deslocamentos no sensor CO-F-17 no sentido do fluxo (direção X) têm comportamento semelhante aos dos outros sensores, mas com uma amplitude bem inferior, de 1,6 mm.

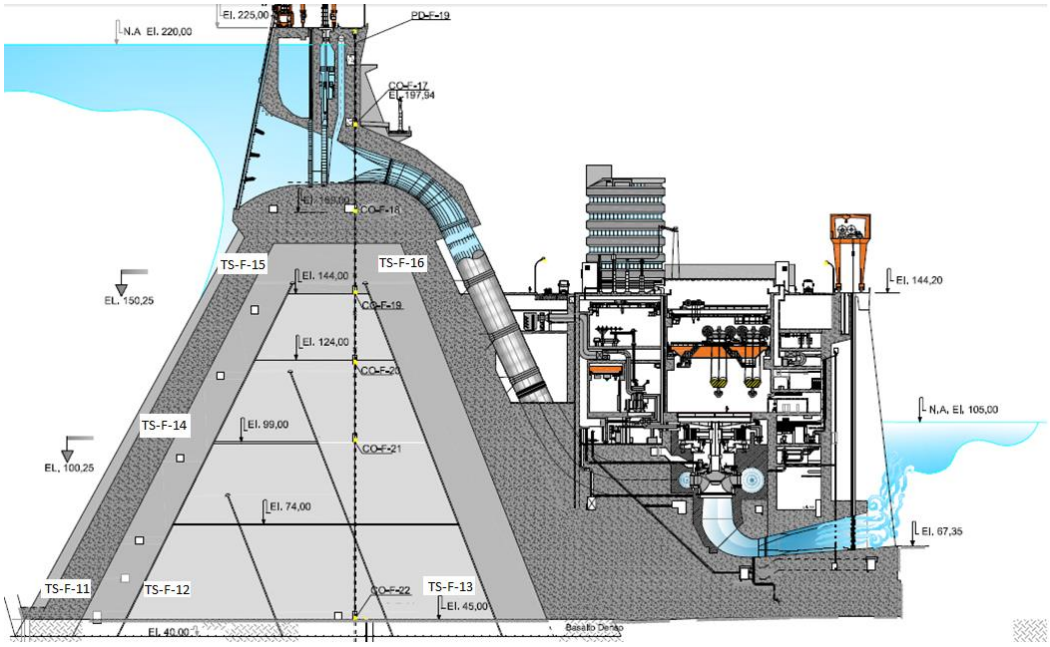

Figura 2: Pêndulo direto do bloco F19/20 (PD-F-19).

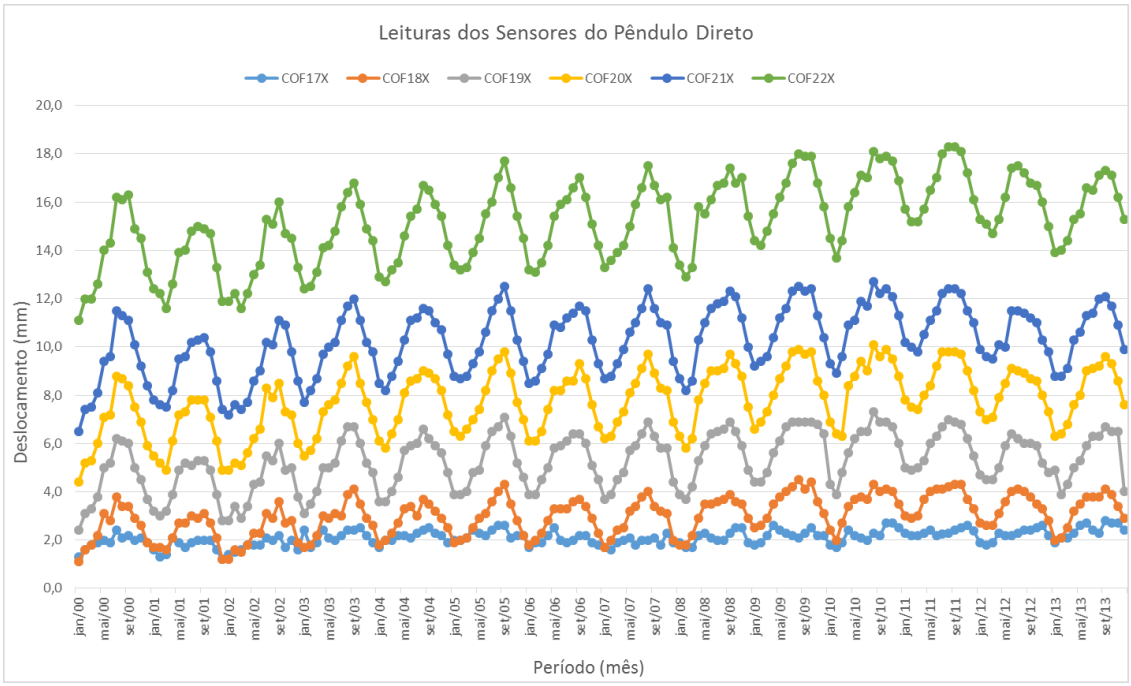

Figura 3: Sensores do pêndulo direto no bloco F19/20.

AJUSTE DO MODELO

O número de observações (leituras dos instrumentos com periodicidade mensal) é de 168 para cada variável. Para a modelagem foi utilizado o software livre R (R CORE TEAM, 2014).

O que precede o ajuste é a justificativa da escolha do método de mínimos quadrados parciais. O método de mínimos quadrados ordinários (modelo de regressão clássico), ao contrário do método de mínimos quadrados parciais, apresenta resultados instáveis para tamanhos de amostra pequenos em relação ao número de variáveis independentes e o alto grau de correlação entre as variáveis independentes (multicolinearidade) aumenta a variância dos coeficientes estimados (MORELLATO, 2010).

O fator de inflação da variância (VIF – Variance Inflation Factor) quantifica a gravidade da multicolinearidade em uma análise de regressão dos mínimos quadrados ordinários. Essa estatística fornece um índice que mede o quanto a variância de um coeficiente de regressão estimado é maior na presença de colinearidade.

O VIF é calculado para cada x_i (variável independente) dado por $VIF = \frac{1}{1-R_i^2}$, onde R_i^2 é o coeficiente de determinação da equação de regressão $x_i = \alpha_1 x_1 + ... + \alpha_{i-1} x_{i-1} + \alpha_{i+1} x_{i+1} + ... + \alpha_p x_p$. Avalia-se a magnitude da multicolinearidade considerando se o valor de $VIF > 10$ então a multicolinearidade é alta.

Em dados de monitoramento de barragens, geralmente existe multicolinearidade. De fato, isso é verificado nessa aplicação. Os fatores de inflação da variância para as variáveis independentes (nível do reservatório e temperaturas dos termômetros de superfície) estão apresentados na Tabela 2. De acordo com os valores da tabela, as variáveis que representam a variação do nível do reservatório (z, z2, z3 e z4, funções do modelo HT_dT), TSF12 e TSF13, como apresentam valores VIF maiores que 10, indicam alta multicolinearidade. A presença da multicolinearidade impede a utilização do modelo de regressão clássico, assim, justifica-se a escolha do método de mínimos quadrados parciais para a estimativa dos parâmetros, uma vez que esse método não é afetado pela presença de multicolinearidade.

Tabela 2: Fator de inflação da variância.

z	z2	z3	z4	TSF11	TSF12	TSF13	TSF14	TSF15	TSF16	t	ln t
155,6	1844.7	3692,1	948,18	1,4	27,1	25,4	1,6	4,8	3,1	6,4	6,4

Pode-se medir a contribuição de uma dada variável independente x_j para a construção de um componente do modelo, calculando os quadrados dos pesos w_{hj}^2. No

entanto utilizam-se valores *VIP* (importância da variável para a projeção) como forma de classificar as variáveis independentes, em termos de seu poder explicativo. Os preditores com $VIP > 1$ são considerados mais relevantes para a construção de *Y*.

Na Tabela 3, apresentam-se os valores *VIP* para todas as variáveis independentes, considerando até 4 componentes para o modelo. Consideram-se quatro componentes, mas, nesse momento, qualquer outro valor maior que 2 seria aceitável, pois observa-se que, a partir de dois componentes, o valor de *VIP* não se altera.

Tabela 3: Importância da variável para a projeção considerando 4 componentes.

	t1	**t2**	**t3**	**t4**
Z	0,68	0,73	0,73	0,73
z2	0,69	0,74	0,74	0,74
z3	0,68	0,74	0,74	0,74
z4	0,67	0,73	0,73	0,73
TSF11	0,84	0,84	0,84	0,84
TSF12	1,41	1,37	1,37	1,37
TSF13	1,40	1,37	1,37	1,37
TSF14	0,31	0,32	0,32	0,35
TSF15	1,44	1,41	1,41	1,40
TSF16	1,24	1,21	1,21	1,22
T	0,93	0,96	0,96	0,96
ln t	0,95	0,94	0,94	0,94

Nota-se que TSF12, TSF13, TSF15 e TSF16 são mais relevantes no modelo. Como t e $\ln t$ apresentaram valores próximos a 1, opta-se por manter essas variáveis no modelo também.

A validação cruzada para o novo modelo (com as variáveis z, z2, z3, z4, TSF11 e TSF14 excluídas) é apresentada na Tabela 4. Nota-se que $Q_h^2 \geq 0,0975$ para a escolha de *h = 3* componentes para a variável COF18X, *h = 1* para a variável COF19X e *h = 2* para as variáveis COF20X, COF21X e COF22X, respectivamente. Assim, o modelo será ajustado considerando *h=3*.

Tabela 4: Validação cruzada considerando até 6 componentes para o modelo.

Q_h^2	COF18X	COF19X	COF20X	COF21X	COF22X
t1	0,754	**0,798**	0,819	0,821	0,788
t2	0,118	0,057	**0,127**	**0,120**	**0,287**
t3	**0,131**	0,083	0,083	0,049	-0,005
t4	-0,005	-0,007	0,009	0,006	0,039
t5	0,035	0,001	-0,002	0,009	0,017
t6	-0,027	-0,016	-0,011	-0,011	-0,017

Outro resultado importante é o coeficiente R^2 e a proporção de variância explicada pelas componentes do modelo. As duas primeiras colunas da Tabela 5 correspondem às variáveis independentes. As colunas três e quatro referem-se às respostas, variáveis dependentes. Com a escolha de $h = 3$, temos 95% e 84% da variância das variáveis independentes e dependentes respectivamente, explicada pelo modelo.

Tabela 5: Variância explicada pelo modelo.

	R2 de X	R2 de X acumulado	R2 de Y	R2 de Y acumulado
t1	0,55724	0,55724	0,797434	0,797434
t2	0,320032	0,877272	0,031556	0,82899
t3	0,069959	0,947231	0,015125	0,844116

Os deslocamentos obtidos pelos sensores do pêndulo direto (COF18X, COF19X, COF20X, COF21X e COF22X) e os deslocamentos previstos pelo ajuste do modelo são dados na Figura 4. Os resíduos são apresentados graficamente na Figura 5.

Nessa aplicação, nem todas as variáveis de previsão (independentes) contribuíram para a interpretação dos deslocamentos, pois os valores de *VIP* na Tabela 3 indicaram apenas as variáveis TSF12, TSF13, TSF15 e TSF16 como mais relevantes para o modelo, e opta-se por manter os termos que modelam os efeitos irreversíveis (t e $\ln t$). Isso apenas confirma informação já conhecida pelos engenheiros de que o deslocamento, na leitura do pêndulo direto, é fortemente influenciado pela temperatura ambiente. Assim, as variáveis z, z2, z3 e z4 foram excluídas, pois não há contribuição relevante da variação do nível do reservatório nos movimentos relativos captados pelo pêndulo direto no bloco F19/20. No entanto deve-se salientar que não ocorreu contribuição relevante da variação do nível do reservatório nesse instrumento avaliado apenas (pêndulo direto do bloco F19/20).

Os termômetros excluídos do modelo (TSF11 e TSF14) têm localização a montante. O TSF11 está instalado próximo à face do bloco, na cota 50,2 m (acima do nível do mar), e apresenta $\Delta t = 2°C$, ou seja, mede indiretamente a temperatura da água do reservatório em uma cota na qual a temperatura praticamente não varia. O TSF14 localiza-se na cota 100,25 m e apresenta uma variação de temperatura maior, aproximadamente $\Delta t = 5°C$. No entanto, por se localizar a montante e a uma

profundidade de, aproximadamente, 120 m, não contribui para a previsão do movimento relativo desse bloco.

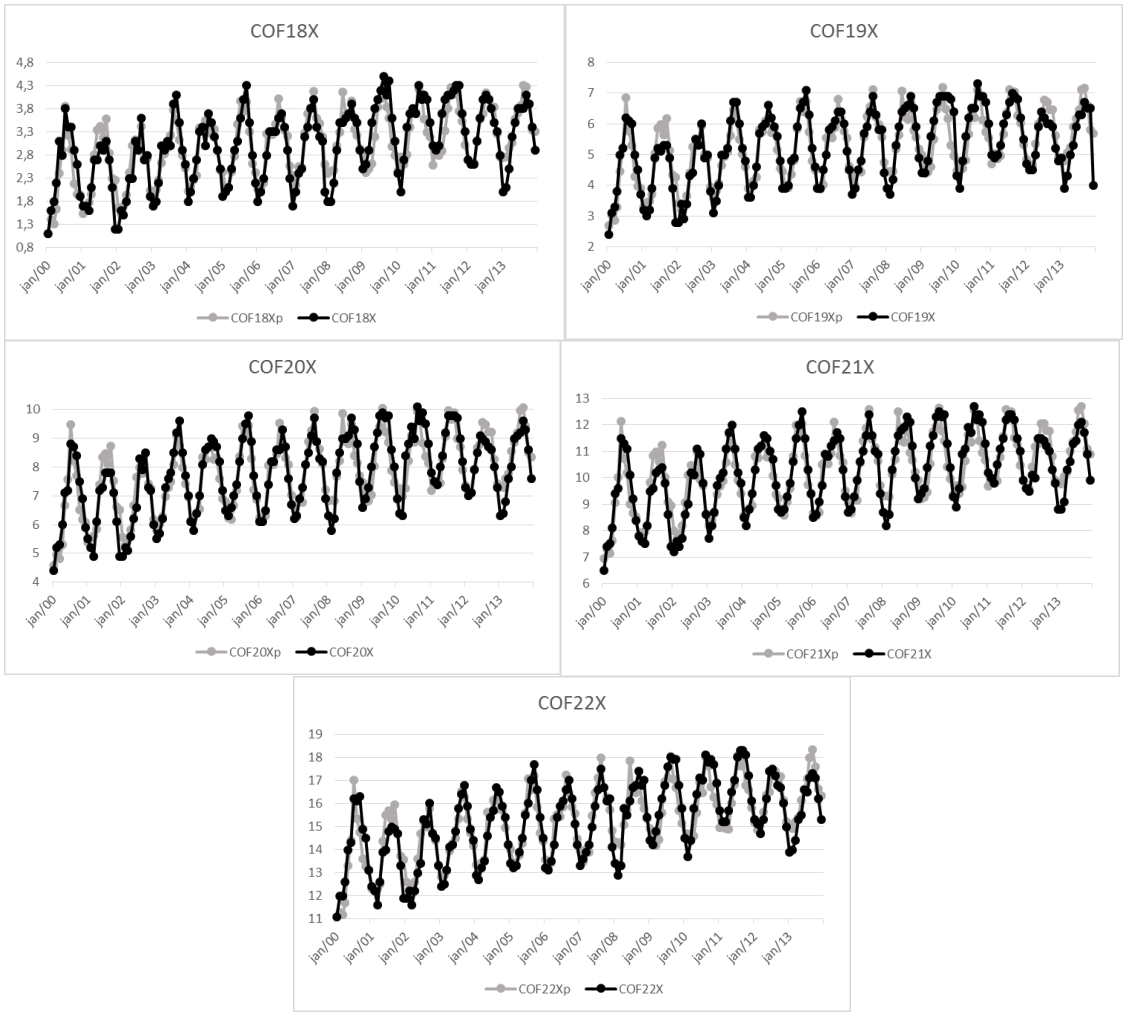

Figura 4: Valores observados nos sensores do pêndulo direto (COF**X) e valores previstos (COF**Xp) pelo modelo.

Os termômetros TSF12 e TSF13, que se mantiveram como variáveis do modelo, estão instalados na cota 50,25 m, mas ambos instalados na face interior do bloco (bloco do tipo gravidade aliviada), expostos à temperatura do ar no interior do bloco, assim, têm maior variação e contribuem na previsão dos deslocamentos do pêndulo direto. Os termômetros TSF15 e TSF16 estão localizados na cota 150,85 m, a montante e a jusante respectivamente. Assim, o TSF15 tem menor variação de temperatura que o TSF16, mas ambos contribuem na previsão dos deslocamentos do pêndulo direto.

Observa-se, nos gráficos dos deslocamentos (Figura 4), uma leve tendência de crescimento, porém esse comportamento é consistente com as tendências de deformação

de uma barragem ao longo do tempo. Finalmente, todos os resíduos do modelo têm média zero e estão distribuídos de forma aleatória.

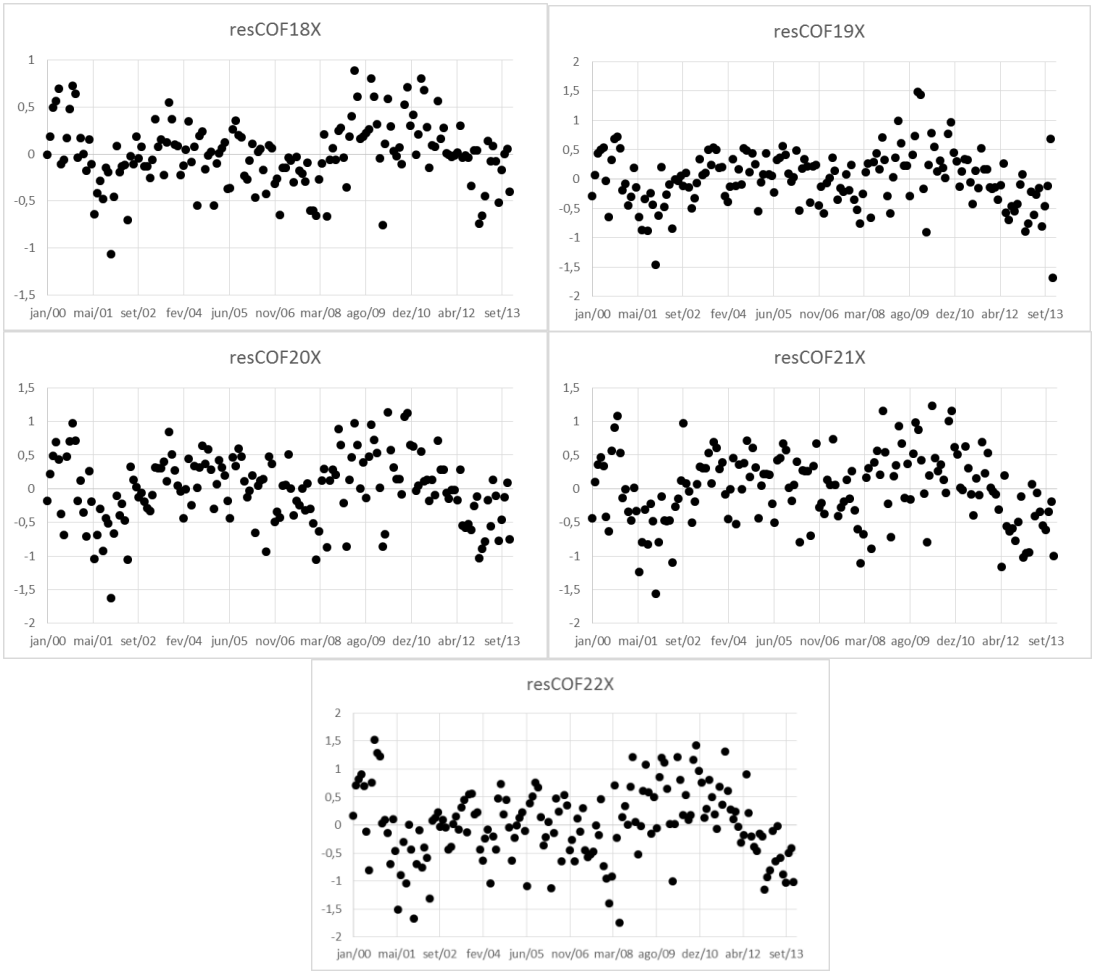

Figura 5: Resíduos do modelo para cada sensor do pêndulo direto.

CONCLUSÃO

O exemplo dado mostra que a regressão por mínimos quadrados parciais é útil para o tratamento de dados de monitoramento de barragens, uma vez que a multicolinearidade, presente nas variáveis independentes desses dados, impede o uso da regressão clássica. O método constrói um modelo que maximiza a correlação entre as variáveis dependentes (respostas) e as variáveis independentes (preditoras) observadas, e a grande vantagem é sua característica multivariada, proporcionando um estudo do comportamento de diversas variáveis simultaneamente.

A análise apresentada identifica que as variações do nível do reservatório e as leituras dos termômetros TSF11 e TSF14, instalados a montante, não contribuem significativamente para a previsão dos movimentos relativos do bloco, medidos pelo pêndulo direto. Assim, reduz o conjunto de 12 variáveis independentes para 6.

O modelo de regressão por mínimos quadrados parciais extrai, do relacionamento entre as 5 variáveis dependentes e 6 variáveis independentes, apenas 3 componentes. Essas 3 componentes explicam aproximadamente 95% da variabilidade das variáveis independentes, e explicam mais que 84% da variabilidade das variáveis dependentes. Isso mostra um bom potencial para o uso da regressão por mínimos quadrados parciais no tratamento de dados de monitoramento de barragens, reduzindo o número de variáveis a serem monitoradas.

Na literatura, uma gama de modelos univariados é proposta para dados de monitoramento de barragens, enquanto que ao estimar, conjuntamente, os parâmetros, no caso de modelos multivariados, obtém-se um ganho de eficiência dos estimadores, e leva-se em conta o relacionamento entre todas as variáveis. Embora os métodos estatísticos sejam frequentemente utilizados para modelar os dados em monitoramento de barragens, muitas pesquisas ignoram a presença de certas correlações entre as variáveis, o que inviabiliza o uso de modelos de regressão clássica. Portanto, investigações de técnicas que admitam correlações entre as variáveis são necessárias em aplicações dessa área.

Com o modelo ajustado, é possível prever as leituras dos sensores do pêndulo direto, conhecendo as variações do nível do reservatório e a temperatura dos termômetros de superfície. Essas previsões, quando comparadas às leituras reais, fornecem informação se houve mudança de comportamento com relação ao comportamento anterior, considerado estável.

Em trabalho futuro, pretende-se prever as leituras dos sensores do pêndulo direto e construir o intervalo de confiança para os estimadores, de modo a estabelecer limites de controle para as novas observações de deslocamentos.

REFERÊNCIAS

AHMADI-NEDUSHAN, B. **Multivariate Statistical Analysis of monitoring data for concrete dams**. Tese de Doutorado do Departamento de Engenharia Civil e Mecânica Aplicada, McGill University. Montreal, p. 211. 2002.

BUZZI, M. F.; DYMINSKI, A. S.; CHAVES NETO, A. **Avaliação das correlações de séries temporais de leituras de instrumentos de monitoração geotécnico-estrutural e temperatura ambiente na barragem de ITAIPU – Caso do pêndulo direto**. XXVIII Congresso Ibero Latino-Americano de Métodos Computacionais em Engenharia (CILAMCE). Porto: [s.n.]. 2007.

CHOUINARD, L.; ROY, V. **Performance of Statistical Models for Dam Monitoring Data**. Joint International Conference on Computing and Decision Making in Civil and Building Engineering. Montreal: [s.n.]. 2006. p. 9.

DE SORTIS, A.; PAOLIANI, P. Statistical analysis and structural identification in concrete dam monitoring. **Engineering Structures**, v. 1, n. 29, p. 110-120, Janeiro 2007.

DENG, N., WANG, J., e SZOSTAK, A. C. (2008) – **"Dam deformation analysis using the partial least squares method"**, 13th FIG International Symposium on Deformation Measurements and Analysis e 4th IAG Symp. on Geodesy for Geotechnical and Structural Engineering, Lisbon.

GARCIA, H.; FILZMOSER, P. **Multivariate Statistical Analysis using the R package chemometrics**. University of Technology: Department of Statistics and Probability Theory. Vienna, p. 71. 2011.

HAIR, J. F. et al. **Análise Multivariada de Dados**. 6ª. ed. São Paulo: Bookman, 2009.

ITAIPU BINACIONAL. ITAIPU BINACIONAL Barragem. **ITAIPU BINACIONAL**, 2015. Disponivel em: <http://www.itaipu.gov.br/energia/barragem>. Acesso em: 04 Fevereiro 2015.

LÉGER, P.; LECLERC, M. Hydrostatic, temperature, time-displacement model for concrete dams. **Journal of engineering mechanics**, v. 133 No. 3, p. 267-277, Março 2007.

LI, F.; WANG, Z.; LIU, G. Towards an Error Correction Model for dam monitoring data analysis based on Cointegration Theory. **Structural Safety**, v. 43, p. 12-20, Julho 2013.

MATA, J. Interpretation of concrete dam behaviour with artificial neural network and multiple linear regression models. **Engineering Structures**, v. 33, n. 3, p. 903-910, Março 2011.

MEVIK, B.-H.; WEHRENS, R. The pls package: principal component and partial least squares regression in R. **Journal of Statistical Software**, v. 18, n. 2, p. 1-24, 2007.

MORELLATO, S. A. **Modelos de regressão PLS com erros heteroscedásticos. Dissertação de Mestrado em Estatística.** Universidade Federal de São Carlos - UFSCar. São Carlos, p. 49. 2010.

R CORE TEAM. **R:** A language and environment for statistical computing, 2014. Disponivel em: <http://www.R-project.org>. Acesso em: 01 Setembro 2014.

TOBIAS, R. D. **An introduction to partial least squares regression**. 20th SAS User Group International Conference (SUGI). Orlando: [s.n.]. 1995.

WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics. **Chemometrics and intelligent laboratory systems**, v. 58, n. 2, p. 109-130, 2001.

XI, G. Y. et al. Application of an artificial immune algorithm on a statistical model of dam displacement. **Computer & Mathematics with Applications** , v. 62, n. 10, p. 3980-3986, Novembro 2011.

YU, H. et al. Multivariate analysis in dam monitoring data with PCA. **Science China Technological Sciences**, v. 53, n. 4, p. 1088-1097, 2010.