13th FIG Symposium on Deformation Measurement and Analysis
4th IAG Symposium on Geodesy for Geotechnical and Structural Engineering

LNEC, LISBON 2008 May 12-15

INTRODUCTION AND ANALYSIS OF COMMONLY USED NON-PARAMETRIC MODELS OF DAM DEFORMATION IN CHINA

Nianwu DENG[1, 3], Jian-Guo WANG[2], Anna SZOSTAK-CHRZANOWSKI [3],
and Yun ZHANG [3]

[1]*State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, China*

[2]*Department of Earth and Space Science and Engineering, York University, Canada*

[3]*Canadian Centre for Geodetic Engineering, University of New Brunswick, Canada*

Abstract: Dam deformation analysis is an important part of dam safety monitoring. Generally dynamic models of dam deformation analysis include: parametric models and non-parametric models. In China, large amount of research activities in data analysis is in developing new non-parametric models such as stepwise regression (SR), partial least-squares (PLS) regression, artificial neural network (ANN), time series (TS), and grey system (GS). The methods have been applied in dam deformation analysis, and have showed good results in modelling and predicting deformation. The principles and merits of above listed non-parametric methods are discussed.

1. INTRODUCTION

A total of 25821 dams were registered by China Commission on Large Dam (CHINCOLD) by August 2002. Since then, more than 100 dams have been built annually. China retains the largest number of dams more than any other countries worldwide. In 1980’s China initiated a program which increased the dam safety evaluation. Dam Safety Supervisory Centre (DSSC) under State Electricity Regulatory Commission (SERC, P. R. China) and Dam Safety Management Centre (DSMC) under The Ministry of Water Resources (MWR, P. R. China) were formed in 1985 and 1988, respectively. Dam safety management has been guided into the path of a legal system. A series of laws and regulations for dam safety management were set off while the standards and specifications for dam safety management were issued. Moreover, the technical level in theoretical and practical respects, for instance sensor (instrumentation) installation, original observation, data analysis and safety evaluation, has comprehensively been improved in China.

Dam original observation is one of the essential links to effective measures in dam safety management. Based on the observation data, one can verify the design performance during construction and life of the structure. Raw observation data provide valuable information for the interpretation of dam behaviour. However, the raw observation data cannot reveal displacement field and trend of the dam deformation. It is necessary to perform a comprehensive analysis that will identify, extract, and generalize the various deformation phenomena from the huge volume of raw data.

Two types of dynamic models have been formulated for data analysis of the dam deformation monitoring non-parametric models, which are based on mathematical and statistical theory, and parametric models, which are based on principles of continuum mechanics.

Non-parametric models are based on mathematical and statistic forecast algorithms. Two types of non-parametric models exist. One type is based on the functional relationship between independent variables – the environmental variables and dependent variables – the displacements. This type of models can interpret the internal operational causes and effects. Such models include: multiple regression (MR), stepwise regression (SR), principal component regression (PCR), partial least square regression (PLSR), and artificial neural network (ANN) etc. The second type of models is based on setting up statistic rules of dependent variables using linear statistic models by themselves, rather than by other environmental variables. They do not model the relationship between causes and effects. This type of models includes: time series (TS) and gray system (GS). The deformation prediction of these models is based on the extracted information from the available raw observations of deformation, although the modelling process is performed in different ways.

Parametric models are based on analysis of the observation data by applying the principle of continuum mechanics. First, the deterministic relationship between dependent variables and independent variables is constructed based on mechanical principles. Next, the linear statistics is applied to adjust the assumptions and parameters through the verification of observations with the calculated values.

In most of data analyses the influence of hydrostatic pressure (including silt pressure, wave pressure, sediment pressure and uplifting pressure) and temperature on the displacements is mainly considered.

The development of the data analysis based on non-parametric models has made rapid progress in China. This paper focuses on introducing some algorithmic developments of the above models such as SR, PLSR, ANN, TS, and GS and applications to dam deformations in China.

2. REVIEW OF NON-PARAMETRIC MODELS IN CHINA

2.1. The Stepwise Regression

2.1.1. Principle of stepwise regression

The traditional regression models include simple linear regression, multiple linear regression, principal component regression and stepwise regression etc. They can be used for the analysis of the relation between dependent variables and independent variables. Among them, the stepwise regression is the commonly applied method in data analysis of dam deformation.

The most commonly used multiple regression modelling approaches are: the forward (entry) method, the backward (removal) method and the (forward) stepwise method. The forward method starts with no predictor variable in the model, and adds the one that gives the highest correlation with or effect on the response variable and so on. Usually one provides a stopping criteria related to a certain ***F***-test statistic. The backward method starts with all predictor variables in the model and removes variable that has the computed value on the removal statistic less than the critical value and so on, The algorithm stops once none of the remaining variables statistically has insignificant effect on the current model. The stepwise regression (Li, 1989) is a combination of the forward and backward methods that removes and adds the predictive variables to the regression model through a sequence of ***F***-tests so that the best subset of the predictors can be identified. The stepwise regression performs its 1st step exactly as in the forward method. At any subsequent step, whenever two or more variables have been included in the model, one it will follow up every forward entry with a backward removal that deletes any variable that statistically becomes insignificant to the up-to-date model. This procedure will be repeated until no new predictor variable needs to be included.

2.1.2. Assessment of stepwise regression

The stepwise regression is a traditional and classic algorithm in which only the variables having significant effect on the response variable are considered in the model. Hence, its analytical results can correspond to reality comparatively well. Stepwise regression method is commonly used all over the world (Nadushan, 2002), and also very often applied in data analysis of dam deformation in China.

The stepwise regression algorithm might add the predictive variables to the model, which are very weakly correlated (Zhou & Li, 2004). It will result in incorrect model so that the optimal model cannot be reached because certain correlation, more or less, exists among them. This is disadvantageous to the model explanation.

The conventional multivariate analysis approaches including SR cannot solve the difficulty brought by the correlation among the predictive variables and has influence on an existence of the type of correlation and its disadvantages to modelling of the stepwise regression. Therefore, more attention should be paid to the selection of environmental variables. Furthermore, the stepwise regression appears incapable of analyzing the short time period of data, especially for the data with the number of observation epochs less than the number of predictor variables.

2.2. Partial least square regression (PLSR)

2.2.1. Principle of PLSR

PLSR is a relatively novel multivariate analysis method, which integrates multiple linear regression, canonical correlation analysis and principal component analysis (Wang, 1999). This method can combine the type of data analysis characterized by modelling and forecasting of deformation with the non-analytical type of the cognitive data analysis. In case that strong correlation exists among independent variables, PLSR can result in more reliable solutions than other type of the regressions. Through the decomposition and filtering of deformation observation data from a monitoring system, PLSR algorithm extracts the group of the predictor variables that are best in explaining of the response variables. PLSR can efficiently overcome the difficulty caused by the colinearity of variables at the data modelling stage, and is also capable of modelling the short time period of observation data.

Assume that one has q dependent (response) variables y_1, y_2… y_q and p independent (predictive) variables x_1, x_2, …, x_p. The data from n observation points can are available denoted as $X = [\ \mathrm{X}_1, X_2, \ldots, X_p\ ]_{n\times p}$ and $\mathrm{Y} = [\ \mathrm{Y}_1, Y_2, \ldots, Y_q\ ]_{n\times q}$ respectively, in which are $X_k = (\ x_{1k}\ , x_{2k}, \ldots, x_{nk}\ )^T$ and $Y_j = (\ y_{1j}, y_{2j}, \ldots\ y_{nj}\ )^T$. As start, the latent component t_1 and u_1 are extracted from X and Y, respectively, as the linear combinations of predictive variables of x_1, x_2, …, x_p and the response variables of y_1, y_2, …, y_q based on the following requirements:

(1). t_1 and u_1 should carry as much information as possible from $\boldsymbol{X}$ and $\boldsymbol{Y}$,

(2). t_1 and u_1 should possess the maximal correlation.

These mean that t_1 and u_1 will represent $\boldsymbol{X}$ and $\boldsymbol{Y}$ as well as possible and t_1 can provide the strongest capability to interpret u_1.

After the extraction of t_1 and u_1, t_1 will be regressed with respect to $\boldsymbol{X}$, and u_1 with respect to $\boldsymbol{Y}$ as well. Then the further latent components t_2 and u_2 will be extracted using the residual information of $\boldsymbol{X}$ and $\boldsymbol{Y}$ after t_1 and u_1. This procedure will be continued until one reaches the latent components t_m and u_m that can provide a satisfied regression together with the past extracted components. And at the end, the regression functions will be given by the original variables.

How to determine a better regression is even more important. In many cases, PLSR does not need all of the latent components to construct the regression model, but only select the first $\boldsymbol{m}$ components for $m \le rank(X)$ under certain cutoff criteria as the PCA does. These $\boldsymbol{m}$ components could construct a satisfied predictive model while the follow-up components cannot make any significant contribution to the interpretation of the response variable vector $\boldsymbol{Y}$. In this case more components will not improve cognition of statistic trends in the data, and may mislead by false predictive conclusion. How many components exactly need to be included in PLSR? The decision for a new latent component can be made through observing the improvement of the predictive ability of the PLSR model each time after the extraction of each latent component. The cross validation can be used for this purpose (Wang, 2002).

2.2.2. Evaluation of PLSR

The PLSR method has been applied to the data analysis of deformation monitoring (Deng, 2001; Xu & Wu, 2001; Qiu & Pan, 2005). In general, the PLSR algorithm enables to reach the highest correlation of the response variables with the predictive variables without deleting any of the variables. It can overcome the difficulty of the colinearity very well while it gives fully expression to the contribution of each of the individual predictors to the response variables. The analytical expression of the regression models allows having the displacements, i.e. the response variables, well interpreted. Furthermore, PLSR is capable of modelling of multiple predictors and multiple response variables. This is very advantageous to the interpretation of the dam deformation with all of the points as a whole globally.

2.3. Artificial Neural Network (ANN)

2.3.1. Principle of ANN

The artificial neural network becomes popular. It requires interdisciplinary approach because it involves a wide range of disciplines such as mathematics, physics, computer science, electronics, and biology etc. The ANN is opening up a vast range of applications.

Some particular types of neural network models are: Feed-forward Back-Propagation (BP) network, Radial Basis Function (RBF) network, and some modified algorithms based on them. They have been applied to the data analysis of dam deformation monitoring. The basic principle of feed-foreword BP network will be given briefly below.

Analogues to a multiprocessor computer system, a Neural network is a form of a system with simple processing elements, a high degree of interconnection, simple scalar messages and adaptive interaction between elements, The feed-forward back propagation (BP) network is a very popular model in neural networks (Rao, 2003). In multi-layer feed forward networks, the processing elements are arranged in layers and only the elements in adjacent layers are connected. A minimum of three layers is needed, which are the input layer, the middle or hidden layer and the output layer (Zhang, 1993; Zhao, 1999). Figure 1 gives a three-layer feed-foreword BP network that is composed of *n* artificial neurons in $X = (x_0, x_1, \ldots, x_{n-1})^T$, $X \in R^n$ in input layer, n_1 artificial neurons in $X' = (x'_0, x'_1, \ldots, x'_{n1-1})^T$, $X' \in R^{n1}$ in hidden layer; *m* artificial neurons in $Y = (y_0, y_1, \ldots, y_{m-1})^T$, $Y \in R^m$ in output layer. w_{ij} and θ_j are weights and thresholds between input and hidden layers, respectively. w'_{jk} and θ'_k are weights and thresholds between hidden and output layers, respectively. Here are $i = 0,1,\ldots,n-1$, $j = 0,1,\ldots,n_1-1$ and $k = 0,1,\ldots,m-1$. The output of each of the artificial neurons must meet the following equations:

$$\begin{cases} y_k = f\left(\sum_{j=0}^{n_1-1} w'_{jk} x'_j - \theta'_k\right) \\ x'_j = f\left(\sum_{i=0}^{n-1} w_{ij} x_i - \theta_j\right) \end{cases} \tag{1}$$

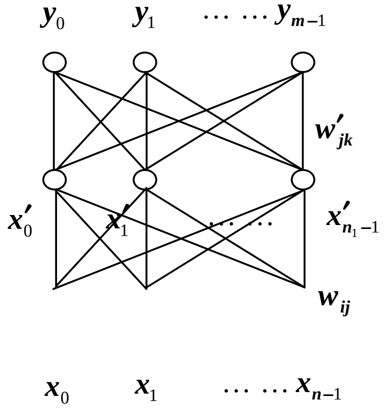

Figure 1 - three-layer feed-foreword BP network

In order to determine the weights w_{ij} and w'_{jk} and thresholds θ_j and θ'_k, a neural network need to be trained by using p real samples: (x^1, y^1), (x^2, y^2), …,(x^p, y^p). Once the weights and thresholds are successfully obtained from the training, they will be tested through the different samples other than the training samples.

2.3.2. Evaluation of ANN model

The application applying the ANN method to the data analysis of dam deformation can be found in (Deng et al, 2001; Zhang et al 2003, etc.). The ANN method can overcome the drawback that the functional models must be linear or nonlinear with original observables, has a strong self-adaptive capability and fault-tolerance, and does not have any limitation to the number of the input and output artificial neurons. One can achieve a satisfied data fitting and forecast effect with the ANN (the forecast effect is not as good as the data fitting effect, but meets the practical requirements). Because of none analytical solution from the ANN analysis, it lacks of the ability of the deformation interpretation linked input artificial neurons to the output artificial neurons; however some authors attempted to provide this type of the interpretation (Xu et al, 2003).

2.4. Time Series (TS) model

2.4.1. Principle of Time Series model

A time series data is a sequence of consecutive measurements taken at (often equally spaced) time intervals about the same phenomenon (Tian, 2001). Time series analysis consists of two groups of methods: frequency domain methods and time domain methods. The deformation analysis often employs the latter ones. Time series analysis attempts to analyze the collected data of a phenomenon in the past towards to forecasting its future behaviour. That is, the regular pattern how a phenomenon changes with the time is revealed based on its historic data, is then extended to extrapolate its future changes in order to forecast the future behaviour of the phenomenon.

A time series of data may be influenced by many factors. The factors, which have their effect on the data over the long period of time, will play the decisive roles and make the time series data show certain trend with regularity. The other factors, which have their impact on the data only over the short period of time, play the non-decisive roles and create certain high frequency irregularity in the data. A time series data can be an composition of any of the combination of the following four attributes: (1) the trend - a series data may show certain constant, increasing, or decreasing trend over time intervals; (2) the seasonality – a time series data may repeatedly show a certain regular pattern over certain fixed time intervals; (3) the circulatory – the data may repeatedly show a certain regular pattern, but not over certain fixed time intervals; and (4) the random – the time series data may show certain irregular fluctuation because of accidental environmental condition changes over time.

There are two categories of prediction approaches for time series: the deterministic time series models and the stochastic time series models. A deterministic time series model will fit a specific deterministic function of time to the time series data. Different types of data behaviours are imitated by different types of functions. At the end, they are composed together to provide a combined expression of the time series. There are the trend forecasting method, smoothed forecasting method and decomposition method and so on. A stochastic time series model explores the functional structure through studying the correlation between the variables in time so that the future trend of a phenomenon can be predicted.

2.4.2. Evaluation of Tine Series Model

The forecasts according the time series data are based on modelling the relations between the deformation variables and the time. For instance, there is certain correlation between the water levels and temperatures in time so that a model composed by them is able to forecast the dam deformation. In practice, small or medium-sized dams lack of certain observation activities for some reasons, for example no temperature observation, the time series analysis methods show good imitation and forecast ability to certain extent (Zhu, 2006).

Although there are good correlations between time intervals and hydrostatic pressure or thermal pressure, correlation between deformation and hydrostatic or thermal pressure are not absolutely replaced by correlation between deformation and time intervals, so TS model has some limitations.

2.5. Grey System Model (GSM)

2.5.1. Principle of Grey System Model

The grey system model, which was first proposed by Julong Deng (1996) in 1982, has successfully been applied to the analysis of uncertain systems for multi-data inputs, discrete and insufficient data. In general, the predictive factors that have influence on dam deformation cannot be unknown clearly. Presently, mostly two types of GM models (GM (1, 1) and GM (1, N)) are being used in data analysis of dam deformation.

Using the GM (1, N) model, given series of measurements ($x_i^{(0)}$) =($x_1^{(0)}$, $x_2^{(0)}$, …, $x_n^{(0)}$), ($x_i^{(1)}$) as a first-order series can be generated from the raw observed data ($x_i^{(0)}$) after repeatedly accumulated generating operation (denoted $x_i^{(1)}(k) = \sum_{m=1}^{k} x_i^{(0)}(m)$). According to the theory, the differential equation of the whitening of $x_i^{(1)}(k)$ over the unit time interval is given by:

$$\frac{dx_i^{(1)}}{dt} + ax_1^{(1)} = b_1 x_1^{(1)} + b_2 x_2^{(2)} + \cdots + b_{N-1} x_N^{(1)} \qquad (2)$$

which is defined as GM (1, N), is the differential equation of first-order with N variables. The parameter vector $\beta = (a, b_1, b_2, \ldots, b_{N-1})^T$ in (2) can be estimated by applying the method of least squares. The imitation and prediction can be carried out by using the inverse of ($x_i^{(1)}$), once the parameters are obtained.

2.5.2. Evaluation of Grey System Model

Grey system model can be applied when the system has limited number of observation data. The grey system model can be advantageous to the data fitting and deformation forecasts with the satisfied accuracies (Yin, et al, 2002).

Using grey system model the degree of the contribution made by individual environmental variables may be obtained, therefore qualitative analysis, not quantitative analysis, can be performed towards to the deformation interpretation.

3. CONCLUSIONS

Dam deformation data analysis is essential to the dam safety monitoring. In China the SR is widely used in data analysis of dam deformation because it has the advantages to be simple and practical. The PLSR is a novel kind of multivariate data analysis method, which integrates multiple linear regression method, canonical correlation analysis with principal component analysis. It can overcome adverse effect of high colinearity between the predictive variables. At same time, PLSR is able to deal with the short-series samples. The ANN does not require that a model must be the linear or nonlinear functions of the original observation variables, and possess a strong self-adaptive and failure-tolerance capacity. The TS can be applied when there is a shortage of information (such as lack temperature information). The GS can only model small scale of samples.

References

Deng, Julong (1996): *the basic method of grey system*, Huazhong University of Science and Technology Press, China, 1996.

Deng, Nianwu; Qiu, Fuqing and Xu, Hui (2001): *Application of BP model to data analysis of earth-rock dams*, Journal of Wuhan University of Hydraulic and Electric Engineering, China, Vol. 34, No. 4, pp.17-20, April 2001.

Deng, Nianwu (2001): *Application of Partial least square regression to dam deformation monitoring data analysis*, Dam Observation and Geotechnical Tests, China, Vol. 25, No. 6, pp.16-18, June 2001.

Li, Zhenzhao (1989): *Data analysis of concrete observation data*, Hydrowater and Hydropower press, China, 1989.

Nadushan,B.A (2002): *Multivariate statistical analysis of monitoring data for concrete dam*, PHD dissertation, December 2002.

Qiu, Xiaodi and Pan, Lin (2005): *A partial least square regression method application to dam safety monitoring modeling and its realization on computer*, Dam observation and Geotechnical Tests, China, Vol.29, No.6, pp.58-61, June 2005.

Rao, M. Ananda and Srinivas, J. (2003): *Neural Networks – Algorithms and Applications*, Alpha Science International Ltd., ISBN 1-84265-131-5, Pangbourne RG8 8UT, UK, Copyright 2003.

Tian, Zhen (2001): *Time Series Theory and Methods*, 2nd edition, Higher education press, China, 2001.

Welsch, W., and Heunecke, O.(2001): *models and terminology for the analysis of geodetic monitoring observations*, Official Report of the Ad-Hoc Committee of FIG Working Group 6.1, http://www.fig.net/pub/figpub/pub25/figpub25.htm.

Wang, Huiwen (1999): *Partial least square regression method and application*, national defense industry press, China, 1999.

Xu, Hongzhong and Wu, Zhongru (2001): *Partial least square regression and its application to dam safety monitoring*, Dam Observation and Geotechnical Tests, China, Vol. 25, No. 6, pp.22-23, June 2001.

Xu, Hongzhong; Wu, Zhongru; Shi, Bin and Wang, Jian(2003): *Neural network method for determining the component proportion of dam effect-variable*. SHUILI XUEBAO/Journal of Hydraulic Engineering, China, Vol. 34, No.6, pp. 111-114, June 2003.

Yin, Zhizheng and Zhang, Jiasheng (2002): *The gray model application for main dam deformation monitoring and forecast*, GX Water resources & Hydropower Engineering, China, 2002(4), pp.13-15.

Zhang, Liming (1993): *Modeling and appliance of artificial neural network*, Fudan University press, China, 1993.

Zhao, Lingmin (1999): *Multi-layer Feed-forward artificial neural network*, Yellow River Hydrowater Press, China, 1999.

Zhang, Xiaochun; Xu, Hui; Deng Nianwu and Chen, Renxi; et al (2003): *Application of a radial function neural network model to data processing technique of dam safety monitoring*, Engineering Journal of Wuhan University, China, Vol. 36, No. 2, pp.33-36, April 2003.

Zhu, Weibing (2006): *The Time-series superposition model for dam monitoring analysis, hydropower Automation and dam monitoring*, China, Vol.30 No.5, pp.52-55, October 2006.

Zhou, Wenfang and Li Min (2004): *Discussion on shortcoming of stepwise regression analysis*, Northwest Water Power, China, pp.49-50, April 2004.

Corresponding author contacts

Nianwu DENG
deng@unb.ca
Department of Geodesy and Geomatics Engineering, University of New Brunswick
Fredericton, N.B., E3B 5A3 Canada