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

LNEC, LISBON 2008 May 12-15

DAM DEFORMATION ANALYSIS USING THE PARTIAL LEAST SQUARES METHOD

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

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

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

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

Abstract: Main problem in data analysis is a construction of mathematical models relating environmental variables and patterns of deformation. In case of the dam environmental variables are temperature, water level in a reservoir, etc. The most commonly used method of data analysis is statistical modelling of the data. The partial least-squares regression (PLSR) is a statistical method which finds a linear model describing some predicted variables in terms of other observable variables. The partial least-squares regression (PLSR) is a multivariate statistical algorithm, which can overcome some of the shortcomings of other approaches, for instance, the multiple correlation among independent variables. PLSR methodology is presented on the example of an earth dam located in central China. A three dimensional deformation analysis for a single point on the dam is performed. The analysis consists of a data fitting, deformation prediction, and contribution analysis of individual factors. The presented in the paper research shows that PLSR results are more reliable and have the better integrity than the other methods.

Key words: Data analysis, deformation, partial least squares regression, monitoring.

1. INTRODUCTION

Data analysis is one of the essential components of interpretation of dam deformation monitoring, because the data from multiple observation epochs must be analysed (Wu, 1990; Li, 1989). Information of the time and spatial connections of the observations, the specific deformation characteristics, and the weak links should be identified so that the deformation process and trends can be identified. The goal of the data analysis is to generate the information which can be used in a physical interpretation of deformations and in prediction of behaviour of the analysed or similar engineering structure.

There may be a large volume of observation data of points located on the surface and within a dam structure. Statistical methods are often employed to model the data in order to describe

the deformation patterns. Commonly, the water level, temperature, time, and other quantities are directly measured and they are considered as observable variables. There are inevitably certain multiple correlations among these variables and deformation so that the traditional statistic approaches may run into numerical difficulties such as model deficiency.

The partial least-squares regression (PLSR) is a statistical method which finds a linear model describing some predicted variables in terms of other observable variables. The partial least-squares regression is a multivariate statistical algorithm, which can overcome some of the shortcomings of other approaches, for instance, the multiple correlation among independent variables. The partial least squares regression has been applied to multivariate data analysis in chemistry, economics, medicine, psychology, and some other disciplines (Ren, 1997; Wang, 1999; Rosipal and Krämer, 2006). PLSR generalizes and combines characteristic features from multivariate regression (MR), canonical correlation analysis (CCA), and principal component analysis (PCA) without imposing their restrictions. It is suitable for the situation when an applied method needs to predict a set of dependent variables from a large set of independent variables, especially in case the independent variables are highly collinear. The solution from PLSR is more effective and more reliable comparing to MR, which in the past was widely employed in this type of applications. PLSR can extract the integrated independent variables that may interpret the dependent variables by decomposing and filtering the data. The data can be modelled with the better fitting and predictive effects and without the limitation to the number of the sample points. PLSR can be also used in case of smaller number of data, as it happens often at the early stage of the monitoring.

2. PARTIAL LEAST SQUARES

Assume that one has q dependent (response) variables $b_1, b_2 \ldots b_q$, and p independent (predictor) variables $a_1, a_2 \ldots a_p$. In order to study the relationship between the response variables and the predictor variables, the data from n observation points are available denoted as $\boldsymbol{A}_{n\times p} = [\, A_1, A_2 \ldots, A_p \,]$ and $\boldsymbol{B}_{n\times q} = [\, B_1, B_2 \ldots, B_q \,]$ (Wang, 1999).

From the ordinary multivariate linear regression, the least squares solution can be given by

$$\hat{B} = A(A^T A)^{-1} A^T B \qquad (1)$$

under the assumption that the sample data is satisfied with the Gauss–Markov theorem. $\boldsymbol{B}$ is called the least-squares solution of $\boldsymbol{B}$ wherever $\boldsymbol{A^T A}$ is invertible. This solution does not exist if $\boldsymbol{A^T A}$ is rank defect, for example, in case there is high collinear in $\boldsymbol{A}$. A rank defect $\boldsymbol{A^T A}$ can practically bring a series of complications. However, the PLSR will be not affected.

Principally, the PLSR extracts new factors from both the variation of $\boldsymbol{A}$ and $\boldsymbol{B}$. In PLSR only the variation of $\boldsymbol{A}$ is used, not as in the case of PCA analysis. These new factors are commonly called latent variables or components that will play the same role as $\boldsymbol{A}$ and $\boldsymbol{B}$. In the beginning of the analysis, the latent components t_1 and u_1 are extracted from $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively, as the linear combinations of $a_1, a_2 \ldots a_p$ and $b_1, b_2 \ldots b_q$ based on the following requirements:

(1) t_1 and u_1 should carry as much information as possible from the raw data $\boldsymbol{A}$ and $\boldsymbol{B}$,

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

These means that t_1 and u_1 will represent $\boldsymbol{A}$ and $\boldsymbol{B}$ 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 $\boldsymbol{A}$, and u_1 with respect to $\boldsymbol{B}$ as well. Then the further latent components t_2 and u_2 will be extracted using the residual information of $\boldsymbol{A}$ and $\boldsymbol{B}$ after t_1 and u_1. This procedure will be continued until one reaches the latent components t_m and u_m, which can provide a satisfied regression together with the past extracted components.

How to determine the better regression equations is even more important. In many cases, PLSR does not need all of the latent components to construct the regression model, but select first m components for $m \leq rank(\boldsymbol{A})$ under certain cut-off criteria as the PCA does. These m components could construct a satisfied predictive model whilst the follow-up components cannot make a significant contribution to the interpretation of the response variable vector $\boldsymbol{B}$. In this case more components will not bring any more beneficial influence on the cognition of statistic trends in the data, and may mislead by false predictive conclusion (Wang , 1999).

One can observe the predictive improvement to decide how many latent components should be selected each time after the extraction of each latent component. Generally, one can construct individual regression equations by taking out each of the observation points one by one sequentially and use the corresponding models to predict those points after h latent components are extracted the goodness of the predictions can be used for the modeling evaluation.

Consider that b_j ($j = 1,2,...,p$) is predicted as $\boldsymbol{B}_{hj(-i)}$ at the i-th data point. For each of the response variables with all of n points the sum of the squares of the prediction errors are defined as

$$PRESS_{hj} = \sum_{i=1}^{n} (B_{ij} - B_{hj(-i)})^2 \tag{2}$$

The total prediction error $PRESS_h$ for B is defined as the total sum of the errors from all of the response variables:

$$PRESS_h = \sum_{j=1}^{p} PRESS_{hj} \tag{3}$$

Obviously, the magnitude of this error is sensitive with the changes of the points and will also vary with the goodness of the predictive ability of the models based on the available data. Besides, a model can be constructed from all the data points based on the h latent components.

The predicted value of each of the response variable for the i-th data point is written as B_{hji}. Then one can define SS_{hj} as the sum of the squares of the residuals:

$$SS_{hj} = \sum_{i=1}^{n} (B_{ij} - B_{hji})^2 \tag{4}$$

The total error is given by

$$SS_h = \sum_{j=1}^{p} SS_{hj} \tag{5}$$

Generally, total prediction error $PPESS_h$ is larger than total error SS_h, SS_h is smaller than SS_{h-1}.

The next step is to compare total error SS_{h-1} with total prediction error $PPESS_h$. SS_{h-1} is the total regressive error based on all of the data points, but using the first h-1 latent components. $PPESS_h$ is derived based on the first h components and affected by the disturbance error of the sample data. If $PPESS_h$ is reasonably smaller than SS_{h-1}, the predictive accuracy is considerably improved by adding one more latent component. As a general rule, it is beneficial if one more latent component is extracted whilst

$PPESS_h / SS_{h-1} \leq 0.95^2$.

Otherwise, it is not necessary to take an account into the next latent component because the total prediction error can not significantly be reduced.

3. EXAMPLE OF AN EARTH DAM

An example given is the three dimensional deformation analysis of an earth dam situated in the central China. The dam is build of the clay soil using the slope wall structure. The dam is 1223 m long, 56 m high, and has the maximum elevation of 162 m (MSL) of its top. Following factors have impact on the dam deformation: geometry of the dam, construction materials, construction method, topography and geological condition of the dam foundation, and the changes of the water level.

Three dimensional analysis of the monitored deformation from 1975 year to 1982 year of a point on the surface of the dam is performed. The monitored point is located on the axis of a dam. In the analysis, X is the horizontal displacement in a perpendicular direction of the dam axis (parallel to the river flowing direction), Y is the horizontal displacement perpendicular to X, and Z is the vertical displacement.

Eight predictor variables were assumed based on the characteristics of the earth dam and practical experience as follows:

- aging effect: θ, $\ln \theta$, $\theta/(\theta+1)$, θ^2, $\theta^{-0.5}$ (θ is the time in year),
- water level: H, H^2, H^3 (H is the upstream water level in meter).

The variance–inflation factors of eight selected variables and deformation are shown in the Table 1.

θ	$\ln \theta$	$\theta/(\theta+1)$	θ^2	$\theta^{-0.5}$	H	H^2	H^3	X	Y	Z
1.8×10^5	1.3×10^6	6.9×10^6	10	3.1×10^6	1.3×10^3	0.06	3.3×10^{-7}	0.05	0.08	0.09

Table 1: The variance-inflation factors of the variables

In general, there is high collinearity among the variables if the variance-inflation factor is bigger than 10 (Wang, 1999). Hence, the ordinary least-squares regression here encounter with difficulties. The cross validation with respect to the response variables is shown in the Table 2.

	X				*Y*		*Z*		
h	1	2	3	4	1	2	1	2	3
Q_{hk}^2	0.626	0.456	0.347	-0.255	0.671	-0.034	0.824	0.730	-0.121

Table 2: The cross validation

The critical value of $Q^2_{hk}(=1\text{-}PRESS_h/SS_{h\text{-}1})$ is equal to 0.0975 at the 95% confidential level. The best predictive models for *X*, *Y* and *Z* can be constructed by selecting h = 3, 1, 2, respectively. The 1st extracted latent components t_1 and u_1 are correlated at –0.939. Their relationship is given in Figure 1.

The correlation among the extracted components and predictor variables are given in the Table 3. As can be seen in the Table 3, 79.7% of the information from the raw data is used after the extraction of the very first latent components t_1 and u_1. Another 19.7% of the information left in the residuals is added to the analysis after the 2nd latent components t_2 and u_2 are extracted. But the extraction of the 3rd latent components t_3 and u_3 can contribute only 0.5% information to the data analysis. In total, 90.9% of the information in *X* can be interpreted by using t_1, t_2 and t_3. t_1 can represent *Y* up to 69.7% while 96.2% of the information in *Z* can be extracted by t_1 and t_2.

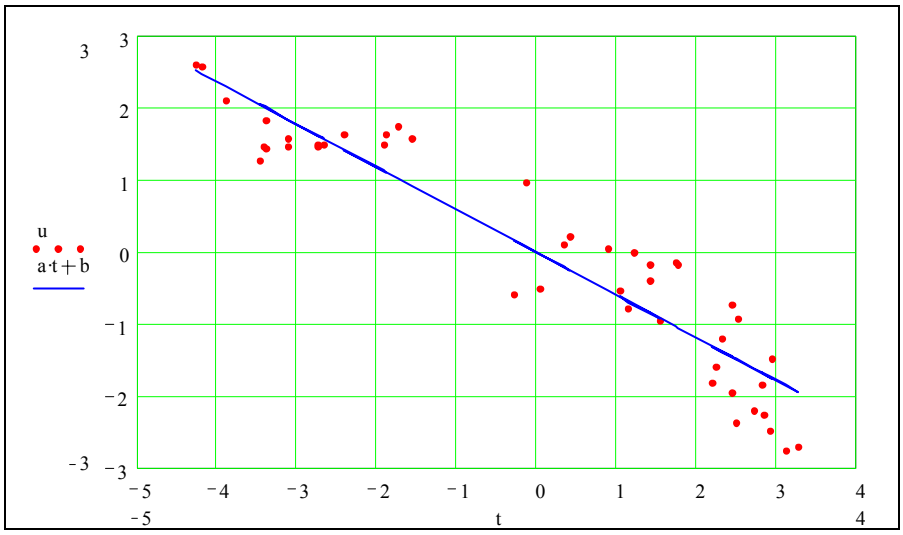

Figure 1: Relation between t_1 and u_1

	θ	ln θ	$\theta/(\theta+1)$	θ^2	$\theta^{-0.5}$	H	H^2	H^3	X	Y	Z
t_1	-0.982	-0.967	-0.946	-0.982	0.953	-0.750	-0.755	-0.760	-0.813	-0.835	0.917
u_1	0.957	0.968	0.967	0.935	-0.968	0.555	0.562	0.569	0.924	0.853	-0.957
t_2	-0.177	-0.255	-0.312	-0.109	0.295	0.661	0.655	0.649	-0.419	0.154	0.349
u_2	-0.127	-0.276	-0.295	-0.034	0.274	0.554	0.548	0.541	-0.536	0.229	0.340
t_3	0.065	-0.024	-0.085	0.154	0.067	-0.021	-0.021	-0.022	-0.269	0.077	0.044
u_3	0.042	-0.019	-0.057	0.111	0.046	-0.014	-0.015	-0.016	-0.375	0.153	0.034

Table 3: The correlation among the extracted components and the predictor variables

The displacements and the fitting curves are given in Figure 2, 5 and 8. The residuals are plotted in Figure 3, 6 and 9. The decomposed fitting values with respect to the aging variables and the water level variables, respectively, are drawn in the Figure 4, 7 and 10.

In the example, all of the predictor variables made their contributions to the interpretation of the displacements in certain extent. The displacements in three directions are positively proportional to the water level. The aging variables had relatively strong influence on the *X* component whilst the water level variables had relative small impacts on it. The effect on the other two components from the water level appears periodically. The displacements in *Y* and *Z* components became bigger with the time (the aging variables). These conclusions are consistent with the deformation trends of an earth-rock dam. All of the regression residuals are small and randomly equal to ZERO, which are characterized by random errors.

The predictive ability of the models was proved through the random sampling from the raw data points (Wang, 1999). First, 45 sample points were used to construct the models and 6 sample points left for the evaluation of the model predictive ability. The goodness of fit for the models is: $\hat{\sigma}_x = 0.998$, $\hat{\sigma}_y = 0.737$, and $\hat{\sigma}_z = 0.686$ while the RMS errors for the 6 validation points were $\hat{\sigma}_x = 0.621$, $\hat{\sigma}_y = 0.652$, $\hat{\sigma}_z = 0.852$.

The stepwise regression was also performed for the same observation data and its results was compared with the results of PLSR (Table 4).

		Stepwise	PLSR
X	Modelling	0.8115	0.4793
	Predictive	1.6068	0.7668
Y	Modelling	0.5871	0.5309
	Predictive	1.0871	0.4336
Z	Modelling	0.4778	0.4601
	Predictive	1.1561	0.6972

Table 4: Comparison between the stepwise regression and PLSR

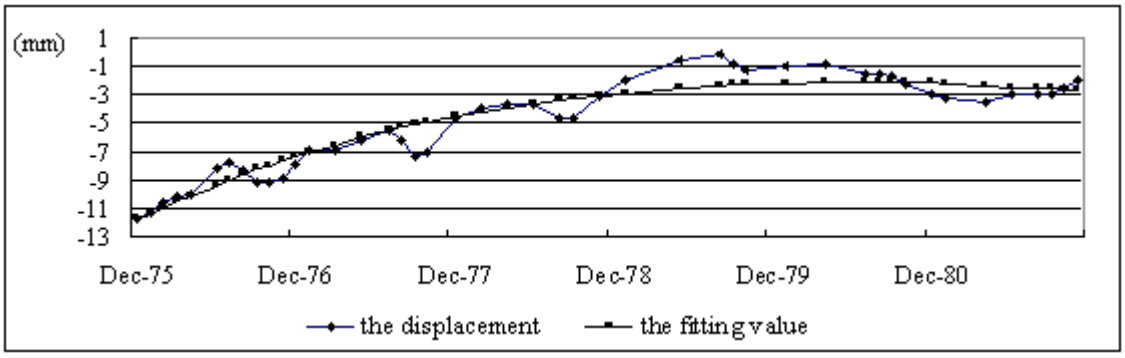

Figure 2: Displacement and the fitting curve of *X*-direction

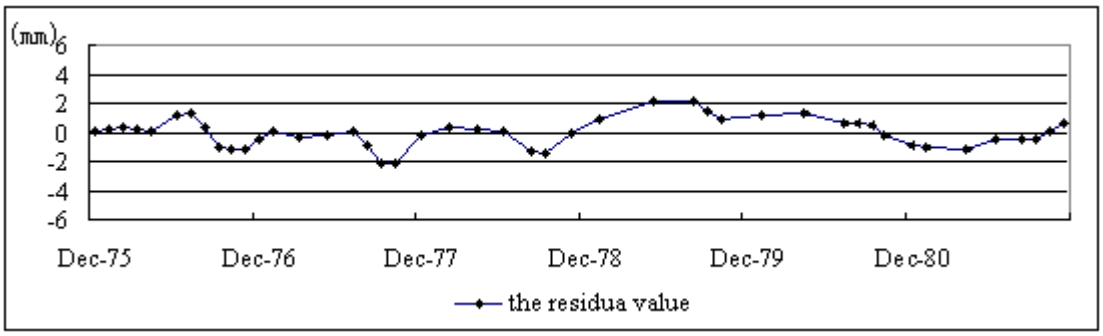

Figure 3: Residual curve of *X*-direction

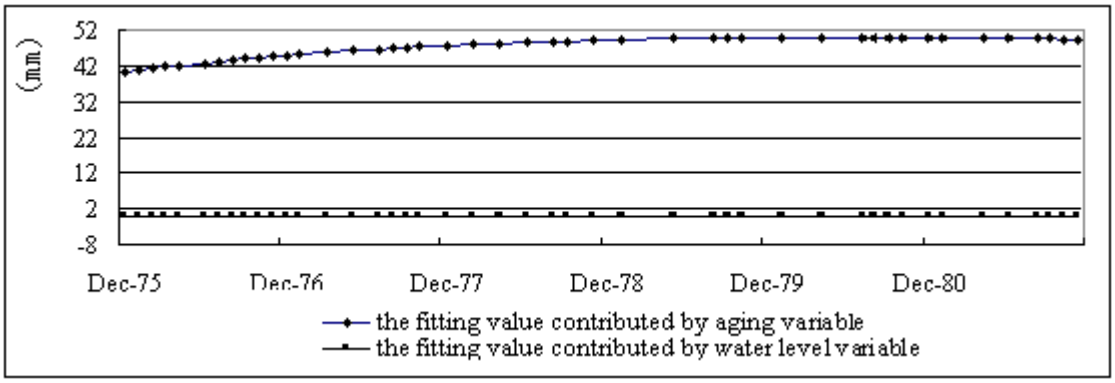

Figure 4: Decompsed fitting curve of *X*-direction

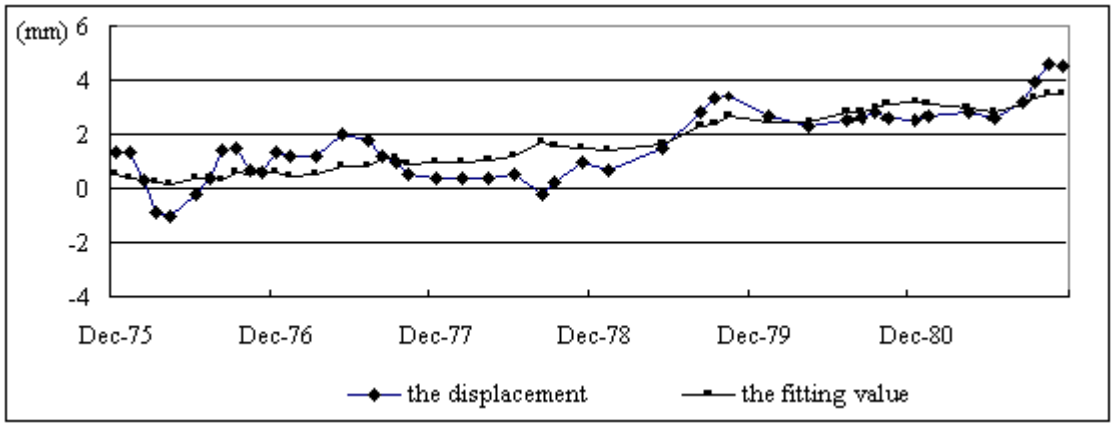

Figure 5: Displacement anf fitting curve of *Y*-direction

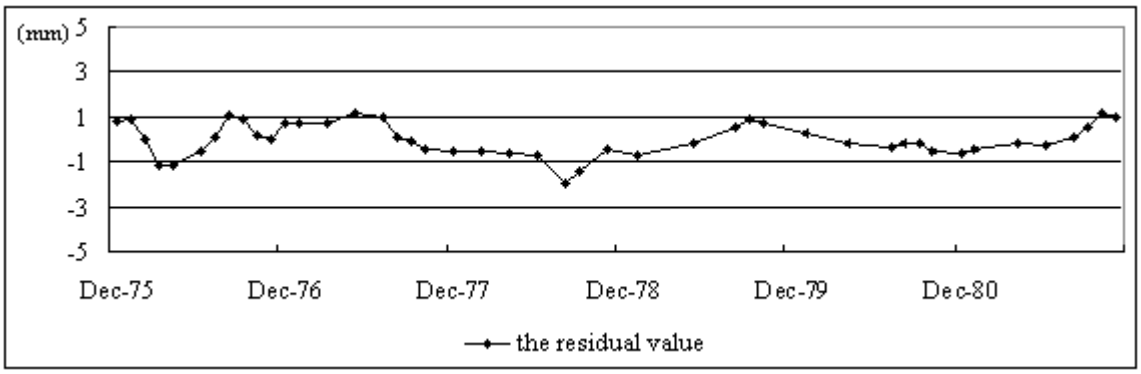

Figure 6: Residual curve of *Y*-direction

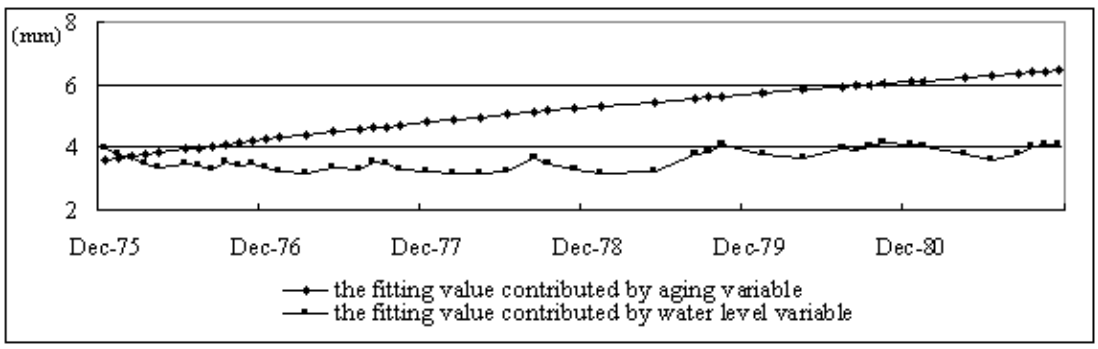

Figure 7: Decomposed fitting curve of *Y*-direction

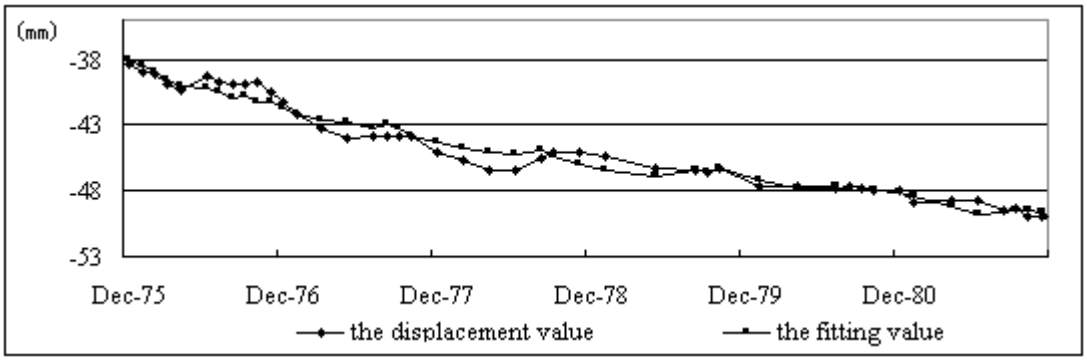

Figure 8: Displacement and the fitting curve of *Z*-direction

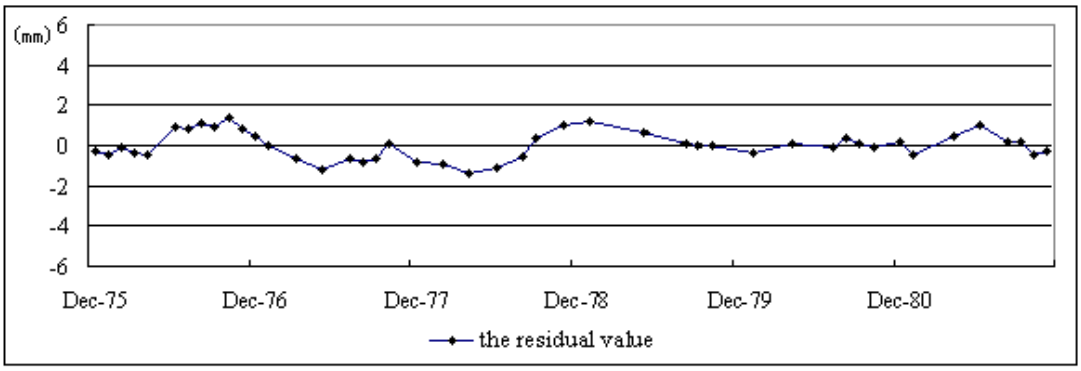

Figure 9: Residual curve of *Z*-direction

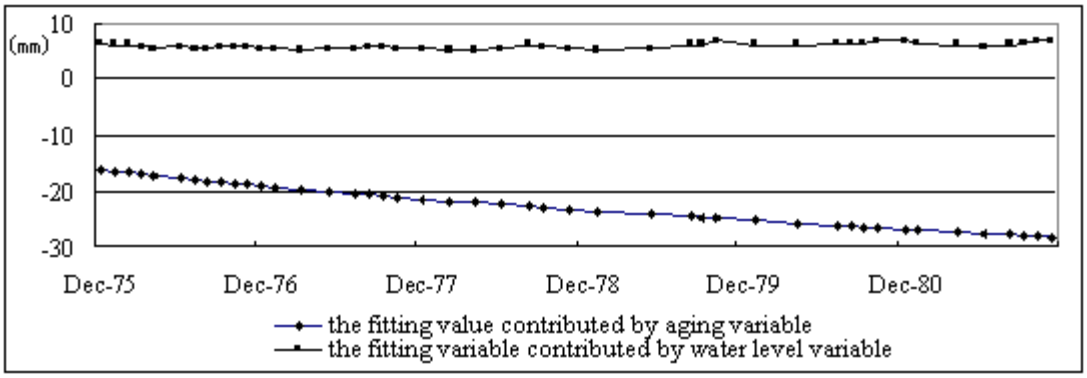

Figure 10: Decomposed fitting curve of *Z*-direction

4. CONCLUSIONS

The partial least squares regression has the better ability to deal with the collinearity among the predictor variables. It can take account into all of the observed predictor variables to construct the models and maximize the correlation between the response variables and the observed predictor variables. The PLSR can also be applied to the multivariate response variables so the spatial deformation analysis becomes possible. The given example shows that the PLSR shows good potential for the processing of dam monitoring data in practice.

References

Wu, Zhong-Ru (1990): *Theory of Safety Monitoring of Hydraulic Constructions andits Applications*, Publishing House of He-Hai University, August 1990. (Chinese)

Li, Zhen-zhao (1989): *Observation Data Analysis of Concrete Dams*，Publishing Houseof Hydraulic and Electric Industry，July 1989. (Chinese)

Ren, Ruo-Si (1997): *Multivariate Data Statistic Analysis- Theory,* Methods and Applications, Publishing House of Defense Industry，June 1997. (Chinese)

Wang, Hui-Wen (1999): *Partial Least Squares Method and Its Applications*, Publishing House of Defense Industry，April 1999. (Chinese)

Rosipal, R. and N. Krämer, (2006): *Overview and Recent Advances in Partial Least Squares*, in Saunders et al (Eds.): SLSFS (Subspace, Latent Structure and Feature Selection Workshop) 2005, LNCS (Lecture Notes in Computer Science) 3940, pp. 34-51, 2006

Corresponding author contacts

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