

# Monitoring of the World's Environmental Risk Based on MPCA

Shu Pang, Yiyun Han, and Xiaodong Yuan

**Abstract**—MPCA (multilinear principal component analysis) and hypothesis testing are used to establish the environmental monitoring system, including both the regional and the variable factors. High-dimensional data make it advantage than the previous models in considering comprehensive high-dimensional variables. So simulation of the system measures the change and change range of the environment, providing a more precise reference about the status of the environment.

**Index Terms**—MPCA, monitoring, environmental risk, hypothesis.

## I. INTRODUCTION

Global warming, air pollution, accelerating extinction of species, eutrophication and so many problems occur in almost everywhere in the world. And it is approaching a state shift in Earth's biosphere [1]. The regional problems are affecting the world, showing the deteriorating and unrest status of the earth.

Now many countries are establishing the regional environmental monitoring system, many organizations are establishing models to evaluate different variables' degree of health and change.

In the context of the times, it is necessary to establish the environmental monitoring system. As the global variable data cannot represent the whole status, and some severe regional environmental problems may be rather dangerous, it is important to establish a system based on regional data, representing the comprehensive changes in environment, including population, climate, pollution and many other factors.

## II. INDEX SELECTION

There are many kinds of environmental variables to monitor, such as PM2.5 for air, BOD for water and so on. This paper does not pay attention to the small index, but to the macro variables. And in order to monitor the world's environment, China, America, England, South Africa and Australia are included, providing regional data. However, the method of monitoring is also applicable to the micro variables mentioned above.

Human nowadays is proved to remain a dominating role in earth global environment [1] and has threatened the ecosystem's function and service [3]. Given the dominating position of human, it is necessary to place human as the

center factor in environmental monitoring. Through the insight about how to summarize the human related factors [4], it is sorted that the effect to earth ecosystems to direct factors and indirect factor.

Table I shows the detailed classification of variables that are chosen to establish the monitoring of the environment of the world.

TABLE I: THE CLASSIFICATION OF ENVIRONMENTAL VARIABLES

Classification	Factor	Index
Direct	Population change	Population growth rate
	Change in economic activity	GDP growth
	Technology	Cost of energy
Indirect	Habitat change	Land use
	overexploitation	Overexploitation
	Biodiversity	Biodiversity
	Pollution	Water and air pollution emission
	Climate change	Average temperature change

## III. ENVIRONMENTAL MONITORING

The monitoring contains three-dimensional factors, which are variable, region and time. And the three-dimensional data are interrelated. The three-dimensional data structure the matrix  $X (I \times J \times K)$ , where  $I$  is the number of time,  $J$  is the number of variable,  $K$  is the number of region.

The traditional two-dimensional factor model can only be used to analyze the factors separately, which is no longer applicable. Taking into account the complex variables, it is a good way to choose MPCA as the method monitoring the environment with the comprehensive variables. MPCA is now widely used in process monitoring, which compares the sample of the process with some standard samples [5]-[6]. As there is no standard sample in environmental process, in this paper, MPCA is used to compare the on-line variables' data of environmental process with the data in the past.

MPCA slice the data matrix along the time direction, and then put the slice data into arrangement from left to right, thus constituting new data matrix  $X (I \times JK)$

According to the Principal Component Analysis, the matrix  $X (I \times JK)$  can be expressed as

$$X = T_R P_R^T + E = \hat{X} + E \quad (1)$$

Here  $TR$  is the score of the principal component,  $PR$  is the loading vector of the principal component, which is also the eigenvector of correlation coefficient matrix of  $X$ , and  $E$  is the residual matrix.

Based on MPCA, the monitoring system uses the statistics

SPE (squared prediction error) and Hotelling- $T^2$  of the three-dimensional data to determine whether there is a system problem occurs.

First, analyze the three-dimensional data integrally, reducing data into fewer dimensions by MPCA, which gets the multivariate principal components. Then project the new data vector from time series onto two orthogonal subspace, and calculate the SPE and Hotelling- $T^2$  between existing data and new data.

Principal component score of the new data is in (2).

$$t_{new,R} = X_{new,R} P_R \quad (2)$$

The residual matrix  $e$  is in (3).

$$e = X_{news} - t_{new,R} P_R \quad (3)$$

The Hotelling- $T^2$  of the  $k$ -th time is in (4). And the  $S$  is the covariance matrix of the principal component score.

$$T^2 = t_{new,R}^* S^{-1} t_{new,R}^T \quad (4)$$

In MPCA, it is no longer appropriate to use SPE of principal component analysis, here the SPE of the  $k$ -time data is in (5).

$$SPE_k = \sum_{c=(k-1)J+1}^{kJ} e(c)^2 \quad (5)$$

By building the statistic SPE and Hotelling- $T^2$  related to the  $k$ -th time new data, it can be measured the mean value and relationship between the new and old data.

According to the Hotelling- $T^2$  and SPE, it can be determined whether there is a big environmental change happened in variable or region at the new time, and whether the environment becomes unstable.

In order to do hypothesis testing, it is necessary to determine the control limits of the statistics. And the control limit of  $k$ -time SPE is in (7).

$$SPE_{\alpha}^k = \frac{v_k}{2m_k} \chi_{2m_k^2/v_k, \alpha}^2 \quad (7)$$

According to the Hotelling- $T^2$ , quantile  $T_{\alpha}^2$  is in (8).

$$T_{\alpha}^2 = \frac{R(I^2 - 1)}{I(I - R)} F_{R, I-R, \alpha} \quad (8)$$

When take  $\alpha=0.05$ , the control limit can be got. When the curve of SPE exceeds the control limit, it means that there is a big change that time, and the environment is unstable in 95% confidence interval.

#### IV. SIMULATIONS

Fig. 1 and Fig. 2 show the simulation of the monitoring system of the world environment based on five countries' data for example. The dotted line is the control limit of SPE or Hotelling- $T^2$  in 95% confidence interval. And the full line is the SPE or Hotelling- $T^2$  that calculated by data. Due to the lack of standard sample, the control limit of the two statistics is rather inaccurate at the beginning time, and become more

and more accurate with the accumulation of data.

From the Fig. 1, the SPE simulation provides a more precise reference about the complex environmental status of the five countries. It is reasonable to believe that in this simple model that based on just five countries' data, the environment is stable as the SPE does not exceed the control limits. Due to lack of data, it cannot simulate the recent two years' status and forecast the status in the future.

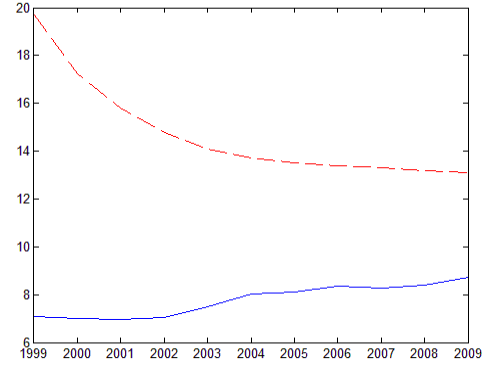


Fig. 1. SPE-time simulation result.

From the Fig. 2, the Hotelling- $T^2$  simulation plays not so good that the Hotelling- $T^2$  calculated by data is too small. And the reason is the matrix  $S$  in (4) is singular to working precision.

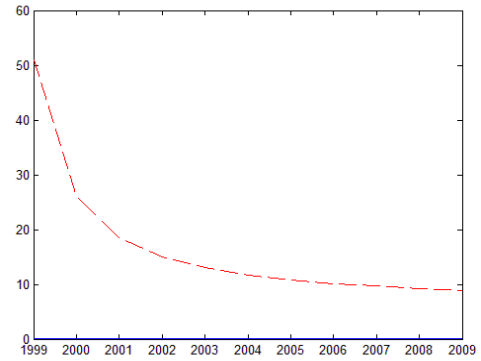


Fig. 2. Hotelling-T2-time simulation result.

#### V. MODEL CHECKING

In order to check the model, it is reasonable to create the erroneous data due to the lack of the data of environmental risk.

First, the environmental variable risk is to be checked. To achieve this, the pollution data from 2007 to 2009 of the USA is multiplied by 10.

Fig. 3 and Fig. 4 show the world's monitoring simulation result of the huge pollution risk of the USA.

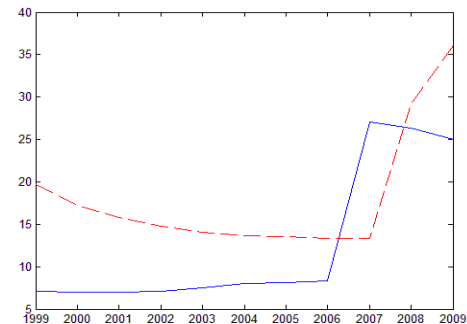


Fig. 3. SPE-time variable risk simulation result.

From the Fig. 3, the SPE line exceeds the control limit between the year 2006 and 2007. It responds immediately to the variable changing risk. And the control limit increases because it is related to the mean and variance of the SPE. So the SPE line goes underneath the limit line in the next two years. The monitoring system can only find the first environmental variable risk.

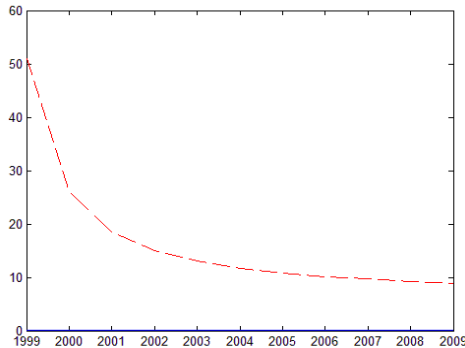


Fig. 4. Hotelling-  $T^2$ -time variable risk simulation result.

From the Fig. 4, the Hotelling- $T^2$  line does not respond to the risk due to the same reason that the matrix  $S$  in (4) is singular to working precision.

Second, the environmental region risk is to be checked. To achieve this, the data from 2006 to 2009 of the USA is multiplied by 10.

Fig. 5 and Fig. 6 show the world's monitoring simulation result of the USA environmental risk.



Fig. 5. SPE-time region risk simulation result.

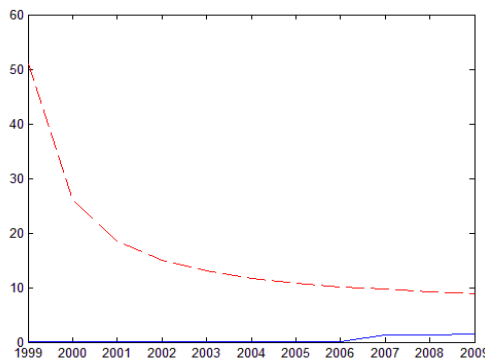


Fig. 6. Hotelling-  $T^2$ -time region risk simulation result.

From the Fig. 5, the SPE line exceeds the control limit between the year 2005 and 2006. It responds immediately to the regional environmental changing risk. But the control limit changes just like before. The monitoring system can only find the first environmental region risk.

From the Fig. 6, the Hotelling-  $T^2$  line increases since year

2006 as a respond to the risk. But it is still not good enough due to the same reason that the matrix  $S$  in (4) is singular to working precision.

## VI. CONCLUSIONS

1) MPCA and hypothesis testing are applicable to establish the environmental monitoring system, including both the regional and the variable factors. With the high-dimensional data, it can show the complex change of the world's environment, in the both way of region and variable. And the monitoring is real time, which bases on the data from ancient times and current time.

2) Simulation of the system measures the change and change range of the environment, providing a more precise reference about the status of the environment, which may be helpful to policy-making.

3) The monitoring system is not accurate enough due to the lack of data, variables and regions. With more and more data are included, the monitoring may be more accurate, and may be used to predict the future of world's environmental changes.

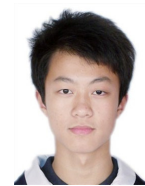
4) The curve of the simulation is in upward trend in shock, which shows that the world's environmental risk is in shock and increasing at the same time. And with more data, the environmental risk may be predicted in the system through time series model, such as ARMA [7].

5) It may be better to use the SPE statistic rather than Hotelling- $T^2$ , as the covariance matrix used to calculate may be singular to working precision.

6) The defect of the monitoring system is that the slow changes of the environment may not be detected and displayed in the curve. So the monitoring system is applicable for limited in monitoring the abruptly shift of the environment.

## REFERENCES

- [1] A. D. Barnosky *et al.*, "Approaching a state shift in Earth's biosphere," *Nature*, vol. 486, no. 7401, pp. 52, 2012.
- [2] P. M. Vitousek, H. A. Mooney, J. Lubchenco *et al.*, "Human domination of Earth's ecosystems," *Science*, vol. 277, no. 5325, pp. 494-499, 1997.
- [3] Millennium Ecosystem Assessment (Program), *Ecosystems and Human Well-Being: Our Human Planet: Summary for Decision Makers*. Ed. Millennium Ecosystem Assessment, Island Press, 2005, vol. 5.
- [4] R. Watson and A. H. Zakri, *UN Millennium Ecosystem Assessment Synthesis Report*, 2005.
- [5] J. H. Yen, "Development of PARAFAC models in off-line and on-line batch process monitoring," *Chemical Engineering*, 2001.
- [6] J. Zhang, "Research on batch process monitor based on multistage MPCA," M.S. thesis, Dept. Chemical Engineering, Beijing University of Chemical Technology, China, 2009.
- [7] S. Wang, X. F. Hu, Z. Zhao, and F. L. Wang, "Research on fault prediction method based on MPCA-AR," *Chinese Journal of Scientific Instrument*, vol. 30, no. 8, pp. 1779-1782, 2009.



**Shu Pang** was born in Hangzhou, China 1991. He is a junior undergraduate, and he is studying in the department of mathematics, Zhejiang University.

His Research interest is mainly in the model and variable selection of high-dimensional data and the statistical inference of the spatial data.

In 2011, he investigated Hangzhou public's satisfaction of food regulatory, and won the third prize in the survey design contest of Zhejiang province. In 2012, he participated in the student's science and technology innovation project of Zhejiang province, studying the price forecasts of stock index futures, using neural networks and statistics.