EnvironmentAsia
The international journal by the Thai Society of Higher Education Institutes on Environment

EnvironmentAsia 14(3) 2021 71-79
DOI 10.14456/ea.2021.24
ISSN 1906-1714; ONLINE ISSN: 2586-8861

Estimation of $PM_{2.5}$ Concentrations in Northern Thailand Using the Gappy Proper Orthogonal Decomposition Method

Kuntalee Chaisee[1], Suttida Wongkaew[2], and Ekkachai Thawinan[2*]

[1] *Data Science Research Center, Department of Statistics, Faculty of Science, Chiang Mai University, Chiang Mai 50200, Thailand*

[2] *Data Science Research Center, Department of Mathematics, Faculty of Science, Chiang Mai University, Chiang Mai 50200, Thailand*

* Corresponding author: ekkachai.thawinan@cmu.ac.th

Received: January 13, 2021; February 26, 2021; Accepted: May 22, 2021

Abstract

In the northern region of Thailand, $PM_{2.5}$ concentration has been the worst air pollution problem in the country for the past several years. This study applies the gappy proper orthogonal decomposition (gappy POD) method to estimate values missing from an incomplete dataset of $PM_{2.5}$ concentration in this area. Although monitoring and collection of $PM_{2.5}$ concentration data provide information about the air quality in the area, the collection process often misses some data, hence obscuring important information. After the POD method is used to extract dominant data information, the gappy POD, based on least squares optimization, is employed to approximate the missing data. The number of POD bases plays an important role in the approximation; thus, investigating the effects of a number of POD bases used in the gappy POD method is needed. The accuracy of the gappy POD method is validated by comparing estimation errors to errors from the baseline interpolation method using inverse distance weighting (IDW). The study shows that the results of the gappy POD method are more accurate than the IDW estimations.

Keywords: Missing data; Proper orthogonal decomposition; Gappy proper orthogonal decomposition; Air pollution; $PM_{2.5}$

1. Introduction

Air pollution, particularly airborne particulate matter (PM), has become a global environmental issue over the past several years. PM_{10} and $PM_{2.5}$ are common fine particles that can cause serious health problems when inhaled. Over the past few years, Thailand has suffered from high PM levels, especially in the northern region. The thick cloud of smog found in northern Thailand has typically been caused by forest fires and agricultural burning (Khamkaew *et al.,* 2016; Oanh *et al.,* 2011). Levels of $PM_{2.5}$ have become severe, causing numerous health problems such as difficulties in breathing, sore throats, and sinus problems. Moreover, it can contribute to respiratory and cardiovascular system health issues such as lung cancer and cardiopulmonary diseases (Brook *et al.,* 2010; Pope *et al.,* 2002). Currently, Chiang Mai University's Climate Change Data Center (CMU-CCDC) (https://www.cmuccdc.org) is running a project to measure and collect $PM_{2.5}$ concentrations. The PM data are valuable because they are essential to the evaluation of air quality. Equipment to monitor air pollution has been installed at many sites in the northern region. This system provides real-time air pollution measures as the monitors are close to the source of the problem. In addition, monitoring, collecting, and analyzing PM levels, especially when they reach harmful levels, provides critical information to those responsible for development of risk

assessments and public health policy. However, data values are often missing from PM records due to either equipment malfunctions or bad connections to the database server. These problems lead to an incomplete dataset, which must be repaired before the data can be used. Therefore, the fulfillment method used to preserve some information should be used to construct a completed dataset.

One effective methodology for rebuilding lost data entails using a set of low dimensional data; however, this is still based on dominant trends of the overall data. The gappy proper orthogonal decomposition (gappy POD) method is a technique recently developed to deal with damaged or missing data. It was developed from a POD approach used in several past research projects (Abbey *et al.,* 1991; Astrid *et al.,* 2008; Gunnink *et al.,* 1996; Jolliffe *et al.,* 2011; Kunisch *et al.,* 1999; Mees *et al.,*1987; Volkwein *et al.,* 2007; Wu *et al.,* 2003). It has been continuously refined and applied in different forms. For example, it is part of the principal component analysis (PCA) used in statistics (Cotta *et al.,* 2020; Hotelling *et al.,* 1933; Pearson *et al.,* 1901) and is included in the empirical orthogonal functions (EOFs) method used in meteorology and geophysics (Dommenget *et al.,* 2002; Gunnink *et al.,* 1996; Abbey *et al.,* 1991; Volkwein *et al.,* 2007; Jolliffe *et al.,* 2011). A key component of the gappy POD is the estimation of missing data based on non-missing data, and projections establish a modified set of selected basis vectors. There has been a wide range of gappy POD applications. For instance, it was used in an aerodynamic flow field to calibrate and illustrate how air flows past a wing (Bui *et al.,* 2004). In chemical engineering, it was applied to the reconstruction of flame kinetics in a spark-ignition engine. In climatology, it was used to reconstruct a temperature field in Tibet (Tsering *et al.,* 2019).

The primary goal of this work is to successfully apply the gappy POD method to reconstruct $PM_{2.5}$ concentration data missing from information provided by CMU-CCDC air pollution monitoring sites. To accomplish this, firstly a complete snapshot of the $PM_{2.5}$ data is constructed. Then this snapshot is used to construct the POD basis, capturing significant characteristics of the data. After that, the missing data is approximated from a linear combination of this set of bases. The coefficients of this representation are obtained using the least square method, and the gappy POD is applied to generate the missing data points. Finally, the missing data are fulfilled from the complete samples assisted by the suitable POD bases. To evaluate the efficiency of the method, the errors from the gappy POD method to the baseline and the inverse distance weighting (IDW) methods are compared. IDW is commonly used in interpolation as it estimates values of unknown points by using weighted distances and some nearby known points.

2. Methodology

2.1 Study location and data

The study area in northern Thailand includes eight provinces: Chiang Rai, Chiang Mai, Lampang, Lamphun, Mae Hong Son, Nan, Phayao, and Phrae. According to the data provided by the CCDC project, there are 174 air monitoring stations located across the region. PM levels in March are usually high, causing health problems. As a result, the data of March 2020 was used in this study. Unfortunately, during this time, 37 stations were inactive and therefore could not provide any data. A stable active station is defined as one with at least 90% uptime data available; otherwise, it is considered a nonstable active station. Based on this criterion, there were 57 stable active stations and 80 nonstable active stations, as shown in Figure 1.

A dataset with the time stamp is called a snapshot. A feature is data from one station that has been included in a snapshot. Therefore, there are 57 features for each snapshot, one from each of the stable active stations. Complete snapshots are needed for employment of the gappy POD method. Unfortunately, missing data occurred randomly at different times and stations. Figure 2 shows the frequency of missing features in each snapshot. The average number of missing features is approximately two, with a maximum of 14 features.

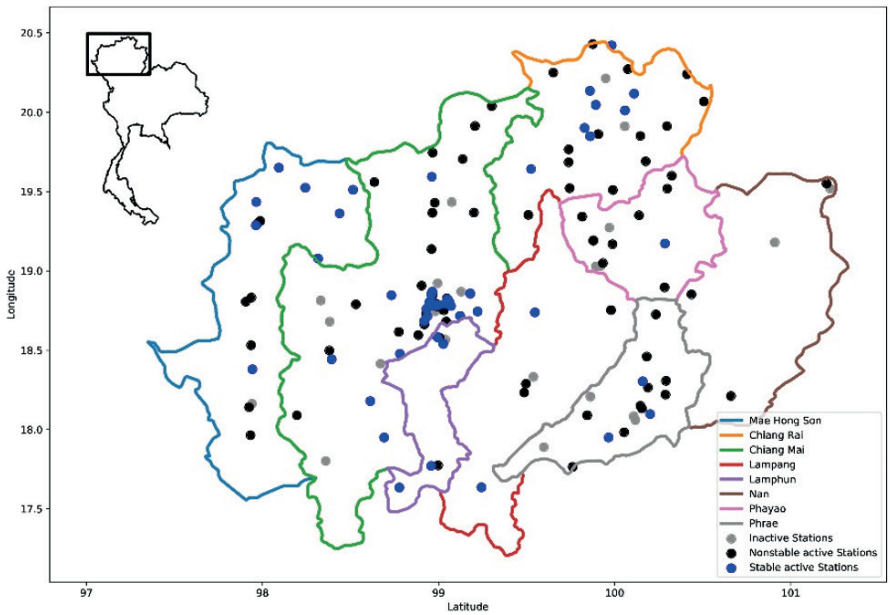


Figure 1. Locations of air pollution monitoring stations

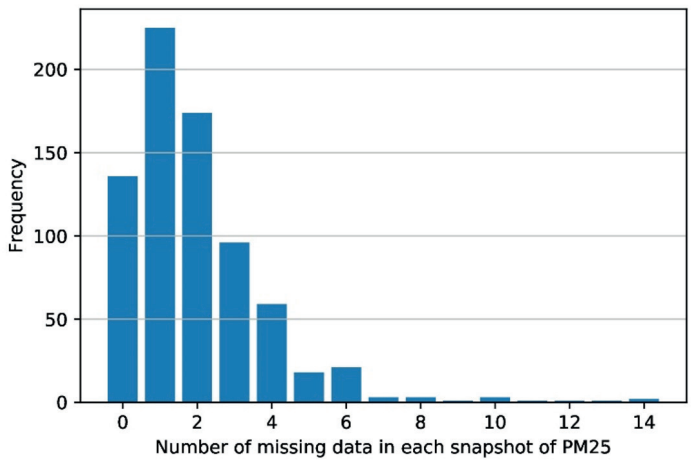


Figure 2. Histogram of $PM_{2.5}$ data from each snapshot

2.2 Numerical methods

2.2.1 Proper Orthogonal Decomposition (POD)

POD is a well-known numerical method broadly applied in finite elements for dimensional reduction. Recently, it has gained popularity in data analysis (Sukuntee *et al.,* 2018; Tsering-xiao *et al.,* 2019). In general, the goal of using the POD is to form a set of bases that can capture significant characteristics from all data. As a result, the POD enables estimation of missing data points by using selected POD bases and therefore reduces computation complexity.

In the following section, fundamental concepts and some notations related to POD construction are provided.

A matrix of complete snapshots is written by $S = [s_1, s_2, \ldots, s_n] \in R^{n \times n_s}$ where column vectors, $s_i = (s_{i1}, s_{i2}, \ldots, s_{in})^\top \in R^n$, i = 1,…, n , are complete data and s_{ij}, $j = 1,\ldots, n$ are data components. The data component s_{ij} is called feature.

The POD basis is obtained by using a Singular Value Decomposition (SVD) of the matrix S; that is, the matrix S can be written as $S = UDV$ where $U = [u_1,\ldots,u_r] \in R^{n\times r}$ and $V = [v_1,\ldots,v_r] \in R^{n_s\times r}$ are orthogonal matrices and $D = diag(\lambda_1,\ldots,\lambda_r) \in R^{r\times r}$, with singular values in decreasing order: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq 0$.

The POD basis is written in a matrix form of $P = [u_1,\ldots,u_{n_p}]$, obtained from the set of column vectors of U where n_p is the chosen number of basis vectors $n_p < r$. With this representation, each snapshot s_i can be approximated by:

$$s_i \approx \sum_{j=1}^{n_p} p_j\left(p_j^T s_i\right) = PP^T s_i.$$

Then, the minimum error of the approximation snapshots generated by using the POD basis is given by:

$$||s_i - PP^T s_i||_2^2 = \sum_{l=n_p+1}^{r} \lambda_l^2,$$

which is the summation of the square of neglected singular values $\lambda_{(n_p+1)},\ldots,\lambda_r$.

2.2.2 Gappy POD

Gappy POD is an extension of the POD method, which uses incomplete information from partially available data. The gappy POD is a projection-based method; however, it can be interpreted as being an interpolation method in some applications.

Suppose that s is a vector of an incomplete snapshot for known k features, where $k < r$. Define I_s^M as an ordered set of missing data and I_s^K as a set of known data of s, respectively. From this setting, the index set;

$$I = \{1,\ldots,r\} = I_s^M \cup I_s^K, \text{ and}$$
$$I_s^M \cap I_s^K = \emptyset.$$

With these index sets of s, it can be decomposed to $s^K = (s_i)_{i\in I_s^K}$, and $s^M = (s_i)_{i\in I_s^M}$ in the ordered indexes. Similarly, define $u_i^M = \left(u_{ij}\right)_{j\in I_s^M}$, and $u_i^K = \left(u_{ij}\right)_{j\in I_s^K}$ as the missing data and known data sub-vectors of each POD basis, respectively.

An approximation of the incomplete snapshot s can be constructed by finding a coefficient vector $a = (a_1,\ldots,a_k)^T$ from following the least square problem:

$$\min_{a\in R^{np}} \|s^K - P_s^K a\|_2,$$

which can be solved and implies that $a_i = \left(u_i^K\right)^T s^K$. Then, missing data can be approximated by

$$s^M = \sum_{i=1}^{n_p} a_i\, u_i^M.$$

The incomplete data of snapshot s then can be fulfilled by using approximated data s^M. Result accuracy can be investigated by randomly removing data from complete snapshots, as seen in the result section of missing data approximation.

2.2.3 Inverse Distance Weighting (IDW)

The IDW is one of the practical mathematical methods for interpolation. The concentration (z) of air pollution at a non-monitoring point (x_0) is calculated using a set of the neighboring monitored values $z_i = z(x_i)$ sampled at locations denoted by x_i. The relationship of interpolation is $z(x_0) = \sum_{i=1}^{n} w_i \cdot z(x_i)$, where $\sum_{i=1}^{n} w_i = 1$ and w_i represents the weights assigned to each point of the neighboring values, and the total of all weights equals one. With this approach, the IDW chooses w_i depending on the inverse of the distance between the unsampled point x_0 and all x_j considered window, such that

$$w_i = \frac{\frac{1}{d(x_0,x_i)}}{\sum_{j=1}^{n}\left(\frac{1}{d(x_0,x_1)}\right)},$$

where $(a,b) = \sqrt{a^2+b^2}$, the distance function, and $P \geq 1$, the parameter used to control the contributions area around sampled points. One can choose $P = 1$. When observed values that are closer to the points of interest are more heavily weighted, a bigger search window can be used to preserve some of the variations of local pollutant levels. Data from all monitoring stations presented in the search window are included in IDW interpolation. In our case, the window was chosen to be the

area of consideration. It is possible to choose smaller windows, but this can leave some areas without an estimated value because of inadequate sensor locations.

3. Results and discussion

This section presents the numerical results from the reconstruction of missing data. The analysis is based on 136 complete snapshots of $PM_{2.5}$ data from 57 stations that are divided into two data sets: a training set and a testing set. The training set consists of 100 samples and is used to construct the POD basis, while the testing set consists of 36 samples and is used for evaluation. Therefore, some data points are randomly removed to construct incomplete snapshots from these sets of data.

3.1 Data reconstruction using POD

Results corresponding to a SVD analysis are presented below. Figure 3 demonstrates that the sizes of singular values dropped drastically at the beginning and remained relatively stable following the 20th value. This suggests that the first 20 vectors of the POD basis are more significant than the remaining vectors.

Next, Figure 4(a) illustrates the relation between the convergence of reconstruction and the number of POD bases by showing the average errors of the reconstruction over the training set with a different number of bases.

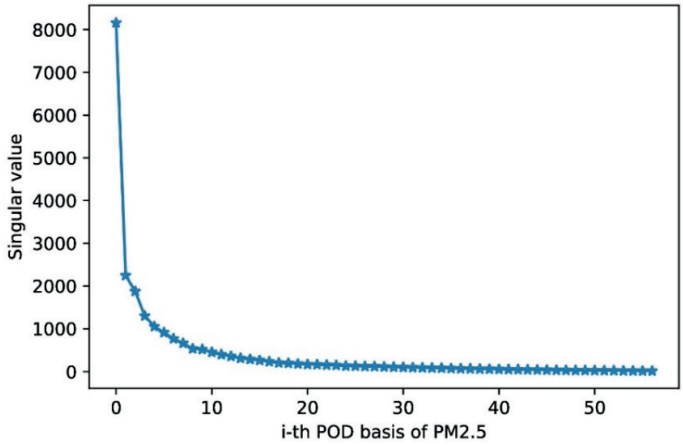
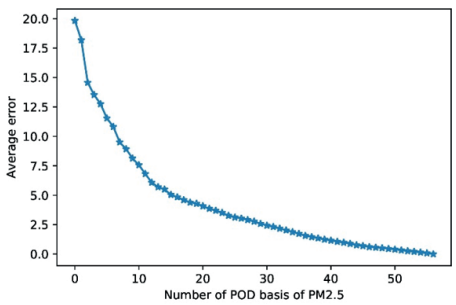
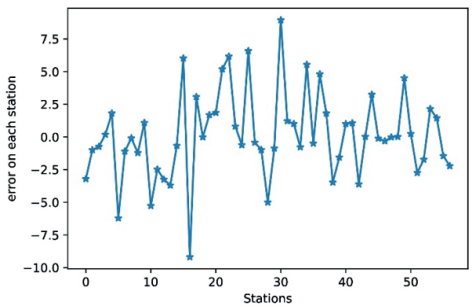


Figure 3. The singular value of the matrix of training snapshot sets



(a) average errors for reconstructing data



(b) errors with 20 POD basis vectors

Figure 4: The average errors with different numbers of bases and overall errors, features of one reconstructed snapshot with 20 POD bases.

Unsurprisingly, as the number of POD bases increases, the average error decreases. However, using many bases might not be necessary because there is only minor error improvement as the number of bases increases. In addition, this increase might lead to an over-fitting of the missing data approximation. The average errors of the reconstruction using 20 POD basis vectors at each station are shown in Figure 4(b).

3.2 Effects of the gappy POD on data reconstruction

Applying the gappy POD poses several questions about the number of significant bases and how many snapshots are needed to construct a POD basis. These questions are investigated by performing two tests for a gappy POD scheme. The first test determines the effect of a significant basis, and the second determines how the number of snapshots used for basis construction affects the average error of various missing data points.

Test 1: the effect of the number of significant bases

In Test 1, different missing values in the training and the testing data sets are investigated with different POD bases. First, some data are randomly removed from the training and test sets. The number of missing data points are 2, 5, 10, 15, and 20, respectively. These missing data are then reconstructed using a different POD basis number. Figure 5 compares the average of the absolute differences between results of reconstruction using different numbers of POD bases and the complete data set. On the left-hand side, the results of the training set are shown, whereas the results of the test set are shown on the right-hand side. Overall, the mean errors in the restored values have been significantly affected by the number of POD bases; that is, the error increases as the number of POD bases increases. Moreover, the optimal number of POD bases for approximation of the missing data is between 10 and 15. It is important to note that the POD basis has been constructed from data in the training set. The average errors with absolute norm tend to be smaller than the ones in the test set.

Initially, some significant bases and snapshots used in the basis construction are investigated as they play an important role in the estimation. Using too many bases does not always lead to the best approximation. According to this Test, 10 and 15 are the optimal numbers of bases to use in the gappy POD.

Test 2: the effect of the number of snapshots used for basis construction

The number of snapshots used for basis construction also affects approximation of the missing data. In this experiment, the number of snapshots used to construct a POD basis varies from 20 to 100. Consistent with the results of Test 1, the number of POD basis (n_p=10) is fixed, and the performance of the gappy POD to restore from 2 to 20 points of missing data is investigated. The results of this experiment are shown in Figure 6 and demonstrate that increases in the number of complete snapshots do not improve the gappy POD method's performance.

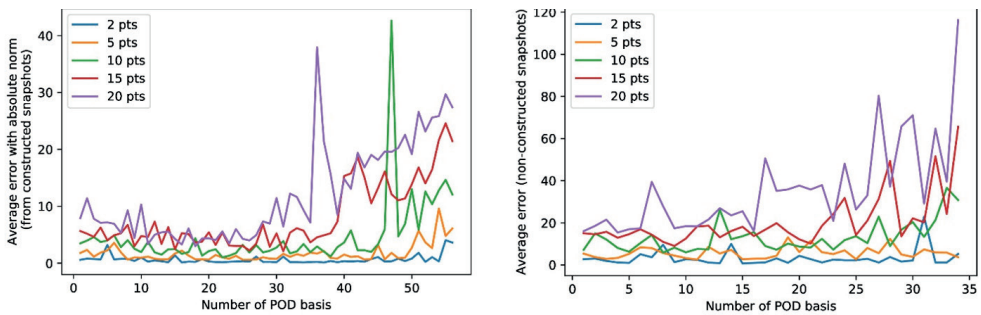


Figure 5. The average errors for the restored missing data of points 2, 5, 10, 15, and 20 for different POD basis numbers.

These results illustrate that the number of snapshots for POD basis construction has a minor effect on mean absolute errors. Nevertheless, the errors increase after using 70 snapshots. Consequently for the next Test, between 60 and 70 snapshots have been considered for constructing sets with a 10 and 15 POD basis.

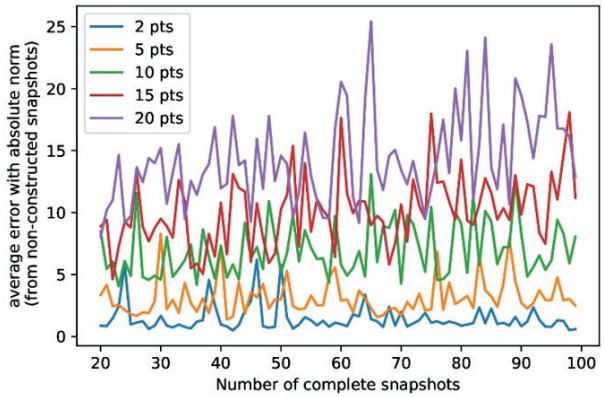


Figure 6. The average errors of the restored missing data using different numbers of snapshots to construct a POD basis.

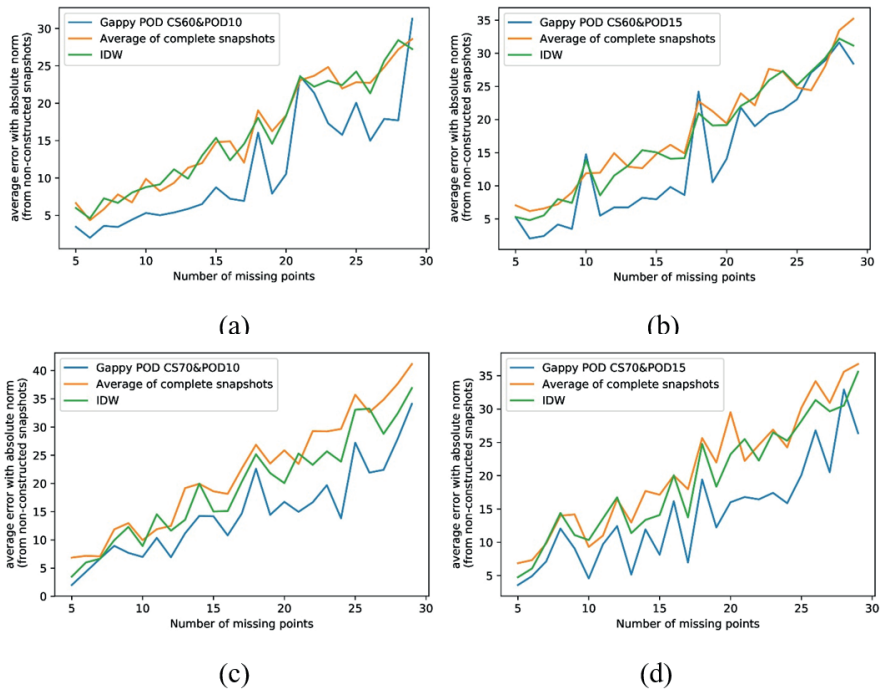


Figure 7. Absolute errors in the data reconstruction of eight missing values: comparing the gappy POD method with the IDW method. Note that the gappy POD method in panel (a) uses 60 complete snapshots and 10 POD basis. Panel (b) results use 60 complete snapshots and 15 POD basis. Panel (c) uses 70 complete snapshots and 10 POD basis, while panel(d) uses 70 complete snapshots and 15 POD basis.

Test 3: Comparing results of different combinations: POD basis vs. snapshots with other methods

This test is used to investigate the performance of the gappy POD method to approximate the missing data in the testing dataset. Four numerical experiments are performed using different numbers of snapshots with POD basis constructions (n_s) and using different numbers of POD basis for the gappy POD method (N_p). These test cases include:

- Case I : n_s=60 and N_p =10,
- Case II : n_s=60 and N_p =15,
- Case III : n_s=70 and N_p =10,
- Case IV : n_s=70 and N_p =15.

Test 3 evaluated the efficiency of the gappy POD in four cases and compared them to baseline methods, including the naive method in which missing values were calculated using an average of the available snapshots and the IDW.

The results of these four cases are shown in Figures 7(a) - 7(d). It can be seen from the numerical results that the mean error in the gappy POD method is much smaller than the mean error in the IDW method in all cases. It is clear that the gappy POD method outperforms the other two methods, especially when n_s = 60 snapshots and N_p = 10 POD basis are employed, see in Figure 7(d).

Numerical tests demonstrate that the gappy POD approach is an efficient tool for dealing with missing data. Figure 7 shows the mean errors of four cases compared with the average of complete snapshots and the approximation provided by IDW. Test III reveals that the gappy POD, including an adequate number of bases, outperforms both the IDW method and the average data. Notably, if ten bases are used, both cases give better results; 60 and 70 snapshots for this dataset.

4. Conclusion

The gappy POD has been employed to approximate the missing $PM_{2.5}$ concentration data. The data was obtained from monitoring stations located in the northern region of Thailand. Numerical results indicate that the gappy POD is an efficient method for reconstruction, able to approximate missing data. An optimal number of bases is essential to improve accuracy. To be exact, our study shows that only about 60 complete snapshots are needed for optimal basis construction, and only about 10 POD bases are required to reconstruct this dataset.

Acknowledgements

This research was supported by Chiang Mai University. The authors are graceful to the Climate Change Data Center (CCDC) project for providing valuable data.

References

Abbey DE, Mills PK, Petersen FF, Beeson WL. Long-term ambient concentrations of total suspended particulates and oxidants as related to incidence of chronic disease in California Seventh-Day Adventists. Environmental Health Perspectives 1991; 94: 43-50.

Astrid P, Weiland S, Willcox K, Backx T. Missing point estimation in models described by proper orthogonal decomposition. IEEE Transactions on Automatic Control 2008; 53(10): 2237-51.

Brook RD, Rajagopalan S, Pope III CA, Brook JR, Bhatnagar A, Diez-Roux AV, Holguin F, Hong Y, Luepker RV, Mittleman MA, Peters A. Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the American Heart Association. Circulation 2010; 121(21): 2331-78.

Bui-Thanh T, Damodaran M, Willcox K. Aerodynamic data reconstruction and inverse design using proper orthogonal decomposition. AIAA journal 2004; 42(8): 1505-16.

Cotta HH, Reisen VA, Bondon P, Prezotti Filho PR. Identification of redundant air quality monitoring stations using robust principal component analysis. Environmental Modeling and Assessment 2020; 25(4): 521-30.

Dommenget D, Latif M. A cautionary note on the interpretation of EOFs. Journal of climate 2002;15(2): 216-25.

Gunnink JL, Burrough PA. Interactive spatial analysis of soil attribute patterns using exploratory data analysis (EDA) and GIS. Spatial Analytical Perspectives on GIS, 1st ed.; Ficher, M., Scholten, H., Unwin, D., Eds. 2019; 13: 87-100.

Jolliffe, I. Principal component analysis. In: International encyclopedia of statistical science, Springer. 2011; 1094–1096.

Hotelling H. Analysis of a complex of statistical variables into principal components. Journal of educational psychology 1933; 24(6): 417.

Khamkaew C, Chantara S, Janta R, Pani SK, Prapamontol T, Kawichai S, Wiriya W, Lin NH. Investigation of biomass burning chemical components over Northern Southeast Asia during 7-SEAS/BASELInE 2014 campaign. Aerosol and Air Quality Research 2016; 16(11): 2655-70.

Kunisch K, Volkwein S. Control of the Burgers equation by a reduced-order approach using proper orthogonal decomposition. Journal of optimization theory and applications 1999; 102(2): 345-71.

Mees AI, Rapp PE, Jennings LS. Singular-value decomposition and embedding dimension. Physical Review A. 1987; 36(1): 340.

Oanh NT, Ly BT, Tipayarom D, Manandhar BR, Prapat P, Simpson CD, Liu LJ. Characterization of particulate matter emission from open burning of rice straw. Atmospheric Environment 2011; 45(2): 493-502.

Pearson K. LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 1901; 2(11): 559-72.

Pope Iii CA, Burnett RT, Thun MJ, Calle EE, Krewski D, Ito K, Thurston GD. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. Jama 2002; 287(9): 1132-41.

Sukuntee N, Chaturantabut S. An Application of Proper Orthogonal Decomposition for Estimating Missing Data of Patients in Different Cause Groups. Thai Journal of Mathematics 2018; 21-33.

Tsering-xiao B, Xu Q. Gappy POD-based reconstruction of the temperature field in Tibet. Theoretical and Applied Climatology 2019; 138(1-2): 1179-88.

Volkwein S. Proper orthogonal decomposition: applications in optimization and control. CEA-EDFINRIA Numerical Analysis Summer School. 2007.

Wu GG, Liang YC, Lin WZ, Lee HP, Lim SP. A note on equivalence of proper orthogonal decomposition methods. Journal of Sound and Vibration 2003; 265(5): 1103-1110.