Data Inspecting and Denoising Method for Data-Driven Stochastic Subspace Identification

Xiaohang Zhou, Lu Cao, Inamullah Khan, and Qiao Li

1Bridge Engineering Department, Southwest Jiaotong University, Chengdu 610031, China
2School of Civil Engineering, National University of Sciences and Technology, Risalpur, KPK 24080, Pakistan

Correspondence should be addressed to Qiao Li; ql3721@163.com

Received 9 October 2018; Accepted 19 November 2018; Published 23 December 2018

Abstract

Data-driven stochastic subspace identification (DATA-SSI) is frequently applied to bridge modal parameter identification because of its high stability and accuracy. However, the existence of abnormal data and noise components may make the identification result of DATA-SSI unreliable. In order to achieve a reliable identification result of the bridge modal parameter, a data inspecting and denoising method based on exploratory data analysis (EDA) and morphological filter (MF) was proposed for DATA-SSI. First, EDA was adopted to inspect the data quality for removing the data measured from malfunctioning sensors. Then, MF along with an automated structural element (SE) size determination technique was adopted to suppress the noise components. At last, DATA-SSI and stabilization diagram were applied to identify and exhibit the bridge modal parameter. A model bridge and a real bridge were used to verify the effectiveness of the proposed method. The comparison of the identification results of the original data and improved data was made. The results show that the identification results obtained with the improved data are more accurate, stable, and reliable.

1. Introduction

Bridge is a critical structure of the whole transportation network, and it is vital for engineers to be aware of its operational state [1]. Modal parameter identification is the first and key step of bridge operational state analysis [2, 3]. Data-driven stochastic subspace identification (DATA-SSI) is one of the favourite techniques for modal parameter identification [4]. DATA-SSI possesses high stability and accuracy because of its ability of considering multiple outputs [5].

DATA-SSI was used by Zhang for extracting modal parameters of an arch bridge model, and the results showed that DATA-SSI was characterized with high precision [6]. DATA-SSI was applied by Altunisik on a scaled girder bridge for extracting modal parameters, and the results showed that the method had a good ability for identifying frequencies and mode shapes [7]. DATA-SSI was adopted by Boonyapinyo to identify the frequencies and damping ratios of a bridge girder model under the excitation of wind, and DATA-SSI was proved to be more reliable than covariance-driven stochastic subspace identification (CO-SSI) [8]. An automated DATA-SSI was developed by Ubertini for filtering the false modes and applied the method on two real bridges, and it was proved that DATA-SSI was better than frequency domain decomposition methods [9]. DATA-SSI was adopted by Brincker on the Great Belt Bridge for modal parameter identification, and the results showed that DATA-SSI was appropriate for identifying closely spaced modes [10].

However, there are many interferences in data acquisition and transmission, such as malfunctioning of sensors, defects of transmission system, and failure of shielding measures and noise components, which would lead to unreliable bridge monitoring data [11]. Sometimes, the measured data are completely distorted or the valuable structural responses are fully submerged in noise, and even DATA-SSI cannot get an acceptable identification result [12]. Moreover, the data collected from the continuous health monitoring system are in huge amount, and the inspecting and denoising of large data sets require a significant amount of time and effort. In order to get a reliable bridge modal parameter identification result, efficient data inspecting and denoising techniques are needed. DATA-SSI is a time-domain method; hence, the data inspecting and denoising techniques are also preferred to be in time domain.
For bridge monitoring data inspecting, exploratory data analysis (EDA) is a potential solution. EDA is a time-domain data visualization tool for exhibiting the data statistical properties; it is adaptive and efficient and needs no prior information [13]. A considerable amount of research on theories and applications of EDA was made [14, 15]. EDA based on boxplot and robust-class selection was applied for geochemical mapping in the research of Bounessah, and the boxplot was proved to be very useful in capturing the distribution, skewness, and outliers of the data [16]. A new EDA technique based on interactive evolutionary computation was proposed by Malinchik and Bonabeau, and a rapid data dimension reduction was realized by the proposed technique [15]. A large number of basic examples of applying EDA were introduced by Mast and Trip to demonstrate the general process of EDA [17]. EDA was used by Vezzoli to evaluate the performance of a large amount of sensors and capture the causes of variation in the data, and EDA was proved to be suitable for dealing with massive amount of data [18]. Three correlation analysis techniques of EDA were adopted by Xiao et al. to inspect a pump’s data and evaluate its working state, and the conclusion was drawn that EDA was a basic but useful data analysing tool [19]. The applications of EDA mentioned above mainly focus on the field of industry and machinery, but its application on bridge monitoring is relatively less.

EDA is an advanced data inspecting tool, but it makes no changes to the data itself. In order to improve the data quality, a powerful data denoising technique is still needed. The commonly used data denoising methods are linear digital filter (LDF) and decomposition-reconstruction method (DRM), but there are many limitations in applying LDF and DRM. The data processed by LDF always have a sudden truncation in frequency domain along with a phase delaying problem; in other words, the data are distorted. For DRM, it is hard to determine a general standard for selecting the components of valuable structural response, and its low computational efficiency is another serious drawback. Recently, morphological filter (MF), which is a kind of time-domain filter, is being widely applied because of its high efficiency and capacity of considering nonlinearity [20, 21]. MF was used by Zou and Liu to get a low distorted image for the target recognizing system, and it was proved that MF was superior to the traditional LDF [22]. The noise source of a low X-ray imaging system was investigated, MF was adopted to eliminate the noise components by Dan et al., and a clean image with useful details was achieved [23]. MF combined with fuzzy principle component analysis was proposed in the work of Baghsah and Kasaei, and the method was proved to be an efficient denoising tool [24]. Yuan and Li adopted MF to detect and remove the noise components in data, and the results showed that MF was effective at various noise levels [25]. A composite MF combined with genetic programming training algorithm was developed by Yang and Li, the method was adopted on simulated and real MRI data to eliminate the noise components, and the results showed that the method was sensitive to noise especially when the noise level was high [26]. According to the aforementioned research studies, MF is an efficient tool for filtering the noise components in data and it is a promising solution for bridge monitoring data denoising. However, there are still problems to be solved in order to make MF an adaptive method, such as the automated size determination of structural element (SE).

In this paper, a data inspecting and denoising method based on EDA and MF for DATA-SSI was proposed. First, EDA was adopted to inspect the data quality in order to find out the abnormal data sets and locate the malfunctioning sensors. Then, MF along with an adaptive SE size determination method was applied to suppress the noise components. Finally, DATA-SSI was adopted to identify the bridge modal parameter. In order to verify the effectiveness of the proposed method, the identification results of the original data and improved data were compared. The overall research framework of this paper is shown in Figure 1.

2. Main Theory

2.1. EDA. It is difficult to inspect the quality of huge amount of data acquired by the continuous bridge health monitoring system, but EDA with efficient data mining ability may be a solution to this problem [27]. The hypothesis of traditional statistical analysis along with prior knowledge is abandoned in EDA, and only the value of the data itself is focused. Visualization tools are used to exhibit the data characteristics and assess the data quality intuitively. Therefore, EDA will aid engineers to explore the details of the bridge monitoring data with higher precision and less computational time. Numerous techniques are available for EDA, such as boxplot, QQ plot, control chart, Andrew’s curves, and histogram [28]. Due to the limited space, only the boxplot will be introduced and adopted to demonstrate the effectiveness of EDA. The boxplot is a basic but effective tool to visualize the distribution and statistical properties of the data and to provide multivariate displays with univariate information [29]. In the boxplot, the data information such as location, spread, skewness, and potential outliers are obviously revealed. More importantly, the characteristics of different data sets can be compared by placing the boxplots side by side.

Five important statistics of a data set are needed to construct a boxplot. They are median ($M$), upper quartile (UQ), lower quartile (LQ), upper limit (UL), and lower limit (LL). In the boxplot, the data are sorted in the descending order; therefore, the sample quartiles $M$, UQ, and LQ can be found easily and separately. UL and LL can be defined as the following equation:

$$UL = UQ + 1.5 \times IQR,$$

$$LL = LQ - 1.5 \times IQR,$$

where the interquartile range $IQR$ is defined as

$$IQR = UQ - LQ.$$

Data points that are larger than UL or smaller than LL are taken as potential outliers. The most common form of the boxplot is shown in Figure 2.

The distribution and skewness of the data set can be estimated by observing the relative location of LQ, UQ, and $M$, and the potential outliers of the data are directly figured
2. Potential outliers. Potential outliers are not considered during the calculation process of the five statistics; hence, the boxplot has a very good resistance to the impact of abnormal data points.

2.2. MF. MF is a nonlinear time-domain filter based on the theory of mathematical morphology [20]. In the theory of mathematical morphology, dilation and erosion are the two basic operations associated with SE; and there are only two parameters shape and size of SE that should be assigned in advance.

Dilation operation on data \( \alpha \) by SE \( \beta \) and erosion operation on data \( \alpha \) by SE \( \beta \) are expressed as \( \alpha \oplus \beta \) and \( \alpha \ominus \beta \), respectively, as shown in the following equation:

\[
\begin{align*}
\alpha \oplus \beta &= \max\{\alpha(i) + \beta(j)\}, \\
\alpha \ominus \beta &= \min\{\alpha(i) - \beta(j)\},
\end{align*}
\]

where \( 0 \leq i \leq (n - 1) \), \( 0 \leq j \leq (k - 1) \), \( \alpha(i) \) is the \( i \)th point of \( \alpha \) with a total number of \( k \), and \( \beta(j) \) is the \( j \)th point of \( \beta \) with a total number of \( n \).

Opening and closing are the two advanced operations that derived from the dilation and erosion and denoted by symbol \( \ominus \) and symbol \( \odot \), respectively. The opening operation is defined as equation (4) while the closing operation is defined as equation (5):

\[
\begin{align*}
\alpha \ominus \beta &= \alpha(\alpha \ominus \beta) \ominus \beta, \\
\alpha \odot \beta &= \alpha(\alpha \odot \beta) \odot \beta.
\end{align*}
\]

However, the opening operation can only process the data points that are smaller than the local mean value while the closing operation is just the opposite; hence, the opening and the closing operations should be combined as equation (6):

\[
a' = \frac{(\alpha \odot \beta + \alpha \ominus \beta)}{2}.
\]

By using Equation (6), the impulse and white noise components can be filtered out from the original data. The flow chart of the operation of MF is presented in Figure 3.

In the application process, the type and size of SE are the two parameters that should be assigned in advance. The commonly used types of SE are triangle and circle, and the triangle SE is more sensitive to white noise [4]. For bridge monitoring data, the main component of noise is white noise; hence, the triangle SE is selected in this paper. The size of SE is a more important parameter. The noise components will not be completely removed when the size of SE is too small; otherwise, the valuable structural response components will be impaired when the size of SE is too big. However, there is no reliable formula for calculating the appropriate size of SE.

In this paper, a simple and practical SE size determination method based on spectrum analysis was proposed. First, the spectrum analysis is adopted on the original data, and the frequency with the highest amplitude is selected. Then, MF is applied on the original data with the size of SE increasing as \( 2n_i + 1 \), and spectrum analysis is adopted during this process to track the amplitude changes of the selected frequency. The filtering process should not impair the valuable components of the data; hence, the increasing of SE size should be terminated before the amplitude of the selected frequency is reduced. However, white noise contains all the frequency bands, and the adopting of MF would inevitably affect all the spectrums along the frequency domain. In this study, the limitation of the amplitude reduction of the selected frequency was set to 10%. In other words, the size of SE is determined when the reduction ratio of the amplitude of the selected frequency reaches 10%. The key steps of the proposed SE size determination method are depicted in Figure 4.

2.3. DATA-SSI. An oscillatory system without deterministic input can be described by DATA-SSI with using the state space [30] as

\[
\begin{align*}
x_{k+1} &= Ax_k + \omega_k, \\
y_k &= Cx_k + \nu_k,
\end{align*}
\]

where \( x_k \) and \( y_k \) are the state vector and the outputs, respectively, of a system at the time instant \( k \), \( A \) and \( C \) are the system matrices, and \( \omega_k \) and \( \nu_k \) are the white noise disturbances.

In this paper, we followed the methods by Khan et al. 2015 to compute the system matrices \( A \) and \( C \) [31]. The Hankel matrix of DATA-SSI can be determined by computing the projection matrix of the output data, and it can be expressed as the following equation:

\[
\text{Hankel} = \begin{bmatrix}
Y_0 & Y_1 & Y_2 & \cdots \\
Y_1 & Y_2 & Y_3 & \cdots \\
Y_2 & Y_3 & Y_4 & \cdots \\
\vdots & \vdots & \vdots & \ddots
\end{bmatrix},
\]
The numbers of block rows and columns in the Hankel matrix are the two important parameters that would directly affect the identification results of DATA-SSI. Moreover, the number of block columns must be larger than that of block rows.

Then, RQ decomposition is adopted on the Hankel matrix, and the two projection matrices \( P_i \) and \( P_{i+1} \) can be obtained via the following equation:

\[
\begin{bmatrix}
R_{11} & 0 & 0 & 0 & Q_1 \\
R_{21} & R_{22} & 0 & 0 & Q_2 \\
R_{31} & R_{32} & R_{33} & 0 & Q_3 \\
R_{41} & R_{42} & R_{43} & R_{44} & Q_4
\end{bmatrix}
\]

\[
P_i = \begin{bmatrix}
R_{31} \\
R_{41}
\end{bmatrix}
\begin{bmatrix}
Q_1 \\
Q_2
\end{bmatrix},
\]

\[
P_{i+1} = \begin{bmatrix}
R_{41} & R_{42} & R_{43}
\end{bmatrix}
\begin{bmatrix}
Q_1 \\
Q_2 \\
Q_3
\end{bmatrix}.
\]

Singular value decomposition is performed on the projection matrix \( P_i \) to get the observability matrix \( O_i \) and the Kalman filter state space sequence \( \hat{S}_i \):

\[
P_i = USV^T = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix}
\sum_{i=1}^{T} 0 \\
0 \\
\sum_{i=0}^{T} = 0
\end{bmatrix} \begin{bmatrix} V_1^T \\
V_2^T
\end{bmatrix},
\]

\[
O_i = U_i S_1^{1/2} T, \]

\[
\hat{S}_i = [O_i]^{*}[P_i].
\]

The similarity transformation \([T]\) can be set equal to the identity matrix and a factorization can be applied to \([P_{i+1}]\), thus obtaining the following equation:

\[
[P_{i+1}] = [O_i]^{*}[\hat{S}_{i+1}],
\]

\[
[\hat{S}_i] = [O_i]^{*}[P_{i+1}].
\]

Hence, the system matrices \( A \) and \( C \) can be obtained by using the Kalman filtered state matrix \( x_i \) and the last block row of the output data matrix \( y_{i+1} \), as shown in the following equation:

\[
A = x_{i+1} x_{i}^{-1},
\]

\[
C = y_{i+1} x_{i}^{-1}.
\]

The eigenvector \( \Psi \) and the eigenvalue \( \Lambda \) can be obtained by the following equation:

\[
A = \Psi \Lambda \Psi^{-1}.
\]

The eigenvalue \( \Lambda \) can be converted from discrete time domain to continuous time domain as the following equation:

\[
\lambda = \frac{\ln(\Lambda)}{\Delta t},
\]

At last, the frequency \( f \), the damping ratio \( \xi \), and the mode shape \( \phi \) can be derived from the following equation:

\[
f = \frac{\sqrt{\text{Re}(\lambda)^2 + \text{Im}(\lambda)^2}}{2\pi},
\]

\[
\xi = \frac{\text{Re}(\lambda)}{\sqrt{\text{Re}(\lambda)^2 + \text{Im}(\lambda)^2}},
\]

\[
\phi = C\Psi.
\]

For bridge structures, there is generally no prior information about the system order that can be known in advance, and an improper system order in the algorithm of DATA-SSI would lead to false modes in the identification results. In order to eliminate the effects of undetermined system order, the stabilization diagram [32] was associated with DATA-SSI in this paper. By increasing the system order gradually, the modal parameters with real physical meaning will continuously emerge in the stabilization diagram; and the stable poles representing the real modal parameters are obtained.

### 3. Applications

In order to verify the data inspecting and denoising method proposed for DATA-SSI, a large-scale cable-stayed model bridge and a real long-span cable-stayed bridge were taken as instances. The processes of modal parameter identification demonstrated in Figure 1 were adopted.

#### 3.1. Model Bridge

A large-scale cable-stayed model bridge was taken as the first instance to verify the proposed method. The overall span arrangement of the model bridge is 6.5 + 19 + 6.5 = 32 m. The height of the pylons is 4.55 m while the height of the piers is 1.9 m. Counterweights are installed in order to make the dynamic characters of the model bridge coincide with that of the real bridge. Seventeen horizontal acceleration sensors are installed along the main girder to acquire the lateral dynamic response of the model bridge. The layout of the model bridge and the arrangement of the acceleration sensors are shown in Figure 5, and the constructed model bridge is shown in Figure 6.
Acceleration responses of the model bridge under the excitation of white noise were collected. The peak ground acceleration (PGA) of the white noise was 0.1 g, and the sampling frequency of the testing system was 256 Hz. The boxplot was adopted to inspect the quality of the original measured data. The inspection results of the 17 sensors are shown in Figure 7.

As shown in Figure 7, the boxplot can intuitively present the distribution and outliers of the measured data. The extents of UL and LL of each sensor along the main girder have a correspondence with the amplitude of the first-order mode shape of the girder. But the distributions of UL and LL of sensor 5# are not in conformity with the envelope of the mode shape. It can be speculated that the measured data from sensor 5# was unreliable; hence, it was neglected in the following processes.

Then, MF was adopted to suppress the noise components in the original data in order to improve the data quality. Due to the limitation on space, only the data measured from sensor 1# was taken as the example to demonstrate the denoising process. By using the SE size determination method proposed in this paper, the SE sizes were determined.
as 5. The original data, the improved data, and the residual along with their Fourier spectrums of sensor #1 are presented in Figure 8.

As can be seen in Figure 8, the amplitude of the measured data is decreased after the process of MF, but there is no significant change between the Fourier spectrums of the original data and the processed data except a small amount of energy loss in the latter one. The waveform of the residual is similar with that of white noise, and its Fourier spectrum has a wide distribution along the frequency domain.

The boxplot was adopted again to inspect the quality of the processed data, as shown in Figure 9. It can be seen that the distributions and numbers of outliers of each sensor are reduced, and it can be concluded that the quality of the measured data is improved by adopting MF.

DATA-SSI was applied to identify the modal parameters of the model bridge. The stabilization diagram was
applied to eliminate the false modes and present the stability of the identification results. In the stabilization diagram, the blue, green, and red points represent the stability of frequency, damping ratio, and mode shape, respectively. The comparison of the stabilization diagrams of the original data and the improved data was performed, as shown in Figure 10. The comparison results of identified frequencies of the original data and improved data along with the calculated frequencies are exhibited in Table 1.

As can be seen in Figure 10 and Table 1, the frequency values of the first two poles in Figure 10(a)) are almost the same with that in Figure 10(b)), and they are all coincident with the calculated values. However, the first two poles in Figure 10(b)) possess much more red points, which means that the identification results are more reliable. There are two poles located at 7.35 Hz and 8.64 Hz in Figure 10(a)), and the two poles mainly consist of blue points. While there is only one pole with a considerable amount of red points located at 8.04 Hz in Figure 10(b)), the identified value is very close to the calculated value. The third-order modal parameters are not identified with the original data, but the mean frequency value of the last two poles in Figure 10(a) is 8.00 Hz, which is also very close to the calculated value. Thus, it can be speculated that the third-order frequency is divided into two parts in Figure 10(a)); and they are concentrated after the data denoising. A conclusion can be drawn that the data quality is significantly improved by adopting EDA and MF.

3.2. Sutong Bridge. Sutong Bridge, a long-span cable-stayed bridge located on the Yangtze River, was taken as another instance to verify the proposed method, as shown in Figure 11. The overall span arrangement of Sutong Bridge is $2 \times 100 \text{ m} + 2 \times 150 \text{ m} + 1088 \text{ m} + 2 \times 150 \text{ m} + 2 \times 100 \text{ m}$. Sections of the first-side span on each side and the main span are selected to be installed with vertical acceleration sensors. The layout of the bridge and the arrangement of the 14 vertical acceleration sensors are shown in Figure 12.

The acceleration responses of the main girder of Sutong Bridge were acquired under the ambient excitation with a sampling frequency of 20 Hz. The boxplots of the 14 vertical acceleration sensors are listed in Figure 13. It can be

---

**Table 1: Comparison of identified and calculated frequency values.**

<table>
<thead>
<tr>
<th>Item</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculated value</td>
<td>2.82</td>
<td>4.28</td>
<td>8.16</td>
</tr>
<tr>
<td>Identified value with original data</td>
<td>2.95</td>
<td>4.24</td>
<td>7.35/8.64</td>
</tr>
<tr>
<td>Identified value with improved data</td>
<td>2.95</td>
<td>4.23</td>
<td>8.04</td>
</tr>
</tbody>
</table>

---

**Figure 9**: Boxplots of improved data.

**Figure 10**: Stabilization diagrams of modal parameter. (a) Original data. (b) Improved data.
deduced that the boxplots of sensors 1#, 2#, 9#, and 12# are abnormal as compared with others. Hence, the data measured from those sensors should be neglected in the following processes.

The sampling frequency of 20 Hz was too low to adopt MF; therefore, the original measured data were resampled from 20 Hz to 256 Hz. Then, MF was adopted to suppress the noise components inside the original data. The data measured from sensor 3# was taken as the example to demonstrate the denoising process. The size of SE was determined as 11 according to the proposed method. After the process of MF, the data were decimated back to 20 Hz. The original data, the improved data, and the residual along with their Fourier spectrums of sensor 3# are presented in Figure 14. The same conclusions of Figure 8 also can be drawn from Figure 14.

The boxplot was adopted again to inspect the quality of the improved data, as shown in Figure 15. For Sutong Bridge, the number of outliers is not significantly reduced after adopting MF. The main reason for this phenomenon is that there are impulse responses of the bridge due to the impact of environmental factors. The amplitude of the impulse responses vary from high to low gradually, and MF cannot filter out this kind of component; as a result, the data points with large amplitude are still taken as outliers in the boxplot.

DATA-SSI combined with the stabilization diagram was applied to identify the modal parameters of Sutong Bridge. The modal parameter identification results of the original data and improved data are demonstrated in Figure 16.

According to the FEM calculation results, the first six frequencies of Sutong Bridge are mainly distributed in the range of 0 to 0.5 Hz. However, there are only two stable poles in the range of 0 to 0.5 Hz in Figure 16(a)), and this number of identified modal parameters is far...
from being enough for bridge operational state analysis. The modal parameter identification results in Figure 16(b)) are significantly improved with the improved data. More poles of modal parameter with low frequency are identified while the third pole located around 0.3 Hz is more stable. The comparison results of identified frequencies of the original data and improved data along with the calculated frequencies are exhibited in Table 2.

As can be seen in Table 2, only the first- and third-order modal parameters are identified with the original data, while the first six order modal parameters are identified with the improved data and the identified values generally align with the calculated ones. Obviously, the modal parameter identification results of the improved data are more accurate than that of the original data. The main reason for causing the above differences is that most of the structural responses in the original data are submerged by
4. Conclusions

A data inspecting and denoising method for DATA-SSI was proposed in this paper. A time-domain data inspecting tool termed EDA was adopted to inspect the data quality. It can efficiently visualize the data quality and locate the malfunctioning sensors. A time-domain filter named MF along with an automated SE size determination method was adopted to suppress the noise components. By adopting the MF technique, the noise components in the original measured data can be suppressed effectively and the valuable structural responses are remained without distortion. Then, the improved data were processed by DATA-SSI to identify the modal parameters. A large-scale cable-stayed model bridge and a real long-span cable-stayed bridge were taken as instances to verify the proposed method. The results show that the data quality is significantly improved by the proposed method, and the modal parameter identification results of DATA-SSI with the improved data are more accurate and reliable.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was funded by the National Basic Research Program of China (no. 2013CB036302).

References

[4] X. Zhou, D. Shan, and Q. Li, “Modal parameter identification of a curved cable-stayed model bridge based on EDA and...


