Time Series Imputation via $L_1$ Norm based Singular Spectrum Analysis

Mahdi Kalantari*, Masoud Yarmohammadi*, Hossein Hassani† and Emmanuel Sirimal Silva‡

*Department of Statistics, Payame Noor University, 19395-4697, Tehran, Iran.
†Research Institute of Energy Management and Planning, University of Tehran, Iran
‡Fashion Business School, London College of Fashion, University of the Arts London, UK

Abstract

Missing values in time series data is a well-known and important problem which many researchers have studied extensively in various fields. In this paper, a new nonparametric approach for missing value imputation in time series is proposed. The main novelty of this research is applying the $L_1$ norm based version of Singular Spectrum Analysis (SSA), namely $L_1$-SSA which is robust against outliers. The performance of the new imputation method has been compared with many other established methods. The comparison is done by applying them to various real and simulated time series. The obtained results confirm that the SSA based methods, especially $L_1$-SSA can provide better imputation in comparison to other methods.

Keywords: Time Series, Basic SSA, $L_1$-SSA, Reconstruction, Missing value, Imputation.

1 Introduction

When dealing with real-world situations, missing values are commonly encountered in time series due to many reasons such as instrument malfunctions or failures to record observations, human mistakes and lost records. Eliminating those values may result in the loss of key information relevant to the inference. Imputation, which is the estimation of missing values, is an important part of the data cleaning process in time series analysis [1].

Most statistical analysis tools could be used after the imputation of missing values. It is noteworthy that imputing missing values alters the original time series; consequently, wrong imputation can severely affect the forecasting performance [2]. To this end, some authors believe that the treatment of missing observations can be more important than the choice of forecasting method [1]. Hence, employing effective and sound imputing algorithms to obtain the best possible imputes is of great importance. Missing data also prevents the production of statistically reliable statements about the variables and often further data analysis steps rely on complete data sets.

Imputation is a widespread area in time series analysis and some methods have been developed for imputing in time series. Examples of some traditional methods can be found in [3–7].
Choosing the proper imputing technique depends on the structure of time series concerned. Different series may require different strategies to impute missing values. An Expectation-Maximisation (EM) algorithm based method for imputation of missing values in multivariate normal time series has been proposed in [8]. This imputation algorithm accounts for both spatial and temporal correlation structures [9]. State-space representation or Kalman filter approach is another suitable method used for imputing, see [10] for more details. The use of ARIMA and SARIMA models for imputation of univariate time series was evaluated in [11]. The missing value estimation in the context of additive outliers and influential observations in time series can be found in [12, Chap. 6].

For maximum likelihood fitting of ARMA models and estimation of ARIMA models with missing values see [13, 14].

A major drawback of standard imputation methods in time series is assuming stationarity for the data, linearity for the model or normality for the errors which can only provide an approximation to the real situation. One solution for overcoming these difficulties is via the employment of nonparametric approaches. Given the advantage of not being restricted by any of the parametric assumptions enables nonparametric methods to provide a much closer representation of the real world scenario [15]. As such, nonparametric methods are extensively used in statistical analyses. The Singular Spectrum Analysis (SSA) technique is a very good example of such methods. Applications of this powerful and nonparametric technique is increasingly wide spread in time series analysis and other fields; for references see e.g. [15–19].

Interestingly, one of the effective applications of SSA is imputation in time series. Some methods for imputation based on SSA have been designed for stationary time series [20, 23] whilst in [21] a more general approach which is applicable to different kinds of time series was proposed. An extension of SSA forecasting algorithms for gap filling was proposed in [24]. In this subspace approach, the structure of the extracted component is continued to the gaps caused by the missing values. In another gap filling method proposed in [25], a weighted combination of the forecasts and hindcasts yielded by the recurrent SSA forecasting algorithm was used. This approach was further enhanced by using bootstrap re-sampling and a weighting scheme based on sample variances in [26].

In this paper, we propose a new approach for missing data imputation in univariate time series within the SSA framework. In this method, missing values are replaced by initial values and then reconstructed repeatedly until convergence occurs. The last reconstructed values are considered as imputed values. It is noteworthy that the idea underlying the iterative algorithm was derived from [21] and was in fact suggested earlier for imputation of gaps in matrices in [22]. The main novelty of the proposed technique is its application of the $L_1$ norm based version of SSA, namely $L_1$-SSA which was introduced in [27]. Recall that the basic version of SSA is based on the Frobenius norm or $L_2$ norm. The main advantages of this newly proposed approach are its robustness against outliers and lack of assumptions relating to the stationarity of time series and normality of random errors. The results from the proposed method are compared with those attained via other established methods such as Interpolation, Kalman Smoothing and Weighted Moving Average. The obtained results confirm that the SSA based methods, especially $L_1$-SSA can provide better imputation in comparison to other methods.

The remainder of this paper is organised as follows. A brief introduction into $L_1$-SSA and the new imputation method are given in Section 2. The other imputation methods are presented in Section 3 in more detail. In addition, this section also evaluates the perfor-
mance of imputation methods via applications which compare them with simulated and
real time series. Finally, Section 4 presents a summary of the study and some concluding
remarks.

2 New Imputation Method

In this section; first, a short description of $L_1$-SSA is presented. Thereafter, we propose
the new imputation method based on $L_1$-SSA.

2.1 A Brief Description of $L_1$-SSA

The SSA technique consists of two complementary stages: Decomposition and Reconstruc-
tion, and both of these include two separate steps [28]. At the first stage we decompose
the series in order to enable signal extraction and noise reduction. At the second stage we
reconstruct a less noisy series and use the reconstructed series for forecasting new data
points [19]. The theory underlying SSA is explained in more detail in [28]. The most
common version of SSA is called Basic SSA [28]. It is notable that the matrix norm used
in Basic SSA is the Frobenius norm or $L_2$-norm. Recently, a newer version of SSA which is
based on $L_1$-norm and therefore called $L_1$-SSA was introduced and it was confirmed that
$L_1$-SSA is robust against outliers [27]. In the following, the steps of $L_1$-SSA are concisely
presented. For more detailed information on $L_1$-SSA, see [27].

Stage 1: Decomposition

Let $Y_N = \{y_1, \ldots, y_N\}$ be the time series and $L$ ($2 \leq L < N - 1$) be some integer called
the window length.

Step 1: Embedding

In this step; firstly, the lagged vectors of size $L$ are built as follows:

$$X_i = (y_i, \ldots, y_{i+L-1})^T, \quad 1 \leq i \leq K,$$

where $K = N - L + 1$. Secondly, the trajectory matrix of the time series $Y_N$ is defined as:

$$X = [X_1 : \cdots : X_K] = (x_{ij})_{i,j=1}^{L,K} = \begin{pmatrix}
    y_1 & y_2 & y_3 & \cdots & y_K \\
    y_2 & y_3 & y_4 & \cdots & y_{K+1} \\
    y_3 & y_4 & y_5 & \cdots & y_{K+2} \\
    \vdots & \vdots & \vdots & \ddots & \vdots \\
    y_L & y_{L+1} & y_{L+2} & \cdots & y_N
\end{pmatrix}$$

Note that $X$ has equal elements on the anti-diagonals $i + j = \text{const}$. Matrices of this type
are called Hankel matrices.
Step 2: Singular Value Decomposition (SVD)

In this step, the Singular Value Decomposition (SVD) of the trajectory matrix $X$ is performed. Suppose that $\lambda_1, \ldots, \lambda_L$ are the eigenvalues of $XX^T$ taken in the decreasing order of magnitude ($\lambda_1 \geq \cdots \geq \lambda_L \geq 0$) and $U_1, \ldots, U_L$ are the eigenvectors of the matrix $XX^T$ corresponding to these eigenvalues. Set $d = \text{rank } X = \max \{i, \text{such that } \lambda_i > 0\}$, the number of positive eigenvalues. If we denote $V_i = X^T U_i / \sqrt{\lambda_i} \quad (i = 1, \ldots, d)$, the SVD of the trajectory matrix $X$ in $L_1$-SSA can be written as:

$$X = X_1 + \cdots + X_d = \sum_{i=1}^{d} w_i \sqrt{\lambda_i} U_i V_i^T,$$

where $X_i = w_i \sqrt{\lambda_i} U_i V_i^T$. The $w_i$ is the weight of singular value $\sqrt{\lambda_i}$. These weights are diagonal elements of diagonal weight matrix $W = \text{diag}(w_1, w_2, \ldots, w_d, 0, 0, \ldots, 0)$ and are computed such that $\|X - U W \Sigma V^T\|_{L_1}$ is minimized; where $U = [U_1 : \cdots : U_L]$, $V = [V_1 : \cdots : V_L]$, $\Sigma = \text{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_d})$ and $\|\cdot\|_{L_1}$ is the $L_1$ norm of a matrix. For more information, see [27].

Stage 2: Reconstruction

Step 3: Grouping

In this step, we partition the set of indices $\{1, \ldots, d\}$ into $m$ disjoint subsets $I_1, \ldots, I_m$. Let $I = \{i_1, \ldots, i_p\}$. Then the matrix $X_I$ corresponding to the group $I$ is defined as $X_I = X_{i_1} + \cdots + X_{i_p}$. For example, if $I = \{1, 2, 7\}$ then $X_I = X_1 + X_2 + X_7$. In signal extraction problems, $r$ leading eigentriples are chosen. That is, indices $\{1, \ldots, d\}$ are partitioned into two subsets $I_1 = \{1, \ldots, r\}$ and $I_2 = \{r + 1, \ldots, d\}$.

Step 4: $L_1$-Hankelization

In this step, we seek to transform each matrix $X_{I_j}$ of the grouping step into a Hankel matrix so that these can subsequently be converted into a time series, which is an additive component of the initial series $Y_N$. Let $\mathcal{H} A$ be the result of the Hankelization of matrix $A$. In $L_1$-SSA, Hankelization corresponds to computing the median of the matrix elements over the “antidiagonal”. This type of Hankelization has an optimal property in the sense that the matrix $\mathcal{H} A$ is the nearest to $A$ (with respect to the $L_1$ norm) among all Hankel matrices of the same dimension [27]. On the other hand, $\|A - \mathcal{H} A\|_{L_1}$ is minimum; so this type of Hankelization is denoted by $L_1$-Hankelization.

$L_1$-Hankelization applied to a resultant matrix $X_{I_j}$ of the grouping step, produces a reconstructed series $\tilde{Y}_N^{(j)} = \{\tilde{y}_1^{(j)}, \ldots, \tilde{y}_N^{(j)}\}$. Therefore, the initial series $Y_N = \{y_1, \ldots, y_N\}$ is decomposed into a sum of $m$ reconstructed series:

$$y_t = \sum_{j=1}^{m} \tilde{y}_t^{(j)}, \quad t = 1, 2, \ldots, N.$$
2.2 New Imputation Algorithm Based on $L_1$-SSA

Prior to presenting the algorithm, we find it pertinent to clarify that we do not change the $L_2$-norm to $L_1$-norm during the construction of projectors. Instead, this change occurs at the Hankelization step. Thus, the decomposition stage results in a correction of the $L_2$ decomposition and is therefore in reality, a $L_1$-$L_2$ decomposition.

Let $Y_i^{(i)} = \{y_1, \ldots, y_{i-1}, \star, y_{i+1}, \ldots, y_N\}$ be the time series where only the $i$th value is missing ($i = 1, \ldots, N$). The symbol ‘$\star$’ stands for the missing value and it is obvious that $i$ is the position of this value. In the iterative $L_1$-SSA imputation method, missing values are replaced by initial values and then reconstructed repeatedly until convergence occurs, as proposed in [21]. The last reconstructed values are considered as imputed values. This imputation algorithm contains the following steps:

Step 1) Set a suitable initial value in place of missing data.

Step 2) Choose reasonable values of $L$ and $r$.

Step 3) Reconstruct the time series where its missing data is replaced with a number.

Step 4) Replace the $i$th value of time series with its $i$th reconstructed value.

Step 5) Repeat steps 3 and 4 until the absolute value of the difference between successive replaced values of the time series by their reconstructed value is less than $\delta$. ($\delta$ is the convergence threshold.)

Step 6) Consider the final replaced value as the imputed value.

3 Empirical Results

In this section; firstly, the other imputation methods are briefly discussed. Secondly, the comparison criteria which are used in this paper are defined. Thirdly, the performance of algorithms for imputation of one missing value are compared via a simulation study. Finally, all of the imputation methods are assessed by applying them to real data.

3.1 Other Imputation Methods

The other imputation algorithms of univariate time series which are used in this paper are as follows:

1. **Iterative Basic SSA**: In this method, the imputation algorithm proposed in Section 2.2 is used for imputation via Basic SSA.

2. **Interpolation**: Linear, spline and Stineman interpolation are used to impute missing values.

3. **Kalman Smoothing**: The Kalman smoothing on the state space representation of an ARIMA model is used for imputation.

4. **LOCF**: Each missing value is replaced with the most recent present value prior to it (Last Observation Carried Forward).
5. **NOCB**: The LOCF is done from the reverse direction, starting from the back of the series (Next Observation Carried Backward).

6. **Weighted Moving Average**: Missing values are replaced by its weighted moving average. The average in this implementation is taken from an equal number of observations on either side of a missing value. For example, for imputation of missing value at location $i$, the observations $y_{i-2}, y_{i-1}, y_{i+1}, y_{i+2}$, are used to calculate the mean for moving average window size 4 (2 left and 2 right). The moving average window size 8 (4 left and 4 right) is taken into account in this paper. The weighted moving average is used in the following three ways:

- **Simple Moving Average (SMA)**: All observations in the moving average window are equally weighted for calculating the mean.
- **Linear Weighted Moving Average (LWMA)**: Weights decrease in arithmetical progression. The observations directly next to the $i$th missing value ($y_{i-1}, y_{i+1}$) have weight 1/2, the observations one further away ($y_{i-2}, y_{i+2}$) have weight 1/3, the next $y_{i-3}, y_{i+3}$ have weight 1/4 and so on.
- **Exponential Weighted Moving Average (EWMA)**: Weights decrease exponentially. The observations directly next to the $i$th missing value have weight $\frac{1}{2^1}$, the observations one further away have weight $\frac{1}{2^2}$, the next have weight $\frac{1}{2^3}$ and so on.

In SSA based imputation methods (Basic SSA and $L_1$-SSA), for reconstruction of simulated series in Section 3.3, the number of leading eigenvalues ($r$) have been selected according to the rank of the corresponding trajectory matrix. All calculations of imputation methods (except SSA) are done with the help of the R package `imputeTS`. For more information see [29]. For Basic SSA computations, the R package `Rssa` is employed. For more details see [30–32].

### 3.2 Comparing Criteria

In this paper, the performance of algorithms for imputation of one missing value are compared by means of the commonly applied accuracy measures of Root Mean Squared Error (RMSE) and Mean Absolute Deviation (MAD). They are defined as follows:

\[
RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} e_i^2},
\]

\[
MAD = \frac{1}{N} \sum_{i=1}^{N} |e_i|,
\]

where $e_i = y_i - \hat{y}_i$ is the imputing error and $\hat{y}_i$ is the imputed value for $y_i$.

The following ratios are used for comparing $L_1$-SSA and other methods:

\[
RRMSE = \frac{\text{RMSE based on } L_1\text{-SSA}}{\text{RMSE based on another method}},
\]

\[
RMAD = \frac{\text{MAD based on } L_1\text{-SSA}}{\text{MAD based on another method}}.
\]
It is clear that if the above ratios are less than 1, then we can conclude that $L_1$-SSA outperforms the competing method of imputation by $1 - \text{RRMSE}$ percent (or $1 - \text{RMAD}$ percent). For comparing Basic SSA and $L_1$-SSA, the Ratio of Absolute Error (RAE) is used:

$$RAE(i) = \frac{|e_i| \text{ based on } L_1\text{-SSA}}{|e_i| \text{ based on Basic SSA}},$$

where $RAE(i)$ denotes the value of RAE after imputing the $i$th missing observation. If $RAE(i) < 1$, then $L_1$-SSA outperforms Basic SSA. Alternatively, when $RAE(i) > 1$, it would indicate that the performance of $L_1$-SSA is worse than Basic SSA. For better comparison, the dashed horizontal line $y = 1$ is added to all figures of RAE.

### 3.3 Simulation Results

The following simulated time series are used in this study:

(a) $y_t = \sin(\pi t/6) + \epsilon_t$

(b) $y_t = \exp(0.01t) + \epsilon_t$

(c) $y_t = 0.1t + \sin(\pi t/6) + \sin(\pi t/3) + \epsilon_t$

(d) $y_t = 0.1t + \sin(\pi t/12) + \sin(\pi t/6) + \sin(\pi t/4) + \sin(\pi t/3) + \sin(5\pi t/12) + \epsilon_t$

where $t = 1, 2, \ldots, 100$ and $\epsilon_t$ is the noise generated by a normal distribution. In each of the simulated series, one observation is removed artificially at different positions to create one missing value. Additionally, three outliers with different magnitude are inserted in each simulated series at non-equidistant positions for assessing the performance of the imputation methods when faced with outliers. It is assumed that the positions of the missing values are not the same as of the outliers.

For SSA imputation, we need two parameters; $L$ and $r$. The window length ($L$) for those cases is chosen as 48, 50, 48 and 48 respectively. For more details and useful recommendations about window length selection, see [17]. The number of the eigenvalues that are required for reconstruction for those cases are 2, 1, 6 and 12 respectively. In the simulation study; firstly, the noise is generated by a normal distribution. Secondly, the generated noise is added to a noiseless time series (e.g. Sine series). Thirdly, the ratio of the comparing criteria (RRMSE and RMAD) are calculated. These three stages are repeated 1000 times and finally, the mean of RRMSE and RMAD are reported.

In Table 1, the different imputation methods are compared in terms of RRMSE and RMAD. Results show that $L_1$-SSA reports better performance in comparison to other methods in all cases. It is noteworthy that Basic SSA is the next best imputation method in all cases.
Table 1: Comparison of imputation methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>case a</th>
<th>case b</th>
<th>case c</th>
<th>case d</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RRMSE</td>
<td>RMAD</td>
<td>RRMSE</td>
<td>RMAD</td>
</tr>
<tr>
<td>Basic SSA</td>
<td>0.63</td>
<td>0.6</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>Linear Inter.</td>
<td>0.26</td>
<td>0.38</td>
<td>0.34</td>
<td>0.52</td>
</tr>
<tr>
<td>Spline Inter.</td>
<td>0.16</td>
<td>0.28</td>
<td>0.21</td>
<td>0.32</td>
</tr>
<tr>
<td>Stineman Inter.</td>
<td>0.26</td>
<td>0.43</td>
<td>0.33</td>
<td>0.51</td>
</tr>
<tr>
<td>Kalman Smoothing</td>
<td>0.32</td>
<td>0.34</td>
<td>0.65</td>
<td>0.67</td>
</tr>
<tr>
<td>LOCF</td>
<td>0.14</td>
<td>0.15</td>
<td>0.25</td>
<td>0.46</td>
</tr>
<tr>
<td>NOCB</td>
<td>0.14</td>
<td>0.15</td>
<td>0.24</td>
<td>0.45</td>
</tr>
<tr>
<td>SMA</td>
<td>0.14</td>
<td>0.12</td>
<td>0.58</td>
<td>0.62</td>
</tr>
<tr>
<td>LMA</td>
<td>0.18</td>
<td>0.16</td>
<td>0.56</td>
<td>0.62</td>
</tr>
<tr>
<td>EWMA</td>
<td>0.24</td>
<td>0.22</td>
<td>0.5</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Figures 1-4 show the plots of the errors for different imputation methods for all cases. From these figures we can conclude that the following results satisfy for all cases:

1. In the LOCF method, the absolute value of the imputation error increases if the missing value has been placed just after the outlier. However in NOCB method, this is true if the missing value has been placed just before the outlier.

2. In interpolation methods (Linear, Spline and Stineman), the absolute values of the imputation error for neighborhoods of the outliers are greater than elsewhere.

In case (a), the wave pattern of the imputation error is visible almost for all methods. Also in the $L_1$-SSA method, the imputation error at the end of series is greater than elsewhere.

Figure 1: Plots of imputation errors in Sine series (case a).

In case (b), the imputation errors show an upward pattern for Kalman smoothing method. Also in Weighted Moving Average methods (SMA, LMA and EWMA), the absolute values of the imputation error for neighbourhoods of the outliers are greater than elsewhere.
Figure 2: Plots of imputation errors in Exponential series (case b).

In case (c) similar to case (a), there is wave pattern in imputation errors almost for all methods. Also in the $L_1$-SSA method, the imputation error at the end of series is greater than elsewhere.

Figure 3: Plots of imputation errors for case c.

In case (d), similar to cases (a) and (c), there is wave pattern in imputation errors almost for all methods. Also in this case, the absolute values of the imputation error for neighbourhoods of the outliers are greater than the rest.
In Figure 5, the plots of absolute errors and RAE for SSA based imputation methods are presented for all cases. Interestingly, it is evident from these figures that $L_1$-SSA has superiority over Basic SSA for imputation of missing values when there are outliers in time series. The solid and dash lines correspond to basic SSA and $L_1$-SSA, respectively.

Figure 5: Plots of absolute errors and RAE for all cases.

3.4 Real Data

In this subsection, the efficiency of imputation methods are compared for imputing of one missing value in real data. To this end, three time series data sets are considered as follows:
1. **War series:** The U.S. combat deaths in the Vietnam War, monthly from January 1966 to December 1971 including 72 observations [33].

2. **Chickenpox series:** Monthly reported number of chickenpox in New York city from January 1931 to June 1972 comprising 498 observations [34].

3. **Measles series:** Number of cases of measles in Baltimore, monthly from January 1939 to June 1972 containing 402 observations [35].

Figure 6 shows the time series plot of these data sets. Here, let us assume that there are two outliers in February and May 1968 in the War series. Also assume that the Chickenpox series includes three outliers in March and April 1949 and March 1953, and that there are three outliers in February 1939, March 1944 and March 1949 in the Measles series.

![Time Series Plots](image)

Figure 6: Time series plot of real data.

For reconstructing, \( L = 21, 51, 28 \) and \( r = 7, 30, 21 \) are used for SSA based imputation in War, Chickenpox and Measles series, respectively. Similar to simulated series, one observation is removed deliberately at different positions to create one missing value. In Table 2, the different imputation methods are compared according to the RRMSE and RMAD criteria. Results indicate that \( L_{1}\)-SSA is the best imputation method. It is noteworthy that based on the RRMSE, the next best method is Basic SSA.
Table 2: Comparison of imputation methods for real data.

<table>
<thead>
<tr>
<th>Method</th>
<th>War series</th>
<th>Chickenpox series</th>
<th>Measles series</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RRMSE</td>
<td>RMAD</td>
<td>RRMSE</td>
</tr>
<tr>
<td>Basic SSA</td>
<td>0.91</td>
<td>0.79</td>
<td>0.98</td>
</tr>
<tr>
<td>Linear Inter.</td>
<td>0.86</td>
<td>0.85</td>
<td>0.78</td>
</tr>
<tr>
<td>Spline Inter.</td>
<td>0.82</td>
<td>0.77</td>
<td>0.83</td>
</tr>
<tr>
<td>Stineman Inter.</td>
<td>0.86</td>
<td>0.84</td>
<td>0.8</td>
</tr>
<tr>
<td>Kalman Smoothing</td>
<td>0.81</td>
<td>0.73</td>
<td>0.94</td>
</tr>
<tr>
<td>LOCF</td>
<td>0.69</td>
<td>0.68</td>
<td>0.39</td>
</tr>
<tr>
<td>NOCB</td>
<td>0.69</td>
<td>0.68</td>
<td>0.39</td>
</tr>
<tr>
<td>SMA</td>
<td>0.85</td>
<td>0.83</td>
<td>0.29</td>
</tr>
<tr>
<td>LMA</td>
<td>0.88</td>
<td>0.89</td>
<td>0.36</td>
</tr>
<tr>
<td>EWMA</td>
<td>0.9</td>
<td>0.91</td>
<td>0.46</td>
</tr>
</tbody>
</table>

Figures 7-9 depict the plots of imputation errors for different imputation methods. It can be seen that the imputation error increases if the missing value is an outlier.

Figure 7: Plots of imputation errors in War series.
Figure 8: Plots of imputation errors in Chickenpox series.

Figure 9: Plots of imputation errors in Measles series.

Figure 10 shows the plots of absolute errors and RAE for SSA based imputation methods in real data. From these plots, it can be deduced that almost always, $L_1$-SSA has better performance than Basic SSA.
4 Conclusion

In this paper, we proposed a new nonparametric approach for missing value imputation of univariate time series within the SSA framework. In the proposed method, the $L_1$ norm based version of SSA, namely $L_1$-SSA, was applied for imputation of missing values in the presence of outliers.

The performance of the new imputation method was compared with many other established methods such as Interpolation, Kalman Smoothing and Weighted Moving Average with respect to RMSE and MAD criteria using both simulated and real world data.

In particular, it was expected that $L_1$-SSA would enable better imputation in comparison to basic SSA when faced with outliers, because $L_1$ norm is less sensitive than $L_2$ norm to the presence of outliers. It is interesting that the comparison of results confirm that almost always $L_1$-SSA outperforms basic SSA.

The results obtained in this study also indicates that the SSA based methods ($L_1$-SSA and basic SSA) can provide better imputation in comparison to other methods when faced with time series polluted by outliers. This was proven via both the simulation and application to real data.

In terms of future research, the capability of $L_1$-SSA for multiple imputation will be considered. The important issue of selecting the optimal parameters of SSA for imputation ($L$ and $r$) has potential for further exploration, and those interested can begin by considering the research in [21] which presents one approach to the choice of SSA parameters for iterative gap-filling.

References


