

Replacement of Missing Data and Outliers Using Wavelet Transform Methods



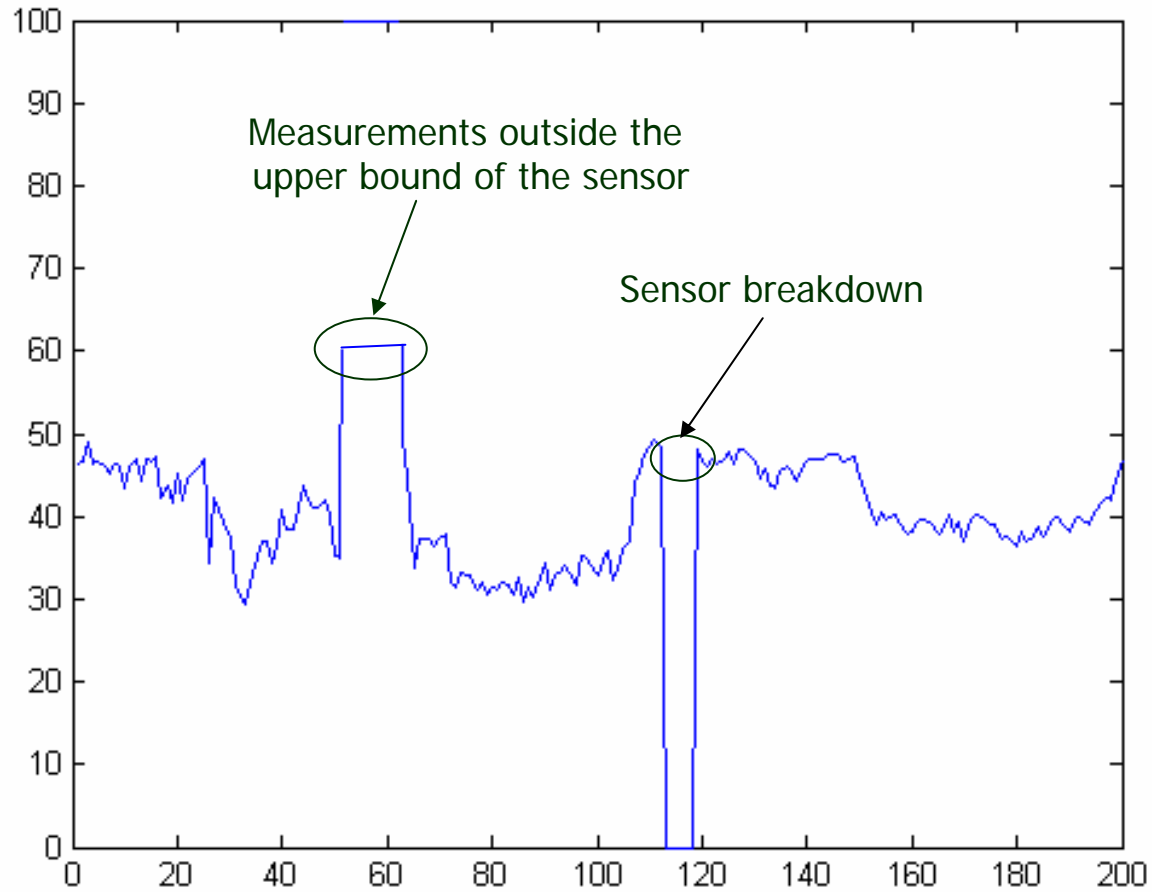
Liqian Zhang, Research Associate
Department of Chemical and Materials Engineering
University of Alberta

-
1. Motivation
 2. Discrete wavelet transform
 3. Proposed algorithms for treatment of missing data
 4. Examples to illustrate reconstruction of missing data
 5. Application to detect and reconstruct outliers
 6. Conclusions

Motivation

- Causes of missing data
 - Failure of measurement devices and/or errors in data management
 - Sensor breakdown
 - Measurement outside the range of the sensor
 - Data acquisition system malfunction
 - Energy blackouts
 - Interruption of transmission lines
 -

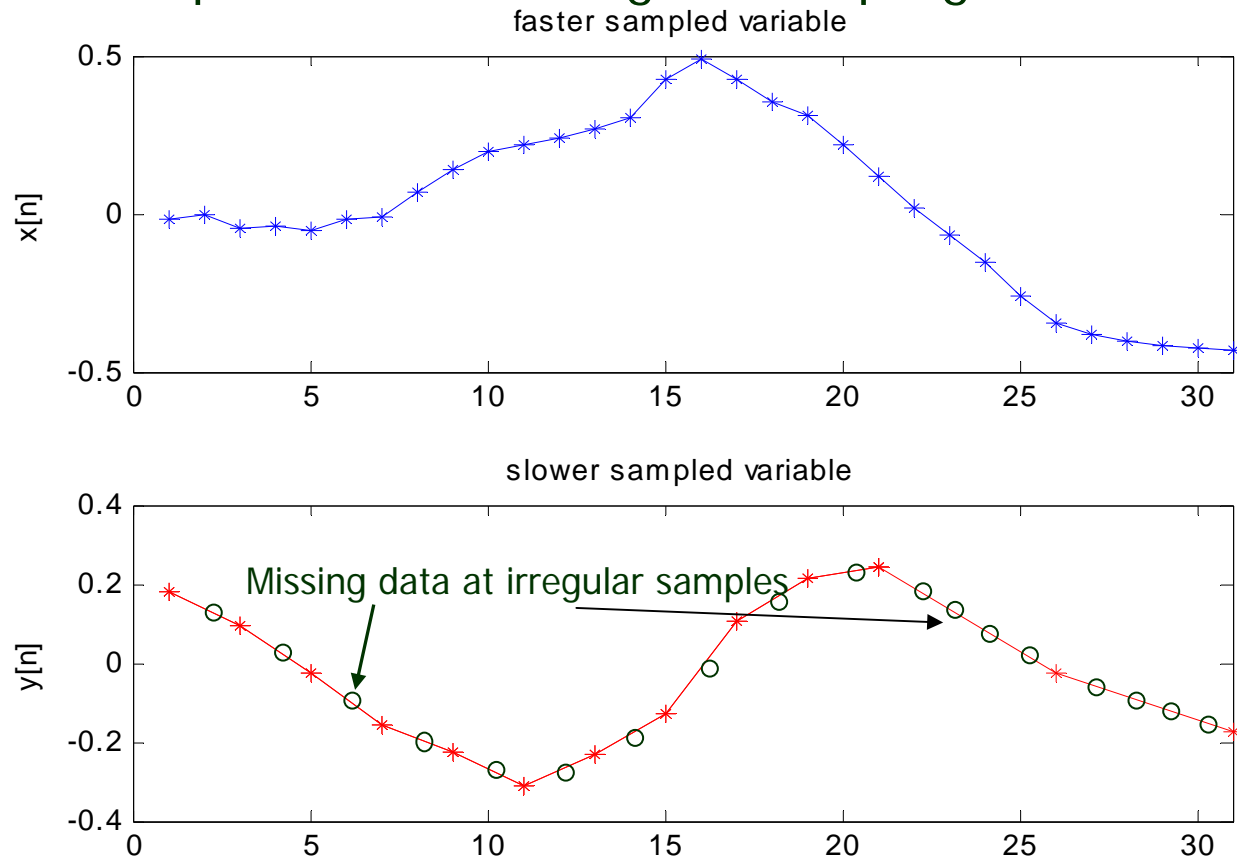
Example of measurement outside the range of the sensor or sensor breakdown



Motivation ... (Causes of missing data)

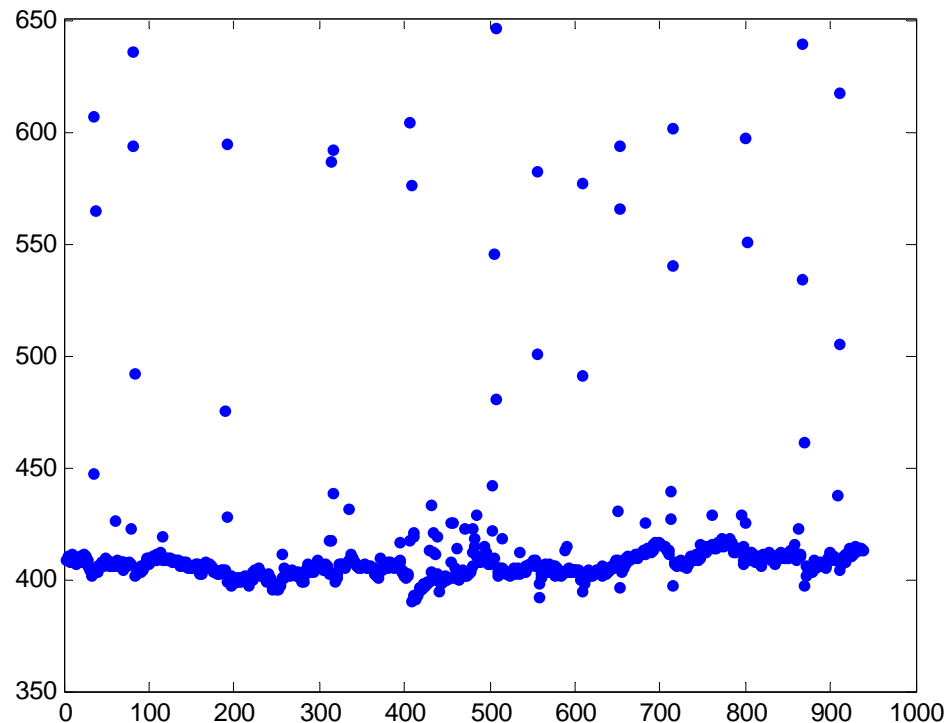
- Data may be missing because of the strategy of sampling: multi-rate or irregular sampling

An example: multi-rate/irregular sampling



Motivation ... (Causes of missing data)

- Outliers can be considered as missing data: v14 from Syncrude Canada



- Data compression (with default settings) is still a common practice in industry

Motivation

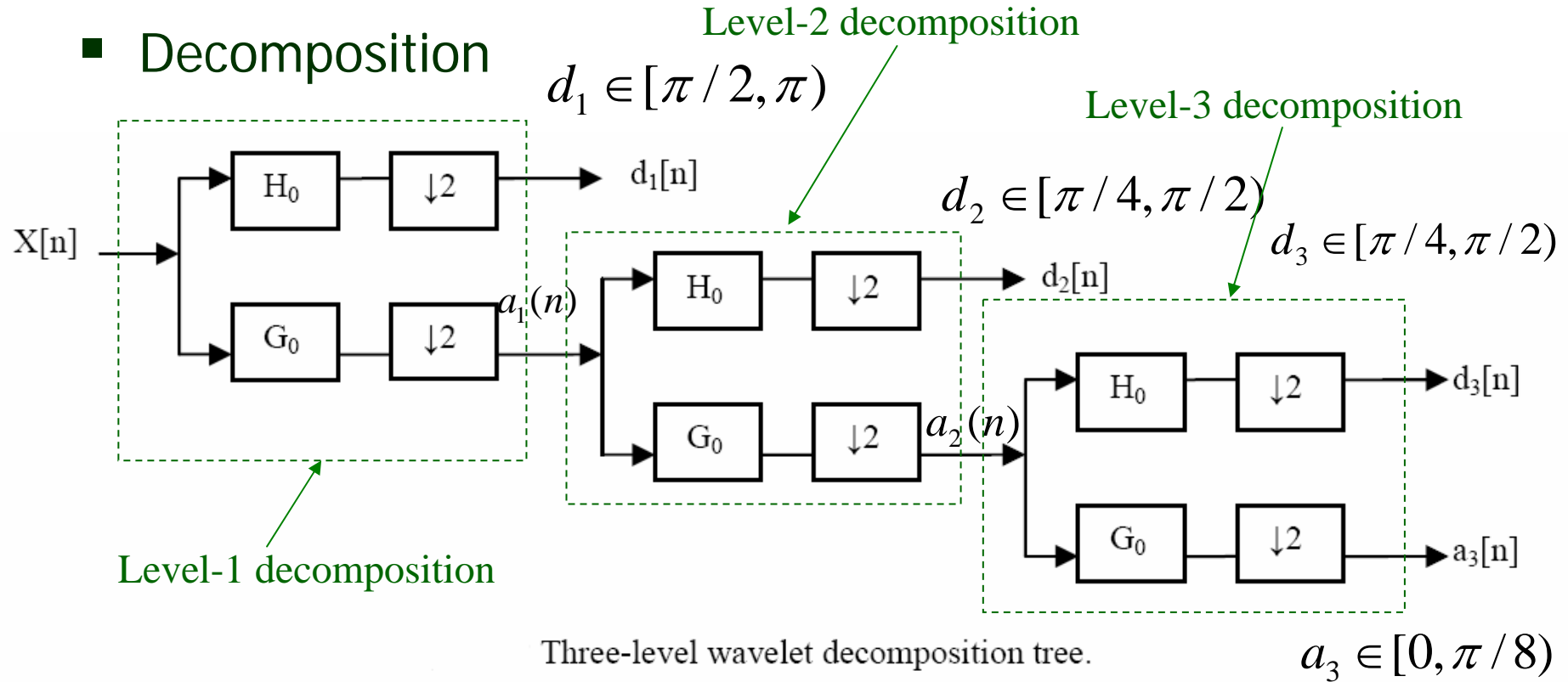
- Data driven methods are now extensively used in process industries for
 - Process identification
 - Process monitoring
- Such methods require well-conditioned data with the following features:
 - Uncompressed data or raw data
 - Properly time synchronized
 - Outlier detection and replacement
 - Reconstruction of missing data

Discrete wavelet transform

- Why use wavelet transform?
 - Existing methods of treating missing data
 - Direct interpolation in the time-domain (with no regard for information in the frequency domain)
 - Spectrum estimation (with no regard for information in the time domain)
 - Wavelet transform shows how the energy of signal varies with time and frequency
 - Applications:
 - For de-noising and compressing signals
 - In biology for cell membrane recognition
 - In finance for detecting quick variation of data
 - For machine condition monitoring and fault diagnosis ,....

Discrete wavelet transform

Decomposition



Discrete wavelet transform

- x is decomposed as

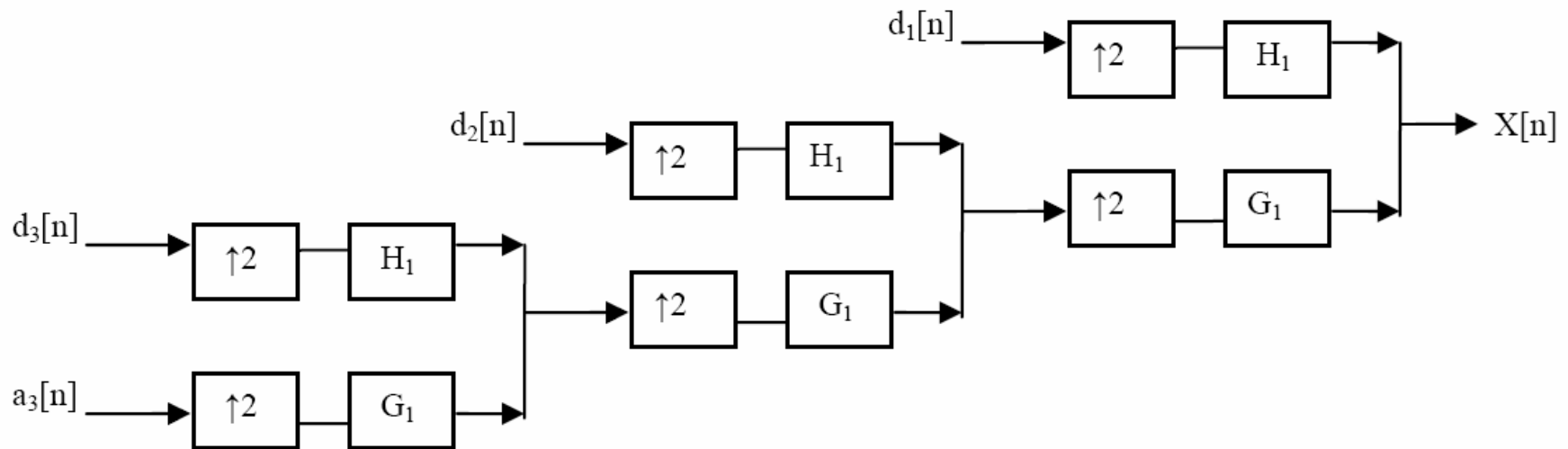
$$x = d_1 + d_2 + d_3 + a_3$$

$$d_1 \in [\pi/2, \pi), d_2 \in [\pi/4, \pi/2), d_3 \in [\pi/8, \pi/4), a_3 \in [0, \pi/8)$$

- DWT offers a good time resolution at high frequencies, and good frequency resolution at low frequencies.

Discrete wavelet transform

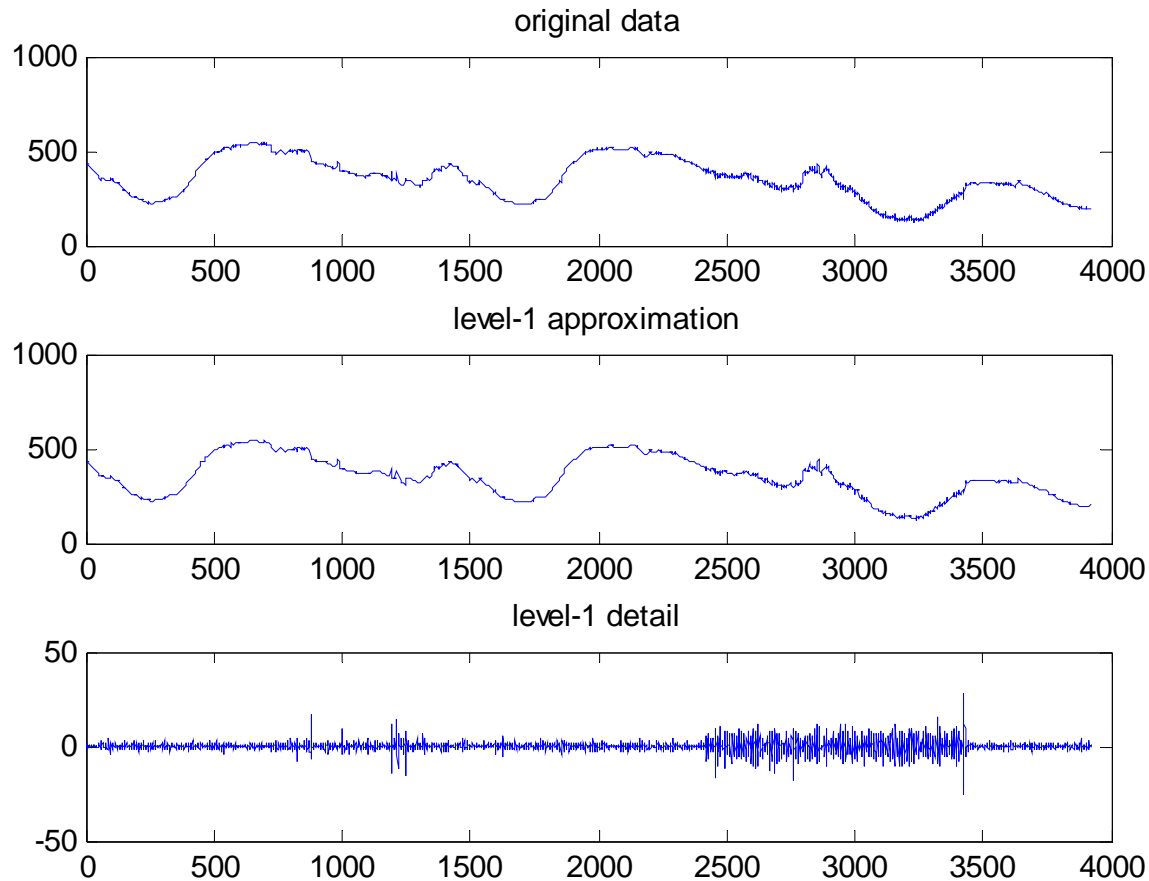
■ Reconstruction



Three-level wavelet reconstruction tree.

Discrete wavelet transform

- An example



Proposed algorithms for treatment of missing data¹³

- Data set considered

$$x = \left[x(0) \quad x(1) \quad \cdots \quad x(N-1) \right]^T$$

- Missing data description

- Regularly sampled

$$x(n) = y(t) \big|_{t=nT}$$

$y(t)$ is a continuous-time signal

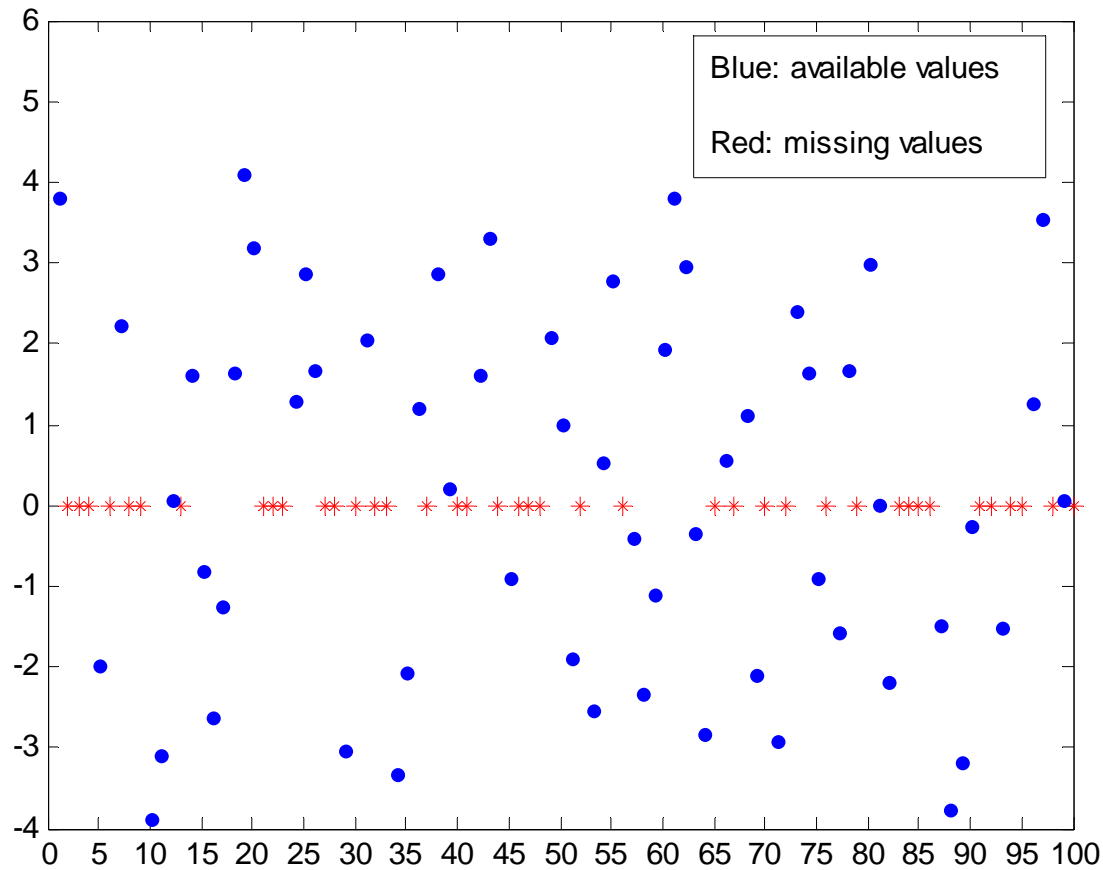
- Samples at some sampling instants are missing

- Two types of missing data

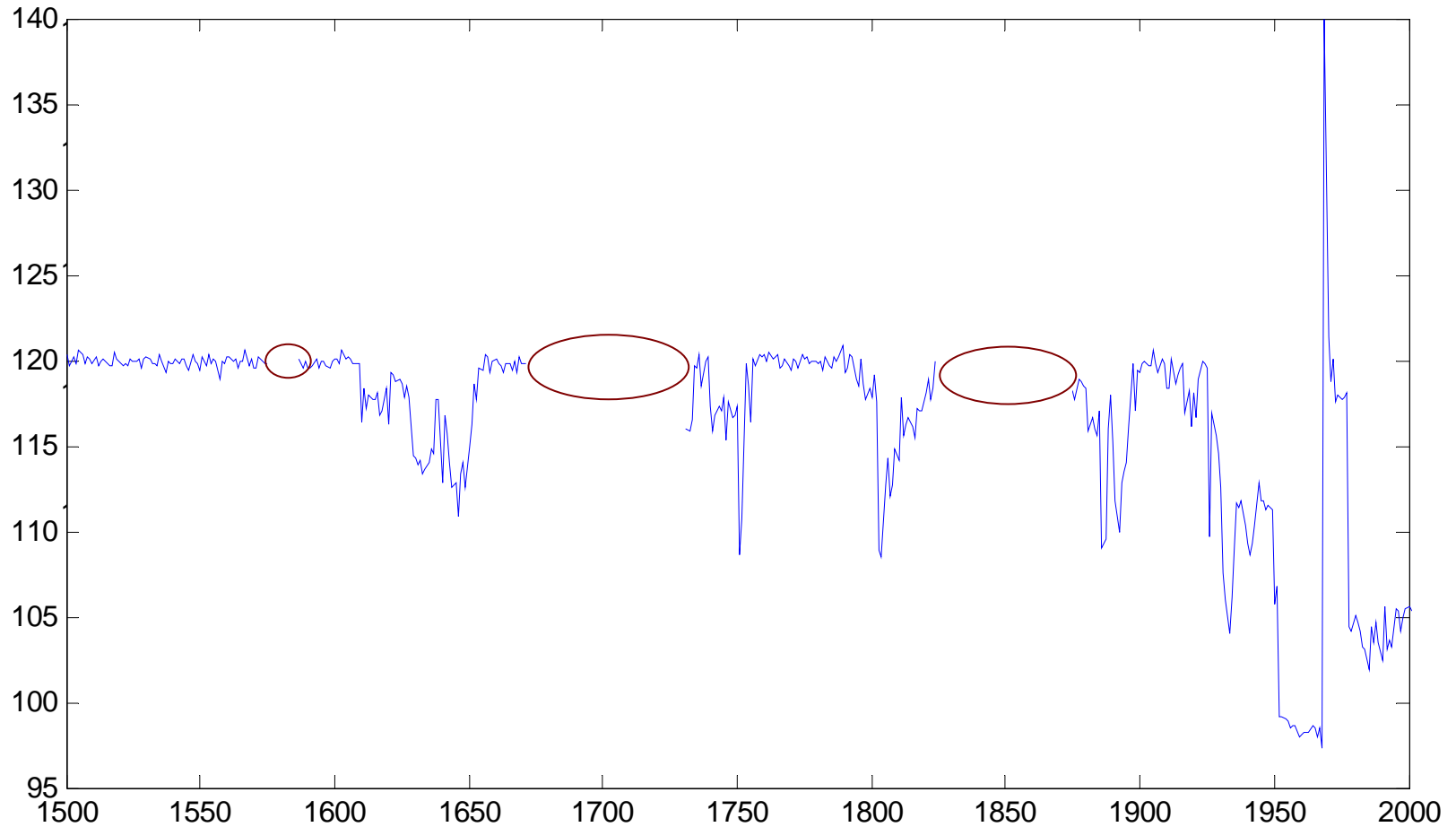
- Type 1 – randomly missing data: data are missing at random time instants
- Type 2 - gapped data: missing data constitutes some gaps

Proposed algorithms for treatment of missing data¹⁴

Type 1 – randomly missing data



Type 2: gapped data – Scaled industrial data from Matrikon



- Algorithm 1: Wavelet transform + EM algorithm

$$x = d_1 + a_1$$

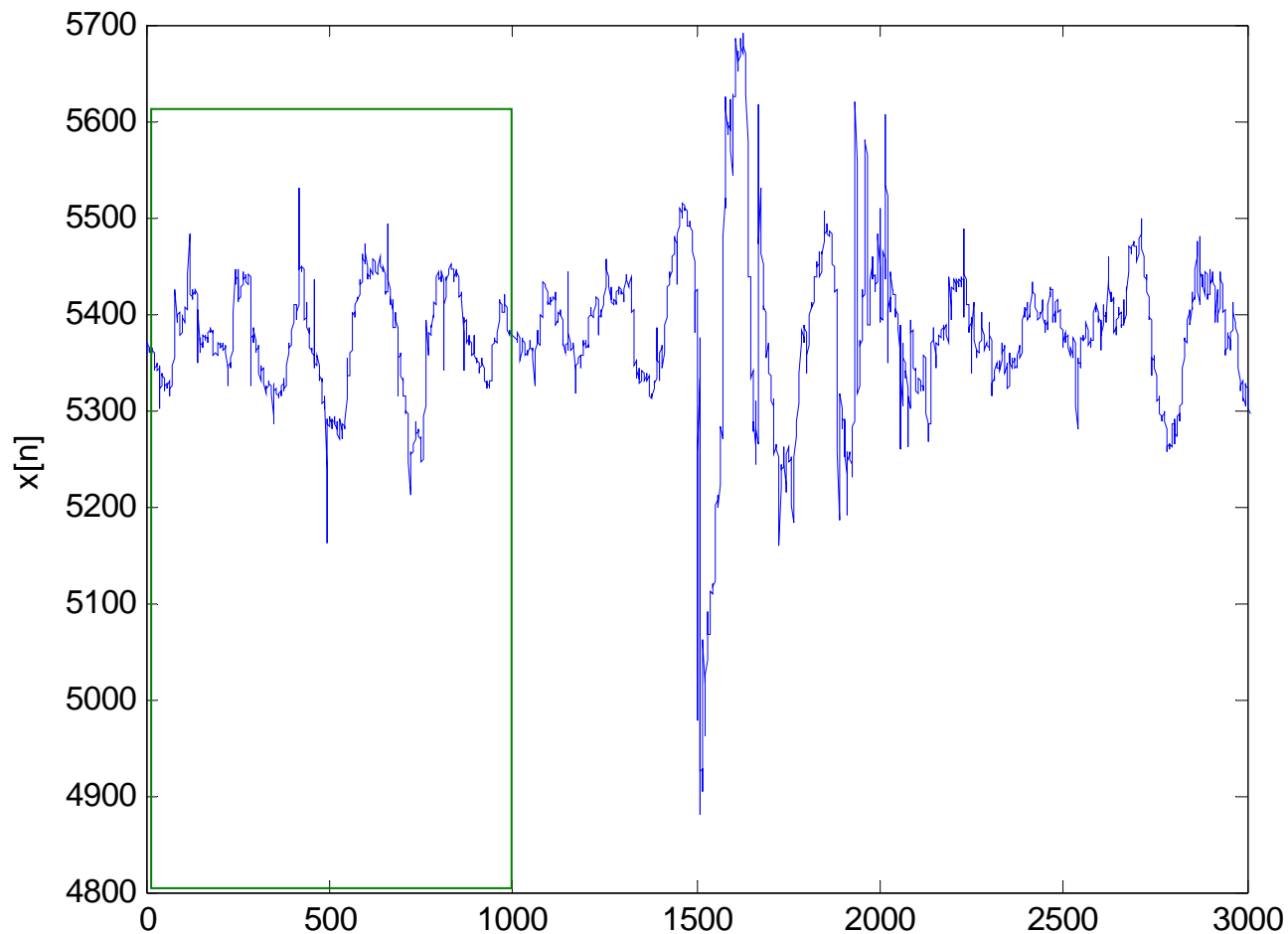
- Larger computational burden with larger data sets
- Sensitive to the initial choice
- Reliable for gapped data and more missing data

- Algorithm 2: Wavelet transform + least squares

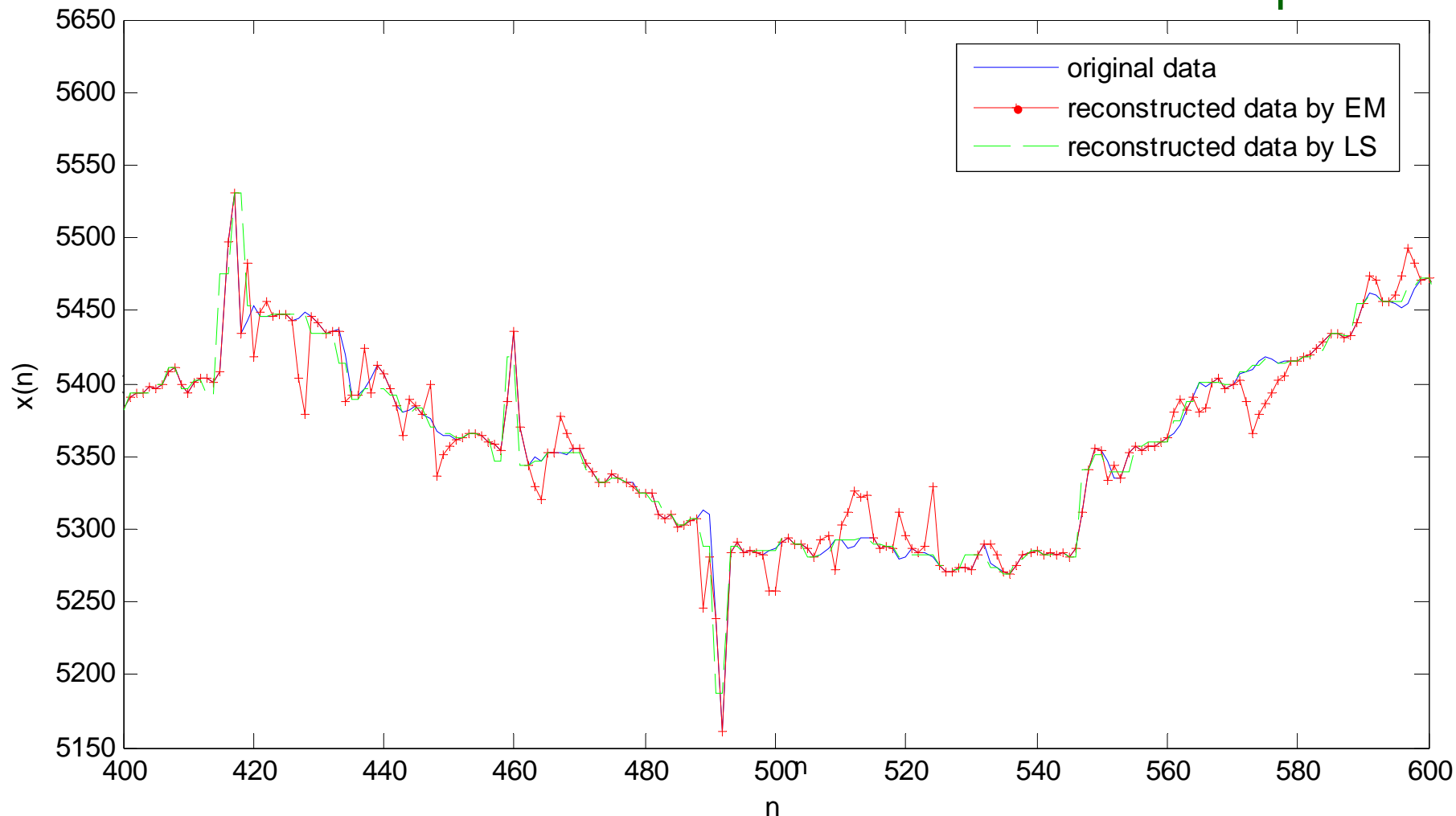
$$x \approx a_1$$

- Computational burden is much less
- No initial choice is needed
- Reliable for randomly missing data, but not for gapped data

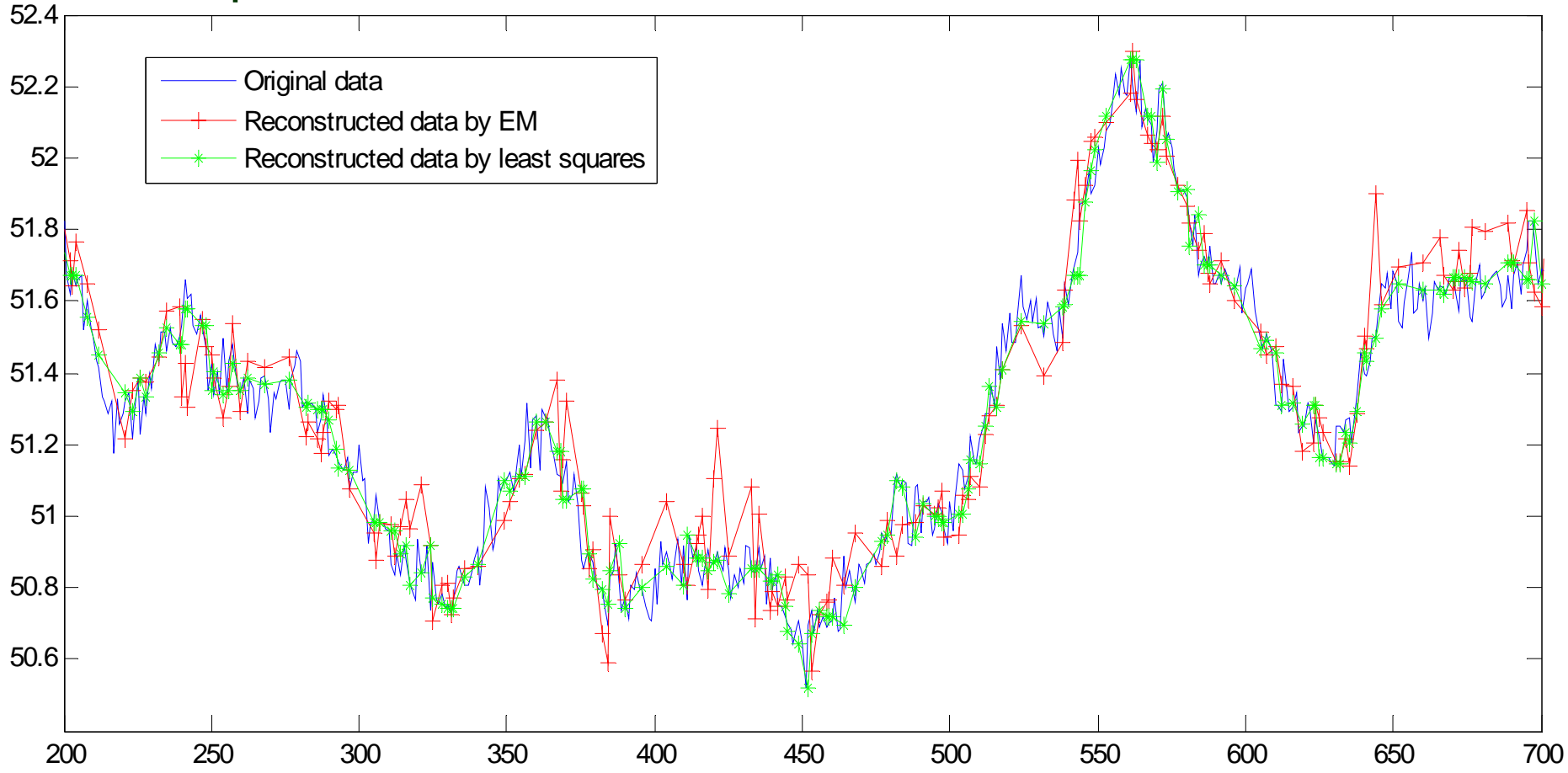
- Example 1: Scaled data from AT Plastics: 40% data were removed randomly and can be considered as missing data



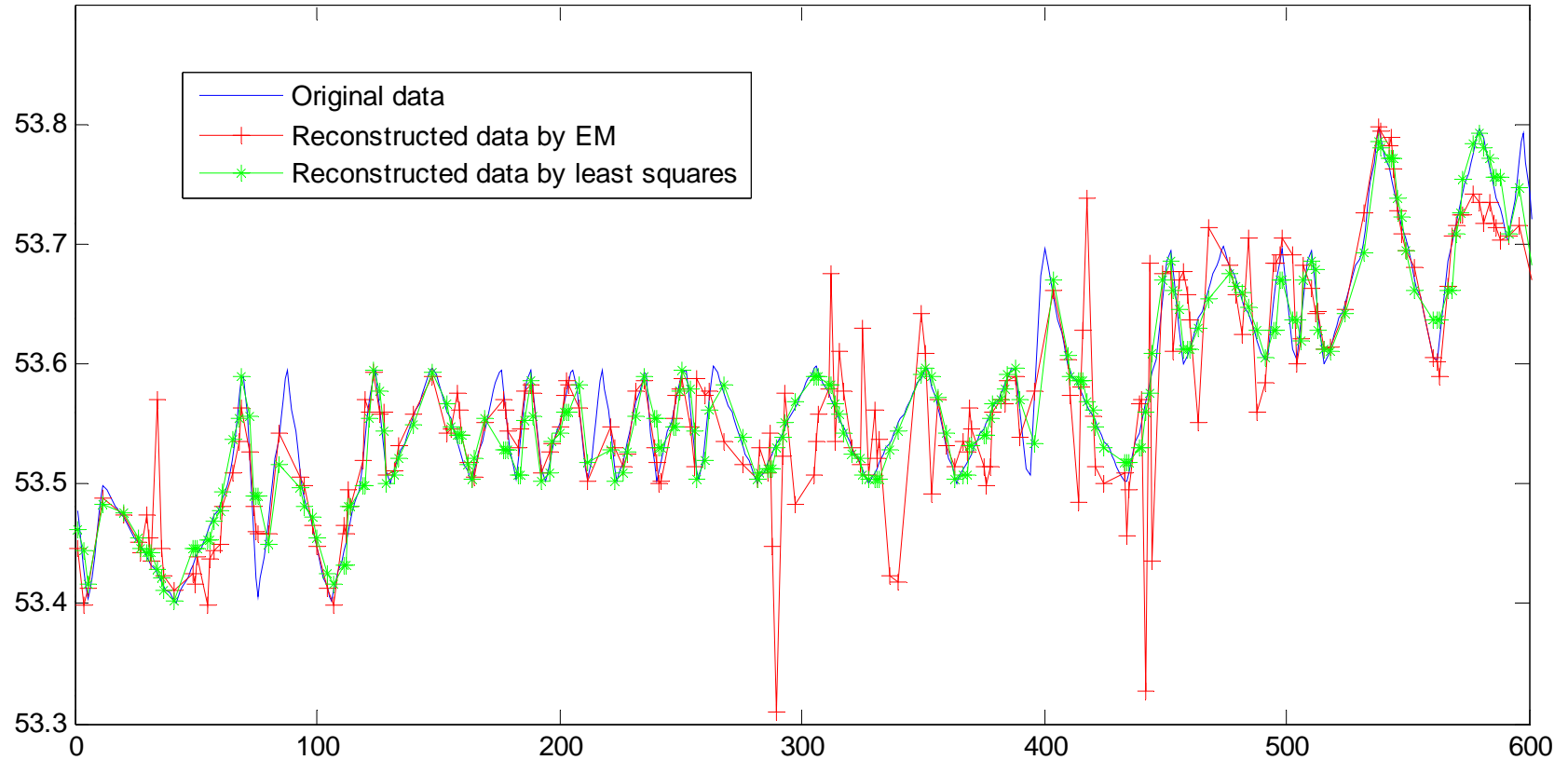
Simulation results for the first 1000 samples



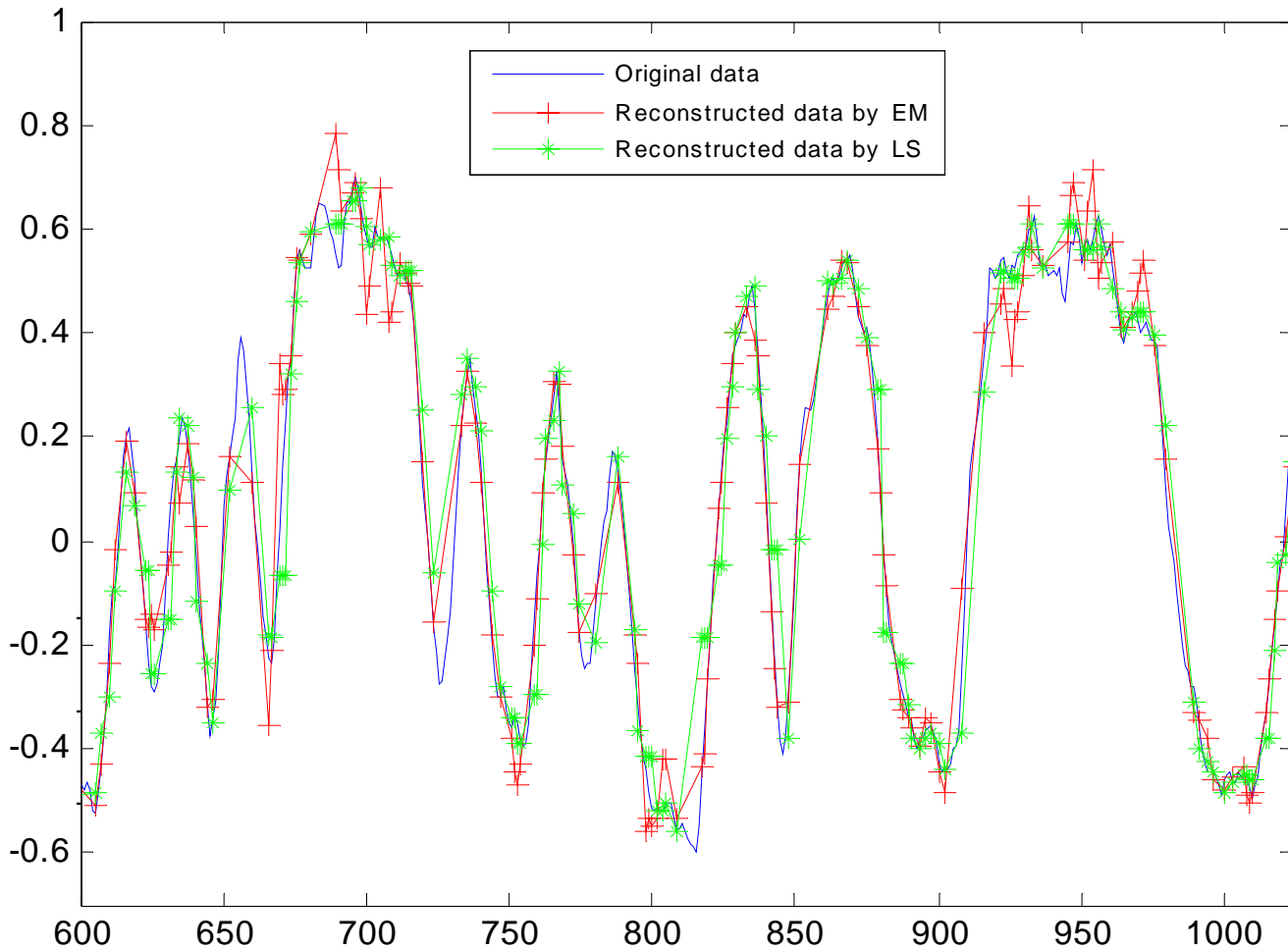
■ Example 2: Scaled industrial data from Suncor



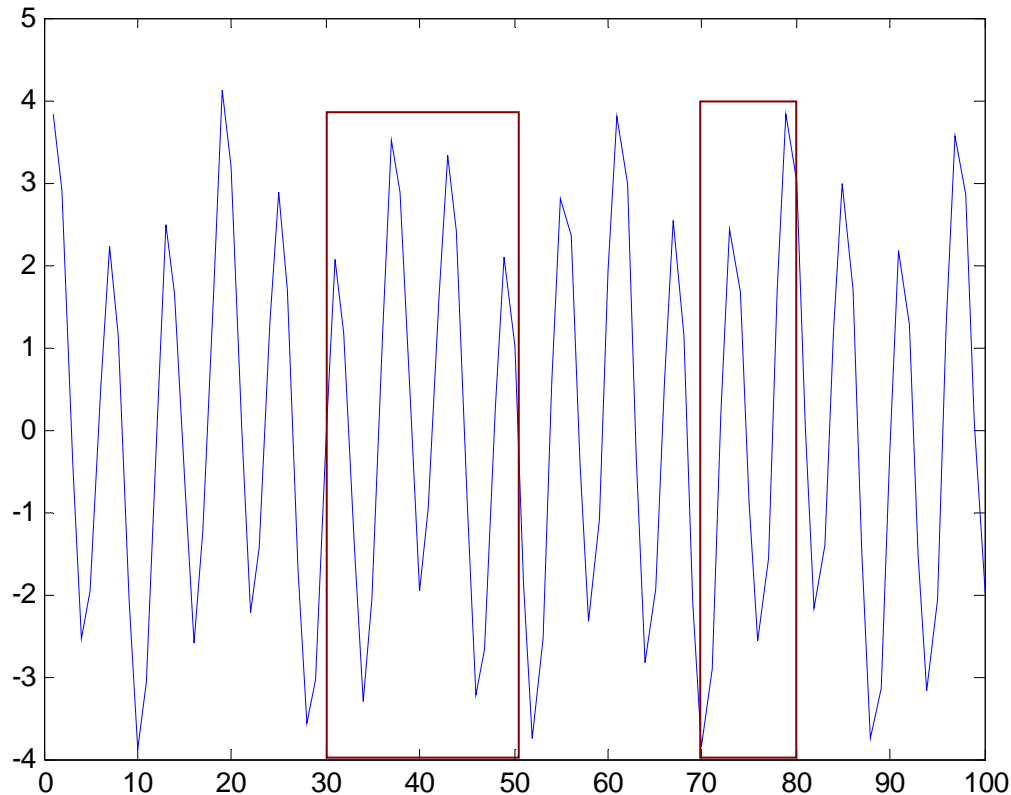
- Example 3: Scaled industrial data from Matrikon

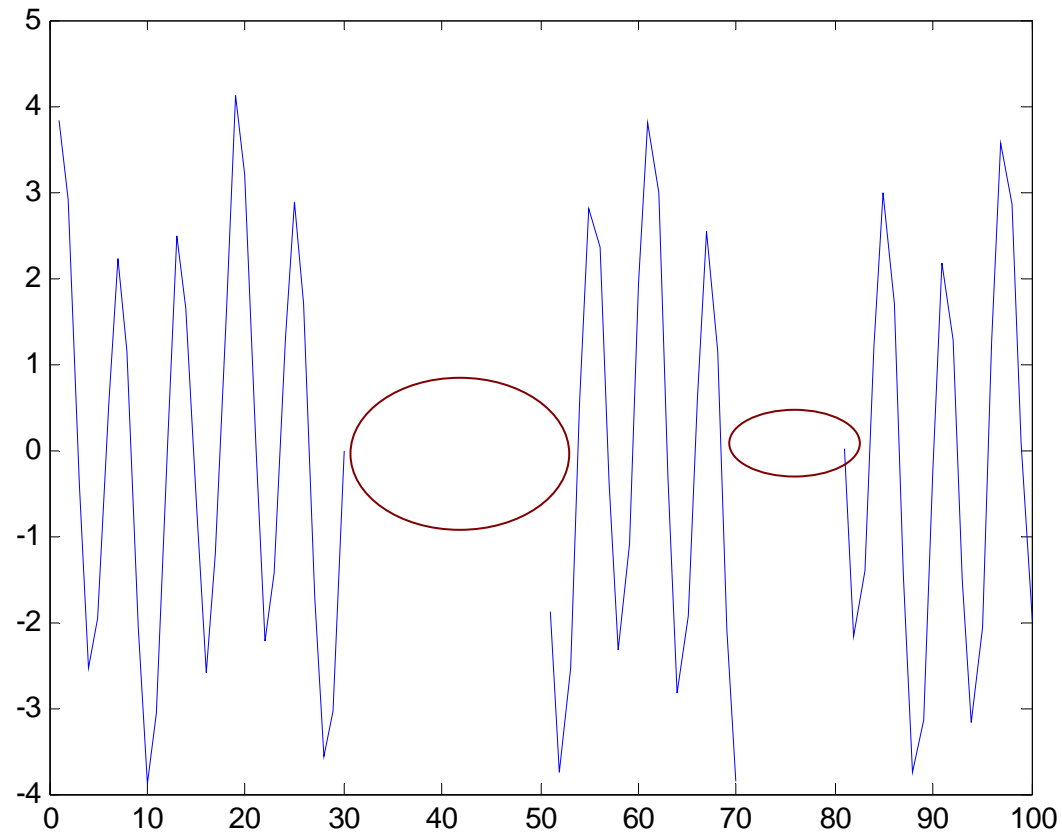


- Example 4: Experimental data from a dryer

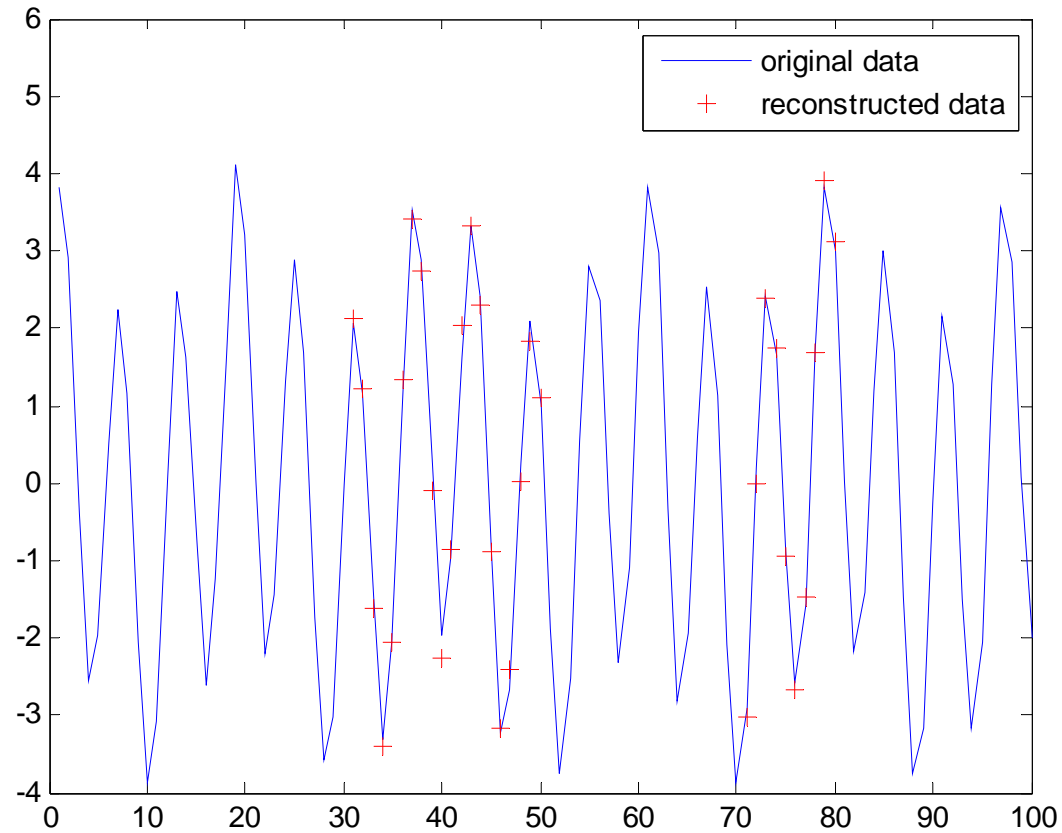


- Example 5: Simulated data: sum of two sinusoidal signals with white noise – remove the samples of 31-50 and 71-80 as gapped data



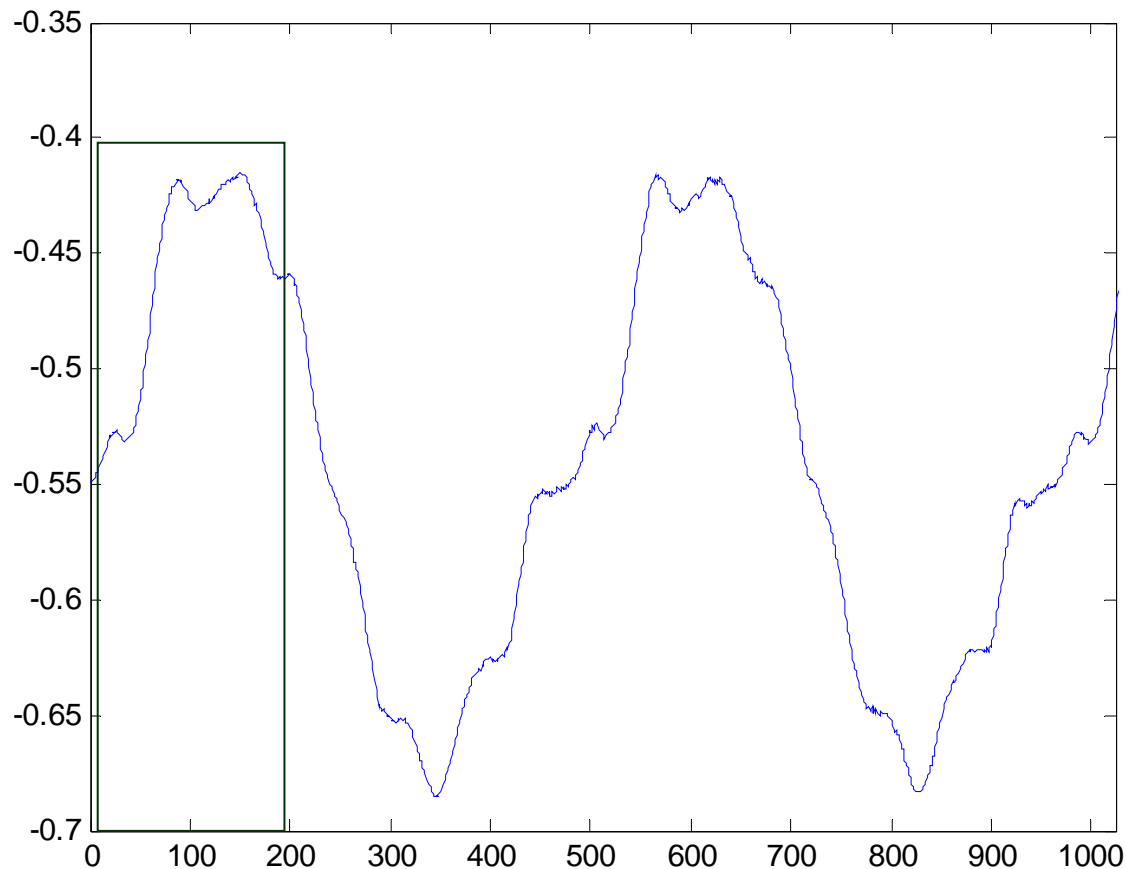


Examples to illustrate reconstruction of missing data

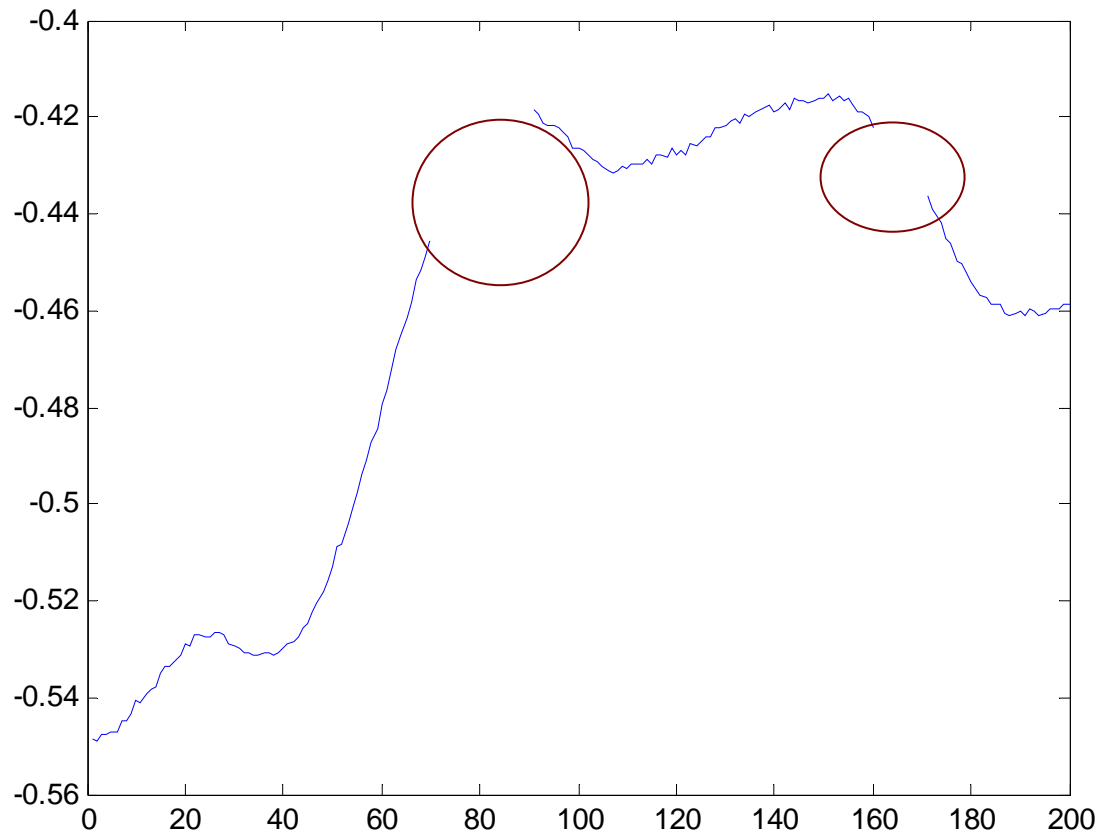


Examples to illustrate reconstruction of missing data

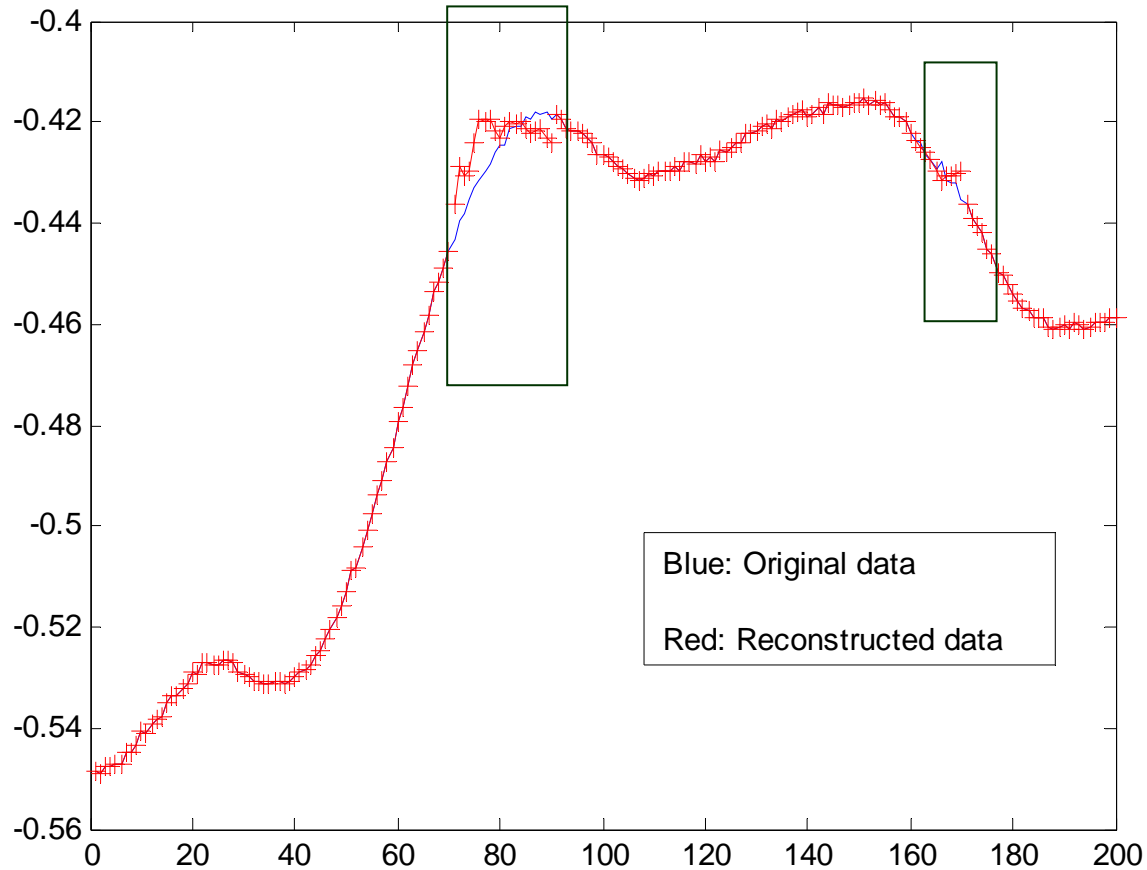
- Example 6: Experimental data from a pilot plant – remove the samples of 71-90 and 161-170 as gapped data



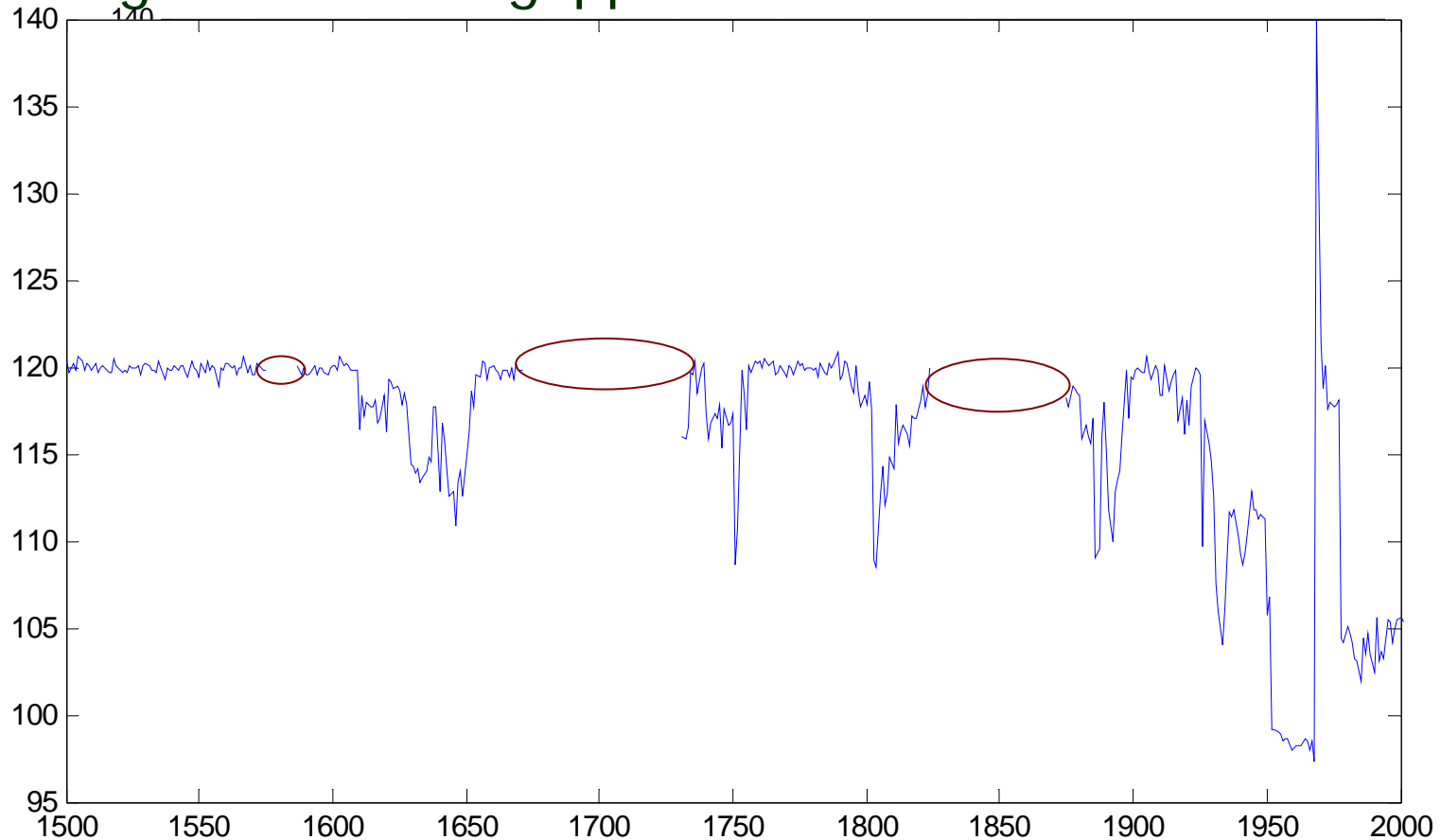
Examples to illustrate reconstruction of missing data²⁶



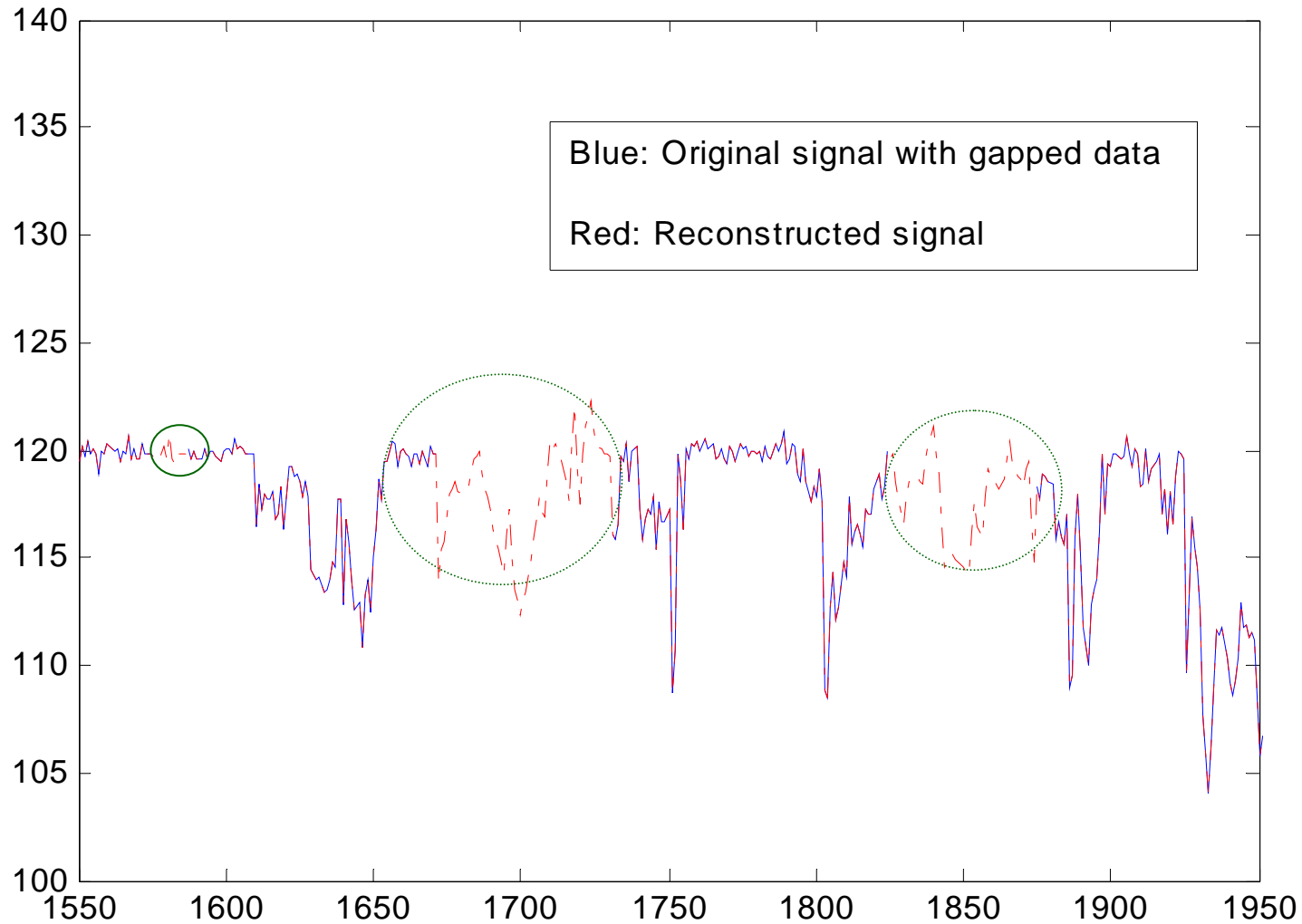
Examples to illustrate reconstruction of missing data²⁷



- Example 7: Scaled industrial data from Matrikon– notice large sections of gapped data

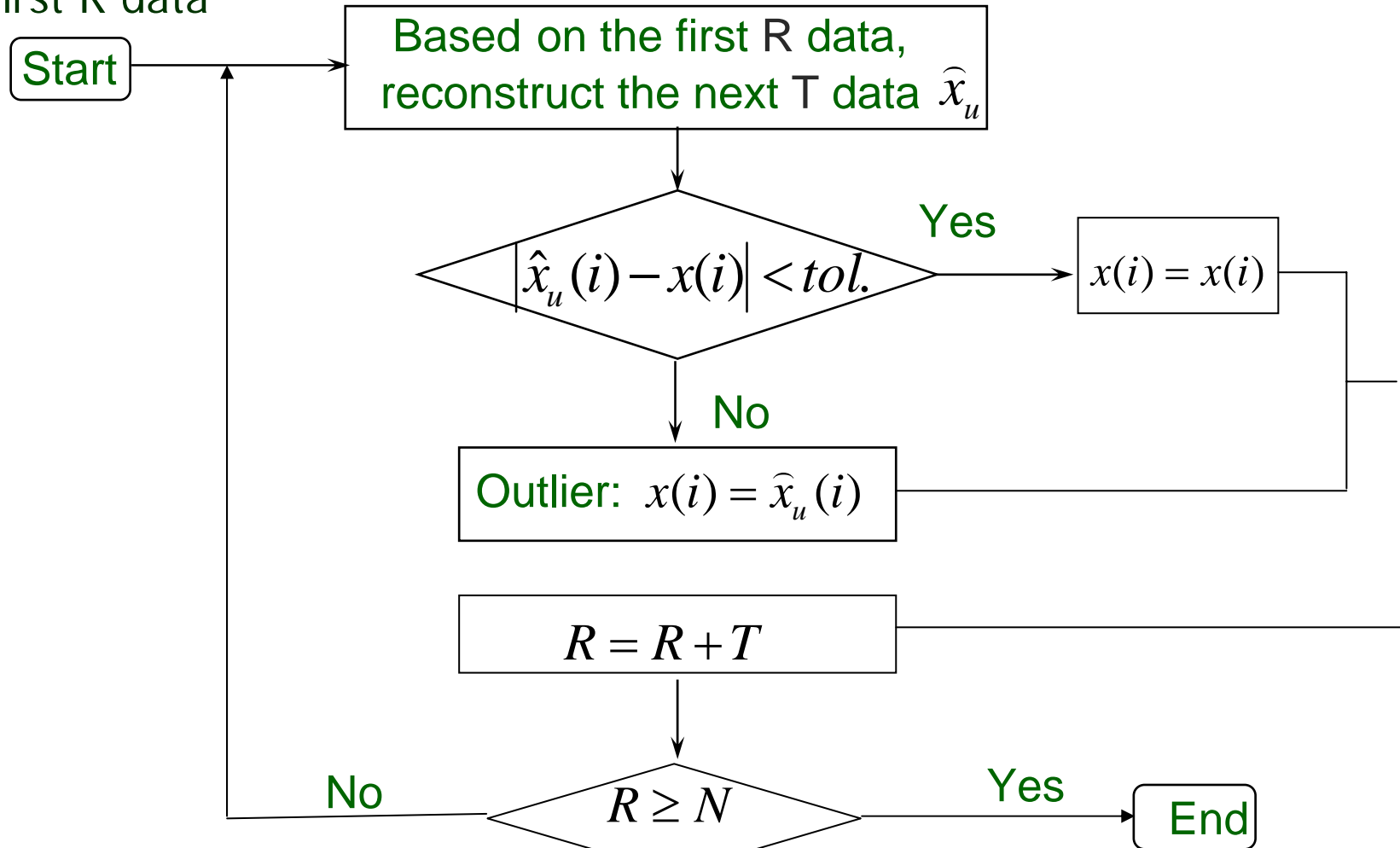


Examples to illustrate reconstruction of missing data²⁹



Application to detect and reconstruct outliers

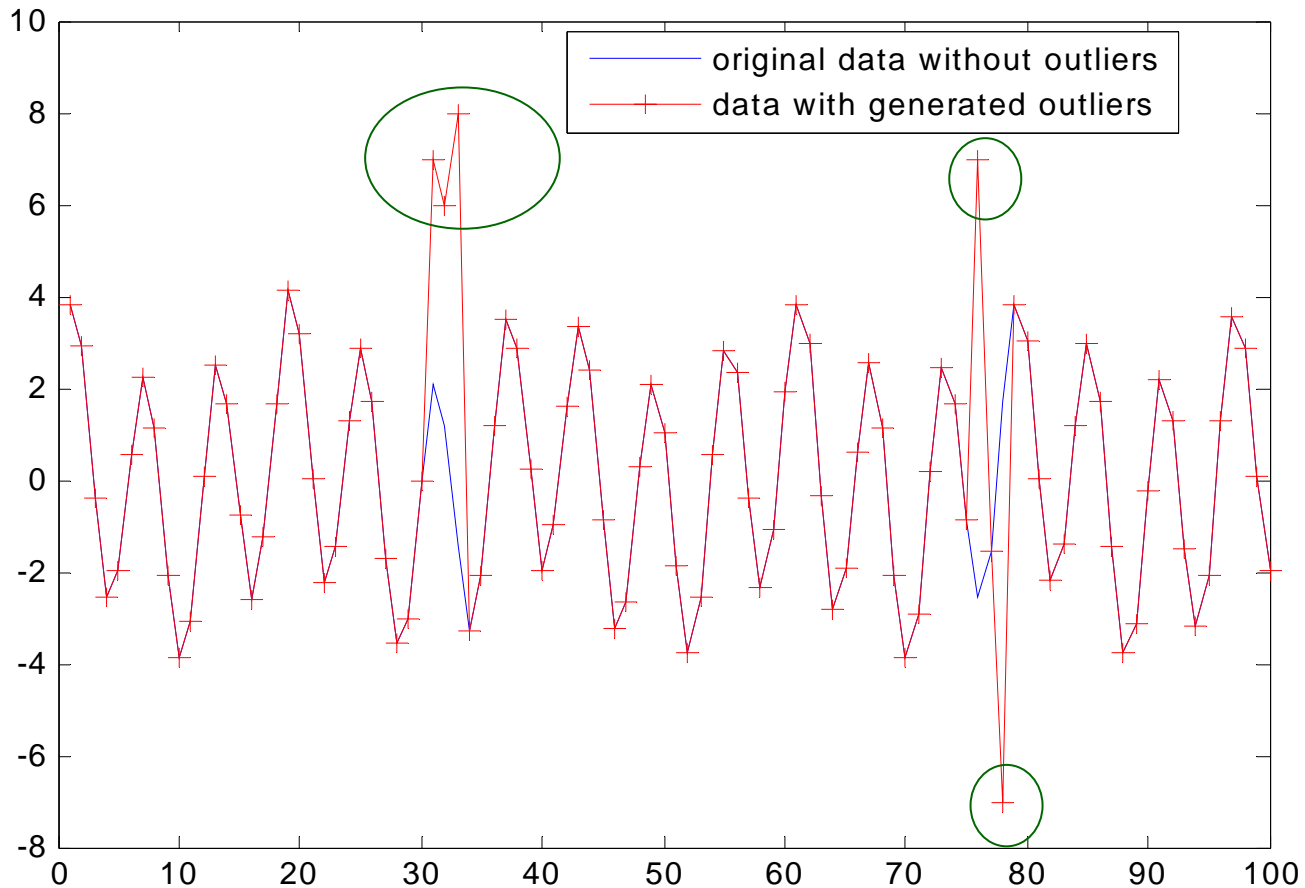
- Choose window size T and suppose that there are no outliers in the first R data



Application to detect and reconstruct outliers

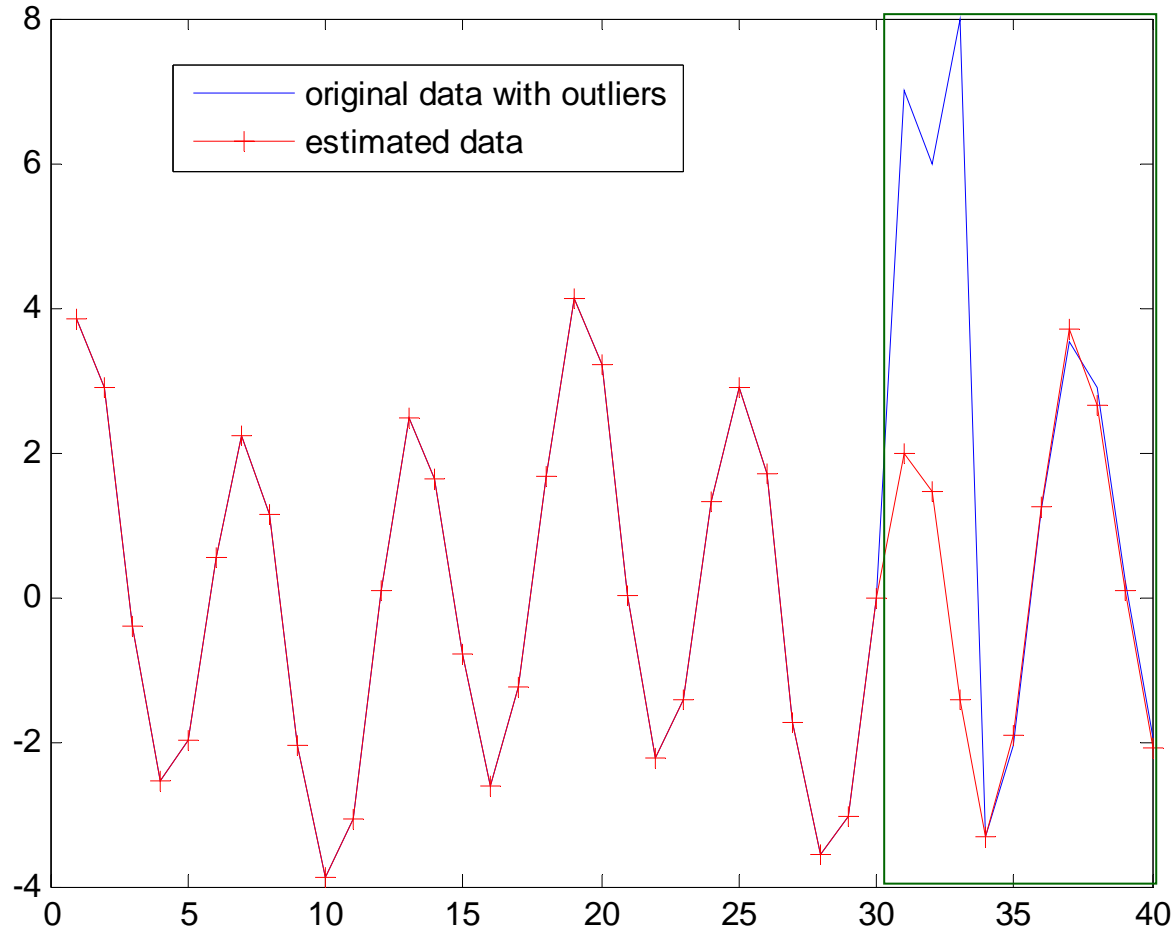
31

- Example 1: generate outliers at 31-33,76,78



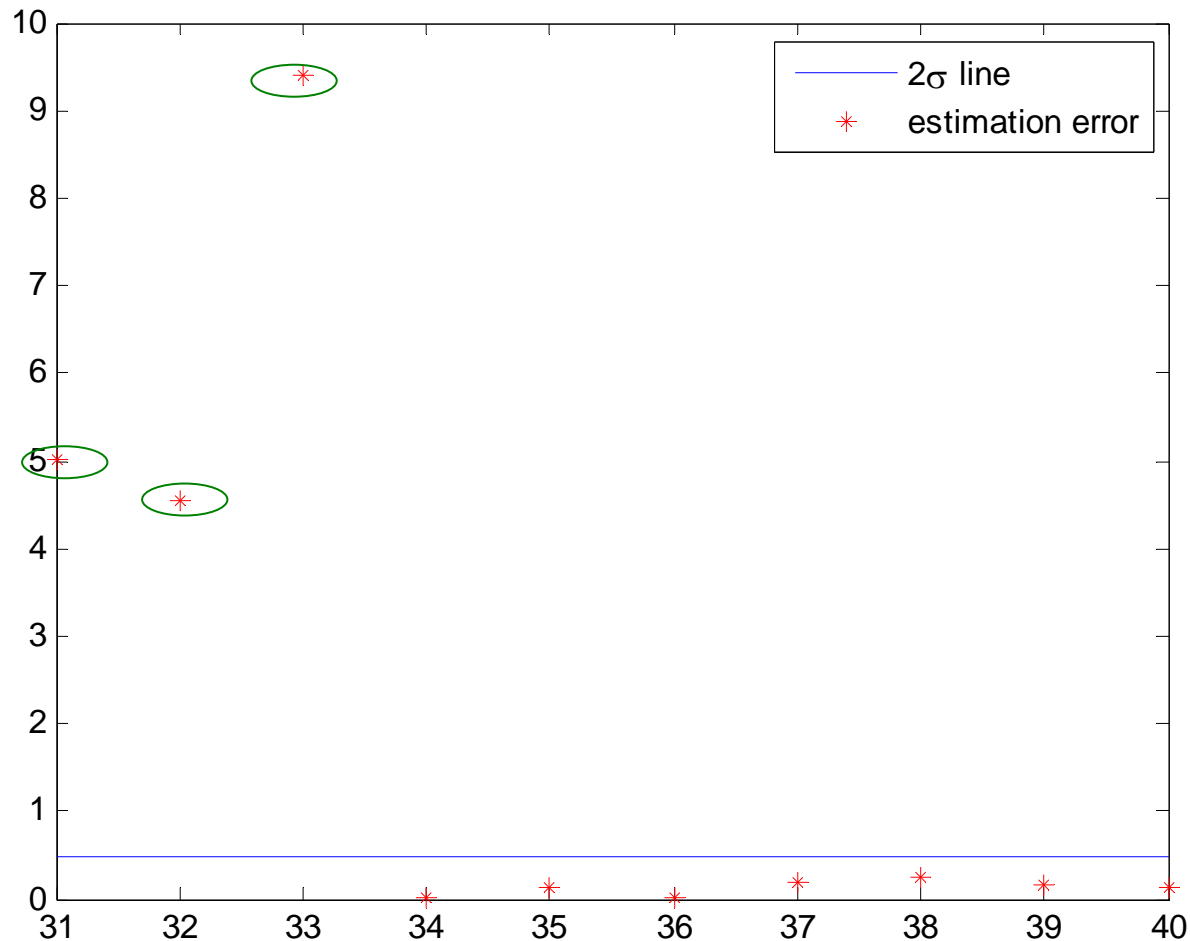
Application to detect and reconstruct outliers 32

- Suppose there is no outlier in the first 30 points, estimate the data over the interval 31-40



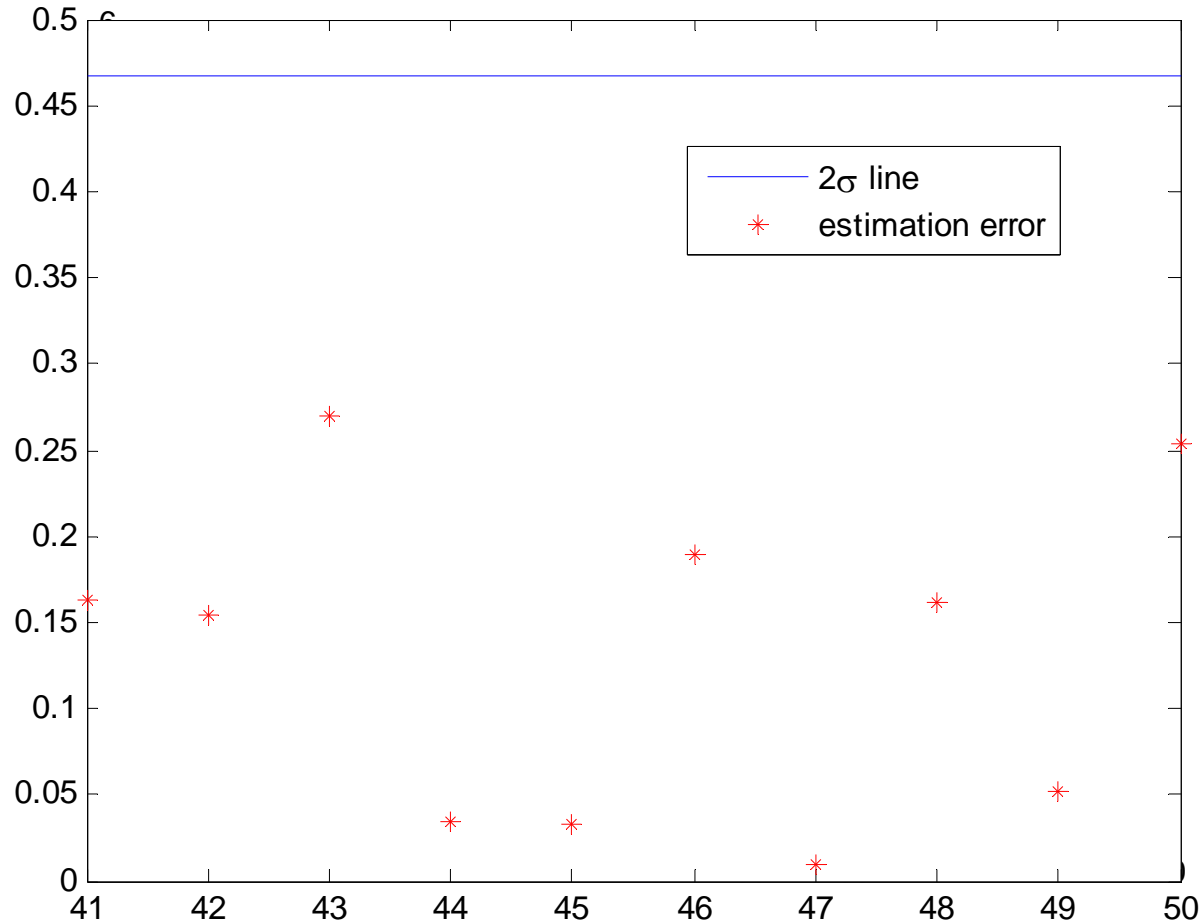
Application to detect and reconstruct outliers ³³

- Detect outliers over 31-40: data at 31-33 are outliers

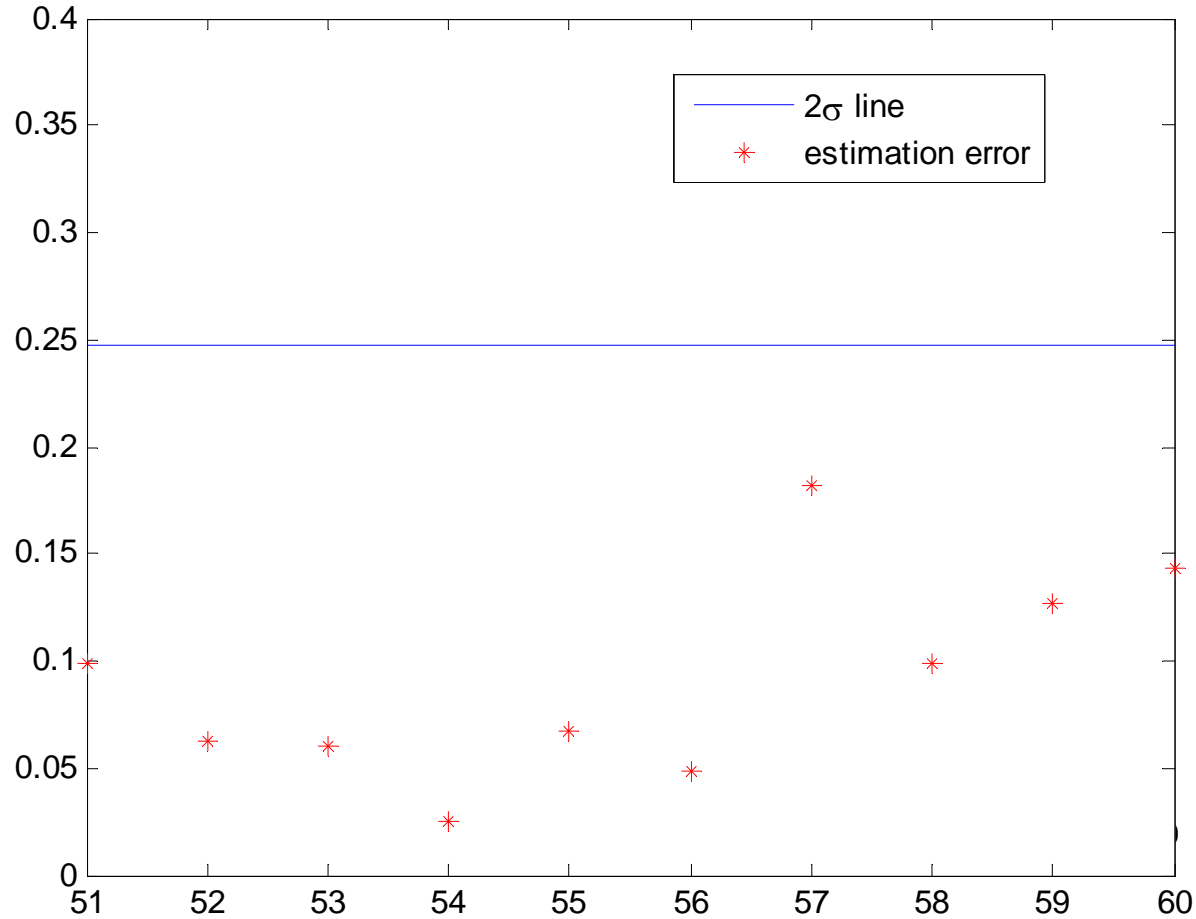


Application to detect and reconstruct outliers ³⁴

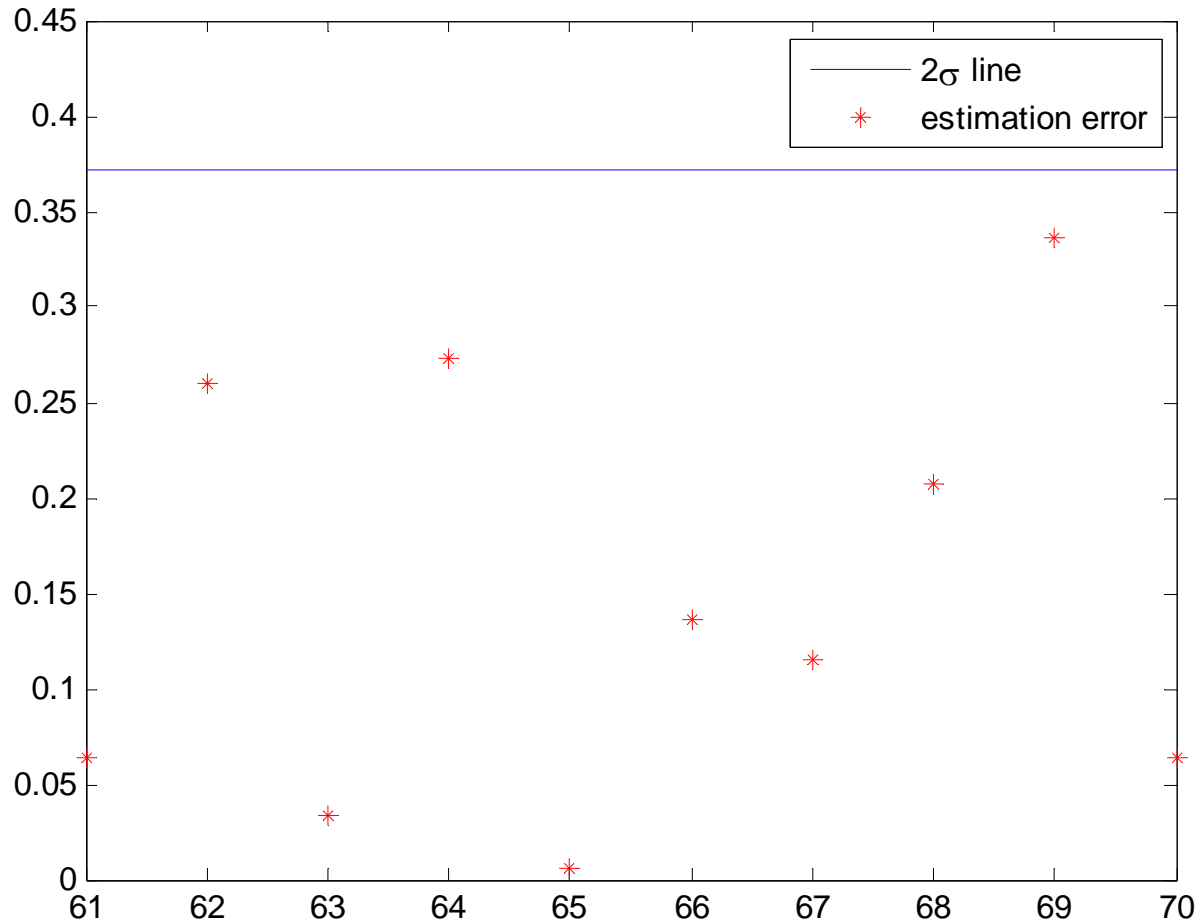
- Replace outliers with estimated values, based on the first 40 samples, estimate samples over interval 41-50



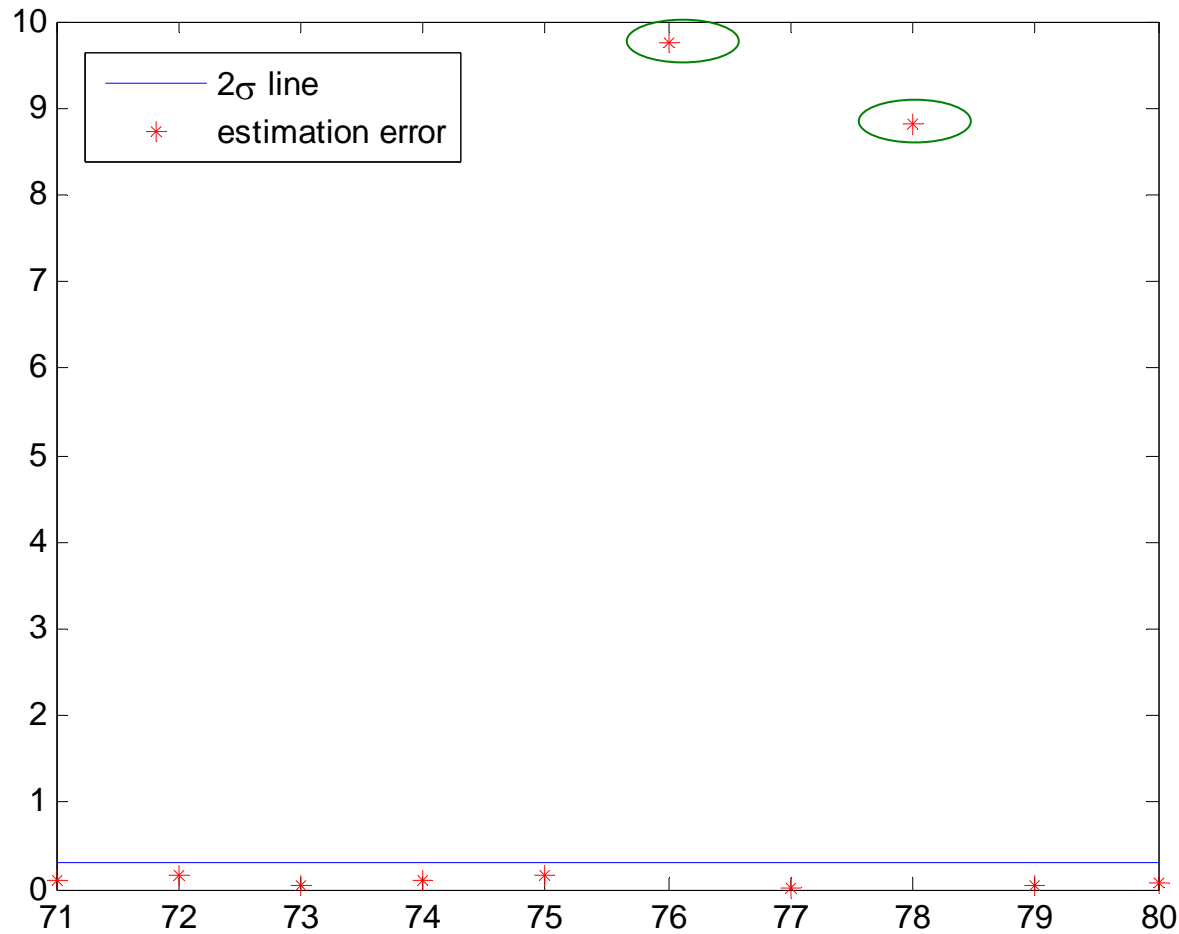
Application to detect and reconstruct outliers



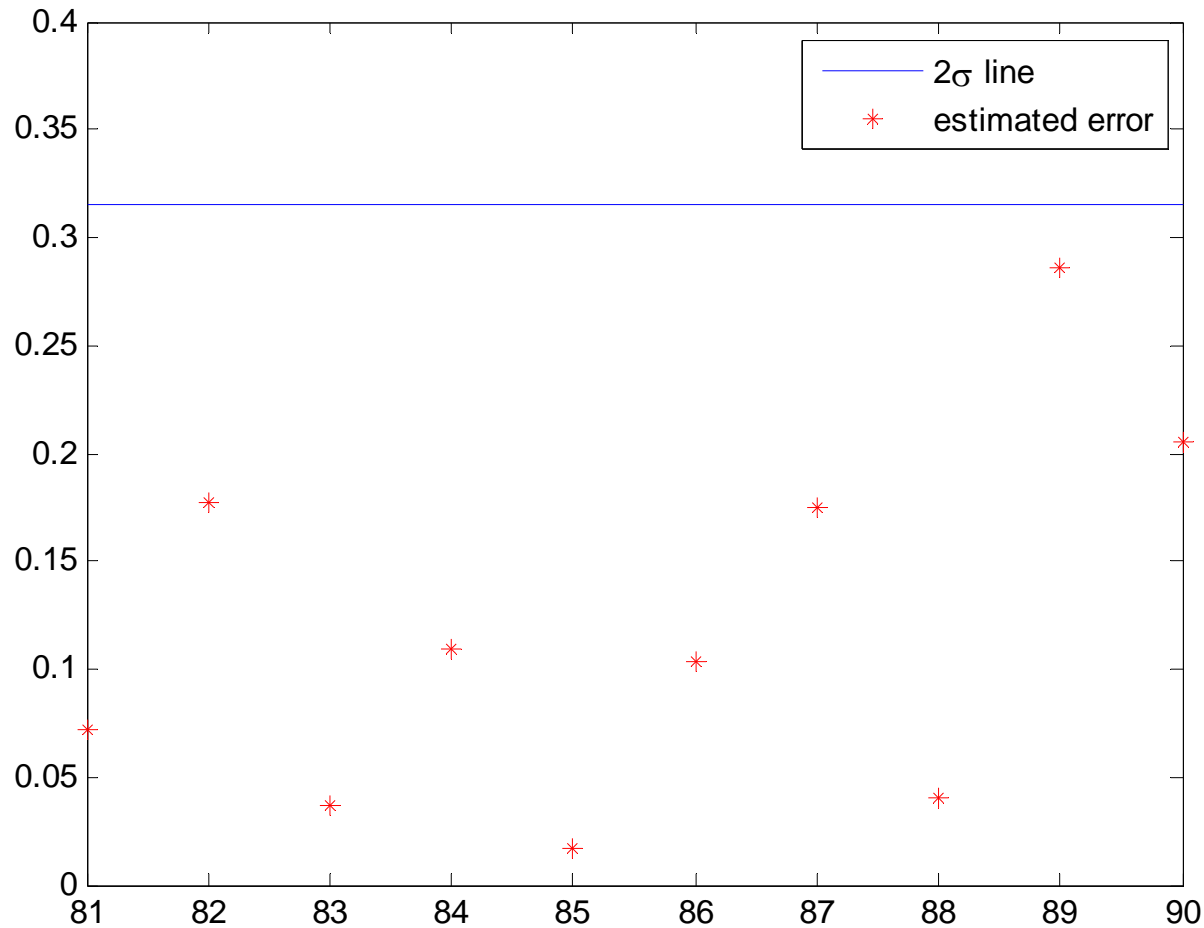
Application to detect and reconstruct outliers ³⁶



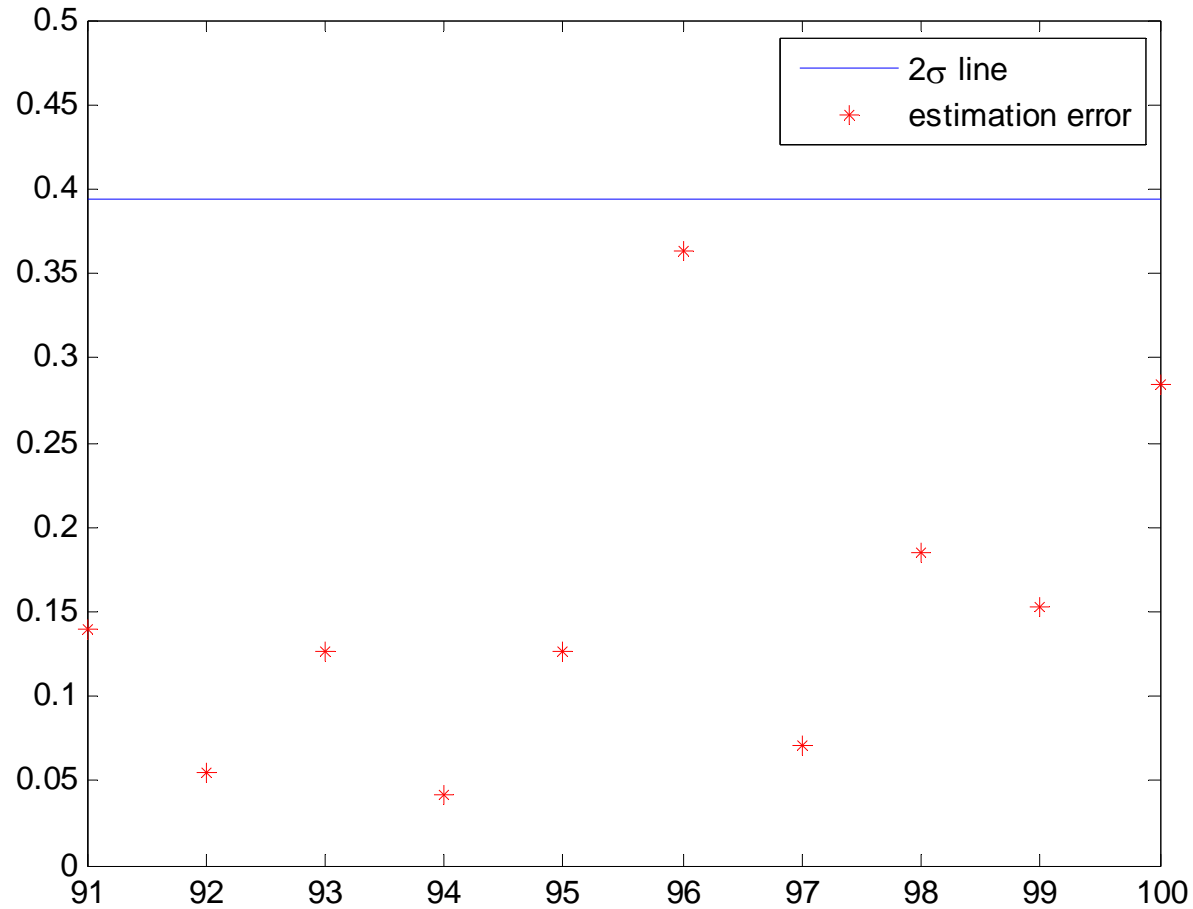
Application to detect and reconstruct outliers 37



Application to detect and reconstruct outliers ³⁸

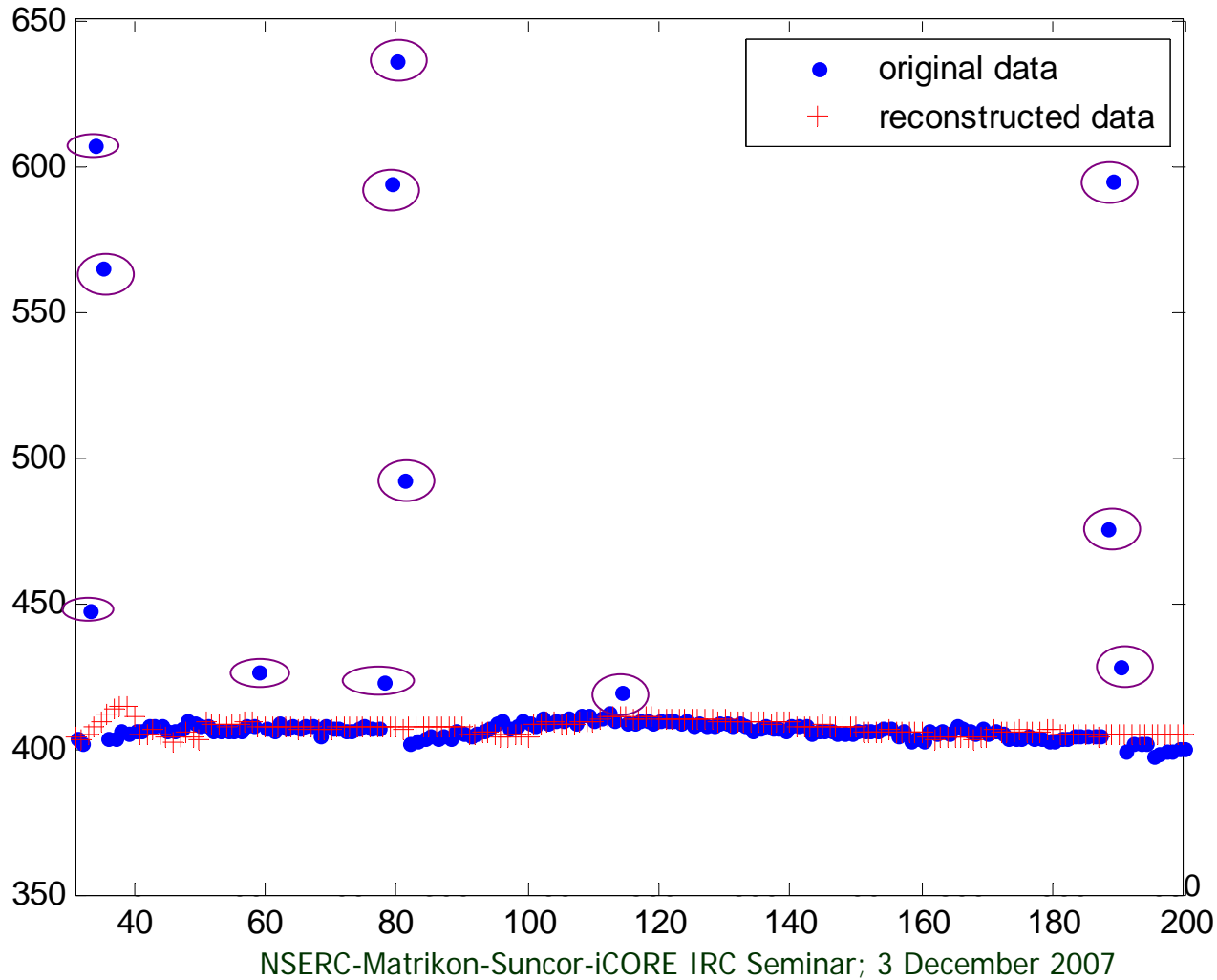


Application to detect and reconstruct outliers ³⁹



Application to detect and reconstruct outliers⁴⁰

■ Example 2: Industrial data V14 tag from Syncrude



Concluding Remarks

1. Two methods of reconstructing missing data based on discrete wavelet transform have been introduced.
2. The two methods have been compared by application to industrial data sets:
 - 1) The LS-based algorithm is simple and reliable if the data are missing randomly.
 - 2) The EM-algorithm is more reliable for gapped data.
3. The proposed method has also been successfully applied for detecting and replacing outliers.

Acknowledgements

- Dr. Sirish L. Shah and Dr. Tongwen Chen
- CPC Group Members
- NSERC-Matrikon-Suncor-iCORE for financial support