Journal of the Korean Geo-Environmental Society
19(1): 5～14. (January 2018) http://www.kges.or.kr
ISSN 1598-0820
DOI https://doi.org/10.14481/jkges.2018.19.1.5

다수 계측 데이터에 대한 복합 이상치 평가 및 검증

Compound Outlier Assessment and Verification for Multiple Field Monitoring Data

전 제 성†
Jesung Jeon

Received: October 26th, 2017; Revised: October 30th, 2017; Accepted: December 1st, 2017

ABSTRACT : All kinds of monitoring data in construction site could have outlier created from diverse cause. In this study generation technique of synthesis value, its regression, final outlier detection and assessment are conducted to distinct outlier data included in extensive time series dataset. Synthesis value having weight factor of correlation between a number of datasets consist of many monitoring data enable to detect outlier by increasing its correlation. Standard artificial dataset in which intentional outliers are inserted has been used for assessment of synthesis value technique. These results showed increase of detection accuracy for outlier and general tendency in case of having different time series models in common. Accuracy of outlier detection increased in case of using more dataset and showing similar time series pattern.

Keywords : Monitoring, Time series data, Error data, Detection of error data, Synthesis value, Regression

요 지 : 건설 현장에서 생산되는 각종 계측 데이터 내에는 다양한 원인에서 생성된 각종 이상 데이터가 포함되어 있다. 본 연구에서는 시계열 데이터 내에 포함된 이상 데이터의 효과적 판정을 위한 합성신호 생성 기법과 그를 이용한 회귀분석, 최종적인 이상 데이터 판단과 평가 등에 관한 연구를 수행하였다. 방대한 데이터로 구성된 다수 데이터셋에 대한 이상 데이터 평가 시 다수의 데이터셋 간의 상관성을 가중치로 한 합성신호는 특정 데이터셋 과의 상관성을 크게 향상 시키는 효과를 보였으며, 이를 통해 효과적인 이상 데이터 판정이 가능하였다. 인위적 이상 데이터가 포함된 인공 오류 데이터를 생성하고 이에 합성신호 기법을 적용한 결과, 이상 데이터 판정 정확도가 크게 증가 하였으며 이러한 결과는 이종 시계열 모델의 경우에서도 동일하게 확인되었다. 이상 데이터 판정의 정확도는 신호 합성에 이용되는 데이터셋 수가 많고 시계열 모델 특성이 유사할수록 크게 증가하였다.

주요어 : 계측, 시계열 데이터, 이상 데이터, 복합 이상 데이터 판정, 신호 합성, 회귀분석

1. 서 론

최근 건설 분야의 시공관리 혹은 유지관리 차원에서 수행되는 각종 구조물 거동 파악 및 이를 이용한 안정성 평가는 가장 중요하고 기본적인 건설 업무로 인식되고 있다. 즉, 모든 건설 단계에서 구조물 거동 감시를 통한 각종 시공관리 및 안전관리가 이루어지고 있으며, 이러한 구조물 거동 감시에 있어 가장 널리 이용되는 방법은 현장 구조물 계측일 것이다. 최근에는 IT기술의 발전과 함께 현장 센싱 및 데이터 전송 등에 대한 기술발전에 의해 건설 분야에서도 다양한 자동계측 기술이 적용되고 있으며 급속도로 그 적용규모와 범위가 확대되고 있다. 자동계측 기술의 확대는 계측 주기를 시간 단위 이내로 단축시켜 다량의 계측 데이터를 양산하고 있으며, 이러한 방대한 양의 계측 데이트는 궁극적으로 거동 감시 및 관리의 가장 기본적이며 핵심적인 자료로 활용되고 있다. 다양한 계측 센서들을 통해 측정된 빅데이터 범주의 계측 데이터들에는 여러 가지 전기적, 물리적 또는 기계적 이유에서 양산되는 다양한 형태의 이상 데이터가 포함되어 있다. 이러한 이상 데이터를 구조물의 거동 예측 및 안전관리 등에 그대로 이용한다면 그 최종 결과의 신뢰도에 큰 영향을 미치게 된다. 즉, 각종 계측 데이터는 적정한 방법의 분석 과정을 통해 다양한 이상 데이터가 제거된 상태에서 구조물의 거동감시 및 안전관리 등에 이용되는 것이 필수적이다(Williams et al., 2002). 과거 건설 분야에서의 일반적인 계측 데이터 분석은 시계열 데이터에 대한 경시변화 이상 패턴을 일단 변동, 다단 변동, 단기 결측, 영구 결측, 단기 급등, 상시 미동 등 제한된 형태로 정의하고, 실제 데이터의 시계열 패턴과 그 형태를 비교하는 과정을 통해 수행

† Department of Construction Information Engineering, Induk University (Corresponding Author : jsjeon@induk.ac.kr)

하였다. 계측 데이터의 경시변화 형태에는 다양한 외부환경 및 기계적, 전기적 오차가 포함되어 있으며 이러한 요인에서 파생되는 이상 데이터는 경시변화의 왜곡 및 판정 결과의 오류를 발생시킬 수 있다. 또한 경시변화 형태에 대한 육안 판정은 판단자의 주관과 기술적 능력에 따라 큰 결과 차이를 보이게 되며, 이러한 고전적 방식으로서는 계측 데이터에서 적정 이상치를 신속하고 정확하게 제거하는 것이 불가능하다. 과거 국내에서 수행된 구조물 안전도 평가를 위한 데이터 마이닝 관련 연구는 계측정보를 바탕으로 한 회귀분석 방법의 적용, 인공신경망 기법의 적용, 해석변수의 불확실성을 고려한 신뢰도 평가 등 추후 거동예측에 관한 내용이 대부분 이었으며, 입력정보 즉 계측결과 자체에 대한 신뢰성 평가 및 분석에 관한 연구는 미비한 실정이었다(Jeon et al., 2015a).

본 연구에서는 건설 분야 구조물에서 활용되는 계측 데이터들의 오류 데이터 선별을 목적으로 다수 데이터셋 대상의 다중 분석방법을 혼합하는 방식의 이상치 평가와 보정 기법에 관한 연구를 수행하였다.

2. 기존 이상치 평가 기법

국외에서는 과거부터 센서 네트워크 분야를 중심으로 각종 원인에서 야기된 다양한 이상 데이터의 형태를 구분(Ramanathan et al., 2006; Ni et al., 2009; Sharma et al., 2010) 하고 다양한 이상 데이터 형태별 이상치들의 평가 기법에 관한 연구가 진행되었다.

가장 쉽게 널리 이용되는 이동 평균법의 경우, 데이터의 시간 간격에 따라 예측치와 실제 데이터 변화 사이에 상당한 시간차가 있을 수 있으며, 데이터의 변화폭이 큰 경우는 그 적용에 문제가 있을 수 있다(Mourad & Bertrand-Krajewski, 2002).

Ramanathan et al.(2006)은 데이터셋 내에서 각 데이터들 간 경시 변화량 및 변화율을 활용하여 이상치를 평가하였으며, Ni et al.(2009)은 이들 분포 값에 대한 평균, 분산, 변화도 및 특정 회귀분석 모델을 이상치 평가에 활용하기도 하였다. 이러한 분석방법에서는 모두 단일 데이터셋을 대상으로 이상치 평가가 수행되므로 시계열 특성이 크게 나타나거나 정상 데이터의 분포비율이 매우 높은 경우가 아니라면 그 결과의 신뢰도가 급격히 낮아지는 단점이 있다. Jeffery et al.(2006)은 개별 데이터셋이 아닌 복수의 데이터셋을 활용하되 데이터 간의 공간적 상관성을 이용하여 오류 데이터 판정의 정확성을 높이는 연구를 수행하였다. Jeon at al.(2015b)는 특정 계측기에서 생성되는 단일 데이터셋 만을 대상으로 한 이상치 평가에는 한계가 있음을 보였으며, 복수 데이터셋 대상의 복합 분석이 최종적인 이상치 판정에 매우 효과적임을 제시하였다. 특정 구간별 데이터들의 상관성에 기초한 대표적 이상치 분석 방법은 Eq. (1)의 선형 최소자승법 회귀분석 모델(Kailath, 1975)을 이용한 방법이다

$$\widehat{u_1}(u_2) = m_{u_1} + \frac{\lambda_{u_1u_2}}{\lambda_{u_2}}(u_2 - m_{u_2}) \quad (1)$$

여기서, m_{u_1}, m_{u_2}는 데이터 셋 u_1, u_2의 평균값, $\lambda_{u_1u_2}$는 u_1, u_2의 공분산, λ_{u_2}는 u_2에 대한 분산, $\widehat{u_1}(u_2)$는 u_2 내의 특정 데이터를 이용하여 산정한 u_1 예측 값을 나타낸다. 그러나 이러한 방법은 2개의 데이터셋 만을 이용하는바, 다수의 데이터셋을 모두 이용하지 못하는 한계를 가지고 있다. 한편 Ni et al.(2009)은 센서 네트워크 분야의 이상치 판정에 적용되는 알고리즘 대부분이 모호한 오류 데이터 형태 및 모델을 기반으로 생성되었기에 다양한 데이터의 이상치 판정에 범용적으로 적용되는데 한계가 있음을 지적하였다.

3. 다수 데이터셋 대상의 복합 이상치 분석

3.1 상관성 기반의 합성신호 생성

건설 과정 중 또는 건설 후 유지관리 차원에서 수행되는 각종 계측은 그 항목과 측정 위치별로 매우 다양한 데이터를 양산한다. 동종의 계측 항목에 대해서도 다양한 위치에서의 다양한 데이터셋이 양산되는데, 특정 데이터셋의 이상치 분석에 있어서는 대상 데이터셋 내부의 데이터만을 이용하는 즉, 단일 데이터셋 만을 이용하는 분석방법과 동종 혹은 이종 데이터셋을 동시에 활용하는 다중 분석방법이 이용될 수 있다(Jeon, 2016). 특정 계측기에서 생성되는 데이터는 공간적 일정 구역 내 타 계측 데이터와 상호 연관성을 갖는 것이 일반적이다. 즉, 계측 데이터의 이상치를 판정하는 데 있어 특정 계측기에서 생성되는 단일 데이터셋 만을 분석하는 것 보다는 상호 연관성을 나타낼 수 있는 타 계측 데이터를 동시에 분석하는 과정을 통해 더욱 효과적인 이상치 판정이 가능하다(Jeon et al., 2015; Park & Jeon, 2015). 동일 독립변수(x)로부터 측정된 n개의 관측치가 Fig. 1에서와같이 서로 다른 선형모델을 갖는다고 가정하면 n개 센서에서 일정 기간 동안 수집된 관측치를 이용하여 새로운 합성 독립 신호를 생성할 수 있다. 이러한 합성 신호와 특정 센서 관측치를 대상으로 선형모델을 추정할 수 있는데, 이때 특정 센서 관측치의 노이즈 e_n 영향은 피할 수 없지만,

타 센서 관측치의 노이즈들이 상호 상쇄됨으로 인해 단일 관측치를 이용하는 경우보다 노이즈 영향을 줄일 수 있게 되며 이로 인해 특정 센서 관측치와 합성 신호 간 상관성은 더욱 높아지게 된다.

실제 건설현장의 계측 상황을 고려할 때, 특정 계측 값에 영향을 미치는 독립변수는 동일한 한 종류가 아닌 수위, 지반조건, 구조물 특성, 계측 위치 등 여러 가지의 다양한 종류가 있는 것이 일반적이다.

Fig. 2와 같이 k개의 물리적인 독립변수로부터 측정된 n개의 센서 관측치가 서로 다른 선형모델을 갖는다고 가정할 때, 센서 관측치 중에는 y_3 처럼 여러 개 독립변수의 영향을 받는 센서 관측치들이 있을 수 있다. 이와 같은 경우, 대상 센서 관측치와 나머지 관측치들과의 상관계수를 구하면, 상관성이 없는 관측치들, 즉 이종 독립변수의 영향을 받은 관측치들 간의 상관계수는 0에 가깝게 되고, 상관계수를 이용한 가중합을 구하면 상관성이 없는 관측치 들은 그 영향이 크게 감소되어 결과적으로 앞서 언급한 y_n과 합성신호 y'와의 선형성이 좋아지게 된다.

즉, 각각의 센서 관측치(데이터셋)가 임의 개수의 물리적 독립변수를 포함한 서로 다른 선형모델을 갖는다고 가정할 때, 각 센서 관측치에는 각기 서로 다른 독립변수의 영향을 받는 다양한 특성이 나타날 수 있다. 특정 독립변수에 대해 일부 그룹의 센서 관측치 들은 큰 상관성을 보이는 반면 다른 센서 관측치들은 낮은 상관성을 보일 수 있다. 이와 같은 경우, 대상 센서 관측치와 나머지 관측치들과의 상관계수를 구하면 상관성이 없는 관측치들, 즉 이종 독립변수 관계를 갖는 센서 관측치 들의 상호 상관계수는 0에 가깝게 된다. 특정 센서 관측치에 대하여, 나머지 센서 관측치들의 합을 구하되 각 관측치들 간 상관계수를 가중치로 한 전체 센서 관측치들의 가중합을 구하면 상관성이 없는 관측치들은 그

$$
\left.\begin{aligned}
y_1 &= a_1x + b_1 + e_1\\
y_2 &= a_2x + b_2 + e_2\\
y_3 &= a_3x + b_3 + e_3\\
&\cdots\cdots\\
y_n &= a_nx + b_n + e_n
\end{aligned}\right\}
\quad y' = (\sum_{i=1}^{n-1} a_i)x + \sum_{i=1}^{n-1} b_i + \sum_{i=1}^{n-1} e_i,\ \text{noise}\ \ e_i \sim N(0,\sigma^2)
$$

$$
y' = a'x + b' \quad (\sum_{i=1}^{n-1} e_i \to 0)
$$

$$
y_n = a_nx + b_n + e_n \quad \longleftrightarrow \quad y_n = \frac{a_n}{a'}y' - \frac{a_nb'}{a'} + b_n + e_n
$$

Fig. 1. Synthesis of monitoring value having same independent variable

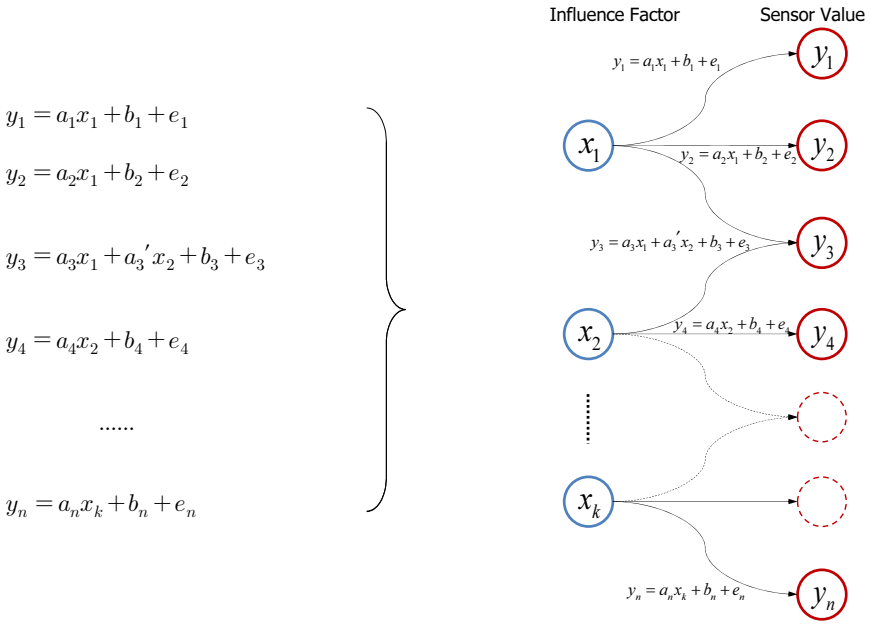

Fig. 2. Synthesis of monitoring value having different kind of independent variables

영향이 크게 감소되어, 특정 센서의 관측치와 합성신호의 상관성이 크게 증가하게 되고 이들 간의 회귀모델을 통해 최종적인 이상치 분석이 가능하게 된다.

3.2 합성신호를 이용한 이상치 평가

본 연구에서는 이와 같은 상관계수 가중치를 도입한 합성신호 기반의 이상치 분석 알고리즘(ASCS)을 개발하였다. 특정 데이터 블록의 크기를 규정한 후, 블록 내 데이터들을 대상으로 합성신호 생성 및 이를 이용한 회귀모델 구성, 회귀모델 검정 및 추정에 의한 이상치 분석 등을 수행하였다(Fig. 3). 이러한 일련의 이상치 분석과정은 전체 데이터 대상의 이동 블록마다 동일한 방식으로 적용되었다.

모든 블록은 동일한 개수의 데이터를 포함하고 있으며, 블록마다 이상치 분석용 합성신호를 생성하게 된다. 특정 데이터셋을 분석 데이터셋으로 설정한 후, 다른 데이터셋은 이용 데이터셋으로 활용하게 된다. 블록 내 데이터를 대상으로 분석 데이터셋과 각 이용 데이터셋과의 상관계수를 구하고 이를 가중치로 한 가중합을 Eq. (2)와같이 최종적인 합성신호로 산정하게 된다.

$$sv_j = \sum_{i=1}^{dsn} (v_{i,j} \times cor_{i,tar}) \tag{2}$$

여기서, sv_j는 특정 블록 내 특정일자(j) 데이터에 대한 합성신호 값(블록 내 데이터=1~n), $v_{i,j}$는 특정 데이터셋(i=1~dsn) 내의 특정 일자(j) 데이터, $cor_{i,tar}$는 특정 블록 내 분석 데이터셋과 타 이용 데이터셋(i)과의 상관계수를 나타낸다.

특정 블록별로 해당 블록 내에서 생성된 합성신호와 분석 데이터셋을 대상으로 회귀모델을 구성하고 이 모델을 통해 특정 시간에서의 예측치를 산정한다.

회귀모델의 유의성 판단은 Table 1과 같은 분산 분석을 통해 수행하였는데, 통계량 MSR/MSE는 자유도 $[k, n-(k+1)]$인 F 분포를 따른다고 알려져 있다. 본 연구에서는 F 통계량을 구한 후, F 분포를 이용한 유의수준 α에서 임계치를 비교하여 F 통계량이 임계치보다 크면 유의하다는 판단을 하였다.

특정 데이터의 이상치 여부는 구성된 회귀모델의 예측치와 실제 관측치의 차이 정도에 따라 판단될 수 있다. 즉, 통계적으로 매우 큰 잔차(관측치-예측치)가 관측될 경우 이는 이상치로 간주될 수 있다.

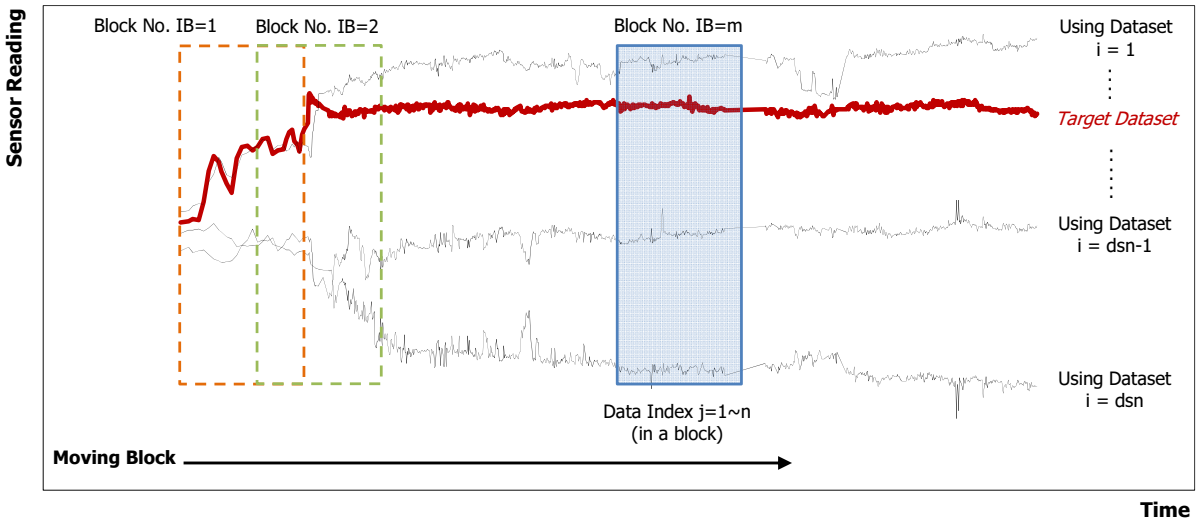

Fig. 3. Synthesis of each signal in moving data block including all dataset

Tabel 1. Analysis of variance for significance verification of regression model

Value	Degree of freedom	Mean square	F-statistic	Threshold
SSR = $\sum(\hat{y_i} - \bar{y})^2$	k	$MSR = \frac{SSR}{k}$	$\frac{MSR}{MSE}$	$F[\alpha; k, n-(k+1)]$
SSE = $\sum(y_i - \hat{y})^2$	n-(k+1)	$MSE = \frac{SSE}{n-k-1}$		
$\sum(y_i - \bar{y})^2$	n-1			

※ n : number of observed value, k : number of independent variable

잔차는 평균이 0이고 표준편차가 σ인 정규분포 특성을 가지나, 본 연구에서는 표준편차가 관측시점에 따라 달라지므로 예측잔차의 정규화된 표현인 스튜던트 잔차(studentized residual)를 사용하였다. j번째 관측치의 잔차를 e_j라고 할 때, 스튜던트 잔차 t_n은 Eq. (5)와 같다. Leverage 포인트는 회귀모델의 독립변수 값이 평균에서 많이 벗어나는 경우 스튜던트 잔차 값을 상대적으로 높게 만들어 이상치로 감지될 확률을 높이는 역할을 한다.

$$e_j = v_{i,j} - \hat{v}_j \tag{3}$$

여기서, e_j는 특정 블록 내 j 번째 데이터에 대한 잔차이며 $\hat{v}_j$는 합성신호 회귀모델을 이용한 특정 블록 내 j 번째 데이터에 대한 예측값을 나타낸다.

$$Lev_n = \frac{1}{n} + \frac{(sv_n - \overline{sv})^2}{\sum_{j=1}^{n}(sv_j - \overline{sv})^2} \tag{4}$$

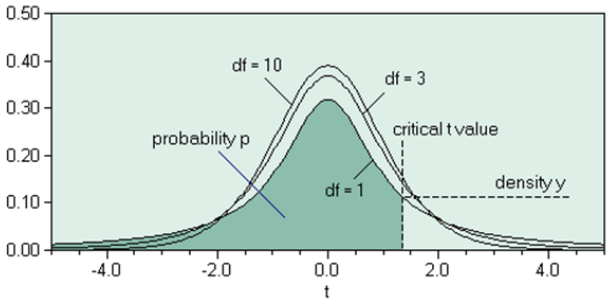

Fig. 4. t-distribution acceptance region

여기서, Lev_n는 특정 블록 내 n 번째 데이터에 대한 레버리지, sv_j는 특정 블록 내 j 번째(특정일자) 데이터에 대한 합성신호 값, $\overline{sv}$는 특정 블록 내 합성신호 평균값을 나타낸다.

$$t_n = \frac{e_n}{\sqrt{MSE(1 - Lev_n)}} \tag{5}$$

여기서, t_n은 특정 블록 내 n 번째 데이터에 대한 스튜던트 잔차, e_n은 특정 블록 내 n 번째 데이터에 대한 잔차를 나타낸다.

관측 데이터수를 n, 회귀모델 독립변수 개수를 k라고 할 때, 스트던트 잔차는 n-k 자유도를 가지는 t 분포를 따르는 것으로 알려져 있는 바, 유의수준을 95%로 하면 Fig. 4에서 채워진 영역이 양측검정에 의해 97.5%가 되고 스튜던트 잔차의 임계치 t값을 구할 수 있게 된다. 센서 관측치의 스튜던트 잔차 t_n를 구하여, 이 값이 위에서 구한 임계치 보다 크면 최종적인 이상치로 판단하였다.

4. 합성신호 기법의 평가 및 검증

4.1 임의 시계열 모델 적용

미지의 물리적인 독립변수 x_1, x_2를 가정하고 임의 직선의 방정식을 통해 6개의 데이터셋($y_1 \sim y_6$)으로 구성된 센서 관측치를 생성하였다. 노이즈는 $N(0,4)$ 정규분포로부터 난수형태로 생성하여 데이터 관측치에 포함 시켰으며, 센서마다 관측치 16개를 생성하였다. Fig. 5는 데이터셋 y_6을 대상

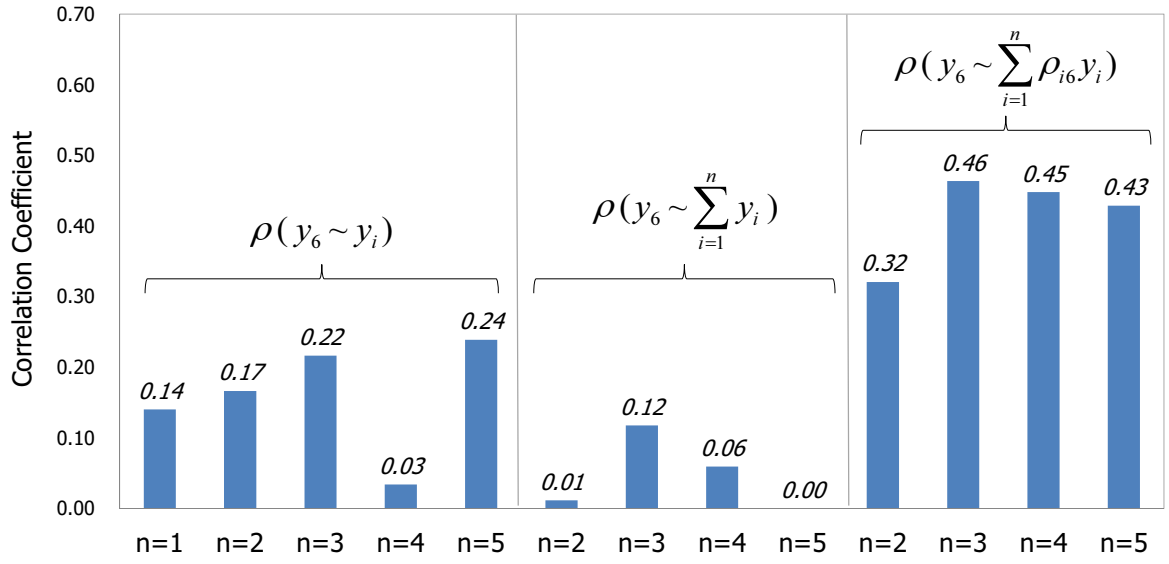

Fig. 5. Variation of correlation by application of synthesis value for virtual data

으로 각각의 개별 타 데이터셋과의 상관계수, 특정 데이터셋에 대한 단순합과의 상관계수 및 상관계수 가중합과의 상관계수를 나타낸다. 관측 데이터셋 y_6을 예측하는 데 있어 타 관측치들을 이용하기 위해서는 y_6와 상관도가 높은 관측 데이터셋을 찾아내야 하지만, Fig. 2에서 보는 바와 같이 y_6와 각 타 데이터셋 들을 대상으로 한 상관도는 노이즈 영향으로 인해 낮게 나타나고 있다. 타 관측치들의 단순합을 이용한 합성신호는 상관도를 더 떨어뜨리는 효과를 보이고 있다. 반면 y_6와 각각의 타 관측 데이터셋들 간 상관계수를 가중치로 사용한 합성신호는 상관도를 높이는 효과를 보이고 있음을 알 수 있다.

선형 최소자승법 회귀분석 모델(Kailath, 1975)에 대한 Eq. (1)을 이용하여 두 데이터셋에 의한 y_6 예측을 수행하였다. Fig. 6은 y_5를 이용하여 y_6를 예측한 결과와 합성신호 기법을 이용하여 y_6을 예측한 결과를 나타낸 것으로서, 특정 데이터셋 만을 이용한 결과에 비해 합성신호 기법(ASCS)을 이용한 예측결과는 향상된 상관성에 기인하여 예측성능도 향상시키고 있음을 알 수 있다.

4.2 인공 데이터 생성 및 적용

시계열 형태의 표준 인공 데이터를 생성하고 이를 이용하여 이상치 평가 기법에 대한 정량적 평가와 검증을 실시하였다. 특정 시계열 모델의 데이터셋은 독립변수와 계수들로 구성된 1차 선형 방정식에 백색잡음(noise)을 추가하는 방식으로 작성하였다. 표준정규분포(μ=0, σ=1.0)를 따르는 역함수 값을 난수형태로 발생하고 여기에 특정 AR 계수값을 적용해 시계열 데이터셋을 작성한 후 이에 대한 2차 차분 값을 특정 독립변수 데이터셋으로 생성하였다. 특정 조건의 인공 데이터는 6개의 유사 시계열 모델에 의한 데이터셋 및 3개의 유사 시계열 모델에 의한 데이터셋, 6개의 데이터셋으로 구성되지만 2종류의 상이한 시계열 모델이 적용된 데이터셋으로 구성하였다. 독립변수 생성과 동일한 과정으로 1차 선형 방정식의 계수들 및 백색잡음을 생성하고, 이들을 모두 종합하여 최종적인 특정 데이터셋을 생성하였다. 백색잡음($ErrStd$)은 평균 0, 표준편차 0.5~2.0 범위에서 난수형태로 생성하였으며, 정상 시계열 데이터셋 내에 그 강도(f_i)를 달리한 인공 오류 데이터를 삽입시켜 최종적인 표준 인공 데이터를 생성하였다. 인공 오류 데이터는 Eq. (6)과같이 해당 정상 데이터를 중심으로 전후 3개씩 총 7개 정상 데이터의 평균값에 특정 오류 강도를 곱하는 방식으로 생성하였다.

$$AD_i = f_i \times \frac{\sum_{j=i-3}^{i+3} DS_j}{7} \tag{6}$$

여기서, AD_i, DS_j는 특정 인공 오류 데이터 및 정상 데이터, f_i는 인공 오류 데이터의 오류 강도를 나타낸다. Fig. 7~Fig. 9는 오류 데이터가 포함된 인공 데이터셋에 대한 이상치 분석 결과를 나타낸다.

6개의 유사 시계열 모델을 모두 이용한 다중 합성신호 분석 결과, 백색잡음의 표준편차가 0.5~1.0 범위에서는 이상

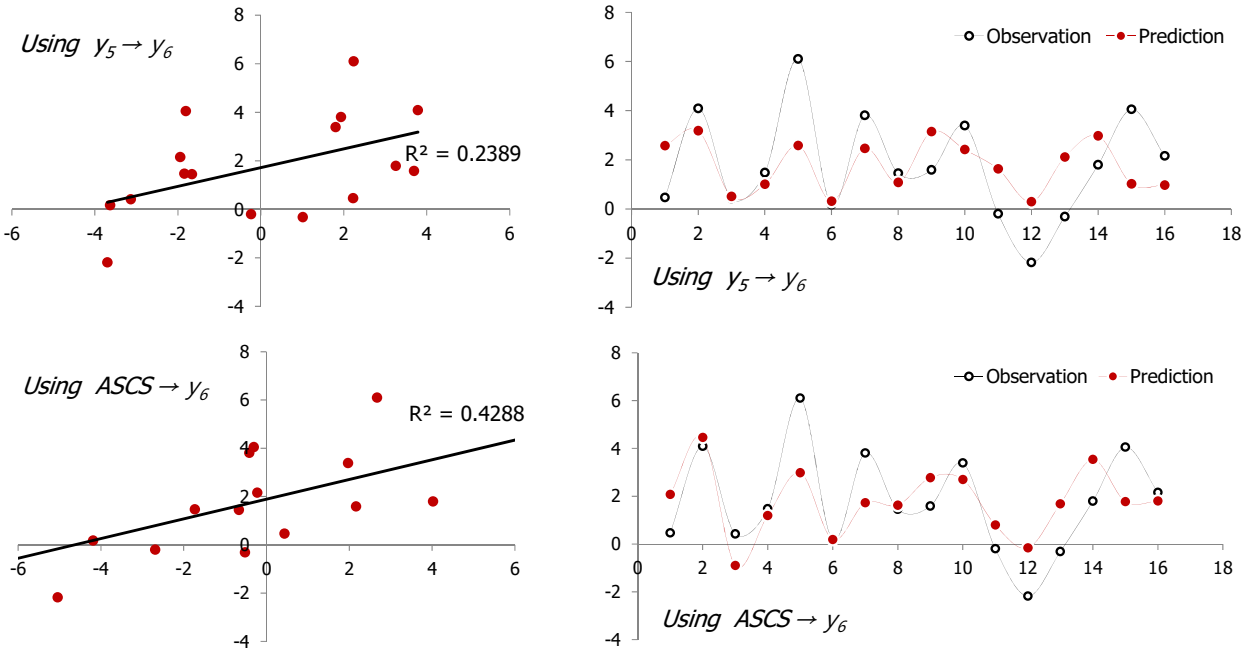

Fig. 6. Prediction of data by linear least square estimation using y_5 and ASCS

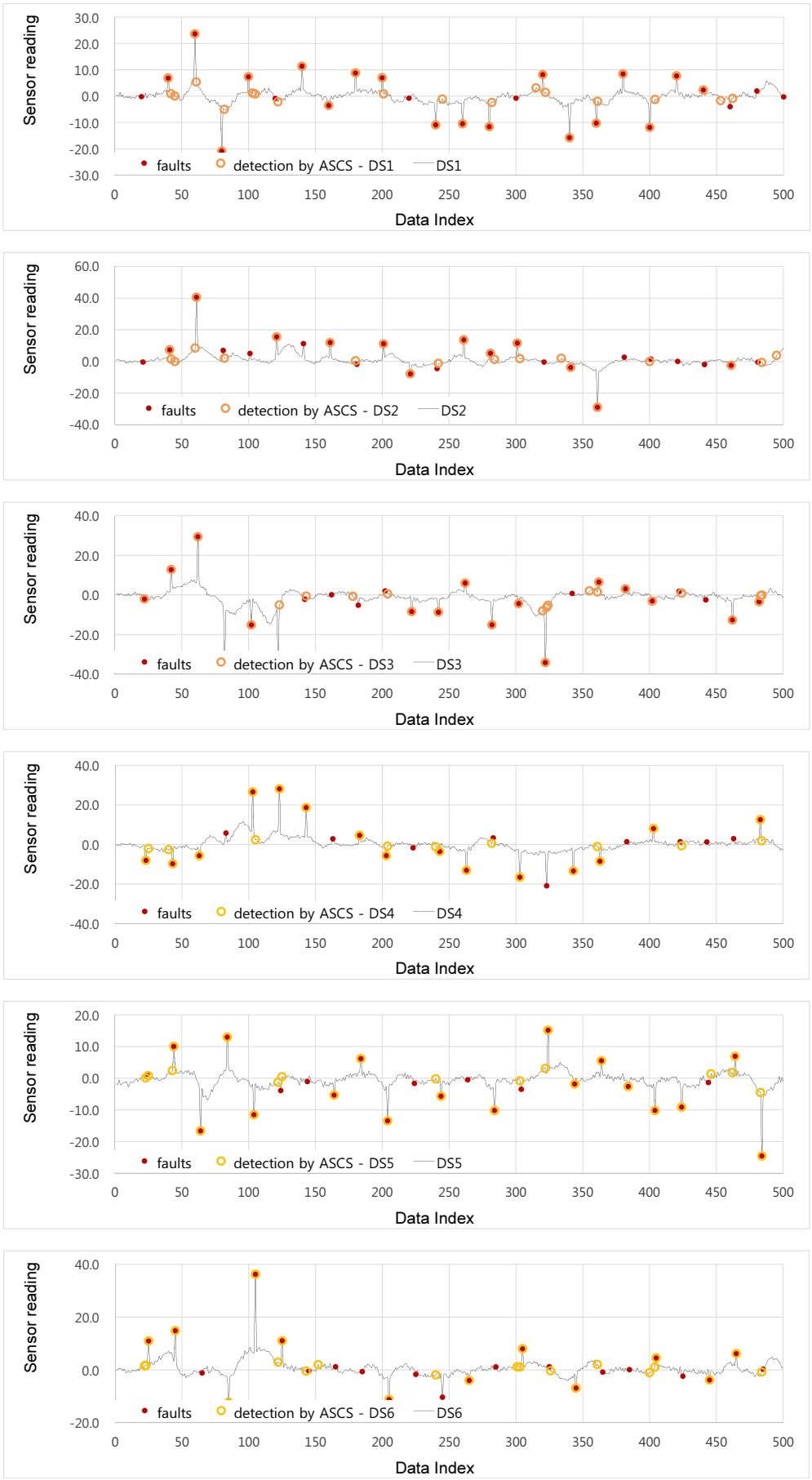

Fig. 7. Outlier detection in artificial dataset under condition of $f_i = 5.0$, $ErrStd = 0.5$ and similar six sort of time series models

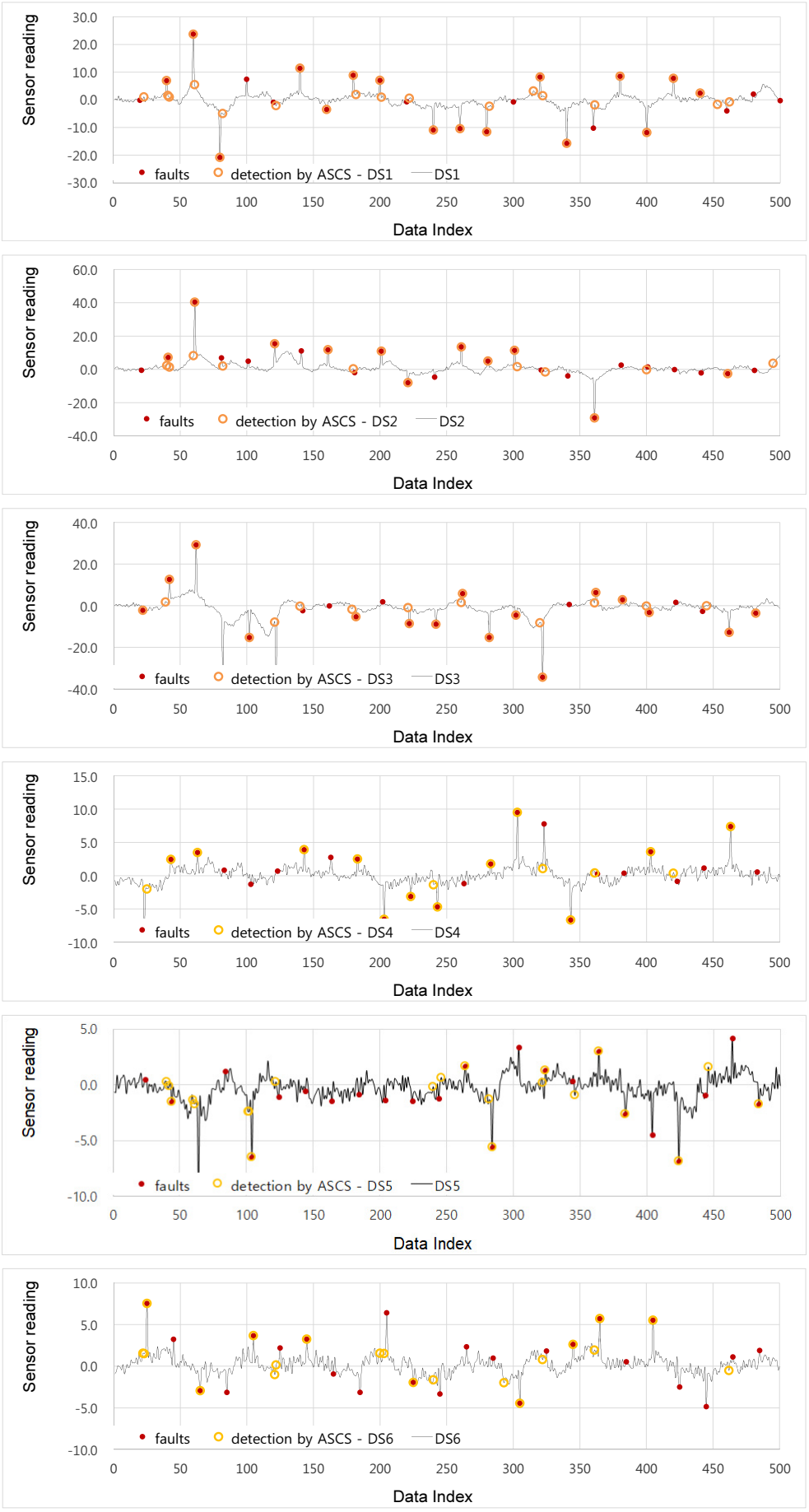

Fig. 8. Outlier detection in artificial dataset under condition of $f_i = 5.0$, $ErrStd = 0.5$ and different two types of time series models

치 판정 정확도가 80%에 육박하거나 상회하는 것으로 나타났으며, 그 정도는 오류 데이터의 강도가 커질수록 증가하는 것으로 나타났다. 백색잡음의 표준편차가 1.5를 넘어가는 경우는 실제로도 정상 데이터와 오류 데이터의 구분이 모호한 경우가 많았으며, 오류 데이터의 강도가 7.5 이상이 되어도 변화가 적은 구간의 정상 데이터로부터 생성된 오류 데이터들은 그 판정에 한계가 있었다. 실제 백색잡음의 표준편차가 1.0 이하이고 오류 데이터의 강도가 7.5 이상에서는 이상치 판정 정확도가 80~90%로 나타났으며, 육안상 구분의 모호함을 고려한다면 실제적인 판정 정확도는 더 클 것으로 예상된다. 각각 3개의 유사 시계열 모델로 구성되며 2가지 이종 시계열 모델이 적용된 인공 데이터셋의 다중 합성신호 분석 결과, 6개의 유사 시계열 모델을 모두 이용한 경우에 비해 이상치 판정정확도는 대략 10% 정도 낮게 나타났다. 3개의 유사 시계열 모델만을 이용한 다중 합성신호 분석 결과, 백색잡음의 표준편차가 0.5~1.0 범위에서는 이상치 판정 정확도가 62% 내외로 나타났으며, 그 정도는 오류 데이터의 강도가 커질수록 증가하는 것으로 나타났다.

5. 결 론

과거부터 건설 분야에서 구조물 시공 및 안전 관리 차원에서 수행되었던 수많은 계측 데이터에는 다양한 원인에 기인한 각종 이상 데이터가 포함되어 있다. 이러한 이상 데이터는 계측결과를 활용한 구조물 거동 예측 및 안전관리 결과의 신뢰성을 떨어뜨리게 된다.

본 연구에서는 각종 시계열 계측 데이터들 중 다양한 원인에서 생성될 수 있는 여러 오류 데이터들을 선별하기 위한 목적으로, 다수의 데이터셋을 복합적으로 활용하는 이상치 분석기법과 그 평가 및 검증에 관한 연구를 수행하였다. 단일 데이터셋 만을 대상으로 한 이상치 분석은 시계열 분석법 등을 통해 수행될 수 있으나, 매우 강하고 일관된 시계열 특성을 보이지 않는 한 적정 이상치 분석에 한계가 있다. 회귀분석을 이용하는 경우는 두 개의 데이터셋 내지는 다수의 데이터셋을 이용할 수 있는 장점이 있으나, 다수 지점에서의 다양한 계측결과 중 상당수 데이터셋은 그들 간 상관성이 매우 낮을 수 있으며, 이러한 경우는 최종적인 이상치 평가결과의 정도가 매우 낮아지게 되는 단점이 있다. 각 데이터셋 간 상관성을 가중치로 한 합성신호 생성 및 회귀분

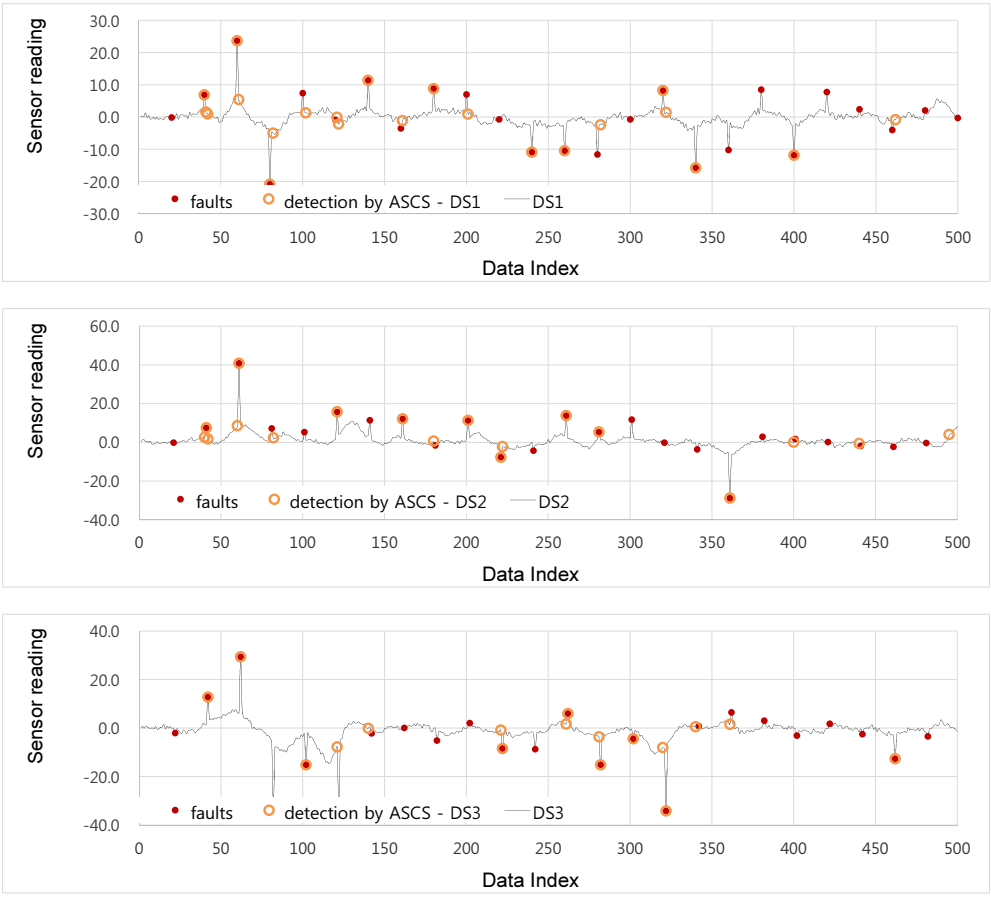

Fig. 9. Outlier detection in artificial dataset under condition of $f_i = 5.0$, $ErrStd = 0.5$ and similar three sort of time series models

석 기법을 이용한 이상치 분석기법에 대하여 임의 시계열 데이터와 오류 데이터가 포함된 표준 인공 데이터를 이용하여 해당 기법에 대한 평가 및 검증을 수행하였다.

임의 시계열 모델을 통한 각 데이터셋 간의 상관성 분석을 실시한 결과, 다수의 데이터셋을 활용하되 그들 간 상관성을 가중치로한 합성신호는 특정 데이터셋과의 상관성을 높이는 특성이 있는 것으로 나타났다.

유사 시계열 모델 및 이종 시계열 모델을 혼합하되 특정 강도의 오류 데이터와 백색잡음을 포함하는 표준 인공 데이터를 생성하였으며, 이를 이용한 합성신호 기법에 대한 평가를 수행하였다. 합성신호 생성에 많은 수의 유사 시계열 모델이 사용될수록 그 판정 정확도는 높아짐을 알 수 있었으며, 이는 실제 동종의 계측기가 많을수록 최종적인 합성신호 분석법의 이상치 판정 정확도가 높아짐을 의미한다.

6개의 동종 및 이종 데이터셋이 포함된 표준 인공데이터의 경우, 백색잡음 표준편차가 1.0 이하이고 오류 데이터의 강도가 7.5 이상에서도 이상치 판정 정확도가 70～90%로 나타났으며, 합성신호 생성에 이용되는 데이터셋 수가 감소할수록 최종적인 이상치 판정 정확도는 대략 10～20% 이상 낮아짐을 알 수 있었다.

감사의 글

본 연구는 인덕대학교의 2015학년도 교내학술연구비 지원에 의해 수행되었습니다.

References

1. Jeffery, S. R., Alonso, G., Franklin, M. J., Hong, W. and Widom, J. (2006), Declarative support for sensor data cleaning, Proc. of 4th International Conference on Pervasive Computing, Ireland, pp. 83~100.
2. Jeon, J. S., Koo, J. K. and Park, C. M. (2015a), Outlier detection in time series monitoring datasets using rule based and correlation analysis method, Journal of the Korean Geo-Environmental Society, Vol. 16, No. 5, pp. 43~53 (in Korean).
3. Jeon, J. S., Shin, D. H. and Kim, K. Y. (2015b), Outlier detection in time series monitoring dataset by adaptive multiple synthesis method, KSCE 2015 Convention, Kunsan, Korea, pp. 27~28 (in Korean).
4. Jeon, J. S. (2016), Development of outlier evaluation technique and operation system for monitoring data, KSCE 2016 Convention, Jeju, Korea, pp. 355~356 (in Korean).
5. Kailath, T. (1975), Square-root algorithms for least-squares estimation, IEEE Trans. Automatic Control, Vol. 20, No. 4, pp. 487~497.
6. Mourad, M. and Bertrand-Krajewski, J.-L. (2002), A method for automatic validation of long time series of data in urban hydrology, Water Science and Technology, Vol. 45, No. 4~5, pp. 263~270.
7. Ni, k., Ramanathan, N., Chehade, M., Balzano, L., Nair, S., Zahedi, S., Pottie, G., Hansen, M. and Srivastava., M. (2009), Sensor network data fault types, ACM Transactions on Sensor Networks, Vol. 5, No. 3, Article 25, pp. 1~29.
8. Park, C. M. and Jeon, J. S. (2015), Regression-based outlier detection of sensor measurement using independent variable synthesis, Journal of Korean Institute of Plant Engineering, Vol. 20, No. 3, pp. 87~93 (in Korean).
9. Ramanathan, N., Balzano, L., Burt, M., Estrin, D., Kohler, E., Harmon, T., Harvey, C., Jay, J., Rothenberg, S. and Srivastava, M. (2006), Rapid deployment with confidence: calibration and fault detection in environmental sensor networks. Tech. Rep. 62, CENS, pp. 1~14.
10. Sharma, A. B., Golubchik, L. and Govindan, R. (2010), Sensor faults: detection methods and prevalence in real-world datasets, ACM Transactions on Sensor Networks, Vol. 6, No. 3, Article 23, pp.1~39.
11. Williams, G. J., Baxter, R. A., He, H. X., Hawkins, S. and Gu, L. (2002), A comparative study of RNN for outlier detection in data mining, IEEE International Conference on Data-mining (ICDM'02), Maebashi City, Japan, CSIRO Technical Report CMIS-02/102, pp. 1~709.