한국해안 · 해양공학회논문집/ISSN 1976-8192(Print), ISSN 2288-2227(Online)
Journal of Korean Society of Coastal and Ocean Engineers 26(5), pp. 285~291, Oct. 2014
http://dx.doi.org/10.9765/KSCOE.2014.26.5.285

수온 관측 자료의 효율적인 이상 자료 탐지
Efficient Outlier Detection of the Water Temperature Monitoring Data

조홍연* · 정신택** · 고동휘** · 손경표***
Hongyeon Cho*, Shin Taek Jeong**, Dong Hui Ko** and Kyeong-Pyo Son***

요 지 : 연안의 수온 모니터링 자료는 이상자료 및 결측을 포함하고 있기 때문에 통계정보를 왜곡할 수 있다. 다양한 이상자료 감지 기법이 제안되고 있으나 결측이 없고 이상자료에 대한 사전정보를 가정하고, 어떤 적용기법은 과도한 계산시간이 소요되기 때문에 적용에 제한이 따른다. 본 연구에서는 방대한 자료에서도 효과적으로 이상자료를 감지할 수 있는 실용적인 Robust 모형을 제안하였다. 이 모형은 계산시간을 크게 저감하는 부분자료 추출기법을 이용한 어림성분 추정과정 및 어림성분으로부터 계산되는 잔차성분으로부터 이상자료를 반복적으로 진단하여 제거하는 부분으로 구성되어 있다. 이 모형의 성능평가는 새만금호에서 5분 간격으로 관측한 2년 동안의 수온 자료를 이용하여 수행하였다. 모형 적용결과, 이상자료가 전체자료에서 차지하는 비율은 1.6-3.7% 정도로 파악되었으며, 전체적으로 대부분의 이상자료가 제거되는 것으로 파악되었다. 또한 어림성분 추정과정의 반복적용은 Long-span 조건을 먼저 적용하는 것이 효과적인 것으로 파악되었다.

핵심용어 : 이상자료, 결측구간, 수온, 새만금호, Robust 모형, 부분자료 추출

Abstract : The statistical information of the coastal water temperature monitoring data can be biased because of outliers and missing intervals. Though a number of outlier detection methods have been developed, their applications are very limited to the in-situ monitoring data because of the assumptions of the a prior information of the outliers and no-missing condition, and the excessive computational time for some methods. In this study, the practical robust method is developed that can be efficiently and effectively detect the outliers in case of the big-data. This model is composed of these two parts, one part is the construction part of the approximate components of the monitoring data using the robust smoothing and data re-sampling method, and the other part is the main iterative outlier detection part using the detailed components of the data estimated by the approximate components. This model is tested using the two-years 5-minute interval water temperature data in Lake Saemangeum. It can be estimated that the outlier proportion of the data is about 1.6-3.7%. It shows that most of the outliers in the data are detected and removed with satisfaction by the model. In order to effectively detect and remove the outliers, the outlier detection using the long-span smoothing should be applied earlier than that using the short-span smoothing.

Keywords : outliers, missing interval, water temperature, Lake Saemangeum, robust model, sub-data re-sampling

1. 서 론

연안에서 다양한 환경센서를 이용한 관측이 수행되면서 방대한 환경인자 자료가 축적되고 있다. 그러나 자료의 축적과 더불어 환경센서의 주기적인 빈번한 관리가 곤란하고 현장에서의 예상할 수 없는 환경변화로 인한 센서작동 불량 등의 문제로 인하여 결측 및 이상자료가 빈번하게 발생하고 있다. 한정된 소수의 관측자료 또는 전담인력의 주기적인 자료관리가 가능한 경우를 제외하고는 관측자료는 이상자료와 결측구간을 포함하는 자료로 제공되고 있기 때문에 자료를 분석하고자 하는 기관이나 개인이 각자의 주관적인 방법으로 처리하고 있는 실정이다. 이상자료는 비정상적인 자료로 정의되며, 자료의 통계정보를 왜곡하기 때문에 객관적이고 적절한 감지기법을 이용하여 감지하여 제거할 필요가 있다(Agresti and Franklin, 2007; Cho and Oh, 2012; Cho et al., 2013). 특히 방대한 자료(big data)는 하나하나 살펴가며 수동으로 제거하는 과정보다는 자동화된 기법의 적용이 필요하기 때문에 자동화된 이상자료 진단-제거기법에 관한 연구가 활발하게 추

*한국해양과학기술원, 해양환경보전연구부(Marine Environments and Conservation Research Division, Korea Institute of Ocean Science and Technology, Ansan PO Box. 29, Seoul, 425-600, Korea, hycho@kiost.ac)
**원광대학교 토목환경공학과, 원광대학교 부설 공업기술개발연구소 연구위원(Corresponding author: Shin Taek Jeong, Department of Civil and Environmental Engineering, Wonkwang University, 460, Iksandae-ro, Iksan, Jeonbuk, 570-749, Korea. Tel:+82-63-850-6714, Fax:+82-63-857-7204, stjeong@wku.ac.kr)
***환경부 자원순환국 자원재활용과(Resource Recycling Division, Ministry of Environment, Government Complex-Sejong 11 Doum 6-ro, Sejong Special Self-Governing City, 339-012, Korea)

진되고 있다(Basu and Meckesheimer, 2007). 이러한 연구 중에는 또한 자료 특성을 고려한 이상자료 진단기법 등에 관한 연구도 수행되고 있다(Ben-Gal, 2005; Hubert and van der Veeken,2008).

그러나 이상자료와 결측의 발생양상과 빈도 등은 매우 다양하기 때문에 어떤 하나의 기법만으로는 다양한 이상자료 감지에는 한계가 있다. 본 연구에서는 미지의 이상자료와 결측구간이 포함되어 있는 관측자료에서 이상자료를 효과적으로 감지하는 기법을 개발하여 적용하였다. 제안된 기법은 관측자료를 다양한 시간규모에서 변화양상을 탐지하는 Robust 평활기법을 조합하여 이상자료를 제거할 수 있으며, 다양한 시간규모에서 변화양상을 신속하게 탐지하기 위한 방법으로 부분적인 자료만을 이용하는 계산시간 저감을 위한 방법도 포함한다. 방대한 자료는 변화양상을 파악하는 방법이 과도한 계산시간을 유발할 수 도 있기 때문에 신속한 방법도 실용적으로 요구된다. 개발된 모형의 성능평가는 새만금호의 4개 지점에서 5분 간격으로 관측된 자료를 이용하여 수행하였다.

2. 이상자료 제거기법 및 자료

이상자료 제거기법은 기본적으로 결측이 없는 완전한 자료(complete)를 전제조건으로 하여 이상자료를 유형별로 분류하고 이에 따른 다양한 기법이 제안되고 있다. 이상자료는 대략 AO(Additive outlier), IO(innovative outlier) 유형으로 분류되고 있으며, 각각은 대부분의 자료에 비하여 크기 자체가 유별난 자료, 자료의 변화양상에서 유별나게 벗어나는 자료를 의미한다(Tsay, 1988; Barnett & Lewis, 1994). 이상자료 감지기법은 감지하고자 하는 자료의 개수에 따라 소수의 한정된 이상자료 감지기법과 다수의 이상자료(outlier patch)를 감지하는 기법으로도 분류되고 있다(Chiang, 2008). 본 연구에서는 시간변화 양상이 뚜렷한 관측 자료에서 신속하게 시간변화 양상을 파악하는 기법을 적용하고, 평균 및 표준편차 정보를 이용하는 일반적인 이상자료 판단기준을 적용하는 단계를 포함하는 실용적인 기법을 제안한다. 기존의 이상자료 감지기법과의 실질적으로 차이는 미지의 이상자료 발생 빈도 및 양상과 결측구간을 포함한 자료에서도 이상자료를 감지할 수 있다는 부분이다.

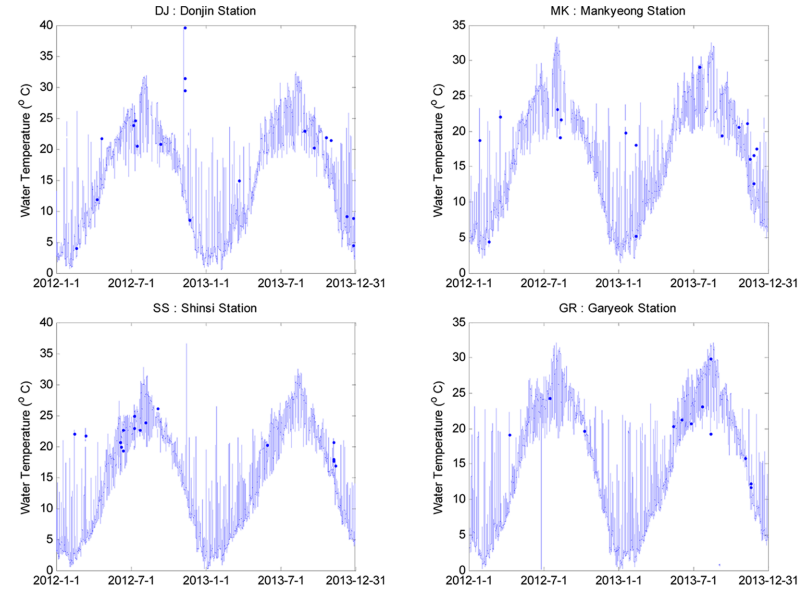
Fig. 1. Monitoring stations of the water temperature in Lake Saemangeum.

2.1 수온 관측자료

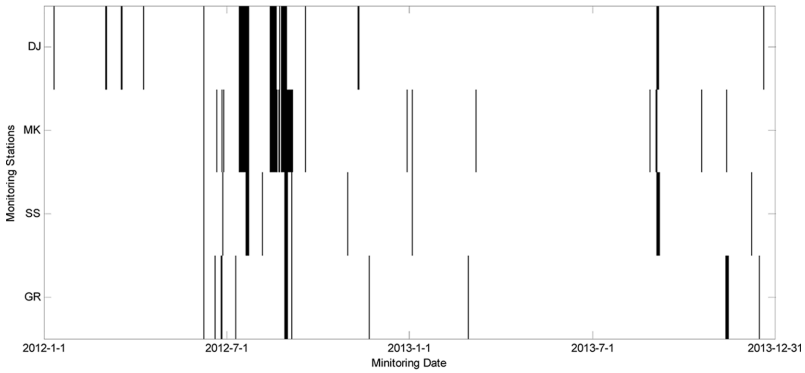
환경부는 새만금호의 수질관리를 위한 목적으로 Fig. 1에 도시된 새만금호 4개 지점(만경, 동진, 신시, 가력 지점)에서 다양한 환경인자를 센서를 이용하여 5분 간격으로 연속 관측하고 있다. 4개 지점의 구체적인 정보는 Table 1과 같다. 2012년부터 2013년까지 2년 동안 관측된 환경인자중 하나인 수온자료를 이용하여 본 연구에서는 이상자료 감지모형의 성능평가를 실시하였다. 전혀 처리과정을 거치지 않은 상태의 자료는 다른 연안 환경 모니터링 자료와 같이 결측과 빈번한 이상자료가 발생하고 있음을 알 수 있다. 본 연구에서는 이상자료 감지모형의 적용을 위한 사전단계로 수온 자료가 가질 수 있는 가능한 충분한 범위(영하 10°C, 영상 40°C 범위)를 벗어나는 자료와 센서에서 결측을 판단하는 지정수치(본 자료의 경우 정수 0) 자료는 제거하고 도시하였다(Fig. 2(a) 참조). 한편 이상 자료 제거 전의 결측정보는 결측정보 행렬의 흑백도시로 전반적인 양상을 파악할 수 있다(Fig. 2(b) 참조). 결측비율은 2년 동안의 완전한 관측 자료 개수에 대한 비율로 정의할 경우, 완전한 전체 자료의 개수는 210,528개 (= 12개/시간 × 24시간/일 × (365 + 366)일)이며, 결측구간의 크기는 동진, 만경 신시, 가력에서 각각 11,250개, 12,074개, 5,207개, 4,366개로 결측비율은 각각 5.2%, 5.7%, 2.5%, 2.1% 정도로 하천 하구지점에 해당하는 동진 및 만경 지점의 결측비율이 새만금호 배수갑문 지점에 비하여 2배 이상 높다. 자료의 결측비율은 이상자료로 진단되는 자료가 일반적으로 제거되기 때문에 이상자료로 진단되는 개수정도가 추가로 증가하게 된다.

Table 1. Basic information of the 4 water quality measurement stations

Station	Observation period	Geographical coordinate (N°, E°)	Items
Garyeak	2011.11~2014.3	35.731, 126.529	pH, DO, TP, TN, Salinity, COD, Water temperature
Dongjin	2011.11~2014.3	35.722, 126.813	pH, DO, TP, TN, Salinity, COD, Water temperature
Mankyeong	2011.11~2014.3	35.907, 126.957	pH, DO, TP, TN, Salinity, COD, Water temperature
Shinsi	2011.11~2014.3	35.815, 126.484	pH, DO, TP, TN, Salinity, COD, Water temperature



(a) Time-series plot of the water temperature data



(b) Missing indicator matrix plot of the water temperature data

Fig. 2. Time-series and missing matrix plots of the water temperature data.

2.2 실용적인 이상자료 감지모형

이상자료 감지기법은 수온 자료의 시간적인 변동양상을 파악하는 과정과 잔차 성분을 이용한 이상자료 진단과정으로 Cho and Oh(2012) 방법을 보완한 방법이다. 이상자료 진단기준은 자료의 분포특성 및 변화 특성에 대한 사전 연구 성과를 이용하는 방법이 가장 적절하나 그러한 변화 특성이 시간 및 공간에 따라 다양한 특성을 보이는 경우가 일반적이기 때문에 이상자료 진단기준 제시는 매우 완화된 기준을 사용하여 정상적인 자료의 제거과정을 최소화하여야 한다. 본 연구에서는 수온의 경우, 정상적인 범위를 확실하게 넘어서는 40°C 이상 영하 10°C 이하 자료를 이상자료로 진단하여 제거하였으며, 자료 제시과정에서 특정 수치(–9999, 0 등)로 제시되는 결측자료는 기준이 명확하기 때문에 그 기준을 이용하여 제시하였다. 정상적인 범위를 확실하게 넘어서는 조건을 제시하지 않은 경우에도, 본 연구에서 제시하는 이상자료 제거기법을 적용하면 정상적인 범위를 벗어나는 자료가 제거되지만, 이 과정은 자료의 전반적인 변화 양상 파악을 위한 도시분석(graphical analysis) 과정을 위한 간단한 절차이다. 비정상적인 범위의 수치는 명확하게 잘못된 자료로 간주할 수도 있으며, 도시분석을 방해하기 때문에 사전에 간단한 기준으로 제거하는 과정이 유용하다고 할 수 있다. 또한 자료에서 가장 최소가 되는 주요 변동양상을 일단위(day unit)로 간주하여 평균보다는 굳건한(robust) 인자로 제시되는 Median 수치와 표준편차보다는 굳건한 통계적인 수치로 제시되는 IQR(inter-quartile range) 또는 MAD(median absolute deviation) 수치(Agresti & Franklin, 2007)를 이용하여 잔차가 기준범위(편차 기준 3.0)를 벗어나는 경우 이상 자료로 간주하여 제거하였다. 잔차가 정규분포 조건을 만족하는 경우, 편차 3.0 조건을 벗어나는 자료는 1.0% 정도이며, 전체 자료에서 1.0% 정도의 특이한 자료를 이상자료로 간주하는 조건에 해당된다고 할 수 있다. 절대적인 이상자료 판단기준은 없다. 이상 자료는 잘못된 자료일 수 도 있으나, 명확한 근거가 없는 경우 특이한 자료로 간주하는 것이 타당하다. 따라서 관측자료에서 이상자료를 제거한다는 의미는 대부분의 자료유형과는 다른 양상을 가지는 소수의 자료를 제거(또는 표시)한다는 의미이며, 특이한 자료 분석의 경우에는 이상 자료 제거에 보다 신중할 필요가 있다.

어림 성분의 변화 양상은 Robust Smoothing 기법을 이용

하여 수행하였다. Robust Smoothing 기법은 자료의 일정한 연속구간(Span)의 자료만을 대상으로 최적 직선(또는 곡선)을 Robust 기법으로 추출하는 방법으로, 비교적 이상 자료에 둔감한 방법이다(Cleveland, 1979). 어림성분은 Span (또는 Bandwidth) 크기에 따라 변화한다. 작은 Span 조건에서는 작은 시간규모까지의 변화 양상을 파악하나 이상 자료의 영향에 민감하고, 큰 Span 조건에서는 작은 변동양상이 무시되기 때문에 부분적인 변화양상을 파악할 수 없다(Fig. 3 참조). 따라서 최적 Span 조건을 도출할 필요가 있으며, 본 연구에서는 $(\text{Bias})^2 + \text{Variance}$ 수치로 제공되는 목적함수가 최소가 되는 Span 조건을 선택하였다. Span 조건의 극단은 Span = 0 조건과 Span = 100%(= 1) 조건이며, 이 조건을 이용하여 Smoothing 기법을 적용하는 경우, 각각 주어진 자료(No variance)와 똑 같은 자료와 평균(No bias)으로 어림 성분이 추정된다. 최적 Span 추정을 위한 계산과정에서 필요한 분산 및 편기는 다음과 같이 식 (1)로 정의된다.

$$T(t_i) = \hat{T}(t_i) + \varepsilon(t_i),\ i = 1, 2, \ldots, N \tag{1}$$

여기서, $T(t_i)$, $\hat{T}(t_i)$는 각각 관측 수온자료, 어림(approximate, smoothed) 수온자료이며, $\varepsilon(t_i)$는 잔차 성분에 해당한다. N은 자료의 개수이다.

관측 자료의 분산(variance)과 편기는 다음과 같이 식 (2)로 정의되며, 여기서 $\bar{T}$는 관측 수온자료의 평균이다.

$$\begin{aligned} Variance[T(t)] &= \frac{1}{(N-1)}\sum_{i=1}^{N}[T(t_i)-\bar{T}] \approx \frac{1}{N}\sum_{i=1}^{N}[T(t_i)-\bar{T}]^2 \\ [Bias\{T(t)\}]^2 &= \frac{1}{N}\sum_{i=1}^{N}[T(t_i)-\hat{T}(t_i)]^2 \end{aligned} \tag{2}$$

식 (1)에서 정의된 어림 수온자료는 Span 크기에 따라 다르기 때문에 최적 Span 조건에 해당하는 어림 수온 추정자료가 필요하며, 그 조건으로는 일반적으로 이용되는 최소제곱 잔차(least squared residual) 조건이 적용될 수 있으며, 그 경우 식 (3)에서 볼 수 있는 바와 같이 최적 Span 조건은 분산과 편기의 제곱을 합한 수치가 최소가 되는 조건에서 결정된다(Silverman, 1998; Storch & Zwiers, 1999).

$$\begin{aligned} \frac{1}{2}\sum_{i=1}^{N}\varepsilon^2(t_i) &= \frac{1}{N}\sum_{i=1}^{N}[T(t_i)-\hat{T}(t_i)]^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}[T(t_i)-\bar{T}+\bar{T}-\hat{T}(t_i)]^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}[(T(t_i)-\bar{T})^2+(\hat{T}(t_i)-\bar{T})^2+2(T(t_i)-\bar{T})(\hat{T}(t_i)-\bar{T})] \\ &= \frac{1}{N}\sum_{i=1}^{N}[(T(t_i)-\bar{T})^2+(\hat{T}(t_i)-\bar{T})^2] \\ &= Variance(T(t_i)) + [Bias(T(t_i))]^2 \end{aligned} \tag{3}$$

자료의 개수가 많은 방대한 자료의 경우 비교적 큰 Span (10-30%) 조건을 적용하는 경우, 과도한 계산시간이 소요되기 때문에 시간적인 측면에서 매우 비효율적이다. 본 연구에서는 이러한 과도한 계산시간을 단축하기 위한 방법으로 전체 자료에서 Span 규모에 상응하는 자료를 부분 추출하여 Robust Smoothing 기법을 적용하는 모형을 개발하였다. 부분 추출자료는 특정 시간간격에서 1-2개의 자료를 추출하는 방법으로 본 연구에서는 일 자료(144개)에서 1개의 자료를 추출하고, 추출된 자료를 이용하여 Robust Smoothing 기법을 적용하기 때문에 자료의 개수는 1/144 정도로 감소하게 되어

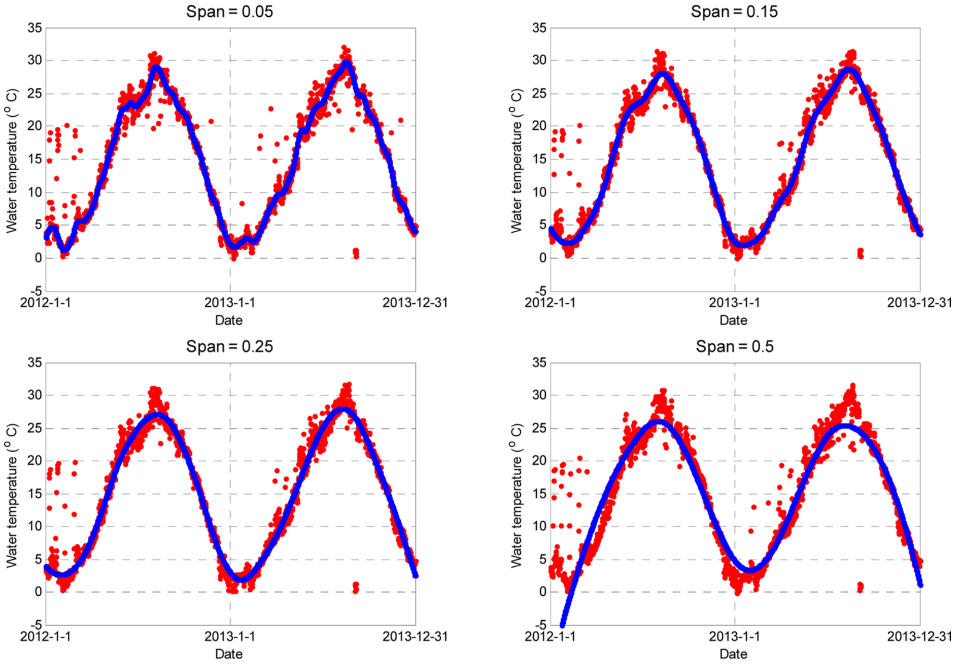


Fig. 3. Smoothing curve with the span percentages.

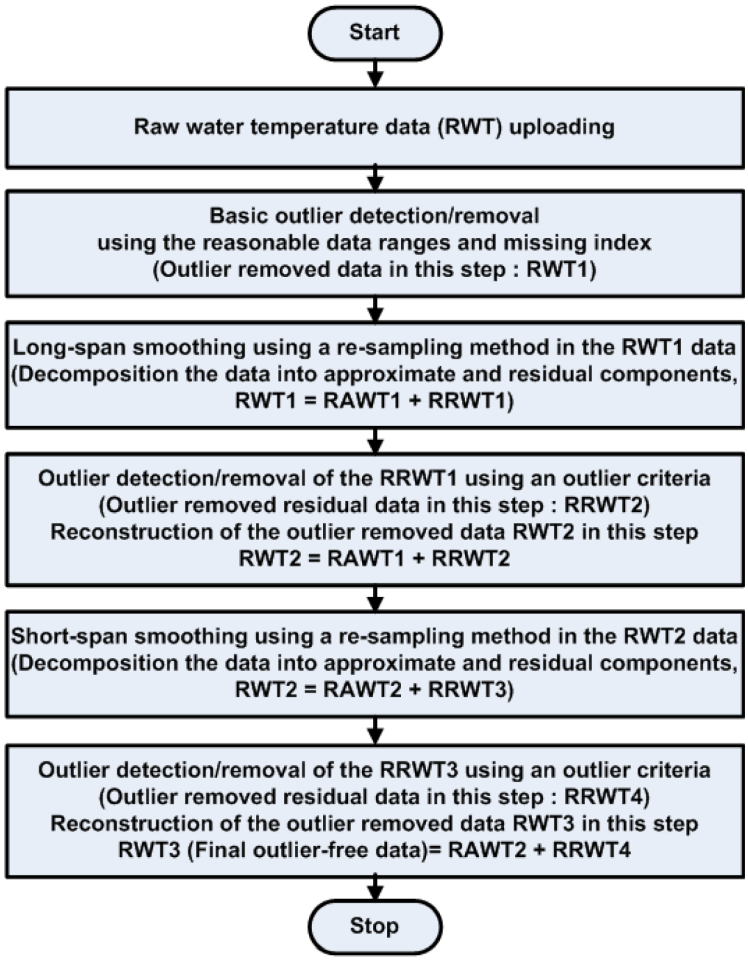


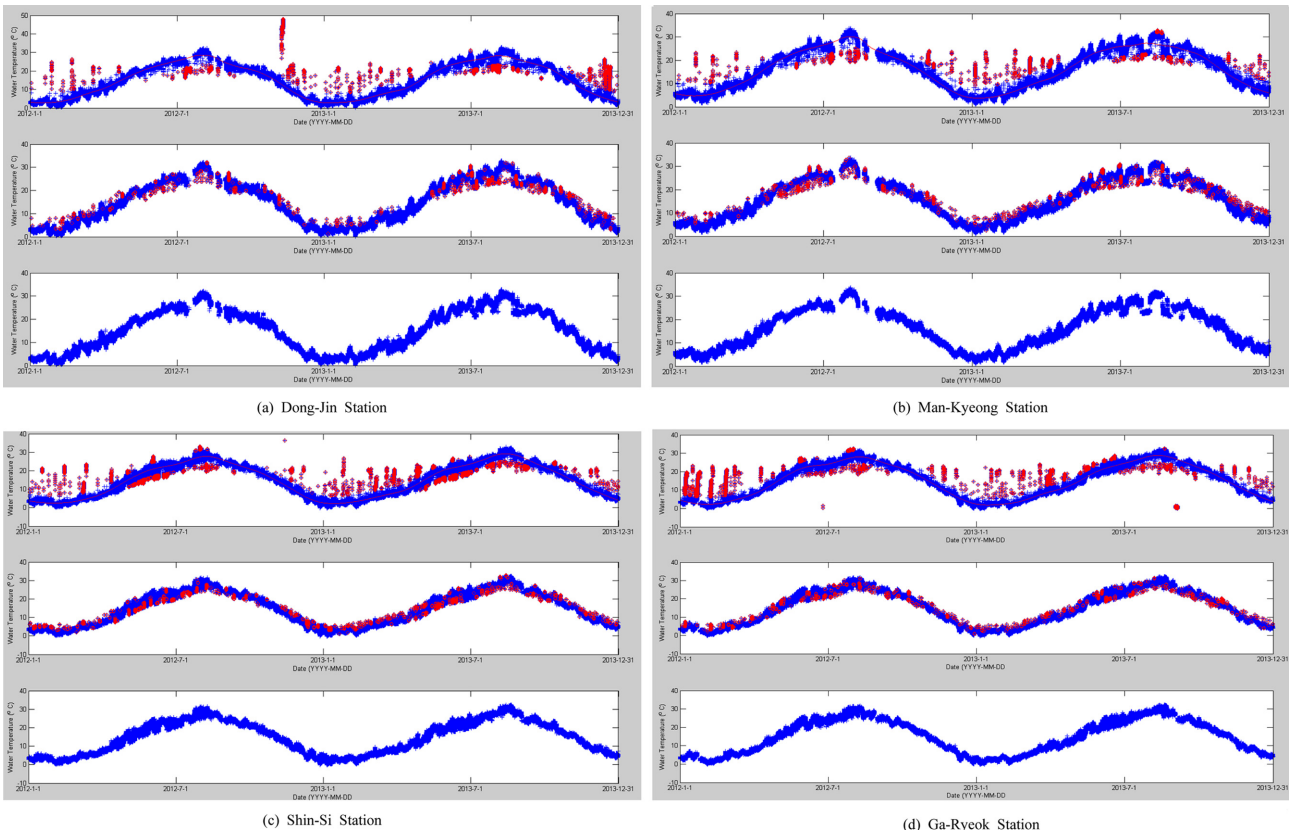
Fig. 4. Flowchart of the outlier detection/removal process.

매우 신속한 계산이 가능하게 된다. 특정 시간규모에서의 자료의 추출시점 변화에 따른 어림성분(Robust Smoothing 기법을 이용하여 추출된 시간적인 변화 성분)의 변동은 미미한 정도로 파악되었다. 잔차성분 자료의 분포는 이상자료를 제외하는 경우, 정규분포와 유사한 분포를 보이고 있는 것으로 파악되었다. 본 연구에서는 잔차성분의 정규성 검정(normality test)은 수행하지 않았다.

일단 주어진 Span 조건에서 자료의 어림성분이 추정되면, 어림성분과 관측자료의 차이를 잔차성분으로 간주할 수 있으며, 잔차성분에서 Robust 통계매개변수에 해당하는 Median 및 IQR 정보를 이용하여 비교적 널리 이용되는 이상자료 판단기준을 적용하여 이상자료를 감지하게 된다. 감지되는 이상자료는 어림성분을 추출하는 Span 조건에 따라 다르다. 큰 Span 조건을 사용하는 경우, 전체적인 변화 양상에서 두드러지게 드러나는 이상자료를 제거할 수 있으며, 작은 Span 조건을 적용하는 경우, 개수가 적지만 빈번하게 발생하는 이상자료를 제거할 수 있다. 어림성분의 추출결과는 결측구간과 이상자료의 개수 및 발생빈도의 영향을 받기 때문에 이러한 영향을 고려하여 이상자료를 추출하는 기법을 적용할 수 도 있으나, 실질적으로 이상자료 및 결측구간에 대한 사전 정보를 이용할 수 없기 때문에 적용에 제한이 따른다. 본 연구에서 개발하여 적용하는 이상자료 감지과정은 다음과 같은 흐름도로 표현되며 (Fig. 4 참조), 각각의 과정에서 자료의 특성을 고려하여 모형의 매개변수를 조정할 수 있다. 보다 세부적인 또는 전반적인 시간규모의 이상 자료제거를 위해서는 적절한 Span 조건 또는 이상 자료 진단조건을 조합하여 추가적으로 실행할 수 있다. Fig. 4의 흐름도에 제시된 약어에서 RWT, RAWT, RRWT는 각각 수온자료, 수온자료의 어림성분 및 잔차성분을 의미하며, 각각의 약어 뒤에 붙은 수치는 이상자료 제거단계를 의미한다. 처리단계에서 어림성분은 유지하면서 잔차성분에서 이상자료를 제거하고, 다시 이상자료가 제거된 잔차성분을 이용하여 이상자료가 제거된 수온자료를 구성하기 때문에 각각의 성분에 따라 처리단계 번호에 차이가 발생한다.

3. 이상 자료 진단모형 적용 결과

2.2 절에서 소개한 이상자료 제거기법을 조합하여 적용하면 다음과 같은 이상 자료 감지 및 제거 전 · 후의 자료 변화 양상을 파악할 수 있다. Robust Smoothing 기법을 이용한 이상자료 탐지는 Span = 15%(전체 자료의 15% 정도[약 31,579개]를 사용하여 어림성분을 추출) 조건에서 어림성분을 추출하여 잔차의 이상 자료를 제1차로 탐지 · 제거하고, 다음



(a) Dong-Jin Station (b) Man-Kyeong Station

(c) Shin-Si Station (d) Ga-Rveok Station

Fig. 5. Outlier detection-removal pattern by the model with a different span. (Red circles = detected outliers, Blue + = Data; Upper panel : Long-span outlier detection model, Middle panel : Short span outlier detection model, Lower panel : Outlier-free data)

Table 2. Basic statistical information before and after outlier removals

Stations		n	MR(%)	Mean	Median	SD	IQR
DJ	BOR	199,544	5.2	14.9	15.2	8.8	16.1
	AOR	194,809	7.5	14.7	14.7	8.8	16.3
MK	BOR	198,454	5.7	16.0	16.4	8.3	14.7
	AOR	194,237	7.3	15.9	16.1	8.3	14.9
SS	BOR	205,321	2.5	14.5	15.1	8.8	16.5
	AOR	197,721	6.1	14.4	14.4	8.9	16.7
GR	BOR	206,162	2.1	15.0	15.5	9.1	17.0
	AOR	198,770	5.8	14.9	15.2	9.1	17.2

Ref. AOR and BOR mean "after and before outlier removal" cases, respectively.

에는 Span = 144 × 3 조건(3일 동안의 자료)에서 어림성분을 추출하여 이상 자료를 제2차로 탐지·제거하였다. 이 방법은 Span 조건을 달리하여 추가로 더 적용할 수 도 있으나, Span 조건은 큰 조건에서 작은 조건 순서로 적용하는 것이 이상 자료의 영향에 보다 둔감하기 때문에 바람직한 결과를 제시한다(Fig. 5 참조). 이상 자료가 정확하게 제거되었는가를 판단하는 정확한 기준은 없다. 다만 이상 자료로 의심되는 자료를 이상 자료로 간주하여 제거한 것이기 때문에 자료 손상이 최소가 되는 범위에서 제거하는 방법이 권장된다.

한편 이상자료가 관측 자료의 통계정보에 미치는 영향을 분석하기 위하여 각각의 단계에서의 기본적인 통계정보에 해당하는 수치를 추정하여 제시하였다(Table 2 참조). 가장 그럴듯한(most likely) 통계정보는 결측구간에 해당하는 미지의 자료의 영향을 받기 때문에 결측구간의 자료를 적절한 방법으로 추정하여 채운 뒤의 통계정보가 가장 그럴듯한 정보로 간주할 수 있으나, 본 연구에서는 결측구간의 자료 보충은 수행하지 않았기 때문에 최종 단계를 적용한 경우의 자료를 기준으로 통계정보 변화 양상을 비교·분석하였다. 이상자료는 부분적으로는 큰 영향(DJ, SS 지점의 Median 차이)을 미치고 있는 것으로 판단할 수 있으나 전체 자료에서 차지하는 비율이 5.8-7.5% 정도이기 때문에 평균 및 표준편차에 미치는 영향은 미미한 수준으로 파악되었다. 이상 자료를 제거하는 경우, 자료의 통계적인 분포 추정이 안정되는 것은 당연한 결과이지만, 이상자료의 적절한 제거를 의미하기도 한다.

4. 토 의

이상 자료 진단과정 그림(Fig. 5)에서 볼 수 있는 바와 같이 자료의 시간적인 변동양상에서 크게 벗어나는 이상 자료는 큰 Span 조건에서 추출된 어림성분에 대한 잔차를 이용하면 이상 자료가 대부분 감지되어 제거되고 있음을 알 수 있다. 이 과정에서 감지되지 않은 국지적인 소규모의 이상 자료도 작은 Span 조건에서 추출된 어림성분으로부터 도출된

잔차를 이용하면 대부분이 감지되어 제거되고 있다. 보다 더 작은 시간규모의 이상 자료는 자료의 작은 시간규모에 대한 시계열 모형을 구축하여 진단하는 방법이 적절할 것으로 판단되나, 과도한 진단은 정상인 자료도 제거될 수 있기 때문에 최적의 모형적용이 필요하지만, 최적 적용에 대해서는 정확한 판단은 곤란한 실정이다. 다만 “가장 그럴듯한” 판단만이 가능하며, 자료 분석 경험자의 주관적인 판단과 모형을 이용한 객관적인 판단의 적절한 조합이 필요하다.

한편 어떤 Span 조건의 조합이 최적 조합인지는 자료의 특성에 따라 다를 것으로 판단되나, 이상자료나 결측구간에 비하여 정상적이고 신뢰할 수 있는 자료가 다수를 차지하고 있다는 조건을 관측 자료가 만족한고 가정하면, Robust Smoothing 모형에 대한 최적 Span 정보를 추출할 수 있으며, 이 Span 정보로부터 출발하여 보다 작은 의미있는 특성 시간규모 수준까지 단계적으로, 대략 2-3단계로 구분하여 적용하면 대부분의 이상자료를 진단하여 제거할 수 있는 것으로 판단된다. 보다 다양한 환경인자 관측자료 및 다양한 결측구간 및 이상자료 규모에 대한 실질적인 검토 및 모형 성능평가에 대한 연구가 필요할 것으로 사료된다.

5. 결론 및 제언

관측 수온자료의 시간적인 변동양상을 파악하는 Robust Smoothing 기법을 이용하여 어림 성분과 잔차 성분을 추출하고, 추출된 잔차 성분을 이용하여 이상 자료를 탐지·제거하는 모형을 개발하여 적용하였다. 개발된 모형을 Span 조건을 달리하여 적용하는 경우 두드러지는 대부분의 이상 자료가 효과적으로 제거되었다. 또한 방대한 자료의 어림성분 추출과정에서 소요되는 과도한 계산시간은 특정 시간규모에 대한 부분적으로 추출된 자료를 이용하는 방법으로 크게 저감하여 실용적인 적용을 가능하게 하였다.

이상 자료는 관측장비의 빈번하고 지속적인 관리가 곤란한 해양 관측 자료에서는 매우 빈번하게 발생하고 있다. 또한 대부분의 경우, 센서를 이용하여 관측이 수행되기 때문에 측적되는 자료가 매우 방대하기 때문에 수작업으로 이상 자료를 제거한다는 것은 실질적으로 불가능하다. 또한 이상 자료가 제거되지 않은 상태의 자료 분석은 잘못된 결과 또는 왜곡된 결과를 도출할 가능성이 매우 크기 때문에 이상 자료 처리기법은 기법 측면에서도 매우 중요하지만 통계적인 추론, 수학적인 기법 등이 포함되어 모형으로 구성되어 적용되고 있다. 이상 자료는 자료의 특성 및 관측센서의 특성과 직결되어 있기 때문에 다양한 환경자료 하나하나에 대한 가장 효과적인 제거기법을 개발하여 적용할 필요가 있을 것으로 사료된다.

감사의 글

본 연구과제는 환경부지정 전북녹색환경지원센터의 연구비 지원에 의해 수행된 연구과제입니다. 연구비 지원에 감사드립니다.

References

Agresti, A. and Franklin, C. (2007). Statistics, The Art and Science of Learning from Data, Pearson Education Inc.

Barnett, V. and Lewis, T. (1994). Outliers in Statistical Data. Third Edition, John Wiley & Sons.

Basu, S. and Meckesheimer, M. (2007). Automatic outlier detection for time series: an application to sensor data, Knowledge and Information Systems, 11(2), 137-154.

Ben-Gal, I. (2005). Outlier detection, Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researcher (Editors: Maimom, O. and Rockach, L), Chapter 1(1-16), Kluwer Academic Publishers.

Chiang, J-T. (2008). The algorithm for multiple outliers detection against masking and swamping effects, International J. of Contemporary Mathematical Sciences, 3(17), 839-859.

Cho, H.Y. Oh, J. Kim, K.O. and Shim, J.S. (2013). Outlier detection and missing data filling methods for coastal water temperature data, *Journal of Coastal Research*, Special Issue, No. 65, pp.1898-1903.

Cho, H.Y. and Oh J., 2012. Outlier detection of the coastal water temperature monitoring data using the approximate and detailed components, J. of the Korean Society for Marine Environmental Engineering, Technical Note, 15(2), 156-162.

Cleveland, W.S., 1979. Robust locally weighted regression and smoothing scatterplots, J. of the American Statistical Association, 74(368), 829-836.

Hubert, M. and van der Veeken, S. (2008). Outlier detection for skewed data, J. of Chemometrics, Special Issue, 22, 235-246.

Silverman, B.W. (1998). Density Estimation for Statistics and Data Analysis, Chap.3, Chapman & Hall/CRC.

Storch, H.v. and Zwiers, F.W. (1999) Statistical Analysis in Climate Research, Sec. 5.3, Cambridge Univ. Press.

Tsay, R.S. (1988). Outliers, level shifts, and variance changes in time-series, J. of Forecasting, 7, 1-20.

원고접수일: 2014년 8월 30일
수정본채택: 2014년 9월 30일
게재확정일: 2014년 10월 16일