한국해양환경공학회지
Journal of the Korean Society
for Marine Environmental Engineering
Vol. 15, No. 2. pp. 156-162, May 2012

기술보고

어림과 나머지 성분을 이용한 연안 수온자료의 이상자료 감지

조홍연† · 오지희
한국해양연구원 해양환경보전연구부

Outlier Detection of the Coastal Water Temperature Monitoring Data Using the Approximate and Detail Components

Hongyeon Cho† and Jihee Oh
Marine Environment & Conservation Research Department, KORDI, Ansan PO Box 29, Seoul 425-600, Korea

요 약

연안 환경모니터링 사업이 확대되면서 방대하게 축적되어 있는 연안 환경모니터링 자료의 통계적 분석을 위해서는 모니터링 자료에서 빈번하게 발생하는 이상 자료의 감지·처리가 우선적으로 필요하다. 본 연구에서는 연안 환경모니터링 자료의 어림성분과 나머지(또는 잔차)성분을 이용한 이상자료 진단기법을 제안하였다. 주기함수를 이용한 조화분석 방법과 국지 회귀함수추정 방법을 이용하여 각각 어림성분과 나머지성분을 추출한 후, 추출된 나머지성분 자료에 범용적인 Grubbs 검정기법 및 수정표본점수기법을 적용하여 이상자료를 진단 · 제거한 후 이상자료가 제거된 자료로 재구성하는 방법이다. 제안된 이 기법을 국립수산과학원 실시간어장정보시스템 제공하는 연안 수온 연속 모니터링 자료에 적용한 결과 이상자료가 성공적으로 제거되는 양상을 보이는 것으로 파악되었다.

Abstracts – Outlier detection and treatment process is highly required as the first step for the statistical analysis of the monitoring data having many outliers frequently occurred in the coastal environmental monitoring projects. In this study, the outlier detection method using the approximate and detail (or residual) components of the (raw) data is suggested. The approximate and detail components of the data can be separated by the diverse filtering and smoothing methods. The decomposition of the data is carried out by the harmonic analysis and local regression curve, respectively. Then, the Grubbs' test and modified z-score method widely used to detect outliers in the data are applied to the detail components of the water temperature data. The new data set is reconstructed after removed the outliers detected by these methods. It can be shown that the suggested process is successfully applied to the outlier detection of the coastal water temperature monitoring data provided by the Real-time Information System for Aquaculture Environment, National Fisheries Research and Development Institute (NFRDI).

Keywords: outlier(이상자료), approximations and details(어림과 나머지), water temperature monitoring data(수온 모니터링 자료), Grubbs test(Grubbs 방법), modified z-score method(수정표본점수기법), residual(잔차)

1. 서 론

최근 다양한 관측센서를 이용한 연안환경 모니터링 사업이 활발하게 수행되면서 방대한 환경자료가 축적되어 부족한 자료 또는 제한된 자료를 이용하여 수행 · 분석된 과거의 연구 성과가 재검토 · 재해석되어 새로운 현상 및 특성이 발견되고 있다. 과거의 컴퓨터 연산능력 향상이 기여한 과학기술 발달에 버금가는 수준으로 최근에는 관측기술의 발달로 연속 환경모니터링 자료가 축적되면서 새로운 개념으로 해양과학분야의 기술발전에 기여하고 있다.

그러나 방대한 모니터링 자료의 축적은 DRIP(Data rich, but information poor) 현상을 유발하여 종합적이고 체계적인 자료 분석을 통한 정보추출이 제한받고 있다. 이 제한요소 중의 하나는 이상 자료의 처리문제이다.

이상 자료는 기존의 수동(manual) 방식 또는 휴대용 관측 장비

†Corresponding author: hycho@kordi.re.kr

를 이용한 간헐적인 관측에서는 인간의 실수 및 기기 검교정(calibration) 문제 등으로 발생하지만, 관측 장비의 계류에 의한 연속관측 과정에서는 주기적인 센서관리 미흡, 안정적인 전원공급 제한, 센서의 오작동 등의 문제로 빈번하게 발생한다.

관측자료의 개수가 적은 경우에는 확실하게 측정범위를 벗어나는 이상 자료는 간단한 범위 처리과정 및 자료 관리자의 판단을 통하여 제거할 수 있으나, 연속적으로 그리고 매우 짧은 시간간격으로 측정되어 자료의 개수가 기하급수적으로 늘어나는 경우 모든 자료를 하나하나 확인하면서 이상 자료를 수작업으로 제거하는 것은 불가능하게 되어 자동화된 또는 체계화된 처리기법을 필요로 한다.

기본적으로 원(原) 자료(raw data) 분석을 선호하는 전문가는 연속적인 환경 모니터링 자료로부터 정보를 추출하고자 하기 때문에 자료 분석을 위한 전처리과정을 불가피하게 거쳐야 한다. 특히, 대부분의 연속 환경모니터링 자료에서 빈번하게 관찰되는 이상자료를 감지하고 처리하는 과정이 필수적으로 요구된다. 일반적으로 자료를 분석하는 연구자는 각자의 경험과 자료특성을 감안하여 이상자료를 처리하여 왔으나 주관적이고 경험적인 요소가 개입되기 때문에 같은 자료인 경우에도 이상자료를 처리한 자료가 서로 다를 수 있기 때문에 분석결과의 차이가 발생할 수 있다. 따라서 객관적인 측면에서 이상자료를 효과적으로 처리하는 기법에 대한 검토가 요구되고 있다.

본 연구에서는 최근 연안 환경모니터링 사업이 확대되면서 방대하게 축적되어 있는 연안 환경모니터링 자료의 통계적 분석을 위한 전처리 과정중의 하나에 해당하는 이상 자료 감지·처리 기법을 제안하는 것을 목적으로 한다. 환경모니터링 자료 중에서 가장 중요한 인자 중의 하나에 해당하는 수온자료(국립수산과학원 제공자료)를 대상으로 본 연구에서 제안한 기법을 적용하고, 이상 자료 제거 기법을 적용한 경우와 적용하지 않은 경우의 통계정보를 비교분석하여 전처리 기법 적용 효과를 분석하였다.

2. 이상 자료의 기본

이상 자료(outlier)는 어떤 자료가 그 자료를 제외한 나머지 자료와 일관성이 없어(inconsistent) 보이는 자료 또는 분명하게 다른(distinctly different) 독특한 특성을 가진 자료로 정의되기도 하고(Barnett & Lewis[1994]; Hair et al.[2010]), 비정상적으로(unusually) 크거나 작은 자료, 극한 자료(extreme value)로 정의되기도 한다(Agresti & Franklin[2007]; Martinez & Martinez[2005]). 따라서 이상 자료는 자료의 통계정보를 왜곡할 수 도 있기 때문에 이상 자료에 대한 정량적인 사전 검토가 필요하다. 이상 자료는 잘못된 자료와 특이한 자료로 구분할 수 도 있다. 모두 판단이 필요하지만, 잘못된 자료는 제거하여야 하며, 특이한 자료는 별도로 처리하거나 제외하여 통계분석을 수행할 수 있도록 표기(marking)하여 관리할 필요가 있는 자료이다.

한편 언어를 이용한 정의와 더불어 이상 자료에 대한 통계적인 기준도 구체적으로 제시되고 있다. 가장 기본적인 정의는 정규분포 또는 기준이 되는 어떤 분포를 가정하고, 평균(m)과 표준편차(SD)의 함수로 정의되는 영역을 벗어나는 자료로 정의한다. 예를 들면, Hair 등[2010]은 표본의 개수가 80개 정도 또는 그 이하에 해당하는 소표본의 경우와 그 이상에 해당하는 대표본의 경우를 구분하여 다음과 같이 이상자료를 정의하고 있다.

소 표본 : m±2.5(SD) 영역을 벗어나는 자료
대 표본 : m±4.0(SD) 영역을 벗어나는 자료

Grubbs[1950] 및 Dixon[1950] 등도 신뢰수준을 포함한 이상 자료 판단을 위한 각각의 통계기준을 제시하고 있다(Garcia[2012]). 기본적으로 이상 자료의 판단기준은 평균을 중심으로 일정 한계범위를 벗어나는 자료를 이상 자료로 간주하는 개념에 기초하고 있다. Grubbs 방법은 Extreme Studentized Deviate($Max(|(x_i-\bar{x})/\sigma_x|), \bar{x}, \sigma_x$=각각 자료 x_i의 평균 및 표준편차) 수치를 한계수치와 비교하여 판단하는 방법이며, Dixon 방법은 자료를 정렬하여 전체 구간에 대한 부분비율 수치를 계산하여 판단하는 방법이다. 수정 표본점수방법은 z-score 계산($z = (x_i-\bar{x})/\sigma_x$)과정에서 표준편차 대신에 MAD(median absolute deviation about the median, $\tilde{x}$) 수치로 계산한 z-score ($z = 0.6745\ (x_i-\tilde{x}\ )$/MAD)수치를 이용하여 이상자료를 판단하는 방법으로 개념에 차이가 있다.

3. 이상 자료 감지기법

이상 자료 진단·처리기법은 자료의 종류만큼이나 다양하다. 다양한 분야에서 다양한 이상 자료 진단 방법이 제시되고 있으나, 모든 자료에서 적용되는 방법은 어떤 특정한 자료보다는 정규분포를 따르는 독립적인 자료조건을 충족하는 경우의 진단방법이 유일하며, 가장 활발하게 연구가 추진되어 그 기준도 매우 유사한, 정립된 단계에 해당한다고 할 수 있다. 따라서 특정한 분야의 특정한 자료에 국한되어 사용되는 매우 복잡한 통계도구 및 모형을 이용한 방법보다는 본 연구 분야 또는 특정 연구 분야에서 분석하고자 하는 자료로부터 범용적인 이상 자료 진단기준을 적용할 수 있는 자료를 추출하는 과정을 추가하여, 추출된 자료를 이용하여 이상 자료를 판단하는 기법이 체계적인 접근방법이라고 판단되며, 활용범위의 확장도 가능할 것으로 판단된다.

따라서 본 연구에서 제안하는 방법은 모니터링 자료로부터 어림성분과 나머지 성분을 추출하여, 나머지 성분을 대상으로 범용적으로 이용되는 Grubbs 검정기법(Grubbs[1950]) 등을 이용하여 이상자료를 판단하는 방법이다. 본 연구에서 제안하는 처리 기법은 실질적인 상황과 통계기법을 조합한 방법으로 다음과 같은 단계로 구성되어 있다.

제1단계: 가시적 감지단계(Visual Detection Process)

연안에서 연속적으로 측정되는 환경인자는 각각의 상식적인 또는 제한적인 범위를 가지고 있다. 따라서 개략적인 또는 한정된 범위

제시에 의한 방법으로 터무니없는 이상 자료를 쉽게 진단·제거할 수 있다. 물리적으로 무의미한 값(범위)이 발생(DO 농도 또는 오염물질 농도의 경우 음수가 발생하거나, pH 농도가 0~14 범위를 벗어나는 경우; 수온이 영하 10도 이하 또는 40도 이상인 경우 등)하거나 발생가능성이 거의 없는 경우의 조건을 제시하여 이상 자료를 제거하는 가장 초보적인 단계의 이상 자료 제거 기법이다. 이 방법은 관측 자료를 도시하는 경우, 매우 크거나 작은 값의 이상 자료가 포함되는 경우 도시범위의 확장으로 정상범위가 축소되어 자료변화 양상의 도식적인 판별이 곤란하게 된다.

제2단계: 자료를 어림과 나머지 성분으로 구분하는 과정(Smoothing Process)

시계열 자료에서 바로 이상 자료를 제거하는 연구가 활발하게 수행되고 있으나, 자료가 가지는 구조적인 특성이 관측항목별로 서로 상이하기 때문에 적용에 제한이 따른다. 본 연구에서는 시계열자료가 아닌 이론적으로 IID(independent, identically distributed) 조건을 따르는 자료에 대한 다양한 이상 자료 제거기법이 활발하게 이용되고 있기 때문에 IID 조건에 유사한 자료를 도출하기 위한 과정으로 관측된 시계열 자료를 전체적인 변화양상을 표현하는 어림(approximate, smooth) 성분과 나머지(잔차, residual) 성분으로 구분하기 위한 과정이다. 이 방법의 적용단계에서는 아직 이상 자료가 제거되지 않은 상태이기 때문에 Robust 기법 적용이 필요하다. 본 연구에서는 아래에 제시된 방법을 이용하여 어림 성분과 나머지 성분을 각각 추출하였다.

(1) Robust Smoothing 방법(RLOESS)

어림 성분을 추정하는 방법은 자료 Smoothing 과정으로, 자료의 변화양상을 국지적으로 가중치를 부여하여 적절한 함수곡선으로 맞추어가는 LOESS 또는 LOWESS (locally weighted regression procedure) 방법이 널리 이용되고 있으며, 이상자료의 영향을 줄이기 위한 Robust LOESS, 즉 RLOESS 방법도 있다(Martinez and Martinez[2005]). 본 연구에서는 이상자료의 영향을 줄이기 위한 RLOESS 방법을 이용하여 수온자료의 어림성분을 추정하였다.

따라서 본 연구에서는 이상자료를 적절하게 감지하기 위해서는 이상자료의 영향을 줄이기 위한 모형 매개변수의 Robust 추정이 필요하다(Rousseeuw & Leroy[2003]).

(2) Harmonic Analysis 방법(HA)

조화분석은 조석의 성분분석에 널리 이용되는 방법이나, 체계적으로 주기성분을 고려한 기온 및 수온자료의 어림 성분 추정으로도 널리 제안·이용되고 있다(Cho et al.[2010]). 이 방법은 조석에 의한 영향이 우세하지 않고, 연 변화 및 계절변화 또는 그 이하의 변동성분이 포함되어 있는 환경인자 및 기상인자의 어림 성분 추정에 유용한 방법이다. 본 연구에서는 뚜렷한 연 변화 양상을 가지는 수온 성분의 어림 성분추정으로 이 방법을 이용하였다.

제3단계: 나머지 성분의 이상 자료 감지·제거 과정

제 2단계에서 추출한 나머지 성분을 대상으로, 기존에 제시된 기본적인 이상 자료 감지 기법을 적용하여 이상 자료를 추출하였다. 이상 자료의 제거 여부는 연구자의 경험에 의존하여야 한다. 본 연구에서는 매우 전통적이고 널리 이용되고 있는 Grubbs 진단기법(95% 유의수준)을 이용하여 나머지 성분의 이상 자료를 진단하였다. 이상 자료로 진단된 자료는 모두 제거하였다. 그러나 전체적인 자료의 분포양상을 해석하는 경우에는 큰 문제가 없을 것으로 판단되나, 극치해석 등을 수행하는 경우에는 이상 자료로 진단되어도 특이한 자료인지, 잘못된 자료인지를 판단하여 처리여부를 결정하여야 한다.

4. 이상 자료 감지기법의 적용

4.1 연안 수온자료(실시간 어장정보 시스템 자료)

국립수산과학원에서는 어업활동에 필요한 어장환경정보 제공 및 수산업진흥을 위한 기반자료 구축을 목적으로 연안의 양식어장 및 어장재해가 빈발한 해역에 실시간 해양환경정보(수온, 염분, DO 농도 등) 자동관측시스템을 구축하여 운영하고 있다(국립수산과학원[2012]; 관측지점은 Fig. 1 참조). 본 연구에서는 비교적 장기간의 자료가 가용한 상태에 있는 백령도, 완도(청산), 영덕(거무역) 지점의 수온 자료를 대상으로 본 연구에서 제안한 이상 자료 감지기법을 적용하였다. 관측 자료는 30분 또는 1시간 간격으로 제공되고 있으며, 관측지점에 따라 이상 자료 및 결측자료(missing data), 관측기간이 크게 차이가 나고 있다.

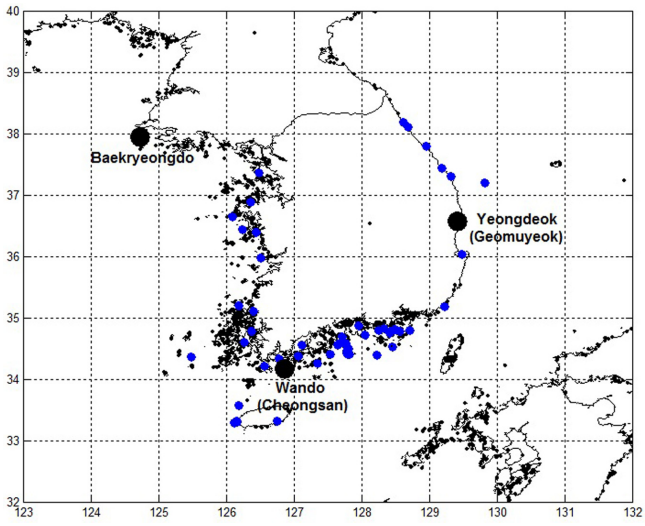

Fig. 1. Location map of the real-time monitoring system of the aqua-culture information (big solid circles : data locations used in this study; Coastline revised from World Vector Shoreline (designed for 1:250,000) data set available through the U.S. National Geophysical Data Center).

4.2 이상 수온자료 감지

국립수산과학원 자료포털에서 다운로드한 자료 그대로를 원자료(raw data)로 가정하고, 제3절에서 제시한 단계를 따라 이상자료를 감지하였다. 제1단계는 가장 기본적인 범위지정에 의한 이상 자료 제거 단계로 범위지정에 의한 이상자료 제거 전·후의 비교 그림이다(Fig. 2 참조). 범위는 하한은 -5 ℃, 상한은 30 ℃ 조건을 사용하였다. 이 범위지정은 간단하게 변경할 수 있으며, 자료의 변동 범위를 감안하여 경험적으로 지정하면 된다. 완도(청산) 지점에서 보이는 바와 같이 정상적인 범위를 크게 벗어나는 잘못된 1~2개 정도의 자료로 인하여 전체적인 자료의 변화범위(-40~160 ℃)가 크게 증가되어 가시적인 수온의 변화양상 파악을 어렵게 하고 있다. 잘못된 자료를 제거하여 자료의 변화구간이 0~30 ℃ 영역으로 제한되는 경우 자료의 변화양상을 시각적으로 쉽게 그리고 보다 뚜렷하게 판단할 수 있다.

범위 지정에 의한 이상 자료를 제거한 후의 과정은 제1단계를 통과한 자료를 이용하여 어림(approximation) 성분과 나머지(잔차) 성분으로 자료를 구분하는 과정이다. 이 과정은 제2단계에서 제시한 RLOESS(Robust LOESS; LOESS=locally weighted regression procedure for fitting a regression curve by smoothing) 방법(국지영역을 지정하는 변수, SPAN=1%)과 조화분석(harmonic analysis, HA) 방법(주기성분은 12개 : 1년 주기부터 1/2년 주기, 1/3년 주기, ..., 1/12년 주기성분까지 이용)을 이용한 자료 변동양상의 근사과정을 통한 어림성분 추출과정과 제1단계를 통과한 자료에서 어림성분을 제외한 나머지 성분추출과정으로 구성된다. 여기서 추출된 나

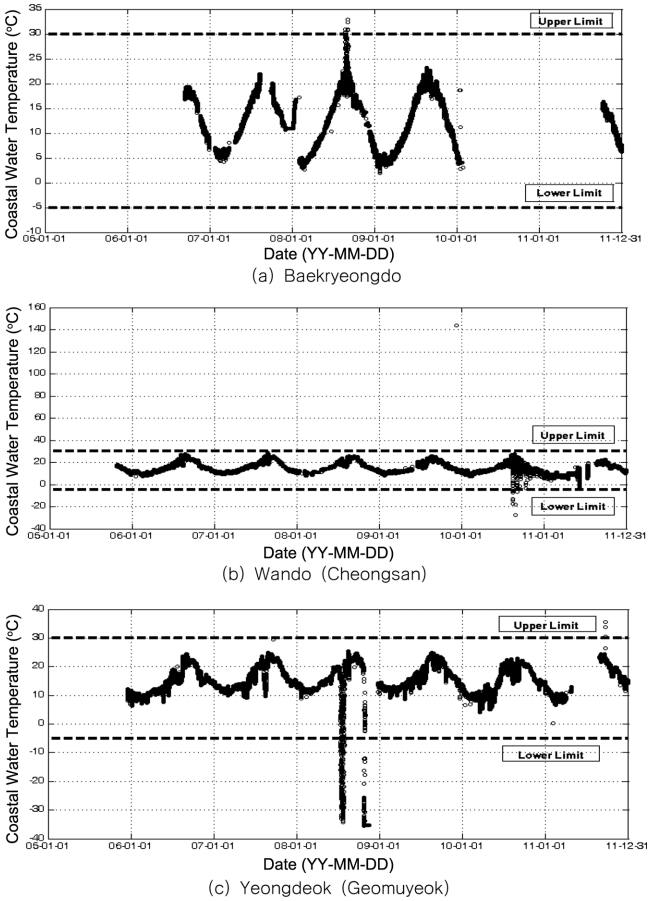

Fig. 2. Outlier detection by the upper and lower limits setting method.

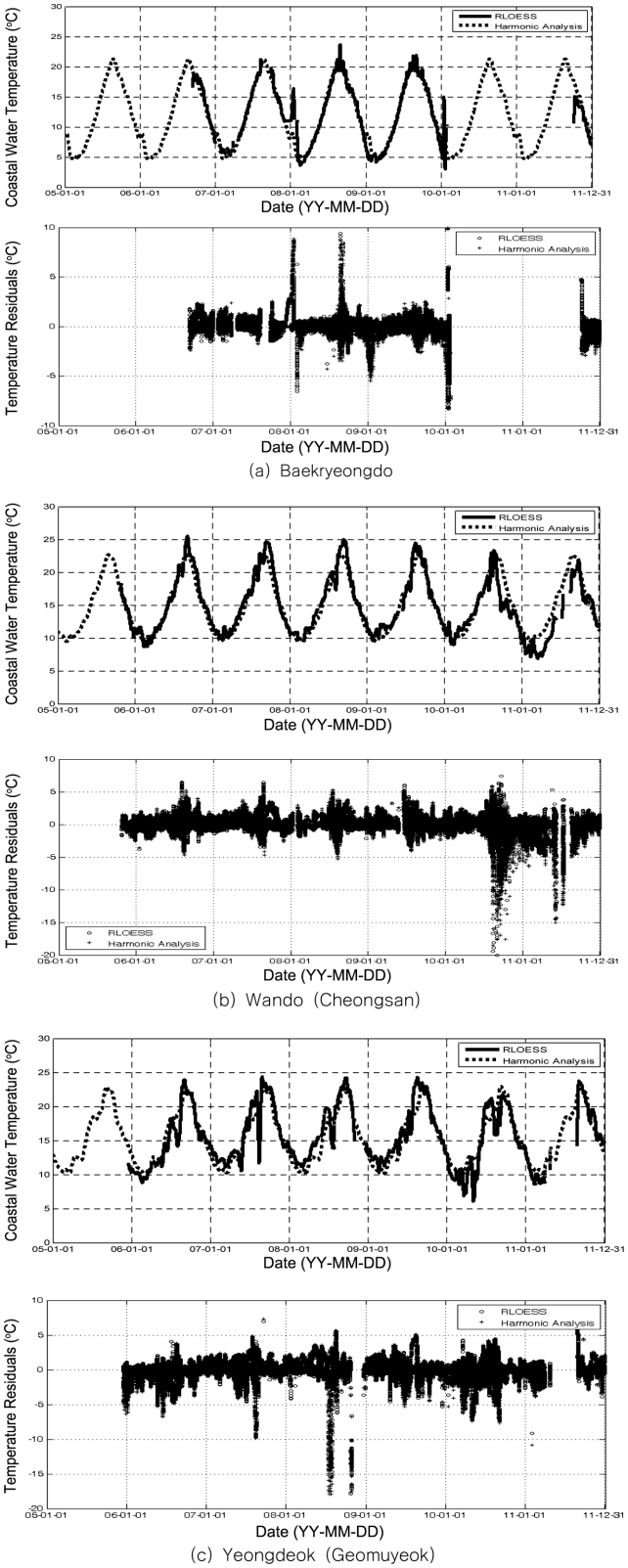

Fig. 3. Time-series plots of the approximate and detail components.

머지 성분을 대상으로 이상 자료 여부를 통계적으로 검정하였다(제3단계). 연안의 수온 변화는 조석의 영향이 중요할 수 있기 때문에 조석성분을 포함하여 조화분석을 수행하였으나, 본 연구영역에

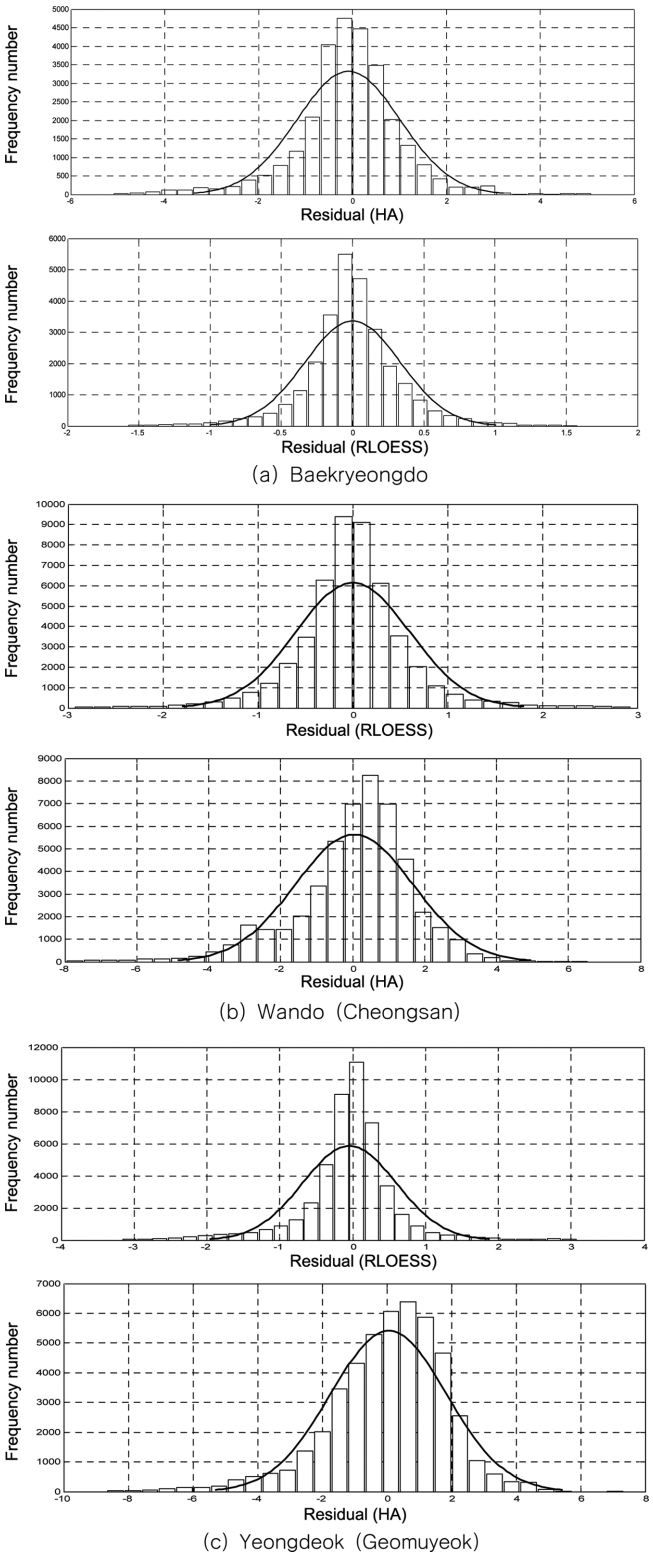

Fig. 4. Histogram of the detail components (Solid line : Normal distribution).

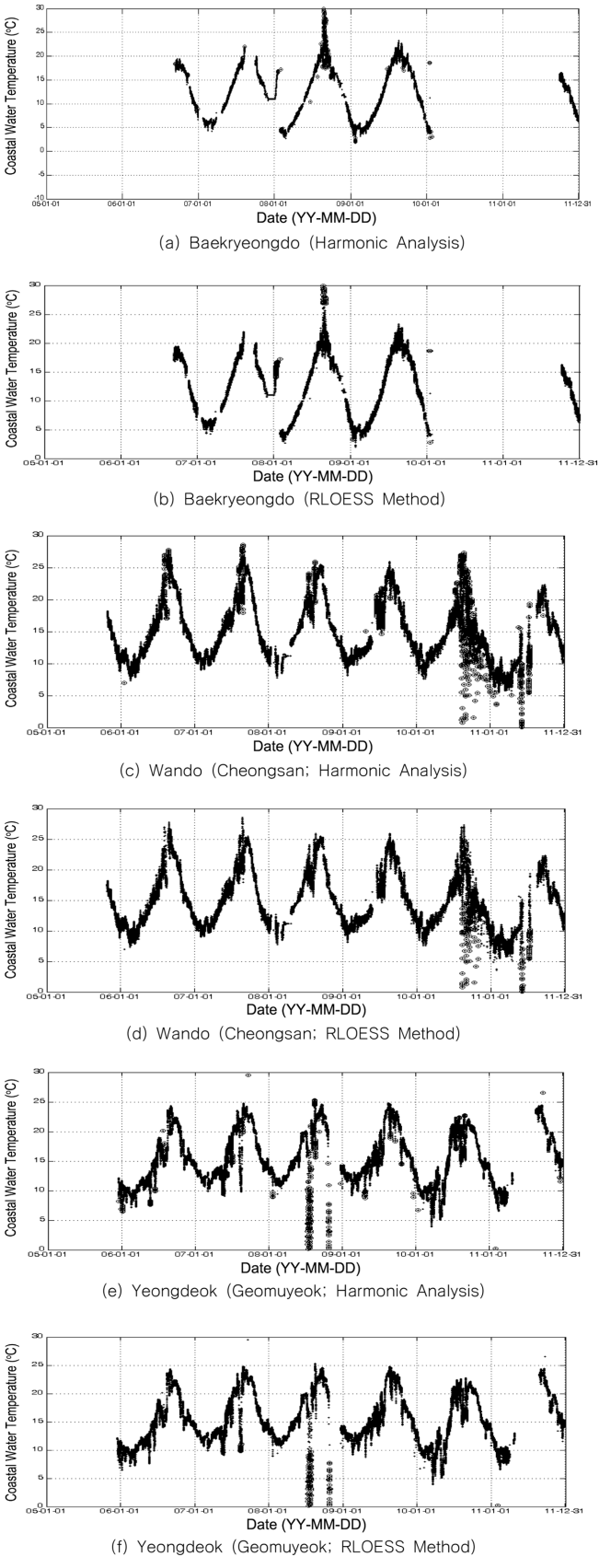

Fig. 5. Outlier and data plots using the harmonic analysis and RLOESS method.

서는 그 영향이 1년 및 1/2년 주기정도의 장주기 성분에 비하여 매우 미미하여 조석성분은 무시하였다.

각각의 방법을 적용하여 추출한 어림 성분과 나머지 성분은 다음과 같다(Fig. 3 참조). 어림성분은 작은 시간규모의 변동성분이 제거되어 보다 평활화(smoothing)된 변화양상을 보이고 있으며, 나머지 성분은 무작위적으로 변동하는 양상을 보이고 있음을 알 수 있다. 나머지 성분은 독립적이고, 정규분포를 따르는 조건을 만족하여야 한다. 제2단계에서 추출된 나머지 성분의 빈도분포와 정규분포를 비교하여 도시하였다(Fig. 4 참조). 그 결과, 나머지 성분이 정규분포와 어느 정도 일치하는 양상을 보이고 있으며, 적절한(robust) 분산매개변수를 선정하는 경우 보다 정규분포에 근사한 분포로 간주할 수 있다고 판단된다. 나머지 성분자료의 분포함수가 반드시 정규분포를 따라야 할 필요는 없고 따르지 않는 경우도 빈번하나, 정규분포에서 크게 벗어나는 경우에는 본 연구에서 사용한 방법에 제한이 따를 것으로 판단된다.

다음은 제3단계에 해당하면서, 가장 핵심이 되는 나머지 성분의 이상자료 감지·제거과정이다. 제2단계의 RLOESS 방법과 HA 방법을 이용하여 추출된 나머지 성분을 Grubbs 기법과 수정 표본점수(z-score) 방법, Dixon 방법 등 범용적으로 이용되는 방법을 이용하여 이상자료 감지에 적용하였으나, 그 감지 결과 차이는 미미하여 Grubbs TEST 적용결과를 대표로 제시하였다. 나머지성분에서 이상자료를 감지·제거하고, 어림 성분을 합하여 이상자료가 제거된 원자료를 재구성하였다. 이상 자료 제거효과는 이상자료를 심볼(⊙)로 표시하여 파악할 수 있도록 도시하였다(Fig. 5 참조).

4.3 이상 자료 제거 영향 분석

이상 자료를 제거한 경우, 자료의 통계적인 정보변화를 비교·분석하였다. 간단한 통계정보를 정리하였으며, 각각의 단계에 따라 본 연구에서 이용한 수온자료의 평균, 표준편차, 중간값, 평균절대편차 등의 통계정보의 변화를 표에 제시하였다(표 1 참조). 표에서 보이는 바와 같이, 영덕(거무역) 지점의 경우 -5 ºC 이하의 수온자료가 1,000개 정도 포함되어 있어 표준편차가 제1단계 범위지정 제거과정 이전·이후에 각각 6.96 ºC, 4.50 ºC로 그 차이가 매우 크게 나타났으나, Robust 추정편차에 해당하는 MAD, Median 변화는

Table 1. Basic statistical information changes of the data before and after outlier removals (SD=standard deviation, MAD=median absolute deviation about the median; R=RLOESS Method, H=Harmonic Analysis Method; BOR, AOR=before and after outlier removals, respectively)

(a) Baekryeongdo

Data type	size	mean	SD	median	MAD
raw data	28,217	12.28	5.09	12.03	4.39
data after step 1	28,209	12.27	5.08	12.02	4.39
residual (R/BOR)	28,209	0.01	0.61	0.00	0.29
residual (H/BOR)	28,209	0.00	1.37	-0.05	0.85
residual (R/AOR)	27,765	0.00	0.34	0.00	0.23
residual (H/AOR)	27,910	-0.08	1.10	-0.06	0.78
data after step 3 (R)	27,765	12.24	5.01	11.97	4.33
data after step 3 (H)	27,910	12.22	5.06	11.89	4.37

(b) Wando (Cheongsan)

Data type	size	mean	SD	median	MAD
raw data	49,669	14.87	4.63	14.06	3.76
data after step 1	49,634	14.87	4.53	14.07	3.75
residual (R/BOR)	49,634	-0.01	0.91	0.00	0.48
residual (H/BOR)	49,634	0.00	1.78	0.28	1.25
residual (R/AOR)	48,911	0.00	0.60	0.00	0.41
residual (H/AOR)	49,433	0.04	1.64	0.28	1.21
data after step 3(R)	48,911	14.84	4.47	14.04	3.70
data after step 3 (H)	49,433	14.90	4.52	14.10	3.74

(c) Yeongdeok (Geomuyeok)

Data type	size	mean	SD	median	MAD
raw data	48,147	14.83	6.96	14.62	4.30
data after step 1	47,460	15.45	4.50	14.71	3.79
residual (R/BOR)	47,460	-0.09	1.07	0.00	0.50
residual (H/BOR)	47,460	0.00	2.01	0.25	1.43
residual (R/AOR)	46,905	-0.04	0.64	0.00	0.42
residual (H/AOR)	47,245	0.06	1.79	0.26	1.36
data after step 3 (R)	46,905	15.48	4.43	14.72	3.75
data after step 3 (H)	47,245	15.50	4.45	14.75	3.76

상대적으로 미미하였다. 이상자료의 개수가 많지 않은 경우, 완도(청산) 및 백령도 지점의 경우 및 범위지정에 의한 이상자료 제거(제1단계) 과정 이후에는 전체적인 통계정보 변화에 미치는 영향은 크지 않은 것으로 파악되었다. 이상자료 제거를 통한 표준편차 및 MAD 감소경향은 미미한 수준이나 예상할 수 있는 바와 같이 감소하는 경향을 보였다. 한편 HA 방법에 의한 방법은 RLOESS 방법보다 큰 표준편차를 보이고 있는 것으로 추정되었으나, 이는 각각의 방법의 영향보다는 기법의 적용을 위한 매개변수(근사성분의 개수 및 Robust 추정구간의 범위)의 영향으로 판단된다.

이상자료가 적절하게 제거되었는가하는 정량적인 판단은 곤란하지만, 이러한 판단이 통계적인 의미의 "가장 그럴듯한(most likely)" 진단 관점에서 보면, 자료도시에서 보이는 눈에 거슬리는 이상자료는 상당부분 제거되어 본 연구에서 제안한 방법이 성능을 발휘하고 있는 것으로 판단된다. 또한, 이상자료 제거 전·후의 나머지 성분의 표준편차 변화를 보면, 이상자료 제거전보다 이상자료 제거 후에 표준편차가 감소하게 되는 예측가능한 양상을 보이고 있어, 본 연구에서 제시한 이상자료 제거방법의 효과가 타당함을 보여준다고 할 수 있다.

5. 결론 및 제언

본 연구에서는 연안 환경모니터링 자료의 어림성분과 나머지성분을 이용한 이상자료 진단기법을 제안하였다. HA 방법과 RLOESS 방법을 이용하여 어림성분과 나머지성분을 추출한 후, 추출된 나머지성분 자료에 Grubbs 검정기법 및 수정표본점수 방법을 적용하여 나머지성분을 진단·제거한 후 이상 자료가 제거된 자료를 재구성하였다. 제안된 이 기법을 국립수산과학원에서 제공하는 연안의 수온 연속 모니터링 자료에 적용한 결과, 이상자료가 성공적으로 제거되는 양상을 보이는 것으로 파악되었으며, 본 기법의 적용성능이 우수한 것으로 파악되었다. 영덕(거무역) 지점의 경우 -5 ℃ 이하의 수온자료가 1,000개 정도 포함되어 있어 표준편차가 제1단계 범위지정 제거과정 이전·이후에 각각 6.96 ℃, 4.50 ℃로 그 차이가 매우 크게 나타났으나, Robust 추정편차에 해당하는 MAD, Median 변화는 상대적으로 미미하였다. 이상자료의 개수가 많지 않은 경우 완도(청산) 및 백령도 지점의 경우 및 범위지정에 의한 이상자료 제거(제1단계) 과정 이후에는 전체적인 통계정보 변화에 미치는 영향은 크지 않은 것으로 파악되었다. 그러나 이상자료는 통계정보의 편이(bias)나 왜곡을 유발할 수 있기 때문에 연안 모니터링 자료의 통계적인 분석을 위해서는 반드시 검토하여 처리하여야 한다.

통계적인 의미의 "가장 그럴듯한(most likely)" 진단 관점에서 보면, 자료도시에서 보이는 눈에 거슬리는 이상자료는 상당부분 제거되어 본 연구에서 제안한 방법이 성능을 발휘하고 있는 것으로 판단된다. 한편 이상자료와 더불어 통계적인 정보의 편이(bias)를 유발하는 결측자료 보충(missing data filling-in or imputation) 등의 처리기법에 대한 연구도 수행되어져야 할 것으로 판단된다.

감사의 글

본 연구는 한국해양연구원 기본연구사업(PE98743)의 지원을 받아 수행되었습니다. 연구비 지원에 감사드립니다. 또한 본 연구에서 사용한 어장환경정보시스템 자료를 제공해주신 국립수산과학원에 감사드립니다.

참고문헌

[1] 국립수산과학원, 2012, 실시간 어장정보시스템. http://portal.nfrdi.re.kr/risa/.

[2] Agresti, A. and Franklin, C., 2007, Statistics, The Art and Science of Learning from Data, Pearson Education, Inc. pp.693.

[3] Barnett, V. and Lewis, T., 1994, *Outliers in Statistical Data*, Third Edition, John Wiley & Sons, Ltd., Chichester, UK, pp.584.

[4] Cho, H.Y., Suzuki, K. and Nakamura, Y., 2010, Hysteresis loop model for the estimation of the coastal water temperatures, -by using the buoy monitoring data in Mikawa Bay, Japan-, Report of the Port and Airport Research Institute, 49(2), pp.123-153.

[5] Dixon, W.J., 1950, Analysis of Extreme Values, The Annals of Mathematical Statistics, 21(4), pp.488-506.

[6] Garcia, F.A.A., 2010, Tests to identify outliers in data series, http://www.se.mathworks.com/matlabcentral/fileexchange/28501, MATLAB Central File Exchange. Retrieved January 19th, 2012.

[7] Grubbs, F.E., 1950, Sample Criteria for Testing Outlying Observations, The Annals of Mathematical Statistics, 21(1), pp.27-58.

[8] Hair, J.F. Jr., Black, W.C., Babin, B.J. and Anderson, R.E., 2010, *Multivariate Data Analysis, A Global Perspective*, Seventh Edition, Chapter 2, Pearson Education, Inc., New Jersey, USA, pp.800.

[9] Martinez, W.L. and Martinez, A.R., 2005, Exploratory Data Analysis with MATLAB, Computer Science and Data Analysis Series, Chapman & Hall/CRC. pp.405.

[10] Rousseeuw, P.J. and Leroy, A.M., 2003, Robust Regression and Outlier Detection, John Wiley & Sons. pp.329.

2012년 1월 19일 원고접수
2012년 4월 3일 심사수정일자
2012년 4월 10일 게재확정일자