이상치 감지

안 태 진 (안성산업대학교 토목공학과 교수)

Q 수문자료에서 이상치를 찾는 이유와 감지방법은 어떤 방법이 있습니까?

A 수문자료계열에서 자료의 변동이나 경향은 크게 자연적인 형상과 인위적인 변화에 의하여 발생하게 된다. 수집된 수문자료계열에서 자료의 일반적인 경향(trend)에 상당히 벗어나는 극치(extreme data)를 이상치라 한다. 통계학에서 이상치는 대다수 관측치에서 상당히 벗어나 있는 관측치를 의미하며, 표본자료의 축척(scale)을 고려할 때 표본자료의 위치(location)로부터 멀리 떨어져 있다. 수문자료계열에서 이상치의 포함 또는 제거는 통계학적 매개변수의 크기에 영향을 미치며 특히 표본자료의 개수가 적을 경우는 더욱 그렇다. 이상치를 포함한 수문자료는 부적절한 통계학적 매개변수를 추정하게 되고 설계 수문량의 불확실성을 초래한다. 따라서 미국수자원위원회(1981)는 수문자료에서 이상치를 감지하기 위하여 일반적인 빈도해석과 유사한 공식을 제안하였다. 이 식에서 빈도계수와 유사한 계수는 10% 유의수준에 관하여 표본자료의 개수에 따른 값을 추천하였으며 이상치 검정을 위한 상한 이상치(high outlier)와 하한 이상치(low outlier)를 계산하도록 하였다. 일반적으로 표본자료로부터 이상치를 제거하면 표본자료의 정규분포 특성은 개선되며, 만약 표본자료의 최소치가 이상치로 판별되어 제거하고 빈도해석을 실시하면 재현기간이 긴 경우의 수문량은 작아진다. 미국수자원위원회는 수문자료계열의 왜곡도계수가 0.4보다 크면 상한 이상치를 먼저 고려하고 왜곡도계수가 -0.4보다 작으면 하한 이상치를 고려하며 왜곡도계수가 ±0.4내 있으면 상한 및 하한 이상치 검정을 동시에 실시하도록 추천하고 있다.

McCormick와 Rao(1995)는 Gumbel 분포방법과 최소중간치자승법(least median of squares method, LMS)으로 홍수량 수문계열의 이상치를 감지하였다. LMS 방법은 수문자료의 극치를 과대평가하지 않는 robust 방법으로 알려져 있으며 잔차 제곱들의 중간치를 최소화하는 방법이다. 이 방법은 작성된 잔차도에서 초기축척과 각 관측치에 가중치를 부여한 후, 최종축척을 계산한 값으로 이상치를 감지하는 범위를 설정하며, 수문량이 설정된 범위를 벗어나면 이상치로 간주된다. LMS 방법을 적용할 때 선형방정식의 변환변수(reduced variate)는 California 공식으로 계산한 재현기간을 이용하였고 확률홍수량은 Gumbel 분포로 추정하였다. 또한 LMS 방법은 확률홍수량에 포함된 이상치뿐만 아니라 변환계수내의 이상치에도 둔감한 방법으로 알려져 있다. Rousseuw와 Leroy(1987)는 LMS와 이상치를 감지하는 알고리즘을 정립하였고, 그 해석 프로그램을 개발하여 보급하고 있다. McCormick와 Rao는 이 방법을 실제수문자료에 적용한 결과, 미국수자원위원회 방법으로는 감지할 수 없는 이상치를 용이하게 감지할 수 있었다고 보고하였다.

〈참고문헌〉

McCormick, D. L. and A. R. Rao (1995), Outlier Detection in Indiana Flood Data, Tech. Rept. CE-EHE-95-4, School of Civil Engineering, Purdue Univ., W. Lafayette, IN 47907.

Rousseeuw, P. J. and Leroy, A. M. (1987), Robust Regression and Outlier Detection, Wiley-Interscience, New York.