Imputation Method를 활용한 수문 결측자료의 보정

Filling in Hydrological Missing Data Using Imputation Methods

강태호*, 홍일표**, 김영오***

Tae-Ho Kang, Il-Pyo Hong, Young-Oh Km

요 지

과거 관측된 수문자료는 분석을 통해 다양한 수문모형의 평가 및 예측과 수자원 정책결정에서 활용된다. 하지만 관측장비의 오작동 및 관측범위의 한계에 의해 수집된 자료에는 결측이 존재한다. 단순히 결측이 존재하는 벡터를 제외하거나, 결측이 존재하는 자료 구간에 선형성이 존재한다는 가정 하에 평균을 활용하기도 했으나, 이로 인하여 자료의 통계특성에 왜곡이 야기될 수 있다. 본 연구는 결측의 보정으로 자료가 보유하는 정보의 손실 및 왜곡을 최소화 할 수 있는 방안을 연구하고자 한다. 자료의 결측은 크게 완벽한 무작위 결측(missing completely at random, MCAR), 무작위 결측(missing at random, MAR), 무작위성이 없는 결측(nonrandom missingness)으로 분류되며, 수문자료는 결측을 포함한 기간이 그 외 기간의 자료와 통계적으로 동일하지는 않지만 결측자료의 추정이 가능한 MAR에 속하는 것이 일반적이므로 이를 가정으로 결측을 보정하였다. Local Lest Squares Imputation(LLSimput)을 결측의 추정을 위해 사용하였으며, 기존에 쉽게 사용되던 선형보간법과 비교하였다. 적용성 평가를 위해 소양강댐 일 유입량 자료에 1 - 5 %의 결측자료를 임의로 생성하였다. 동일한 양의 결측자료에 대해 100개의 셋을 사용하여 보정의 불확실성 범위를 적용된 방법에 대해 비교·평가하였으며, 결측 증가에 따른 보정효과의 변화를 검토하였다. Normalized Root Mean Squared Error(NRMSE)를 사용하여 적용된 두 방법을 평가한 결과, (1) 결측자료의 비가 낮을수록 간단한 선형보간법을 사용한 보정이 효과적이었다. (2) 하지만 결측의 비가 증가할수록 선형보간법의 보정효과는 점차 큰 불확실성과 낮은 보정효과를 보인 반면, (3) LLSimpute는 결측의 증가에 관계없이 일정한 보정효과 및 불확실성 범위를 나타내는 것으로 드러났다.

핵심용어 : Local Lest Square Imputation, 선형보간법, 수문 결측자료

1. 서론

추가적인 보정 과정이 적용되지 않은 기상 및 수문 관측자료는 관측의 실패, 관측장비 오작동, 관측 범위의 한계에 의해 결측 및 이상치를 포함하게 된다. 수문 및 기상현상의 분석 및 예측 모형의 보정을 위해 요구되는 관측자료를 이상치의 보정 과정 없이 사용하게 된다면 정확성 및 신뢰도가 낮은 연구결과를 피하기 어렵게 된다. 따라서 본 연구는 수문자료의 보정에 대해 Kim et al. (2005)이 타 분야에서 제안한 LLSimpute/PC와 LLSimpute/L2 기법의 적용가능성을 검토하여 정보의 손실 및 왜곡을 최소화 할 수 있는 새로운 방안을 제시하고자 한다.

2. 연구방법

* 한국건설기술연구원 수자원연구부 연구원·E-mail : kangth@kict.re.kr
** 정회원·한국건설기술연구원 수자원연구부 책임연구원·E-mail : iphong@kict.re.kr
*** 정회원·서울대학교 건설환경공학부 부교수·E-mail : yokim05@snu.ac.kr

본 연구는 Kim et al. (2005)의 LLSimpute/PC와 LLSimpute/L2 기법을 1974년부터 2006년까지의 소양강댐 일별 관측유입량 보정을 위해 적용하였다. 결측자료의 보정 정확성을 판단하기 위해 결측자료가 존재하지 않는 일 유입량자료에 임의로 결측자료를 생성하여 보정·평가하였다.

2.1 결측자료의 생성

자료의 결측은 크게 완벽한 무작위 결측(missing completely at random, MCAR), 무작위 결측(missing at random, MAR), 무작위성이 없는 결측(nonrandom missingness)으로 분류된다(Kim et al., 2005). 수문학적인 모델링에서 대부분의 결측자료는 MAR의 범주에 속하므로, 본 연구에서는 결측자료가 MAR에 속하는 것으로 가정하여 결측을 생성하였다. 결측의 생성을 위해 우선적으로 0부터 1사이의 값을 가지는 uniform 분포로부터 난수를 발생시킨다. 이 값은 0과 1 사이의 값을 가지게 되므로 결측자료의 생성을 목적으로 하는 범위에 포함되는 자료의 수를 곱해주게 되면 공간적으로 결측자료의 위치에 대한 정보를 얻을 수 있게 된다.

Random number * Total number of data = Location of Missing data (1)
0 < Random number =< 1

2.2 전가방법

전가방법(imputation method)의 기본적이 아이디어는 자료의 행렬에서 결측의 추정을 위해 기록된 다른 행의 결측되지 않은 부분과 일종의 유사성을 사용하는 것이다. 그러므로 결측은 유사한 벡터를 사용한 선형 조합으로 추정된다. 유사한 벡터를 구분하기위한 가장 흔한 방법은 Euclidean norm, correlation, variance minimization이다. Troyanskaya et al. (2001)는 Euclidean norm을 유사성을 고려하기 위한 적절한 방법으로 제안하였다. 선택된 벡터들은 상관성이 높은 순으로 정리되며, 열에서 결측된 값들과 결측되지 않은 값들 사이의 선형 회귀를 찾기 위한 최적의 k를 결정하여 k번째까지의 벡터를 정리한다. 예를 들어, 행 g_1에서 첫 번째 자료가 결측이라면, 결측되지 않은 전체의 행렬에서 벡터 g_1과 유사한 벡터들을 찾는다. 이러한 k개의 유사한 벡터에 기초하여 결측자료를 예측할 수 있다. 벡터 g_1과 k번째 가까운 벡터는 다음의 식과같이 정의되며, 식에서 q는 벡터의 길이를 나타내는 것으로 각각의 행에 대해 결측자료의 수에 따라 다른 값을 가진다.

$$g_i^T \in \Re^{k\times(q-1)},\ 1 \le i \le k \tag{2}$$

$$\mathrm{A} \in \Re^{k\times(q-1)},\ \mathrm{b} \in \Re^{k\times 1}\ \ \&\ \mathrm{w} \in \Re^{(q-1)\times 1} \tag{3}$$

여기서 행렬 **A**는 k개의 행과 (q - 1)개의 열들을 나타낸다. 벡터 b는 k개 벡터들의 첫 번째 값들을, 벡터 **W**는 고려되는 결측자료를 포함하는 벡터의 결측자료를 제외한 다른 값들을 나타낸다. 다음으로 least square regression method가 a를 찾기 위해 적용된다.

$$\min_x \left\| \mathrm{A}^{\mathrm{T}}\mathrm{x} - \mathrm{w} \right\|_2,\ \ \alpha = \mathrm{b}^{\mathrm{T}}\mathrm{x} \tag{4}$$

2.3 k 벡터의 결정

k개의 유사한 벡터를 선택하기위해 두 가지 방법이 적용되었다. 첫 번째는 L2-norm으로 결측자료가 존재하는 벡터와 가까운 벡터일수록 상관성이 높다는 가정 하에 가까운 순으로 k개의 벡터를 나열하는 방법이다. 두 번째는 Pearson correlation coefficient로 결측이 존재하는 벡터와 k개의 벡터간의 상관계수를 구해 상관관계가 큰 순으로 k개의 벡터를 나열하는 방법이다. Pearson 상관계수는 결측이 존재하는 벡터에서 결측자료

를 제외한 나머지 값들에 대한 벡터 $g_1'=(g_{12}, ..., g_{1n})^T$와 보정을 위해 사용되는 k개의 벡터 중 j번째 벡터인 $g_j'=(g_{j2}, ..., g_{jn})^T$를 사용하여 계산된다.

$$r_{1j} = \frac{1}{(n-1)} \sum_{i=2}^{n} \left(\frac{g_{1i} - \overline{g}_1}{\sigma_1} \right) \left(\frac{g_{ij} - \overline{g}_j}{\sigma_j} \right) \tag{5}$$

여기서 $\overline{g_j}$와 σ_j는 각각 g_j'의 평균과 분산이다. 상관계수가 음의 값을 가지는 경우도 높은 상관성을 가지는 것이므로 상관계수의 절대값이 1에 가까운 순으로 k개의 벡터를 나열하게 된다.

현재까지 최적의 k를 결정하기 위한 이론적인 결과는 없기(Kim et al., 2005) 때문에 식 6과 같이 결측값을 제외한 값에서 임의로 결측 a를 정하고 이를 다양한 수의 k값에 대해 적용함으로써 최적을 k를 구하는 방법을 사용되었다. 즉, 임의로 정한 결측 a를 구하기 위해 LLSimpute를 적용하고, 보정된 결과를 실제값과 비교하여 다양한 k 값 중 최적의 k_{opt}를 찾는 방법이다.

$$\begin{pmatrix} \mathbf{g}_1^T \\ \mathbf{g}_{s_1}^T \\ \vdots \\ \mathbf{g}_{s_k}^T \end{pmatrix} = \begin{pmatrix} \text{miss} & \alpha & \mathbf{w}_2 & \mathbf{w}_3 & \mathbf{w}_4 & \text{miss} \\ B_{1,1} & \mathbf{b}_1 & A_{1,2} & A_{1,3} & A_{1,4} & B_{1,2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ B_{k,1} & \mathbf{b}_k & A_{k,2} & A_{k,3} & A_{k,4} & B_{k,2} \end{pmatrix} \tag{6}$$

k_{opt}를 찾기 위한 기준으로는 식 7에 정리되어있는 NRMSE(Normalized Root Mean Squared Error)가 사용되었다. 식 7에서 y_{guess}와 y_{ans}는 각각 결측에 대한 추정값과 참값이며, std(·)는 표준편차를 의미한다.

$$\text{NRMSE} = \sqrt{\text{E}[(y_{guess} - y_{ans})^2]} / \text{std}(y_{ans}) \tag{7}$$

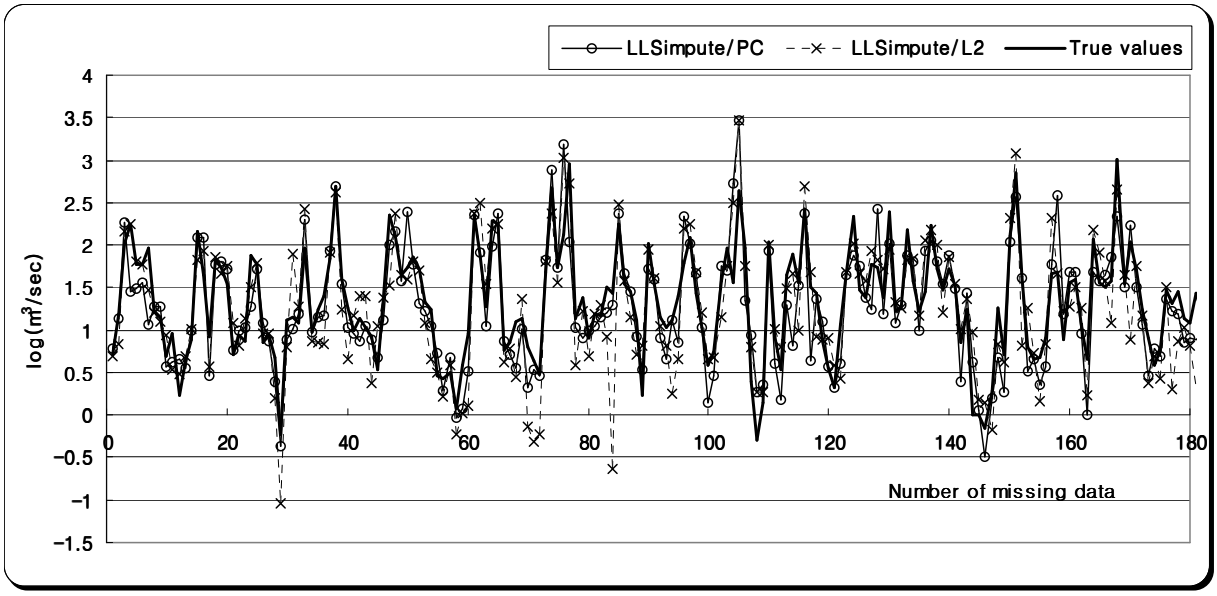
3. 적용 및 평가

3.1 결측의 생성

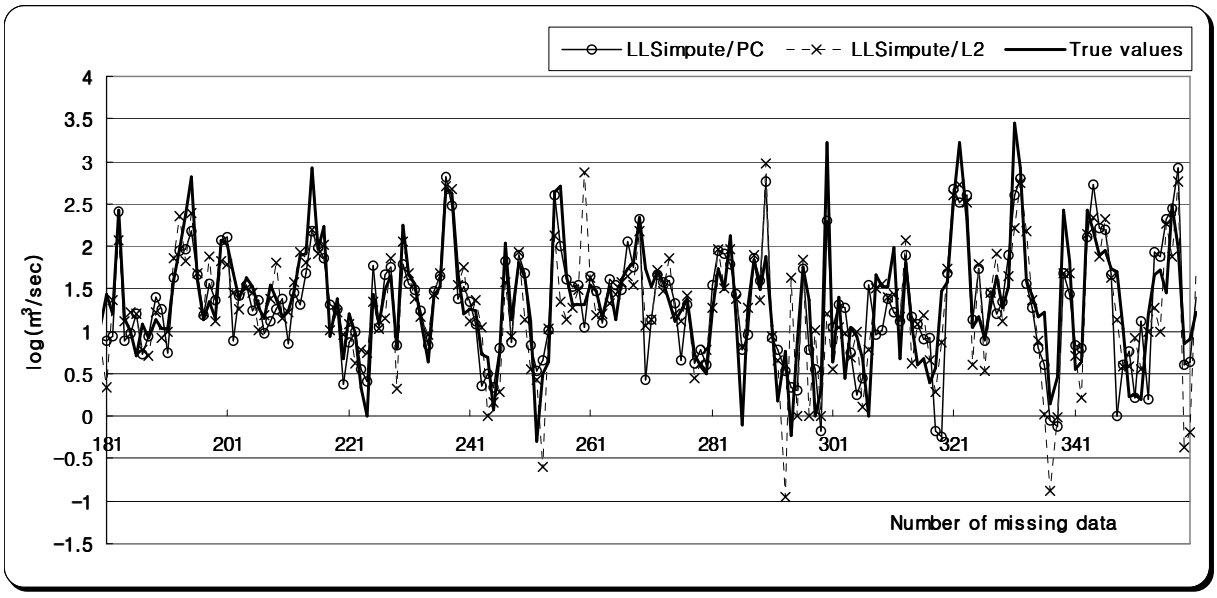
1974년부터 2004년까지 12053개의 소양강댐 일 유입량자료에 대해 1, 2, 3, 4, 5 %의 결측자료를 생성하였다. 즉, 1 ~ 5 % 각각에 대해 120, 241, 362, 482, 603개의 결측자료를 범위 내에서 무작위로 선정하게 된다. 또한 1 ~ 5 %에 대한 하나의 셋으로는 모집단에 대한 경향을 파악하기 어려움으로, 각각의 퍼센트에 대해 100개의 셋을 구성하여 모집단의 특성을 알아보고자 하였다.

3.2 결측의 보정

보정의 정도를 판단하기 위해 결측자료에 대한 참값과 LLSimpute/L2, LLSimpute/PS, 선형보간(linear interpolation)을 통해 보정된 자료를 비교하였다. 1 ~ 5 %에 대해 각각 100개 셋을 생성하였으므로 결과를 나타내기 위해 대표되는 결과를 선정하여 그림으로 나타내었다. 그림 1은 결측이 3 % 존재하는 경우의 첫 번째 셋에 대한 결과이다. 유량이 적은 경우와 많은 경우를 동시에 비교하기 위해 유량에 로그를 취한 결과를 사용하여 두 가지 방법을 비교하였다. 그림 1에서 LLSimpute/L2의 보정결과가 LLSimpute/PC의 경우 보다 실제값과 0.5 log(m^3/sec) 이상의 차이를 보이는 경우가 빈번하였다. 특히 적은 양의 유량추정에 있어 이러한 양상이 두드러지게 나타났다. 즉, 결측이 존재하는 벡터와 가까운 벡터일수록 상관성이 높다고 가정하는 LLSimpute/L2는 LLSimpute/PC와 비교하여 이상치를 보일 수 있는 가능성이 높다는 것을 보여준다.



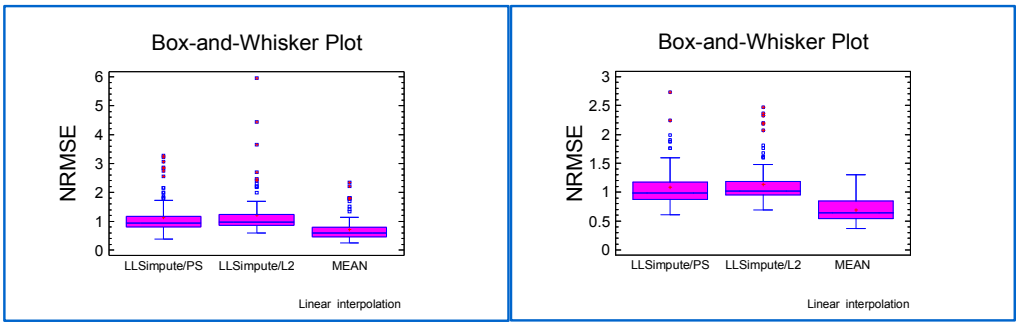
(a) 1 ~ 180번째 결측



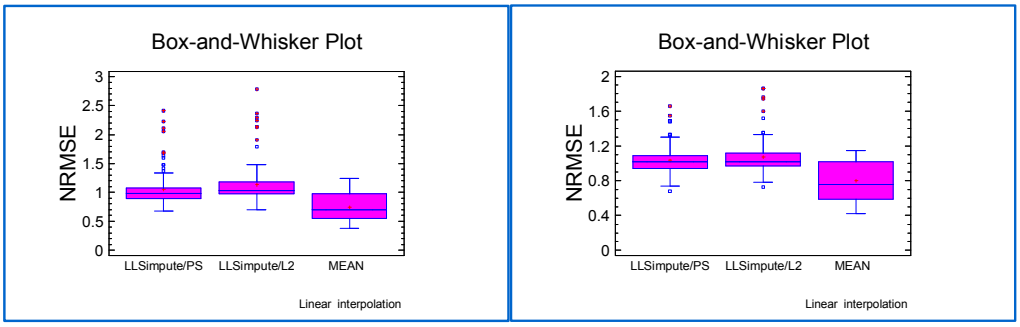
(b) 181 ~ 360번째 결측

그림 1. 전가방법을 사용한 3 % 결측의 보정 결과

다음으로 적용된 각 방법의 100개 셋에 대한 NRMSE 평가결과를 Box-plot으로 도시하였다(그림 2). 전체 자료에 대한 결측자료의 비가 작은 경우 선형보간을 통해 보정한 결과가 다른 두 방법에 비해 좋은 결과를 보였다. 하지만 결측의 비가 증가할수록 선형보간에 의한 보정 정도는 임의로 구성한 100개 각 셋의 사이에서 점차적으로 불균일한 결과를 보였다. 이러한 결과는 점차적으로 결측의 비율이 증가하는 경우 선형보간에 의해서는 일정한 보정 정확성을 기대하기 힘들다는 것을 보여준다. 이에 반해 LLSimpute/L2와 LLSimpute/PC에 의한 보정은 결측자료의 비가 증가하여도 선형보간에 비해 상대적으로 일정한 수준의 보정 결과를 기대할 수 있다. 적용된 두 전가방법 중 상관계수에 따라 결측의 추정에 사용될 벡터를 선정하는 LLSimpute/PC가 LLSimpute/L2와 비교하여 100개 셋에 대해 미소하지만 작은 NRMSE의 평균을 보였으나 방법 간 보정 정확성의 뚜렷한 차이는 없었다.



(a) 1 % 결측 (b) 3 %의 결측



(c) 4 % 결측 (d) 5 % 결측

그림 2. 100개 셋에 대한 1, 3, 4, 5 %의 결측 보정 후 NRMSE를 사용한 평가결과

4. 결론

유량자료에 존재하는 결측을 보정하기 위한 방안으로 선형보간법, LLSimpute/L2, LLSimpute/PC의 3가지 방법을 1974년부터 2006년까지의 소양강댐 일 유입량자료에 대해 임의로 1 ~ 5 %의 결측자료를 생성하여 검토하였다. 적용결과 LLSimpute/L2와 LLSimpute/PC는 결측의 비율이 증가하여도 보정의 정확성 및 불확실성 범위가 일정하게 유지된 반면, 선형보간법의 경우 보정 불확실성이 증가하고 평균적인 정확성이 낮아졌다. 적용된 두 전가방법 중 상관계수에 따라 결측의 추정에 사용될 벡터를 선정하는 LLSimpute/PC가 단순히 결측에 가까운 벡터를 사용하는 LLSimpute/L2와 비교하여 이상치를 보일 수 있는 가능성이 높았다. 본 연구에서 분석한 결과는 결측의 양에 따른 보정효과를 검토한 것으로 방법 간의 세부적인 분석을 위해서는 결측자료가 연속으로 존재하는 다양한 경우를 분류하여 분석하는 등과 같이 다양한 결측 양상에 대한 보정효과의 추가적인 검토가 요구된다.

감 사 의 글

본 연구는 국토해양부의 '앙상블 모형을 이용한 확률적 유량예측' 연구사업 지원에 의해 수행되었습니다.

참 고 문 헌

1. Kim, H., Golub, G. H., and Park, H. (2005). Missing value estimaion for DNA microarray gene expression data: Local least squares imputaion, *Bioinformatics,* 21, pp. 187-198.
2. Little, J. L., and Rubin, D. A. (1987). *Statistical Analysis with missing data*, John Wiley, New York.

3. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Fastie, T., Tibshirani, R., Botatein, D., and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays, *Bioinformatics,* 17, pp. 1-6.