다중회귀분석을 이용한 강우량 결측치 보정

Completion of the Missing Rainfall Data by a Multi-regression method

이명우* , 이봉희** , 김형수*** , 심명필****

Lee Myoung Woo , Lee Bong Hee , Kim Hung Soo , Shim Myung Pil

요 지

강우자료의 구축은 수문해석에 있어 가장 기본적이며 중요한 단계라 할 수 있다. 하지만 수문 관측 자료의 경우 결측치가 존재하여 그에 대한 보정이 필요한 경우가 종종 발생하게 된다. 따라서 수문자료의 분석을 수행하기에 앞서 우선 자료에 대한 검정을 실시하고, 결측치가 존재할 경우는 이를 보정하여 분석을 수행하여야 한다. 본 연구에서는 다변량통계기법의 하나인 다중회귀분석을 이용하여 강우 결측치를 보정하였다.

본 연구에서는 다중공선성과 자기상관에 대하여 고려한 다중회귀모형을 구성하였다. 모형의 구성 시 모든 결측지점에 적용이 가능하지 않아 일반성이 떨어짐을 확인 할 수 있었지만, 모형이 구성될 경우 통계적 적합도와 유의수준을 확인 할 수 있는 장점이 있었으며, 다중회귀모형이 구성되는 경우 좋은 보정 결과를 주는 것을 확인 할 수 있었다.

핵심용어 : 강우결측치보정, 다중회귀분석, 다변량통계기법, 다중공선성

1. 서 론

수문자료의 빈도해석에서 양질의 자료의 구축은 가장 기본적이며 중요한 작업이다. 이러한 수문자료의 결측치 보정에 관해 국내·외에서 많은 연구가 진행되었다. 김응석 등(1999)은 산술평균법(arithmetic average method), 년정상강우량법(normal ratio method), 수정년정상강우량방법(modified normal ratio method), 역거리법(inverse-distance method), 거리고도비율법(the ratio of distance and elevation method), 선형계획법(Linear programing), 크리깅방법(simple kriging method)등 여러 보정 방법을 비교한 바 있으며, Chang 등(2005)은 퍼지이론(fuzzy theory)을 이용하여 강우량을 산정한 바 있다. 본 연구에서는 강우의 보정에 다중회귀분석을 사용하였다. 회귀분석(regression analysis)은 통계기법 중 가장 많이 사용되는 방법 중 하나이다. Sokol(2003)은 레이더로부터 유도된 강우와 관측강우를 결합하여 시간강우량을 결정하는 방법을 선형회귀분석을 이용하여 개발하였다. 이동률 등(2001)은 지하수 함양량에 의한 월유출량의 영향을 평가하고, 이를 다중회귀모형의 독립변수로 이용하여 장기 월유출량의 예측을 시도한 바 있다. 정세웅(2003)은 8년간의 일별 수질자료와 댐 방류량 자료를 이용하여 겨울철동안 NH_3-N 농도를 예측할 수 있는 다중회귀분석 모형

* 정회원·동부엔지니어링 수자원환경부 사원·E-mail : moo97@dongbueng.co.kr
** 정회원·동부엔지니어링 수자원환경부 과장·E-mail : waterpia95@dongbueng.co.kr
*** 정회원·인하대학교 환경토목공학부 부교수·E-mail : sookim@inha.ac.kr
**** 정회원·인하대학교 환경토목공학부 교수·E-mail : shim@inha.ac.kr

을 개발한 바 있고, 조홍제 등(2000)은 울산수위관측소 지점의 저수위 유량을 나타내는 다중회귀식을 개발한 바 있다. 그리고 김성원(2000)은 다중회귀분석모형을 적용하여 일 유출량을 예측한 바 있다. 또한 윤강훈과 김태균(2004)은 특점지점의 유출과 수위예측을 위하여 선행유량과 강우레이더 예측 강우를 변수로 하는 통계학적 다중선형회귀 홍수예측모형을 개발한 바 있다.

2. 다중회귀분석

2.1 다중회귀분석

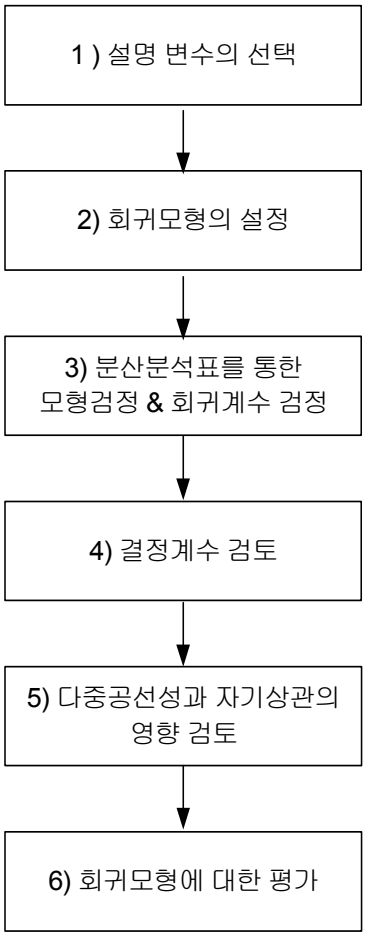

그림 1. 다중회귀모형의 적용 과정

회귀분석은 주로 기술(description), 추론(inference) 그리고 예측(prediction)하는 것을 목적으로 하고 있으며, 독립변수와 종속변수와의 관계를 설명하는 기법으로 가장 많이 사용되는 통계학적 방법 중 하나이다. 또한 회귀분석에서는 종속변수와 독립변수의 관계가 완벽하지 않다는 기본가정에서 시작한다. 본 연구에서 적용한 다중회귀모형의 구성과정은 그림 1과 같다. 종속변수 y를 매트릭스 형식으로 표현하면 다음 식(1)과 같이 나타낼 수 있다.

$$y = Xb + \epsilon \tag{1}$$

여기서 y와 ϵ은 (n×1)의 차수를 갖는 벡터이고 b는 [(p+1)×1]의 차수를 갖는 벡터이다. 여기서 X는 [n×(p+1)]의 매트릭스이며, 이는 b_0를 계산하기 위한 더미(dummy)변수를 포함하기 때문이다. b 매트릭스는 최소제곱법을 통하여 다음 식(2)와 같이 산정할 수 있다.

$$\hat{b} = (X'X)^{-1}X'y \tag{2}$$

회귀모형의 적정성을 평가하기 위한 기준으로는 다음과 같은 조건들이 있다(Farnum 등, 1989).

1) 종속변수를 가능한 적은 독립변수로 나타낼 수 있는 간단한 모형

2) 종속변수에 대한 독립변수의 유용성 검정을 위한 통계적인 F-검정을 만족하는 모형

3) 모형 독립변수의 회귀계수가 통계적인 t-검정을 만족하는 모형

4) 결정계수(R^2)가 높은 모형

5) 잔차(residual)의 검정을 만족하는 모형

2.2 다중공선성과 자기상관의 효과 검토

공선성(collinearity)은 두 독립변수간 관련성을 나타내는 표현이며, 하나의 독립변수가 다른 여러 독립변수들과 상관관계가 높으면 다중공선성(multicollinearity)을 갖는다고 말한다. 큰 다중공선성은 회귀분석의 신뢰성을 저하시킬 수 있으며, Condition Index(CI)는 이러한 다중공선성을 나타내는 하나의 지표로 사용된다. 자료행렬 X는 특이값분해(singular value decomposition)를 통해 상관관계를 갖지 않는 단위벡터들로 구성된 행렬로 분해될 수 있다. 대각행렬의 요소는 상관관계를 갖지 않는 단위벡터들의 표준편차로서 이를 행렬 X의 특이값이라 한다. CI는 자료행렬 X의 각 열벡터가 단위벡터가

되도록 정규화한 후 이를 특이값분해하여 구한 대각행렬의 요소들 중 가장 큰 값 $d_{\max}$와 가장 작은 값 $d_{\min}$의 비율(CI=$d_{\max}/d_{\min}$)로 나타내며, 30이 넘을 경우 다중공선성이 심각함을 의미한다.

또한 자기상관의 영향이 있을 경우 회귀계수의 유의성에 대한 추론이 과대평가 될 수 있으므로 자료의 자기상관에 대한 검토가 필요하다. 다시 말하면, 회귀계수가 유의하지 않음에도 불구하고 유의하다는 결론을 내릴 가능성이 높아진다는 것이다. 자기상관여부를 확인하는 통계적 방법으로는 식 (3)과 같은 Durbin-Watson(DW) 통계량이 있다. 식 (3)에서 구한 DW가 2.0에 접근 할 경우 자기상관이 0에 가깝다는 것을 의미한다.

$$DW=\frac{\sum_t (e_t - e_{t-1})^2}{\sum_t e_t^2} \qquad (3)$$

3. 대상유역의 선정과 모형의 구성

3.1 대상유역 및 호우사상

본 연구에서는 국제수문개발계획(IHP)의 시험유역인 보청천의 탄부소유역을 대상유역으로 선정하였다. 탄부소유역에는 4개의 강우 관측소가 위치하고 있으며, 각 관측소별 강우의 지속기간이 24시간인 연 최대치 계열 22개 자료를 “2004년 국제수문개발계획(IHP) 보고서”(건설교통부, 2005)로부터 획득하여 본 연구에 사용하였다.

3.2 결측치 보정과 자료의 구성

본 연구에서 IHP보고서로부터 획득하여 사용한 강우자료에는 결측치가 존재하며, 이러한 결측치 보정을 하기위하여 다중회귀분석을 적용하였다. 그림 2는 강우관측소를 평면상에 나타낸 것이며 그림 3은 고도를 고려하여 3차원으로 나타낸 것이다.

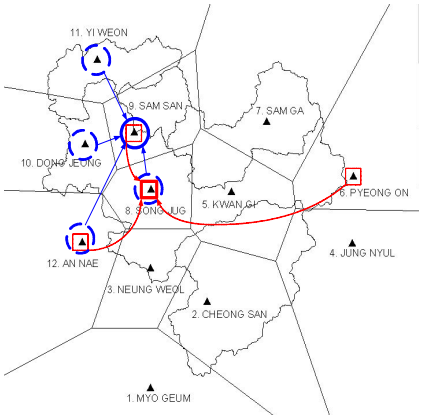

그림 2. 보청천 유역의 강우관측소 & 회귀모형

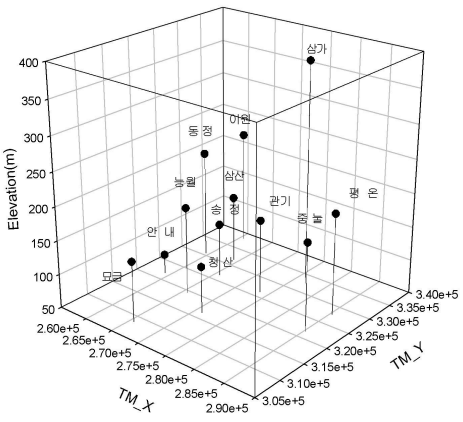

그림 3. 강우관측소의 3차원 도시

표 2는 각 강우관측소 강우량의 상관계수행렬을 나타낸 것이다. 삼가, 송죽, 삼산관측소의 경우 강우자료의 검토결과 결측치가 존재하는 것으로 나타났다. 송죽과 삼산관측소에서는 높은 상관관계를 갖는 강우관측소가 존재함을 확인 할 수 있었으나, 삼가관측소의 경우 높은 상관관계를 보이는

강우관측소를 확인할 수 없었다. 이는 높은 고도와 가장 가장자리에 위치하는 삼가관측소의 지리적 영향에서 기인한 것으로 판단된다.

이러한 분석결과를 바탕으로 결측치를 잘 설명할 수 있는 독립변수를 선정하였다. 독립변수의 선정은 위의 상관계수 뿐 아니라 다중공선성의 유무와 시계열의 영향을 고려하여 수행하였다. 또한 회귀모형전체는 유의수준 0.01을 만족하는 모형을 선택하였고, 각 회귀계수에 대한 검정은 유의수준 0.1을 만족하는 모형을 선정하였다. 다음의 회귀모형은 단계별변수선택법을 적용한 뒤 위의 조건들은 검토하여 선정한 회귀모형을 나타낸다. 회귀모형의 평가를 위한 통계치는 표 3과 같고 구성된 회귀모형의 결과는 그림 4, 5와 같다.

표 2. 보청천유역의 강우관측소간 상관계수

	묘금	청산	능월	중눌	관기	평온	삼가	송죽	삼산	동정	이원	안내
묘금	1.000	0.349	0.350	0.295	0.357	0.494	0.718	0.286	0.405	0.226	0.338	0.125
청산	0.349	1.000	0.653	0.670	0.636	0.448	0.521	0.542	0.404	0.271	0.403	0.554
능월	0.350	0.653	1.000	0.584	0.678	0.426	0.569	0.671	0.647	0.582	0.610	0.561
중눌	0.295	0.670	0.584	1.000	0.840	0.479	0.484	0.867	0.727	0.617	0.708	0.896
관기	0.357	0.636	0.678	0.840	1.000	0.339	0.532	0.882	0.818	0.714	0.621	0.817
평온	0.494	0.448	0.426	0.479	0.339	1.000	0.693	0.495	0.353	0.337	0.624	0.380
삼가	0.718	0.521	0.569	0.484	0.532	0.693	1.000	0.599	0.619	0.515	0.655	0.445
송죽	0.286	0.542	0.671	0.867	0.882	0.495	0.599	1.000	0.877	0.847	0.829	0.945
삼산	0.405	0.404	0.647	0.727	0.818	0.353	0.619	0.877	1.000	0.863	0.661	0.751
동정	0.226	0.271	0.582	0.617	0.714	0.337	0.515	0.847	0.863	1.000	0.811	0.764
이원	0.338	0.403	0.610	0.708	0.621	0.624	0.655	0.829	0.661	0.811	1.000	0.788
안내	0.125	0.554	0.561	0.896	0.817	0.380	0.445	0.945	0.751	0.764	0.788	1.000

표 3. 회귀모형의 평가

통계치 / 모형	CI Index	DW	1차 자기상관계수	R^2	R^2_{adj}
삼산	19.529	1.926	0.019	0.897	0.868
송죽	7.500	1.878	0.033	0.9715	0.966

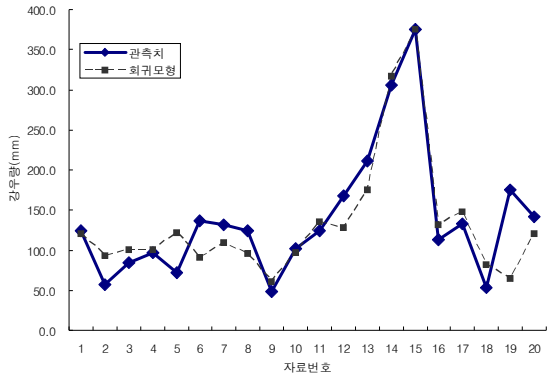

그림 4. 삼산관측소 회귀모형의 결과와 관측치의 비교

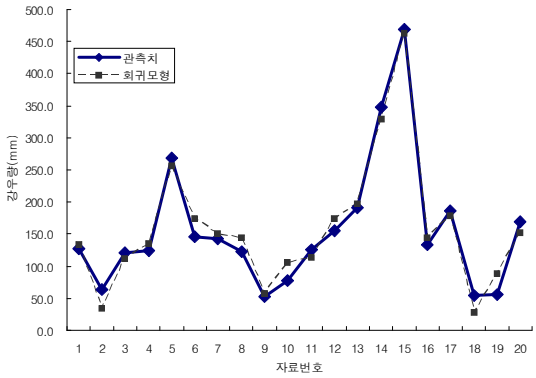

그림 5. 송죽관측소 회귀모형의 결과와 관측치의 비교

삼산관측소와 송죽관측소의 경우 관측치와 회귀모형을 통해 산정한 결과치의 비교·검토결과 그림 4와 그림 5에 나타난 바와 같이 비교적 관측치를 잘 모의하는 것으로 나타났으며, 삼가관측소의

경우 다른 관측소와 높은 상관계수를 보이는 관측소가 존재하지 않고 가장 높은 고도에 위치해 있는 등의 이유로 통계적 조건에 알맞은 회귀모형을 선정할 수 없었다. 따라서 거리와 고도를 이용한 거리고도비율법(the ratio of distance and elevation method)을 이용하여 결측치를 보정해야 할 것으로 판단된다.

5. 결 론

본 연구에서 결측치를 보정하기 위하여 다중회귀분석을 수행하였다. 다중회귀모형의 구성시 결정계수 R^2의 검토만으로 모형을 평가해서는 안 되며, 다중공선성과 자기상관의 효과를 고려하여 모형을 구성해야 한다. 다중공선성이나 자기상관은 회귀모형의 신뢰성을 떨어뜨리는 악영향을 미치기 때문이다. 따라서 본 연구에서는 다중공선성과 자기상관을 고려하여 종속변수인 강우관측소를 결정하여 회귀모형을 구성하였다.

다중회귀분석을 통한 결측치 보정은 공식의 일반화를 이끌어 내기는 어려우나 주어진 통계적 절차를 따라 통계적으로 적합도와 유의수준을 확인해 볼 수 있었으며, 좋은 보정결과를 주는 장점이 있다. 또한 다중회귀분석의 적용으로 각 대상유역에 알맞은 회귀모형을 결정하는 방향제시가 될 수 있을 것으로 판단된다. 또한 다중공선성에 대한 검토로 독립변수의 상관관계에 대하여 검토해 볼 수 있으므로, 적정 독립변수 선택에 도움이 될 것으로 기대된다.

참 고 문 헌

1. 건설교통부(2005). 2004년 국제수문개발계획(IHP) 연구보고서, 건설교통부.
2. 김성원(2000). 다층신경망모형에 의한 일 유출량의 예측에 관한 연구, 한국수자원학회 논문집, 제33권, 제5호, pp. 537-550.
3. 김응석, 김형수, 김중훈(1999). 점 강우량 결측시 보정방법에 관한 비교 연구, 한국수자원학회 학술대회지, '99년 한국수자원학회 학술발표회 논문집, pp. 374-381.
4. 김중훈, 김태균, 김응석(1995). 산악 지역을 고려한 점강우량 결측시 보정 방법, 대한토목학회 1995년도 학술발표회 논문집(II), 대한토목학회, pp. 169-172.
5. 윤강훈, 김태훈(2004). 레이더 예측 강우를 이용한 다중회귀 예측모형의 적용가능성 평가, 대한토목학회 논문집, 제24권, 제4B호, pp. 295-300.
6. 이동률, 윤용남, 안재현(2001). 월 유출량 예측 변수로서 지하수 함양량의 이용, 한국수자원학회 논문집, 제34권, 제3호, pp. 275-285.
7. 정세웅(2003). 일별 암모니아성 질소(NH_3-N)농도 예측을 위한 다중회귀모형 개발, 한국수자원학회 논문집, 제36권, 제6호, pp. 1047-1058.
8. 조홍제, 황재호, 문성준(2000). 태화강 감조부의 저수위 수위-유량곡선 개선, 한국수자원학회 논문집, 제33권, 제5호, pp. 635-645.
9. Chang, C.L., Lo, S.L. and Yu, S.L.(2005). Applying Fuzzy Theory and Genetic Algorithm to Interpolate Precipitation, Journal of Hydrology, ELSEVIER, Vol. 314, pp. 99-104.
10. Farnum, N.R, and Stanton, L.W.(1989). Applied Quantitative Forecasting Methods, PWS-KENT Publishing Company, Boston.
11. Sokol, Z.(2003). Utilization of Regression Models for Rainfall Estimated Using Radar-Derived Rainfall Data and Rain Gauge Data, Journal of Hydrology, ELSEVIER, Vol. 278, pp. 144-152.