국내 지역 홍수빈도해석을 위한 기법 제안: Bayesian-GLS 회귀

Proposing a Technique for Regional Flood Frequency Analysis: Bayesian-GLS Regression

정대일*, 제리 스테딘저**, 김영오***, 성장현****

Dae Il Jeong, Jery R. Stedinger, Young-Oh Kim, Jang Hyun Sung

요 지

국내 홍수빈도 분포의 매개변수 추정에서 지점추정(at-site estimate) 방법은 유량 자료의 부족으로 발생하는 표본오차(sampling error)가 크기 때문에 충분한 유량 자료를 보유한 지점에 한하여 제한적으로 사용되고 있다. 대안으로 동질성을 가진 유역의 유량 자료를 모아 지역 매개변수를 추정하는 지수홍수법(Index Flood Method)이 제안되기도 하였으나, 이질성이 큰 우리나라의 유역특성 때문에 적용이 쉽지 않다. Stedinger와 Tasker가 1986년 제안한 GLS(Generalized Least Square) 기법은 유역을 동질지역으로 구분할 필요가 없으며 지점들간의 상관관계와 이분산성을 고려할 수 있어, 국내 홍수빈도 해석을 위해서 꼭 도입해야할 기법으로 생각된다. 본 연구에서는 기존의 GLS 기법의 단점을 보완한 Bayesian-GLS 기법을 이용하여, 국내 대유역에 골고루 위치하며 댐의 영향을 받지 않는 31개 지점의 연최대 일유량 시계열의 L-변동계수(L-moment coefficient variation)와 L-왜도계수(L-moment coefficient skewness)를 추정할 수 있는 회귀모형을 제안하였다. 위 회귀모형을 구성하기 위한 유역특성으로는 유역면적, 유역경사, 유역평균강우 등을 사용하였다. Bayesian-GLS (B-GLS) 적용 결과를 OLS(Ordinary Least Square) 및 Bayesian-GLS 기법에서 지점간의 상관관계를 고려하지 않는 Bayesian-WLS(Weighted Least Square)와 비교 평가하여 그 우수성을 입증하였다. 따라서 본 연구에서 제안된 B-GLS에 의한 지역회귀모형은 국내의 미계측유역이나 또는 관측 길이가 짧은 계측유역의 홍수빈도분석을 위해 매우 유용할 것으로 기대된다.

핵심용어: Bayesian Generalized Least Square, L-변동계수, L-왜도계수, 연최대 일유량, 유역경사, 유역면적, 유역평균강우, 회귀모형

1. 서 론

미국과 영국 등 선진국에서는 이미 40년 전부터 수문빈도분석의 필요성을 인지하였고, 국가적인 차원에서 주목을 받아 학문적인 연구와 실용화가 진행되었다. 당시 선진국 학자들의 가장 큰 고민은 짧은 자료 기간을 이용하여 100년 이상의 재현기간을 가진 홍수량을 산정하여야 하는 것이었으며, 이제야 30여년의 체계적인 자료를 보유하게 된 국내의 실정이 그들의 당시 고민과 비슷하다고 할 수 있다.

국내 홍수량 자료는 길이가 짧아 시간적 길이를 공간적으로 보완하는 지역화 기법이 필요하다. 하지만 동질성을 가진 유역의 유량 자료를 모아 지역 매개변수를 추정하는 지수홍수법이 제안되기도 하였으나, 이질성이 큰 우리나라의 유역특성 때문에 동질성이라는 가정조차 만족시키지 못하는 경우가 다반사이다.

* Post Doctoral Fellow, School of Civil and Environmental Engineering, Cornell University · E-mail: dj64@cornell.edu
** Professor, School of Civil and Environmental Engineering, Cornell University · E-mail: jrs5@cornell.edu
*** 서울대학교 건설환경공학부 부교수 · E-mail: yokim05@snu.ac.kr
**** 서울대학교 공학연구소 연구원 · E-mail: kon26@snu.ac.kr

Stedinger와 Tasker(1986)가 제안한 GLS(Generalized Least Square)기법은 유역을 동질지역으로 구분할 필요가 없으며, 지점들간의 상관관계와 이분산성을 고려할수 있는 지역화기법으로서 유량 자료의 여건상 국내 홍수빈도 해석을 위해서 꼭 도입해야할 기법으로 판단된다.

최근 Reis 등(2005)은 기존의 GLS기법을 개선하여 모형오차와 추정된 매개변수의 불확실성을 표현할 수 있는 Bayesian GLS 기법을 제안하였으며, 미국의 Muskingum 유역과 브라질의 Tibagi 유역의 적용사례를 통해 그 효용성을 확인한 바 있다. 따라서 본 연구에서는 B-GLS 기법을 이용하여 L-왜도계수(L-moment coefficient of variance; L-CV)와 L-변동계수(L-moment coefficient of skewness; L-CS)를 유역특성인자(예; 유역면적, 유역경사, 유역평균강우)를 이용하여 추정하는 지역회귀모형을 구축하였다. 본 연구를 통해 구축된 지역회귀모형은 국내의 미계측유역이나 또는 관측 길이가 짧은 계측유역의 홍수빈도분석을 위해 매우 유용할 것으로 기대된다.

2. 기본개념

지역회귀기법은 유역의 물리적 특성을 설명변수로 확률홍수량을 구하는 경우와 확률분포형의 매개변수를 구하는 경우가 있으며, 본 연구에서는 확률분포형의 매개변수를 구하는 것이 목적이다. 일반적으로 사용되던 OLS(Ordinary Least Square)기법에 의한 매개변수 지역화는 오차항의 등분산성과 독립성의 가정이 위배되는 단점에 항상 노출되어 있었으므로, 이러한 문제점을 극복하기 위해 1980년대 이후 WLS(Weighted Least Square)와 GLS 기법을 이용한 지역화 기법이 제안되었다(Stedinger와 Tasker, 1986)

목적변수 y_i(L-변동계수 또는 L-왜도계수)는 유역의 특성을 나타내는 설명변수들(X_{i1}, ..., X_{ik})의 선형관계와 오차항(δ_i)으로 이루어진 식(1)과 같이 표현될 수 있다.

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} + \delta_i \qquad i = 1, 2, ..., n \tag{1}$$

여기서, n은 관측지점의 개수를 의미한다. 지점 i로부터 지점추정된 y_i의 추정값 $\widehat{y_i}$에는 표본오차(η_i)가 포함되어있으므로 참값인 y_i는 아래와 같이 표현될 수 있다.

$$\widehat{y_i} = y_i + \eta_i \qquad i = 1, 2, ..., n \tag{2}$$

표본오차 η_i의 분산은 각 지점간의 자료의 길이와 지점간의 상관관계에 의해 결정되며 식(1)과 식(2)를 이용하여 GLS모형을 행렬식으로 나타내면 식(3)과 같다.

$$\widehat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{\delta} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3}$$

여기서, $\mathbf{X}$는 $[n \times k+1]$행렬, $\boldsymbol{\beta}$는 $[k+1]$ 벡터로 GLS모형의 매개변수를 의미한다. $\boldsymbol{\eta}$와 $\boldsymbol{\delta}$는 각각 표본오차와 모형오차 벡터를 의미한다.

총 오차인 $\boldsymbol{\varepsilon}$는 평균이 0이며, 식(4)와 같은 공분산행렬로 표현할 수 있다.

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = \Lambda(\sigma_\delta^2) = \sigma_\delta^2 I + \Sigma(\widehat{y}) \tag{4}$$

여기서, σ_δ^2는 모형오차의 분산을 의미하며, $\Sigma(\widehat{y})$는 표본오차의 공분산행렬을 의미한다. 서로 다른 두지점(i, j)의 목적변수 y_i, y_j의 상관관계를 고려하지 않을 경우 GLS기법은 WLS기법이 되며, 표본오차의 공분산행

렬의 대각행렬이 모두 모형오차의 분산 σ_δ^2와 같은 경우 OLS기법이 된다. GLS기법에 의한 β의 추정값 b와 분산 $\Sigma[b]$은 식(5)를 통해 계산할 수 있다.

$$b=[X^T\Lambda(\sigma_\delta^2)^{-1}X]^{-1}X^T\Lambda(\sigma_\delta^2)^{-1}\hat{y} \qquad \Sigma[b]=[X^T\Lambda(\sigma_\delta^2)^{-1}X]^{-1} \tag{5}$$

이와 같이 GLS 기법은 모형오차의 분산을 모우멘트법이나 최우도법을 이용하여 추정할 수 있으나, 추정된 모형오차의 분산이 0 또는 거의 0과 같은 비현실적인 값들을 제시하는 경우가 빈번하였으므로, Reis 등 (2005)은 Bayesian 기법을 이용하여 보다 합리적인 모형오차의 분산을 추정하고, 이에 따른 β의 사후분포 (posterior distribution)를 제시할 수 있는 B-GLS기법을 제안하였다. Bayesian 분석법은 β와 σ_δ^2 사전분포의 구체적인 정보를 필요로 한다. 매개변수의 사전분포를 구하는데 있어 평균이 0이고 비교적 큰 분산을 갖는 다변량 정규분포(multivariate normal distribution)를 이용하며 이는 사전정보가 거의 없는 지역의 경우, 확률밀도함수는 상대적으로 평활(flat)해 지며 사전정보가 많은 경우에는 지수분포를 띠게 된다.

구축된 지역회귀모형의 정확성을 평가하기 위해 AVP(Average Variance of Prediction)와 Pseudo R^2, ERL(Effective Record Length)을 사용하였다. AVP는 구축된 모형이 목적변수의 참값을 얼마나 잘 예측하는지를 분산을 이용해 평균적으로 설명하는 척도이며 식(4)를 이용해 계산할 수 있다. AVP가 작을수록 모형의 예측 능력이 우수하다고 할 수 있다.

$$AVP_{GLS}=\widehat{\sigma_\delta^2}+\frac{1}{n}\sum_{i=1}^{n}x_i(X^T\hat{\Lambda}^{-1}X)^{-1}x_i^T \tag{6}$$

여기서, x_i는 각 지점의 설명변수들의 열벡터를 의미한다.

기존의 결정계수 R^2는 회귀식의 제곱합(SSR, sum of squares from regression)과 잔차의 제곱합(SSE, sum of squares due to residual error)을 구분하여 계산되나, WLS나 GLS기법의 경우 SSE와 SSR이 표본오차와 모형오차를 모두 포함하고 있으므로 GLS기법의 가치를 설명하기에는 적당하지 않다. Reis et al. (2006)은 k개의 설명변수를 이용한 GLS모형의 모의오차가 설명변수를 사용하지 않은 경우에 비해 얼마나 개선되는 지를 평가할 수 있는 Pseudo-R^2를 식(7)과 같이 제안하였다.

$$\overline{R}^2=\frac{n[\hat{\sigma}_{\delta^2}(0)-\hat{\sigma}_{\delta^2}(k)]}{n\hat{\sigma}_{\delta^2}(0)} \tag{7}$$

여기서, $\hat{\sigma}_{\delta^2}(k)$는 k개의 설명변수를 가진 GLS모형의 모형오차분산을 의미한다.

마지막으로 ERL는 목적변수(L-변동계수, L-왜도계수)에 대해 GLS모형의 AVP와 지점분석에 의한 오차분산의 비에 평균자료길이를 곱하여, GLS모형의 AVP가 지점분석의 분산보다 얼마나 작은지를 자료의 길이를 통해 보여주는 것으로 AVP가 작을수록 ERL의 길이는 길어진다.

3. 적용 및 결과

B-GLS기법을 이용한 지역회귀모형을 구축을 위해, 전국의 유량관측 지점을 대상으로 10년 이상의 유량자료를 보유하고 있으며, 상류 댐에 의한 조절유량에 영향을 받지 않는 31개의 지점을 선정하였다. 선정된 31개 지점의 평균 자료길이는 22년이었다. 그림 1은 선택된 31개 지점의 연최대 일유량 시계열에서 지점(at-site) 추정된 L-변동계수와 L-왜도계수, 그리고 각 지점의 자료길이를 보여주고 있다. 추정값들의 지점간 분산이 큼을 확인할 수 있으며, 자료의 길이에 의한 추정값들의 표본오차가 클 것임을 짐작할 수 있다.

지역회귀모형의 설명변수로서는 먼저, 관측지점이 위치한 대유역을 구분하기 위해 이변수 지표(Z_1, Z_2)를 이용하여, 한강유역(1, 0), 낙동강유역(0, 1), 금강과 섬진강유역(0, 0)로 표현하였다. 다음으로 유역면적에 자연로그를 취한 Ln(A), 유역경사, 평균 강우량(MP), 연최대 일강우량의 표준편차 등 6가지를 사용하였으며, 설명변수의 모든 조합에 대해 회귀모형을 구축하고 정확성을 비교하였다.

목적변수인 L-왜도계수, L-변동계수의 표본오차의 공분산행렬 $\Sigma(\hat{y})$을 추정하기 위해 국내 홍수자료를유연성이 큰 GEV(Generalized Extreme Value)로 가정하였으며, Jeong et al.(2007)의 연구결과에 근거하여 각 지점의 자료길이와 지점간의 상관관계를 이용해 추정하였다. 이 과정에서 31개의 각 지점간의 상관계수가 필요하나, 두 지점간의 자료길이가 짧은 경우 표본 상관계수를 이용하면 표본오차로 인해 물리적으로 비현실적인 음수의 상관관계가 추정되어 표본오차의 공분산행렬 $\Sigma(\hat{y})$이 음의 값을 가질 수 있다. 따라서 본 연구에서는 표본 상관계수를 사용하지 않고, 두 지점간의 자료길이가 20년 이상인 지점들의 표본상관계수와 거리와의 관계를 식(8)과 같이 유도하여 사용하였다.

$$\rho_{ij}=0.45\exp(-d_{ij}/80) \tag{8}$$

여기서, ρ_{ij}는 두 지점의 연최대 일유출량 간의 상관계수를 d_{ij}는 두 지점간의 거리를 의미한다. 그림 2는 자료길이가 20년 이상인 각 지점들의 표본상관계수와 거리, 그리고 식(8)의 관계식을 나타내고 있다.

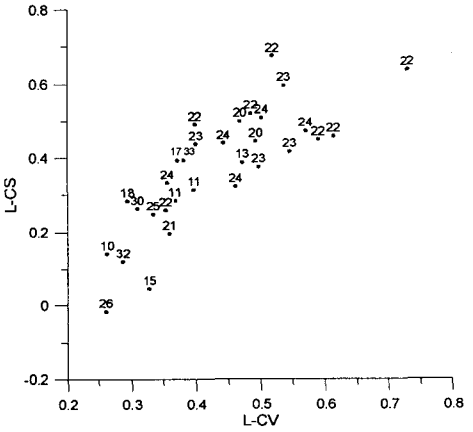

그림 1. 각 지점의 L-변동계수와 L-왜도계수 추정값 및 자료길이

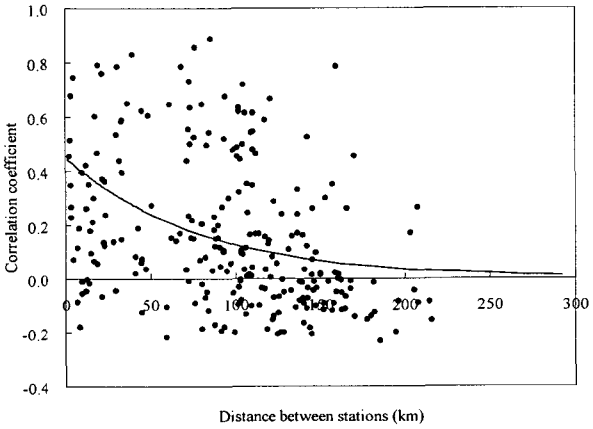

그림 2. 두 지점간의 표본 상관계수와 거리

표 1은 L-변동계수와 L-왜도계수에 대한 OLS, B-WLS, B-GLS기법 중에서 6개의 설명변수로 구축가능한 모든 회귀모형 중에서 가장 우수한 모형의 매개변수 추정값과 모형오차분산, AVP, Pseudo-R^2, ELR 등을 비교한 것이다. 각 지점의 설명변수들은 평균이 0이 되도록 모두 조정하여, β_0가 L-변동계수, L-왜도계수의 지역평균값을 갖도록 하였다. L-변동계수의 B-GLS모형은, 대수로그를 취한 유역면적(Ln(A))과 유역 평균강우(MP)를 설명변수로 이용할 경우 가장 우수하였으며, AVP는 0.0057로서 이중 75 %를 모형오차분산이 차지하고 있었다. 반면, L-왜도계수의 B-GLS모형은, Ln(A)만을 설명변수로 이용할 경우 가장 우수하였으며, AVP는 0.0094로서 이중 약 60 %를 모형오차분산이 차지하고 있었다. OLS와 B-WLS, B-GLS의 비교결과, OLS 기법에서 모형오차의 분산과 AVP가 가장 커 열등하였다. B-WLS 기법은 AVP와 ERL 등 정확성 지표면에서 볼 때 B-GLS 기법만큼 우수하였으나, 목적변수의 지점들 간의 상관관계를 고려하지 않음으로서 AVP값을 잘못 제시하고 있다.

L-변동계수 B-GLS모형의 ERL는 21년으로서, 길이가 21년인 관측 자료로부터 지점 추정된 값과 동일한 정확성을 나타냄을 의미하며, 31개 지점 자료의 평균길이 22년과도 비슷하였다. 따라서 미국의 Bulletin 17B(IACWD, 1982)에서 제안한 것과 같이 지점 추정값과 지역회귀모형 추정값의 가중평균을 이용하는 것이 바람직할 것으로 생각된다. L-왜도계수의 ERL은 51년으로서 본연구에서 사용한 자료의 평균길이 22년보다 두 배 이상 길어 지역회귀모형의 추정값이 지점 추정값에 비해 매우 정확함을 확인하였다.

4. 결 론

본 연구에서는 L-왜도계수, L-변동계수를 유역의 특성인자를 이용한 B-GLS기법으로 지역회귀모형을 구축하였다. 구축된 B-GLS모형은 일반적인 OLS와 지점간의 상관관계를 고려하지 않는 B-WLS를 이용한 지역회귀모형과 비교하여 우수함을 확인하였다. L-변동계수의 경우 유역면적과 유역 평균강우를 설명변수로 채택하였으며, ERL이 21년으로 모형 구축에 사용된 31개 지점의 평균 자료길이 22년과 비슷하여, 지역회귀모형의 추정값과 지점 추정값을 가중평균해 사용하는 방안이 바람직해 보인다. L-왜도계수의 경우 유역면적만을 설명변수로 채택하였으며, ERL가 51년으로서 지역회귀모형의 추정값이 지점 추정값에 비해 훨씬 정확함을 확인하였다. 본 연구에서 제안한 B-GLS 지역회귀모형은 국내의 미계측유역이나 관측 길이가 짧은 계측유역의 홍수빈도분석을 위해 매우 유용할 것으로 사료된다.

		β_0	Ln(A)	MP	σ_δ^2	표본분산 평균	AVP_{GLS}	$\bar{R}^2$(%)	ERL (years)
L-변동계수	OLS	0.4314 (0.0154)	-0.0410 (0.0105) [0.0]	-0.1436 (0.0413) [0.1]	0.0073	0.0007	0.0080	42	15
	B-WLS	0.4306 (0.0169)	-0.0410 (0.0114) [0.1]	-0.1404 (0.0485) [0.5]	0.0040 (0.0023)	0.0009	0.0049	54	24
	B-GLS	0.4220 (0.0285)	-0.0416 (0.0116) [0.1]	-0.1307 (0.0522) [1.3]	0.0043 (0.0021)	0.0015	**0.0057**	**45**	**21**
L-왜도계수	OLS	0.3679 (0.0267)	-0.0472 (0.0181) [0.9]		0.0221	0.0014	0.0235	16	21
	B-WLS	0.3792 (0.0757)	-0.0492 (0.0206) [1.1]		0.0059 (0.0047)	0.0016	0.0074	31	65
	B-GLS	0.3405 (0.0489)	-0.0535 (0.0188) [0.6]		0.0060 (0.0044)	0.0035	**0.0094**	**37**	**51**

* 여기서 (·)는 해당 변수의 표준오차를 [·]는 p-value(%)를 의미함

감 사 의 글

본 연구는 건설교통부 한국건설교통기술평가원의 이상기후대비시설기준강화 연구단에 의해 수행되는 2005 건설기술기반구축사업(05-기반구축-D03-01)에 의해 지원되었습니다.

참 고 문 헌

1. Jeong, D. I, Stedinger, J. R., and Martins, E. S. (2007). "Variance and cross-site correlations among L-CV and L-CS for GEV and GLO distributions." *Water Resources Research*, submitted.
2. Reis, D. S., Jr., Stedinger, J. R., and Martins, E. S. (2005). "Bayesian generalized least squares regression with application to log Pearson type 3 regional skew estimation." *Water Resources Research*, 41, W10419, doi:10.1029/2004WR003445.
3. Reis, D. S., Jr., Stedinger, J. R., Martins, E. S. (2006). "Operational Bayesian GLS regression for regional hydrologic analyses." *manuscript*, Cornell University, NY, USA.
4. Stedinger, J. R., and Tasker, G. D. (1986). "Regional hydrologic analysis, 2 model-error estimation, estimation of sigma and Log-pearson type III distribution." *Water Resources Research*, 22(10), pp. 1487-1499.