Journal of Environmental Science International
22(3); 303~308; March 2013

pISSN: 1225-4517 eISSN: 2287-3503
http://dx.doi.org/10.5322/JESI.2013.22.3.303

ORIGINAL ARTICLE

공간 데이터와 시계열 데이터로부터 유도된 공분산행렬을 결합한 강수량 결측값 추정 모형

성 찬 용*
계명대학교 환경대학 환경계획학과

Development of a Model Combining Covariance Matrices Derived from Spatial and Temporal Data to Estimate Missing Rainfall Data

Chan Yong Sung*
Department of Environmental Planning, Keimyung University, Daegu 704-701, Korea

Abstract

This paper proposed a new method for estimating missing values in time series rainfall data. The proposed method integrated the two most widely used estimation methods, general linear model(GLM) and ordinary kriging(OK), by taking a weighted average of covariance matrices derived from each of the two methods. The proposed method was cross-validated using daily rainfall data at thirteen rain gauges in the Hyeong-san River basin. The goodness-of-fit of the proposed method was higher than those of GLM and OK, which can be attributed to the weighting algorithm that was designed to minimize errors caused by violations of assumptions of the two existing methods. This result suggests that the proposed method is more accurate in missing values in time series rainfall data, especially in a region where the assumptions of existing methods are not met, i.e., rainfall varies by season and topography is heterogeneous.

Key words : Missing rainfall data, Geostatistics, General linear model, Ordinary kriging

1. 서 론

시계열 강수량 데이터는 수문분석을 비롯한 환경과 관련된 여러 분야의 연구에 필수적인 자료이다(Michaud와 Sorooshian, 1994). 하지만 측정 장비의 고장 등으로 인해 시계열 강수량 데이터에는 결측값이 많은데, 이러한 결측값들은 수문분석의 정확성을 낮출 뿐 아니라(Choi 등, 2010), transfer function 모형과 같이 시계열 데이터에 바탕을 둔 여러 수문분석기법을 적용할 수 없게 한다(Sung과 Li, 2010).

시계열 강수량 데이터의 결측값를 추정하는 여러 기법들이 제안되었는데, 그 중 가장 널리 적용되는 기법으로 general linear model(GLM)을 들 수 있다. GLM은 결측이 발생한 지점의 강수량과 인근 지점들의 강수량 간의 관계를 결측이 발생하지 않은 시점의 실제 관측값들로부터 유도된 공분산행렬을 이용하여 결측값을 추정하는 기법이다. GLM은 상대적으로 간단할 뿐 아니라, 결측값을 추정하려는 위치에서 다른 시점에 측정된 관측값들을 이용한다는 장점이 있지만, 결측값을 추정하려는 시점에서의 강수량 간의 관

Received 27 October, 2012; **Revised** 29 December, 2012;
Accepted 12 February, 2013
***Corresponding author** : Chan Yong Sung, Department of Environmental Planning, Keimyung University, Daegu 704-701, Korea
Phone: +82-53-580-5913
E-mail: cysungg@kmu.ac.kr

계를 고려하지 못하는 단점이 있다. 즉 GLM은 모든 시점의 강수량이 독립동일분포(independent and identical distribution)라고 가정하는데, 우리나라와 같이 강수량의 공간분포가 계절에 따라 크게 변하는 지역에서는 이와 같은 가정이 위배될 때가 많다.

강수량 결측값을 추정하는 다른 방법으로는 공간통계기법(geostatistics)의 하나인 ordinary kriging(OK)이 있다. OK도 인근 지점의 측정값과의 관계에 바탕을 두고 결측값을 추정한다는 점에서 GLM과 유사하지만, GLM이 결측값이 발생한 시점과 다른 시점의 측정값들로부터 유도된 공분산행렬을 이용하는데 반해, OK는 결측값이 발생한 시점의 강수량의 인근 지점들에서 측정된 강수량의 공간적 분포로부터 공분산행렬을 유도한다는 차이가 있다(Jeffrey 등, 2001; Tabios와 Salas, 1985), 하지만 OK의 가정 또한 우리나라처럼 산지가 많고 지형이 복잡하여 강수량의 공간적 분포가 분명하게 나타나지 않는 지역에서는 만족되기 쉽지 않다(Bacchi와 Kottegoda, 1995; Beek 등, 1992).

본 연구에서는 GLM과 OK의 장점을 통합하는 새로운 결측강수량 추정기법을 제시하고, 이 기법의 신뢰성을 실제 강수량 데이터를 이용하여 검증해 보았다. 앞에서 설명했듯이 GLM과 OK 모두 결측값을 추정하려는 지점과 인근 지점의 측정된 강수량들의 공분산행렬을 각각의 기법 고유의 가정에 기반하여 유도한 다음, 이들 행렬을 이용하여 결측값을 추정한다. 본 연구가 제시하는 기법의 기본 개념은, GLM과 OK가 각각 다른 가정에 기반하여 유도한 공분산을 두 기법의 가정이 특정 시점의 실제 강수량 분포를 설명하는 정도에 비례하게 가중평균하여 새로운 공분산행렬을 유도한 다음 이를 이용하여 결측강수량을 추정하는 것이다. 본 연구에서는 새로운 추정기법을 이론적으로 설명하고, 이 기법의 신뢰성을 형산강 유역의 13개 관측소에서 2010년 1년 동안 관측한 일간 강수량 데이터를 이용하여 검증해 보았다.

2. 재료 및 방법

2.1. Ordinary Kriging을 수정한 새로운 결측강수량 추정기법

공간통계기법에는 공간상의 변수의 평균을 어떻게 가정하는지에 따라 simple kriging(SK), ordinary kriging(OK) 등으로 나눌 수 있다. 이 중 SK는 강수량의 평균이 알려져 있고 분석 대상지 내에서는 일정하다 가정하고 강수량을 추정하는 반면, OK는 평균은 모르지만 분석 대상지 내 일정하다 가정하고 강수량을 추정한다. 이 외에도 분석 대상지 내 강수량 평균이 어떤 추세가 있다고 가정하는 universal kriging (UK)이나 강수량에 영향을 미치는 다른 변수를 이용하는 cokriging (CK) 등이 있지만, 상대적으로 복잡하지 않으면서도 추정정확도가 높은 OK가 강수량 추정에 가장 널리 사용된다(Zimmerman 등, 1999; Schabenberger와 Gotway, 2005).

OK는 시점 t 에서 지점 s_0의 강수량 $\hat{Z}_t(s_0)$을 다음과 같이 정의한다.

$$\hat{Z}_t(s_0) = \lambda_t' \boldsymbol{Z_t}(s) \tag{1}$$

여기서 λ_t'는 시점 t 에 대한 $n \times 1$ 가중치 벡터, $\boldsymbol{Z_t}(s)$는 시점 t 에서 인근의 n 지점들의 강수량들로 구성된 $n \times 1$ 벡터이다. λ_t'는 불편추정조건 $\lambda_t' \mathbf{1} = 1$을 만족하고 예측오차의 분산 $Var\left[\hat{Z}_t(s_0) - Z_t(s_0)\right]$를 최소화하도록 하는 식 (2)를 풀어 얻는다.

$$\text{argmin } Q = Var\left[\hat{Z}_t(s_0) - Z_t(s_0)\right] - 2m\lambda_t' \mathbf{1} \tag{2}$$

여기서 $2m$은 라그랑지 승수이다. 식 (2)를 λ_t'와 m에 대해 편미분하면,

$$\frac{\partial Q}{\partial \lambda_t} = 2\boldsymbol{\Sigma_t}\lambda_t - 2\sigma^2 - 2ml = 0 \tag{3}$$

과

$$\frac{\partial Q}{\partial m} = 2\left(\lambda_t' \mathbf{1} - \mathbf{1}\right) = 0 \tag{4}$$

를 얻는다. 여기서 $\boldsymbol{\Sigma_t}$는 확률변수 $\boldsymbol{Z_t}(s)$ 간의 $n \times n$ 공분산행렬, σ^2는 $\boldsymbol{Z_t}(s)$와 $Z_t(s_0)$ 간의 $n \times 1$ 공분산 벡터이다. 식 (3)과 (4)를 연립하여 λ_t'를 얻은 후, 이를

식 (1)에 대입하면 $\hat{Z}_t(s_0)$을 추정할 수 있다.

여기까지 과정은 GLM과 OK가 동일하지만, 두 기법의 차이는 $\boldsymbol{\Sigma}_t$를 유도하는 방식에 있다. GLM은 결측값이 발생하지 않은 시점에서 관측된 강수량 데이터들을 이용하여 시계열 전체에 대해 하나의 $\boldsymbol{\Sigma}_{GLM}$을 유도하는 데 반해, OK는 결측값이 발생한 지점 인근에서 관측된 강수량들의 공간적 관계를 이용하여 결측값이 발생한 시점에 대해 각각 $\boldsymbol{\Sigma}_{OK_t}$를 유도한다. OK는 두 관측지점 간 강수량의 편차는 관측지점 사이의 거리의 함수라 가정하고 $\boldsymbol{\Sigma}_{OK_t}$를 유도하는데, 두 관측지점의 강수량 편차제곱을 semivariance라 하고, semivariance를 관측지점 간의 거리에 대한 표현한 함수를 semivariogram이라 한다. 일반적으로 semivariance는 관측지점 사이의 거리가 멀어짐에 따라 증가하다가 일정 거리에 도달하면 더 이상 증가하지 않는데, semivariance가 더 이상 증가하지 않는 지점까지의 거리를 range라 하고, range에 도달하였을 때의 semivariance 값을 sill이라 한다(Fig. 1). 이론적으로 semivariogram 모형은 원점을 지나야 하지만, 즉 같은 지점에서 측정된 강수량의 편차는 0이어야 하지만, 대부분의 경우 실제 측정된 강수량에서 유도된 semivariogram은 원점을 지나지 않는다. 거리가 0일 때의 편차가 생기는 이유는 강수량 측정 오류 등으로 발생하는데 이 편차를 nugget이라 한다. 다양한 형태의 semivariogram 모형 중, 본 연구에서는 실제 데이터와 가장 잘 일치하는 exponential 모형,

$$\hat{\gamma}(h)=\sigma_t^2\exp\left(-3\frac{h}{\phi_t}\right) \tag{5}$$

을 적용하였다. 여기서 $\hat{\gamma}(h)$는 시점 t 에서 거리 h 만큼 떨어진 두 지점에 대한 semivariance, σ_t^2와 $\hat{\phi}_t$는 시점 t에서의 파라미터들이다. 강수량 데이터의 2차 정상성(second order stationarity)을 가정하면 $\gamma_t(h)=C_t(0)-C_t(h)$가 되어, $\boldsymbol{\Sigma}_{OK_t}$의 행렬요소인 거리 h 만큼 떨어진 두 지점간의 공분산 $C_t(h)$를 유도할 수 있다.

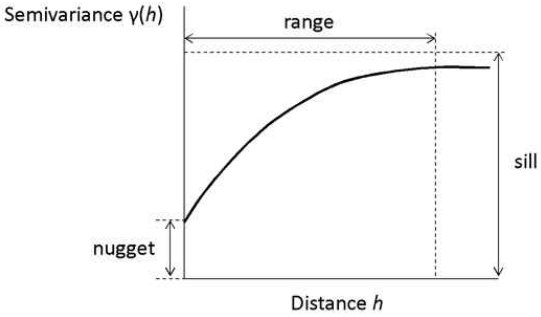

Fig. 1. Conceptual diagram of a semivariogram.

본 연구가 제시한 수정OK는 GLM과 OK에서 유도된 공분산행렬들을 두 기법의 추정 신뢰도에 비례해서 가중 평균하여 새로운 공분산행렬 $\boldsymbol{\Sigma}_{t,\,MidifiedOK}$를 얻은 다음, 이를 식 (3)에 대입하여 강수량을 추정한다(Sung, 2012). 각 기법의 추정 신뢰도는 각 모형에 대해 Nash와 Sutcliffe의 모형적합도(R^2)를 이용하여 계산한다(ASCE, 1993). OK는 결측이 발생한 모든 시점에 대해 각각 하나씩 $\boldsymbol{\Sigma}_{OK_t}$을 유도하기 때문에 모형적합도($R^2_{OK_t}$)도 추정하려는 결측값 수만큼 계산되지만, GLM은 분석 기간 전체에 대해 하나의 $\boldsymbol{\Sigma}_{GLM}$만을 유도하기 때문에 모형적합도($R^2_{GLM}$)도 하나만 계산된다.

2.2. 추정기법 신뢰도 검증

본 연구에서는 수정OK의 결측강수량 추정의 신뢰도 검증을 위해, 형산강 유역의 중앙에 위치한 검단 관측소(관측소 코드: 21014080)에서 실제로 관측한 일간 강수량을 GLM과 OK, 수정OK를 이용하여 각각 추정한 후 추정정확도를 비교하였다. 분석을 위해 2010년 1월 1일부터 12월 31일까지 1년간 검단관측소와 인근 12개 관측소에서 측정한 일간 강수량 데이터를 국가수자원관리종합정보시스템(http://www.wamis.go.kr)에서 얻어 사용하였다. 2010년 1년 동안의 일평균 강수량은 검단관측소가 2.45 mm였고, 13개 관측소 중 강수량이 가장 적었던 곳은 기북관측소의 2.01 mm, 가장 많았던 곳은 덕동관측소의 2.74 mm였다(Table 1). 형산강 유역은 경상북도 포항시, 경주시, 경상남도 울산시에 걸쳐 있고, 유역의 전체 면적은 1,395 km^{2}이다(Fig. 2).

Table 1. Average daily precipitation in Geomdan and 12 neighboring rain gages

Rain gages	Gage code	Daily average precipitation (mm)	Precipitation 7.7.2010 (mm)	Precipitation 7.21,2010 (mm)
Boolgooksa	21014120	2.45	102	21
Cheongbook	21014110	2.6	76	34
Deokdong	21014130	2.74	87	29
Doodong	21014090	2.61	97	28
Geomdan	21014080	2.66	111	16
Geoncheon	20014040	2.07	97	35
Gibook	21014050	2.01	62	39
Gige1	21014010	2.02	80	55
Gige2	21014100	2.14	65	49
Gyeongju1	21014020	2.33	95	15
Gyeongju2	21014070	2.2	74	40
Oksan	21014060	2.21	64	55
Pohang	21014140	2.61	69	40

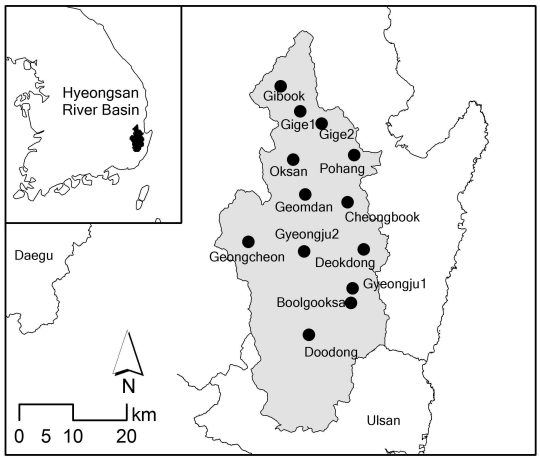

Fig. 2. Location of Geomdan and 12 neighboring rain gages.

수정OK의 신뢰도는 다음의 과정을 통해 검증하였다. 우선 GLM의 모형적합도(R^2_{GLM})를 분석기간 전체에 대해 교차검증을 통해 계산하였다. 교차검증이란 전체 대상기간 중 하루 검단 측정소의 일간 강수량이 결측되었다고 가정하고 나머지 관측값으로부터 유도된 GLM모형을 이용하여 그날의 강수량을 추정하여 이 추정값을 실제 관측값과 비교하는 과정을 의미한다. 본 연구에서는 전체 365개 관측값에 대해 교차검증을 실시하여 GLM의 모형적합도(R^2_{GLM})를 계산하였다. 다음 대상기간 내 각각의 시점에 대해 OK의 semivariogram의 모형적합도($R^2_{OK_i}$)를 계산하고 이를 R^2_{GLM}과 비교하여 $\boldsymbol{\Sigma_{GLM}}$과 $\boldsymbol{\Sigma_{OK_i}}$를 가중평균한 $\boldsymbol{\Sigma}_{t,\,ModifiedOK}$를 얻은 다음, 이를 이용하여 검단 관측소의 일간 강수량을 추정하였다. 인근 12개 지점의 강수량 데이터가 모두 0이면 식 (5)의 파라미터 값이 결정되지 않기 때문에, 그 날의 검단 관측소의 강수량도 0으로 추정하였다. 모든 분석은 강수량 데이터에 1을 더한 후 로그변환하여 시행하였다. 수정OK의 결측 강수량 추정 신뢰도를 분석하기 위해, 전체 시점에 대한 모형적합도 $R^2_{ModifiedOK}$를 계산하고 이를 $\boldsymbol{\Sigma_{GLM}}$와 $\boldsymbol{\Sigma_{OK}}$와 비교하였다. $\boldsymbol{\Sigma}_{t,\,OK}$은 R언어의 geoR 패키지을 이용하여 유도하였다 (Ribeiro와 Diggle, 2001).

3. 결과 및 고찰

수정OK는 GLM이나 OK보다 정확하게 일간 강수량 데이터의 결측값을 추정하는 것으로 나타났다. 검단측정소에 대한 일간 강수량 추정 결과를 보면, 수정

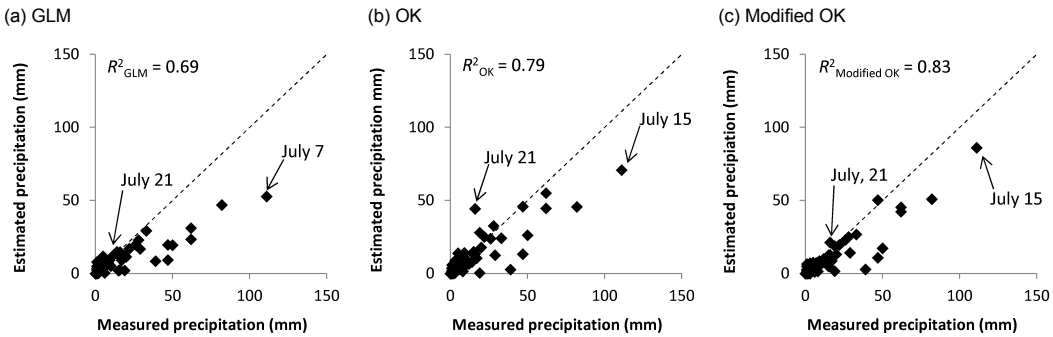

Fig. 3. Estimated daily precipitation values by GLM, OK, and Modified OK.

OK의 모형적합도가 0.83으로 GLM의 0.69나 OK의 0.79보다 높았다(Fig. 3). 특히 수정OK는 강수량이 많은 날 기존 추정법보다 더 정확하게 강수량 결측값을 추정하는 것으로 나타났다. 일반적으로 수문분석은 비가 많이 올 때 더 필요하다는 점을 고려하면, 강수량이 많은 날에 오차가 적다는 것은 수정OK의 큰 장점이라 하겠다.

보다 자세한 분석을 위해, OK의 추정오차가 상대적으로 큰 7월 21일과 상대적으로 적은 5월 21일에 대한 semivariogram을 비교하여 보았다(Fig. 4).

7월 21일은 OK의 추정오차가 GLM의 추정오차보다 큰 날이었다. 이날 13개 관측소의 평균 강수량은 35 mm였고, 검단관측소의 강수량은 평균보다 낮은 16 mm였다(Table 1). Fig 4a에서 볼 수 있듯이 이날은 측점 지점 사이의 거리와 semivariance 간의 관계가 명확하게 나타나지 않았고, semivariogram의 $R^2_{OK_{7월21일}}$도 0.14로 낮았다. 반면 GLM은 이날의 강수량을 21.3 mm로 추정하여 오차가 OK보다 적었기 때문에, $\Sigma_{t,OK}$보다 Σ_{GLM}을 더 많이 가중하여 $\Sigma_{t,ModifiedOK}$을 유도하였다.

이에 반해 7월 7일은 OK의 추정오차가 GLM에 비해 상대적으로 적은 날이었다. 이날은 분석 대상기간 중 강수량이 가장 많은 날로, 13개 지점의 평균 강수량은 83 mm, 검단관측소의 강수량은 111 mm였다(Table 1). 이날의 GLM의 추정강수량은 52.8 mm이고, 7월 21일에 비해 OK의 추정강수량은 70.8 mm여서 OK의 추정오차가 상대적으로 적었다. Fig 4b의 semivariogram 또한 이날의 강수량은 관측소간 강수량 차이가 거리가 비례해 커졌음을 보여준다($R^2_{OK_{7월7일}}$=0.37). 따라서 이날은 7월 21일 보다 $\Sigma_{t,OK}$에 더 많은 가중치를 주어 $\Sigma_{t,ModifiedOK}$를 유도하였다.

4. 결 론

본 연구에서는 공간 데이터로부터 공분산행렬을 유도하는 OK와 시계열 데이터로부터 공분산행렬을 유도하는 GLM을 통합하여 강수량 결측값을 추정하는 수정OK을 제시하고, 이 기법의 신뢰도를 실제 데이터를 가지고 검증해 보았다. 분석 결과 수정OK는 GLM이나 OK보다 더 정확하게 강수량 결측값을 추정하는 것으로 나타났다. 이 결과는 GLM이나 OK의 기본 가정이 실제 측정된 강수량의 분포와 다른 때가 많아 둘 중 하나를 이용하여 추정된 강수량의 오차가 큰 반면, 수정OK는 두 기법의 기본 가정 중 실제 강수량 분포와 더 유사한 기법에 더 많은 가중치를 주어 공분산을 추정하기 때문에 추정의 정확도가 높기 때문으로 분석되었다. 즉, 수정OK는 결측값이 발생한 시점의 강수량의 분포가 GLM의 가정에 위배될 때는 OK에 가중치를 주어 공분산을 유도하고, 반대로 OK의 가정에 위배될 때는 GLM에 가중치를 주어 공분산을 유도하기 때문에, GLM이나 OK보다 더 정확한 추정값을 얻을 수 있는 장점이 있었다.

마지막으로 본 연구가 제안한 수정OK의 정확도를 개선하기 위한 향후 연구과제를 제안하고자 한다. 수

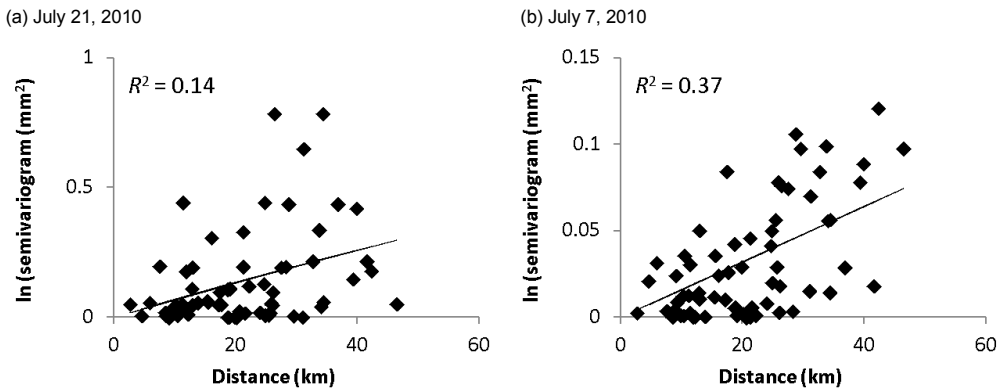

Fig. 4. Empirical semivariogram clouds with fitted semivariogram models.

정OK는 결측값의 추정과 같이 추정하려는 지점에 다른 관측값이 있을 때에만 적용할 수 있다. 이는 GLM이 결측값이 발생한 지점의 강수량과 인근 지점의 강수량간의 공분산을 실제 관측값에서 유도하기 때문으로, 가용한 데이터를 최대한 이용한다는 점에서는 수정OK의 장점으로 볼 수 있기도 하지만, 결측이 발생한 지점의 시계열 데이터가 없는 경우에는 적용할 수 없다는 점에서 모형의 한계라 할 수 있다. 따라서 앞으로 이러한 한계를 극복하여 관측값이 없는 지역에 대한 강수량 추정이 가능하도록 수정OK를 확장하는 연구가 요구된다. 예를 들어, 강수량 데이터 외 측정가능한 여러 기후학적 변수를 이용하여 강수량을 추정하려는 지점과 인근 지점들 간의 강수량의 공간적 분포를 나타내는 공분산행렬을 유도하고 이를 실제 강수량 데이터로부터 유도한 공분산행렬과 결합하는 기법을 개발한다면, 강수량 관측값이 없는 지점에 대해서도 수정OK와 비슷한 추정기법을 적용할 수 있을 것이다.

또한 본 연구가 제시한 수정OK를 다양한 기후와 지역에 대해 검증해 보아야 할 것이다. 수정OK의 장점은 GLM이나 OK 중 기본 가정이 특정 시점에 대한 강수량의 공간적 분포와 더 잘 맞는 기법을 선택적으로 가중하여 결측값을 추정한다는 데 있다. 따라서 수정OK의 적용가능성은 지형이나 계절 등에 따라 달라질 것으로 예측된다. 수정OK가 어떤 계절, 그리고 어떤 기후적 특성이 있는 지역에 적합한지에 대한 경험적 연구가 요구된다. 본 연구가 앞으로 공간데이터와 시계열데이터를 결합하여 강수량을 추정하는 기법을 개발하는 데 도움이 될 수 있기를 기대한다.

참 고 문 헌

ASCE Task Committee on Definition of Criteria for Evaluation of Watershed Models of the Watershed Management Committee, Irrigation and Drainage Division, 1993, Criteria for evaluation of watershed models, J. Irrig. Drain. E.-ASCE, 119, 429-442.

Bacchi, B., Kottegoda, N. T., 1995, Identification and calibration of spatial correlation patterns of rainfall, J. Hydrol., 165, 311-348.

Beek, E. G., Stein, A., Janssen, L. L. F., 1992, Spatial variability and interpolation of daily precipitation amount, Stoch. Hydrol. Hydraul., 6, 304-320.

Choi, Y. J., Kim, Y. S., Lee, G. H., Kim, J. C., 2010, The verification of application of distributed runoff model according to estimation methods for the missing rainfall data, J. Environ. Sci., 19, 1375-1384.

Jeffrey, S. J., Carter, J. O., Moodie, K. B., Beswick, A. R., 2001, Using Spatial interpolation to construct a comprehensive archive of Australian climate data, Environ. Modell. softw., 16, 309-330.

Michaud, J. D., Sorooshian, S., 1994, Effect of rainfall-sampling errors on simulations of desert flash floods, Water Resour. Res., 30, 2765-2775.

Ribeiro, P. J., Diggle, P. J., 2001, geoR: a package for geostatistical analysis, R-NEWS 1, 15-18.

Schabenberger, O., Gotway, C. A., 2005, Statistical methods for spatial data analysis, Chapman & Hall, Boca Raton, FL.

Sung, C. Y., 2012, Estimating missing rainfall data in urban areas using hybrid approach of geostatistics and generalized least square estimation, J. Nakdong River Environ. Res. Inst., 16, 301-320.

Sung, C. Y., Li, M. H., 2010, The effect of urbanization on stream hydrology in hillslope watersheds in central Texas, Hydrol. Process., 24, 3706-2717.

Tabios, G., Salas, J. D., 1985, A comparative analysis of techniques for spatial interpolation of precipitation, Water Resour. Bull., 21, 365-380.