비매개변수적 다변량 핵밀도 추정법을 이용한 수문학적 응용

○ 차영일[1], 문영일[2], 최채복[3]

1. 서 론

최근의 이상기후로 인하여 수문량에 대한 정확하고 적절한 분석은 더욱 중요해지고 있다. 수문학자들은 통계적인 추정 문제로써 지하수 수위나 강수, 다른 매개변수 사이의 함수적인 관계로써 강우와 유출 그리고 저수지로 유입되는 월 유량과 같은 추계학적인 시간계열의 모의발생에 대하여 많은 관심을 가져 왔다. 그러나 대부분의 경우 하나의 변량만을 갖고 있지 않기 때문에, 여러 개의 변량을 갖는 결합 확률밀도 함수의 형태를 보고 해석하는 것이 타당할 것이다. 지금까지의 일반적인 방법은 자료에서 매개변수적인 수문 모형을 적용해온 것이 사실이다. 이런 매개변수적인 접근은 매개변수의 구조를 우연히 알 수 있다면 매우 정확하고 합당하고 효과적일 수 있다. 그러나 이런 매개변수의 구조를 알아낸다는 것은 어려운 일이며, 또한 상대적으로 부족한 자료를 이용하여 공간과 시간적인 관계를 고려한 다변량 매개변수의 구조를 알아낸다는 것은 매우 힘든 일이다. 따라서, 지금까지의 매개변수적 수문분석은 그 해석적인 적용의 어려움으로 다변량 분석보다는 단변량 분석에 치중하는 경향이 있었다. 그러나 최근의 이상기후 상황에서는 한가지 원인과 결과의 수문량에 의한 수문분석 보다는 여러 가지의 수문량을 변량으로 이용하는 것이 더욱 정확할 것이며 또한 수문분석 도구로서 여러 가지 이점이 있다 할 수 있다. 특히, 우리나라와 같이 수문 자료가 충분히 구축되지 않은 상태에서 그 수문 자료의 시간적 공간적인 구조를 해석할 경우 비매개변수적 다변량 핵밀도 추정법(nonparametric multivariate kernel density estimation)을 이용하면 매개변수적 방법으로 할 수 없는 많은 응용과 그에 따른 적절한 결과를 도출해 낼 수 있을 것이다. 따라서, 본 연구에서는 비매개변수적 다변량 핵밀도 추정법에 대한 특성을 살펴보고 임의의 수문자료를 사용하여 적용성을 분석하였다.

1) 서울시립대학교 토목공학과 박사과정
2) 서울시립대학교 토목공학과 부교수
3) 서울시립대학교 토목공학과 석사과정

2. 다변량 핵밀도함수

단변량 핵밀도함수는 모든 실수 x에 대하여 식 (1)과 같이 정의된다 (Silverman, 1986).

$$f(x)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-X_i}{h}\right) \tag{1}$$

여기서, X_1, X_2, ……, X_i는 실관측치, K(·)는 단변량 핵함수, h는 광역폭이다. 그러나, 다변량 핵밀도함수 f(u)는 식 (2)와 같이 벡터의 형태로 d차원으로 나타내어진다(Silverman, 1986).

$$f(u)=\frac{1}{n}\sum_{j=1}^{n}K(y) \tag{2}$$

여기서,

$$y=\frac{(u-u_i)^T S^{-1}(u-u_i)}{h^2} \tag{3}$$

K(y)는 다변량 핵함수이고, $u=[u_1,u_2, \ldots ,u_d]^T$는 가정된 밀도함수를 가지는 d차원의 확률 벡터이다. $u=[u_1,u_2, \ldots ,u_{di}]^T$는 n 표본 벡터이고, h는 핵함수의 광역폭, 그리고 S는 d×d인 u_i의 공분산 행렬이다. 본 연구에서는 다음과 같이 주어지는 다변량 Gaussian 확률밀도함수 K(u)를 사용하였다.

$$K(y)=\frac{1}{(2\pi)^{d/2}h^d\det(S)^{1/2}}\exp(-y/2) \tag{4}$$

K(y)의 값은 u와 u_i간의 거리에 근거한 관측치 u_i에 주어지는 가중치를 표현한다. 따라서 최종적으로 비매개변수적 다변량 핵밀도 추정법은 다음 식 (4)로 정의된다.

$$f(u)=\frac{\det(S)^{-1/2}}{nh^d}\sum_{i=1}^{n}K\left(h^{-2}(u-u_i)^T S^{-1}(u-u_i)\right) \tag{5}$$

여기에 사용된 거리는 공분산을 이해하기 위해 제안된 Euclidean 거리이다. 식 (2)로 부터 핵함수는 추정지점의 인접부에서 관측치의 상대빈도에 대한 가중평균이다. 핵함수 K(·)는 상대적인 가중치를 h는 계산된 평균값에 걸친 광역폭의 범위를 규정짓는다. 공분산 행렬 S의 역할은 좌표상에서 선형 상관의 가능여부를 인식하게 한다. 이는 핵함수의 결과를 적당히 가늠하게 할 수 있게 하고, 회전된 좌표에서 변동 비율의 양에 있어서 폭을 다양하게 한다.

광역폭 h를 선택하는 많은 방법들이 있지만 통계 문헌에 제시된 가장 좋은 방법으로는 d=1에서 선택하는 Sheather와 Jones(1991)방법과, d=2에서 선택하는 Wand와 Jones(1994)의 방법을 들 수 있는데, 계산의 부담으로 인해 f(u)에서 MISE(mean integrated square error)를 최소화함으로써 광역폭을 자동적으로 선택할 수 있는 방법도 제시되었다. 이것은 이론적으로 최선의 선택은 아니지만, 그 수행에 있어서는 좀더 까다로운 선택방법과 비교해 볼 때, 계산시간에 있어서 매우 효율적이다. 근사적인 최적의 광역폭은 다음식의 MISE를 최소로 하는 값이다.

$$h_{opt}^{d+4} = d\beta\alpha^{-2}\left\{\int(\nabla^2 f)^2\right\}^{-1} n_1 \tag{6}$$

여기서 f는 표준분포(standard density)이고 상수 α, β는 다음과 같다.

$$\alpha = \int t_1^2 K(t)dt \ , \quad \beta = \int K(t)^2 dt \tag{7}$$

위의 식 (6)에서 만약 Φ가 단위 d변량 정규분포라면 다음식과 같다.

$$\int(\nabla^2\Phi)^2 = (2\sqrt{\pi})^{-d}\left(\frac{1}{2}d + \frac{1}{4}d^2\right) \tag{8}$$

식 (8)에 의해서 주어진 값은 단위 분산을 가지는 정상적으로 분포된 자료의 완화(smoothing)를 위한 최적의 광역폭을 구하기 위해 식 (6)에 대입한 후, MISE를 최소로 하는 Gaussian 핵함수의 최적의 광역폭은 다음과 같다(Silverman, 1986).

$$h = \left(\frac{4}{d+2}\right)^{\frac{1}{d+4}} \times n^{-\frac{1}{d+4}} \tag{9}$$

3. 적용 및 결과

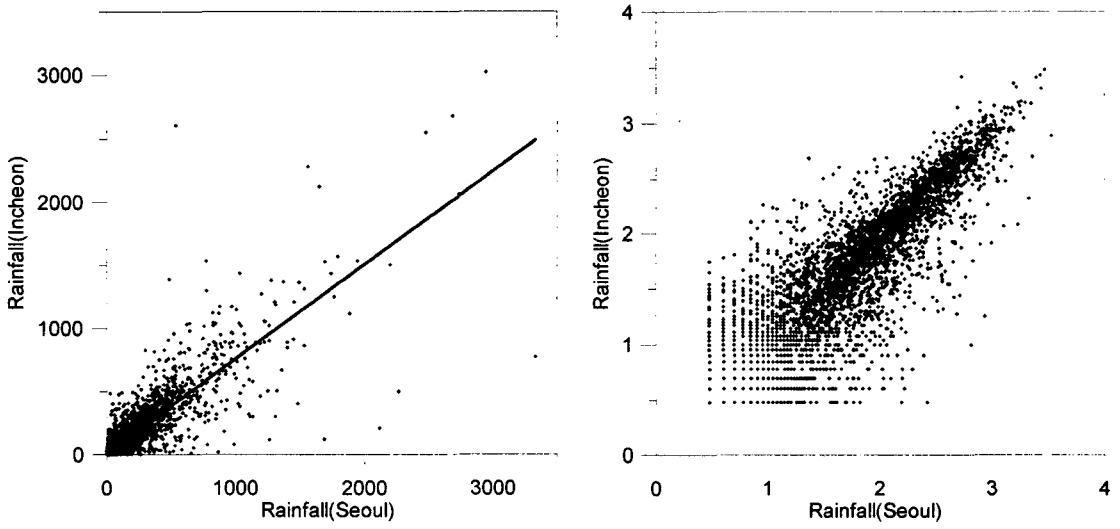
서론에서 전술한 것과 같이 비매개변수적 다변량 핵밀도 추정법은 공간-공간, 시간-공간 등 여러 가지 형태의 변량을 분석할 수 있다. 그러나, 이번 연구에 적용된 자료는 자료의 획득이 비교적 용이하고 신뢰성이 있는 기상청산하 관측소의 자료 중 거리상으로 인접하지는 않았지만 비교의 목적으로 서울지점과 인천지점의 일 강우에 대하여 적용하였다. 두 지점의 일 강우 자료 중 모두 0.1mm이하는 무강우로 간주하여 그 이상의 자료를 가지고 수문 계열을 작성하여 분석에 이용하였고 그 내용은 다음 표 1과 같다.

그림 1은 두 수문계열의 자료를 도시한 그래프로 x축은 서울 강수량, y축은 인천 강수량이고 그림 1의 (b)는 자료의 도시를 위하여 상용 로그를 취해 변환한 후 도시하였다.

강우자료 계열의 모의발생을 위한 강우모형은 크게 지점 강우모형(point rainfall models)과 다변량 강우모형(multivariate rainfall model) 등으로 구분되어 질 수 있다. 지점 모의발생모형의 경우는 대상 지점의 한 지점에 대한 강우특성을 해석하는 데는 용이 할 수 있겠으나, 유역전반이나 인근 지점의 특성을 해석하는데는

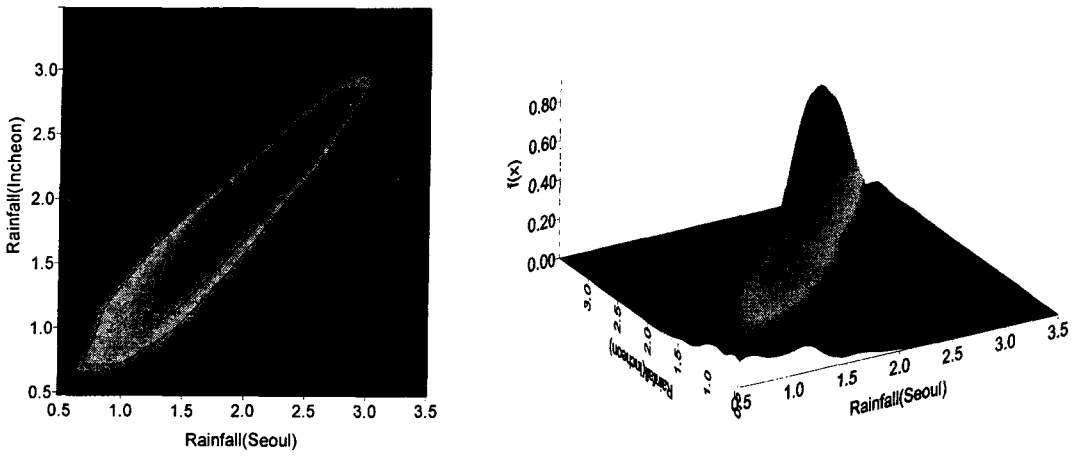
표 1 지점별 강우자료

유 역	지점명	자료관측기간	자료크기	총 자료수
한 강	서 울	1954 ~ 1999	46년	3822
	인 천	1954 ~ 1999	46년	



(a) 강수량분포(0.1mm)와 회귀선 (b) 변환자료

그림 1 서울지점과 인천지점의 강수일에 대한 일 강수량 분포



(a) Contour Map (b) 3D Map

그림 2 서울과 인천의 결합확률밀도함수

많은 어려움이 따른다. 따라서 유역내 각 지점의 공간적 특성을 고려할 수 있는 다변량 강우모형을 이용하여 해석을 하는 것이 좋을 것이다. 그러나, 다변량 강우모의발생모형은 각각의 확률분포형과 회귀식을 구하여 모의 발생을 실시하여야 하는 어려움이 있고 강우의 분포들을 선형적으로 가정하여야 했다. 이런 다지점 강우모의발생 모형의 경우 다변량 핵밀도 추정법을 이용한다면 해석이 용이하고 적용이 쉬울 것이다.

그림 1의 (a)는 선형이 아닌 강우를 선형으로 가정하는 기존의 회귀식을 이용한 방법이고, 그림 2는 비매개변수적 다변량 핵밀도 추정법을 이용하여 비선형인 두 지점간의 강우의 결합확률밀도함수를 구하여 강수량을 모의 발생하는데 적용할 수 있다. 그림 1과 그림 2의 모든 경우에서 서울강수량의 크기가 커지면 인천강수

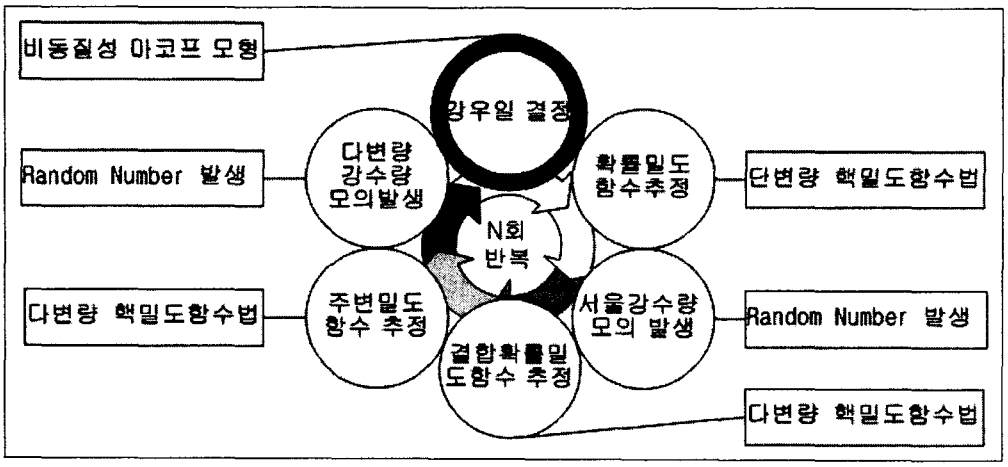


그림 3 다변량 핵밀도함수 추정법에 의한 강수량 모의발생 절차

량도 같이 증가하는 것을 볼 수 있다. 그림 1은 선형 회귀분석을 실시하여만 해석이 가능하지만, 그림 2는 그 자체가 결합확률밀도 함수가 되기 때문에 다른 회귀분석 등의 절차가 필요 없이 유역내 각 지점의 공간적 특성을 고려하여 해석이 가능하다. 그림 3은 다변량 핵밀도 추정법에 의한 다지점강수량 모의발생에 대한 간단한 절차를 나타낸 것이다. 다변량 핵밀도 추정법에 의한 강수량 모의발생은 먼저 비동질성 마코프 모형을 이용하여 강수일을 결정한 후, 강수일에 대한 단변량 핵밀도함수의 의해 1차 지점의 강수량을 모의하고, 다변량 핵밀도함수에서 그 강수량에 대한 비매개변수적 주변밀도함수를 설정하여 최종적으로 원하는 강수량을 모의발생 시킨다.

4. 결론

수문분석에 있어서 대부분의 경우 수문계열은 하나의 변량만을 갖지 않고 여러 개의 변량을 갖기 때문에 다변량 분석이 필요하다. 그러나, 지금까지의 매개변수적 방법은 수학적으로 다변량 결합확률밀도 함수를 구하기가 어려웠기 때문에 선형적인 회귀식과 같은 방법으로 비선형인 수문변량을 해석해온 것이 사실이다. 이런 경우 본 연구에서 예로 든 다지점 강우모의발생과 같이 다변량 핵밀도함수 추정법을 이용한다면 지점간의 강우사상을 선형으로 가정하지 않고도 비선형적인 방법으로 자료의 특성을 왜곡하지 않고 그대로 모의 발생할 수 있을 것이다. 또한 비매개변수적 다변량 핵밀도 추정법은 공간-공간, 시간-공간 등의 자료계열에 대한 제약이 거의 없이 바로 적용이 가능하고 선형성 등과 같은 가정이 필요없이 자료계열의 특성이 그대로 적용되므로, 기존의 방법보다 합리적이며 적용성에 있어서도 단순하다 할 수 있겠다. 따라서 앞으로의 수문학적인 분석에 있어서 기존의 방법과 병용하여 분석을 실시한다면 더 좋은 결과를 얻을 수 있을 것이다.

5. 참고문헌

Sheather, S. F., and Jones, M. C.(1991). "A reliable data-based band-width selection method for kernel density estimation", J. Roy. Statistical Soc. 53(B):683~690.

Silverman, B. W.(1986), "Density estimation for statistics and data analysis." New York: Chapman and Hall.

Wand, M. P. and M. C. Jones(1993). "Comparison of smoothing parametrization in vivariate kernel density estimation." JASA 88(422):520-528.