韓國水資源學會論文集
第36卷 第6號 · 2003年 12月
pp. 1025~1033

수위-유량곡선을 위한 비매개 변수적 Kernel 회귀모형

Nonparametic Kernel Regression model for Rating curve

문 영 일* / 조 성 진** / 전 시 영***

Moon, Young-Il / Cho, Sung Jin/ Chun, Si Young

Abstract

In common with workers in hydrologic fields, scientists and engineers relate one variable to two or more other variables for purposes of predication, optimization, and control. Statistics methods have improved to establish such relationships. Regression, as it is called, is indeed the most commonly used statistics technique in hydrologic fields; relationship between the monitored variable stage and the corresponding discharges(rating curve). Regression methods expressed in the form of mathematical equations which has parameters, so called parametric methods. some times, the establishment of parameters is complicated and uncertain.

Many non-parametric regression methods which have not parameters, have been proposed and studied. The most popular of these are kernel regression method. Kernel regression offer a way of estimation the regression function without the specification of a parametric model.

This paper conducted comparisons of some bandwidth selection methods which are using the least squares and cross-validation.

Keyword : nonparametric, regression, hydrologic application, rating curve.

요 지

수공구조물의 설계를 비롯하여, 수자원 분야의 기술적 설계의 기초는 수문자료의 처리와 분석에 중심을 두고 있다고 할 수 있다. 수문 자료의 분석방법 중 가장 보편적이면서도 중요한 방법은 자료들의 관계를 도식적으로 규명하는 회귀분석이다. 수위-유량 관계곡선과 같은 수문 자료에 대한 기존의 매개변수적 회귀모형이 갖는 단점은 자료의 특성에 따라, 복수의 회귀식이 산정되거나 동일자료에 대해서도 서로 다른 회귀식이 산정됨으로써 신뢰할 수 있는 회귀곡선을 만들기가 어렵다는 것이다. 이에 비해 주어진 자료에 의해 도출되는 kernel 회귀모형은 자료의 특성과 경향성을 적절히 표현해 줄 수 있는 방법이다.

본 논문에서는 비매개변수적 방법인 kernel 회귀모형을 분석하고, kernel 회귀모형의 중요 인자인 bandwidth의 선택 방법에 따른 kernel 회귀모형의 특성에 대해 비교 분석하였다.

* 서울시립대학교 도시과학대학 토목공학과 부교수
Associate Prof., Dept. of Civil Engineering, University of Seoul, Seoul, Korea, 130-743
(E-mail : ymoon@uos.ac.kr)

** 서울시립대학교 토목공학과 박사수료
Ph.D. Student, Dept. of Civil Engineering, University of Seoul, Seoul, Korea, 130-743

*** 원광대학교 공과대학 토목환경도시공학부 교수
Professor, Division of Civil, Environmental & Urban Engineering, Wonkwang University, Iksan, Chonbuk, Korea, 570-749

핵심용어 : 비매개변수적, 회귀모형, 수문학적 응용, 수위-유량곡선

1. 서 론

수문학에서 다루게 되는 폭풍과 강우, 수위와 유량, 지하수오염과 비점오염원 등과 같은 일련의 사건에 대하여, 그 사건을 지배하는 인자들간의 관계를 규명하는 것은 공학적으로 매우 중요한 일이다. 이러한 변량들의 관계를 해석하는 가장 일반적인 방법이 회기분석(regression)이다. 회기모형의 경우 일반적으로 하나의 독립변수를 갖는 단순회기모형(simple regression)과 복수의 독립변수를 갖는 다중회기모형(multiple regression)으로 나누게 된다. 단순회귀모형은 아래와 같은 식으로 나타낼 수 있다.

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad (1)$$

여기서, X , Y 는 각기 독립과 종속변수이며 β_0, β_1 은 회기 상수, ε 은 오차항이다. 이러한 모형의 회기분석은 자료로부터 도출되는 회기상수로부터 회귀식을 만들어 이 회귀상수와 회귀식으로 자료의 특성을 표현하게 된다. 이러한 모형을 일반적으로 매개변수적 회기분석이라고 한다. 매개변수적 회귀분석의 경우 정확한 가정, 즉 회귀상수의 도입이 수반되어야 하며, 또한 이질적인 자료의 추정이 힘든 경우가 있고, 이러한 경우 단일모형으로는 적정한 모형을 구함에 있어 어려움이 따른다. 하천의 수위-유량곡선의 작성이 대표적인 예이며, 지역적인 지하수위 해석 및 오염도 해석 등의 적용에 무리가 있을 수 있다(조성진 등, 1999).

2. 매개변수적 회귀모형의 적용

기존에 사용중인 매개변수적 회귀모형의 수문학적 응용은 서론에서 언급한 바와 같이 다양하다. 매개변수적 회귀모형을 직접적으로 적용하는 경우는 수위-유량 관계곡선(stage-discharge relation curve)이 대표적이라고 할 수 있다.

수위-유량관계곡선은 적정한 관계식으로 나타내게 되는데 유량관측소에서의 연속적인 수위기록으로부터 연속적인 유량자료를 필요로 할 경우에 수위-유량 관계곡선을 수립하게 된다. 다만 수위와 유량의 관계를 일반적인 수위-유량관계식으로 나타내는 것은 부정확성을 내포하게 되며, 그 요인을 몇 가지로 나누면 다음과 같다(윤태훈, 1997).: (i)수위관측소 지점에서 하상의 세굴, 퇴적, 준설 등으로 조절단면의 불안정성 (ii)수위 유량관계에 영향을 줄 수 있는 식생 등으로 인한 계절적인 변화의 영향 (iii)홍수파의 급상승과 급하강에 따른 지연효과 (iv)조석작용으로 인한 흐름의 변화 (v)수위 및 유량의 측정오차 등이다. 이와 같은 상황에서는 수위-유량관계를 수정할 필요가 있으며 수위만 가지고 유량이 직접 결정되지 않고 수면경사와 같은 제 3의 매개변수를 통해서 유량이 결정되는 경우가 존재한다. 위에서 언급한 변수들을 포함하는 다중 회귀분석(multiple regression)을 적용하여 단수회귀모형의 부정확성을 보완 할 수도 있으나, 수위-유량 관계식의 활용면에서 어려움이 있다고 할 수 있다.

일반적으로 하천의 수위-유량 관계곡선은 다음과 같은 식(2) 또는 식(3)으로 나타낼 수 있다.

$$Q = a\,(H+b)^n \qquad (2)$$

$$Q = cH^2 + dH + e \qquad (3)$$

여기서, Q는 유량 (m^3/s), H는 수위를 나타내며 a, b, c, d, e, n 은 지역에 따라 변하는 상수 값으로 회귀분석을 통하여 구해야 하는 상수들이다. 만약 수위-유량 관계곡선이 산술 좌표에서 단순한 곡선으로 도시되지 않는다면, Q와 H의 값을 대수좌표에 도시하는 것이 편리하다. 수위-유량방정식의 대수 형태는 식(2)의 경우 다음과 같이 표현되며, 이 식은 직선으로 도시될 수 있으므로, 기울기의 변화가 더욱더 분명하게 나타난다.

$$\log Q = \log a + n \log (H+b) \qquad (4)$$

수위-유량관계곡선의 선정은 식(2)와 식(3) 또는 다른 관계식을 이용한 회귀분석을 통하여, 그 식들 중 상관계수가 가장 큰 것을 그 지점의 수위-유량 관계식으로 결정하는 것이 일반적 방법이다. 만약 점들이 매끄러운 단일 곡선을 그리지 못한다면, 수위-유량관계를 지배하는 수로조절 특성에 변화가 있음을 의미하며 대부분의 자연하천에서 발생되는 현상이다.

단면의 불규칙성과 하상 변동 등으로 인하여 관측자

료를 하나의 관계식으로 표시하기 곤란할 때는 단면이 급격히 변하거나 하상 변동이 예상되는 수위를 기준으로 하여 2~3개의 곡선으로 나누어 표현하는 것이 일반적이다.

따라서, 수위-유량 관계식을 하나의 일관성 있는 회귀식으로 나타낼 수 없을 때 불규칙성이 나타나는 구간이나 수위에 따라서 각각의 회귀식을 구하여 사용해 오고 있다.

이러한 매개변수적 회귀모형은 고수위와 저수위의 수리특성을 반영하여, 매개변수식을 작성하게 되므로, 지배적인 특성을 최대한 회귀식에 적용하는 것이 가능하여, 분석된 회귀식은 수위로부터 유량을 산정하는데 있어서 적용이 편리하다. 그러나, 고수위와 저수위부의 회귀식을 분리 적용하는 불편함과, 관계식을 여러 구간으로 나누어 회귀분석 할 경우에 전체적인 자료의 특성을 반영하지 못하게 되고 경향성(trend)을 저해하며 각 구간 또는 수위에 따라 다른 회귀식을 적용해야 하는 불편함을 가지고 있다.

3. Kernel 회귀모형의 특성

3.1 Kernel 회귀모형

비매개변수적 회귀분석(non-parametric regression)은 매개변수를 통하지 않고 자료로부터 알고자하는 지점의 추정값을 추정한다. 이러한 비매개변수적 회귀분석은 주어진 자료의 특성으로부터 잡음(noise)을 제거 또는 감소시킬 수 있으며, 이로 인하여 자료의 해석에 있어 보다 원자료에 근접하는 회귀모형을 구할 수 있다는 장점을 지닌다. 비매개변수적 회귀분석은 기본 회귀분석이 해석하기 어려운 자연계의 이질적이고, 다중변수, 시간과 공간적인 변수를 지니게 되는 자료들에 대한 유용한 해석방법이라고 할 수 있다(Wand 등, 1995).

자료로부터 회귀모형을 산정해내는 비매개변수적 회귀모형은 원자료의 분포특성을 최대한 회귀할 수 있는 장점을 지니고 있으나, 기존 방법에 비하여 이론적 접근이 쉽지 않다.

최근에 많이 사용되는 비매개변수적 회귀방법(non-parametric regression estimator)은 kernel estimators, nearest neighbor methods, smoothing spline 등이 있으며, 이와 같은 회기모형은 기존의 회기분석 방법에 비하여 보다 쉽게 적정한 회귀모형을 제시해 준다. Smoothing spline의 경우 관측치에 가까운 추정값를 구할 수 있지만, 자료의 경향성을 나타내는 것이 부족한 것으로 알려져 있다. 이에 비하여 kernel regression의 경우는 자료의 분포 및 경향성을 근접하게 나타내는 장점이 있다고 할 수 있다. 따라서 기존의 회귀분석에서 해석하기 어려운 자연계의 이질적이고, 다중의 변수를 지니며, 시간과 공간적인 변수를 지니게 되는 자료들에 대한 유용한 해석방법이다.

Kernel regression estimator의 가장 기본적인 형태를 나타내면 다음과 같다.

$$f(x)=\frac{\sum_{i=1}^{n} K\left(\frac{(x-x_i)}{h}\right)y_i}{\sum_{j=1}^{n} K\left(\frac{(x-x_j)}{h}\right)} \qquad (5)$$

여기서, h 는 광역폭(bandwidth), n 는 자료개수, (x_i, y_i)는 주어진 관측자료이며, x 는 추정하고자하는 값, $K(\cdot)$ 는 kernel 함수이다.

식(5)를 연속적인 값으로 나타내기 위해 발전시키면 다음과 같다.

$$\hat{f}(X)=\frac{1}{n}\sum_{j=1}^{n}\int_{s_{j-1}}^{s_j} K\left(\frac{(X-s)}{h}ds\right)Y_j \qquad (6)$$

여기서, $s_j=\frac{(X_{j+1}+X_j)}{2}$ 이다.

식(5)는 식(6)과 같은 개념이며, 식(5)에 비하여 상대적으로 bias가 작아지는 경향이 있다. 특히 이 식은 공간적으로 비균일한 자료에 있어서 가장 적합한 추정값을 제시해 주는 것으로 알려져 있다.

Kernel 함수는 관측자료에 가중치를 부여하며, 가중치는 kernel 함수의 모양에 의해 결정되고, 그 폭은 bandwidth에 의해 결정된다. kernel regression의 경우 어떤 kernel함수를 결정하느냐에 따른 추정치의 변화는 크지 않다고 알려져 있다. 이에 비하여 추정치는 bandwidth에 매우 민감한데, 이 bandwidth의 결정은 회귀오차를 최소화하는 여러 가지 기법에 의하여 결정된다.

3.2 Kernel Function

Kernel regression에 있어서 kernel 함수는 관측값에 가중치로 작용하여 bandwidth내의 관측값으로부터 임의의 지점의 추정치를 찾아주는 역할을 한다. 관측값에 대한 가중치는 Kernel 함수의 모양에 의해 결정이

되며, 그 모양에 따라 여러 가지 종류가 있다. kernel 함수의 일반적인 특징은 다음과 같다.

$$\int_{-\infty}^{\infty} K(u)du = 1 \quad (7)$$

$$\int_{-\infty}^{\infty} uK(u)du = 0 \quad (8)$$

$$\int_{-\infty}^{\infty} u^2K(u)du = \alpha \quad (9)$$

(α 는 0이 아닌 상수)

3.3 Bandwidth

바람직한 회귀모형을 추정하기 위해서 자료가 나타내는 일련의 신호(signal)를 찾아내야 하는데 일반적으로 이 신호(signal)는 매끄럽다고 할 수 있다. 즉, 바람직한 회귀곡선을 "잔차의 제곱(오차)이 작으면서 매끄러운 곡선이다." 라고 할 수 있는 것이다(강근석, 김충락, 1999).

이 개념을 식으로 나타내면 다음과 같다.

① $\Sigma [y_i - f(x_i)]^2$: 적합도 (goodness of fit)

② $\int_a^b f''(x)^2 dx$

: 매끄러움 정도 (smoothness, roughness)

①과 ② 작게 하는 것이 올바른 회귀분석이라고 할 수 있다.

균형적인 조율을 위해서, $0 < q < 1$ 을 가정하고, 식을 만들면 다음 식 10과 같다.

$$(1-q)\sum_{i=1}^{n}[y_i - f(x_i)]^2 + q\int_a^b [f''(x)]^2 dx \quad (10)$$

여기서, $\frac{q}{(1-q)} = \lambda$ 라고 하면 위 식은 다음 식 (11)과 같이 변형될 수 있다.

$$S(\lambda) = \sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda\int_a^b [f''(x)]^2 dx \quad (11)$$

λ 는 잔차제곱과 매끄러운 정도의 균형을 나타내는 값으로 bandwidth를 의미한다. λ 가 아주 작으면, 들쭉날쭉해지고, λ 가 아주 커지면 직선에 가까워진다. 즉, Kernel regression 에 있어서 bandwidth h로 사용되는 λ 는 Kernel 함수의 bandwidth를 의미하며, 이 크기에 따라서 관측값에 대한 kernel 함수의 가중치가 결정이 되므로 smoothing parameter로 작용하여 전체적인 회귀모형에 영향을 미치게 된다.

4. Bandwidth 추정

적정한 Kernel regression의 수행을 위한 bandwidth λ 의 추정법으로 가장 많이 사용되는 것이 교차확인(Cross Validation)이다. 이는 회귀모형의 MSE (Mean Squared Error)를 최소화 할 수 있는 bandwidth를 설정하는 개념에서 출발하는 것이다.

4.1 Least Squares method

LS 방법은 자료와 추정된 값의 MSE를 최소로 하는 λ를 찾아내는 것으로 다음 식과 같이 정의된다.

관측값으로부터 추정된 f의 $x = x_j$에서의 추정치를 $f_\lambda(x_j)$ 라 하면 $LS(\lambda)$ 는 다음과 같다.

$$LS(\lambda) = \frac{1}{n}\sum_{j=1}^{n}(y_j - f_\lambda(x_j))^2 \quad (12)$$

여기서, y_i 는 i 번째 관측자료.

식 (12)는 식 (11)의 적합도(goodness of fit)부분과 같은 개념이다. 따라서, 자료의 매끄러운 정도를 나타내는데 무리가 따를 수 있다.

4.2 Least Squares Cross Validation

n개의 관측치중 j 번째를 제외한 n-1 개의 관측치로 추정된 f의 $x = x_j$에서의 추정치를 $f_{\lambda(j)}$ 라 하면 $CV(\lambda)$ 는 다음 식 (13)과 같이 정의된다.

$$CV(\lambda) = \frac{1}{n}\sum_{j=1}^{n}(y_j - f_{\lambda(j)}(x_j))^2 \quad (13)$$

소거 정리에 의하여 식 (13)을 변형하면, 다음의 식 (14)와 같다.

$$CV(\lambda) = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{y_j - f_\lambda(x_j)}{1-h_{jj}}\right)^2 \quad (14)$$

λ 가 어떤 값으로 주어졌을 때 $f(x)$ 의 추정치 $\hat{f}_\lambda(x)$ 라고 하면, 다음과 같은 식을 사용할 수 있다.

$$\hat{f}_\lambda(x) = H_\lambda y \quad (15)$$

여기서, H_λ ; hat matrix , h_{jj}는 H_λ 의 {j,j}

즉, Cross-Validation 방법은 추정치 n-1개의 관

측치만 써보고, 1개의 관측값으로 확인(validation)을 해보는 기법이다. $CV(\lambda)$ 값을 최소로 하는 λ 값이 최적의 bandwidth가 되며, Least squares cross-validation은 MSE의 bias를 적게 조절하여 주는 방법이다. 즉, 회귀모형의 MSE에 대한 기대값을 취할때 발생되는 항 $(2n^{-1}\sigma^2 tr[H_\lambda])$을 보정하여 bias를 조정하는 기법이다(Eubank, 1988).

4.3 General Cross Validation

GCV는 LSCV의 개념과 같은 방식이지만, h_{jj} 대신에 $tr(H_\lambda)/n$ 을 사용한 것이다(Wand 등, 1995). $GCV(\lambda)$ 는 다음 식 (16), (17)과 같이 정리 할 수 있다.

$$GCV(\lambda) = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{y_j - f_\lambda(x_j)}{1 - tr(H_\lambda)/n}\right)^2 \tag{16}$$

$$GCV(\lambda) = \frac{\frac{1}{n}\left(\sum_{j=1}^{n}(y_j - f_\lambda(x_j))^2\right)}{\left(\frac{1}{n}\,tr[I - H_\lambda]\right)^2} \tag{17}$$

$$= n \times \frac{\left(\sum_{j=1}^{n}(y_j - f_\lambda(x_j))^2\right)}{(EDF)^2} \tag{18}$$

여기서, $tr(H_\lambda)$ 는 $\sum_{j=1}^{n} h_{jj}$,

$\left(\sum_{j=1}^{n}(y_j - f_\lambda(x_j))^2\right)$; residual sum of squares,

$EDF = tr[I - H_\lambda]$; equivalent degrees of freedom.

LSCV와 개념적으로 동일선상에 있는 GCV의 경우, LSCV에 가중치 $\left(\frac{1}{n}tr[I - H_\lambda]\right)^{-2}$ 를 부여하여 예측모형에 대한 MSE의 bias를 줄여 주며, noise가 큰 자료의 경우 bandwidth의 결정에 있어서 LSCV보다 적합한 것으로 알려져 있다(Eubank, 1988).

4.4 bandwidth 추정 방법에 따른 회귀모형 결과분석

Kernel regression에 있어 bandwidth 결정방법에 의한 회귀결과를 비교하기 위하여, 임의의 함수를 설정하여 생성된 자료에 normal 분포로부터 무작위 추출한 잡음(noise)을 첨가하여 모형 적용을 위한 자료군을 생성하였다. 사용된 함수는 step function과 지수-sin함수로, 적용함수 및 첨가된 잡은(noise)은 표 1과 같다.

Kernel function은 동일하게 Quadratic kernel을 사용하였으며, 회귀모형을 구한 결과는 그림 1 ~ 그림 8과 같다. 그림에서 점선은 잡음을 첨가하기 이전의 원함수를 나타내며 굵은 실선은 kernel regression에 의한 회귀모형을 나타낸다. 회귀모형이 원함수에 근접하는 정도를 비교할 수 있다.

적용결과, bandwidth 결정방법에 차이에 있어서 GCV에 의해 bandwidth를 결정한 모형이 LS에 의한 결과에 비하여 bias의 경우LS에 비하여 큰 경우가 있으나, 전반적으로 작은 경향을 보이며, MSE의 경우는 GCV가 작고, 특히 그림 4, 8과 같이 잡음(noise)이 크면 MSE가 커져, LS는 bandwidth h값을 매우 작게 설정하게 되어 부적합한 회귀모형을 제시하게 되지만, GCV의 경우 상대적으로 MSE가 작고, 그림 3, 7과 같이 적절한 모형을 회귀해내고 있다.

5. Kernel 회귀에 의한 수위-유량곡선의 작성

일반적으로 수문학에서 사용되는 회귀분석은 자료처리를 위한 통계적 수단으로서 매개변수적 방법을 사용한다. 통계적인 기법으로서 kernel regression은 비교적 널리 사용되고 있으나, 수문학적인 분야에서 kernel

표 3. 적용함수

Data	True Function	Noise	Sample size
1	f(t) = 2 0.5<t≤1 f(t) = 1 0≤t≤0.5	N(0,0.04)	100
2	f(t) = 2 0.5<t≤1 f(t) = 1 0≤t≤0.5	N(0,1)	100
3	$f(t) = e^{-t}\sin(2\pi t)$ 0≤t≤1	N(0,0.01)	100
4	$f(t) = e^{-t}\sin(2\pi t)$ 0≤t≤1	N(0,0.25)	100

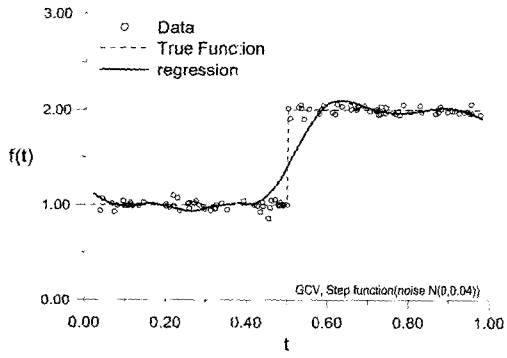

그림 1. GCV를 선택한 Kernel regression 결과(Step function, noise(N(0,0.04))

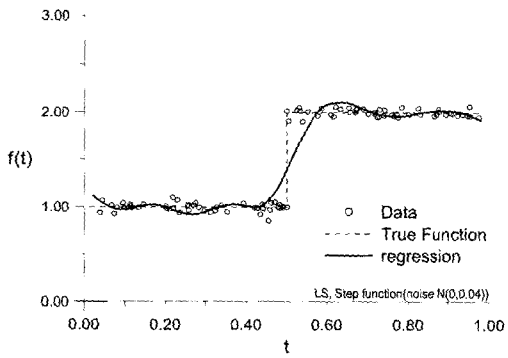

그림 2. LS를 선택한 Kernel regression 결과 (Step function, noise(N(0,0.04))

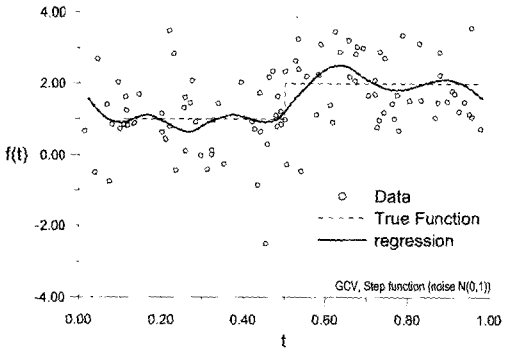

그림 3. GCV를 선택한 Kernel regression 결과(Step function, noise(N(0,1))

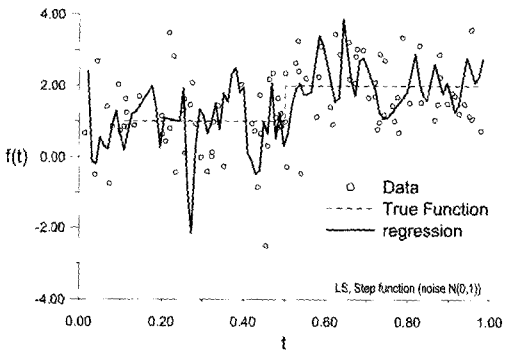

그림 4. LS를 선택한 Kernel regression 결과 (Step function, noise(N(0,1))

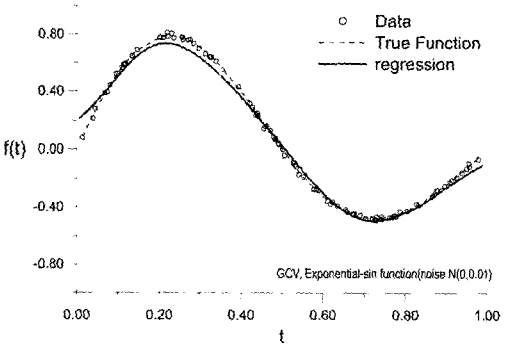

그림 5. GCV를 선택한 Kernel regression 결과 Exponential-sin function (noise(0,0.1))

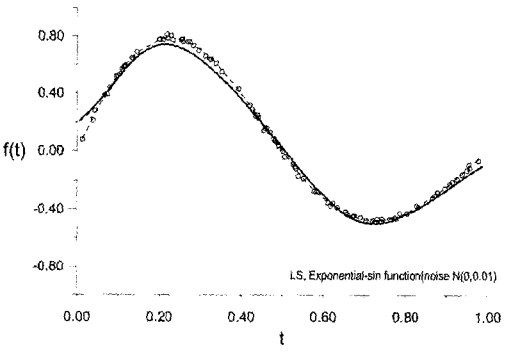

그림 6. LS를 선택한 Kernel regression 결과 Exponential-sin function(noise(0,0.1))

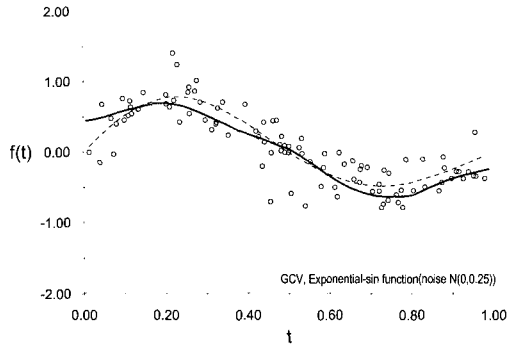

그림 7. GCV를 선택한 Kernel regression 결과 Exponential-sin function (noise(0,0.25))

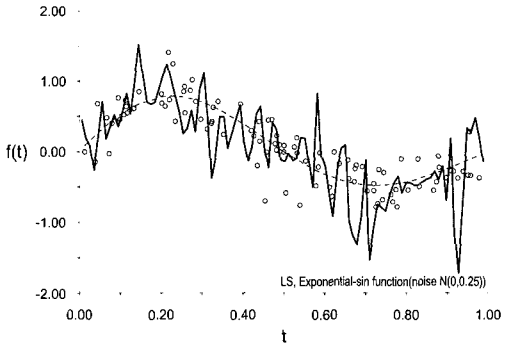

그림 8. LS를 선택한 Kernel regression 결과 Exponential-sin function (noise(0,0.25))

표 4. 공주 수위-유량 곡선식 분석표

분석 자료	수위구분	자료수	유도된 곡선식	결정계수 (r^2)	비고
98년	0.1≤H≤0.99	7	$Q=150.000(H+0.400)^{1.000}$	0.41	모든자료 사용
				0.92	H=0.1과 H=0.9를 제외할 시
	0.99≤H≤7.00	23	$Q=82.809\times(H+0.618)^{1.958}$	0.96	모든자료 사용

(자료:금강수계 유량측정 조사 보고서, 금강홍수통제소, 1998)

표 5. GCV와 LS에 따른 Kernel regression 결과

Function	Cross-vaild.	Bias	MSE
Step function N(0.0.04)	GCV	-0.0034	0.012
	LS	-0.0057	0.013
Step function N(0,1)	GCV	-0.002	0.054
	LS	-0.020	0.605
Exponential -Sin N(0,0.01)	GCV	-1.422	2.843
	LS	-1.416	2.847
Exponential -Sin N(0,0.25)	GCV	-1.468	3.077
	LS	-1.471	3.292
funtion No. 5 N(0,0.01)	GCV	-0.543	0.857
	LS	-0.544	0.863
funtion No. 6 N(0,0.25)	GCV	-0.573	0.893
	LS	-0.566	1.073

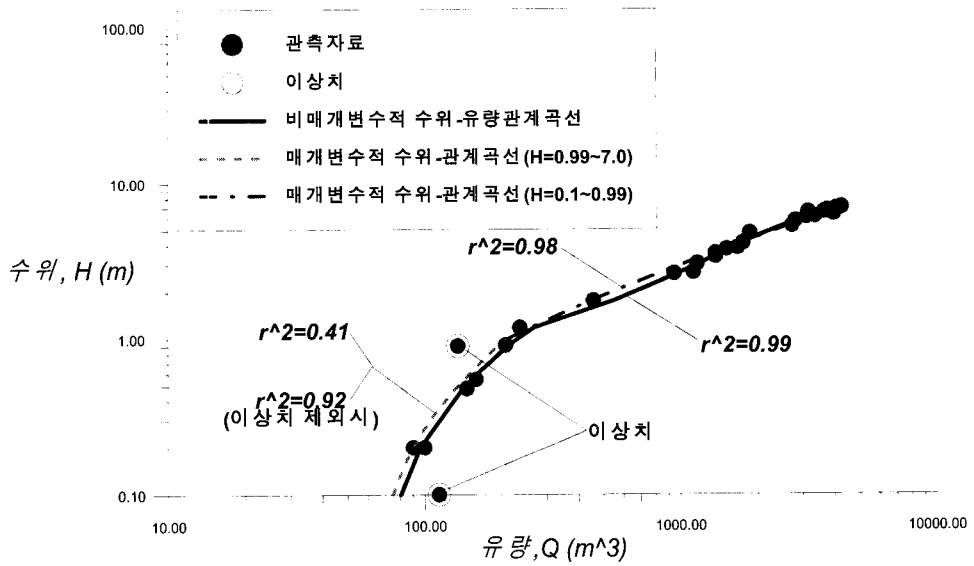

그림 9. 금강수계 공주지역 수위-유량 곡선

regression을 수위-유량 곡선의 작성에 적용한 것은 처음이다. 이에 따라 Kernel 회귀모형의 수문학적 적용성을 시험하기 위해 금강 홍수통제소에서 98년 측정한 금강수계 공주지역의 수위 및 유량관측 자료(건교부 금강홍수통제소, 1998)를 이용하여 kernel 회귀모형을 적용하고, 매개변수적인 방법으로 산정된 값과 비교하였다.

매개변수적 방법에 의해 산정된 회귀식은 표 2와 같으며, 산정된 회귀식에 의한 회귀곡선과 kernel 회귀분석에 의한 회귀곡선은 그림 9와 같다.

표 2와 같이 98년과 99년 공주지역의 수위-유량관계식은 하상 변화의 영향을 상대적으로 많이 받게 되는 저수위 부분($0.1m \le H \le 0.99m$)과 홍수파의 영향을 받게 되는 고수위 부분($0.99m \le H \le 7.00m$)을 나누어 두 개의 관계식으로 수위-유량관계식을 개발하였다. 고수위 경우는 결정계수 r^2=96%로 수위-유량관계가 양호하게 나타났으나 저수위경우는 자료군에서 다소 벗어난 H=0.1m와 H=0.9m인 측정값를 무시하고 회귀분석을 하였을 때 r^2=92%로 나타났다. 반면에 저수위자료 모두 사용시는 r^2=41%로 추정된 회귀식이 자료의 분산을 표현하는데 부족하였다.

대체적으로 기존의 매개변수적 수위-유량곡선과 비매개변수적 곡선이 비교적 비슷한 거동을 나타내 주고 있으나, 매개변수적 회귀분석에 의한 수위-유량 관계식은 수위 H=0.99m를 기준으로 자료를 두 구간으로 나누어 회귀식을 구하였으며 또한 자료군에서 벗어난 H=0.1m와 0.9m 측정값은 제외하였으며 비매개변수적 회귀분석은 일관된 하나의 회귀관계식이다. 그림 9와 같이 비매변수적 방법은 하나의 곡선으로 r^2= 99%를 보여주고 있다.

6. 결 론

비매개변수적 회귀모형 산정방법중의 하나인 kernel regression은 자료의 분포 및 경향성을 반영한 적절한 회귀모형을 제시해준다.

Kernel 회귀모형의 특성을 좌우하는 인자인 bandwidth 선정방법 중 LS의 경우 잡음(noise)이 큰 자료의 처리를 하게 될 경우, 오차의 bias가 커지게 되어 적정한 bandwidth를 선정하지 못하나, GCV의 경우는 Cross-Validation을 통해 오차의 bias를 줄임으로써 적합한 bandwidth를 선정하게 되는 것을 알 수 있다.

수문학에서 사용되는 매개변수적 회귀모형은 일반적으로 수위-유량 곡선이 나타내게 되는 문제는 저수위와 고수위에서의 수위-유량곡선이 서로 다른점, 즉 자료에 따라 복수의 회귀식을 산정해야 하는 문제나, 자료군에서 벗어난 자료의 선택, 회귀 구간을 나누는 방법에 따라 동일 자료라 하더라도 회귀식이 일관되지 않아 신뢰

할 수 있는 대표성을 지닌 회귀식의 유도가 쉽지 않았다는 것이다. 그러나 비매개변수적 방법을 사용하면 제외되는 자료 없이 관측된 자료를 사용하여 회귀모형을 산정하며 구간을 나누지 않고 곡선을 유도하므로 보다 일관되고 자료를 대표하는 관계식을 얻을 수 있는 장점을 지닌다.

Kernel 회귀모형의 특성을 이해하고, 적절한 band-width 선정방법인 GCV를 이용한 Kernel regression은 수문 기초자료의 효율적인 자료 처리 방법으로서 수자원분야에 광범위하게 이용될 수 있을 것이다.

감사의 글

이 논문은 2001년도 서울시립대학교 학술연구조성비에 의하여 연구되었습니다. 또한 본 논문의 세 번째 저자는 2001년도 원광대학교 교비지원을 받아 본 연구성과에 기여하였으며, 연구비 지원에 감사드립니다

참 고 문 헌

강근석, 김충락(1999). 회귀분석, 교우사, pp. 370~390.

문영일, 조성진, 김동권(2000). "수문학적 응용을 위한 비매개변수적 회귀모형 산정." 학술발표회 논문집(III), 대한토목학회, pp. 111~114.

조성진, 문영일, 권현한(1999). "비매개변수적 다항식을 이용한 수위-유량 관계곡선." **학술발표회 논문집(III)**, 대한토목학회, pp. 29~32.

조성진, 문영일, 황성환, 박대형, 권현한(2001). "General cross-validation과 Least squares method에 의한 비매개변수적 회귀모형의 특성." **학술발표회 논문집(III)**, 한국수자원학회, pp. 33~36.

건설교통부 금강홍수통제소(1998). **금강수계 유량측정조사 보고서.**

윤태훈(1997). 응용 수문학. 청문각.

Cleveland, W.S., and S.J. Devlin(1988). "Locally weighted regression : An approach to regression. analysis by local fitting." *J. Amer. Stat. Assn.*, Vol. 83 (403), pp. 596-610.

Cleveland, W.S., S.J. Devlin, and E. Grosse(1988). "Regression by local fitting." *J. Econometrics*, 37, pp. 87-114.

Eubank, R.L(1988). *Spline Smoothing and Nonparametric Regression*. DEKKER, New York.

Moon, Young-Il(1997). "A Nonparametric Nonlinear Time Series Forecasting Model Application To Selected Hydrologic Variables in Korea." *American Geophysical Union*, Fall Meeting Vol. 78 #46.

Scott, David W.(1992). *Multivariate Density Estimation*. JOHN WILEY & SONS, New York.

Wand, M.P. and Jones, M.C.(1995). *Kernel Smoothing*. CHAPMAN & HALL, New York.

(논문번호:03-03/접수:2003.01.15/심사완료:2003.11.08)