

# 분포형 CN 기반 토지피복별 유출가중치를 이용한 오염부하량 능형회귀모형 개발

## Development of Ridge Regression Model of Pollutant Load Using Runoff Weighted Value Based on Distributed Curve-Number

송철민\* · 김진수\*\*†

Song, Chul Min · Kim, Jin Soo

### Abstract

The purpose of this study was to develop a ridge regression (RR) model to estimate BOD and TP load using runoff weighted value. The concept of runoff weighted value, based on distributed curve-number (CN), was introduced to reflect the impact of land covers on runoff. The estimated runoff depths by distributed CN were closer to the observed values than those by area weighted mean CN. The RR is a technique used when the data suffers from multicollinearity. The RR model was developed for five flow duration intervals with the independent variables of daily runoff discharge of seven land covers and dependent variables of daily pollutant load. The RR model was applied to Heuk river watershed, a subwatershed of the Han river watershed. The variance inflation factors of the RR model decreased to the value less than 10. The RR model showed a good performance with Nash-Sutcliffe efficiency (NSE) of 0.73 and 0.87, and Pearson correlation coefficient of 0.88 and 0.93 for BOD and TP, respectively. The results suggest that the methods used in the study can be applied to estimate pollutant load of different land cover watersheds using limited data.

**Keywords:** Flow duration interval; multicollinearity; ridge regression; runoff weighted value

## 1. 서론

오염부하량 예측 모형은 크게 물리적 분포모형과 통계학적 모형으로 구분할 수 있다. 물리적 분포모형은 유역 내의 인간 활동이나 자연조건 변화 등의 공간적인 분포특성을 오염부하량 계산에 반영시킬 수 있다는 장점이 있으나(Lee et al., 2011), 소유역별 강우, 유량, 수질, 토지이용 등의 방대한 자료가 요구된다.

통계학적 모형은 물리적 분포모형에 비해 비교적 모형의 운영방법이 단순하며, 주로 최소제곱법 다중회귀모형(이하 다중회귀모형)을 이용한다. 그러나, 다중회귀모형은 독립변수들 간에 강한 상관관계를 나타내는 다중공선성(multicollinearity)이 존재할 경우에는 회귀계수의 영향력을 설명할 수 없는 문제점을 가지고 있다(Mekanik et al., 2013). 그 대안으로 Hoerl and Kennard (1970)는 능형회귀모형(ridge regression model)

을 제안하였다. 이 모형은 다중공선성을 개선하여 회귀식의 독립변수의 손실을 방지한다.

능형회귀모형을 수문 현상 및 수질 문제에 적용한 사례로서 Shih and Gervin (1980)은 미국 Okeechobee 호수질의 Landsat 조사에 능형회귀분석을 이용하였다. 이태리 Venice Lagoon 수질 문제에 있어서 플랑크톤의 생물지표 가능성에 대하여 능형회귀분석이 이용되었다(Bianchi et al., 2003). 캐나다 Moisie 강의 수온을 예측하는 데 능형회귀모형과 다중회귀모형이 적용되었다(Ahmadi-Nedushan et al., 2007). Wang and Xu (2008)는 리모트센싱을 이용하여 미국 Pontchartrain 호의 염도변화를 예측하기 위한 능형회귀모형을 개발하였다. 또한, Lima and Lall (2010)은 주기적 능형회귀모형을 이용하여 브라질 수력발전 적지를 선정하기 위한 월별 하천유량 예측을 실시하였다. 국내에서 Seong (1988)은 지역의 홍수량 빈도분석에서 유전알고리즘을 이용하여 능형회귀모형을 검증하였다. 하지만, 국내외적으로 능형회귀모형을 수문 및 수질 문제에 적용한 사례는 많지 않은 실정이다.

또한, 토지피복을 고려한 한 선행연구로서는 Isik et al. (2013)이 CN과 인공신경망을 이용하여 토지피복 변화에 따른 일별 유출량을 모의하였다. Moon et al. (2014)과 Deshmukh et al. (2013)은 CN을 토지피복별 면적가중평균 방법을 통해 연구유역의 대표 CN을 구하여 평균유효우량을 산정한 바 있다. Shin et al. (2001)은 논, 밭 및 산지에서 실측한 유량 및 수

\* Dept. of Policy for Watershed Management, The Policy Council for Paldang Watershed

\*\* Dept. of Agricultural and Rural Engineering, Chungbuk National University

† Corresponding author

Tel.: +82-43-261-2573 Fax: +82-43-271-5922

E-mail: [jskim@cbnu.ac.kr](mailto:jskim@cbnu.ac.kr)

Received: October 24, 2017

Revised: January 4, 2018

Accepted: January 5, 2018

질자료를 다중회귀모형에 반영하여 원단위를 산정하여 보고 하였다. 그러나 앞서 언급한 모형들은 유역 대표 CN을 산정하여 이용하거나, 일부 토지피복에 대한 실측자료를 이용하였기 때문에 모든 토지피복에 대한 영향을 모형에 적용할 수 없는 한계를 갖고 있다. 이에 각 토지피복별 유출특성을 설명할 수 있는 모형의 필요성이 대두되고 있다.

본 연구에서는 분포형 CN을 이용하여 토지피복별 유출특성을 반영할 수 있는 유출가중치와 일유출량을 산정하고, 이를 이용하여 오염부하량을 예측할 수 있는 능형회귀모형을 개발하고자 한다.

## II. 재료 및 방법

### 1. 연구유역

연구유역은 환경부가 수질오염총량관리제 시행을 위한 단위유역 중 한강수계 내 흑천 A (E 127° 31' 55.39", N 37° 27' 37.69") 단위유역이다. 흑천(지방 2급)은 경기도 양평군 청운면 신론리에 위치한 성지봉(EL. 791 m)에서 발원하여 남한강 분류에 합류한다. 연구유역은 토지피복의 변화가 거의 없으며 도시화가 되지 않은 전형적인 농촌지역이다(Fig. 1). 연구유역의 면적은 314.0 km<sup>2</sup>, 하천장은 42.9 km, 평균고도는

201.4 m이며, 평균경사는 27°이다. 연구유역의 토지이용형태를 보면 산림이 75.9 %로 가장 큰 비율을 차지하며, 논 8.3 %, 밭 6.9 %, 도심 3.5 %, 순이다(Table 1).

### 2. 연구자료

본 연구에서 활용한 자료는 수질, 유량, 강수, 토지피복도와 토양도 등 5개 자료이다. 자료 수집기간은 2007년 1월 1일부터 2014년 12월 31일까지이다. 2007년부터 2011년까지의 5년간 자료는 오염부하량 모형을 만드는데 사용되었으며, 2012년부터 2014년까지의 3년간 자료는 모형의 예측 결과를 검증하는데 활용되었다.

모형개발에는 모두 온라인상에서 제공하는 자료를 활용하였다. 수질, 유량, 토양도는 물환경정보시스템 (<http://water.nier.go.kr>), 토지피복도는 환경공간정보서비스 (<https://egis.me.go.kr>), 강수자료는 기상청 (<http://kma.go.kr>)에서 제공된다. 수질자료는 8일 간격의 BOD 및 TP 농도자료이다. 유량자료는 일단위 자료이며, 강수자료는 연구지역 내·외에 위치하고 있는 8개 관측지점의 일단위 자료이다.

토지피복도 및 토양도의 공간해상도는 30 m × 30 m, 축척은 1:5,000의 raster (grid) 포맷이다. 토지피복도는 7개 항목(수역, 도심, 나지, 초지, 산림, 논, 밭)으로 되어 있고, 토양도는 토성 등 59개 항목을 포함하고 있다.

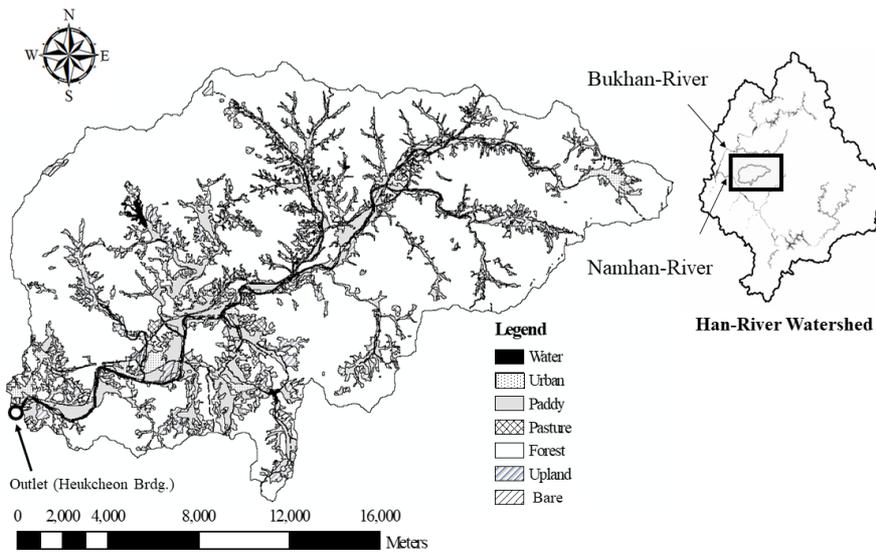


Fig. 1 Location of study area

Table 1 Land cover of study area

Land cover	Water	Urban	Bare	Pasture	Forest	Paddy	Upland	Total
Area (km <sup>2</sup> )	8,32	10,87	7,10	1,36	238,39	25,99	21,99	314,03
Percentage (%)	2,7	3,5	2,3	0,4	75,9	8,3	7,0	100,0

### 3. 연구방법

#### 가. 다중회귀모형과 능형회귀모형

먼저 유량지속곡선 (flow duration curve, FDC)을 5개 구간으로 구분하여 각 구간별 BOD 및 TP의 일오염부하량을 식 (1)과 같이 다중회귀모형으로 표현하였다.

$$L = \beta_0 + \beta_1 Q_1 + \beta_2 Q_2 + \beta_3 Q_3 + \dots + \beta_i Q_i \quad (i = 1, 2, 3, \dots, p) \quad (1)$$

여기서  $\beta_i$ 는 회귀계수,  $L$ 는 일오염부하량,  $Q_i$ 는 토지피복별 일유출량,  $p$ 는 토지피복종류의 개수를 나타내며, 회귀모형에 사용된 변수명은 Table 2와 같다.

여기서 회귀계수는 토지피복이 물리적으로 오염부하량에 영향을 미치지 때문에 0이 아닌 것으로 가정하였다. 토지피복별 일유출량은 토지피복 분포가 복잡하여 실측자료 수집이 어렵기 때문에 SCS (soil conservation service)의 CN을 기반으로 한 토지피복별 유출가중치 (runoff weighted value) 개념을 도입하여 산정하였다. 일별 강수량은 8개 관측지점의 자료를 ArcGIS 10.1 (ESRI, 2012)에서 역거리가중법 (inverse distance weight)을 통해 산정하였다.

다중회귀모형의 회귀계수는 식 (2)를 이용하는데, 다중공선성이 존재할 경우 행렬  $X^T X$ 는 특이행렬 (singular matrix)이 되기 때문에 회귀계수는 불안정한 결과를 초래한다. 다중공선성을 판단하는 방법으로 식 (3)과 같이 각 독립변수의 분산팽창지수 (variance inflation factor, VIF)을 이용하는데 이는 잔여분산 (residual variance)의 역수 값을 의미한다.

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad (2)$$

$$VIF_j = \frac{1}{1 - R_j^2} \quad (3)$$

여기서  $R_j^2$ 는  $j$ 번째 독립변수를 종속변수로 가정한 후 이를 나머지 독립변수들에 회귀시켜 얻은 결정계수를 의미한다. 일반적으로  $VIF > 10$ 이면 심각한 다중공선성이 존재하는 것으로 판단한다. 이에 다중공선성 문제를 해결하기 위해 Hoerl and Kennard (1970)은 능형회귀모형의 식 (4)를 제안하였다.

이때 회귀계수 식 (2)는 식 (5)로 수정되었다.

$$L = \hat{q}_0 + \hat{q}_1 Q_1 + \hat{q}_2 Q_2 + \hat{q}_3 Q_3 + \dots + \hat{q}_i Q_i \quad (i = 1, 2, 3, \dots, p) \quad (4)$$

$$\hat{q} = (X^T X + kI)^{-1} X^T y \quad (k \geq 0) \quad (5)$$

여기서,  $\hat{q}$ 는 능형회귀계수,  $k$ 는 능형매개변수,  $I$ 는 단위행렬이다. 능형회귀모형은 다중회귀모형에 비하여 편의 (biased) 되어 있으나 분산을 줄여 안정된 회귀계수를 갖는 특징을 갖는다. 능형회귀모형의 평균제곱오차 (MSE)가 다중회귀모형의 평균제곱오차보다 작게 되도록  $k$ 값을 선택하는 것이 중요하다. 본 연구에서의  $k$ 는 식 (6)과 같은 Horel et al. (1975)이 제시한 방법을 이용하였다.

$$k = \frac{p\hat{\sigma}^2}{\hat{\beta}\hat{\beta}} \quad (6)$$

여기서,  $p$ 는 토지피복종류의 개수,  $\hat{\sigma}^2$ 은 모형오차분산 (error variance of model)에 대한 불편추정량 (unbiased estimator)이다.

다중회귀분석은 통계 상용 프로그램인 SPSS ver. 19 (IBM Corp., 2010)을 이용하였고, 능형회귀분석은 통계프로그램인 R (The R Foundation, 2016)을 이용하였다.

#### 나. 유량지속곡선과 부하지속곡선

유량지속곡선은 어떤 유량을 초과하는 날의 연중 비율을 나타내는 곡선을 말한다. 유량지속곡선에서의 일유출초과율 (percent of day flow exceeded)은 식 (7)과 같이 나타낸다 (Vogel and Fenessey, 1994).

$$\text{일 유출초과율 (\%)} = \frac{\text{유량순위}}{\text{실측횟수}} \times 100 \quad (7)$$

여기서, 유량순위는 최고 유량으로부터의 내림차순의 순위를 나타낸다. 본 연구에서는 유량지속곡선을 5개 구간으로 구분하였는데, 일 유출초과율 <10 %, <40 %, <60 %, <90 %

Table 2 Summery of variable names

Constituent	Daily pollutant load		Daily runoff discharge						
	BOD	TP	Water	Urban	Bare	Pasture	Forest	Paddy	Upland
Variable name	$L_{BOD}$	$L_{TP}$	$Q_1$	$Q_2$	$Q_3$	$Q_4$	$Q_5$	$Q_6$	$Q_7$

및 ≤100 %의 구간을 각각 high flow, moist condition, mid-range flow, dry condition 및 low flow라고 한다 (Cleland, 2002).

부하지속곡선 (load duration curve)은 목표수질과 유량지속곡선의 곱으로 나타내며, 미국에서는 오염총량제도를 개발하는 데 부하지속곡선을 이용했다 (US EPA, 2007). 부하지속곡선은 오염물질의 허용부하능력을 나타내며, 유량에 따른 수질변화를 고려할 수 있는 장점이 있다. 실측부하량이 부하지속곡선을 상회하는 경우 허용부하를 초과했다고 판단한다.

다. 분포형 CN 기반 유출가중치 추정

유량지속곡선의 5개 구간별로 토지피복별 일유출가중치 및 일유출량을 산정하였다. SCS의 유출방정식은 식 (8)과 (9)와 같다.

$$S = \frac{25400}{CN} - 254 \tag{8}$$

$$q = \frac{(P - 0.2S)^2}{P + 0.8S} \quad (P > 0.2S) \tag{9}$$

$$q = 0 \quad (P \leq 0.2S)$$

여기서  $P$ 는 강우량,  $S$ 는 잠재저류량,  $q$ 는 잠재유출량이며,  $CN$ 은 국토해양부 설계홍수량 산정요령 (MLTM, 2012)에서 제시한  $CN$ 값.

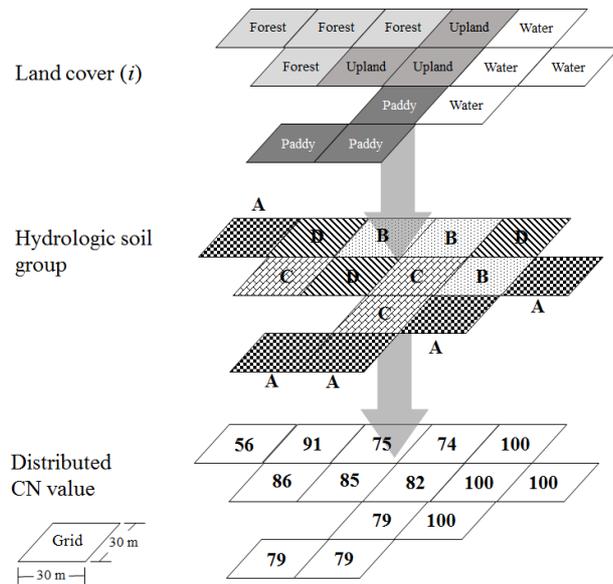


Fig. 2 Schematic for distributed CN value of grid associated with land cover and hydrologic soil group

본 연구에서는 ArcGIS 10.1 (ESRI, 2012)을 이용하여 토지피복을 7개로 분류하고, 토양도를 수문학적 토양군 4개(A, B, C 및 D)로 분류하여 30 m x 30 m 그리드 (grid)의 분포형  $CN$ 을 산정하였다. 분포형  $CN$ 의 토지피복별 잠재저류량  $S_i$ 와 잠재유출량  $q_i$ 은 식 (8)과 (9)를 각각 식 (10)과 (11)과 같이 수정하여 산정되었다.

$$S_i = \sum_{h=1}^n \left( \frac{25400}{CN_{II,h}} - 254 \right) \tag{10}$$

$$q_i = \frac{(P_i - 0.2S_i)^2}{P_i + 0.8S_i} \quad (P_i > 0.2S_i) \tag{11}$$

여기서  $h$ 는 토지피복  $i$ 에 해당하는  $CN$ 이 할당된 그리드의 개수를 나타낸다 (Fig. 2). 식 (10)에서 일별 선행토양수분조건 (antecedent soil moisture condition)이 다른 경우에는  $CN_{II}$ 을 수정하여 사용한다.

토지피복별 일유출가중치 ( $W_i$ )는 총 일유출량에 대한 각 토지피복별 일유출량의 비로 식 (12)와 같다.

$$W_i = \frac{q_i + S_i}{\sum_{i=1}^7 (q_i + S_i)} \tag{12}$$

식 (11)의 토지피복별 일유출량 ( $Q_i$ )은 식 (13)과 같이 토지피복별 일유출가중치와 유역말단의 일유출량 ( $Q_T$ )의 곱으로 산정되었다.

$$Q_i = W_i \times Q_T \tag{13}$$

라. 모형 평가

모형을 평가하기 위하여 식 (14)~(16)과 같이 Nash and Sutcliffe (1970)가 제안한 효율계수 (Nash-Sutcliffe efficiency, NSE), Pearson 상관계수 (Pearson correlation coefficient,  $r$ ) 및 RMSE (root mean square error)을 이용하였다.

$$NSE = 1 - \frac{\sum_{i=1}^n (Y_i^{obs} - Y_i^{\sim})^2}{\sum_{i=1}^n (Y_i^{obs} - Y^{mean})^2} \tag{14}$$

$$\gamma = \frac{\sum_{i=1}^n (C_i - \bar{C}_i)(y_i - \bar{y}_i)}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (C_i - \bar{C})^2}} \tag{15}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - C_i)^2} \quad (16)$$

여기서,  $Y^{mean}$  및  $\bar{C}$ 는 실측의 산술평균,  $Y_i^{obs}$  및  $C_i$ 는 실측값,  $Y_i^{sim}$  및  $y_i$ 는 예측값,

$\bar{y}_i$ 은 예측값의 평균을 의미한다. NSE와  $r$ 는 1에 가까울수록, RMSE는 작을수록 실측값과 예측값이 가깝게 되어 모형의 성능은 우수한 것으로 판단한다.

### III. 결과 및 고찰

#### 1. 유량지속곡선 및 부하지속곡선

연구기간 동안의 측정횟수는 1,826회로 연구유역 말단에

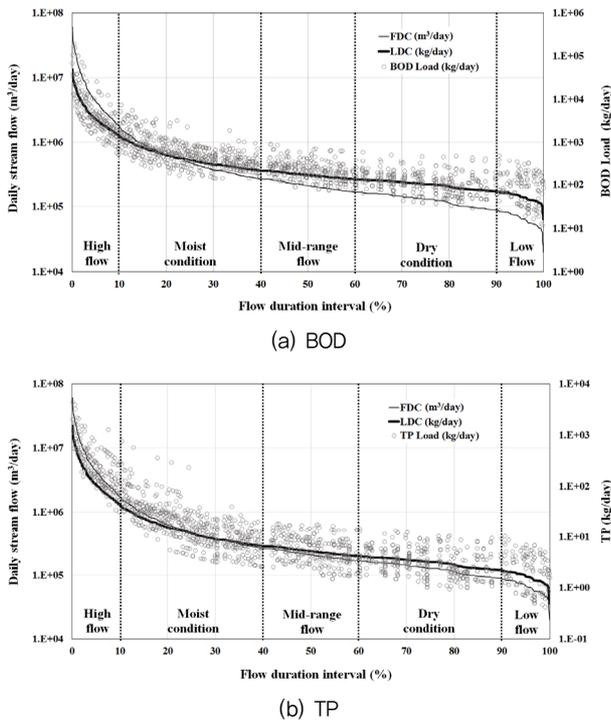


Fig. 3 Flow and load duration curves for study area

Table 3 Number and rate of data exceeding allowable load for flow duration interval

Constituent	Target concentration (mg/L).	Flow duration interval					Total
		High flow	Moist condition	Mid-range flow	Dry condition	Low flow	
BOD	0.8	93 (51.1) <sup>a</sup>	376 (64.9) <sup>a</sup>	177 (53.0) <sup>a</sup>	221 (40.3) <sup>a</sup>	89 (48.9) <sup>a</sup>	956 (52.4) <sup>a</sup>
TP	0.025	161 (88.5) <sup>a</sup>	339 (58.5) <sup>a</sup>	145 (43.4) <sup>a</sup>	187 (34.1) <sup>a</sup>	79 (43.4) <sup>a</sup>	911 (49.9) <sup>a</sup>

<sup>a</sup>The values in parenthesis indicate percentage.

서의 유량지속곡선, 부하지속곡선 및 실측 오염부하량을 나타내면 Fig. 3과 같다. 유량지속곡선은 0.02~60.8 백만 m<sup>3</sup>/day의 범위로 high flow 구간에서 유량변화가 크게 나타났다. 연구유역에 위치한 흑천 A의 목표수질은 BOD가 0.8 mg/L, TP가 0.025 mg/L이며, 목표부하지속곡선의 범위는 BOD가 15.9~48,679.1 kg/day, TP가 0.5~1,521.2 kg/day로 나타났다.

관측된 오염부하량의 허용부하량 초과 개수와 비율은 Table 3과 같다. 허용부하량을 넘는 평균 초과율은 BOD가 52 %, TP가 50 %로 비슷하게 나타났다. 또한, 각 유량지속구간에서의 최대 초과율은 BOD는 moist flow에서 65 %, TP는 high flow에서 89 %로 나타났으나, 최소 초과율은 BOD와 TP 모두 dry condition에서 40 %와 34 %로 나타났다. 이와 같이 BOD와 TP 부하량의 구간별 초과율이 고유량 구간에서 높고 저유량 구간에서 낮게 나타난 것은 연구유역이 비점원 오염에 의하여 크게 영향을 받는 것을 의미한다 (NEDP, 2003; US EPA, 2007).

#### 2. 분포형 CN

토지피복별, 수문학적 토양군별로 산정한 분포형 CN의 결과는 Table 4와 같다. 분포형 CN은 본래 토지피복을 7개 항목, 수문학적 토양군 (HSG)을 4개 항목, 등 전체 28개 항목으로 구분하였으나, 연구유역에서 수문학적 토양군 D는 존재하지 않아 총 21개 항목으로 나타났다. 수문학적 토양군 중 A가 32.8 %, B가 62.2 %, C가 5.0 %로 B가 가장 높은 비율을 나타냈다. 21개 항목 중 산림의 수문학적 토양군 B가 49.5 %로 가장 높게 나타났고, 초지의 수문학적 토양군 C가 0.1 %로 가장 낮게 나타났다.

본 연구지역에서 면적가중평균 CN 및 분포형 CN을 이용한 유출고와 실측 유출고의 관계는 Fig. 4와 같다. 일반적으로 효율계수가 0.65 이상인 경우 모형은 우수한 성능을 갖는 것으로 판단한다 (Moriassi et al., 2007). 분포형 CN과 면적가중평균 CN의 효율계수 (NSE)는 각각 0.80과 0.61로, 분포형 CN에 의한 예측값이 실측값에 더 가깝게 접근하는 것으로 나타났다.

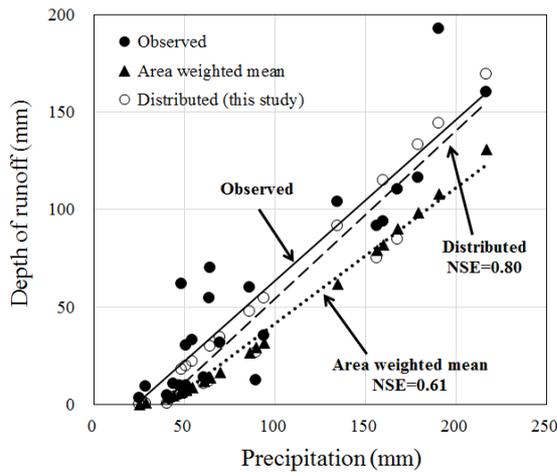
**Table 4** Distributed curve number and number of grids for land cover and hydrologic soil group

Land cover		Water	Urban	Bare	Pasture	Forest	Paddy	Upland	Percentage (%)
HSG-A (7.6-11.4) <sup>a</sup>	CN <sup>b</sup>	100	89	77	30	56	79	63	32.8
	Number of grids	3,403 (1.0) <sup>c</sup>	3,938 (1.1) <sup>c</sup>	2,901 (0.8) <sup>c</sup>	535 (0.2) <sup>c</sup>	83,761 (24.0) <sup>c</sup>	10,639 (3.0) <sup>c</sup>	9,369 (2.7) <sup>c</sup>	
HSG-B (3.8-7.6) <sup>a</sup>	CN <sup>b</sup>	100	92	86	58	75	79	74	62.2
	Number of grids	3,917 (1.1) <sup>c</sup>	7,015 (2.0) <sup>c</sup>	4,176 (1.2) <sup>c</sup>	792 (0.2) <sup>c</sup>	172,777 (49.5) <sup>c</sup>	15,326 (4.4) <sup>c</sup>	12,962 (3.7) <sup>c</sup>	
HSG-C (1.3-3.8) <sup>a</sup>	CN <sup>b</sup>	100	94	91	71	86	79	82	5.0
	Number of grids	1,927 (0.6) <sup>c</sup>	1,127 (0.3) <sup>c</sup>	807 (0.2) <sup>c</sup>	186 (0.1) <sup>c</sup>	8,339 (2.4) <sup>c</sup>	2,913 (0.8) <sup>c</sup>	2,107 (0.6) <sup>c</sup>	
HSG-D (0.0-1.3) <sup>a</sup>	CN <sup>b</sup>	100	95	94	78	91	79	85	0.0
	Number of grids	0 (0.0) <sup>c</sup>	0 (0.0) <sup>c</sup>	0 (0.0) <sup>c</sup>	0 (0.0) <sup>c</sup>	0 (0.0) <sup>c</sup>	0 (0.0) <sup>c</sup>	0 (0.0) <sup>c</sup>	
Total number of grids		9,247 (2.7) <sup>c</sup>	12,080 (3.5) <sup>c</sup>	7,884 (2.3) <sup>c</sup>	1,513 (0.4) <sup>c</sup>	264,877 (75.9) <sup>c</sup>	28,878 (8.3) <sup>c</sup>	24,438 (7.0) <sup>c</sup>	100.0

<sup>a</sup>The values in parenthesis indicate range of infiltration rate

<sup>b</sup>Condition of CN is AMCII (P)0.2S)

<sup>c</sup>The values in parenthesis indicate percentage



**Fig. 4** Comparison of area weighted mean CN and distributed CN for depth of runoff

### 3. 유량지속구간과 토지피복에 따른 유출가중치

Table 5는 유량지속구간과 토지피복에 따른 유출가중치의 평균값과 범위, S-A의 값에 대한 순위(rank)를 나타냈다. 여기서 유출가중치는 유역 전체 유출량에 대한 토지피복별 유출량의 비를 나타내며, S-A 값은 토지피복별 잠재저류량과 면적의 곱을 의미한다.

high flow 구간의 토지피복별 유출가중치의 순위는 산림, 밭, 논, 도심, 수역, 초지, 나지 순으로 나타났으며, moist condition

이하의 4개 구간의 순위는 산림, 밭, 논, 초지, 나지, 도심, 수역 순으로 모두 같게 나타났다. 또한, moist condition 이하의 4개 구간에서는 토지피복별 유출가중치 순위가 S-A의 순위와 같게 나타났다. 이는 이 구간에서의 유출량이 잠재저류량과 토지피복별 면적에 크게 영향받기 때문으로 생각된다.

### 4. 능형회귀모형

자료의 정규성(normality)을 확보하기 위해 토지피복별 일유출량( $Q_i$ ) 및 BOD와 TP 일부하량( $L_{BOD}$ ,  $L_{TP}$ )을 자연로그(natural logarithm) 값으로 변환하였으며, 이들의 관계를 다중회귀분석한 결과는 Table 6과 같다. 각 유량지속구간별 BOD 및 TP 부하량 예측을 위한 회귀모형은 high flow와 moisture condition에서는 모든 독립변수를 사용하면서도 유의성( $p < 0.001$ )이 인정되었으나, 다른 구간에서는 특정의 독립변수를 제거한 후에 비로소 유의성이 인정되었다.

또한, 다중회귀식에서의 일유출량의 분산팽창지수는 Table 7과 같다. 대부분 10 이상의 값을 보여 심각한 다중공선성이 존재하는 것으로 나타났다.

이에 식 (4)를 이용하여 능형회귀분석을 실시하였으며 이에 의한 능형회귀모형의 결과는 Table 8과 같다. 또한 능형회귀에서의 일유출량의 분산팽창지수는 Table 9와 같다. 모두 10이하의 값을 보여 다중공선성의 문제가 개선되는 것으로 나타났다.

Table 8의 유량지속구간별 능형회귀모형을 이용하여 2012

**Table 5** Runoff weighted values for land cover and flow duration interval

Flow duration interval		Water	Urban	Bare	Pasture	Forest	Paddy	Upland	Total
High flow	Weighted value	0,026 (0,000 -0,136) <sup>a</sup>	0,027 (0,000 -0,102) <sup>a</sup>	0,016 (0,002 -0,043) <sup>a</sup>	0,024 (0,002 -0,043) <sup>a</sup>	0,791 (0,567 -0,879) <sup>a</sup>	0,055 (0,005 -0,118) <sup>a</sup>	0,061 (0,025 -0,071) <sup>a</sup>	1,000
	Rank	5	4	7	6	1	3	2	-
Moist condition	Weighted value	0,003 (0,000 -0,130) <sup>a</sup>	0,008 (0,002 -0,104) <sup>a</sup>	0,010 (0,003 -0,041) <sup>a</sup>	0,016 (0,006 -0,057) <sup>a</sup>	0,846 (0,572 -0,878) <sup>a</sup>	0,048 (0,019 -0,106) <sup>a</sup>	0,069 (0,033 -0,071) <sup>a</sup>	1,000
	Rank	7	6	5	4	1	3	2	-
Mid range flow	Weighted value	0,000 (0,000 -0,018) <sup>a</sup>	0,007 (0,005 -0,015) <sup>a</sup>	0,011 (0,005 -0,011) <sup>a</sup>	0,014 (0,013 -0,033) <sup>a</sup>	0,847 (0,834 -0,860) <sup>a</sup>	0,050 (0,019 -0,052) <sup>a</sup>	0,071 (0,057 -0,071) <sup>a</sup>	1,000
	Rank	7	6	5	4	1	3	2	-
Dry condition	Weighted value	0,000 (0,000 -0,007) <sup>a</sup>	0,008 (0,005 -0,009) <sup>a</sup>	0,011 (0,008 -0,011) <sup>a</sup>	0,013 (0,013 -0,026) <sup>a</sup>	0,846 (0,845 -0,859) <sup>a</sup>	0,051 (0,036 -0,052) <sup>a</sup>	0,071 (0,063 -0,071) <sup>a</sup>	1,000
	Rank	7	6	5	4	1	3	2	-
Low flow	Weighted value	0,000 (0,000 -0,007) <sup>a</sup>	0,008 (0,006 -0,011) <sup>a</sup>	0,011 (0,009 -0,011) <sup>a</sup>	0,013 (0,013 -0,017) <sup>a</sup>	0,846 (0,839 -0,855) <sup>a</sup>	0,051 (0,044 -0,052) <sup>a</sup>	0,071 (0,068 -0,071) <sup>a</sup>	1,000
	Rank	7	6	5	4	1	3	2	-
S-A <sup>b</sup>	Value	0	296,838	413,045	482,047	31,689,563	1,949,813	2,671,866	37,503,172
	Proportion	0,000	0,008	0,011	0,013	0,845	0,052	0,071	1,000
	Rank	7	6	5	4	1	3	2	-

<sup>a</sup>The value in parenthesis indicates range of runoff weighted values

<sup>b</sup>This indicates the product of potential maximum soil moisture retention and land cover area

**Table 6** Multiple regression analysis of pollutants for flow duration interval

Constituent	Flow duration interval	Removal of variables	R <sup>2</sup>	R <sup>2</sup> <sub>adj</sub>	F-value	p
BOD	High flow	None	0,619	0,604	40,426	<0,001
	Moist condition	None	0,545	0,539	97,710	<0,001
	Mid-range flow	Q <sub>5</sub>	0,143	0,128	9,123	<0,001
	Dry condition	Q <sub>3</sub> , Q <sub>5</sub>	0,214	0,206	29,489	<0,001
	Low flow	Q <sub>3</sub> , Q <sub>5</sub> , Q <sub>6</sub>	0,156	0,137	8,156	<0,001
TP	High flow	None	0,621	0,606	40,750	<0,001
	Moist condition	None	0,629	0,625	138,555	<0,001
	Mid-range flow	Q <sub>5</sub>	0,132	0,116	8,317	<0,001
	Dry condition	Q <sub>3</sub> , Q <sub>5</sub>	0,185	0,177	24,582	<0,001
	Low flow	Q <sub>3</sub> , Q <sub>5</sub> , Q <sub>6</sub>	0,149	0,130	7,763	<0,001

**Table 7** Variance inflation factor for multiple regression

Flow duration interval	Independent variable						
	Q <sub>1</sub>	Q <sub>2</sub>	Q <sub>3</sub>	Q <sub>4</sub>	Q <sub>5</sub>	Q <sub>6</sub>	Q <sub>7</sub>
High flow	3,1	60,4	352,8	7,5	202,8	190,0	208,2
Moist condition	1,2	98,2	1352,2	89,5	2154,8	1072,1	1775,3
Mid-range flow	1,1	305,3	5639,9	20,4	-	3419,	92,9
Dry condition	1,0	97,9	-	599,7	-	4641,2	6814,9
Low flow	1,0	32,3	-	28,3	-	-	93,0

Table 8 Ridge regression model for flow duration interval

Constituent	Flow duration interval	<i>k</i>	Ridge regression model
BOD	High flow	$2.72 \times 10^{-1}$	$L_{BOD} = -4.3211 + 0.0423Q_1 - 0.4203Q_2 + 1.0229Q_3 + 0.0434Q_4 + 0.4537Q_5 - 0.4361Q_6 + 0.2964Q_7$
	Moist condition	$2.23 \times 10^{-1}$	$L_{BOD} = 0.9619 + 0.0003Q_1 - 0.3780Q_2 + 1.2609Q_3 + 0.7575Q_4 - 0.5043Q_5 - 0.6662Q_6 + 0.3892Q_7$
	Mid-range flow	$4.36 \times 10^{-2}$	$L_{BOD} = 0.1107 + 0.0003Q_1 - 1.3807Q_2 + 8.3565Q_3 + 1.1760Q_4 + 2.0450Q_5 - 4.2177Q_6 - 4.5900Q_7$
	Dry condition	$5.15 \times 10^{-4}$	$L_{BOD} = -169.1931 + 0.004Q_1 - 6.3428Q_2 + 28.2111Q_3 - 12.8219Q_4 + 69.1067Q_5 - 32.8236Q_6 - 44.3062Q_7$
	Low flow	$3.62 \times 10^{-8}$	$L_{BOD} = -30562.0100 + 0.0022Q_1 + 135.5532Q_2 - 3165.9680Q_3 - 525.8086Q_4 + 10446.8800Q_5 + 5670.4460Q_6 - 12560.8500Q_7$
TP	High flow	$0.61 \times 10$	$L_{TP} = -6.7924 + 0.0083Q_1 - 0.0318Q_2 + 0.0650Q_3 + 0.1290Q_4 + 0.2895Q_5 + 0.1762Q_6 + 0.3041Q_7$
	Moist condition	$3.14 \times 10^{-1}$	$L_{TP} = -10.5314 + 0.0015Q_1 - 0.2541Q_2 + 1.3444Q_3 + 0.4912Q_4 + 0.3625Q_5 - 0.9928Q_6 + 0.4582Q_7$
	Mid-range flow	$1.07 \times 10^{-2}$	$L_{TP} = -51.2054 + 0.0008Q_1 - 3.1168Q_2 + 13.6091Q_3 - 0.4675Q_4 + 21.9718Q_5 - 5.2334Q_6 - 25.2415Q_7$
	Dry condition	$1.16 \times 10^{-3}$	$L_{TP} = 1.8560 + 0.0008Q_1 - 12.6775Q_2 + 66.9071Q_3 - 6.0629Q_4 + 22.5659Q_5 - 65.9340Q_6 - 3.7743Q_7$
	Low flow	$5.17 \times 10^{-8}$	$L_{TP} = -22000.0400 + 0.0015Q_1 + 96.7476Q_2 - 2267.0810Q_3 - 380.5130Q_4 + 7522.6320Q_5 + 4062.5020Q_6 - 9033.8320Q_7$

Table 9 Variance inflation factor for ridge regression

Constituent	Flow duration interval	Independent variable						
		$Q_1$	$Q_2$	$Q_3$	$Q_4$	$Q_5$	$Q_6$	$Q_7$
BOD	High flow	3.0	5.0	2.7	6.6	1.6	1.4	1.6
	Moist condition	1.2	8.3	1.1	7.5	1.7	8.5	1.3
	Mid-range flow	1.4	9.5	1.3	2.2	4.0	8.8	4.1
	Dry condition	1.0	6.1	2.9	1.5	3.7	3.7	2.4
	Low flow	1.0	1.2	7.0	1.0	5.8	4.4	1.3
TP	High flow	3.1	5.6	2.8	6.8	1.7	1.5	1.7
	Moist condition	1.2	8.7	1.1	7.9	1.8	9.1	1.4
	Mid-range flow	1.0	1.3	2.0	2.2	5.2	1.3	5.5
	Dry condition	1.0	5.8	2.5	1.4	3.3	3.2	2.2
	Low flow	1.0	4.1	1.7	2.4	1.8	1.5	4.6

년부터 2014년까지의 오염부하량의 예측값과 실측값을 비교 하면 Fig. 5와 같다. 예측값이 실측값의 경향을 전반적으로 잘 따르며, 저유량보다는 고유량에서 실측값에 더 가깝게 접근 하는 것으로 나타났다.

또한, 3년 연평균 부하량의 실측값과 예측값, 효율계수, Pearson 상관계수 (*r*), RMSE의 결과는 Table 10과 같다. 연 평균 BOD 부하량은 실측값이 236천 kg, 예측값이 251천 kg 으로 약 15천 kg의 차이를 보였고, 연평균 TP 부하량은 실측

값이 18천 kg, 예측값이 12천 kg으로 약 6천 kg의 차이를 보 였다. 각 모형의 효율계수 (NSE)는 BOD가 0.73, TP가 0.87 로서 모형 성능이 우수하다는 판단기준인 0.65 이상의 값을 보였다. 각 모형의 Pearson 상관계수 (*r*)는 BOD 0.88, TP 0.93로 예측값과 실측값 간의 선형성이 매우 높게 나타났고, 각 모형의 RMSE는 BOD가 1.98 kg/day, TP가 1.82 kg/day 로 나타났다. 이상의 결과로부터 능형회귀모형의 성능은 우 수하며, BOD보다 TP에서 더 우수한 것으로 나타났다.

Table 10 Results of ridge regression model

Constituent	Annual mean load (kg/day)		NSE	r	RMSE (kg/day)
	Observed	Predicted			
BOD	235,535.7	251,514.6	0.73	0.88	1.98
TP	18,244.1	11,692.0	0.87	0.93	1.82

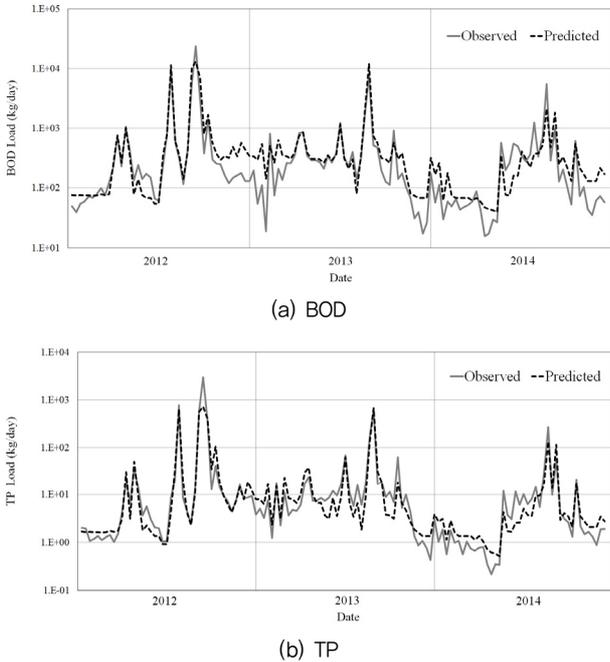


Fig. 5 Comparison between observed and predicted values of pollutant load

위성영상 (Landsat TM) 자료를 이용한 호수 수질 문제에서 자료의 정확도가 높은 수준(double level)에서의 능형회귀기법은 다중회귀분석에 비하여 전체평균제곱오차 (total mean square error)를 13~20% 줄이는 것으로 보고되었다 (Shih and Gervin, 1980). 캐나다 Moisie 강의 수온 예측에 있어서 능형회귀모형에서 다중공선성은 해결되었고, 다중회귀모형에 비하여 RMSE는 크고 NSE는 작으나 그 차이는 작은 것으로 보고되었다 (Ahmadi-Nedushan et al., 2007). Wang and Xu (2008)는 호수의 염도변화 예측에서 다중회귀모형에서는 위성영상자료인 독립변수의 VIF가 12.6~24.0으로 다중공선성 문제가 나타났는데 반하여, 능형회귀모형에서는 독립변수의 VIF가 <8.3로 다중공선성 문제가 개선되는 것으로 보고하였다. 브라질 하천에서의 월유량 예측에서 능형회귀모형은 2개의 선형회귀모형에 비하여 모든 예측 선행시간에서 결정계수( $r^2$ )이 가장 높고, 대부분의 예측 선행시간에서 Pearson 상관계수( $r$ )가 가장 높은 값을 보여, 모형 결과를 개선시키는 것

로 나타났다 (Lima and Lall, 2010).

이와 같이 기존의 연구 결과에서도 능형회귀모형은 다중회귀모형에 비해 다중공선성 문제를 개선하면서 오차가 작거나 오차가 큰 경우에도 큰 차이를 보이지 않는 것으로 나타났다.

#### IV. 결 론

본 연구에서는 토지피복에 따라 유출현상이 다르게 나타나는 현상을 고려하여 분포형 CN 기반으로 토지피복별 유출가중치라는 개념을 도입하고 토지피복별 일유출량으로부터 오염부하량을 예측하는 능형회귀모형을 개발하였다.

분포형 CN을 이용한 유출고는 면적가중평균 CN을 이용한 경우에 비하여 예측값이 실측값에 가깝게 접근하는 것으로 나타났다. 개발된 능형회귀모형은 다중회귀모형보다 분산팽창지수 (VIF)를 줄여 다중공선성 문제를 개선하는 것으로 나타났다. 또한, 능형회귀모형은 높은 효율계수 및 Pearson 상관계수를 보여 모형 성능이 우수한 것으로 나타났다. 이는 능형회귀모형이 다중회귀모형에 비해 다중공선성 문제를 개선하면서 작은 오차를 보이거나 오차가 큰 경우에도 큰 차이를 보이지 않는 기존의 연구 결과와 일치한다.

본 연구에서의 방법론은 제한된 자료로 토지피복 변화에 따른 오염부하량 예측을 가능하게 하여, 수질관리계획 수립 시 활용성이 높을 것으로 생각된다.

#### REFERENCES

- Ahmadi-Nedushan, B., A. St-Hilaire, T. B. M. J. Ouarda, L. Bilodeau, É. Robichaud, N. Thiémonge, and B. Bobée, 2007. Predicting river water temperatures using stochastic models: case study of the Moisie River, *Hydrological Processes*. 21: 21-34. doi: 10.1002/hyp.6353.
- Bianchi, F., F. Aciri, F. B. Aubry, A. Boldrin, E. Camatti, D. Cassin, and A. Comaschi, 2003. Can plankton communities be considered as bio-indicators of water quality in the Lagoon of Venice?, *Marine Pollution Bulletin*, 46(8): 964-971. doi: 10.1016/S0025-326X(03)00111-5.

3. Cleland, B. R., 2002. *TMDL development from the "Bottom Up"-Part II: using duration curves to connect the pieces.*, America's clean water foundation. August 15, 2002. National TMDL Science Policy 2002-WEF Specialty Conference, Phoenix, AZ, USA. doi: 10.2175/193864702785072687.
4. Deshmukh, D. S., U. C. Chaube, A. E. Hailu, D. A. Gudeta, and M. T. Kassa, 2013. Estimation and comparison of curve numbers based on dynamic land use land cover change, observed rainfall-runoff data and land slope, *J. of Hydrology*, 492: 89-101. doi: 10.1016/j.jhydrol.2013.04.001.
5. ESRI (Environmental Systems Research Institute), 2012. *ArcGIS*, Redlands, CA, USA.
6. Hoerl, A. E. and R. W. Kennard, 1970. Ridge regression: biased estimation for non orthogonal problems, *Technometrics*, 12: 69-82.
7. Hoerl, A. E., R. W. Kennard, and K. F. Baldwin, 1975. Ridge regression; some simulations, *Communications in Statistics*, 4:105-124. doi: 10.1080/03610927508827232.
8. IBM Corp. 2010. *IBM SPSS Statistics for Windows, Version 19.0*. Armonk, NY, USA.
9. Isik, S., L. Kalin, J. E. Schoonover, P. Srivastava, and B. G. Lockaby, 2013. Modeling effects of changing land use/cover on daily streamflow: an artificial neural network and curve number based hybrid approach, *J. of Hydrology*, 485: 103-112. doi: 10.1016/j.jhydrol.2012.08.032.
10. Lee, J. D., S. J. Lee, C. H. Hur, and S. D. Kim, 2011. Extraction of soil wetness information and application to distribution-type rainfall-runoff model utilizing satellite image data and GIS, *J. of Korean Society for Geospatial Information System*, 19(3): 23-32 (in Korean).
11. Lima, C. H. R. and U. Lall, 2010. Climate informed monthly streamflow forecasts for the Brazilian hydropower network using a periodic ridge model, *J. of Hydrology*, 380: 438-449. doi: 10.1016/j.jhydrol.2009.11.016.
12. Mekanik, F., M., A. Imteaz, S. Gato-Trinidad, and A. Elmahdi, 2013. Multiple regression and artificial neural network for long-term rainfall forecasting using large scale climate modes, *J. of Hydrology*, 503: 11-21. doi: 10.1016/j.jhydrol.2013.08.035.
13. MLTM, 2012. *Design Flood Estimation Techniques*, Ministry of Land Transport and Maritime Affairs (in Korean).
14. Moon, G. W., J. Y. Yoo, J. H. Ahn, and T. W. Kim, 2014. Comparative analysis of estimation methods for basin averaged effective rainfall using NRCS-CN method, *J. of Korean Society of Civil Engineers*, 34(2): 493-503 (in Korean). doi: 10.12652/Ksce.2014.34.2.0493.
15. Moriasi, D. N., J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith, 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, *American Society of Agricultural and Biological Engineers*, 50: 885-900. doi: 10.13031/2013.23153.
16. Nash, J. E. and J. V. Sutcliffe, 1970. River flow forecasting through conceptual models part I - A discussion of principles, *J. of Hydrology*, 10(3): 282-290. doi: 10.1016/0022-1694(70)90255-6.
17. NDEP (Nevada Division of Environmental Protection), 2003. *Load Duration Curve Methodology for Assessment and Total Maximum Daily Load Development*.
18. The R Foundation, 2016. *R Ver. 3.2.4*, Redlands, CA, USA.
19. Shih, S. F. and J. C. Gervin, 1980. Ridge regression techniques applied to LANDSAT investigation of water quality in Lake Okeechobee, *Journal of the American Water Resources Association*, 16(5): 790-796. doi: 10.1111/j.1752-1688.1980.tb02489.x.
20. Shin, E. S., J. Y. Choi, and D. H. Lee, 2001. Estimation of non-point source pollutant unit-loads in surface runoff considering land-use and basin characteristics, *J. KSWQ*, 17(2): 137-146 (in Korean).
21. Suh, H. S., K. S. Yang, N. Y. Kim, H. Y. Kim, and M. K. Kim, 2013. *SPSS (PAWS) Regression Analysis 3rd edition*. Hannarae Academy, Seoul, Republic of Korea (in Korean).
22. US EPA, 2007. *An approach for using load duration curves in the development of TMDLs*, EPA 841-B-07-006:1-68.
23. Vogel, R. M. and N. M. Fennessey, 1994. Flow-duration curves, 1: New interpretation and confidence intervals, *J. of Water Resource Planning and Management*, 120(4): 485-504. doi: 10.1061/(ASCE)0733-9496(1994)120:4(485).
24. Wang, F. and Y. J. Xu, 2008. Development and application of a remote sensing-based salinity prediction model for a large estuarine lake in the US Gulf of Mexico coast, *J. of Hydrology*, 360: 184-194. doi: 10.1016/j.jhydrol.2008.07.036.