

Research Article

Open Access

## 유전자 알고리즘과 회귀식을 이용한 오염부하량의 예측

박윤식<sup>1\*</sup>

<sup>1</sup>퍼듀대학교 농공학과

### Estimation of Pollutant Load Using Genetic-algorithm and Regression Model

Youn Shik Park<sup>1\*</sup> (<sup>1</sup>Department of Agricultural and Biological Engineering, Purdue University, West Lafayette in 47907, USA)

Received: 22 January 2014 / Revised: 5 February 2014 / Accepted: 7 February 2014

Copyright © 2014 The Korean Society of Environmental Agriculture

This is an Open-Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

#### Abstract

**BACKGROUND:** Water quality data are collected less frequently than flow data because of the cost to collect and analyze, while water quality data corresponding to flow data are required to compute pollutant loads or to calibrate other hydrology models. Regression models are applicable to interpolate water quality data corresponding to flow data.

**METHODS AND RESULTS:** A regression model was suggested which is capable to consider flow and time variance, and the regression model coefficients were calibrated using various measured water quality data with genetic-algorithm. Both LOADEST and the regression using genetic-algorithm were evaluated by 19 water quality data sets through calibration and validation. The regression model using genetic-algorithm displayed the similar model behaviors to LOADEST. The load estimates by both LOADEST and the regression model using genetic-algorithm indicated that use of a large proportion of water quality data does not necessarily lead to the load estimates with smaller error to measured load.

**CONCLUSION:** Regression models need to be calibrated and validated before they are used to interpolate pollutant

loads, as separating water quality data into two data sets for calibration and validation.

**Key words:** Genetic-Algorithm, LOADEST, Pollutant Load Estimation, Regression Model

#### 서론

강우시 하천이나 강으로 유입되는 비점원오염원은 부영양화나 수생태계의 파괴 등의 원인 중 하나이며, 이러한 현상은 자연적인 현상일 수도 있으나 농경지나 도시화와 같은 인위적인 요소도 적지 않은 영향을 끼친다(Carey *et al.*, 2011). 수질 개선 등의 문제를 해결하기 위해서는 이러한 문제를 해결하기 위해서는 비점원오염원을 정량화하는 것이 필요하며 이를 위해 수문 모형 등을 이용할 수 있다(Park *et al.*, 2007). 수문 모형을 검증하기 위해서는 측정된 수질 자료를 이용한 보정 및 검정 과정이 필요하며, 오염부하량 산정에 있어서도 장기적인 수질자료가 필요하다. 이러한 수질자료의 사용에 있어 유량 자료가 함께 요구되나, 수질 자료의 측정은 유량자료의 측정에 비해 상대적으로 많은 시간과 노력, 비용을 요구하기 때문에 유량 자료에 비해 측정 빈도가 낮다. 수질자료는 월 단위나 주 단위 등과 같이 일정한 간격으로 측정되거나, 강우사상에 기반을 두거나, 혹은 일정한 간격과 강우사상을 함께 고려하여 측정된다(King and Harmel, 2003). 측정 빈도가 낮은 수질자료를 보간 하거나 미계측 지점에 대한 오염부하량을 예측하기 위해서 회귀식이 사용되고 있으며,

\*교신저자(Corresponding author): Youn Shik Park  
Phone: +1-765-491-5268; Fax: +1-765-496-1115;  
E-mail: park397@purdue.edu

단순한 형태의 회귀식부터 로그함수로의 치환하거나 다양한 인자를 고려할 수 있는 복잡한 회귀식까지 다양하다(Kronvang and Bruhn, 1996; Robertson and Roerish, 1999; Horowitz, 2003; Robertson, 2003; Coynel *et al.*, 2004; Johnes, 2007; Henjum *et al.*, 2010).

Load Estimator (LOADEST; USGS Report, 2004)는 유량자료와 측정된 수질자료를 이용하여 회귀식을 보정하여, 수질자료가 측정되지 않은 지점에 대한 오염부하량을 산정할 수 있는 프로그램이다. Carey 등(2011)은 1992년에서 2006년 동안의 아질산염질소, 질산염, 총인의 실측자료를 통해 LOADEST를 보정 및 검증한 결과 실측치에 가까운 예측치를 모의하였다. 또한 LOADEST는 부유유사(Dornblaser and Striegl, 2009; Das *et al.*, 2011), 총질소(Oh and Sankarsaubramanial, 2011) 등의 수질자료에 대해 이용된 바 있다. 이처럼 LOADEST는 오염부하량을 예측하기 위해 사용되어지는 반면, 연간 오염부하량 예측에 대해서만 평가가 이루어졌거나, 보정과정에 대한 평가나, 검증과정이 이루어지지 않았다.

LOADEST는 11개의 회귀식을 가지고 있으며, 이 중 2개는 특정 기간에 대한 오염 부하량을 산정하기 위한 것이며, 그 외 9개의 회귀식이 주어진 유량과 수질 자료에 대해 모형의 보정 및 오염 부하량을 예측하기 위한 것이다. 이 9개의 회귀식은 유량만을 이용해 오염 부하량을 산정하는 회귀식부터 유량과 측정시간을 이용하는 회귀식까지 다양하다. LOADEST의 회귀식은 로그로 치환된 오염부하량을 산정하기 위한 것이다. LOADEST의 회귀식들은 다양한 형태의 유량과 측정시간을 고려하도록 되어 있으나, 유량에 대해서는 로그로 치환하는 데에 비해, 측정 시간에 대해서는 로그로 치환되지 않은 값을 사용하고 있다. 비점오염원은 강우시 많은 양이 유출이 되기 때문에 유량과 높은 상관관계를 가질 확률이 크지만, 비점오염원은 계절과 같은 시간에 따라 변동할 수 있기 때문에, 오염 부하량을 산정하는 데에 있어 유량과 측정시간은 모두 중요하다(Robertson, 2003).

따라서 본 연구에서는 유량과 측정시점을 대등하게 고려하여 오염부하량을 예측하는 회귀식과 LOADEST에 의한 예측치와 실측치를 비교하여, 회귀식을 이용한 일부하량 산정 방법을 분석하는 데에 있다.

### 재료 및 방법

#### 수질자료의 이용

본 연구에서 개발된 회귀식의 검정을 위해, United States Geological Survey (USGS, <http://water.usgs.gov/owq/data.html>)의 수질 자료가 사용되었다. 유역의 비점오염원 발생은 유역내 특성뿐만 아니라, 유역내 인위적인 요소 또한 영향을 줄 것으로 판단된다. 오염부하 산정을 위한 회귀식은 유역의 강우나 지형, 인위적인 요소 등과 같은 특성은 고려할 수 없다. 따라서 이러한 영향을 최소화하기 위하여, USGS의 수질 자료 중, 장기간에 걸쳐 측정되고, 비교적 유역 면적이

작은 지점에 대해 측정된 수질 자료가 사용되었다. 총 4개 지점(Table 1)에서 측정된 수질 자료가 사용되었으며, 각 지점은 44 km<sup>2</sup>에서 360 km<sup>2</sup>의 유역면적을 가졌으며, 약 10년에서 20년에 걸쳐 각 수질 자료에 대해 135개에서 1,437개의 수질 자료가 측정되었다. 각 지점에 대해서 수질 자료의 유무에 따라 4-5개의 질소와 인에 관련된 수질 자료가 사용되었다 (Table 2).

수문 모형이나 회귀식 등에 대해 적용성을 판단할 때에는 보정뿐만 아니라 검증 결과도 함께 분석이 되어야 하기 때문에 본 연구의 회귀식 역시 보정 및 검증 과정을 거쳤다. 보정과 검증 과정을 위해서는 실측치를 분할해야 하며, 일반적으로는 자료의 기간에 의해 분할하여 보정과 검증 과정을 거치나, 회귀식의 경우 보정 시 사용된 자료의 범위를 벗어나 예측하게 될 경우, 도시화 등과 같은 유역의 장기간에 걸친 변화 등에 의해 회귀식에 의한 예측치에 대한 불확실성이 커질 것으로 예상되었다. 따라서 본 연구에서는 실측 자료를 분할하는 데에 있어, 보정자료와 검증 자료의 기간이 같도록 할 필요가 있었으며, 이에 실측 자료의 측정된 날짜와 무관하게, 각 자료의 홀수 번째 자료를 보정자료로, 각 자료의 짝수 번째 자료를 검증 자료로 사용하였다. 따라서 각 실측 수질자료는 회귀식에 대한 보정과 검증과정에서 거의 같은 기간과 같은 개수의 수질 자료가 사용이 되었다.

**Table 1. United States Geological Survey stations for water quality data**

USGS station number	Site name	Drainage area (km <sup>2</sup> )	Period
03353637	Little Buck Creek near Indianapolis, Indiana	44	1994.1.3. ~2003.9.2.
05427948	Pheasant Branch at Middleton, Wisconsin	47	1974.7.17. ~2012.2.22.
03303280	Chio River at Cannelton Dam at Cannelton, Indiana	251	2000.2.23. ~2009.12.7.
04063700	Popple River near Fence, Wisconsin	360	1980.10.21. ~2010.6.22.

**Table 2. United States Geological Survey water quality data code and description**

Code	Parameter name
00600	Total nitrogen, water, unfiltered, milligrams per liter
00613	Nitrite, water, filtered, milligrams per liter as nitrogen
00631	Nitrate plus nitrite, water, filtered, milligrams per liter as nitrogen
00665	Phosphorus, water, unfiltered, milligrams per liter as phosphorus
00671	Orthophosphate, water, filtered, milligrams per liter as phosphorus

**일부하량 예측을 위한 회귀식의 이용**

LOADEST는 11개의 회귀식을 가지고 있으며 회귀식 번호를 지정할 경우 지정된 회귀식에 의해 회귀식 계수들을 보정하고, 이렇게 보정된 계수를 이용하여 미계측 지점에 대한 오염부하량을 산정한다. 하지만 회귀식 번호를 0으로 지정할 경우, 회귀식 번호 1-9번 중 주어진 수질자료를 통한 회귀식의 보정 결과가 Akaike Information Criterion에 의해, 예측치가 실측치에 가장 가까운 회귀식에 의해 미계측지점에 대한 오염부하량을 예측한다. 식 (1)에서 식 (9)에서 보이는 바와 같이, 식 (1)과 식 (2)의 경우 오염부하량을 유량에 의해서만 산정하는데, 이는 수질자료가 유량자료와 매우 큰 상관관계를 보일 경우 사용될 수 있으며, 식 (3)부터 식(9)의 경우는 유량과 함께 측정시간(계절적 변화)이 함께 수질자료가 상관관계가 있을 때 사용될 수 있다. 앞서 언급한 바와 같이, LOADEST의 회귀식들은 유량과 측정시간을 모두 고려할 수 있으나, 회귀식에 의해 로그로 치환된 오염부하량을 예측하는 것이다. 다시 말해, 회귀식의 계수 보정에서 회귀식에 의해 산정된 값들은 지수함수에 의해 다시 오염부하량으로 산정이 된다(USGS Report, 2004). 하지만, 로그로 치환된 오염부하량을 산정하는 데에 있어, 유량은 로그로 치환하여 고려하는 반면, 측정시간에 관한 항은 로그로 치환되지 않고 보정과정을 거치기 때문에 이 과정에서 유량과 측정시간은 대등한 위치에 있다고 보기 어렵다. 이러한 유량과 측정시간에 대한 고려가 실측치에 가까운 예측치를 위한 회귀식 계수 보정을 가능하게 할 수도 있지만, 수질자료는 수질자료의 종류와 측정지점에 따라 달라질 수 있기 때문에, 유량과 측정시간을 대등하게 고려하는 회귀식 역시 필요할 것으로 판단되었다.

일부하량(kg):

$$a_0 + a_1 \ln Q \tag{1}$$

$$a_0 + a_1 \ln Q + a_2 \ln Q^2 \tag{2}$$

$$a_0 + a_1 \ln Q + a_2 \text{dtime} \tag{3}$$

$$a_0 + a_1 \ln Q + a_2 \sin(2\pi \text{dtime}) + a_3 \cos(2\pi \text{dtime}) \tag{4}$$

$$a_0 + a_1 \ln Q + a_2 \ln Q^2 + a_3 \text{dtime} \tag{5}$$

$$a_0 + a_1 \ln Q + a_2 \ln Q^2 + a_3 \sin(2\pi \text{dtime}) + a_4 \cos(2\pi \text{dtime}) \tag{6}$$

$$a_0 + a_1 \ln Q + a_2 \sin(2\pi \text{dtime}) + a_3 \cos(2\pi \text{dtime}) + a_4 \text{dtime} \tag{7}$$

$$a_0 + a_1 \ln Q + a_2 \ln Q^2 + a_3 \sin(2\pi \text{dtime}) + a_4 \cos(2\pi \text{dtime}) + a_5 \text{dtime} \tag{8}$$

$$a_0 + a_1 \ln Q + a_2 \ln Q^2 + a_3 \sin(2\pi \text{dtime}) + a_4 \cos(2\pi \text{dtime}) + a_5 \text{dtime} + a_6 \text{dtime}^2 \tag{9}$$

$a_0 - a_6$  = 회귀식 계수

Q = 유량

dtime = 연중 시간을 0~1의 소수로 환산한 값

따라서 본 연구에서는, 오염부하량을 산정하는 회귀식의 기본 형식에 로그로 치환하지 않고, 유량과 시간을 고려할 수 있도록 하였다(식. 10). 본 연구에서 제안되고 사용된 회귀식

의 첫 번째 항은 유량에 관련된 오염부하를 고려하기 위함이며, 두 번째 및 세 번째 항은 시간에 따라 변동하는 오염부하를 고려하기 위함이다. 다시 말해, 본 연구의 회귀식은 유량과 시간을 동등하게 고려할 수 있도록 하였다.

일부하량(kg/d):

$$L = a_0 Q^{a_1} + [1 + \sin(a_2 \text{dtime} + a_3)] a_4 Q + [1 + \sin(a_5 \text{dtime} + a_6)] a_7 Q \tag{10}$$

L = 일부하량(kg/d)

Q = 유량

dtime = 연중 시간을 0~1의 소수로 환산한 값

$a_0 - a_7$  = 회귀식 계수

LOADEST는 모형내 회귀식을 보정하기 위해 최우추정법(Maximum Likelihood Estimation; MLE)과 개선된 최우추정법(Adjusted Maximum Likelihood Estimation; AMLE) 그리고 오차절대값최소화추정법(Least Absolute Deviation; LAD)에 의해 회귀식을 보정하며, 회귀식에 의한 예측치와 실측치의 차이가 표준정규분포를 따를 때에는 AMLE나 MLE에 의한 오염부하량을 사용하고 그렇지 않을 시에는 LAD에 의한 오염부하량을 사용하도록 제안하고 있다(USGS Report, 2004). 이러한 AMLE, MLE, 그리고 LAD 는 회귀식을 보정하는 데에 있어 최적화 기법 중 하나이며, 본 연구에서는 식 (10)의 회귀식 계수들을 보정하기 위해 유전자 알고리즘을 사용하였다. 유전자 알고리즘은 크게 선택, 교차, 변이의 과정을 가진다. 유전자 알고리즘은 복잡한 문제를 효과적으로 해결하기 위해서, 여러 분야에 걸쳐 사용되어 왔다(Togan and Daloglu, 2008). 따라서 본 연구에서는 유전자 알고리즘을 이용하여 회귀식의 계수들을 보정하였다.

**결과 및 고찰**

총 19개의 유량과 수질자료가 본 연구의 유전자 알고리즘에 의한 회귀식과 LOADEST 보정 및 검정에 이용되었다. 회귀식과 LOADEST의 보정 및 검정 결과를 판단하기 위해 Nash-Stucliff Efficiency (NSE; 식 11)과 결정계수(Coefficient of Determination;  $R^2$ ; 식 12)가 사용되었다. Santhi 등(2001)은 NSE와  $R^2$ 가 각각 0.5와 0.6일 때를 기준으로 모형에 의한 예측치를 판단한 바 있다. 따라서 본 연구의 회귀식와 회귀식의 계수들을 보정하기 위한 유전자 알고리즘에 의한 예측치의 정확도를 판단하기 위해서 NSE와  $R^2$ 가 함께 고려되었다. NSE와  $R^2$ 는 실측치( $O_i$ ), 실측치 평균( $\bar{O}$ ), 예측치( $P_i$ ), 예측치 평균( $\bar{P}$ )에 의해 계산된다.

**Nash-Stucliff Efficiency:**

$$NSE = 1 - \frac{\sum_{i=1}^n (O_i - P_i)^2}{\sum_{i=1}^n (O_i - \bar{O})^2} \tag{11}$$

Coefficient of Determination:

$$R^2 = \left( \frac{\sum_{i=1}^n (O_i - \bar{O})(P_i - \bar{P})}{\sqrt{\sum_{i=1}^n (O_i - \bar{O})^2} \sqrt{\sum_{i=1}^n (P_i - \bar{P})^2}} \right)^2 \quad (12)$$

LOADEST 보정 및 검정 결과

Table 3에서 보이는 바와 같이, LOADEST의 보정 과정에서 실측치와 예측치의 NSE는 -1.20 (USGS 지점 05427948, 수질자료 00671)에서 0.97 (USGS 지점 03303280, 수질자료 00631)으로 한 개의 수질자료를 제외한 예측치에서 0.5보다 큰 값을 보였으며(Fig. 1), R<sup>2</sup>는 0.53 (USGS 지점 05427948, 수질자료 00671)에서 0.97 (USGS 지점 03303280, 수질자료 00631)으로 한 개의 수질자료를 제외한 예측치에서 0.6보다 큰 값을 보였기 때문에(Fig. 2) LOADEST의 보정이 잘 된 것으로 판단되었다. LOADEST의 검정 과정에서 실측치와 예측치의 NSE는 -1.25 (USGS 지점 05427948, 수질자료 00671)에서 0.94 (USGS 지점 03353637, 수질자료 00665)로 19개의 수질자료에 대해 13개의 수질자료에 대해서 0.5보다 큰 값을 보였다(Fig. 1). R<sup>2</sup>는 0.36 (USGS 지점 04063700, 수질자료 00631)에서 0.98 (USGS 지점 03353637, 수질자료 00600)으로 19개의 수질자료에 대해 15개의 수질자료에 대해서 0.6보다 큰 값을 보였다(Fig. 2). 19번의 LOADEST의 보정 및 검정 결과 중에서 NSE가 0.5보다 크고 R<sup>2</sup>가 0.6보다 큰 예측치는 총 13개(68%)였으며, 따라서 LOADEST의 보정과 회귀식에 의한 오염부하량 예측은 비교적 잘 이루어진 것으로 판단되었다.

USGS의 4개 지점들은 각기 다른 유역 면적과 각기 다른 실측치의 개수를 가지고 있다. 같은 종류의 수질자료(00665)에 대해 USGS 지점 04063700과 USGS 지점 03303280에 대한 보정 및 검정 결과를 비교해볼 때, 역시 많은 양의 실측치를 통한 검보정이 반드시 실측치에 유사한 예측치를 얻을 수 있다고 판단하기는 힘든 것으로 보인다. 또한, USGS 지점 05427948의 수질자료 00665의 경우 719개의 수질자료(보정시)와 718개의 수질자료(검정시)를 이용하여 보정과정에서는 NSE와 R<sup>2</sup>가 모두 0.7보다 크고, 검정과정에서는 NSE와 R<sup>2</sup>가 모두 0.6보다 컸기 때문에 보정 및 검정 과정에서 예측치가 실측치에 매우 근사하였다. 하지만 더 적은 양의 수질 자료가 이용된 USGS 지점 03303280의 수질자료 00665의 경우 67개의 수질자료를 통한 보정 및 검정 과정에서 모두 NSE와 R<sup>2</sup>가 모두 0.7보다 컸기 때문에 이 역시 많은 양의 수질자료가 반드시 실측치와 가까운 예측치를 의미한다고 보기는 힘든 것으로 판단된다. Robertson (2003)는 수질 자료 측정 빈도에 따른 23 종류의 수질 자료와 회귀식을 이용해 오염부하량 평가를 한 바 있으며, 많은 양의 수질 자료가 반드시 실측치에 더 가까운 예측치를 의미하지 않는다고 제안한 바와 일치한다.

Table 3. Nash-Studiffe Efficiency and coefficient of determination of LOADEST in calibration and validation

Station number (Parameter code)	Calibration			Validation		
	NSE	R <sup>2</sup>	NUM*	NSE	R <sup>2</sup>	NUM*
03303280 (00600)	0.92	0.92	67	0.88	0.91	67
03303280 (00613)	0.54	0.55	67	0.45 <sup>N</sup>	0.48 <sup>R</sup>	66
03303280 (00631)	0.97	0.97	67	0.81	0.85	66
03303280 (00665)	0.72	0.75	67	0.76	0.86	67
03303280 (00671)	0.69	0.73	67	0.50	0.65	67
03353637 (00600)	0.89	0.95	67	0.97	0.98	66
03353637 (00613)	0.71	0.75	67	0.37 <sup>N</sup>	0.83	66
03353637 (00631)	0.73	0.78	67	0.61	0.92	66
03353637 (00665)	0.51	0.91	67	0.94	0.96	66
03353637 (00671)	0.97	0.97	67	0.60	0.87	66
04063700 (00600)	0.91	0.92	78	0.41 <sup>N</sup>	0.48 <sup>R</sup>	77
04063700 (00613)	0.94	0.94	69	0.85	0.89	69
04063700 (00631)	0.51	0.59	95	0.20 <sup>N</sup>	0.36 <sup>R</sup>	94
04063700 (00665)	0.71	0.72	139	0.45 <sup>N</sup>	0.48 <sup>R</sup>	138
04063700 (00671)	0.60	0.61	81	0.78	0.82	81
05427948 (00600)	0.74	0.74	101	0.83	0.83	100
05427948 (00631)	0.56	0.62	79	0.83	0.83	79
05427948 (00665)	0.71	0.83	719	0.64	0.68	718
05427948 (00671)	-1.20 <sup>N</sup>	0.53 <sup>R</sup>	138	-1.25 <sup>N</sup>	0.64	138

NSE: Nash-Studiffe efficiency

\*NUM: Number of water quality samples

R<sup>2</sup>: Coefficient of determination

N: NSE less than 0.5

R: R<sup>2</sup> less than 0.6

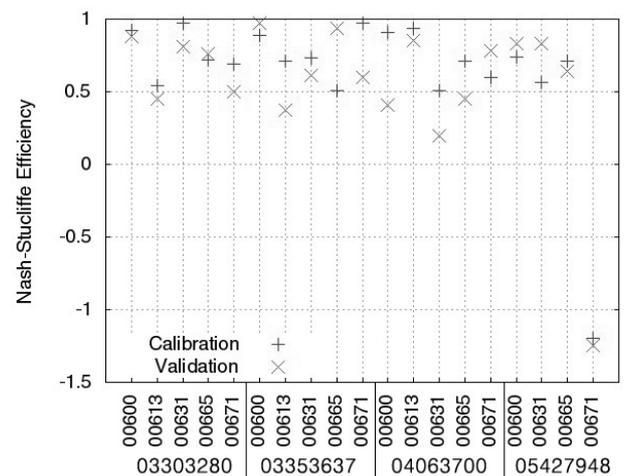


Fig. 1. Nash-Studiff Efficiency of LOADEST in calibration and validation.

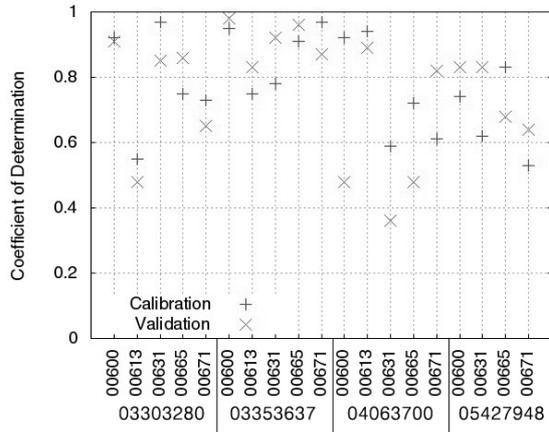


Fig. 2. Coefficient of determination of LOADEST in calibration and validation.

유전자 알고리즘에 의한 회귀식 보정 및 검정 결과

Table 4에서 보이는 바와 같이, 회귀식의 보정 과정에서 실측치와 예측치의 NSE는 0.53 (USGS 지점 03303280, 수질자료 00613)에서 0.96 (USGS 지점 03353637, 수질자료 00600)으로 모두 0.5보다 큰 값을 보였으며(Fig. 3), R<sup>2</sup>는 0.59 (USGS 지점 03303280, 수질자료 00613)에서 0.96 (USGS 지점 03353637, 수질자료 00600)으로 대부분 0.6보다 큰 값을 보였기 때문에(Fig. 4) 유전자 알고리즘에 의한 회귀식은 보정이 잘 된 것으로 판단되었다. 회귀식의 검정 과정에서 실측치와 예측치의 NSE는 -0.99 (USGS 지점 05427948, 수질자료 00631)에서 0.90 (USGS 지점 03353637, 수질자료 00600)으로 19개의 수질자료에 대해 LOADEST와 비슷한 12개의 수질자료에 대해서 0.5보다 큰 값을 보였다(Fig. 3). R<sup>2</sup>는 0.37 (USGS 지점 04063700, 수질자료 00665)에서 0.97 (USGS 지점 03353637, 수질자료 00600)으로 19개의 수질자료에 대해 LOADEST와 비슷한 14개의 수질자료에 대해서 0.6보다 큰 값을 보였다(Fig. 4). 19번의 유전자 알고리즘에 의한 회귀식의 보정 및 검정 결과 중에서 NSE가 0.5보다 크고 R<sup>2</sup>가 0.6보다 큰 예측치는 총 10개(53%)로, LOADEST 보다 3개 적은 수질 자료에 대해 예측치의 보정 및 검정 과정에서 실측치에 가까운 일별 오염부하량을 보였다.

회귀식을 이용한 일별 오염부하량 예측 역시, LOADEST와 비슷한 결과 및 결론을 도출할 수 있었다. 같은 종류의 수질자료(00671)에 대해 USGS 지점 05427948과 USGS 지점 03303280에 대한 보정 및 검정 결과를 비교해볼 때, 많은 양의 실측치를 통한 검보정이 반드시 실측치에 유사한 예측치를 얻을 수 있다고 판단하기는 힘든 것으로 보인다. 특히, USGS 지점 05427948의 수질자료 00665의 경우 719개의 수질자료(보정시)와 718개의 수질자료(검정시)를 이용하여 보정 과정에서는 NSE와 R<sup>2</sup>가 모두 0.8보다 크고, 검정과정에서는 NSE와 R<sup>2</sup>가 모두 0.7에 가까웠기 때문에 보정 및 검정 과정에서 예측치가 실측치에 매우 근사하였다. 하지만 USGS 지점 03353637의 수질자료 00600의 경우 67개의 수질자료(보정시)와 66개의 수질자료(검정시)를 이용하여 보정 및 검정

과정에서 모두 NSE와 R<sup>2</sup>가 모두 0.9보다 컸기 때문에 이 역시 많은 양의 수질자료가 반드시 실측치와 가까운 예측치를 의미한다고 보기는 힘든 것으로 판단된다.

Table 4. Nash-Stucliffe Efficiency and coefficient of determination of the regression model using genetic-algorithm in calibration and validation

Station number (Parameter code)	Calibration			Validation		
	NSE	R <sup>2</sup>	NUM*	NSE	R <sup>2</sup>	NUM*
03303280 (00600)	0.90	0.91	67	0.86	0.90	67
03303280 (00613)	0.53	0.59 <sup>R</sup>	67	0.26 <sup>N</sup>	0.43 <sup>R</sup>	66
03303280 (00631)	0.95	0.95	67	0.76	0.82	66
03303280 (00665)	0.65	0.74	67	0.65	0.84	67
03303280 (00671)	0.82	0.82	67	0.56	0.66	67
03353637 (00600)	0.96	0.96	67	0.90	0.97	66
03353637 (00613)	0.88	0.89	67	0.79	0.80	66
03353637 (00631)	0.78	0.79	67	0.29 <sup>N</sup>	0.88	66
03353637 (00665)	0.94	0.95	67	0.68	0.92	66
03353637 (00671)	0.94	0.95	67	0.49 <sup>N</sup>	0.88	66
04063700 (00600)	0.92	0.93	78	0.51	0.52 <sup>R</sup>	77
04063700 (00613)	0.91	0.91	69	0.69	0.80	69
04063700 (00631)	0.67	0.73	95	-0.72 <sup>N</sup>	0.46 <sup>R</sup>	94
04063700 (00665)	0.76	0.78	139	0.22 <sup>N</sup>	0.37 <sup>R</sup>	138
04063700 (00671)	0.67	0.69	81	0.33 <sup>N</sup>	0.78	81
05427948 (00600)	0.63	0.66	101	0.57	0.77	100
05427948 (00631)	0.76	0.78	79	-0.99 <sup>N</sup>	0.74	79
05427948 (00665)	0.84	0.86	719	0.67	0.68	718
05427948 (00671)	0.79	0.80	138	0.53	0.55 <sup>R</sup>	138

NSE: Nash-Stucliffe efficiency

\*NUM: Number of water quality samples

R<sup>2</sup>: Coefficient of determination

N: NSE less than 0.5

R: R<sup>2</sup> less than 0.6

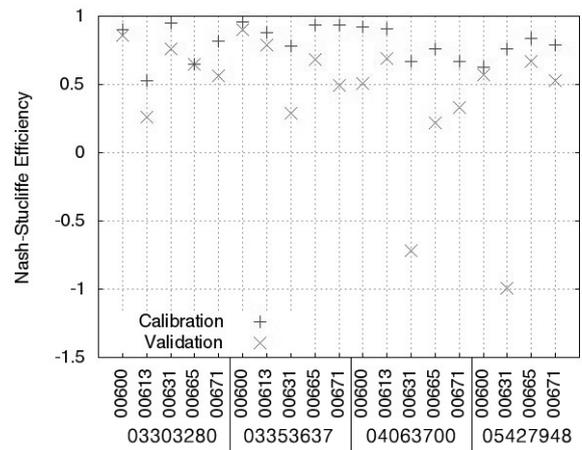


Fig. 3. Nash-Stucliff Efficiency of the regression model using genetic-algorithm in calibration and validation.

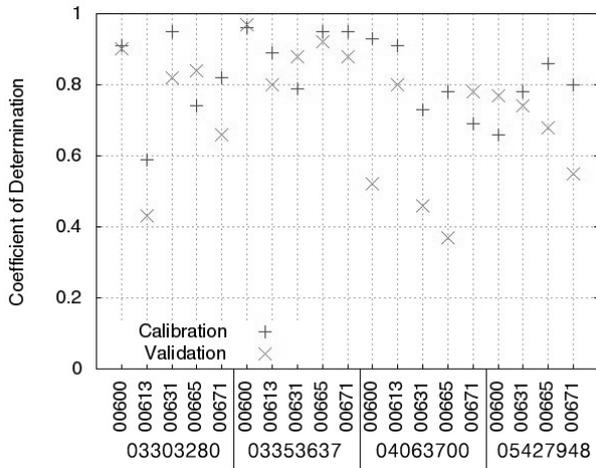


Fig. 4. Coefficient of Determination of the regression model using genetic-algorithm in calibration and validation.

## 결론

본 연구에서 사용된 수질자료는 장기간에 걸쳐 총 4개 지점에서 측정된 5종류였으며, LOADEST와 유전자 알고리즘을 이용한 회귀식의 보정 및 검정을 위해, 실측자료는 모두 두 개의 자료로 분할되어 사용되었다. 일반적으로 많은 양의 수질자료를 통해 회귀식을 보정해야 실측치에 가까운 예측치를 구할 수 있다고 여겨지는 반면, 본 연구에서 사용된 수질 자료에 대해 LOADEST와 유전자 알고리즘을 이용한 회귀식 모두 많은 양의 수질 자료가 반드시 실측치와 가까운 예측치를 보이지는 않았다. 또한, LOADEST와 유전자 알고리즘을 이용한 회귀식 모두, 보정과정에서 실측치와 작은 차이를 보였다고 해서 반드시 검정과정에서 실측치와 작은 차이를 보이지 않았다. 이는 유량 자료에 비해 측정 빈도가 낮은 수질 자료를 회귀식에 의해 일부하량이나 연부하량을 산정할 때, 회귀식의 보정과정에서 사용된 실측치와 예측치가 작은 오차를 보인다고 할지라도 미세측지점에 대해서는 오차가 클 수도 있으며 따라서 일부하량이나 연부하량 역시 실제 발생된 오염부하량과는 적지 않은 오차를 보일 수 있다는 것을 가리킨다. 따라서 다른 수문모형과 마찬가지로, 미세측지점에 대해 오염부하량 산정을 위한 회귀식 사용은 보정과정뿐만 아니라 검정과정을 통해 평가가 된 후 사용되어야 할 것으로 판단된다.

또한, LOADSET와 유전자 알고리즘을 이용하는 회귀식에 의한 보정 및 검정 결과를 볼 때, 두 회귀식에 대해 모두 실측치와 큰 오차를 보이는 수질자료가 있는 반면, 선택적으로 실측치와 예측치가 높은 상관관계를 보인 수질자료가 있었다. 그러므로 측정된 수질자료를 이용하여 미세측지점에 대해 보간할 때, 어느 하나의 모형을 일괄적으로 사용하기 보다는, 실측치를 분할하여 보정과 검정 과정을 통해 보다 실측치에 가까운 예측치를 보인 회귀식을 사용하는 것이 미세측지점에 대한 수질 자료 예측이 효과적일 것으로 보인다.

회귀식은 일반적으로 유량과 수질자료만을 요구하기 때문에,

수문 모형에 비해 적은 노력과 시간을 요구한다. 본 연구에 사용되었던 LOADEST와 유전자 알고리즘을 이용한 회귀식 역시 유량과 수질자료만을 이용하여 보정과 검정과정에서 이루어졌으며, 적지 않은 예측치가 실측치와 가까운 것으로 나타났다. 그러므로 회귀식의 사용은 다른 수문 모형의 사용보다 효율적인 것으로 보인다. 하지만 이와 동시에, 수문모형은 모형 자체의 여러 변수를 통해 실측치에 보다 가까운 예측치를 이끌어 낼 수 있는 기회가 주어지는 반면, 회귀식은 이에 대한 방법이 제한적이다. 이와 같은 점을 고려할 때, 회귀식에 의한 예측치의 사용에 있어 충분한 검정이 주어져야 하며, 이러한 보정 및 검정 과정이 충분히 이루어진다면 오염부하량을 산정하는 데에 있어 보다 효율적으로 이용될 수 있을 것으로 판단된다.

## References

- Carey, R.O., Migliaccio, K.W., Brown, M.T., 2011. Nutrient discharges to Biscayne Bay, Florida: trends, loads, and a pollutant index, *Sci. Total Environ.* 409, 530-539.
- Coyne, A., Schafer, J., Hurtrez, J., Dumas, J., Etcheber, H., Blanc, G., 2004. Sampling frequency and accuracy of SPM flux estimates in two contrasted drainage basins, *Sci. Total Environ.* 330, 233-247.
- Das, S.K., Ng, A.W.M., Perera, B.J.C., Assessment of nutrient and sediment loads in the Yarra river catchment, 19th International Congress on Modelling and Simulation, Perth, Australia, 12-16 December 2011, <http://mssanz.org.au/modsim>, pp. 3490-3496.
- Dornblaser, M.M., Striegl, R.G., 2009. Suspended sediment and carbonate transport in the Yukon river basin, Alaska: Fluxes and potential future responses to climate change, *Water Resour. Res.* 45, W06411, doi:10.1029/2008WR007546
- Henjum, M.B., Hozalski, R.M., Wennen, C.R., Novak, P.J., Arnold, W.A., 2010. A comparison of total maximum daily load (TMDL) calculations in urban streams using near real-time and periodic sampling data, *J. Environ. Monit.* 12, 234-241.
- Horowitz, A.J., 2003. An evaluation of sediment rating curves for estimating suspended sediment concentrations for subsequent flux calculations, *Hydrolog. Proc.* 17, 3387-3409.
- Johnes, P.J., 2007. Uncertainties in annual riverine phosphorus load estimation: Impact of load estimation methodology, sampling frequency, baseflow index and catchment population density, *J. Hydrol.* 332, 241-258.
- King, K.W., Harmel, R.D., 2003. Considerations in selecting a water quality sampling strategy, *Trans. ASAE.* 46, 63-73.

- Kronvang, B., Bruhn, A.J., 1996. Choice of sampling strategy and estimation method for calculating nitrogen and phosphorus transport in small lowland streams, *Hydrolog. Proc.* 10, 1483-1501.
- Oh, J., Sankarasubramanian, A., 2011. Interannual hydroclimatic variability and its influence on winter nutrients variability over the southeast United States, *Hydrol. Earth Sys. Sci. Discuss.* 8, 10935-10971.
- Park, Y., Kim, J., Park, J., Hong, J.H., Choi, D.H., Kim, T., Choi, J., Ahn, J., Kim, K.S., Lim, K.J., 2007. Evaluation of SWAT applicability to simulation of sediment behaviors at the Imha-Dam watershed, *J. Korean Soc. Water Qual.* 23, 467-473.
- Robertson, D.M., Roerish, E.D., 1999. Influence of various water quality sampling strategies on load estimates for small streams, *Water Resour. Res.* 35, 3747-3759.
- Robertson, D.M., 2003. Influence of different temporal sampling strategies on estimating total phosphorus and suspended sediment concentration and transport in small streams, *J. Amer. Water Resour. Assoc.* 39, 1281-1308.
- Santhi, C., Arnold, J.G., Williams, J.R., Dugas, W.A., Srinivasan, R., Hauck, L.M., 2001. Validation of the SWAT model on a large river basin with point and nonpoint sources, *J. Amer. Water Resour. Assoc.* 37, 1169-1188.
- To an, V., Dalo lu, A.T., 2008. An improved genetic algorithm with initial population strategy and self-adaptive member grouping, *Comput. & Struct.* 86, 1204-1218.