J. Korea Water Resour. Assoc. Vol. 54, No. 12 (2021), pp. 1305-1316
doi: 10.3741/JKWRA.2021.54.12.1305
pISSN 1226-6280
eISSN 2287-6138

Development of empirical formula for imbalanced transverse dispersion coefficient data set using SMOTE

Lee, Sunmi[a] · Yoon, Taewon[b] · Park, Inhwan[c]*

[a]Master Course, Department of Civil Engineering, Seoul National University of Science and Technology, Seoul, Korea
[b]Undergarduate Student, Department of Civil Engineering, Seoul National University of Science and Technology, Seoul, Korea
[c]Assistant Professor, Department of Civil Engineering, Seoul National University of Science and Technology, Seoul, Korea

Paper number: 21-104
Received: 26 October 2021; Revised: 22 November 2021; Accepted: 25 November 2021

Abstract

In this study, a new empirical formula for 2D transverse dispersion coefficient was developed using the results of previous tracer test studies, and the performance of the formula was evaluated. Since many tracer test studies have been conducted under the conditions where the width-to-depth ratio is less than 50, the existing empirical formulas developed using these imbalanced tracer test results have limitations in applying to rivers with a width-to-depth ratio greater than 50. Therefore, in order to develop an empirical formula for transverse dispersion coefficient using the imbalanced tracer test data, the Synthetic Minority Oversampling TEchnique (SMOTE) was used to oversample new data representing the properties of the existing tracer test data. The hydraulic data and the transverse dispersion coefficients in conditions of width-to-depth ratio greater than 50 were oversampled using the SMOTE. The reliability of the oversampled data was evaluated using the ROC (Receiver Operating Characteristic) curve. The empirical formula of transverse dispersion coefficient was developed including the oversampled data, and the performance of the results were compared with the empirical formulas suggested in previous studies using R^2. From the comparison results, the value of R^2 was 0.81 for the range of $W/H < 50$ and 0.92 for $50 < W/H$, which were improved accuracy compared to the previous studies.

Keywords: Transverse dispersion coefficient, SMOTE, Empirical formula, Imbalanced data

SMOTE를 이용한 편중된 횡 분산계수 데이터에 대한 추정식 개발

이선미[a] · 윤태원[b] · 박인환[c]*

[a]서울과학기술대학교 건설시스템공학과 석사과정, [b]서울과학기술대학교 건설시스템공학과 학부과정, [c]서울과학기술대학교 건설시스템공학과 조교수

요 지

본 연구에서는 과거 추적자실험결과를 이용하여 2차원 횡분산계수에 대한 새로운 추정식을 개발하고 추정식을 이용한 횡 분산계수 산정결과의 정확도를 검증했다. 다수의 추적자실험이 하폭 대 수심비가 50보다 작은 조건에서 수행되었기 때문에 기존 추적자실험결과만을 이용하여 개발한 추정식은 하폭 대 수심비가 50보다 큰 조건의 하천에 적용하는데 한계를 보인다. 따라서 특정 수리조건에 편중된 횡 분산계수 자료로부터 횡 분산계수 추정식을 개발하기 위해 SMOTE (Synthetic Minority Oversampling TEchnique)를 적용하여 기존 자료의 특성을 반영한 새로운 데이터를 생성했다. SMOTE 기법으로 하폭 대 수심비가 50보다 큰 조건에 대한 수리량과 횡 분산계수 데이터를 생성하였으며, ROC (Receiver Operating Characteristic) 곡선으로부터 생성된 데이터의 신뢰성을 검증했다. 새롭게 생성된 데이터를 포함하여 횡 분산계수 추정식을 개발했고, 추정식을 이용하여 계산한 횡 분산계수의 R^2(결정계수)를 계산하여 기존 연구에서 제안한 추정식과의 정확도를 비교했다. 그 결과, 본 연구에서 개발한 추정식을 이용하여 계산한 횡 분산계수의 R^2가 $W/H < 50$인 조건에서 0.81, $50 < W/H$인 조건에서 0.92를 나타내어 기존 추정식과 비교하여 향상된 정확도를 나타냈다.

핵심용어: 횡 분산계수, SMOTE (Synthetic Minority Oversampling TEchnique), 추정식, 데이터 편중

***Corresponding Author.** Tel: +82-2-970-6507
E-mail: ihpark@seoultech.ac.kr (I. Park)

1. 서 론

하천의 유속구조 및 오염물질의 혼합특성에 대해 이해하는 것은 수질오염사고로부터 수환경을 보호하기 위해 매우 중요한 일이다. 자연하천에 유입된 오염물질의 혼합은 전단류에 의한 분산으로 해석할 수 있다(Fischer *et al*., 1979). 전단류 분산은 혼합양상에 따라 중간역, 원역 혼합으로 구분할 수 있으며, 오염물질의 연직 혼합이 완료된 이후 하폭방향으로 완전히 혼합되기 전까지의 중간역 혼합해석은 2차원 이송-분산 방정식을 적용할 수 있다. 2차원 이송-분산 방정식은 다음 식과 같다.

$$\frac{\partial C}{\partial t}+u\frac{\partial C}{\partial x}+v\frac{\partial C}{\partial y}=D_L\frac{\partial^2 C}{\partial x^2}+D_T\frac{\partial^2 C}{\partial y^2} \quad (1)$$

여기서 C는 수심적분 된 농도, u, v는 각각 수심적분 된 종, 횡방향 유속, D_L은 종 방향 분산계수, D_T는 횡 방향 분산계수이다. Eq. (1)의 종, 횡 분산계수는 전단류에 의한 오염물질의 혼합능을 파악할 수 있는 주요 매개변수이다. 종, 횡 분산계수는 하천의 흐름구조와 지형 특성에 주로 의존한다. 따라서 하천의 사행에 따른 이차류의 발달, 하안에 존재하는 사수역, 하상의 불규칙성 등 복잡한 흐름특성을 보이는 자연하천에서 종, 횡 분산계수의 결정에 많은 어려움이 있다(Seo *et al*., 2005).

분산계수를 결정하는 방법에는 추적자 실험을 통해 취득한 농도 자료를 이용하는 관측법(observation method)과 기본 수리량을 바탕으로 분산계수를 산정하는 추정법(prediction method)으로 크게 나눌 수 있다. 관측법으로서는 주로 모멘트법(change of moment method)이 개발되어 적용되어 왔다. 가장 간단한 모멘트법으로는 단순 모멘트법(simple moment method, SMM)으로 이는 Sayre and Chang (1968)에 의해 제안되었으며, 그 후 Krishnappan and Lau (1977), Fischer *et al*., 1979, Webel and Schatzmann (1984), Nokes and Wood (1988), Rutherford (1994)가 횡 분산계수를 산정하는데 이용하였다. Holley (1971)는 횡방향 유속의 영향을 고려하기 위하여 범용 모멘트법(generalized moment method, GMM)을 제안하였고 이는 Boxall *et al*. (2003)에 의해 사행수로의 횡 분산계수 산정에 적용된 바 있다. 횡 방향 유속뿐만 아니라 곡선좌표계를 도입하여 하천의 만곡효과를 고려할 수 있는 곡선모멘트법(curvilinear moment method, CMM)이 Yotsukura and Sayre (1976)에 의해 제안되었고 Almquist and Holley (1985)가 실헐실 하천의 횡 분산계수 산정에 활용하였다. 모멘트법은 주로 횡 분산계수의 산정에 활용되어 왔기 때문에 이후 Baek *et al*. (2006)은 종, 횡 분산계수를 동시에 선정할 수 있는 2차원 추적법(2D routing procedure, 2D-RP)을 개발했다. 2D-RP는 불규칙한 하폭 및 하상을 고려할 수 없는 단점이 있어서 Baek *et al*. (2006) 및 Seo *et al*. (2005)은 자연하천에서 보다 범용적으로 추적법을 적용하기 위해 유관 개념을 도입한 2차원 유관추적법(two-dimensional stream-tube routing procedure, 2D ST-RP)법을 제시하였다(Han *et al*., 2017). 추정법은 전단류 분산이론에 따라 이론적으로 분산계수를 유도하는 이론식(theoretical formula)과 다수의 분산계수 산정결과를 바탕으로 회귀분석을 통해 개발된 추정식(empirical formula)으로 분류할 수 있다. 하지만 이론식의 난해함을 경험적 방법론으로 간략화한다거나, 이론적 배경에 바탕을 둔 추정식을 개발하기도 하므로, 두 가지 방법론은 상충되기 보다는 상호 보완적인 관계에 있다(Baek and Seo, 2007).

상기 서술한 바와 같이 오염물질의 혼합예측을 위한 분산계수는 관측법과 추정법을 적용하여 결정할 수 있다. 하지만 관측법의 경우 하천에서 추적자 실험결과로부터 산정된 분산계수 자료가 매우 제한적이기 때문에 이송-분산 방정식의 해석을 위한 적용에 한계가 있다. 따라서 실무적 관점에서 오염물질 혼합예측을 위한 분산계수는 추정법에 따라 이론식과 추정식으로부터 산정할 필요가 있다. 추정법으로서 횡 혼합에 지배적인 영향을 주는 하천의 유속구조를 감안하여 개발된 이론식(Baek *et al*., 2006)과 하천에서 비교적 쉽게 측정이 가능한 기본 지형인자와 수리량에 근거한 추정식이 제안된 바 있다(Fischer, 1969; Yotsukura *et al*., 1970; Yotsukura and Sayre, 1976; Gharbi and Verrette, 1998). 그러나 이러한 이론식과 추정식은 하천의 사행특성이나 이차류 등의 영향을 적절하게 해석하지 못하고 특정 하천 자료에 의존하여 유도된 식이기 때문에 모든 하천에 범용적으로 적용이 어려운 단점을 가지고 있다. 또한, 추적자실험을 통하여 산정한 추정식은 5~20개의 데이터를 활용하여 산정했기 때문에 수리특성이 다른 하천에 적용하는 경우 정확도가 떨어지는 경향이 있다(Baek and Seo, 2007). 기존 분산계수 산정법의 한계를 극복하고자 머신러닝을 활용한 분산계수 예측 연구가 2009년부터 Noori *et al*. (2009)에 의해 제안되었지만, 이는 상대적으로 데이터 수가 많은 1차원 종분산계수에 적용한 것으로 분석 데이터 수가 상대적으로 부족한 2차원 분산계수산정 연구에 적용된 바 없다. 따라서 기존 연구의 한계를 보완한 2차원 분산계수 산정을 위한 연구가 필요하다.

본 연구에서는 기존 추적자 실험결과를 활용하여 2차원 횡분산계수 추정식을 개발하였으며, 추정식 개발에 필요한 현장실험자료 부족 문제를 극복하기 위해 SMOTE (Synthetic Minority Oversampling TEchnique)기법을 활용했다. SMOTE는 편중된 자료 그룹에 대해 소수그룹으로 구분되는 자료특성

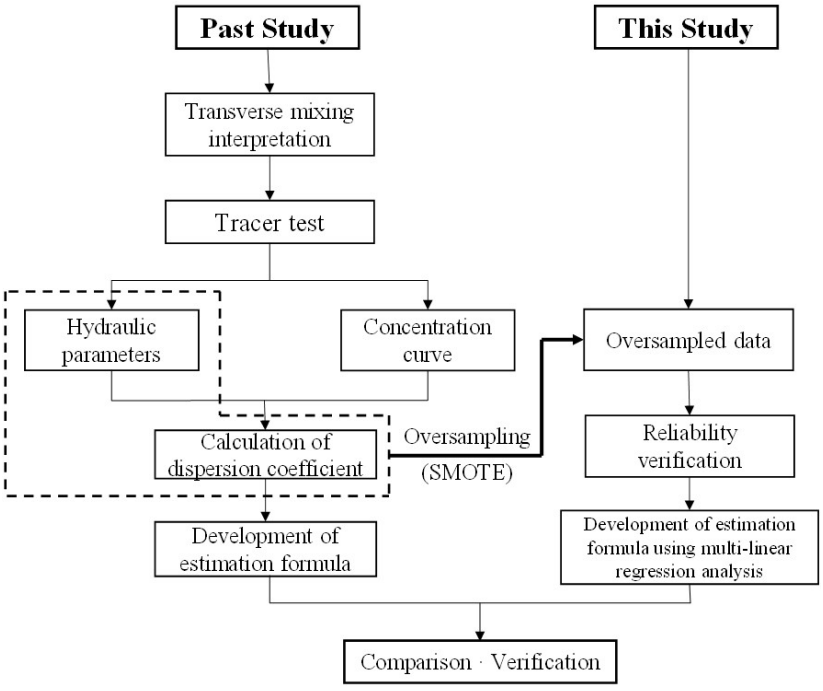

Fig. 1. Flow chart of the research method for estimating transverse dispersion coefficient using SMOTE

을 반영한 데이터 생성 기법이다. 2차원 횡 분산계수 추정식 개발을 위해 국내외 추적자 실험자료 53개를 수집하였다. 53개의 추적자 실험 데이터를 분석하고, SMOTE를 활용하여 소수 그룹에 대한 데이터 수를 증폭시켰다. 그리고 추정식 개발에 적용하기 위해 증폭된 데이터의 신뢰성을 검증하였다. 또한 본 연구에서 개발한 추정식과 기존 연구에서 제안된 추정식을 비교하여 본 연구방법의 신뢰성을 검증하였다. Fig. 1은 상기 서술한 연구절차에 대한 흐름도를 나타낸다.

2. 이론적 배경

2.1 횡 분산계수 추정식

자연하천에서 오염물의 거동에 영향을 미치는 인자는 크게 유체의 역학적 성질, 하천의 흐름 특성, 그리고 지형인자로 분류할 수 있다. 유체의 역학적 성질은 밀도, 점성계수 등이 있으며, 하천의 흐름 특성에 포함되는 인자는 유속, 전단유속, 수심 등이 있다. 마지막으로 지형인자로서 하폭, 하상형상, 사행도 등이 있다. 이러한 자연하천에서 오염물 거동에 영향을 미치는 인자와 횡 분산계수와의 관계를 수학적으로 표현하면 다음과 같다.

$$D_T = f_1(\rho,\ \mu,\ U,\ U_*,\ H,\ W,\ r_c,\ S_f,\ S_n) \tag{2}$$

여기서 ρ는 유체의 밀도, μ는 유체의 점성계수, U는 단면평균 유속, U_*는 마찰유속, H는 단면 평균 수심, W는 하폭, r_c는 곡률반경, S_f는 에너지경사, S_n는 하천의 사행도이다. Eq. (2)의 각 항들은 Buckingham π이론을 사용하여 다음 식과 같이 중요한 물리적 의미를 갖는 무차원항으로 나타낼 수 있다(Seo *et al*., 2005).

$$\frac{D_T}{HU_*} = f_2\left(\frac{Ud\rho}{\mu},\ \frac{U}{U_*},\ \frac{W}{H},\ S_f,\ S_n\right) \tag{3}$$

여기서 $\frac{Ud\rho}{\mu}$는 Reynolds수, $\frac{W}{H}$는 하폭 대 수심비, $\frac{U}{U_*}$는 유속 대 마찰유속의 비에 대한 무차원 변수이다. S_f의 영향은 사행도에 의한 영향보다 작고, $\left(\frac{U}{U_*}\right)$에 포함시킬 수 있으므로 제외할 수 있다. 또한 자연하천에서는 주로 난류흐름이 발생하므로 Reynolds 수에 의한 영향은 미소하므로 제외할 수 있다. $\left(\frac{U}{U_*}\right)$의 경우 하천 사행에 의한 이차류에 따른 에너지 손실도 내포되어 있는 것으로 판단되나 이차류에 대한 횡 분산 증가를 별도로 고려하기 위하여 사행도 (S_n)를 적용할 수 있다. 사행특성을 감안하기 위해 곡률반경 대신 사행도를 선택하였는데, 그 이유는 만곡부가 교호하는 사행하천의 경우 여러 값의 곡률 반경을 가지기 때문에 이를 종합적으로 고려할 수 있는 사행도를 선택했다(Jeon *et al*., 2007). 따라서 Eq. (3)에서

Table 1. Empirical formulas for transverse dispersion coefficient using tracer test results

Reference	Empirical Formulas
Yotsukura *et al*. (1968)	$\frac{D_T}{HU_*}=0.6$
Fischer (1969)	$\frac{D_T}{HU_*}=C\left(\frac{U}{U_*}\right)^2\left(\frac{H}{R_c}\right)^2$ (In the laboratory channel, C=25)
Yotsukura and Sayre (1976)	$\frac{D_T}{HU_*}=0.4\left(\frac{U}{U_*}\right)^2\left(\frac{W}{R_c}\right)^2$
	$\frac{D_T}{HU_*}=0.02\left(\frac{U}{U_*}\right)^2\left(\frac{W}{R_c}\right)^2$
Bansal (1971)	$\frac{D_T}{HU_*}=0.002\left(\frac{W}{H}\right)^{1.498}$
Fischer *et al*. (1979)	$\frac{D_T}{HU_*}=0.3\sim0.9$
Gharbi and Verrette (1998)	$D_T=0.0035\left[\frac{\left(\frac{Q}{H}\right)^{1.75}\left(\frac{W}{H}\right)^{0.25}}{D_L^{0.75}}\right]+0.0005$
Deng *et al*. (2001)	$\frac{D_T}{HU_*}=0.145+\left(\frac{1}{3530}\right)\left(\frac{U}{U_*}\right)\left(\frac{W}{H}\right)^{1.38}$
Jeon *et al*. (2007)	$\frac{D_T}{HU_*}=0.03\left(\frac{U}{U_*}\right)^{0.46}\left(\frac{W}{H}\right)^{0.3}S_n^{0.73}$
Baek and Seo (2013)	$\frac{D_T}{HU_*}=\left(77.88\frac{U}{U_*}\frac{H}{R_c}\right)^2\left\{1-\exp\left(-\frac{1}{77.88\frac{U}{U_*}\frac{H}{R_c}}\right)\right\}$
Seo *et al*. (2016)	$\frac{D_T}{HU_*}\sim\left(\frac{U}{U_*}\right)\left(\frac{W}{R_c}\right)$

자연하천에서의 횡 혼합에 중요하게 영향을 미치는 무차원 인자들만 정리를 하면 다음과 같다.

$$\frac{D_T}{HU_*}=f\left(\frac{U}{U_*},\ \frac{W}{H},\ S_n\right) \tag{4}$$

많은 기존의 연구들이 Eq. (4)의 무차원 수리인자를 이용하여 횡 분산계수 산정을 위한 추정식을 제안했다. Table 1은 기존 연구에서 제안한 횡 분산계수 추정식을 정리한 표이다.

2.2 횡 분산계수 산정을 위한 선행 추적자실험 결과

기존 추적자 실험결과를 이용한 횡 분산계수의 추정식 개발을 위해 26개의 국외 추적자실험 자료와 27개의 국내 추적자 실험 자료를 수집했다. 기존 추적자 실험자료를 정리한 표는 부록에 수록하였으며, 이 중 10개의 데이터(No. 42~51)는 자연하천이 아닌 실험수로, 2개의 데이터(No. 52~53)는 실규모 사행수로에서 실험한 연구 결과이다.

Appendix 1에서 무차원 횡 분산계수(D_T/HU_*)와 Eq. (4)

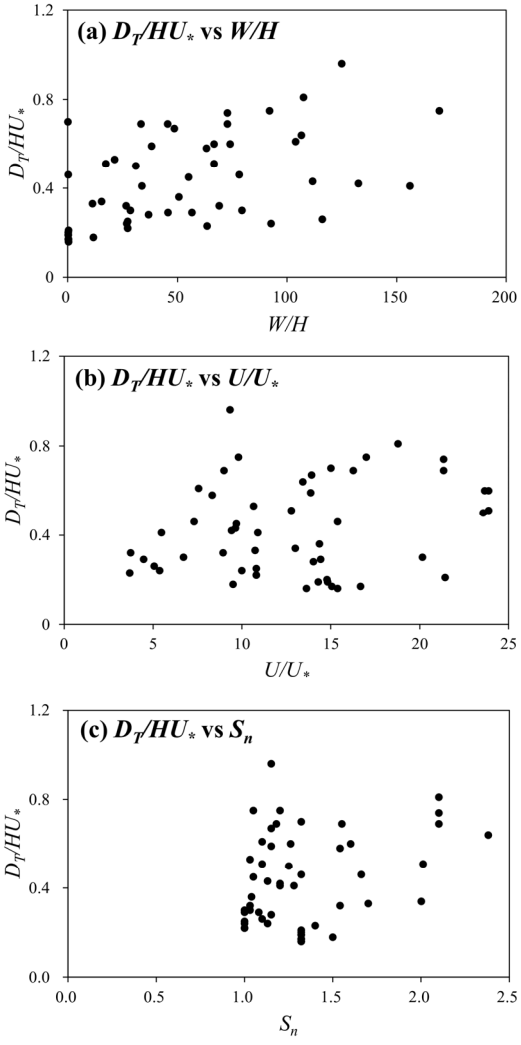

Fig. 2. Relations between the dimensionless transverse dispersion coefficient and the hydraulic parameters

에서 제시한 무차원 수리인자의 관계를 Fig. 2에 도시하였다. D_T/HU_*와 무차원 변수들의 관계를 보면, 비록 데이터의 산포도는 크지만 S_n, W/H 그리고 U/U_*의 증가에 따라 D_T/HU_*가 상승하는 경향을 보이고 있다. D_T/HU_*는 0.16~0.96의 분포를 보이며, 이중 자연하천에 대한 D_T/HU_*의 범위는 0.22~0.96, 실험수로에 대해서는 0.16~0.70의 범위를 나타내어 자연하천과 실험수로에 대한 값의 차이가 크진 않았다. U/U_*의 데이터 범위도 자연하천과 실험수로에서 각각 3.7~23.9, 9.5~21.4의 범위를 보여 수로 규모에 의한 차이가 크지 않았으나, W/H의 경우에는 자연하천에서 15.4~169.5, 실험수로에서 0.1~11.7의 범위를 나타내어 두 데이터 그룹의 차이가 나타났다. S_n은 자연하천과 실험수로에서 각각 1~2.38, 1.32~1.70의 범위를 보였는데 1~1.5의 데이터가 전

체의 77%를 차지하여 데이터의 활용범위가 제한적이었다. 따라서 데이터의 분포가 고르게 나타나는 W/H와 U/U_*를 D_T/HU_*의 추정식 개발에 활용했다.

2.3 SMOTE를 이용한 데이터 Oversampling

편향된 데이터 세트를 학습함에 따라 발생하는 예측 성능 저하 문제를 해결하기 위해 부족한 데이터를 생성하는 오버샘플링(oversampling) 기법을 적용할 수 있다(Chawla *et al*., 2002; Douzas *et al*., 2018). 오버샘플링 기법은 학습데이터 세트의 균형을 맞추기 위해 소수그룹(minority class sample)의 샘플 수를 다수그룹(majority class sample) 수준으로 증폭시켜 학습에 필요한 충분한 데이터를 확보하는 기법이다(Zhu *et al*., 2017). 오버샘플링 기법에는 K 근접이웃 이론을 기반으로 하는 SMOTE (Synthetic Minority Oversampling TEchnique)와 샘플 간 거리에 따라 가중치를 두어 새로운 샘플을 합성하는 MWMOTE (Member Weighted Minority Oversampling TEchnique) 기법이 있다(Chawla *et al*., 2002; Barus *et al*., 2014). 이중 SMOTE 기법은 데이터 과적합 문제를 완화할 수 있는 방식으로 가장 많이 사용되고 있는 기법이다(Nitesh, 2002; Mahamud *et al*., 2016). 이에 따라 SMOTE 기법은 편향된 관측데이터 보간을 통해 홍수 예측모델의 정확도 향상을 위해 활용된 바 있다(Wu *et al*., 2020; Snieder *et al*., 2021).

본 연구에서 사용하는 횡 분산계수 데이터 세트는 전체 샘플 수가 적고 $W/H<50$ 조건에 치우쳐 있기 때문에 오버샘플링 기법 중 SMOTE를 사용하여 데이터를 증폭시켰다. SMOTE에 의한 소수그룹 샘플의 데이터 합성은 K 근접이웃 이론(K-Nearest Neighbor, KNN)을 기반으로 한다(Fig. 3). SMOTE는 KNN기법에 따라 소수그룹 내 한 표본에 가장 가까운 K개의 이웃을 연결하여 소수그룹의 데이터 특성을 따르는 합성샘플(synthetic sample)을 생성한다. KNN 기법은 우선 소수그룹에서 기준샘플을 랜덤으로 선택하여(Fig. 3에서 x_1) 기준샘플로부터 유클리드 거리가 가장 가까운 K개의 샘플(Fig. 3에서 $x_2 \sim x_6$)을 찾는다. K개의 샘플 중에서 무작위로 하나의 샘플을 선택하는데 그 샘플을 KNN 샘플이라고 한다. 기준샘플과 KNN 샘플 간 거리에 0~1사이에 생성된 난수를 곱하여 두 샘플을 잇는 선을 따라 합성샘플을 생성한다. Fig. 3에서 KNN 샘플로 x_2가 선택되면 x_1과 x_2 사이의 거리에 무작위 난수를 곱하여 합성한 새로운 샘플인 e가 생성된다. 이러한 과정을 반복하여 소수그룹의 샘플 수가 다수그룹의 샘플 수와 같아지도록 데이터를 증폭(oversampling)한다.

SMOTE를 통해 합성한 소수 샘플의 활용을 위해 신뢰성 검증이 필요하다. 일반적으로 머신러닝 알고리즘의 성능은 예

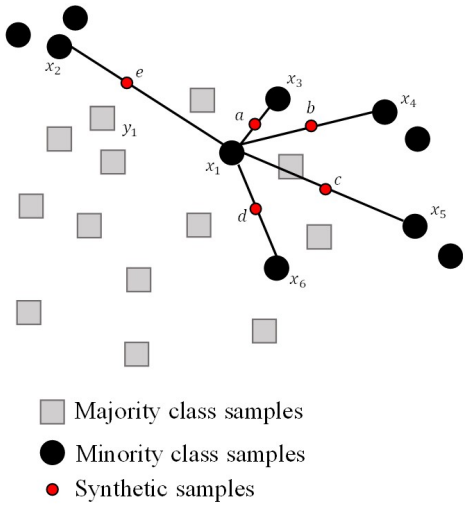

Fig. 3. Conceptual diagram of the K-Nearest Neighbors technique for data oversampling

측 정확도를 사용하여 평가된다. 그러나 데이터가 불균형적이거나 그 수가 현저히 작을 경우 그 방법은 적절하지 않다. 따라서 Fig. 4의 ROC (Receiver Operating Characteristic) 곡선을 사용하여 판별 모형의 성능을 평가한다(Swets, 1988). ROC 곡선은 민감도와 특이도가 어떤 관계에 있는지 표현한 그래프이다. FPR (False Positive Rate)와 TPR (True Positive Rate)은 ROC 곡선에서 각각 x, y 축에 표시되는 값이다. 여기서 Positive는 판단자가 '그렇다'라고 판별했다는 의미이고, True와 False는 각각 '판단을 올바르게 했다'와 '판단을 올바르지 않게 했다'는 의미를 갖는다. 다시 말해 TP (True Positive)는 예측 결과가 '그렇다'고 판단한 것이고 실제로 올바른 값을 판단한 것이며, FP (False Positive)는 '그렇다'고 판단했지만 실제로 올바른 값을 판단하지 못하여 잘못 판단한 것을 의미한다. 이는 Fig. 5의 오차행렬(Confusion Matrix)에서 쉽게 이해할 수 있다. 이때 문턱값(threshold)를 어떻게 설정하느냐에 따라 TPR과 FPR이 달라질 수 있다. 예를 들어 문턱값이 낮은 경우 TPR과 FPR이 모두 높아지고 문턱값이 높은 경우 TPR과 FPR이 모두 낮아지게 된다. 즉, 문턱값이 변함에 따라서 TPR과 FPR이 비례적으로 증가하거나 감소한다. 따라서 모형의 성능평가를 위해 문턱값의 변화에 관계없는 ROC 곡선의 면적을 계산하여 모델의 전반적인 성능을 확인한다.

Fig. 4의 AUC (Area Under the Curve)는 ROC 곡선의 면적을 계산한 값으로서 분류 성능 지표로 사용된다(Bradley, 1997). AUC의 값은 문턱값에 대해 변화하는 ROC 곡선과 달리 오버샘플링 된 데이터의 신뢰성 검토에 있어 문턱값의 영향을 받지 않는다. 이때 ROC 곡선은 낮은 FPR에 대해 1에 가까운 높은 TPR을 보일 때 자료가 신뢰성을 갖는 것으로 평가한

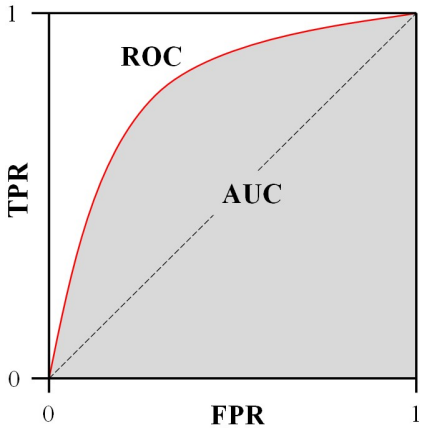

Fig. 4. Conceptual diagram of the ROC curve and the AUC

Predicted Data			
True	False		
TP (True Positive)	FN (False Negative)	True	Observed Data
FP (False Positive)	TN (True Negative)	False	

Fig. 5. Confusion Matrix for ROC curve

다. 따라서 ROC 곡선이 직사각형에 가까운 곡선이 되어 AUC의 값이 1에 가까워지면 신뢰성 있는 데이터가 생성되었다고 판단할 수 있게 된다.

3. 횡 분산계수 추정식 개발

3.1 횡 분산계수의 오버샘플링

2차원 횡 분산계수의 추정식 개발을 위해 Appendix 1의 추적자실험 자료를 이용하였으며, 추정식의 정확도 향상을 위해 SMOTE를 사용하여 추적자실험 데이터 수를 증폭시켰다. SMOTE를 이용하여 추정식 개발에 필요한 D_T/HU_*, W/H, U/U_*의 데이터 합성을 위해 공통적 특성을 갖는 데이터 그룹을 분류해야 한다. 이를 위해 W/H의 범위에 따른 횡 분산계수 추정식의 적합도를 분석한 과거 연구(Baek and Seo, 2017)를 토대로 $W/H<50$, $50\leq W/H<100$, $100\leq W/H$의 세 구간으로 클래스를 분류하여 오버샘플링을 진행했다. SMOTE를 진행할 때, 학습에 필요한 훈련데이터 세트와 결과를 테스트해보는 시험 세트를 무작위로 분리한다. 이 연구에서는 훈련데이터 세트의 비율을 전체의 85%로 지정했다. 따라서 총 53개

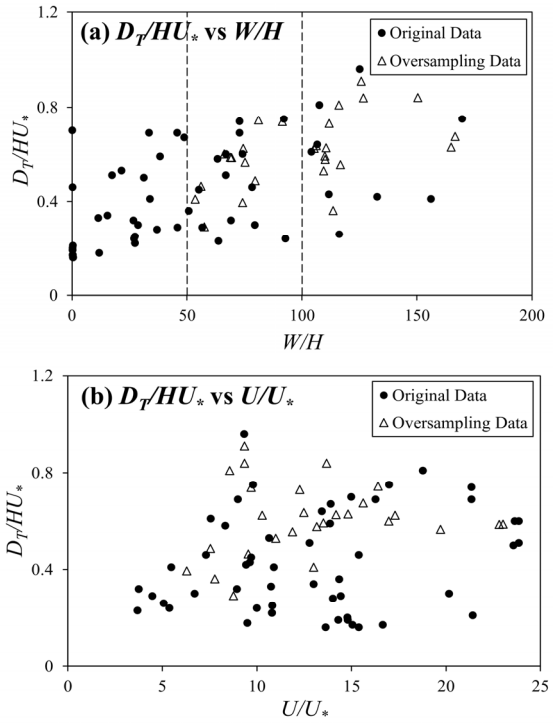

Fig. 6. Relations between the oversampled dimensionless transverse dispersion coefficient and the hydraulic parameters

의 원 데이터 중에 85%인 45개로 오버샘플링을 진행했다. Fig. 6은 기존 데이터와 오버샘플링 된 데이터를 함께 도시한 그래프이다. $W/H<50$의 훈련데이터 수는 24개로 다수그룹으로 분류되었으며, 소수 그룹에 해당하는 $50\leq W/H<100$, $100\leq W/H$의 데이터가 각각 12개에서 24개, 9개에서 24개로 증폭되어 전체 데이터 수는 기존의 53개 데이터에 증폭된 27개를 포함하여 80개가 되었다. U/U_*는 W/H의 데이터 분류 기준에 따라 다수그룹과 소수그룹으로 분류되었으며, $50\leq W/H<100$와 $100\leq W/H$에 해당하는 U/U_*의 데이터가 증폭되었다.

Fig. 6에서 오버샘플링 된 데이터의 활용을 위해 증폭된 데이터들이 해당하는 각 소수그룹의 데이터 특성을 반영할 수 있어야 한다. 따라서 오버샘플링 된 데이터의 신뢰성을 검증하기 위해 합성 데이터 그룹에 대한 ROC 곡선을 Fig. 7에 도시하였다. 그 결과 전체 데이터에 해당하는 검증 결과의 평균을 나타내는 micro average AUC가 0.91, 각 그룹 별 데이터에 해당하는 검증 결과의 평균을 다시 그룹의 개수로 평균을 내는 macro average AUC가 0.88을 나타낸다. 보통 그룹별 불균형 데이터세트를 검증하는 데 전체 샘플의 수를 고려하는 micro average가 더 효과적인 평가지표이다(Jurafsky and Martin, 2017). 따라서 오버샘플링 된 데이터가 기존 데이터의 경향을

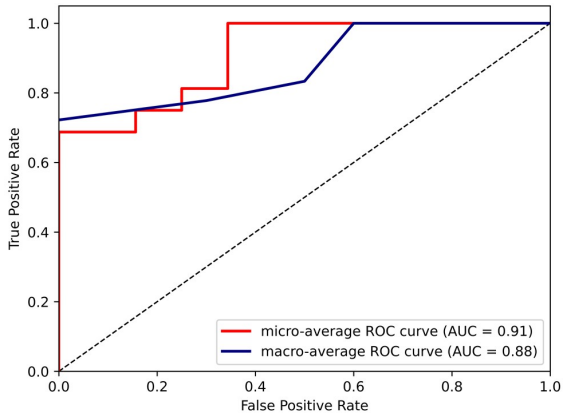

Fig. 7. Results of ROC curves and AUCs

적절히 반영하는 것으로 판단할 수 있다.

3.2 횡 분산계수 추정식 개발

SMOTE를 통해 증폭된 데이터를 활용하여 D_T/HU_*에 대한 추정식을 개발했다. 추정식은 Eq. (4)와 같이 W/H, U/U_*을 이용한 비선형다중회귀식의 형태로 결정했다.

$$\frac{D_T}{HU_*} = a\left(\frac{W}{H}\right)^b\left(\frac{U}{U_*}\right)^c \tag{5}$$

여기서, a, b, c는 회귀상수이다. Eq. (5)를 Eq. (6)와 같이 선형화한 후 파이썬(Python)에 포함 된 라이브러리인 싸이킷런(scikit-learn)을 적용했다. 싸이킷런은 최소제곱법(least squares method)을 사용하여 회귀상수를 산정하며, 전체 데이터의 70%를 훈련데이터 세트, 30%를 검증 데이터 세트로 나누어 가장 적합한 선형방정식을 찾는 알고리즘이다.

$$\ln\left(\frac{D_T}{HU_*}\right) = \ln(a)+b\ln\left(\frac{W}{H}\right)+c\ln\left(\frac{U}{U_*}\right) \tag{6}$$

Appendix 1의 실험데이터와 SMOTE를 통해 오버샘플링 된 데이터를 사용하여 횡 분산계수 추정식에 대한 회귀상수를 산정하였으며, 그 결과는 Eq. (7)과 같다.

$$\frac{D_T}{HU_*} = 0.0703\left(\frac{W}{H}\right)^{0.2002}\left(\frac{U}{U_*}\right)^{0.4514} \tag{7}$$

Eq. (7)의 추정식에서 U/U_*의 회귀상수가 W/H보다 더 크게 산정되었으며, 이는 D_T/HU_*가 U/U_*의 변화에 더 큰 영향을 받는다는 것을 뜻한다. Eq. (7)을 이용하여 W/H와 U/U_*의

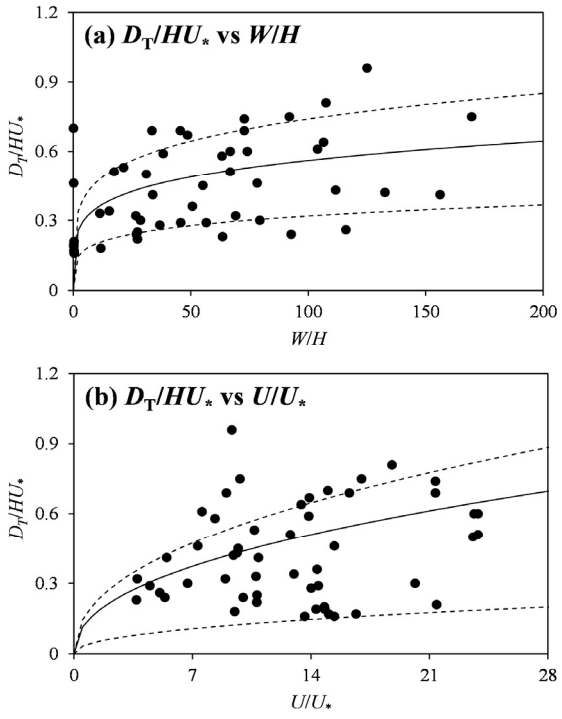

Fig. 8. Variations of the transverse dispersion coefficients against to the non-dimensional parameters

변화에 따른 횡 분산계수 산정결과를 분석했다. Fig. 8은 Appendix 1의 실험데이터를 이용하여 계산한 횡 분산계수와 실험데이터를 비교한 그래프이다. Fig. 8(a)의 실선은 U/U_*의 평균값에 대한 횡 분산계수의 변화를 나타내며, 두 점선은 U/U_*의 최대, 최소값에 대한 횡 분산계수 변화를 나타낸다. 그리고 Fig. 8(b)의 실선은 W/H의 평균값을 이용한 횡 분산계수의 변화, 두 점선은 W/H의 최대, 최소값에 대한 횡 분산계수 변화를 나타낸다. 따라서 Fig. 8의 점선은 Eq. (7)을 이용하여 산정 가능한 횡 분산계수 범위를 뜻한다. 그런데 일부 실험데이터가 Eq. (7)의 추정식 산정 한계를 벗어난 결과가 나타났다. 이러한 결과는 높은 산포도를 갖는 원 데이터를 비선형 회귀모형으로 나타내는 것에 한계가 발생함을 보여준다.

Fig. 8(a)의 실선과 점선의 간격으로부터 U/U_*의 변화에 대한 D_T/HU_* 추정식의 민감도를 확인할 수 있다. 임의의 W/H에 대해 $12.9 < U/U_* < 23.9$ 범위 내 D_T/HU_*의 변화(ΔD_u)는 $3.7 < U/U_* < 12.9$에 대한 변화(ΔD_l)와 유사하게 나타났다. 반면, Fig. 8(b)에서 볼 수 있듯 임의의 U/U_*에 대한 D_T/HU_*의 변화는 $0.1 < W/H < 51.7$ 보다 $51.7 < W/H < 169.5$의 범위에서 더 크게 나타났다. 이러한 결과는 W/H에 대한 회귀상수가 U/U_*에 비해 상대적으로 작기 때문에 W/H에 대한 민감도가 평균값 이상의 값에 대해 D_T/HU_*의 변화에 큰 영향을 주지

않기 때문에 발생했다. W/H가 상대적으로 작은 소하천에서는 하천 양안이 유속구조의 변화에 미치는 영향이 중·대하천과 비교하여 더 크다. 따라서 W/H가 평균값보다 큰 조건에서는 W/H의 변화가 유속구조의 변화에 미치는 영향이 감소하며, 오염물질의 횡 혼합 변화 또한 감소하는 것으로 판단할 수 있다. 이에 따라 본 연구에서 제안한 추정식이 W/H의 변화가 오염물질 횡 혼합의 변화를 적절히 반영하고 있음을 보여준다.

3.3 횡 분산계수 추정식의 검증

본 연구에서 제안한 횡 분산계수 추정식의 우수성을 검증하기 위해 Table 1의 추정식 중 Appendix 1의 실험데이터를 활용 가능한 Bansal (1971), Deng *et al*. (2001), Jeon *et al*. (2007)의 계산결과와 비교했다. Fig. 9은 추정식을 이용하여 계산한 횡 분산계수와 Appendix 1의 실험데이터를 비교한 결과이다. Fig. 9에서 대각선 상에 데이터가 위치한 경우 추정식이 실험데이터를 잘 산정함을 의미하며, Bansal (1971)의 추정식은 횡 분산계수를 과대산정하는 경향이 있음을 보여준다. Deng *et al*. (2001)의 추정식 또한 과대산정하는 경향을 보였는데, 이 두 경험식은 W/H에 대한 회귀상수가 U/U_*보다 높게 산정된 특징을 갖는다. 반면, Jeon *et al*. (2007)은 과대산정된 결과가 거의 나타나지 않았으나 일부 실험데이터에 대해 과소산정하는 경향을 보였다. 본 연구에서 제안한 추정식의 계산결과는 과소 또는 과대산정하는 편중된 결과를 거의 보이지 않았으나 대각선 주변에 산포하는 한계를 나타냈다.

Fig. 9에서 비교한 세 추정식과 본 연구에서 제안한 추정식의 계산 정확도에 대한 정량적 분석을 수행했다. Baek and Seo (2017)은 $W/H<50$인 조건에서는 Bansal (1971), Deng *et al*. (2001)의 추정식이 적용가능하며, $W/H>50$인 조건에서 Jeon *et al*. (2007)의 추정식을 적용 가능함을 제안한 바 있다. 이에 따라 W/H의 범위를 구분하여 추정식의 계산 정확도를 비교했다. 계산 정확도 비교를 위해 다음 식과 같이 회귀식의 적합도를 나타내는 척도인 결정계수(R^2)를 이용했다.

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{y_i}-\bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2} \tag{8}$$

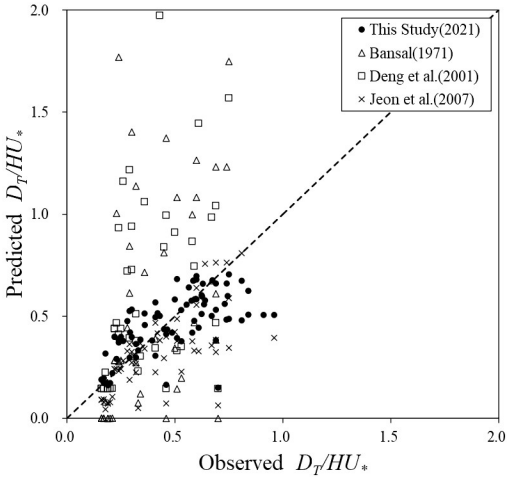

Fig. 9. Comparisons of the transverse dispersion coefficient formulas

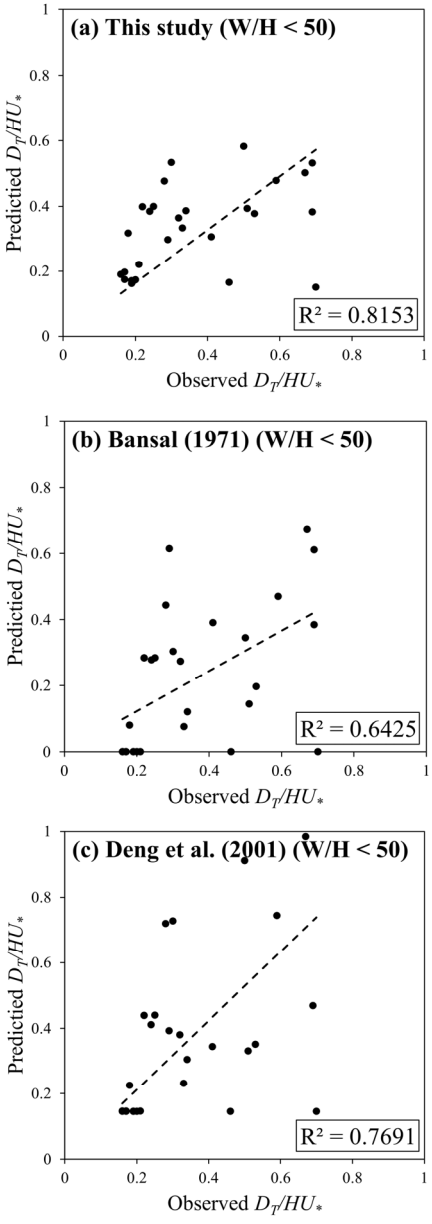

Fig. 10. R^2 of the proposed empirical formulas of the transverse dispersion coefficient ($W/H<50$)

여기서 $\hat{y_i}$는 추정식으로부터 계산된 횡 분산계수, $\bar{y}$는 관측한 실험데이터의 평균, y_i는 횡 분산계수 실험데이터, n은 샘플의 수이다. R^2가 1에 가까울수록 추정식의 적합도가 높음을 나타낸다. Fig. 10은 $W/H<50$인 조건에서 본 연구와 Bansal (1971), Deng *et al*. (2001)의 횡 분산계수 산정 결과를 비교한 그림이다. 본 연구에서 제안한 추정식의 R^2는 0.82로 계산되어 Bansal (1971), Deng *et al*. (2001)의 추정식보다 향상된 정확도를 나타냈다. 그리고 Fig. 11은 $W/H>50$인 조건에서 Jeon *et al*. (2007)의 횡 분산계수 산정결과와 비교한 결과이다. 그 결과, Jeon *et al*. (2007) 추정식의 R^2는 0.91, 본 연구의 R^2는 0.92로 나타나 다소 개선된 결과를 보였다. 이러한 결과는 S_n을 추정식에 반영하지 않아도 오버샘플링을 통해 추정식의 정확도를 향상시킬 수 있음을 보여준다. 또한 소수그룹의 특성을 반영한 데이터를 생성함으로써 횡 분산계수 추정값이 편중되는 현상을 완화하는 결과를 나타냈다.

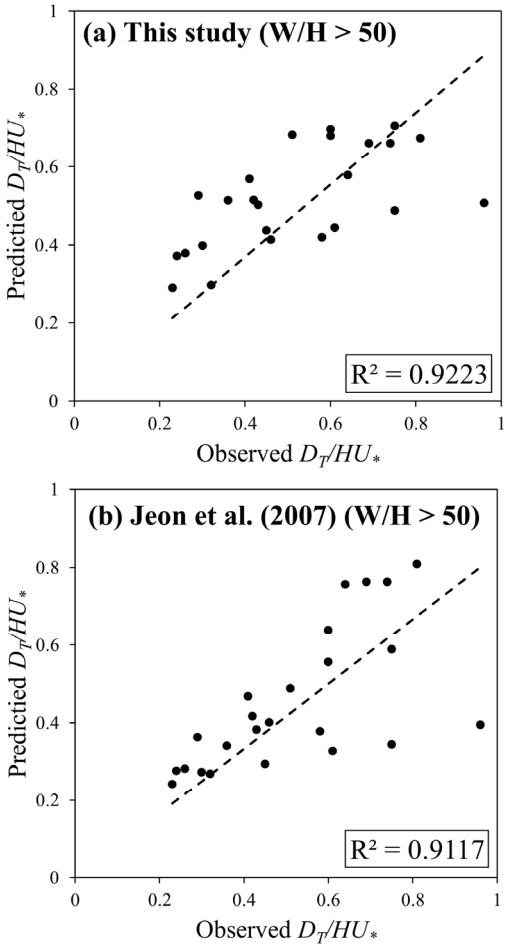

Fig. 11. R^2 of the proposed empirical formulas of the transverse dispersion coefficient ($W/H > 50$)

4. 결 론

본 연구에서는 하천 내 오염물질의 2차원 거동해석에 필요한 횡 분산계수 산정을 위해 무차원 수리인자를 활용한 횡 분산계수 추정식을 개발했다. 횡 분산계수의 추정식 개발을 위해 자연하천과 실험수로에서 수행된 추적자 실험결과를 분석하였으며, W/H와 U/U_*를 추정식 개발에 이용했다. 기존 연구에서 수집된 추적자 실험결과는 $W/H<50$인 조건에서 다수 수행되었으며, 다양한 수리조건에 적용 가능한 횡 분산계수 추정식 개발을 위해 $50< W/H$인 데이터에 대한 보완이 필요하다. 이를 위해 SMOTE 알고리즘을 이용하여 $50< W/H$에 해당하는 소수그룹에 대한 D_T/HU_*, W/H, U/U_* 데이터를 오버샘플링하였으며, 전체 데이터의 약 53%에 해당하는 데이터를 생성했다. 오버샘플링 된 데이터에 대한 신뢰성 검증을 위해 ROC로부터 AUC를 계산하였으며, AUC가 0.91이 되어 새롭게 생성된 데이터가 소수그룹의 데이터 특성을 적절히 반영하고 있음을 확인했다.

오버샘플링 된 데이터와 기존 실험데이터로부터 다중선형 회귀분석을 통해 횡 분산계수 추정식을 개발했다. 새로 개발된 추정식은 W/H보다 U/U_*에 대한 민감도가 높게 나타냈다. 특히 W/H가 기존 추적자실험데이터의 평균값인 51.7보다 큰 경우에는 W/H에 대한 민감도가 큰 폭으로 감소했다. 본 연구에서 제안한 추정식과 기존 연구의 추정식을 비교하였으며, 이때 Baek and Seo (2017)가 W/H의 범위에 따라 다른 추정식 적용이 가능함을 제안한 바와 같이 W/H의 범위를 구분하여 추정식의 R^2를 비교했다. 본 연구에서 제안한 추정식의 R^2는 $W/H<50$인 조건에서 0.81, $50< W/H$인 조건에서 0.92로 나타나 타 연구 결과보다 개선된 정확도를 보였다. 따라서 오버샘플링으로부터 소수그룹의 특성을 반영한 데이터를 생성함으로써 추정값이 다수그룹의 데이터 특성에 편중되지 않은 추정식 개발이 가능함을 알 수 있다. 또한 S_n을 고려하지 않아도 추정값의 정확도를 향상시킬 수 있음을 보였다. 하지만 Baek and Seo (2013)의 연구에서 보인 바와 같이 하천의 곡률에 대한 수리인자($P=\frac{U}{U_*}\frac{H}{R_c}$)가 횡 분산계수와 높은 상관관계를 나타냈기 때문에 향후 연구에서 이를 포함한 오버샘플링 및 추정식 개발 결과의 성능 검토가 필요하다.

본 연구에서는 적은 샘플을 활용한 횡 분산계수 추정식 개발의 한계를 극복하기 위해 오버샘플링을 통해 데이터를 증폭하여 새로운 추정식을 제안했다. 그러나 부족한 샘플 수로 인해 데이터 증폭에 한계가 있었으며, 이에 따라 기존식과 비교한 추정식의 정확도 향상에도 한계가 있었다. 또한 추적자 실험결과의 데이터 산포도가 크기 때문에 추정식의 정확도 향상

에 한계가 있었으며, 이는 단일 회귀식으로 정확도 높은 횡 분산계수 추정식 개발에 제한이 있음을 보여준다. 따라서 향후 연구에서는 머신러닝 기법을 활용하여 선형회귀식의 한계를 보완할 수 있는 횡 분산계수의 추정식 개발이 필요하다.

감사의 글

본 결과물은 환경부의 재원으로 한국환경산업기술원의 미세플라스틱 측정 및 위해성평가 기술개발사업의 지원을 받아 연구되었습니다(과제번호: 2021003110003).

References

Almquist, C.W., and Holley, E.R. (1985). *Transverse mixing in meandering laboratory channels with rectangular and naturally varying cross sections*. Technical Report CRWR-205, University of Texas, Austin, TX, U.S.

Baek, K.O., and Seo, I.W. (2007). "Evaluating coefficient of transverse dispersion induced by shear flow." *Journal of the Korean Society of Civil Engineers B*, KSCE, Vol. 27, No. 1B, pp. 21-28.

Baek, K.O., and Seo, I.W. (2013). "Empirical equation for transverse dispersion coefficient based on theoretical background in river bends." *Environmental Fluid Mechanics*, Vol. 13, No. 5, pp. 465-477.

Baek, K.O., and Seo, I.W. (2017). "Estimation of transverse dispersion coefficient for two-dimensional mixing in natural streams." *Journal of Hydro-environment Research*, Vol. 15, pp. 67-74.

Baek, K.O., Seo, I.W., and Jung, S.J. (2005). "2-D mixing of instantaneous pollutants in meandering channels : II. Determination and analysis of dispersion coefficients." *Journal of the Korean Society of Civil Engineers B*, KSCE, Vol. 25, No. 6B, pp. 463-471.

Baek, K.O., Seo, I.W., and Jung, S.J. (2006). "Evaluation of transverse dispersion coefficient in meandering channel from transient tracer tests." *Journal of Hydraulic Engineering*, Vol. 132, No. 10, pp. 1021-1032.

Bansal, M.K. (1970). *Dispersion and reaeration in natural stream*. Ph. D. dissertation, Univesite de Kansas Laurence, KS, U.S.

Bansal, M.K. (1971). "Dispersion in natural streams." *Journal of the Hydraulics Division*, ASCE, Vol. 97, No. 11, pp. 1867-1886.

Barus, S., Islam, M.M., Yao,X., and Murase, K. (2014). "MWMOTE - Majority weighted minority oversampling technique for imbalanced data set learning." *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 2, pp. 405-425.

Beltaos, S. (1980). "Transverse mixing tests in natural streams." *Journal of the Hydraulics Division*, ASCE, Vol. 106, No. HY10, pp. 1607-1625.

Beltaos, S., and Day, T.J. (1978). "A field study of longitudinal dispersion." *Canadian Journal of Civil Engineering*, Vol. 5, pp. 572-585.

Boxall, J.B., Guymer, I., and Mariion, A. (2003). "Transverse mixing in sinuous natural open channel flows." *Journal of Hydraulic Research*, IAHR, Vol. 41, No. 2, pp. 153-165.

Bradley, A.P. (1997). "The use of the area under the ROC curve in theevaluation of machine learning algorithms." *Pattern Recognition*, Vol. 30, No. 7, pp. 1145-1159.

Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. (2002). "SMOTE : Synthetic minority over-sampling technique." *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321-357.

Deng, Z., Singh, V.P., and Bengtsson, L. (2001). "Longitudinal dispersion coefficient in straight rivers." *Journal of Hydraulic Engineering*, Vol. 127, No. 11, pp. 919-927.

Douzas, G., Bacao, F., and Last, F. (2018). "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE." *Information Sciences*, Vol. 465, pp. 1-20.

Engmann, J.E.O., and Kellerhals, R. (1974). "Transverse mixing in an ice-covered river." *Water Resources Research*, Vol. 10, pp. 775-784.

Fischer, H.B. (1969). "The effect of bends on dispersion coefficients in streams." *Water Resources Research*, Vol. 5, pp. 496-506.

Fischer, H.B. (1973). "Longitudinal dispersion and turbulent mixing in open- channel flow." *Annual Review of Fluid Mechanics*, Vol. 5, pp.59-78.

Fischer, H.B., List, E.J., Koh, R.C.Y., Imberger, J., and Brooks, N.H. (1979). *Mixing in inland and coastal waters*. Academic Press, NY, U.S.

Gharbi, S., and Verrette, J. (1998). "Relation between longitudinal and transversal mixing coefficients in natural streams." *Journal of Hydraulic Research*, IAHR, Vol. 36, No. 1, pp. 43-53.

Han, E.J., Kim, Y.D., Baek, K.O., and Seo, I.W. (2017). "Analytical and experimental study on dispersion and diffusion by tracer test." *Water for Future*, Vol. 50, No. 6, pp. 58-65.

Holley, E.R. (1971). *Transverse mixing in rivers*. Laboratory Report, No. S-132, Delft Hydraulics Lab, Netherlands.

Holley, E.R., and Abraham, G. (1973). "Field tests on transverse mixing in rivers." *Journal of Hydraulic Division*, ASCE, Vol. 99, No. HY12, pp. 313-2331.

Holley, F.M.Jr., and Nerat, G. (1983). "Field calibration of stream-tube dispersion model." *Journal of Hydraulic Engineering*, ASCE, Vol. 109, No. 11, pp. 1455-1470.

Jeon, T.M., Baek, K.O., and Seo, I.W. (2007). "Development of an empirical equation for the transverse dispersion coefficient in natural streams." *Environmental Fluid Mechanics*, Vol. 7, pp. 317-329.

Jurafsky, D., and Martin J.M. (2017). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*, 3rd ed, Pearson Eucation, London, UK, p. 67.

Krishnappan, B.G., and Lau, Y.L. (1977). "Transverse mixing in meandering channels with varying bottom topography." *Journal of Hydraulic Research*, IAHR, Vol. 15, No. 4, pp. 351-371.

Lau, Y.L., and Krishnappan, B.G. (1981). "Modeling transverse mixing in natural streams." *Journal of the Hydraulic Division*, ASCE, Vol. 107, No. HY2, pp. 209-226.

Mahamud, K.R.K., Zorkeflee, M., and Din, A.M. (2016). "Fuzzy distance-based undersampling technique for imbalanced flood data." *Proceedings of the Knowledge Management International Conference, UUM,* Chiang Mai, Thailand, pp. 509-513.

Nitesh, V.C., Kevin W.B., Lawrence, O.H., and Philip, W.K. (2002). "SMOTE: synthetic minority over-sampling technique." *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321-357.

Nokes, R.I., and Wood, I.R. (1988). "Vertical and lateral turbulent dispersion: Some experimental results." *Journal of Fluid Mechanics*, Vol. 187, pp. 373-394.

Noori, R., Karbassi, A., Farokhnia, A., and Dehghani, M. (2009). "Predicting the longitudinal dispersion coefficient using support vector machine and adaptive neuro-fuzzy inference system techniques." *Environmental Engineering Science*, Vol. 26, No.10, pp.1503-1510.

Rutherford, J.C. (1994). *River mixing*, John Wiley and Sons, Chichester, UK.

Sayre, W.W. (1979). "Shore-attached thermal plumes in rivers." *Modelling in rivers*, Edited by Shen, H.W., Wiley-Interscience, London, UK, pp.15.1-15.44.

Sayre, W.W., and Chang, F.M. (1968) *A laboratory investigation of open channel dispersion processes for dissolved, suspended, and floating dispersants*. Professional Paper, No. 433-E. US Geological Survey, U.S., pp. 1-71.

Seo, I.W., Baek, K.O., and Jeon, T.M. (2006). "Analysis of transverse mixing in natural streams under slug tests." *Journal of Hydraulic Research*, Vol. 44, No. 3, pp. 350-362.

Seo, I.W., Choi, H.J., Kim, Y.D., and Han, E.J. (2016). "Analysis of two-dimensional mixing in natural streams based on transient tracer tests." *Journal of Hydraulic Engineering*, Vol. 142, No. 8, pp. 1-16.

Seo, I.W., Jeon, T.M., and Baek, K.O. (2005). "Development of empirical equation of transverse dispersion coefficient for analysis of 2-D mixing in natural streams." *Journal of the Korean Society of Civil Engineers B*, KSCE, Vol. 25, No. 4B, pp. 247-255.

Shin, J., Seo, I.W., and Baek, D. (2020). "Longitudinal and transverse dispersion coefficients of 2D contaminant transport model for mixing analysis in open channels." *Journal of Hydrology*, Vol. 583, pp. 1-15.

Snieder, E., Abogadil, K., and Khan, U.T. (2021). "Resampling and ensemble techniques for improving ANN-based high-flow forecast accuracy." *Hydrology and Earth System Sciences*, Vol. 25, pp. 2543-2566.

Swets, J.A. (1988). "Measuring the accuracy of diagnostic systems." *American Association for the Advancement of Science*, Vol. 240, No. 4857, pp.1285-1293.

Webel, G., and Schatzmann, M. (1984). "Transverse mixing in open channel flow." *Journal of Hydraulic Engineering*, ASCE, Vol. 110, No. 4, pp. 423-435.

Wu, Y., Ding, Y., and Feng, J. (2020). "SMOTE-Boost-based sparse Bayesian model for flood prediction." *EURASIP Journal on Wireless Communications and Networking*, Vol. 78, pp.1-12.

Yotsukura, N., and Cobb, E.D. (1972). *Transverse diffusion of solutes in natural streams*, Professional Paper, No.582-C, U.S. Geological Survey, U.S., pp. 1-19.

Yotsukura, N. Fischer, H.B., and Sayre, W.W. (1970). *Measurement of mixing characteristics of the Missouri River between Sioux City, Iowa and Plattsmouth, Nebraska*. U.S. Geological Survey Water-Supply Paper, Washington D.C, U.S.

Yotsukura, N., and Sayre, W.W. (1976). "Transverse mixing in natural channels." *Water Resources Reseach*, Vol. 12, No. 4, pp. 695-704.

Yotsukura, N., Sayre, W.W., and Alsaffar, A.M. (1968). "Discussion of The mechanics of dispersion in natural streams by HB Fischer." *Journal of the Hydraulics Division*, Vol. 95, pp. 1009-1038.

Zhu, F., Lin, Y., and Liu, Y. (2017). "Synthetic minority oversampling technique for multiclass imbalance problems." *Pattern Recognition*, Vol. 72, pp. 327-340.

Appendix 1. Summary of tracer test results for estimation of transverse dispersion coefficients

No.	River	W (m)	H (m)	U (m/s)	U_* (m/s)	r_c (m)	W/H	U/U_*	S_n	D_T/HU_*	Reference
1	Mississippi River	177.8	3.90	1.30	0.080	-	45.59	16.27	1.18	0.69	Bansal (1970)
2	Missouri River	206.0	2.78	1.74	0.073	-	74.10	23.87	1.26	0.60	Yotsukura *et al*. (1970)
3	Missouri River	183.0	2.74	1.75	0.074	-	66.79	23.65	1.60	0.60	
4	Missouri River	183.0	2.74	1.74	0.073	-	66.79	23.87	1.10	0.51	Yotsukura and Cobb (1972)
5	Atrisco Feeder Canal	18.3	0.67	0.67	0.062	-	27.31	10.81	1.00	0.22	
6	Bow River	104.0	1.00	1.05	0.139	-	104.00	7.56	1.10	0.61	
7	South River	18.3	0.40	0.18	0.040	-	45.75	4.48	1.00	0.29	
8	Atrisco Feder Canal	18.3	0.67	0.66	0.061	-	27.31	10.82	1.00	0.25	Fischer (1973)
9	Atrisco Feder Canal	18.3	0.68	0.63	0.063	-	26.91	10.00	1.00	0.24	
10	Atrisco Feder Canal	18.3	0.67	0.67	0.062	-	27.31	10.81	1.00	0.22	
11	Bernardo Conveyance Canal	20.0	0.70	1.25	0.062	-	28.57	20.16	1.00	0.30	
12	Waal River	266.0	4.70	0.82	0.057	-	56.60	14.44	1.08	0.29	Holley and Abraham (1973)
13	Ijssel River	69.5	4.00	0.97	0.076	1923	17.38	12.78	2.01	0.51	
14	Waal River	266.0	5.25	1.06	0.074	-	50.67	14.36	1.04	0.36	
15	Missouri River	210.0	5.49	1.47	0.106	-	38.25	13.87	1.15	0.59	Engmann and Kellerhals (1974)
16	Missouri River	204.0	5.54	1.50	0.107	-	36.82	14.02	1.15	0.28	
17	Missouri River	201.0	4.13	1.28	0.092	-	48.67	13.91	1.15	0.67	
18	Lesser Salve River	43.0	2.80	0.65	0.050	-	15.36	13.00	2.00	0.34	Beltaos and Day (1978)
19	Missouri River	214.0	1.99	1.39	0.074	792	107.54	18.78	2.10	0.81	Sayre (1979)
20	Missouri River	214.0	2.94	1.58	0.074	792	72.79	21.35	2.10	0.74	
21	Missouri River	214.0	2.94	1.58	0.074	792	72.79	21.35	2.10	0.69	
22	Athavasca River	320.0	2.05	0.86	0.079	-	156.10	10.90	1.20	0.41	Beltaos (1980)
23	Athavasca River	373.0	2.20	0.95	0.056	-	169.55	17.00	1.20	0.75	
24	Athavasca River	252.0	1.90	0.49	0.052	-	132.63	9.42	1.20	0.42	
25	Grand River	59.2	0.51	0.35	0.069	310	116.08	5.07	1.10	0.26	Lau and Krishnappan (1981)
26	Isere River	70.0	2.25	1.40	0.059	1612	31.11	23.57	1.25	0.50	Holley and Nerat (1983)
27	Cheongmi Creek (C-Expt 1)	44.5	0.48	0.34	0.063	397	92.71	5.37	1.13	0.24	Seo *et al*. (2006)
28	Sum River (S-Expt 1)	54.0	0.69	0.34	0.047	381	78.26	7.31	1.66	0.46	
29	Sum River (S-Expt 3)	54.0	0.68	0.31	0.046	3624.2	79.41	6.71	1.03	0.30	
30	Hongcheon River (H-Expt1)	58.6	0.55	0.54	0.040	437.5	106.55	13.43	2.38	0.64	
31	Hongcheon River (H-Expt2)	69.9	1.10	0.21	0.057	559	63.55	3.69	1.40	0.23	
32	Hongcheon River (H-Expt3)	67.0	0.97	0.20	0.053	355	69.07	3.75	1.54	0.32	
33	Daegok Creek (DG-R1)	12.0	0.45	0.17	0.019	880.3	26.67	8.95	1.03	0.32	Seo *et al*. (2016)
34	Daepo Creek (DP-R1)	9.20	0.43	0.65	0.061	308.4	21.40	10.66	1.03	0.53	
35	Gam Creek (GA-R1)	24.80	0.45	0.64	0.066	919.1	55.11	9.70	1.05	0.45	
36	Gam Creek (GA-R2)	23.00	0.25	0.50	0.051	919.1	92.00	9.80	1.05	0.75	
37	Gam Creek (GA-R3)	45.00	0.36	0.56	0.060	824.9	125.00	9.33	1.15	0.96	
38	Gam Creek (GA-R4)	33.50	0.30	0.53	0.055	316.7	111.67	9.64	1.13	0.43	
39	Han Creek (HA-R1)	16.90	0.50	0.23	0.042	1133.9	33.80	5.48	1.28	0.41	
40	Miho Creek (MH-R1)	42.50	1.27	0.27	0.030	221.3	33.47	9.00	1.55	0.69	
41	Miho Creek (MH-R2)	31.00	0.49	0.40	0.048	345.6	63.27	8.33	1.54	0.58	
42	R101	0.10	1.00	0.30	0.020	-	0.10	15.00	1.32	0.70	Baek *et al*. (2005)
43	R151	0.15	1.01	0.20	0.013	-	0.15	15.39	1.32	0.46	
44	R152	0.15	1.01	0.40	0.027	-	0.15	14.82	1.32	0.19	
45	R211	0.21	1.01	0.14	0.010	-	0.21	14.30	1.32	0.19	
46	R212	0.21	1.01	0.29	0.019	-	0.21	15.05	1.32	0.17	
47	R213	0.21	1.01	0.43	0.029	-	0.21	14.79	1.32	0.20	
48	N301	0.30	0.99	0.10	0.006	-	0.30	16.67	1.32	0.17	
49	N302	0.30	0.99	0.20	0.013	-	0.30	15.39	1.32	0.16	
50	N303	0.30	0.99	0.30	0.014	-	0.30	21.43	1.32	0.21	
51	R402	0.40	1.00	0.15	0.011	-	0.40	13.64	1.32	0.16	
52	AMC 315-R1	5.51	0.47	0.57	0.060	-	11.72	9.50	1.50	0.18	Shin *et al*. (2020)
53	AMC 317-R1	5.87	0.52	0.43	0.040	-	11.29	10.75	1.70	0.33	