Korean Journal of Remote Sensing, Vol.32, No.3, 2016, pp.221~234
http://dx.doi.org/10.7780/kjrs.2016.32.3.2

ISSN 1225-6161 (Print)
ISSN 2287-9307 (Online)

Article

GOCI 위성영상과 기계학습을 이용한 한반도 연안 수질평가지수 추정

장은나* · 임정호*† · 하성현* · 이상균* · 박영규**

*울산과학기술원 도시환경공학부, **해양과학기술원 해양순환·기후연구센터

Estimation of Water Quality Index for Coastal Areas in Korea Using GOCI Satellite Data Based on Machine Learning Approaches

Eunna Jang*, Jungho Im*†, Sunghyun Ha*, Sanggyun Lee* and Young-Gyu Park**

*Department of Urban and Environmental Engineering, Ulsan National Institute of Science and Technology

**Korea Institute of Ocean Science and Technology

Abstract : In Korea, most industrial parks and major cities are located in coastal areas, which results in serious environmental problems in both coastal land and ocean. In order to effectively manage such problems especially in coastal ocean, water quality should be monitored. As there are many factors that influence water quality, the Korean Government proposed an integrated Water Quality Index (WQI) based on *in situ* measurements of ocean parameters(bottom dissolved oxygen, chlorophyll-a concentration, secchi disk depth, dissolved inorganic nitrogen, and dissolved inorganic phosphorus) by ocean division identified based on their ecological characteristics. Field-measured WQI, however, does not provide spatial continuity over vast areas. Satellite remote sensing can be an alternative for identifying WQI for surface water. In this study, two schemes were examined to estimate coastal WQI around Korea peninsula using *in situ* measurements data and Geostationary Ocean Color Imager (GOCI) satellite imagery from 2011 to 2013 based on machine learning approaches. Scheme 1 calculates WQI using estimated water quality-related factors using GOCI reflectance data, and scheme 2 estimates WQI using GOCI band reflectance data and basic products(chlorophyll-a, suspended sediment, colored dissolved organic matter). Three machine learning approaches including Random Forest (RF), Support Vector Regression (SVR), and a modified regression tree(Cubist) were used. Results show that estimation of secchi disk depth produced the highest accuracy among the ocean parameters, and RF performed best regardless of water quality-related factors. However, the accuracy of WQI from scheme 1 was lower than that from scheme 2 due to the estimation errors inherent from water quality-related factors and the uncertainty of bottom dissolved oxygen. In overall, scheme 2 appears more appropriate for estimating WQI for surface water in coastal areas and chlorophyll-a concentration was identified the most contributing factor to the estimation of WQI.

Key Words : Water Quality Index, GOCI, machine learning

Received May 11, 2016; Revised May 18, 2016; Accepted May 21, 2016.

† Corresponding Author: Jungho Im (ersgis@unist.ac.kr)

요약 : 우리나라는 대규모 산업단지와 대도시들이 연안에 집중되면서 연안의 오염이 날로 심각해지고 있다. 이러한 연안 오염을 모니터링하기 위해서 위성 영상을 이용한 연안 수질평가지수 모니터링 연구가 수행될 필요가 있다. 수질평가지수란 저층 산소포화도, 엽록소 농도, 투명도, 용존무기질소 및 용존무기인 농도를 수질평가 항목으로 구성하여 해양환경관리법에 따른 해양환경기준을 통해 해역별로 기준을 설정하여 산출하는 지수이다. 이 연구는 한반도 주변의 연안지역을 대상으로 2011년부터 2013년까지의 현장관측 자료 및 Geostationary Ocean Color Imager (GOCI) 위성 영상을 이용하여 연안 표층 해수에 대한 기계학습 기반의 두 가지 수질평가지수 추정 기법을 개발하였다. 첫 번째 방법으로는 GOCI 반사도를 이용하여 추정된 수질평가 항목들로 수질평가지수를 계산하였고, 두 번째 방법은 GOCI 반사도 및 산출물(엽록소 농도, 총 부유물질, 용존유기물)을 이용하여 수질평가지수를 추정하였다. 기계학습으로는 Random Forest(RF), Support Vector Regression (SVR), Cubist를 사용하였다. 수질평가 항목 추정에서 투명도의 정확도가 가장 높게 나타났으며, 모든 수질평가 항목 추정에서 세 가지 기계학습 중 RF의 정확도가 가장 높았다. 하지만 추정된 수질평가 항목들로 계산한 수질평가지수는 추정된 수질평가 항목들의 오차와 저층 산소포화도의 불확실성으로 인해 정확도가 높지는 않았다. 반면 GOCI 반사도와 산출물을 이용하여 추정한 수질평가지수는 현장 관측 기반 수질평가지수와 비교했을 때 첫 번째 방법보다 정확도가 높게 나타났다. 또한 엽록소 농도가 수질평가지수 추정에 가장 중요한 변수로 나타났다.

1. 서 론

우리나라 연안은 인간 생활 활동의 중심지로 생활, 해운, 여가 등 많은 역할을 한다. 갯벌을 중심으로 풍부한 생태계 자원을 보유하고 있으며 대규모 산업단지들과 항만 등의 경제적、상업적 의미가 큰 시설들이 입지하고 있다. 특히 우리나라는 급속한 경제성장 이후 대도시와 대규모 산업단지들이 해안에 조성되어 있고 간척사업과 같은 각종 개발로 인해 대량의 산업폐수와 생활하수 등의 오염물질들이 하천과 강을 통하여 바다로 유입되어 수질오염이 심각하다. 연안 개발 이외에도 적조발생, 기름유출, 태풍 등과 같은 해양재해가 발생하고 있다(Park *et al*., 2013). 일부 지역의 경우 겨울철을 제외한 대부분의 시기에 적조현상이 나타나 수산 피해가 증가하고 있다. 기름 유출 사고는 장기간에 걸쳐 해양 환경에 여향을 미치기 때문에 꾸준한 모니터링이 필요하다. 우리나라는 삼면이 바다로 둘러 쌓여있고 산업이 발달하여 바다의 활용도가 높아지는 만큼 바다로 오염물질 유입이 심각해지고 있으므로, 지속적인 바다 환경 보호와 연안 환경 관련 대책을 마련하기 위해서 연안 수질 모니터링이 필수적이다.

정부는 전국 주요 4대강(한강, 낙동강, 금강, 영산강) 유역에 수질자동측정망을 설치하고 운영함으로써 연안으로 유입되는 물의 수질을 관리하고 있다. 자치단체별로 목표 수질을 설정하고 이를 유지할 수 있도록 오염물질의 배출 총량을 관리하는 수질오염총량관리제를 실시하고 있다. 국토해양부에서는 해양환경관리법에 따라 오염상태가 심각한 5개 연안해역(시화호-인천연안, 광양만, 마산만, 부산연안, 울산연안)을 '특별관리해역'으로 지정하고 오염총량관리제가 적용되고 있다. 우리나라는 연안 환경의 효율적인 관리를 위해 해양환경관리법에 따른 해양환경기준을 통해 수질평가지수(Water Quality Index, WQI)를 도입하여 해역별로 기준을 설정하여 산출하고 있다(Ministry of Land, Transport and Maritime Affairs, 2011). 저층산소포화도(Dissolved Oxygen, DO), 엽록소 농도(Chlorophyll-a, Chla), 투명도(Secchi Depth, SD), 용존무기질소 농도(Dissolved Inorganic Nitrogen, DIN), 용존무기인 농도(Dissolved Inorganic Phosphorus, DIP)가 수질평가 항목으로 구성되어 있다. 각각의 항목들은 Table 1을 이용하여 Table 2의 값을 기준으로 점수가 매겨지며 식(1)을 통해 수질평가지수가 계산된다. 우리나라 생태구역은 Fig. 1와 같이 나뉜다.

$$WQI = 10 \times DO + 6 \times (Chla + SD) / 2 + 4 \times (DIN + DIP) / 2 \quad (1)$$

하지만 수질평가지수를 이용한 연안 수질 평가는 직접 채수를 하여 추가적인 실험을 통해 수질평가 항목들의 양을 계산해야 할 뿐만 아니라, 날씨의 영향과 한정적인 조사지점으로 인해 시공간적 분포를 보기 어렵다.

Table 1. The scoring scheme of each parameter used to calculate WQI

Score	WQI parameter	
	Chla(μg/L), DIN(μg/L), DIP(μg/L)	DO(%), SD(m)
1	≤ reference value(RV)	≥ RV
2	< RV + 0.10 X RV	> RV × 0.10 X RV
3	< RV + 0.25 X RV	> RV × 0.25 X RV
4	< RV + 0.50 X RV	> RV × 0.50 X RV
5	≥ RV + 0.50 X RV	≤ RV × 0.50 X RV

* Reference value of each parameter by division(Table 2)

Table 2. Reference value of each parameter by ocean division

Division	Chla(μg/L)	DO(%)	DIN(μg/L)	DIP(μg/L)	SD(m)
East Sea	2.1	90	139	19	8.6
Straits of Korea	6.3		221	34	2.6
Southwest Sea	3.7		231	25	0.6
Middle West Sea	2.2		427	31	0.9
Jeju	1.6		164	15	8.0

하지만 위성영상을 이용한 수질 모니터링은 광범위한 지역을 주기적으로 관측할 수 있다는 점에서 효율적이다. 대부분의 국내외 위성기반 수질 모니터링 연구에서는 용존유기물, 엽록소 농도, 부유물질, 녹조 등 수질과 연관된 항목들을 추정하였으나(Harvey *et al.*, 2015; Hunter *et al.*, 2010; Kim *et al.*, 2012; Kim *et al.*, 2014; Lee and Lee, 2012; Min *et al.*, 2012), 수질평가지수를 위성으로 모니터링한 사례는 없었다. 본 연구에서는 2011년부터 2013년까지 관측된 현장자료와 Geostationary Ocean Color Imager (GOCI) 위성을 이용하여 기계학습 기반의 두 가지 수질평가지수 추정 기법을 개발하였다. 첫 번째 방법으로는, GOCI 반사도 자료로 수질평가지수 항목들(엽록소, 용존무기질소, 용존무기인, 투명도)을 추정한 뒤 추정된 수질평가지수 항목을 통해 수질평가지수를 계산하였다. 두 번째 방법으로는 GOCI 반사도로부터 직접적으로 수질평가지수를 추정하였고, GOCI 산출물(용존유기물, 엽록소, 총 부유물질)로부터도 수질평가지수를 추정하였다.

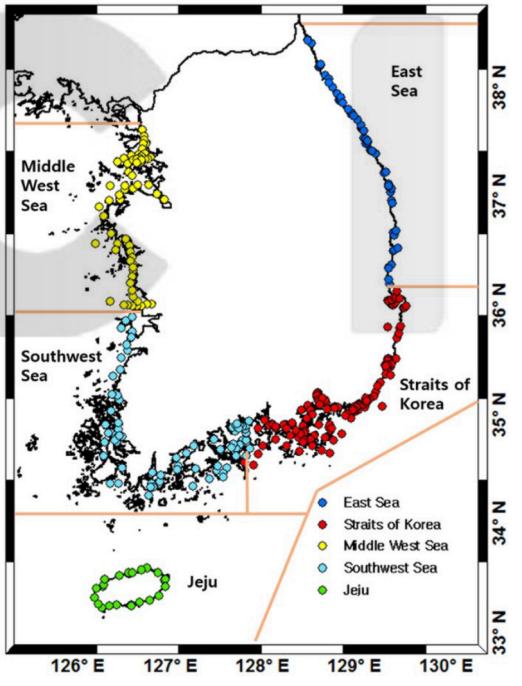

Fig. 1. Ocean divisions and the location of *in situ* measurements

2. 연구지역 및 연구자료

1) 연구지역 및 현장자료

연구지역은 한반도 주변의 연안 지역이다(Fig. 1). 국가해양환경정보통합시스템(Marine Environment Information System, MEIS) 해양환경측정망에서 2011년부터 2013년까지 2월, 5월, 8월, 11월동안 우리나라 생태구역별 5개 연안(대한해협, 동해, 서해중부, 서해남부, 제주해역)을 대상으로 현장 자료를 사용하였다(Table 3). Fig. 1은 우리나라 5개 권역과 각각의 권역에서 현장

Table 3. *in situ* measurements and GOCI satellite data used in this study

	Data	Dates
in situ measurements	Chla(μg/L), SD(m), DIN(μg/L), DIP(μg/L), bottom DO(%)	2011-2013 (Feb, May, Aug, Nov)
GOCI data	Reflectance data(Band 1-6) Band ratio(Band1/2, Band1/3, Band 1/4, Band 1/5, Band 1/6, Band 2/3, Band 2/4, Band 2/5, Band 2/6, Band 3/4, Band 3/5, Band 3/6, Band 4/5, Band 4/6, Band 5/6) Chla, SS, Colored dissolved organic matter(CDOM)	

자료 관측 정점 위치를 나타낸다. 각 해역별 현장 자료 관측 정점 개수는 대한해협 137개, 동해 63개, 서해중부 63개, 서해남부 79개, 제주해역 30개로 총 372개이다. 현장 관측 엽록소 농도, 투명도, 용존무기질소, 용존무기인 및 저층 산소포화도를 이용하여 권역별 기준값(Table 1과 2)에 의해 각 지점별 현장 관측 기반 수질평가지수를 기준자료로써 계산하였다(식(1)).

2) GOCI 위성 자료

GOCI는 2010년 6월에 발사되었고 대한민국 천리안 위성에 탑재된 해양 관측위성이다. 세계최초 정지궤도 해양관측위성으로, 500 m의 공간해상도로 2,500 km × 2,500 km 영역의 한반도를 포함한 동중국해와 일본을 관측하고 있다. GOCI는 8개의 밴드(412, 443, 490, 555, 660, 680, 745, 865 nm)(Table 4)로 오전 9시부터 오후 4시까지 매시간 촬영하여 하루에 8개의 이미지를 제공한다. 이러한 높은 시간해상도는 해양을 주기적으로 관측할 수 있다는 점에서 매우 효율적이다. GOCI는 한반도 주변 해양생태계, 해양환경 및 기후변화를 모니터링하며, 반사도, 용존유기물, 엽록소, 총 부유물질, 적조지수 등 다양한 산출물들을 제공하고 있다. 한국해양위성센터(http://kosc.kiost.ac.kr/)로부터 현장 관측 자료와 상응하는 날짜와 시간의 GOCI LB1 영상을 다운받은 뒤 천리안 해양관측위성 자료처리시스템(GOCI Data Processing System, GDPS)를 통해 현장 관측 자료와 같은 위치의 반사도, 엽록소, 총 부유물질, 용존유기물을 산출하였다.

Table 4. Specification of the GOCI spectral bands

Band	Band center(nm)	Bandwidth(nm)
Band 1	412 nm	20 nm
Band 2	443 nm	20 nm
Band 3	490 nm	20 nm
Band 4	555 nm	20 nm
Band 5	660 nm	20 nm
Band 6	680 nm	10 nm
Band 7	745 nm	20 nm
Band 8	865 nm	40 nm

3. 연구방법

1) 연구방법

이 연구에서는 기계학습을 이용하여 GOCI 위성 자료로부터 수질평가 항목과 수질평가지수를 추정하였다. 위성자료를 이용하기 때문에 수질평가 항목과 수질평가지수는 표층해수에만 국한된다. 기계학습은 회귀분석이나 간단한 경험적 관계식보다 해양의 다양한 비선형적인 특성을 반영하기에 적합하다. 이 연구에서는 Random Forest (RF), Support Vector Regression (SVR), Cubist 3가지 기계학습을 사용하였다. 각 기계학습 기법에 대한 상세 내용은 다음 절에서 설명하였다. 본 연구에서는 두 가지 위성기반 수질평가지수 추정 방법을 제시하였다. 첫 번째 수질평가 항목 추정 방법으로는, 6개의 GOCI 반사도(중심파장 412, 443, 490, 555, 660, 680 nm; Table 4)와 15개의 각 반사도비(Table 3)를 입력변수로 사용하여 기계학습을 통해 수질평가 항목인 엽록소 농도, 용존무기질소, 용존무기인, 투명도를 추정하였다. 해수 표층 정보를 제공하는 위성 자료로는 저층 산소포화도의 추정이 불가능하기 때문에, 저층 산소포화도의 경우 각 권역별 현장 관측 저층 산소포화도 점수를 평균한 값으로 사용하였다. 각 권역별 평균 저층 산소포화도 점수는 서해중부와 제주해역은 2점, 동해, 대한해협과 서해남부 3점이다. 추정된 수질평가 항목으로 권역별 기준에 따라 점수를 매긴 후 계산한 수질평가지수와 현장 관측 자료를 기반으로 계산한 수질평가지수를 비교하였다. 두 번째 수질평가지수 추정 방법으로는 첫 번째 방법에서 사용한 반사도와 반사도비를 사용하여

기계학습을 통해 추정한 수질평가지수와 GOCI 산출물인 용존유기물, 엽록소 농도, 총 부유물질을 사용하여 기계학습을 통해 추정한 수질평가지수를 현장 관측 기반 수질평가지수와 비교하였다.

2) 기계학습

기계학습은 컴퓨터가 훈련 자료를 통해 스스로 학습하여 알고리즘을 만들어 낸다. 본 연구에서는 세 가지 기계학습 기법인 Random Forest (RF), Support Vector Regression (SVR), Cubist를 사용하였다. RF는 Classification And Regression Trees (CART)를 기반으로 한 결정나무(decision tree)의 확장 개념으로, 수많은 독립적인 결정나무를 구축한다. RF는 전체 훈련 자료 중 무작위로 선택된 일부만 사용하여 모델을 구축하고, 나무의 각 노드에서 무작위로 입력변수를 선택하여 사용함으로써 독립적인 결정나무를 만들어 내고 이를 통해 결정나무의 단점으로 알려진 과적합(overfit)을 방지한다(Breiman, 2001). 각 결정나무 구성에 사용되지 않은 훈련 자료(out-of-bag; OOB)는 내부 모델을 평가하고 각 입력변수의 상대적 중요도를 평가하는데 사용된다. 이렇게 구축된 나무들은(weighted) 다수결(majority voting) 또는 평균(averaging) 기법을 통해 취합되어, 분류(classification)의 경우 가장 많이 나온 값을, 회귀(regression)의 경우 평균값을 최종값으로 출력한다. RF가 제공하는 상대적인 변수중요도는 각 변수의 값을 무작위로 넣어서 모델을 생성했을 때 얼만큼 정확도가 감소하는지를 나타내는 것으로, 정확도가 많이 감소할수록 더 중요한 변수임을 알 수 있다. 이 연구에서는 R 통계 소프트웨어(version 3.1.3)의 'random forest' 패키지를 이용하여 RF를 수행하였다.

SVR은 입력자료를 가장 적합한 클래스로 나누는 최적의 초평면(hyperplane)을 찾는 것을 바탕으로 한다(Mountrakis *et al.*, 2011). 클래스들을 분리하는 수많은 후보 초평면들 가운데 각각의 초평면으로부터 각 점들에 이르는 거리의 최소값이 최대가 되는 최대 마진 초평면을 찾는다. SVR은 일반적으로 훈련자료를 원래의 차원에서 고차원으로 변환시켜 초평면을 보다 수월하게 찾고자 한다. SVR은 훈련자료를 고차원으로 변환 시킬 때 쓰이는 커널함수를 선택하는 것이 가장 중요하다. 커널함수는 linear, polynomial, Gaussian, sigmoid, spectral angle 등이 있으며(Kim *et al.*, 2014), 이 연구에서는 Radial Basis Function (RBF)를 사용하였다. SVR은 소량의 데이터로도 높은 정확도를 낼 수 있지만, 알맞은 커널함수 선택이 중요하고 모수화하기 힘든 단점이 있다.

Cubist는 RuleQuest Research 에 의해 만들어진 상용 소프트웨어로, 수정된 회귀나무(modified regression tree)를 기반으로 한다(RuleQuest Research, 2012). Cubist는 훈련 자료로부터 규칙 기반 다중회귀모델들을 만들어, 각 규칙에 따라 다중회귀모델이 적용된다. 여러 규칙이 동시에 만족되는 경우에는 규칙에 만족하는 모든 다중회귀모델 결과의 평균값이 최종값으로 출력된다. Cubist는 규칙에 따른 다중회귀모델을 'if *condition* then *linear formula*'의 형태로 제공하기 때문에 다른 기계학습보다 모델을 적용하고 분석하기 쉽다. Cubist의 attribute usage는 각 노드에서 분류될 때 각각의 변수가 규칙과 다중회귀모델을 만드는데 얼마나 자주 사용되는지를 보여주는 것으로, 이는 RF와 비슷하게 변수 중요도의 역할을 한다.

훈련자료의 수가 적은 관계로, 모든 훈련자료들을 이용하여 각 기계학습의 알고리즘을 만든 뒤 교차검증(10-fold cross validation)을 통해 검증하였다. 검증 결과는 Root Mean Squared Error (RMSE)와 relative RMSE(%)를 이용하여 분석하였다.

4. 연구결과 및 토의

1) GOCI 반사도를 이용한 수질평가 항목 추정

Fig. 2는 GOCI 반사도와 세가지 기계학습을 이용하여 수질평가 항목들을 추정한 결과이며, Fig. 3은 RF에서 상대적으로 변수 중요도가 가장 높은 7개 변수들을 나타낸 것이다. 모든 수질평가 항목에서 RF의 보정 정확도는 가장 좋았지만 검증 결과는 Cubist가 가장 좋았다. 규칙을 찾기에 훈련자료의 수가 충분치 않아 훈련자료의 대표성이 떨어지기 때문에 보정 정확도는 높으나 검증 결과는 좋지 않게 나타났다. 반면 Cubist는 규칙을 찾기에 훈련자료가 적어 적은 수의 조건과 규칙으로 모델을 구성하기 때문에 보정 정확도는 낮을 수 있으나 포괄적인 규칙으로 인해 검증 정확도가 RF만큼 낮게 나

오지는 않는다. 세 가지 기계학습 모두 모든 수질평가 항목을 과소추정하였다. 추정된 수질평가 항목의 범위는 현장 관측 자료의 범위보다 줄어들었는데 이는 일반적인 경험 모델링의 한계로 볼 수 있다. 수질평가 항목

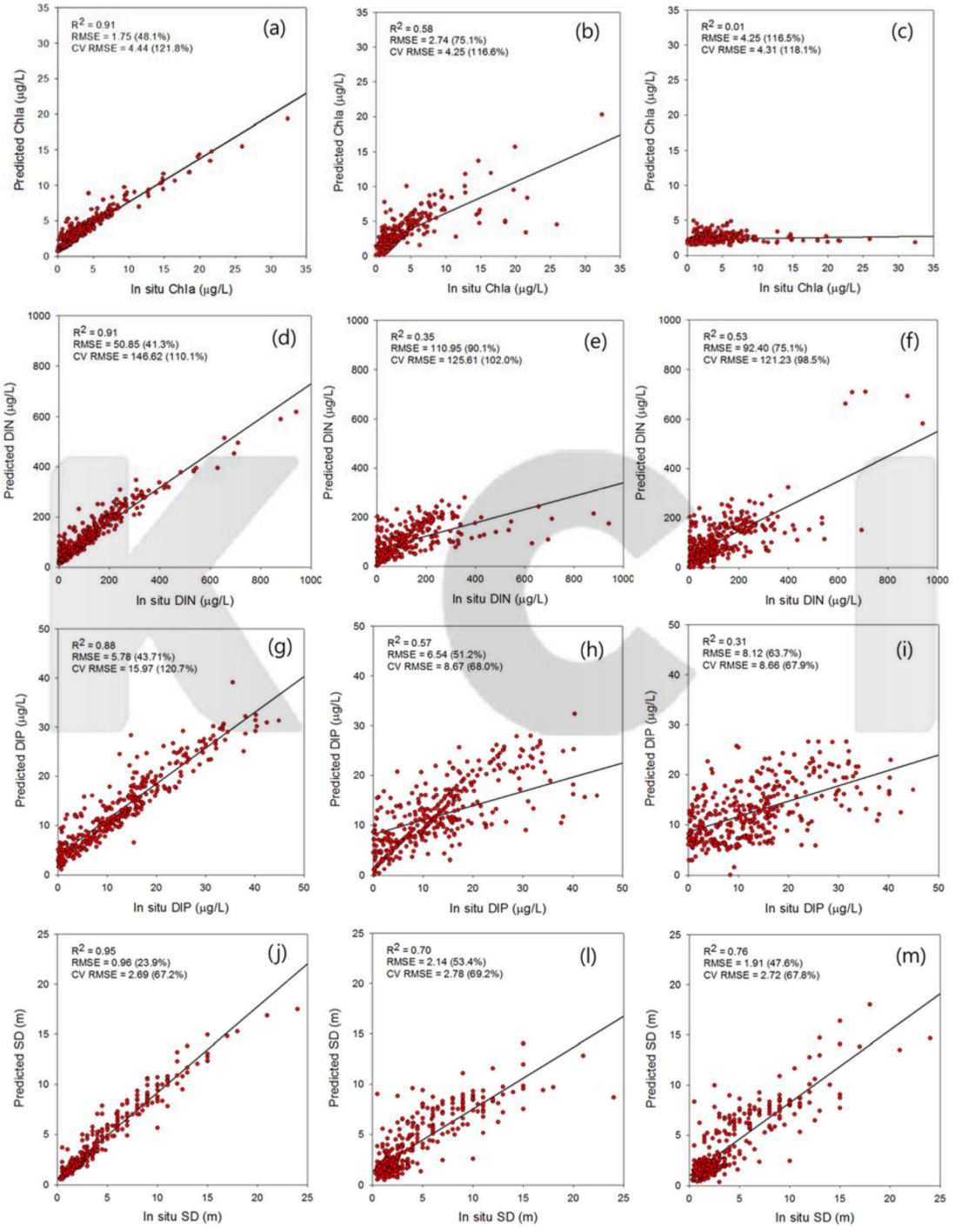

Fig. 2. Scatterplots between the *in situ* measurements and predicted WQI parameter values by model :(a) Chla using RF,(b) Chla using SVR,(c) Chla using Cubist,(d) DIN using RF,(e) DIN using SVR,(f) DIN using Cubist,(g) DIP using RF,(h) DIP using SVR,(i) DIP using Cubist,(j) SD using RF,(l) SD using SVR, and(m) SD using Cubist

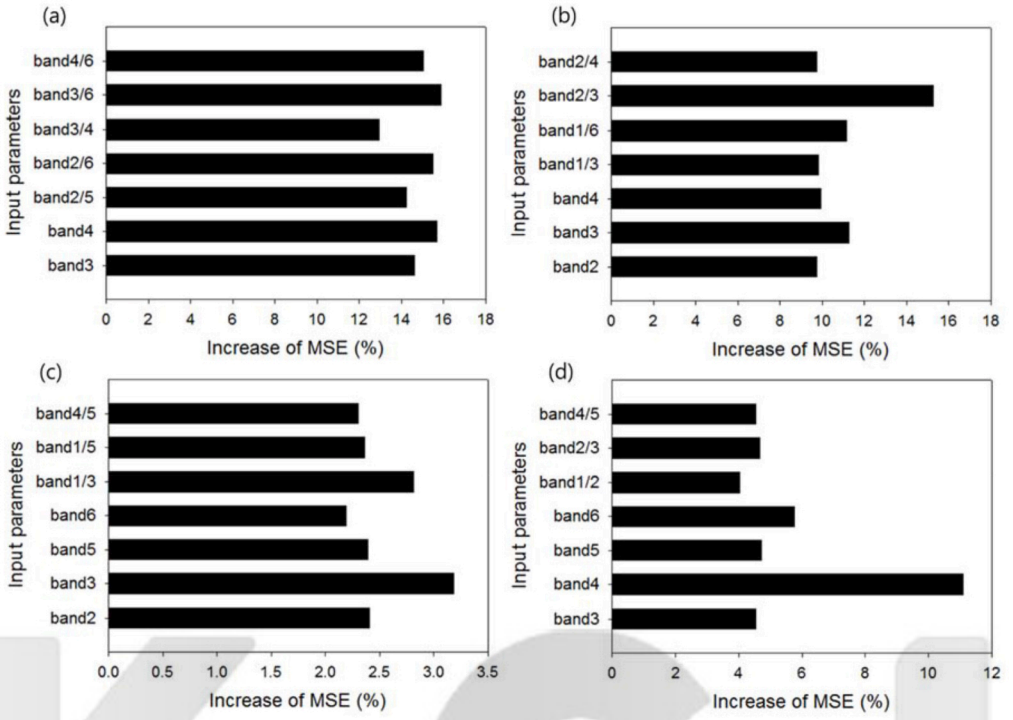

Fig. 3. Relative parameter importance calculated from RF(a) for Chla,(b) for DIN,(c) for DIP, and(d) for SD. The higher the increase of mean squared error(MSE), the more contributing to the model the parameter is.

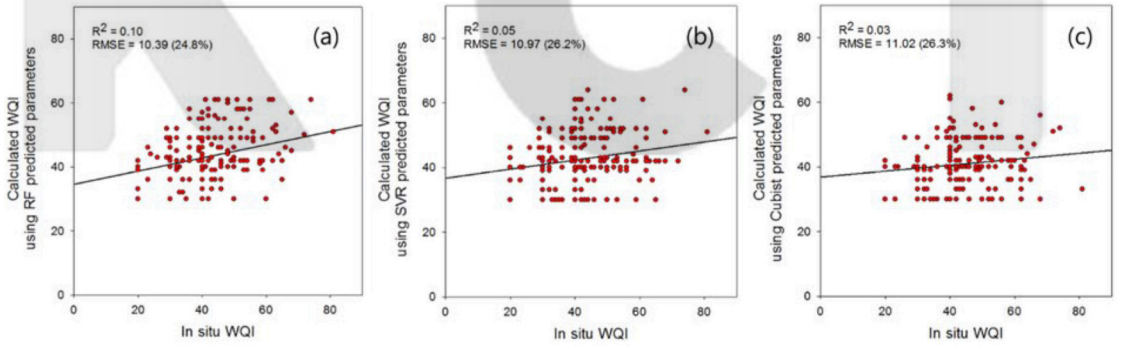

Fig. 4. Scatterplots between the WQI values calculated using *in situ* measurements and the WQI values calculated using the parameters predicted by three machine learning approaches.

중에서는 엽록소 농도의 추정 정확도가 가장 낮고 투명도의 정확도가 가장 높게 나타났다. 밴드 3과 4(중심 파장 490 nm, 555 nm; 녹색 영역)가 대부분의 수질평가 항목 추정에서 중요한 변수로 나타났으며, 이는 GDPS에서 수질과 관련된 GOCI 산출물인 엽록소 농도와 총 부유물질을 산출하는데 주로 쓰이는 반사도이다. 엽록소의 경우 Kim *et al.*(2014)에서 GOCI를 이용하여 엽록소를 추정하는 과정에서 나온 변수 중요도와 비슷한 결과를 보였다. 또한 엽록소의 주요 변수로 나온 반사도와 밴드 비들은 SeaWiFS, MODIS-Aqua 및 GlobColour 등에서 엽록소 농도를 산출하는데 사용된다(Johnson *et al.*, 2013). 해수면의 투명도는 부유물질과 엽록소의 농도에 따라 결정이 된다. 투명도의 변수 중요도에서는 GDPS에서 GOCI 총 부유물질 산출물을 계산하는데 쓰이는 밴드 4(중심파장 555 nm)가 가장 중요한 변수로 나타났으며 엽록소 농도와 관계가 있는 밴드 3번과 5번(중심파장 490 nm, 660 nm)도 주요 변수로 나타났다.

Fig. 4는 GOCI 반사도를 이용하여 추정한 수질평가

항목을 이용하여 계산한 수질평가지수와 현장 관측 기반 수질평가지수를 비교한 결과이다. 계산된 수질평가지수의 정확도는 세가지 기계학습 모두 비슷하게 나타났다. 수질평가 항목들이 현장관측 자료보다 과소추정

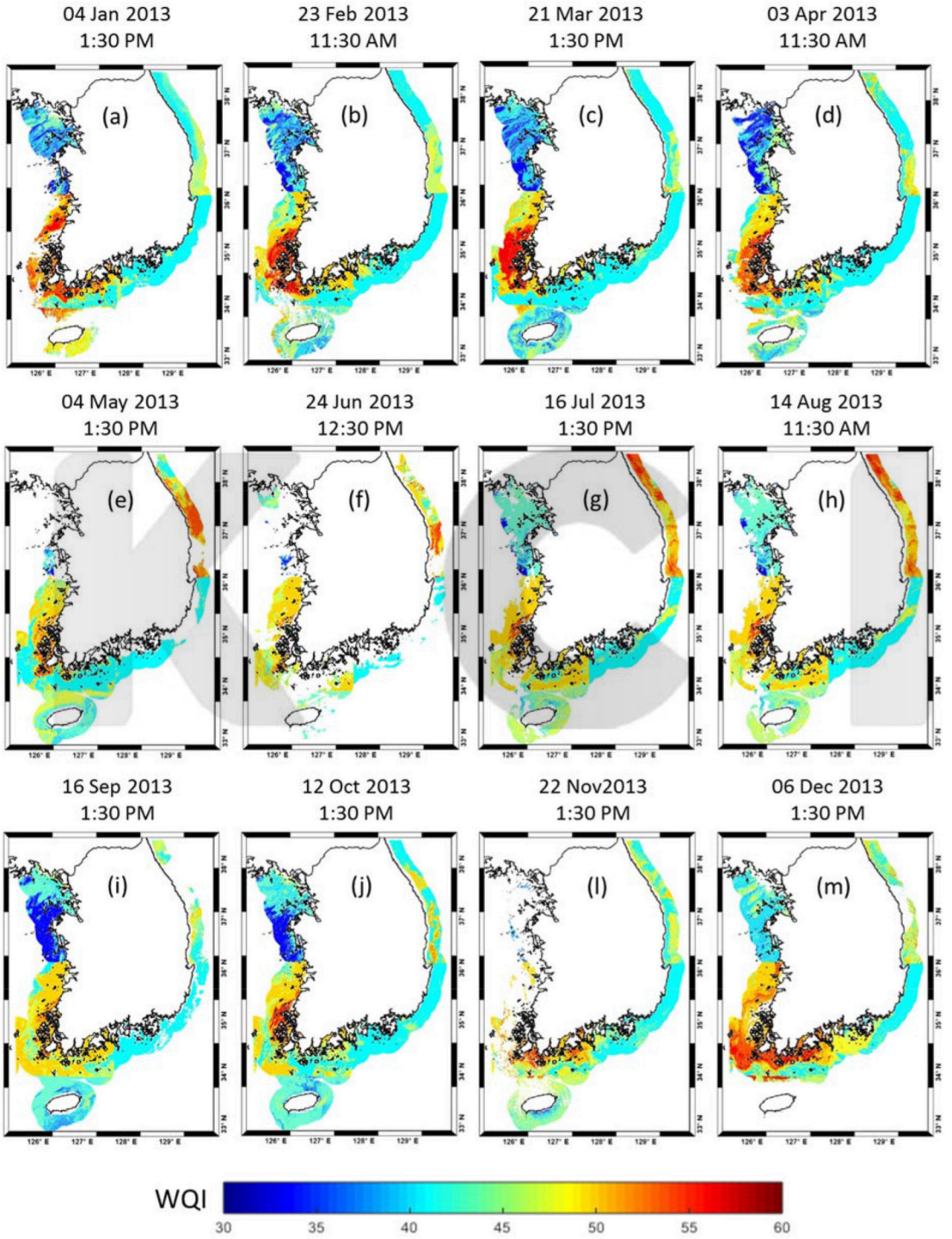

Fig. 5. Spatial distribution of the WQI calculated using the parameters predicted by RF model for the selected date/time of each month in 2013.

되었기 때문에 추정된 수질평가 항목으로 계산한 수질평가지수의 범위는 현장 관측 기반 수질평가지수의 범위보다 작게 나타났다. 수질평가지수를 계산하는 식(식(1))에서 저층 산소포화도의 가중치가 가장 높지만, 저층 산소포화도는 위성 영상을 이용해 추정할 수 없기 때문에 이 방법에서는 해역별 평균값을 사용하였다. 추정된 수질평가 항목들이 가지는 오차에 저층 산소포화도의 불확실성까지 더해져 추정된 수질평가지수의 큰 오차를 유발하였다.

Fig. 5는 수질평가 항목을 추정할 때 정확도가 가장 높았던 RF 모델을 이용하여 추정한 수질평가 항목으로 계산한 수질평가지수의 공간적 분포이다. 월간 분포는 평균이 아닌 매월 가장 구름이 없는 특정 날짜 영상을 이용하였지만 여전히 일부 날짜에서는 구름 때문에 수질평가지수의 공간 분포를 볼 수 없는 지역이 있다. 수질평가지수 분포 지도에서 볼 수 있듯이 권역별로 수질평가 항목 점수를 매기는 기준이 달라 값이 비슷하더라도 권역별로 점수가 달라지기 때문에 권역별 경계선이 뚜렷이 나타난다. 본 연구에서 개발한 수질평가지수 추정 모델은 내륙으로부터 20 km 이내의 샘플들을 이용하여 연안지역에 최적화되어 있어서 외해 지역의 분포는 불확실성이 매우 크다고 볼 수 있기 때문에 Fig. 5와 같이 육지에서 30 km까지 해역의 수질평가지수 분포를 추정하여 나타내었다. 각 월의 한 시점에서의 분포이기 때문에 일반화 하기는 어렵지만 다른 해역보다 서해중부 지역의 수질평가지수가 낮게 나타났으며, 서해남부 지역에서 높게 나타났다. 이는 서해남부 지역이 다른 연안에 비해 섬과 인간활동이 많아 다른 지역보다 오염이 더 심하다.

2) GOCI 반사도와 산출물을 이용한 수질평가지수 추정

Fig. 6는 RF를 통해 6개의 GOCI 반사도와 15개의 반사도 비, 총 21개의 변수를 사용한 경우와 GOCI 산출물인 엽록소 농도, 총 부유물질, 용존유기물을 변수로 사용하여 수질평가지수를 추정한 결과이다. SVR과 Cubist 모델은 두 가지 경우 모두에서 수질평가지수를 평균값과 가까운 값(40~42)으로 추정하였다. 이는 입력 변수들과 수질평가지수 사이에서 규칙을 찾기에 샘플 개수가 적어 평균값을 기준으로 약간의 오차를 내는 모델보다 더 나은 모델을 만들지 못한 것으로 보인다. RF로 추정한 수질평가지수는 기계학습으로 추정한 수질평가 항목으로 계산한 수질평가지수(Fig. 4)보다 정확도가 훨씬 높게 나타났다. 직접적으로 수질평가지수를 추정하는 방법이 첫 번째 방법의 중간과정에서 생기는 오차를 줄였다. 하지만 여전히 경험적 모델링의 한계로 추정된 수질평가지수의 범위가 현장 관측 기반 수질평가지수의 범위보다 적게 나타났다. 수질평가지수를 추정하였을 때, 반사도를 이용한 경우는 산출물을 이용했을 때보다 더 높은 정확도를 보이는데, 이는 입력변수 개수의 차이와 산출물의 정확도로 인해 발생하였다. 반사도를 이용한 경우 밴드 6개와 각각의 밴드비 15개, 총 21개의 입력변수를 사용하였고, 산출물을 이용한 경우는 3개의 입력변수를 사용하였다. 이로부터 수질평가지수와 관련된 입력변수가 많을수록 더 정교한 모델이 만들어진다고 볼 수 있다. 또한 GOCI 산출물이 실제 현장관측 자료보다 과소추정되는 것을 보이는데, 이 또한 수질평가지수 추정 정확도를 낮추는 요인 중 하나이다.

Fig. 7은 RF로 추정한 상대적인 변수 중요도이다. 반사도를 이용한 방법은 입력변수가 21개이기 때문에 정확도가 높은 7개를 뽑았다(Fig. 7(a)). 산출물을 이용한 수질평가지수에서 엽록소 농도가 가장 중요한 변수로 나타났으며, 수질평가지수를 추정하는데 쓰이는 주요 반사도 또한 엽록소 추정에 쓰이는 주요 반사도(Fig. 3(a))와 비슷하였다. 이는 엽록소 농도가 수질평가지수를 계산하는데 쓰이는 수질평가 항목 중 하나이며 수질과 밀접한 연관이 있는 것으로 보인다. 실제 많은 연구에서 수질 모니터링을 위해 엽록소 농도를 주요 변수로 관측해 왔다(Harvey *et al.*, 2015; Kim *et al.*, 2014; Le *et al.*, 2013; Lumb *et al.*, 2011; Novoa *et al.*, 2012).

Fig. 8은 RF와 GOCI 반사도를 이용한, Fig. 9는 RF와 GOCI 산출물(용존유기물, 엽록소 농도, 총 부유물질)을 이용하여 추정한 수질평가지수의 월별 시공간적 분포이다. 첫 번째 방법과는 달리 권역별 기준이 없기 때문에 권역별 경계면이 나타나지 않았다. RF로 추정한 수질평가지수는 첫 번째 방법으로 추정한 수질평가지수(Fig. 5)보다 서해남부와 동해는 낮게, 서해중부, 대한해협 및 제주해역은 높게 나타났다. GOCI 산출물을 이용하여 추정한 수질평가지수(Fig. 9)의 공간적 분포는 세 가지 GOCI 산출물의 공간적 분포와 유사하게 나타

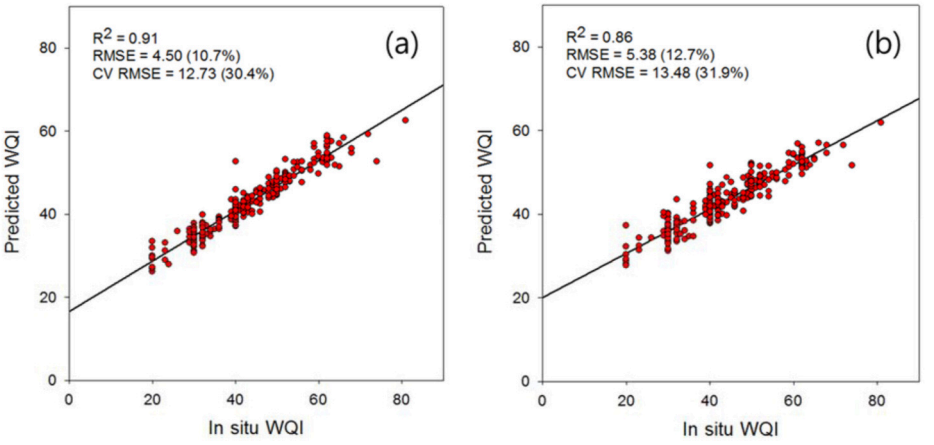

Fig. 6. Scatterplots between the WQI calculated using *in situ* measurements and the WQI predicted using GOCI-derived(a) band reflectance and band ratio data and(b) Chla, CDOM, and SS by RF.

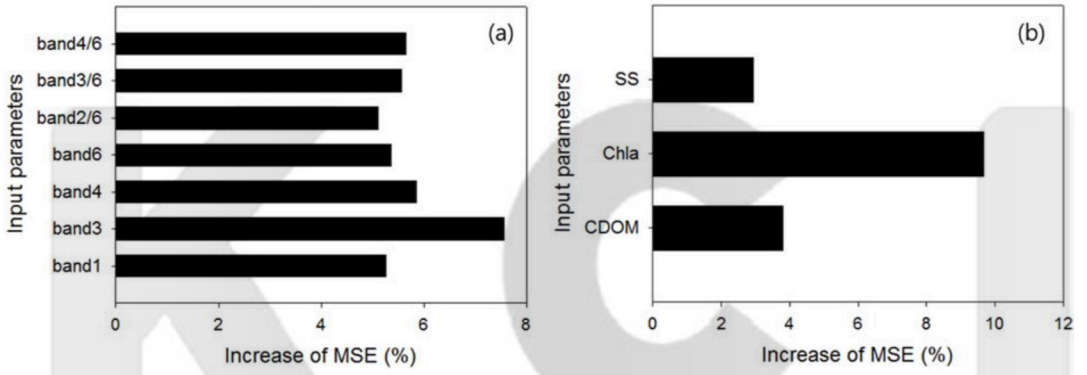

Fig. 7. Relative parameter importance calculated from RF using GOCI-derived(a) band reflectance data and(b) basic parameters(SS, Chla, and CDOM). The higher the increase of mean squared error(MSE), the more contributing to the model the parameter is.

났다. GOCI 반사도를 이용해 추정한 수질평가지수의 공간적 분포(Fig. 8)가 현장 관측 자료로 계산된 수질평가지수의 공간적 분포와 가장 유사하였다. GOCI 산출물을 이용해 추정한 수질평가지수의 5월 4일, 6월 24일, 9월 16일 공간적 분포(Fig. 9(e), (f), (i))는 첫 번째 방법의 결과(Fig. 5)와 GOCI 반사도를 이용해 추정한 수질평가지수(Fig. 8)와 다른 결과로 동해안 북쪽지역이 높은 수질평가지수를 보인다. 입력 변수로 들어간 용존유기물, 엽록소 농도 및 부유물질의 공간적인 분포를 관찰한 결과 수질평가지수가 높게 나타난 동해안 지역의 엽록소 농도와 부유물질이 다른 지역보다 낮게 나타났다. 이는 훈련자료와 입력변수가 이 지역의 특성을 나타낼 수 있는 알고리즘을 만들기에 부족하여 RF에서 적합한 알고리즘을 만들어내지 못한 것으로 보인다.

수질평가지수는 주어진 하나의 식(식(1))을 통해 한반도 연안의 수질을 평가한다. 권역별 수질평가 항목들의 기준을 다르게 두지만 인간활동의 정도와 유입물질의 양 등 지역적 특수성이 하나의 식을 통해 표현하기는 힘들다. 단일 수질평가지수를 사용함으로써 수질평가지수 계산과 관리에는 편리하지만 각 권역별 지역적 특성을 모두 담아내지 못한다는 한계점이 있다.

본 연구의 가장 큰 한계점은 정교한 모델을 만들기 위한 훈련자료가 부족하며 대부분 육지에 매우 근접해 있다는 점이다. 대부분의 현장 관측 자료가 연안(Fig. 1)에 위치하여 GOCI 위성 자료를 사용할 때 육지로 마스킹 되는 지점이 많았고, 구름으로 인해 쓸 수 없는 위성영상도 많았다. 이로 인해 GOCI 반사도와 산출물을 이용하여 수질평가지수를 추정하였을 때 SVR과 Cubist는 모델을 제대로 만들어내지 못하고 단순 평균값으로 결과를 산출하였다. 또한 수질평가지수 항목 중 하나인 저

층 산소포화도는 표층을 관측하는 위성으로 추정하지 못한다는 한계점도 있다. 하지만 기계학습을 이용해 수질평가지수를 추정함으로써 단순한 경험식으로 표현하지 못하는 입력자료와 수질평가지수 사이 다양한 비

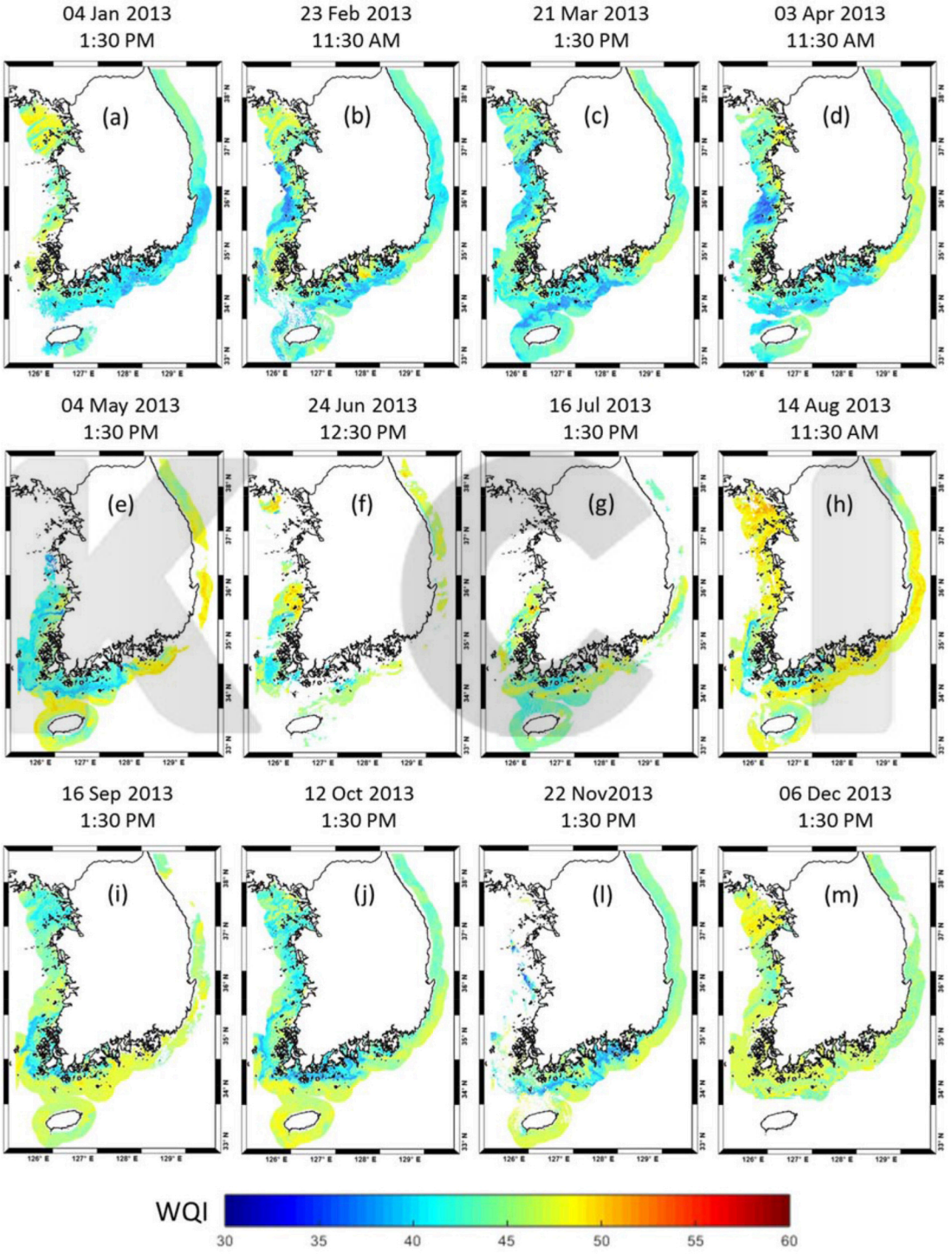

Fig. 8. Spatial distribution of the WQI valued predicted by RF model using GOCI band reflectance for the selected date/time of each month in 2013.

선형적인 특성을 담아냄으로써 적은 훈련자료로 높은 정확도를 나타내었다. 또한 기계학습이 제공하는 변수 중요도를 통해 수질평가지수에 영향을 미치는 변수들도 분석 가능하다. 더 많은 현장관측 자료를 통해 훈련

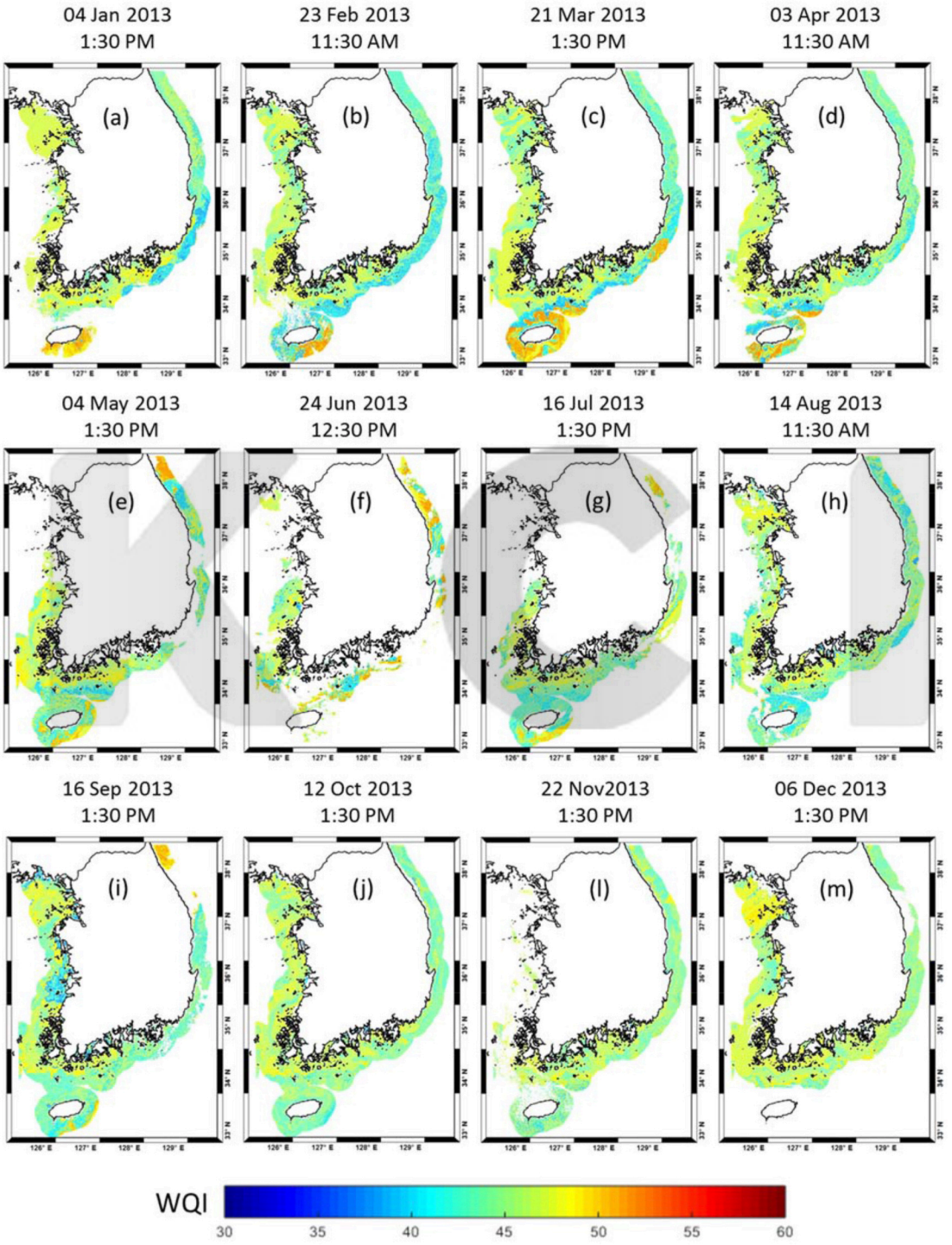

Fig. 9. Spatial distribution of the WQI values predicted by RF model using GOCI products(CDOM, chla, and SS) for the selected date/time of each month in 2013.

자료의 수가 많아지면 더 정교한 수질평가지수 모델을 개발 할 수 있을 것으로 판단된다.

5. 결 론

이 연구는 한반도 주변의 연안지역을 대상으로 2011년부터 2013년까지의 현장관측 자료 및 GOCI 반사도 자료와 산출물을 이용하여 기계학습 기반의 두 가지 수질평가지수 추정 기법을 개발하였다. 세 가지 기계학습 중 RF의 정확도가 가장 좋았으며, SVR과 Cubist는 GOCI 반사도와 산출물을 이용한 수질평가지수 추정 모델을 제대로 산출해내지 못하였다. 수질평가 항목 추정에서는 투명도의 정확도가 가장 높게 나타났으며, 녹색 영역의 반사도(밴드 3과 4; 중심파장 490 nm, 555 nm)가 대부분의 수질평가 항목 추정에서 중요한 변수로 나타났다. 이는 청색 영역의 반사도가 수질 관련 산출물인 엽록소 농도, 총 부유물질 등을 추정하는데 쓰이는 것과 연관이 있는 것으로 보인다. 추정된 수질평가 항목으로 계산한 수질평가지수는 저층 산소포화도의 불확실성과 추정된 수질평가 항목들의 오차로 인해 정확도가 낮았다. 반면 GOCI 반사도와 산출물을 이용하여 RF를 통해 만들어진 수질평가지수 추정 모델은 정확도가 높았다. 첫 번째 방법과 달리 직접적으로 수질평가지수를 추정함으로써 중간 과정에서 생기는 오차들이 없어지면서 정확도가 높아진 것으로 보인다. GOCI 반사도를 이용한 경우가 산출물을 이용한 경우보다 수질평가지수를 더 정확하게 산출하였다. 수질평가지수 추정에서 엽록소 농도와 첫 번째 방법에서 엽록소 농도 추정에 쓰였던 주요 반사도가 중요한 변수로 나타났으며, 이는 엽록소 농도가 수질평가지수와 밀접한 연관이 있음을 보여준다.

두 가지 수질평가지수 추정 모델을 통해 한반도 연안 수질평가지수의 특성을 파악할 수 있었고, 한반도 연안 수질평가지수의 공간적 분포를 분석할 수 있었다. 하지만 제공되는 GOCI 자료에서 많은 연안 현장 관측 지점이 육지로 마스킹되었고 구름 등으로 인해 사용할 수 있는 현장 관측 자료가 적었다. 또한 수질평가 항목 중 하나인 저층 산소포화도는 해수면을 관측하는 위성으로는 추정할 수 없다는 한계점이 있었다. 향후 더 많은 현장 관측 자료를 이용하고, 저층 산소포화도를 대체할 수 있는 다른 표층 수질평가 항목을 추가한다면 더 정확한 수질평가지수 추정 모델을 개발할 수 있을 것으로 기대한다.

사 사

이 연구는 해양수산부(Ministry of Ocean and Fisheries)의 '위성기반 한반도 주변해역 해양 탄소 추정모델 개발'과 GOCI 2단계 활용 연구의 지원을 받아 수행되었습니다.

References

Breiman, L., 2001. Random forests. *Machine learning*, 45(1): 5-32.

Harvey, E.T., S. Kratzer, and P. Philipson, 2015. Satellite-based water quality monitoring for improved spatial and temporal retrieval of chlorophyll-a in coastal waters. *Remote Sensing of Environment*, 158: 417-430.

Hunter, P.D., A.N. Tyler, L. Carvalho, G.A. Codd, and S.C. Maberly, 2010. Hyperspectral remote sensing of cyanobacterial pigments as indicators for cell populations and toxins in eutrophic lakes. *Remote Sensing of Environment*, 114(11): 2705-2718.

Johnson, R., P.G. Strutton, S.W. Wright, A. McMinn, and K.M. Meiners, 2013. Three improved satellite chlorophyll algorithms for the Southern Ocean. *Journal of Geophysical Research: Oceans*, 118(7): 3694-3703.

Kim, Y.H., J. Im, H.K. Ha, J.K. Choi, and S. Ha, 2014. Machine learning approaches to coastal water quality monitoring using GOCI satellite data. *GIScience & Remote Sensing*, 51(2): 158-174.

Kim, Y.J., H.C. Kim, Y.B. Son, M.O. Park, W.C. Shin, S.W. Kang, and T.K. Rho, 2012. Verification of

CDOM algorithms based on ocean color remote sensing data in the East Sea. *Korean Journal of Remote Sensing*, 28(4): 421-434. (in Korean with English abstract)

Le, C., C. Hu, D. English, J. Cannizzaro, and C. Kovach, 2013. Climate-driven chlorophyll-a changes in a turbid estuary: Observations from satellites and implications for management. *Remote Sensing of Environment*, 130: 11-24.

Lee, K.H., and S.H. Lee, 2012. Monitoring of Floating Green Algae Using Ocean Color Satellite Remote Sensing. *Journal of the Korean Association of Geographic Information Studies*, 15(3): 137-147 (in Korean with English abstract).

Lumb, A., T.C. Sharma, and J.F. Bibeault, 2011. A review of genesis and evolution of water quality index(WQI) and some future directions. Water Quality, *Exposure and Health*, 3(1): 11-24.

Min, J.E., J.H. Ryu, S. Lee, and S. Son, 2012. Monitoring of suspended sediment variation using Landsat and MODIS in the Saemangeum coastal area of Korea. *Marine Pollution Bulletin*, 64(2): 382-390.

Ministry of Land, Transport and Maritime Affairs, 2011. *The marine environmental standards on the Marine Environment Management Act*, Ministry of Land, Transport and Maritime Affairs, South Korea.

Mountrakis, G., J. Im, and C. Ogole, 2011. Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(3): 247-259.

Novoa, S., G. Chust, J.M. Froidefond, C. Petus, J. Franco, E. Orive, S. Soane, and A. Borja, 2012. Water quality monitoring in Basque coastal areas using local chlorophyll-a algorithm and MERIS images. *Journal of Applied Remote Sensing*, 6(1): 063519-1.

Park, S., and S.R. Lee, 2013. Marine disasters prediction system model using marine environment monitoring. *The Journal of Korean Institute of Communications and Information Sciences*, 38(3): 263-270.

RuleQuest Research., 2012. *RuleQuest Research data mining tools*. from http://www.rulequest.com/.