랜덤포레스트 알고리즘을 활용한 유역환경 특성과 하천수질의 관계 분석

박유진*, 박세린, 이정아***, 이상우******
*건국대학교 일반대학원 산림조경학과 박사과정, **건국대학교 산림조경학과 박사, ***고려대학교 환경생태공학부 부교수,
******건국대학교 산림조경학과 정교수

1. 서론

하천생태계에서 수질은 하천환경의 건강성과 생물 서식 환경을 결정하는 중요한 요소이며 유역환경 변화에 민감하게 반응하여 유역의 다양한 인자들이 하천에 미치는 영향을 파악하는데 핵심적인 역할을 한다(환경부, 2019; Wang et al., 2021). 하천을 둘러싼 유역환경은 과도한 토지이용, 도시화, 농업활동과 같은 인간활동에 의해 끊임없이 교란되고 있으며 기후변화로 인한 극단적 기상현상의 증가 또한 유역환경의 교란을 가속화하고 있다(Baker, 2006; Park et al., 2020). 이러한 유역환경의 교란은 유역에서 하천으로 유입되는 오염물질 및 오염량의 발생을 증가시키며 하천수질이 저하되는 등의 원인으로 확대된다(Cha et al., 2021; Park et al., 2021). 유역에서 발생하는 오염원의 종류 및 오염물질 이동에 관여하는 지형학적 특성에 따라 교란이 하천수질에 미치는 영향이 달라지므로 유역환경 특성에 따른 적절한 유역관리 방안이 필요할 것으로 보인다(Pratt and Chang, 2012). 이에 본 연구는 랜덤포레스트(random forest) 알고리즘을 활용하여 하천수질 변수에 영향을 미치는 유역환경 변수를 도출하고 하천수질 변수와 유역환경 특성을 나타내는 여러 변수 간의 관계를 분석하고자 한다.

2. 연구방법

2.1 연구대상지 및 변수 선정

연구대상지는 한강 수계와 한강 외 기타수계(안성천, 한강서해, 한강동해 등)가 포함되어 있는 한강 대권역으로 선정하였다. 한강 대권역은 30개의 중권역, 290개의 표준유역으로 구성되어 있으며 25,953km^2의 면적을 차지하고 있다. 하천수질에 영향을 미치는 유역환경 변수는 Table 1과 같이 총 20개의 변수를 선정하였다. 하천수질값은 환경부에서 운영하고 있는 물환경측정망 중 한강대권역에 위치한 수질측정망의 2019년도 BOD, T-N, T-P의 농도의 평균값을 활용하였다. 유역환경 변수는 하천망도, 유역경계, 수치표고자료를 기반으로 ArcMap을 활용하여 수질측정망 지점의 소유역 230개를 선정하고 토지피복자료, 토양환경지도, 인구 통계자료 등을 활용해 각 유역환경에 대한 특성을 나타낼 수 있는 변수값을 도출하였다. 유역환경 변수 중 오염원 변수는 전국오염원조사자료를 활용하였으며 기상요소는 수문가뭄기상 정보시스템의 유역별 강수량, 증발산량을 활용하여 도출하였다.

Table 1. 하천수질 및 유역환경 변수

구분	변수	출처
하천수질 지표	BOD, T-N, T-P	환경부 물환경정보시스템
지형학적 특성	유역면적(km^2), 유역 둘레(km), 평균 경사도(%), 평균 고도(m)	한강 홍수 통제소, Arcmap
오염원	환경기초시설 개수, 가축밀도	환경부 전국오염원조사자료
도시화 정도	인구밀도, 도로밀도	국가공간정보포털
토지이용	도시지역(%), 농업지역(%), 산림지역(%)	환경공간정보서비스
토양 특성	침식 정도(%), 배수 정도(%)	흙토람
기상요소	연강수량 합계(mm), 월 최대강수량 평균(mm), 수온(°C), 증발산량(mm)	환경부 물환경정보시스템 기상청 수문기상 가뭄정보 시스템

2.2 분석방법

하천수질에 영향을 미치는 주요 유역환경 변수를 도출하고 변수들 간의 관계를 파악하기 위하여 랜덤포레스트(random forest) 알고리즘을 활용하였다. 랜덤포레스트 알고리즘은 강력한 머신러닝 기법 중 하나로 여러 개의 의사결정 모델을 조합해 예측값에 대한 오차를 줄이고 모델에 대한 정확도를 개선할 수 있으며 설명변수와 종속변수 간의 비선형성 및 복잡한 관계를 파악할 수 있다(Breiman, 2001; Park, 2021). 랜덤포레스트 알고리즘은 설명변수에 대한 영향을 평가하여 중요도 지수(variable of importance index)와 부분의존도 그래프(partial dependence plots)를 제시하여 설명변수가 종속변수에 미치는 중

요도 및 영향력을 파악할 수 있다. 본 연구에서는 랜덤포레스트 모델을 구축하기 위해 BOD, T-N, T-P의 하천수질값을 종속변수, 유역환경 변수를 설명변수로 선정하였다. 랜덤포레스트 모델 분석은 RStudio의 "randomforest" 패키지를 활용하였으며 랜덤포레스트 내의 의사결정 트리 개수는 500개, 트리마다 사용할 설명변수의 개수는 기본값을 사용하였다.

3. 연구결과 및 고찰

하천수질 변수인 BOD, T-N, T-P에 대한 랜덤포레스트 분석 결과 세 가지 모델 모두 수온이 하천수질을 예측하는데 가장 중요한 유역환경 변수로 나타났으며 평균 경사도와 평균 고도, 증발산량, 배수좋음 등급의 비율이 공통적으로 중요한 유역환경 변수로 나타났다. 앞서 중요한 영향을 미치는 변수 외에도 도시지역과 산림지역의 비율 또한 BOD와 T-P의 하천수질을 예측하는데 중요한 유역환경 변수로 나타나 BOD와 T-P의 수질농도를 예측하는데 중요한 영향을 미치는 변수는 거의 유사한 것으로 나타났다. T-N의 경우, 하천수질을 예측하는데 월 최대강수량 평균값과 연강수량 합계값이 중요한 유역환경 변수로 나타나 기상요소와 관련된 변수들이 많은 영향을 미치는 것으로 확인되었다.

BOD, T-N, T-P의 하천수질에 공통적으로 중요한 영향을 미치는 유역환경 변수에 대한 부분의존도 그래프는 비슷한 패턴을 보이는 것으로 확인되었다. 수온은 13.5℃ ~18.5℃일 경우 BOD, T-N, T-P의 수질이 급격히 악화되는 것으로 나타났으며 18.5℃ 이상부터는 큰 변화가 없는 것으로 분석되었다. 평균 경사도는 증가할수록 하천수질이 지속적으로 개선되는 것으로 나타났으며 평균 경사도 20% 이상부터는 큰 변화가 없는 것으로 나타났다. 평균 고도는 약 20m 이상일 때부터 고도가 증가함에 따라 하천수질이 향상되는 것으로 나타났고 배수등급이 좋은 토양 비율은 60-77%의 비율을 나타낼 때 하천수질이 급격히 개선되었으며 77% 이상부터는 변화가 거의 없는 것으로 나타났다. 도시지역 비율의 경우 3-20%의 비율에서 비율이 증가함에 따라 하천수질이 악화되는 것으로 나타났으며 산림지역은 비율이 증가함에 따라 하천수질이 지속적으로 향상되는 것으로 나타났다.

4. 결론

본 연구는 랜덤포레스트 알고리즘을 통해 BOD, T-N, T-P의 하천수질 변수에 영향을 미치는 주요 유역환경 변수를 도출하고 부분의존도 그래프를 통해 주요 유역환경 변수값의 증감에 따른 하천수질의 변화양상을 파악하였다. 랜덤포레스트 알고리즘은 과적합 문제 회피를 통해 모델의 정확도를 향상시켜 전체적인 하천수질 농도를 예측하고 중요한 변수를 도출하는데 용이하나 하천 각각의 하천수질에 대한 유역환경 변수의 영향을 파악하기는 어렵다. 따라서 향후 모델의 해석력을 향상시키는 알고리즘을 통해 지역적인 스케일에서 개별적인 하천수질에 영향을 미치는 유역환경 지표를 도출하고 변수 간의 관계를 고찰할 수 있는 연구가 수행되어야 할 것으로 판단된다. 본 연구결과는 수질과 유역환경 특성에 따른 하천 수질관리 방안마련 및 유역관리 계획의 근거자료로 활용할 수 있다.

참고문헌

1. Baker, A.(2006) Land use and Water Quality. In Encyclopedia of Hydrological Sciences. Chichester, UK: John Wiley & Sons, Ltd. p. 17.
2. Breiman, L.(2001). Random forests. Machine Learning 45: 5-32.
3. Cha, Y., J. Shin, B. Go, D. S. Lee, Y. Kim, T. Kim and Y. Park(2021) An interpretable machine learning method for supporting ecosystem management: Application to species distribution models of freshwater macroinvertebrates. Journal of Environmental Management 291: 112719.
4. Ministry of Environment(2019) Guide to Diagnosis of Health Deterioration in River Aquatic Ecosystem. Sejong, Korea: MOE.
5. Park, S., S. Kim and S. Lee(2021) Evaluating the relationships between riparian land cover characteristics and biological integrity of streams using random forest algorithms. International Journal of Environmental Research and Public Health 18(6): 3182.
6. Park, Y., S. W. Lee and J. Lee(2020) Comparison of fuzzy AHP and AHP in multicriteria inventory classification while planning green infrastructure for resilient stream ecosystems. Sustainability 12(21): 9035.
7. Pratt, B. and H. Chang(2012) Effects of land cover, topography, and built structure on seasonal water quality at multiple spatial scales. Journal of Hazardous Materials 209: 48-58.
8. Wang, F., Y. Wang, K. Zhang, M. Hu, Q. Weng, and H. Zhang(2021) Spatial heterogeneity modeling of water quality based on random forest regression and model interpretation. Environmental Research 202: 111660.