랜덤포레스트기법을 이용한 분변성대장균 예측모델 개발

Development of fecal coliform prediction model using random forest method

서일원*, 최수연**

Seo, Il Won, Choi, Soo Yeon

요 지

하천에서의 분변성대장균은 분변성 오염 정도를 나타내는 지표로서, 이 농도가 높을수록 오염된 하천수와의 접촉을 통한 호흡기, 소화기 및 피부 관련 질병의 발발 확률이 높다고 알려져 있다. 따라서 하천에서의 수영, 수상스키 등과 같은 입수형 친수활동을 할 때, 분변성대장균 농도가 농도 기준 이하인지를 확인하고 이러한 정보를 친수활동에 이용할 필요가 있다. 그러나 분변성대장균의 경우, 현재 자동수질측정망에서 측정되고 있는 다른 수질인자들과는 달리 실시간 측정이 불가능하다고 알려져 있다. 분변성대장균을 측정하는데 있어 최소 18시간 이상이 필요하며, 이러한 분변성대장균 측정 방식은 하천 이용자들이 안전한 친수활동을 영위하는데 있어 적절한 수질정보를 제공하지 못한다. 그러므로 분변성대장균을 예측하는 모델을 개발하고, 이를 이용하여 실시간 분변성대장균 정보를 생성하여 하천 이용자들에게 제공할 필요가 있다.

본 연구에서는 친수활동이 활발하게 이루어지는 곳 중 하나인 북한강의 대성리 지점에 대해 데이터 기반 모델을 이용하여 분변성대장균을 예측하였다. 데이터 기반 모델은 물리 기반 모델에서 필요한 지형데이터나 비점오염원 등의 초기 오염물의 양에 대한 데이터를 필요로 하지 않고, 대신 독립변수로 사용되는 기상 및 수질데이터를 필요로 한다. 이러한 기상 및 수질데이터는 기존 기상관측소, 수질관측소에서 매일 자동으로 측정되기 때문에 데이터 기반 모델은 물리 기반 모델에 비해 입력데이터를 구성하기가 쉽다는 장점을 지닌다. 이러한 데이터 기반 모델 중 분류 모델은 회귀 모델과 달리 분변성대장균 농도가 일정 수질기준 이상을 넘는지를 바로 예측할 수 있다. 본 연구에서는 분류 모델 중 높은 예측력을 가진다고 알려진 랜덤포레스트(random forest) 기법을 이용하여 분변성대장균 예측 모델을 개발하였다. 분변성대장균 예측 모델은 주어진 기상 및 수질조건에 대해 분변성대장균이 200 CFU/100ml가 넘는지를 예측하였다. 예측된 분변성대장균이 기준을 넘는 경우를 2등급, 넘지 않는 경우를 1등급으로 명명하였다.

모델을 개발하기 위하여 북한강 대성리 인근 측정소에서 2010년부터 2015년까지 측정된 기상 및 수질데이터를 수집하였다. 수집한 데이터를 훈련 및 검증데이터로 샘플링하였으며, 이 때 샘플링한 데이터가 기존 데이터가 가지고 있던 등급별 비율을 유지하기 위하여 층화샘플링을 하였다. 본 연구에서는 샘플링에 의한 불확실성을 줄이기 위하여 랜덤하게 50번 샘플링된 각각의 훈련데이터에 대해 모델을 개발하였다. 50개의 모델의 검증 결과를 종합한 결과, 전체 예측률은 0.139로 나타났다.

핵심용어 : 분변성대장균, 수질기준, 데이터기반모형, 랜덤포레스트 기법, 북한강, 층화샘플링

* 정회원・서울대학교 공과대학 건설환경공학부 정교수・E-mail : seoilwon@snu.ac.kr

** 비회원・서울대학교 공과대학 건설환경공학부 박사과정・E-mail : jjasa1103@snu.ac.kr