EGU2020-5414, updated on 28 Oct 2020
https://doi.org/10.5194/egusphere-egu2020-5414
EGU General Assembly 2020

Predictive modelling of groundwater nitrate pollution at a regional scale using machine learning and feature selection

Aaron Cardenas-Martinez[1], **Victor Rodriguez-Galiano**[1], Juan Antonio Luque-Espinar[2], and Maria Paula Mendes[3]

[1]Universidad de Sevilla, Geografía Física y Análisis Geográfico Regional, 41004 Sevilla, España (vrgaliano@us.es)
[2]Instituto Geológico y Minero de España (IGME), Granada, España
[3]Civil Engineering Research and Innovation for Sustainability (CERIS), Instituto Superior Técnico, Universidade de Lisboa, Portugal

The establishment of the sources and driven-forces of groundwater nitrate pollution is of paramount importance, contributing to agro-environmental measures implementation and evaluation. High concentrations of nitrates in groundwater occur all around the world, in rich and less developed countries.

In the case of Spain, 21.5% of the wells of the groundwater quality monitoring network showed mean concentrations above the quality standard (QS) of 50 mg/l. The objectives of this work were: i) to predict the current probability of having nitrate concentrations above the QS in Andalusian groundwater bodies (Spain) using past time features, being some of them obtained from satellite observations; ii) to assess the importance of features in the prediction; iii) to evaluate different machine learning approaches (ML) and feature selection techniques (FS).

Several predictive models based on an ML algorithm, the Random Forest, were used, as well as, FS techniques. 321 nitrate samples and respective predictive features were obtained from different groundwater bodies. These predictive features were divided into three groups, regarding their focus: agricultural production (phenology); livestock pressure (excretion rates); and environmental settings (soil characteristics and texture, geomorphology, and local climate conditions). Models were trained with the features of a year [YEAR (t_0)], and then applied to new features obtained for the next year – [YEAR(t_{0+1})], performing k-fold cross-validation. Additionally, a further prediction was carried out for a present time – [YEAR(t_{0+n})], validating with an independent test. This methodology examined the use of a model, trained with previous nitrates concentrations and predictive features, for the prediction of current nitrates concentrations based on present features. Our findings showed an improvement in the predictive performance when using a wrapper with sequential search for FS when compared to the use alone of the Random Forest algorithm. Phenology features, derived from remotely sensed variables, were the most explanative features, performing better than the use of static land-use maps or vegetation index images (e.g., NDVI). They also provided much more comprehensive information, and more importantly, employing only extrinsic features of groundwater bodies.