Mapping nitrate concentrations in upper groundwater using Random Forest.

Job Spijker[1], Astrid Vrijhoef[1], Dico Fraters[1], Martine Hoogsteen[1]

[1] Centre for Environmental Quality, National Institute for Public Health and the Environment, Bilthoven, the Netherlands.

E-mail: job.spijker@rivm.nl

Abstract

The aim of this study was to create a map of nitrate concentrations leaching from the root zone of Dutch agricultural soils. We used nitrate data from a national monitoring network and the Random Forest algorithm as prediction and interpolation method. A large set of spatial auxiliary data, like soil types, groundwater levels and crop types, was used as dependent variables. The explained variance and statistical errors indicate that the interpolation and map visualisation is suitable for interpretation of the spatial variability of the nitrate concentrations in the Netherlands.

Keywords: nitrate, ground water, machine learning, random forest, mapping, monitoring

1. Introduction

The upper meter of groundwater is most susceptible to influences from agricultural practices and is therefore monitored in the Netherlands. Results of the monitoring are aggregated at farm type and four (soil) regions and presented as yearly averages (Boumans e.a. 2005).

While this is sufficient to evaluate the effectiveness of policy measurements, i.e. the changes in agricultural practices for each of the four regions, the current and future challenges regarding sustainable development of the agricultural system require that an overview delineated into smaller areas should be available. A map with the spatial distribution of nitrate concentrations in the upper groundwater can function as an evidence based management tool for regional policy regulations.

The overall objective of this study is to create a predictive model, based on the monitoring data at farm level, to estimate the nitrate groundwater concentrations as result of leaching from the root zone at a more local spatial level than the four regions and visualise this estimation in a map.

2. Method

Interpolation methods without the prerequisite of spatial autocorrelation are preferred because the result does not rely on distance between measurements but are mode dependent on local features within the auxiliary spatial data (Hengl e.a. 2018)

We created a predictive model using Random Forest (RF): a data driven machine learning technique. We used many spatial datasets as auxiliary data, i.e., soil type, land use, groundwater levels, crop types, pollutant emission, and altitude. This auxiliary data is providing the dependent variables. The nitrate monitoring data was used to train the model.

3. Results

Preliminary results show that with the trained model we explained about 45 to 50 % ot the observed variance in nitrate concentrations. Concentrations can be predicted with an RMSE of 18-20 mg L^{-1} and a mean absolute error ranging from 13 to 15 mg L^{-1}. A semivariogram indicated no spatial dependence between the residuals of the prediction.

4. Discussion and Conclusion

RF is a suitable method to predict nitrate concentrations when sufficient auxillary data is present. The explained

variance and statistical errors indicate that the map visualisation is suitable for interpretation of the spatial variability at a more local scale than the four regions. In our forthcoming paper we will publish the detailed map and explain our approach in detail.

Acknowledgements

This research was commisioned by the Dutch Ministry of Agriculture, Nature and Food Quality

References

Boumans, L.J.M., Fraters, D., Van Drecht, G., 2005. Nitrate leaching in agriculture to upper groundwater in the sandy regions of the Netherlands during the 1992–1995 period. Environmental Monitoring and Assessment 102, 225–241. https://doi.org/10.1007/s10661-005-6023-5

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B.M., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6, e5518. https://doi.org/10.7717/peerj.5518