PAPER

A machine learning based modelling framework to predict nitrate leaching from agricultural soils across the Netherlands

Job Spijker, Dico Fraters and Astrid Vrijhoef
National Institute for Public Health and the Environment, PO Box 1, 3720 BA, Bilthoven, The Netherlands

E-mail: job.spijker@rivm.nl

Keywords: mapping, random forest, interpolation, monitoring, Netherlands

Abstract

Throughout recent decades, the excessive use of animal manure and fertiliser put a threat on the quality of ground and surface waters in main agricultural production areas in Europe and other parts of the world. Finding a balance between agricultural production and environmental protection is a prerequisite for sustainable development of ground and surface waters and soil quality. To protect groundwater quality, the European Commission has stipulated a limit value for NO₃⁻ of 50 mg l⁻¹. Member states are obliged to monitor and regulate nitrate concentrations in groundwater. In the Netherlands, this monitoring is carried out by sampling nitrate concentrations in water leaching from the root zone at farm level within the national Minerals Policy Monitoring Program. However, due to the costly procedure, only a limited number of about 450 farms can be sampled each year. While this is sufficient for providing a national overview of nitrate leaching, as a result of current and future challenges regarding the sustainable development of the agricultural system, Dutch policymakers need to gain insight into the spatial distribution of nitrate at smaller spatial scales. This study aimed to develop a predictive modelling framework to create annual maps with full national coverage of nitrate concentrations leaching from the root zone of Dutch agricultural soils, and to test this model for the year 2017. We used nitrate data from a national monitoring program and combined them with a large set of auxiliary spatial data, such as soil types, groundwater levels and crop types. We used the Random Forest (RF) algorithm as a prediction and interpolation method. Using the model, we could explain 58% of variance, and statistical errors indicate that the interpolation and map visualisation is suitable for interpretation of the spatial variability of nitrate concentrations in the Netherlands. We used the variable importance from the RF and the partial dependency of the most important variables to get more insight into the major factors explaining the spatial variability. Our study also shows the caveats of data-driven algorithms such as RF. For some areas where no training data was available, the model’s predictions are unexpected and might indicate a model bias. The combination of visualisation of the spatial variability and the interpretation of variable importance and partial dependence results in understanding which areas are more vulnerable to NO₃⁻ leaching, in terms of land use and geomorphology. Our modelling framework can be used to target specific areas and to take more targeted regional policy measurements for the balance between agricultural production and protecting the environment.

Introduction

Finding a balance between agricultural production and environmental protection is a prerequisite for sustainable development of ground and surface waters and soil quality. Groundwater is frequently used as a source for drinking water or, when drainage systems are installed, directly affects surface water ecosystems. Throughout recent decades, the excessive use of animal manure and fertiliser put a threat on the quality of ground and surface waters in main agricultural productions areas in Europe and other parts of the world (Van Grinsven et al 2012, Wang and Li 2019).
To protect groundwater quality, the European Commission has stipulated a limit value for \(\text{NO}_3^-\) of 50 mg \(\text{l}^{-1}\). Member states are obliged to monitor and regulate nitrate concentrations in ground and surface water (European Commission 1991). The Dutch government has decreed monitoring of water leaching from the root zone of agricultural land (root zone leachate). This monitoring is carried out within the Dutch national Minerals Policy Monitoring Program (LMM). This root zone leachate is most susceptible to influences from agricultural practices such as crop rotation and the application of animal manure and artificial fertiliser (Boumans et al 2005). This monitoring network divides the Netherlands into four soil type regions based on their geomorphological and dominant soil type and characteristics (figure 1); these are referred to as regions in this article.

Results of the LMM are usually reported as status or trends of the nitrate concentration for each of the four soil regions and for each main farm type within a region, see for example https://lmm.rivm.nl/ (in Dutch) and Fraters et al (2020). Main farm types include arable, dairy, and (intensive) animal farming. Reporting the quality of root zone leachate at an aggregated level, such as soil regions, is sufficient to evaluate the effectiveness of the overall policy measurements regarding, e.g. the reduction of the usage of manure and fertiliser as laid down in action programs. However, current and future challenges regarding the sustainable development of the agricultural system require an overview delineated into smaller areas. For example, the National Plan of Action on Supporting the Transition to Circular Agriculture (Ministry of Agriculture, Nature and Food Quality 2019) postulates that cooperation between stakeholders and policy regulations should be more oriented towards regional challenges and targets, instead of national targets. Therefore, Dutch policymakers need a map with a nationwide overview of the spatial distribution of nitrate concentrations in the root zone leachate, including deep insight into factors which can explain the spatial variability in terms of natural and human influencing factors. This could function as an evidence-based management tool for the aforementioned regional policy regulations.

Traditionally, maps showing the spatial variability of measures concentrations are made with interpolation techniques based on spatial autocorrelation of the data, for example, ordinary kriging (e.g. McBratney et al 2003), regression (e.g. Boumans et al 2005) or inverse distance (e.g. Cohen et al 2012). Such spatial models are generally used for interpolation of univariate data. While auxiliary data can be incorporated in the interpolation model, for example, as regression models or by co-kriging (e.g. Kempen et al 2015), these models also rely on spatial autocorrelation. Interpolation methods without the prerequisite of spatial autocorrelation are preferred because their result does not rely on ‘a crude distance-weighted blend of geographically neighbouring measurement’ (Kirkwood et al 2016). Without spatial autocorrelation, predicted concentrations solely depend on local spatial features within the auxiliary data, instead of their position relative to measurements.

Auxiliary data can consist of other continuous scaled data, such as measurements of soil data or other spatial interpolations, but they can also consist of categorical data, for example, land use or crop type categories. Also, both predictor and auxiliary data are not necessarily statistically independent or do not necessarily follow a normal distribution. In the Netherlands, a large number of spatial datasets is available, containing data about, among others, soil type, land use, drainage class, average groundwater levels, crop types, pollutant emission and...
altitude (Kadaster 2020). Many of these datasets can be used as auxiliary data for a prediction model. Some of this data, such as crop type or pollutant emissions, is a result of farm management and can thus be potentially used to understand the impact of policy and regional management on nitrate leaching.

To create a predictive modelling framework using this large amount of available auxiliary data, a data-driven machine learning technique is preferable because of its efficiency in time and effort and relative low complexity compared to kriging or regression (Hengl et al 2018). Random Forest (RF) is a machine-learning technique which is often used in environmental and soil studies (Grimm et al 2008, Heung et al 2014, Rodriguez-Galiano et al 2014, Kirkwood et al 2016, Hengl et al 2017, Camera et al 2017). It is a decision tree-based algorithm which creates an estimate using an ensemble of a predefined number of trees, and it can be used for both classification and regression (Breiman 2001). Compared with other machine learning methods, its efficiency and resistance to over-fitting is an advantage, while having comparable accuracy in prediction (Heung et al 2016).

The aim of this study was to develop a predictive modelling framework to predict nitrogen concentrations that are lost from Dutch agricultural soils due to leaching from the rootzone. We evaluate the model predictions for the year 2017 and create a map with full national coverage of subsurface nitrate concentrations and we discuss the environmental implications.

The predictions from the modelling framework can contribute to the drafting of the aforementioned regional policy regulations, while the modelling framework itself can be of use in other countries or regions, providing the availability of the data.

Methods

Study area

The Netherlands is geologically part of the subsiding North Sea Basin, enclosed by the Brabant Massif in the South and the Rhenish Massif in the East. Due to the sedimentary infill of the basin, the subsurface across the Netherlands mainly consists of unconsolidated sediments. Altitude in the Netherlands ranges from a few metres below sea level to around 300 metres at the highest point in the south-east corner of the country. Due to the sedimentary nature of the Quaternary sediments, the Dutch subsurface contains large shallow aquifers of saline and fresh water. The phreatic zone usually starts around 1 metre depth in low-altitude areas, to a few metres in areas with higher altitudes. In general, the freshwater zone ranges from 50 m near the coast to 300 m in the middle and, due to the shallower Quaternary base, to around a few metres below surface in the east and south-east, near the border of the country (de Mulder et al 2003). Within the LMM, four soil type regions are distinguished: the Clay Region mainly consists of marine and fluviatile, marine or glacial origin (de Mulder et al 2003). Soils in the Netherlands are arbitrarily defined on the Dutch soil map as the upper 120 cm (Locher and de Bakker 1990). Dutch soils are usually grouped into five soil types: marine and fluviatile clay, peat, sand and loess. While this classification overlaps with the geogenic origin of the sediments, these five soil types are mainly based on soil texture and therefore often differ from lithogenic classification (Stiboka 1965, Van der Veer 2006). Within the LMM, four soil type regions are distinguished: the Clay Region mainly consists of marine and fluviatile clay soil; the Peat, Sand, and Loess Regions are roughly based on the peat, sand, and loess soil respectively.

Due to the unconsolidated nature of the Quaternary sediments, the Dutch subsurface contains large shallow aquifers of saline and fresh water. The phreatic zone usually starts around 1 metre depth in low-altitude areas, to a few metres in areas with higher altitudes. In general, the freshwater zone ranges from 50 m near the coast to 300 m in the middle and, due to the shallower Quaternary base, to around a few metres below surface in the east and south-east, near the border of the country (de Mulder et al 2003).

Excluding open water, in 2015 the land area in the Netherlands was 3.37 million ha; 66% is in use for agricultural activities, and 14% for forest and nature areas. For the agricultural area, 1.85 million ha is used for production, 58% is in use for cattle, and 25% is used for arable land (Statistics Netherlands 2020).

Data collection and preparation

For our predictive modelling we used data from the Dutch national LMM monitoring program on nitrate concentrations in water leached below the rooting zone from agricultural soils across the Netherlands. The LMM monitors among others water quality of leachate at around 450 commercial farms. Farms are a stratified random sample from a group of 1,500 farms participating in Farm Data Accountancy Network (FADN). LMM strata are based on soil type, region and district, farm type and farm economic size class. LMM represents at least 80% of the agricultural used area in each of the four soil type regions (Van Vliet et al 2017, Fraters et al 2021).

At each farm, samples are taken at 16 locations within agricultural fields. The number of locations per field depends on field size; locations within a field are randomly selected with a selection tool. Geohydrological conditions determine the sampling method. The types of water sampled as leaching water are (1) soil moisture in the unsaturated zone beneath the root zone, at a depth of between 1.5 and 3.0 metres below the surface if the groundwater is at more than 5 metres below the surface (mainly in the Loess Region, 1.5% of the Dutch agricultural area); (2) drain water if the majority of the fields are drained by tile drains (only in the Clay Region, 42%); (3) the top metre of the phreatic groundwater if the groundwater is less than 5 m below the surface and farms do not comply with drain water monitoring (Peat and Sand Region, 9% and 47% respectively).
Most samplings are carried out during the winter period. Soil moisture and the upper metre of groundwater are sampled once a year. However, at farms in the Clay Region and the low, wet parts of the Sand Region, groundwater is sampled twice. The 16 individual samples are used to make two mixed samples, each consisting of 8 individual samples. Tile drain water is sampled four times a year. One mixed sample, containing 16 individual samples, is made per sampling round. All mixed samples are analysed in the laboratory.

The methodology of the LMM is detailed in Fraters et al (2001), Boumans et al (2005), and Hooijboer and Fraters (2011). Locations and individual data of these farms are confidential under an agreement with participants in the monitoring network. Due to these restrictions related to privacy, it is not possible to provide a map with farm locations nor show individual measurements. However, the design of the LMM provides for a representative distribution of farms across the study area.

A summary of the monitoring data for the year 2017 is given in Table 1. Figure 2 shows the nitrate concentrations in box-and-whisker plots. The distribution of nitrate concentrations is skewed to the right, with maximum values around 110 mg l$^{-1}$. The Sand and Loess regions show the highest median concentrations, respectively 27 and 48 mg l$^{-1}$. The data from the Clay Region shows a lower median concentration of 16 mg l$^{-1}$, while in the Peat Region, the median nitrate concentrations are the lowest at 1.3 mg l$^{-1}$. The variance in nitrate concentrations is comparable for the Sand, Clay and Loess regions but evidently smaller in the Peat Region.

The LMM uses yearly averaged nitrate concentrations for each farm as a basis for the status and trends reports. In this study, we used the same farm annual mean concentration for NO$_3^-$, To relate farm management and cropping to the measured nitrate concentrations, nitrate concentrations are attributed to farm management in the previous year, with the exception of soil moisture. This is based on soil leaching time. Nitrate concentrations measured in tile drain water and groundwater sampled in winter 2016–2017, groundwater sampled in summer 2017 and soil moisture sampled in winter 2017–2018 are attributed to the farm management data in 2016.

Farm annual nitrate means were selected, and locations and acreage of farms were added to the dataset. Extreme outliers in the dataset were indicated when their value exceeded the median plus or minus a threshold of 10 times the median absolute deviation (MAD) of the total dataset, according to the method of Leys et al (2013) but choosing a far more conservative threshold of 10 compared to the suggested value of 2.5 by Leys et al (2013). We excluded these outliers from the dataset based on the assumption that such outliers result from very localised factors, or sampling and laboratory errors for which no information is included in the auxiliary data. Hence, we expect that these extreme values cannot be sufficiently predicted and such prediction is beyond the scope of our study. However, these outliers will result in extreme values for the residuals. Since the model performance is based on a mean square error, these extreme values will strongly impact the calculation of this error.

Since nitrate concentrations are averaged on farm scale, model parameters were estimated on an analogous spatial scale. Geographically, a sample can be considered as point information while a farm is an area or polygon. Therefore, auxiliary data should be considered using equal geographical units, i.e. an area. For categorical feature data, such as soil type, land use or crop type, the relative fraction of the total surface of the farm was calculated, according to Boumans et al (2005 et al). For numerical feature data, an average per farm was used.

Table 2 shows the auxiliary data used for this study. Each dataset was transformed into a 25 m raster grid, with the same projection and extent, as covariate layer for the model.

For some categorical datasets, a different classification was created to reduce the number of available classes. Crops were grouped into main crop types such as vegetables, potatoes, beets, grains and grass. From the soil map attributes, both detailed soil types as well as main soil groups were used.

We used these maps as spatial co-variables and assumed that the variability in these maps covered spatial autocorrelation and geographical. No additional information about location (i.e. coordinates) or distance is used for training the RF model.

**Model setup and optimisation**

In our data driven modelling framework that we developed, we used the Random Forest (RF) algorithm for the spatial prediction of nitrate concentrations. Our procedure is based on the approach described by Hengl et al (2017). Data...
Table 2. List of auxiliary data used. Each dataset is either a polygon (p) or raster (r) map. Polygon maps can contain multiple attributes. A more detailed table can be found in the Supporting Information accompanying this paper.

| Dataset                  | Variable name                  | Source                                                                 | Description                                                                                     |
|--------------------------|--------------------------------|----------------------------------------------------------------------|-------------------------------------------------------------------------------------------------|
| Crop type (p)            | brp_gewaspercelen_2015          | National inventory of crop types                                      | Crops categorised using crop categories for the monitoring network                               |
| Crop category (p)        | brp_categorien_2015            | Derived from crop type                                                |                                                                                                 |
| N Emission (r)           | Erspatial_Landbouw_bodem_ntot_2015_rast | National Pollutant Release and Transfer Register (PRTR)               | Emission from the agricultural sector of N-total to soil (bodem), air (lucht) and water (water). |
| NH4 Emission (r)         | Erspatial_Landbouw_lucht_nh4_2015_rast, Erspatial_Landbouw_water_nh4_2015_rast | PRTR                                                                 | Emission of NH4 from the agricultural sector to air                                             |
| P Emission (r)           | Erspatial_Landbouw_bodem_ptot_2015_rast, Erspatial_Landbouw_water_ptot_2015_rast | PRTR                                                                 | Emission of P-total to soil (bodem) and water.                                                   |
| Elevation (r)            | ahn25                          | National Digital Elevation model                                      |                                                                                                 |
| Landuse (r)              | lgn7                           | National Landuse model, version 7                                     |                                                                                                 |
| Groundwater level (r)    | ghg_1998-2006_l1               | National hydro-logical model                                          | Estimated average high groundwater levels                                                       |
| Organic matter (r)       | Orgstof,....cm                 | Organic matter at soil layers 5, 10, 25, 40, 60, 80, 100, 120 cm depth | Data supplied by Wageningen Agricultural University                                             |
| Soil map (p)             | BodemHoofd, bodemGwt, bodemEerste_bod, bodemCode, | Data from national soil map, soil type, groundwater regime, soil category and subcategory | Selection of soil map attributes                                                                |
| Clay fraction (r)        | lutum                          | Clay fraction upper soil                                              | Data supplied by Wageningen Agricultural University                                             |
| Soil regions (p)         | hgr                            | LMM Soil region map                                                   | See figure 1                                                                                    |

* http://www.nationaalgeoregister.nl/geonetwork/srv/dut/catalog.search#/metadata/b812a145-b4fe-4331-8dcb-d914327a87ff.
* http://www.emissieregistratie.nl/erpubliek/bumper.en.aspx.
* http://www.nationaalgeoregister.nl/geonetwork/srv/dut/catalog.search#/metadata/5x1adb58-0fb6-425d-b06f-577a56c0a4228.
* http://www.nationaalgeoregister.nl/geonetwork/srv/dut/catalog.search#/metadata/DD8407B0-E7D7-4EDF-807B-CCF7C04DC870.
* http://www.nhi.nl/nl/index.php/data/.
* https://www.rivm.nl/landelijk-meetnet-effecten-mestbeleid/organisatie-van-lmm/regionale%20aanpak.
handling, training of the model, prediction and map generation are all implemented in the R statistical software (R Core Team 2020). Figure 2 gives a schematic overview of the framework.

For monitoring data and auxiliary data, the intersecting co-variables are aggregated at farm level. This results in a large table, the feature matrix, with the concentration of NO$_3^-$ as the independent variable and many dependent variables. This feature table, our training dataset, was used to train the RF model. The training dataset contained 450 cases ($n$) and 895 dependent variables ($p$). From this set, 4 cases were removed as outliers and 2 cases did not have nitrate concentrations.

The Ranger R package is used for providing the RF algorithm (Wright & Ziegler 2017). Important parameters for the RF algorithm are the number of trees and the number of variables randomly sampled for each split. Starting values for the number of trees are generally set at 1500. The number of variables for each split is based on the number of dependent variables, $p$, and $p/3$ is often used as starting value (see Breiman 2001). The calculated out-of-bag Mean Square Error (oobMSE) was used as a measure of the accuracy of trained models. Using the oobMSE, in an interactive approach, optimal values for the number of trees and the number of variables randomly sampled for each split where chosen. The initial model was trained using the default values for the number of trees and the number of variable splits at each node. We varied the number of trees and splits until no relevant decrease of the oobRMSE and the RMSE of the cross validation (see below) was observed.

To obtain more insight into the model prediction, we used the variable importance in conjunction with partial dependency plots (Greenwell 2017). For the most important variables, we used the partial dependency to study how these variables are related to the prediction.

To obtain an independent measure of accuracy of the model, we used a separate 10-fold cross validation. In this cross validation, the samples were randomly divided into 10 groups. Over 10 iterations, 9 groups were selected to train the model while 1 group was selected for validation, and each group was selected once as validation group. Accuracy was calculated using the Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Median Absolute Error (MedAE) of the difference between observed and predicted values.

To create a map of predicted nitrate concentrations in the spatial domain, we created a spatial grid with a resolution of $500 \times 500$ m. The area of a single $500 \times 500$ m cell is a practical choice considering the resolution of the auxiliary data and the area of the average farm. The surface of each pixel used for prediction is the surface of the agricultural parcels within the pixel. The same applies for farms, where the surface of the farm is related to the surface of the parcels within a farm. The grid was created using a mask covering the agricultural area (farm parcels) in the Netherlands.

The masked $500 \times 500$ m grid was intersected with the same auxiliary data used for creating the feature data matrix, which was used to train the model. For each grid cell, we calculated surface fractions of categorical auxiliary data and averages of numerical auxiliary data, and subsequently used the resulting data matrix as target.
features for prediction. We used the trained RF model to predict nitrate concentrations using the target matrix. Nitrate concentrations were only predicted for the 500x500m pixels if they contained an area with agricultural parcels. If no agricultural parcels intersected with the pixel, no nitrate concentration was predicted. Predictions for each cell were visualised on the map.

Results and discussion

To understand the environmental implications of a predictive framework and to explain the spatial patterns on map as a results of this framework, it is important to have a good understanding of the underlying model. Therefore we will not only discuss the model outcome but also try to put the model accuracy and prediction in the context of our data.

To optimise the our model we iterated over the amount of trees and splits using our training data. The final model was based on 1500 trees and 0.25*p/3 splits, and it explained 58% of the variance while the oobMSE was 485. The 10-fold cross validation confirmed this performance. Explained variance based on the cross validation was a slightly higher 59%, the MSE was 504 and the MAE and MedAE were 15 mg l\(^{-1}\) and 11 mg l\(^{-1}\) respectively.

If we included the four outliers, with concentrations larger than 200 mg l\(^{-1}\), the predictions were comparable, the correlation coefficient of predictions with and without outliers was 0.98. Median and MedAE were the same. As expected, the MSE of the model with the outliers included increased significantly to 879, and hence, a lower explained variability of 53%. Examination of the residuals did indeed confirm that the outliers resulted in extreme residual values which explain the change in MSE.

The prediction variance is presented in figure 4, which shows the observed values versus the fitted values from the k-fold cross validation. To overcome the restriction of not showing individual measurements, the data of figure 4 is ranked based on predicted value. Next, for each group of 10 subsequent predictions the average prediction and observed value is calculated and the mean and standard deviation (error bars) is depicted. Figure 4 and analysis of the individual residuals (not shown) show that for observed concentrations above 70 mg l\(^{-1}\), the prediction is close to the 1:1 line; above these values, the prediction tends to underestimate the observations.
Both the explained variance and the oobMSE indicate that the RF model explained a large proportion of the observed variance in the monitoring network but that the variance of predictions should be taken into account when interpreting the results. Considering the overall variance in the dataset, the MSE and MAE showed that the prediction had a relatively large confidence interval. While we consider this sufficient for assessing spatial patterns at a regional scale, an assessment at smaller scales (e.g. farm scale) can only be performed if one takes the higher uncertainty into account.

Figure 5 shows variable importance. The most important variables are related to the the percentage of grassland on the farm (lgn7.1: land use category 1, grassland; and brp_gewaspercelen_2016.265: crop category 265, permanent grass). The partial dependency plot (figure 6) shows that lower nitrate concentrations are predicted on farms with a high percentage of grass. There are two reasons for this. Firstly, farms with a high percentage of grassland are often located in peaty, low-altitude areas. In these areas, measured nitrate concentrations are relatively low, while, for example, in more sandy, higher-altitude regions, the nitrate concentrations are relatively high and cropping is more diverse (Fraters et al 2021). Secondly, also on sandy soils, a similar nitrogen surplus on grassland results in lower nitrate concentration than on arable land due to higher denitrification in grassland soils (Fraters et al 2015). Altitude (ahn25) also has high importance: up to 25 m altitude the model predicts an increasing nitrate concentration which concurs with the aforementioned partial relation with grasslands.

The partial dependency plots (figure 6) also show that increasing P emission to surface water (variable ERspatial_Landbouw_water_ptot_2016_rast) results in lower NO₃⁻ concentration. P emission to surface water occurs especially at fields with high groundwater levels (Schoumans 2015). In fields with high groundwater levels, the denitrification rate is often high and, therefore, nitrate concentrations relatively low (Fraters et al 2015). The N emission to surface water (variable ERspatial_landbouw_water_ntot_2016_rast) also shows lower NO₃⁻ concentrations with increasing N emission, but only for lower emission values. With further increasing of the N emission, the NO₃⁻ concentration increases also. It is important to note that the P and N emissions are values which results from the National Water Quality Model, which consists of several nationwide submodels (National Pollutant Release and Transfer Register 2019). Interpretation of a prediction based on a modelled value, without knowing the true value of the modelled variable, is cumbersome. However, since all models are wrong, using such variables can still be very useful in terms of prediction and explaining variance (see, e.g. Box 1976).

Other important variables are geomorphological or geogenic parameters reflecting the soil types and landscape features, e.g. clay content (lutum), groundwater level (ghg_1998_2006_II). Some collinearity between these parameters exists. Areas with low altitude in the Netherlands are characterised by clayey and peaty soils and shallow groundwater levels. Higher-altitude areas are often Pleistocene sandy areas with deep groundwater.

Figure 4. Plot of observed versus predicted values from the k-fold cross validation. Each group of 10 subsequent predictions, the mean prediction and observed value are plotted, together with the standard deviation (error bars). For explanation, see text.
Figure 5. Variable importance of the RF model. Variable names refer to names in table 2. For categorical variables, the category number is in the dot-suffix of the variable.

Figure 6. Partial dependency plots for the 10 most important variables. Variable names refer to names in table 2. For categorical variables, the category number is in the dot-suffix of the variable.
levels. The dependency on organic matter (orgstof\textsubscript{005cm}, orgstof\textsubscript{100cm}, orgstof\textsubscript{150cm}, at resp. 5, 100, and 150 cm depth) is probably caused by geomorphological and landscape features, while in theory organic matter also has a more mechanistic relation with explaining nitrate concentrations. Due to the nature of the RF algorithm, the prediction accuracy does not suffer from this dependency between geomorphological and geogenic variables (Breiman 2001, Tomaschek \textit{et al} 2018).

The combination of variable importance and partial dependency provides insight into the relative importance of the predictor variables from the auxiliary data and show the nature of the relationship between the predictor and the target variable. Our results show that the NO\textsubscript{3}\textsuperscript{−} concentrations depend on a combination of geomorphology and land use. The land use parameters are interesting because they can indicate how farm management affect the NO\textsubscript{3}\textsuperscript{−} concentrations. Unfortunately, farm management data was not available and the variables used in this study, cropping and P emission to surface water, are too coarse or complicated to interpret. However, we envisage that if we could improve the model by adding farm management data, a clearer delineation can be made between how natural factors affect NO\textsubscript{3}\textsuperscript{−} concentration in relation to the control by farm management.

To create the final map, the features of the dataset used for prediction are calculated the same way as the features of the training dataset. However, the surface of the used pixels, 500 × 500 m (25 ha), is half the surface of the median farm area, which is 43 ha. While this, in theory, might influence the prediction, we consider the values of the independent variables to be relative to the surface of the area of either the farm or the predicted pixel. Considering the variance of the prediction, we assume that this difference in surface area is not relevant for the prediction. The predictions are shown in a map, see figure 7.

The map in figure 7 shows that the highest predicted concentrations occur in the southern sandy area, generally above the 50 mg l\textsuperscript{−1} threshold for groundwater. The lowest concentrations occur in the peaty areas in the West and North of the country. The predicted spatial pattern of nitrate concentrations is in accordance with earlier reported nitrate predictions (Boumans \textit{et al} 2008) and the online reports from the LMM (see https://lmm.rivm.nl/, in Dutch). If we overlay the predictions with the sample locations, the spatial patterns of measured nitrate concentrations concur with the predictions (data not shown).

\textbf{Figure 7.} Map of nitrate concentrations in the root zone leachate for the year 2017. Areas A, B, and C are areas with high predicted concentrations, marked for discussion.
One unexpected result of our model is the prediction of high concentrations in the Veluwe area, area B in figures 1 and 7. This area is located on glacial deposits with higher altitudes which consist mostly of nature areas and extensive agriculture. Also, no measurements are available from the LMM on these high-altitude areas in the Veluwe. In the North Limburg area, area C in figures 1 and 7, the predicted high concentrations are expected and concur with measurements from the same area. Both areas have in common that they are located on high altitude (80–100 m) sandy soils with diverse cropping. However, the North Limburg area consists of periglacial deposits and (reclaimed) peat, and has intensive agriculture. The geomorphological, land use, and farm management (e.g. manure application rate) differences between areas B and C are not part of the auxiliary data. Since there are no measurements in area B, the model is not trained on other variables which might differentiate between these areas; hence the model predicts the same high concentrations in area B as in area C. The same goes for area A, a higher altitude sandy dune region. No measurements are available here either, and predicted concentrations are relatively high. This shows the caveats of machine-learning models such as RF, which are based on associations in the data. Our auxiliary data and measurements do not cover the differences between areas A, B and C. Hence, the model is biased to use the same range of the predictions for each area, contradictory to our expectations.

In general, while the spatial pattern of nitrate concentrations at the national scale is clear from the map, a more detailed assessment of nitrate concentrations at smaller local scales should be executed with caution, considering aforementioned caveats and the prediction variance. As a minimum, using the predictions at more local scale than this variance, either the RMSE or absolute errors should be taken into account. To reduce the prediction variance, additional data can be added to the framework which, for example, represents the local variation in land use or land management. Land management data in particular were not available at the national scale for this research. For our future research, we will investigate how we can reduce the prediction uncertainty with additional data.

The combination of visualisation of the spatial variability and the interpretation of variable importance and partial dependence results in understanding which areas are more vulnerable to NO3− leaching, in terms of land use and geomorphology. Mitigating nitrate loss to groundwater through policy measurements regarding the reduction of manure and fertiliser use can result in a trade-off for agricultural production. Our modelling framework can be used to target specific areas, distinguished by a specific combination of geomorphology and land use factors, to take more precise policy measurements. This can be beneficial because general nationwide policy measurements also target areas were nitrate emission to groundwater is of no concern. When using the results of this study for smaller spatial areas, one must consider the prediction error in relation to how the results are applied to certify that the error is acceptable for the considered application.

Conclusions

Using our RF predictive modeling framework, we created a map of nitrate concentrations leached from the root zone of agricultural soils across the Netherlands for the year 2017. With our model, we interpolate the nitrate concentrations measured at the farm level on a national scale. In our model, the most important variables for the prediction are variables related to the present of grasslands (land use, crops), and variables related to altitude, soil (soil type, clay and organic matter content), groundwater level and N and P emissions to surface water. The explained variance and statistical errors indicate that the map visualisation is suitable for interpretation of the spatial variability at a smaller local scale than the four soil type regions of the LMM. Although predictions for specific areas for which no measurement data are available during the training stage of the model showed unexpected results, probably due to model bias. The profound insight provided by the RF variable importance and partial dependence of the most important variables, gives local policymakers opportunities for creating more regional targeted action plans for the balance between agricultural production and protection of the environment.

In our future research we will apply the modelling framework on different years to see how the spatial patterns in the Netherlands change over time. Also we foresee that we will use the framework to predict the spatial variability of other parameters measured in the LMM monitoring network, for example dissolved phosphate or dissolved organic carbon.

We believe that our framework can be applied to other regions within Europe and maybe elsewhere. The most challenging part of our framework is that measurements must be linked to land use and farm management data, using the same spatial scale in both the training and the prediction (target) datasets. As the data are derived from the LMM monitoring program we use farm average nitrate concentrations to train our model and model predictions are on the farm scale as well. We presume that for others regions where, for example, only surface water is monitored at the outlet of (micro-) catchments, a different approach is needed. In such case (micro-) catchments can be considered as spatial scale for training and prediction.
One of the problems we encountered in our study is the interpretation of the uncertainty, or model error, of Machine Learning algorithms in terms of policy applications. For future research we recommend to specifically address this uncertainty. We propose that, besides the quantitatively calculated error terms, the uncertainty should also be expressed in more qualitative terms. This would help policy makers with the interpretation of the results and would give them confidence in how to apply the results.

Acknowledgments

This study was performed within the Dutch National LMM monitoring program, financed by the Ministry of Agriculture, Nature and Food Quality (RIVM Project M350601). We thank Annemieke van der Wal for her comments on the draft version of this paper. We also like to thank an anonymous reviewer for their suggestions, which improved this paper.

Data availability statement

The data generated and/or analysed during the current study are not publicly available for legal/ethical reasons but are available from the corresponding author on reasonable request.

ORCID iDs

Job Spijker https://orcid.org/0000-0001-9276-2640
Dico Fraters https://orcid.org/0000-0003-1558-1920

References

Boumans L, Fraters D and van Drecht G 2008 Mapping nitrate leaching to upper groundwater in the sandy regions of The Netherlands, using conceptual knowledge Environ. Monit. Assess. 137 243–9
Boumans L J M, Fraters D and Van Drecht G 2005 Nitrate leaching in agriculture to upper groundwater in the sandy regions of the Netherlands during the 1992–1995 period Environ. Monit. Assess. 102 225–41
Breiman L 2001 Random forests Mach. Learn. 45 5–32
Box G E P 1976 Science and statistics
Cohen D R, Rutherford N F, Morisseau E and Zissimos A M 2012 Geochemical patterns in the soils of Cyprus Sci. Total Environ. 420 250–62
de Mulder E F, Geluk M C, Ritsema I L, Westerhoff W E and Wong T E 2003 De ondergrond van Nederland (The underground of the Netherlands, in Dutch) (Groningen: Wolters Noordhoff)
European Commission 1991 Council Directive 91/676/EEC of 12 December 1991 concerning the protection of waters against pollution caused by nitrates from agricultural sources
Fraters B, Hooijboer A E J, Vrijhoef A, Plette A C C, Van Duijnhoven N, Rozemeijer J C, Gosseling M, Daatselaar C H G, Roskam J L and Begeman H A L 2021 Agricultural practices and water quality in the Netherlands: status (2016–2019) and trends (1992–2019). The 2020 Nitrate Report with the results of the monitoring of the effects of the EU Nitrates Directive action programmes National Institute for Public Health and the Environment, Bilthoven, the Netherlands, RIVM report 2020–0184 (https://rivm.nl/bibliotheek/rapporten/2020-0184.pdf)
Fraters B, Boumans L J M, Van Leeuwen T C and De Hoop W D 2001 Monitoring nitrogen leaching for the evaluation of the Dutch minerals policy for agriculture in clay regions TheScientificworld J (S2) 758–66
Fraters D, van Leeuwen T, Boumans L and Reijis J 2015 Use of long-term monitoring data to derive a relationship between nitrogen surplus and nitrate leaching for grassland and arable land on well-drained sandy soils in the Netherlands Acta Agriculturae Scandinavica, section B—Soil & Plant Science 65 144–54
Greenwell B M 2017 pdp: An R package for constructing partial dependence plots The R Journal 9 421–36
Grimm R, Behrens T, Mäker M and Ellenbeer H 2008 Soil organic carbon concentrations and stocks on Barro Colorado Island—Digital soil mapping using Random Forests analysis Geoderma 146 102–13
Hooijboer A E J and Fraters B 2011 Nitrogen leaching to groundwater and surface water in the Dutch peat region Groundwater Quality Management in a Rapidly changing world. Proc. of the 7th Int. Groundwater Quality Conf. Zurich (Switzerland, 13–18 June 2010) (IAHS Publication) 342, pp 358–61
Hengl T et al 2017 SoilGrids250m: global gridded soil information based on machine learning PLoS One 12 e0169748
Hengl T, Nussbaum M, Wright M N, Heuvelink G B M and Graier B 2018 Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables PeerJ 6 e5518
Heung B, Bulmer C E and Schmidt M G 2014 Predictive soil parent material mapping at a regional-scale: a Random Forest approach Geoderma 214–215 141–54
Heung B, Ho H C, Zhang J, Kruddy A, Bulmer C E and Schmidt M G 2016 An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping Geoderma 265 62–77
Kadaster 2020 National georegister - PDOK [WWW Document] URL https://www.nationaalgeoregister.nl/ (accessed 7.1.20)
Kempen B, Brus D J and de Vries F 2015 Operationalizing digital soil mapping for nationwide updating of the 1:50,000 soil map of the Netherlands Geoderma 241–242 313–29
