A machine learning approach to understand nitrate leaching in Iowa watersheds

by

Ishan Nalinkant Patel

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Industrial Engineering

Program of Study Committee:
Guiping Hu, Major Professor
Michael Castellano
Cameron MacKenzie

Iowa State University

Ames, Iowa

2022

DEDICATION

I dedicate this thesis to my beautiful family.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

NOMENCLATURE

CV	Cross-Validation
LightGBM	Light Gradient Boosting Machine
ML	Machine Learning
RF	Random Forest
MARB	Mississippi Atchafalya River Basin

ACKNOWLEDGMENTS

I would like to thank my committee chair, Dr. Guiping Hu, and my committee members, Dr. Michel Castellano and Dr. Cameron MacKenzie, for their guidance and support throughout the course of this research.

In addition, I would also like to thank my friends, colleagues, the department faculty, and staff for making my time at Iowa State University a wonderful experience.

ABSTRACT

As one of the corn belt states of the US, Iowa has corn and soybean as the main row crops, which are the main source of nitrate leaching. Many agencies such as USEPA focus on the hypoxic zone in the Gulf of Mexico caused by nitrate leaching in the croplands of MARB. Researchers have utilized quantitative methods such as regression, simulation, and qualitative methods to calculate nitrate load. Since machine learning aims to understand the structure of the data and fit that data to models to predict future outcomes, it can be a great way to tackle this problem because it can predict future outcomes and provide additional insights from the data. The time-series dataset used in this study focused on predicting Nitrate yield (kg NO_3-N ha^{-1} cropland) for 29 watersheds of Iowa, for which the data was collected from 2001 to 2018. The objective of this study was to find relationships between the nitrate yield with the independent variables from the dataset, which can explain the trend and help understand future nitrate leaching in the state of Iowa. The same model can identify potential causes and relationships for different datasets from different states. Walk Forward Cross-Validation approach was used for this study, which focuses on solving time-series analysis problems. The RRMSE value of the trained model for the test year 2018 was 23.68%, with an R^2 score of 77.06%. The model suggested that the most important features were annual discharge, rain, corn to soybean ratio, and other variables. The Partial Dependency Plots (PDP) explain their relationship with the target variable, nitrate yield. The relationship from PDP shows an underlying aspect of what value ranges contribute to the sudden changes in nitrate yield and how this finding can help policymakers and environmental agencies understand the problem further.

CHAPTER 1. INTRODUCTION

Nitrate leaching is a naturally occurring process where nitrate leaves the soil and either mixes with surface water or passes through the different soil layers and merges with groundwater. Nitrate is an essential component of modern agricultural applications, and it is not a problem when it is within the root zone of the crops, but once it leaves there, it becomes an environmental pollutant, which goes on to merge with ponds, rivers, and other water bodies, ultimately to the ocean. The Mississippi Atchafalaya River Basin (MARB) is one of the largest river basins in the world, and 24 states contribute directly or indirectly to the river basin, which then leads to the Gulf of Mexico. The nitrate losses of the Midwest crop region and the US Corn Belt of the MARB are important factors behind the hypoxic zone in the northern part of the Gulf of Mexico. By 2035, The United States Environmental Protection Agency (USEPA) plans to reduce the average annual size of the hypoxic zone from ~ 15,000 km^2 to ~5,000 km^2.

The nitrate leached in the MARB damages the aquatic ecosystem. It pollutes the drinking water as the nitrate levels in ground and surface waters are a vital parameter of groundwater pollution. Higher levels of nitrate can be toxic to newborns. The maximum nitrate limit in drinking water has been set to 10 mg NO_3-N/l by the USEPA. Nitrate leaching is a big issue because it pollutes the groundwater and helps eutrophicate phosphate by carrying sulfate and immobilizing iron (Smolders et al., 2010)**.**

Nitrate leaching is a complex process that has many reasons behind it. The introduction of synthetic N fertilizer to the US Corn Belt saw a 50-300% increase in crop productivity, leading to cropping methods that focus on warm-season annual crops such as maize and soybean (Hatfield et al., 2009). Perennial crops have low nitrate loss (Randall et al., 1997), and annual crops, on the contrary, have high nitrate losses (David et al., 2010). Part of the reason behind this is the root depth of these crops, as perennial crops have longer roots, which means they can hold on to the nitrate more than the annual crops with shallow roots. Another reason behind the difference in nitrate loss is that annual crop like maize and soybean has fallow periods. The land is no longer cultivated during the fallow periods, and the absence of plant demand leads to higher nitrate loss (Martinez-Feria et al., 2018). So fallow periods in conjuncture with excessive Nitrate fertilizer have increased the region's Nitrate loss.

The soil profile is also significant for nitrate leaching, as a coarse profile with less depth to the water below would be more susceptible to nitrate leaching. For example, sand is more vulnerable to nitrate loss than clay. Weather variables such as temperature and rain affect the nitrate leaching tremendously, as a higher amount of rain can drive the nitrate through the soil. (Lu et al., 2020) has found that extreme precipitation drives the increase in nitrate load in the Gulf of Mexico. Also, the interannual weather variability in conjunction with the cropping timeline can be vital. For example, most nitrate leaching occurs during the winter as the annual crops like maize and soybean would have been under the fallow period. During the early application of N fertilizer in spring and summer, the nitrate can be lost very quickly if there is an excess of rain. The annual discharge is the total water being discharged at one location of a watershed, which shows the volume rate of water flow being transported through a watershed. The value of annual discharge is highly correlated with nitrate yield as higher discharge would bring more suspended solids, dissolved chemicals, and biological materials with it.

For many decades various national and local governing agencies have been working to reduce nitrate loss from cropping systems to ground and surface waters. (Schilling & Wolter, 2009) used the SWAT model to find the load reduction strategies for the Des Moines River. (Dybowski et al., 2020) have created an interactive website that calculates Nitrate leaching load based on various inputs collected from farmers in Poland's puck commune area. (Danalatos et al., 2022) used the 5-year moving average method and Monte Carlo simulations to measure long-term trends in nitrate losses for Iowa watersheds and found that the interannual variability in weather is strongly responsible for the changes in nitrate yield. (Gentry et al., 2014; McIsaac et al., 2016) used regression to calculate nitrate load and understand interannual variation in nitrate yield in the rivers of Illinois.

Ecological predictions such as nitrate leaching can be effectively carried out using Machine Learning (ML). With adequate data of good quality, a machine learning model can predict future outcomes and provide essential insights that can be useful to solve the underlying problems. Time series analysis of the ecological predictions problem is challenging, but a rich dataset with various spatial and temporal variables helps learn the model better. (Spijker et al., 2021) took the ML approach to map N concentration across the Netherlands. (Ransom et al., 2017) used a hybrid ML approach to predict nitrate concentration in California. Hence, forecasting the nitrate leaching in the state of Iowa could provide a valuable understanding of the problem and be helpful for decision making.

There needs to be a balance between the output of the manmade applications to increase agricultural benefits vs. the impact on the environment caused by them. The ML model can help predict the future outcome of the nitrate yield and help understand the problem better, as it can provide more details on how the target variable is related to the independent variables. The study's objective was to find the relationships between the target variable, nitrate yield, and the independent variables such as weather, discharge, soil, etc. While the Flow Weighted average Nitrate Concentration (FWNC) is a crucial variable, government agencies are more interested in the nitrate load and insights to reduce the same. Nitrate Yield is defined as nitrate load in corn and soybean planting areas. The feature importance and partial dependency plots can portray a clear picture to understand this relationship of how independent variables affect Nitrate yield. These variables and their effect on the target variable vary geographically.

The outcome of this study would be helpful for Iowa and the neighboring states of the corn belt region, where the weather and other features share some similarities. The same approach with the different datasets can show different outcomes, which would then be helpful for that region. Feature importance and partial dependency plots have been used widely in ecological prediction problems, giving insight into what variables affect the most to the target variable (Shahhosseini et al., 2020; Spijker et al., 2021). USEPA aims to reduce the hypoxic zone in the Gulf of Mexico, which is highly dependent on the annual nitrate losses from the Midwest crop regions. Therefore, insights from this study can help local and regional agencies to monitor the trends of the independent variables and focus their efforts and resources on important tasks.

The remainder of this paper talks about the dataset used for this study and various machine learning methodologies. Then, the model performance and insights generated from the analysis are discussed, along with challenges and future improvements, and the paper concludes with findings.

CHAPTER 2. MATERIALS AND METHODS

The dataset contains long-term water quality data with NO_3-N concentrations and monthly Nitrate load measurements which were taken on the first week of the month. The data was curated by the Iowa Department of Natural Resources Ambient Water Monitoring Program (IDNR 2017-2021). The discharge values were taken from the United States Geological Survey stream gauges (sensors that record the measurements, USGS and University of Iowa). Three target variables were calculated from this monthly and daily data: annual flow-weighted average NO_3--N concentration (FWNC; mg NO_3-N L^{-1}) and annual NO_3-N load (Kg NO_3-N watershed^{-1} year^{-1}). Annual load is affected by the size of watersheds, and therefore normalizing it with area gives the nitrate yield, which is not dependent on the size of watersheds. FWNC was dropped because, while spatial analysis of FWNC gives more information on specific watersheds' performance, nitrate yield is used as the target variable because nitrate load has been found as the leading cause behind the size of the hypoxic zone in the Gulf of Mexico (Jones et al., 2018; Rabalais et al., 2002). The below equation calculates nitrate NO_3-N yield (kg NO_3-N ha^{-1} cropland). Since Corn and Soybean make up most of the cropland area for all watersheds throughout the years, the Nitrate yield was calculated with corn and soybean cropland area.

$$nitrate\ yield = \frac{nitrate\ load}{\frac{(corn\ planting\ area + soybean\ planting\ area)}{100} \times size}$$

Data Set

The dataset contains data for 29 watersheds over 18 years and, therefore a total of 522 observations. The auxiliary data consists of soil parameters, size of the watershed, cropland area, discharge, maize and soy yields, and weather variables. Each watershed is different in size and has different values of the target variables as well as the independent variables. The FWNC value ranges from 0 to 22 mg NO_3-N/l with a mean of 6.8 mg NO_3-N/l, and the yield ranges from 0 to 121.27 kg NO_3-N/ha with a mean of 28.32 kg NO_3-N/ha. The size of the watershed varies from 89 to 20,155 km^2. The data has been collected from the year 2001 to 2018, out of which the year 2001 to 2017 has been used to train the model, and the year 2018 has been kept as a holdout data to test the model's performance.

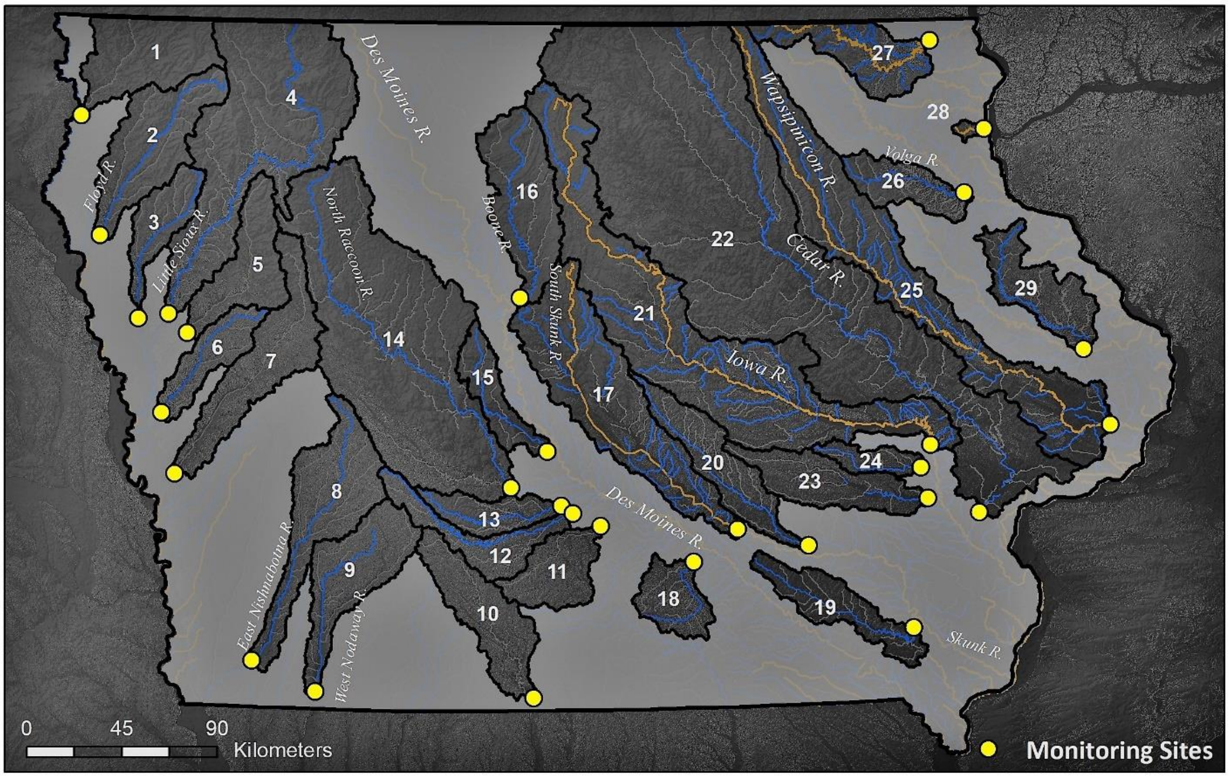

Figure 1: Map of 29 watersheds that are analyzed in this study; the yellow dots show the location of the sample collection site

Data Preprocessing

Data preprocessing is an important task that is performed before training the ML model to reduce complexity. There are 69 independent variables and three interrelated target variables (FWNC, Annual load, and yield). Figure 2 shows that the yield across all the watersheds has increased over time, and a trend seems to be increasing. The year 2012, as seen in figure 2 (b), has extremely low nitrate yield values compared to other years. There was a drought during the year 2012, and therefore the Nitrate yield values for all watersheds were considerably low compared to other years. Since this case is considered an outlier from the ML model training perspective, early model testing presented a high test RRMSE value for 2012. Therefore, the year 2012 is removed from the data set.

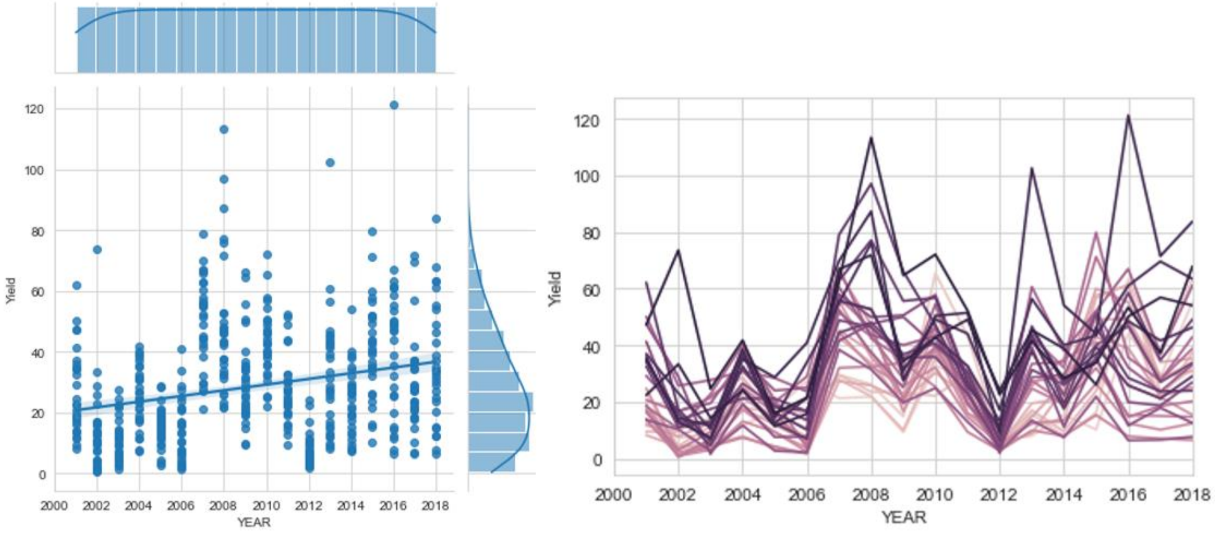

Figure 2: Nitrate Yield trends during the years 2001-2018, (a) trend line for an increase in Nitrate Yield, (b) Watershed wise Nitrate yield per year

The Weather variables, which consist of radiation, rain, and temperature, are provided every month, from which radiation and rain are summed up for the annual value, and temperature is taken as an annual mean value throughout the year. The new feature space also includes average and summed weather variables corn to soybean planting size, shown below.

$$corn\ to\ soybean\ ratio = \frac{corn\ size}{soybean\ size}$$

'Sand' was removed because of the high Pearson correlation with other soil features, as high correlation variables similarly affect the target variable. Thus, the dataset has become denser than the original dataset. For watershed 3, the years 2008, 2009, and 2010 were missing

FWNC and nitrate load values. The missing values were replaced with the mean value for that watershed for training the model across the years. After feature selection and feature construction, there are only 17 features left that are used to simplify the training set for the model to combat the curse of dimensionality, which are as follows: annual discharge, root depth, soil profile, maize yields, manure maize, soy yields, croplands, tile drainage, corn to soybean ratio, annual radiation , rain, average temperature, bulk density, clay, Ksat, silt, Soil organic matter.

Walk Forward Cross-Validation

After data preprocessing, since the data has a form of time series, it does not hold the assumption of being independent and identically distributed (IID). Therefore, traditional cross-validation methods such as K-Fold can no longer be applied for time series analysis. The walk forward cross-validation approach described by (Hyndman & Athanasopoulos, n.d.) works well on time series analysis. The training data is split into parts that roll forward with time, and each particular training set is followed by a validation set which is further ahead in time. This way, the model is no longer accessing the future data, unlike traditional splitting techniques in K-fold. Two methods were introduced in (Hyndman & Athanasopoulos, n.d.) for walk forward cross-validation, sliding window, and expanding window cross-validation. The sliding window approach consists of the same amount of training set each time that rolls forward, whereas in the expanding window approach, the training set size increases after each split. In both approaches, the size of the validation set remains constant, which also rolls forward with the training set.

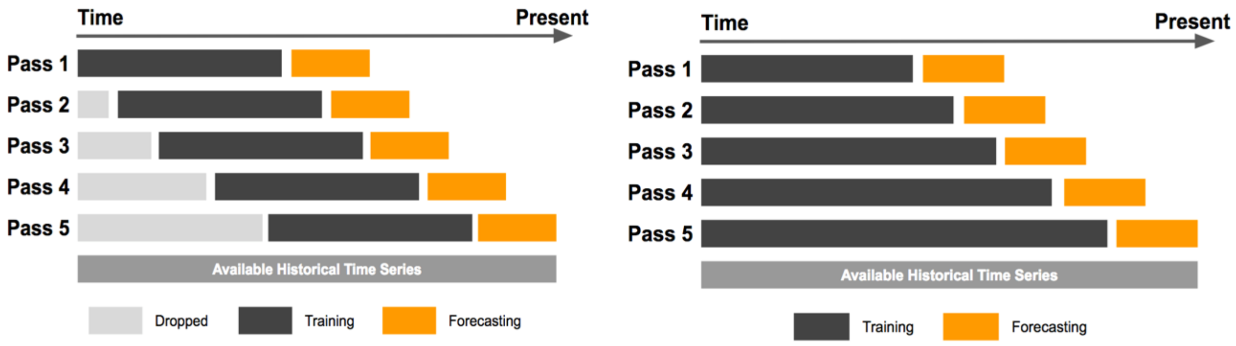

Figure 3: Walk Forward Cross-Validation approach. (a) Sliding Window (b) Expanding window

Machine learning Models

Tree-based ensemble machine learning models such as Random Forest and LightGBM were used as a model in this study as they have been extensively used for solving environmental problems (Chaibi et al., 2021; Hengl et al., 2018; Zhong et al., 2021).

Random Forest

Bootstrap aggregating or bagging is an ensemble technique that reduces the variance and generalizes the model by randomly creating samples of data from the whole dataset with replacement and trains multiple models independently. The final prediction is the average of these learners (Breiman, 1996). Random Forest is a special case of bagging. Multiple individual decision trees are based on random values and the number of predictors taken for split candidates in each iteration (Breiman, 2001). Random Forest uses data points that are excluded from the bootstrapping procedure (out-of-bag observations) to compute errors and therefore performs better than bagging (Cutler et al., 2007).

Light Gradient Boosting (LightGBM)

Gradient boosting is another tree-based ensemble method that combines weak learners iteratively. With each iteration, the model learns from the errors of the previous model and improves. Microsoft proposed LightGBM in 2017; LightGBM is a faster tree-based model, which, unlike other conventional tree-based models, does not grow tree-level wise; instead, it grows leaf wise. (Ke et al., 2017) proposed two new ideas for improved speed and performance: first is gradient-based one side sampling to select data observations with high gradient and exclude the rest, and second, Exclusive Feature Bundling (EFB) to bin mutually exclusive features to reduce the dimension of the data which improves the performance and speed of the model.

Random forest and LightGBM are taken as the choice of models because both perform well with low data and are easy to understand. Random Forest needs minimal hyperparameter tuning and is more robust to overfitting. LightGBM requires more data and needs complex hyperparameter tuning. Although both perform well, LightGBM can overperform Random Forest with proper tuning and with more data. Grid search method of hyperparameter tuning was used under Walk Forward Cross-Validation to reduce the prediction errors such as RRMSE discussed below.

Performance Metrics

Root Mean Squared Error (RMSE)

Root mean squared error (RMSE) is defined as the square root of the average squared deviation of predictions from actual values (*Evaluating Machine Learning Models [Book]*, n.d.).

$$RMSE = \sqrt{\frac{\Sigma_i (y_i - \hat{y}_i)^2}{n}}$$

Relative Root Mean Squared Error (RRMSE)

Relative root means squared error is the RMSE error normalized by the average of actual values and is used as a percentage value where lower RRMSE values are favorable.

$$RRMSE = \frac{RMSE}{\bar{y}}$$

Mean Absolute Error (MAE)

The mean absolute error of a model with respect to a test set is the mean of the absolute values of the individual prediction errors over all instances in the test set (Sammut & Webb, 2010).

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n}$$

Coefficient of Determination R^2 score

Coefficient of determination R^2 is a statistical measure that tells the proportion of variance of the target variable, which is explained by the predictor variables. R^2 explains to what extent the variance of one or more variables can explain the variable of the target variable, and it is a measure of the goodness of fit of a model. It can be defined as below:

$$R^2 = 1 - \frac{RSS}{TSS}$$

Where RSS is the sum of the square of residuals.

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Adjusted R^2 is the modified version of R^2, which has been adjusted for the number of independent variables in the model. Adjusted R^2 gives a percentage of variation that is explained by the independent variables that affect the target variable, whereas R^2 does not accommodate that. Therefore, Adjusted R^2 penalizes if the model uses more independent variables which do not affect the target variable. So for simplifying, the adjusted R^2 score is referred to as R^2 in this study. The formula used is shown below, where n is the number of samples in the data and k is the number of independent variables used.

$$R_{adj}^2 = \left[\frac{(1-R^2)(n-1)}{n-k-1}\right]$$

Mean Directional Accuracy (MDA)

Mean Directional Accuracy (MDA) is a measure that calculates the probability that the prediction model can identify the future direction of the time series (Cicarelli, 1982). MDA coupled with forecast performance metrics such as RRMSE can help compare the performance of the model and understand the capture of trends.

$$MDA = \frac{\sum_t 1_{sgn(y_t - y_{t-1}) == sgn(\hat{y}_t - y_{t-1})}}{N}$$

Where y_t and $\hat{y}_t$ are actual and predicted values respectively at time t, 1 is the indicator function, and sgn(·) denotes the sign function.

CHAPTER 3. RESULTS AND DISCUSSION

Numerical Results

The feature importance and the partial dependency plots (PDP) are used to understand how the various independent variables affect the nitrate yield. The RRMSE value and the adjusted R^2 value suggest the error and the goodness of fit of the model overall. Table 1 shows the results of various scenarios, and the machine learning models were used in two different CV methods discussed in the above chapter.

Table 1: RRMSE and R^2 values of Cross-Validation and optimized ML models

CV method and model used	RRMSE %	Adjusted R^2 %	MAE kg NO_3-N/ha	MDA %
Expanding window RF	29.65	67.46	9.41	75
Expanding window LightGBM	23.68	77.06	7.14	89.28
Sliding window RF	26.71	70.81	8.51	92.85
Sliding window LightGBM	26.66	70.92	7.58	82.14

Sliding Window CV approach

The Sliding window cross-validation has a limited (6 years) training set for each fold, and it tests on the following one-year validation set. The sliding window approach has similar results with RF and LightGBM, with RRMSE values of 26.71% and 26.66%, respectively, and R^2 values of 70.81% and 70.92%, respectively. RF model captures the trend better because the MDA value is 92.95%. Both models perform very similarly because the sliding window approach does not contain enough data for the ML model to learn, as the trend of the whole time-series data is lost because of the truncated training set during each CV fold.

Expanding Window CV approach

In contrast to sliding window CV, the expanding window has a training set that increases as the cross-validation method continue. Therefore with each fold, the model gets more training data. This helps the model to learn underlying trends within the data. RF and lightGBM RRMSE values with expanding window are 29.65% and 23.68%, respectively, while the goodness of fit measure R^2 values were 77.06% and 67.46%. A high value of R^2 with LightGBM suggests that the model is a good fit for the nitrate yield, and the model can explain the variance within the nitrate yield. MDA value of 89.28% suggests that the model can predict the correct direction of the trend 89.28% of the time. LightGBM needs more data than RF and can outperform RF with extensive hyperparameter tuning. The sliding window is often used when there is high-frequency data with hourly and daily data points; expanding window performs well with yearly datasets. Therefore, LightGBM with expanding window cross-validation outperforms Random with sliding window cross-validation approach. The RRMSE and R^2 values show that the model explains a significant portion of the variance in the nitrate yield, and the MAE value shows that the predictions are close, and the model has a large confidence interval on at least the annual level.

Feature Importance based on Expanding Window approach using LightGBM

The feature importance shown below, constructed from the LightGBM model, shows interesting insights into the problem.

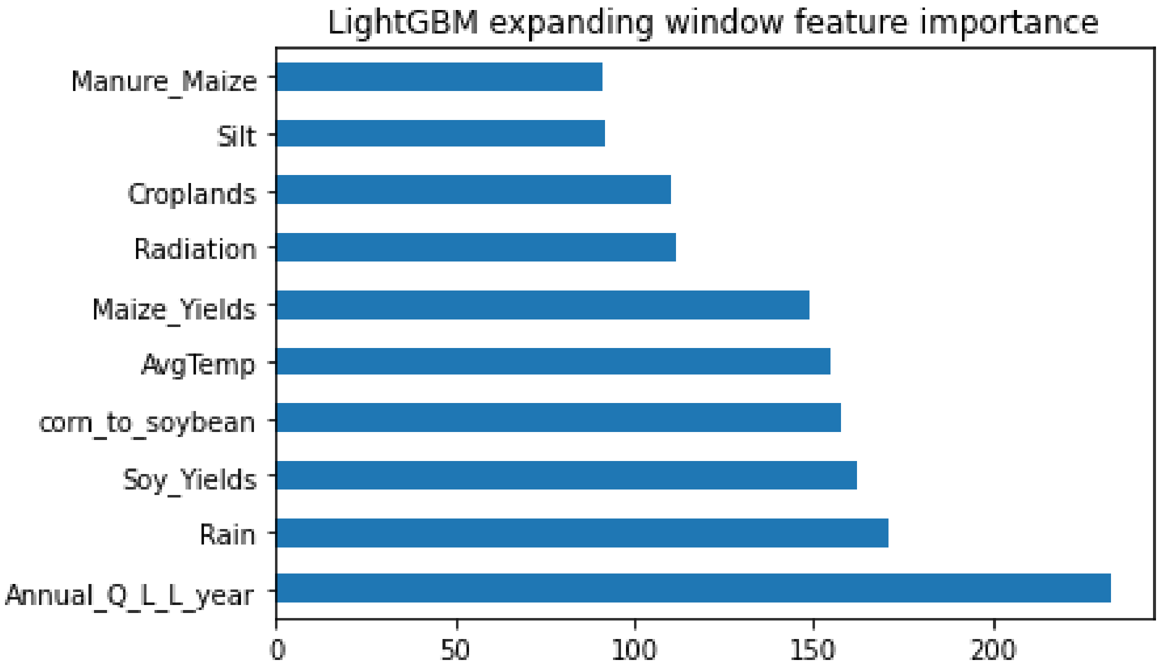

Figure 4: Feature importance constructed from LightGBM expanding window cross-validation approach

The x-axis of the graph above has the numerical value, which shows the relative importance of the feature or variable compared to other variables. The higher the value, the more influential the feature is. In other words, the feature contains more information that explains the relationship with the target variable. The first variable is the discharge in liters and the second variable is the rain in millimeters. The fundamental reason for these two variables being the most important variables for nitrate leaching is that the water carries the nitrate to the ground. The more the water content in the soil, the more nitrate will leach. Next is soybean yield, which is the amount of soy being produced per acre in bushels. The following variable is the corn to soybean ratio, which is the ratio of corn planting area to the soybean planting area. It is important because corn and soybean have different characteristics and require different agricultural methods. This goes directly with the corn to soybean ratio. The weather variables temperature and radiation, along with maize yield or corn yield, come after that, which affects the nitrate leaching. The figure shows the top 10 variables that most affect the nitrate yield.

Partial Dependency Plots (PDP) of Optimized LightGBM

The partial dependency plots are constructed based on the feature importance, and they show further insights from the model, which are beneficial to understanding nitrate leaching. The annual discharge value seems crucial as from the beginning, as annual discharge increases, the nitrate yield increases drastically until reaching the value of around 2.0 x 10^{12} liters, and then it stagnates. Rain is another important variable, and the PDP shows that nitrate yield does not vary until 800 mm of rain but increases sharply around 860 mm and follows a step relationship where it increases and stagnates, and repeats the cycle. The amount of soybean yield also has a step relationship with nitrate yield. Nitrate yield increases and decreases with an increase in soybean yield. The PDP of nitrate yield and corn to soybean ratio is also descriptive. There are multiple steps on which nitrate yield increases sharply with an increase in corn to soybean ratio, specifically at 1.2, 1.6, and 1.7. This shows that the corn planting area is more susceptible to nitrate leaching compared to the soybean planting area, and since the size of the watershed normalizes the nitrate yield, this observation is valid for all watersheds. Average temperature affects the nitrate yield, too, with an average temperature below 8 degrees showing very high nitrate yield and then behaving like a bell curve. This resonates with the literature that nitrate losses are high during winter weather with lower temperature, so annual croplands for corn and soybeans are not cultivated and therefore has a fallow period. There is also some nitrate leaching during spring and early summer, which results from applying N fertilizer followed by above-average rain. Maize yield or corn yield affects the nitrate yield similarly to soybean yield and has a step relationship with nitrate yield. These findings strongly suggest that weather variables can explain the trend in nitrate yield very well. The variables under human control, such as corn to soybean planting area ratio and soy and corn yields, can be optimized to minimize the nitrate losses. The effect of corn to soybean planting area ratio on nitrate yield can be a crucial factor in tackling the challenges of reducing nitrate leaching as a ratio lower than1.2 with adequate weather conditions can significantly drive down the nitrate losses. The annual discharge is another variable that can help reduce nitrate losses. These findings can help in decision-making for local authorities and agricultural professionals.

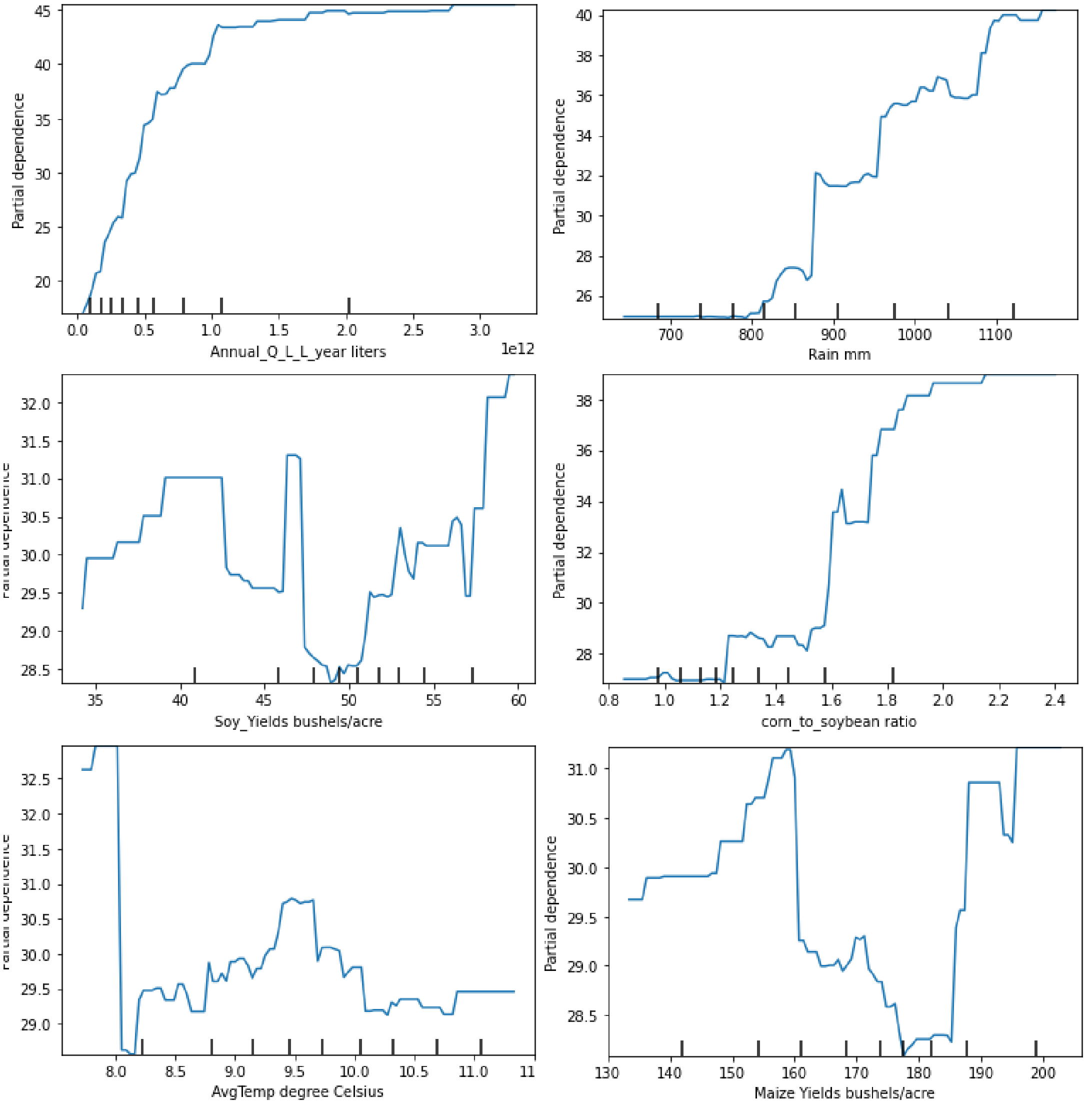

Figure 5: Partial Dependency Plots derived from Expanding Window approach using LightGBM

The challenges faced with this study can suggest future work. The dataset contains scattered daily, and monthly measurements data, but only yearly measurements were used since there is no consistency. In the 13 watersheds with monthly measurements from 2001 to 2008, the installation of automatic sensors allowed daily measurements from 2012 to 2018; the sensors are operated and maintained by the USGS and the University of Iowa. Although these measurements were collected on an irregular basis during winter, the collection point freezes and the sensors are taken out. Machine learning models require detailed and granular data to perform well. Daily or monthly measurements would drastically improve the ML model's performance and reduce errors. Yearly measurements lose some of the temporal details from the data, and the model cannot learn that part during training. Data granularity is of paramount importance for time series analysis in predicting future outcomes. Another important aspect of having detailed data is that more complex ML models can be utilized, such as Neural Network models, particularly Long Short Term Memory (LSTM) Recurrent Neural Network RNN based models. These models require more data. RNN feeds the output back in, and that way, they have a sense of memory by which the layers know and learn from the sequential data. There is solid evidence of a relationship between various weather variables' values at a particular time of a year, and consolidating them together loses the temporal aspect of weather variables as well. Having monthly or daily target variables would enable this approach as well.

CHAPTER 4. CONCLUSION

The nitrate leaching prediction is a complex problem affected by many geographical and agricultural aspects. Adverse weather conditions and aggressive use of animal manure, N fertilizer, and drainage have exacerbated the problem. This study utilized machine learning models to predict future nitrate yield in Iowa and gathered important findings from the analysis that showed different relationships between nitrate yield and other independent variables.

The time-series analysis of the nitrate leaching problem consists of temporal and spatial data, which can train machine learning models to predict the future nitrate yield. This study utilized two tree-based machine learning models, Random Forest and LightGBM, to predict nitrate yield for 2018 and derive insights into the underlying relationship between independent variables and nitrate yield. Since time-series data is not independent and identically distributed, random splitting and traditional cross-validation methods cannot be used. Instead, a new walk forward cross-validation method with two different approaches was used to optimize the model trained on data from 29 watersheds from 2001 to 2017.

The model was tested on the year 2018, and the lowest RRMSE value of 23.68% was achieved using LightGBM with expanding window, walk forward cross-validation. The findings showed that the annual discharge, the annual rain, corn to soybean planting area ratio, soy and corn yield, and average annual temperature affect the nitrate yield the most. The increase in annual discharge seems to increase nitrate yield linearly, whereas rain, corn to soybean ratio, soybean, and corn yield has a step relationship. Controlling annual discharge and corn to soybean planting area ratio can help drive down the nitrate yield. The average annual temperature has a distinct relationship with nitrate yield. Temperatures below 8 degrees Celsius show a very high amount of nitrate leaching and then increase to 9.5 degrees Celsius before decreasing. This decrease in higher annual average temperature shows that nitrate leaching is low during hot weather months. These insights suggest that policymakers and farmers can take steps accordingly to minimize the nitrate yield.

This study is subject to a couple of limitations that can suggest future research directions. First, data granularity is very important, and working with annual data points leads to losing the interannual aspects of independent variables. The interannual variability in independent variables can explain the interannual variability in the target variable nitrate yield (Danalatos et al., 2022). Second, more complicated machine learning and deep learning models can be used with denser data. Monthly or daily data can help the model deep dive, find the convoluted relationships, and show interesting findings. Long Short-Term Memory, RNN based learners, can be used better to understand the underlying dependency between the target variables as they feed the output back into the model to learn and predict the future better. Third, the year 2012, with drought, was removed from training because machine learning models suffer from skewing based on outlier data and cannot understand the natural reasoning behind a drought. More complicated models can accommodate outliers and will not skew during the training process.

REFERENCES

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chaibi, M., Benghoulam, E. M., Tarik, L., Berrada, M., & Hmaidi, A. E. (2021). An Interpretable Machine Learning Model for Daily Global Solar Radiation Prediction. *Energies*, *14*(21), 7367. https://doi.org/10.3390/en14217367

Cicarelli, J. (1982). A new method of evaluating the accuracy of economic forecasts. *Journal of Macroeconomics*, *4*(4), 469–475. https://doi.org/10.1016/0164-0704(82)90065-9

Cutler, D. R., Edwards Jr., T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random Forests for Classification in Ecology. *Ecology*, *88*(11), 2783–2792. https://doi.org/10.1890/07-0539.1

Danalatos, G., Wolter, C., Archontoulis, S., & Castellano, M. (2022). Nitrate losses across 29 Iowa watersheds: Measuring long-term trends in the context of interannual variability. *Journal of Environmental Quality*. https://doi.org/10.1002/jeq2.20349

David, M. B., Drinkwater, L. E., & McIsaac, G. F. (2010). Sources of Nitrate Yields in the Mississippi River Basin. *Journal of Environmental Quality*, *39*(5), 1657–1667. https://doi.org/10.2134/jeq2010.0115

Dybowski, D., Dzierzbicka-Glowacka, L. A., Pietrzak, S., Juszkowska, D., & Puszkarczuk, T. (2020). Estimation of nitrogen leaching load from agricultural fields in the Puck Commune with an interactive calculator. *PeerJ*, *8*, e8899. https://doi.org/10.7717/peerj.8899

Evaluating Machine Learning Models [Book]. (n.d.). Retrieved March 28, 2022, from https://www.oreilly.com/library/view/evaluating-machine-learning/9781492048756/

Gentry, L. E., David, M. B., & McIsaac, G. F. (2014). Variation in Riverine Nitrate Flux and Fall Nitrogen Fertilizer Application in East-Central Illinois. *Journal of Environmental Quality*, *43*(4), 1467–1474. https://doi.org/10.2134/jeq2013.12.0499

Hatfield, J. L., McMullen, L. D., & Jones, C. S. (2009). Nitrate-nitrogen patterns in the Raccoon River Basin related to agricultural practices. *Journal of Soil and Water Conservation*, *64*(3), 190–199. https://doi.org/10.2489/jswc.64.3.190

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

Hyndman, R. J., & Athanasopoulos, G. (n.d.). *Forecasting: Principles and Practice (3rd ed)*. Retrieved March 28, 2022, from https://Otexts.com/fpp3/

Jones, C. S., Nielsen, J. K., Schilling, K. E., & Weber, L. J. (2018). Iowa stream nitrate and the Gulf of Mexico. *PLOS ONE*, *13*(4), e0195930. https://doi.org/10.1371/journal.pone.0195930

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Lu, C., Zhang, J., Tian, H., Crumpton, W. G., Helmers, M. J., Cai, W.-J., Hopkinson, C. S., & Lohrenz, S. E. (2020). Increased extreme precipitation challenges nitrogen load management to the Gulf of Mexico. *Communications Earth & Environment*, *1*(1), 21. https://doi.org/10.1038/s43247-020-00020-7

Martinez-Feria, R. A., Castellano, M. J., Dietzel, R. N., Helmers, M. J., Liebman, M., Huber, I., & Archontoulis, S. V. (2018). Linking crop- and soil-based approaches to evaluate system nitrogen-use efficiency and tradeoffs. *Agriculture, Ecosystems & Environment*, *256*, 131–143. https://doi.org/10.1016/j.agee.2018.01.002

McIsaac, G. F., David, M. B., & Gertner, G. Z. (2016). Illinois River Nitrate-Nitrogen Concentrations and Loads: Long-term Variation and Association with Watershed Nitrogen Inputs. *Journal of Environmental Quality*, *45*(4), 1268–1275. https://doi.org/10.2134/jeq2015.10.0531

Rabalais, N. N., Turner, R. E., & Wiseman, W. J. (2002). Gulf of Mexico Hypoxia, A.K.A. "The Dead Zone." *Annual Review of Ecology and Systematics*, *33*(1), 235–263. https://doi.org/10.1146/annurev.ecolsys.33.010802.150513

Randall, G. W., Huggins, D. R., Russelle, M. P., Fuchs, D. J., Nelson, W. W., & Anderson, J. L. (1997). Nitrate Losses through Subsurface Tile Drainage in Conservation Reserve Program, Alfalfa, and Row Crop Systems. *Journal of Environmental Quality*, *26*(5), 1240–1247. https://doi.org/10.2134/jeq1997.00472425002600050007x

Ransom, K., Nolan, B., Traum, J., Faunt, C., Bell, A., Gronberg, J., Wheeler, D., Rosecrans, C., Jurgens, B., Schwarz, G., Belitz, K., Eberts, S., Kourakos, G., & Harter, T. (2017). A hybrid machine learning model to predict and visualize nitrate concentration throughout the Central Valley aquifer, California, USA. *Science of The Total Environment*, *601–602*, 1160–1172. https://doi.org/10.1016/j.scitotenv.2017.05.192

Sammut, C., & Webb, G. I. (Eds.). (2010). Mean Absolute Error. In *Encyclopedia of Machine Learning* (pp. 652–652). Springer US. https://doi.org/10.1007/978-0-387-30164-8_525

Schilling, K. E., & Wolter, C. F. (2009). Modeling Nitrate-Nitrogen Load Reduction Strategies for the Des Moines River, Iowa Using SWAT. *Environmental Management*, *44*(4), 671–682. https://doi.org/10.1007/s00267-009-9364-y

Shahhosseini, M., Hu, G., & Archontoulis, S. V. (2020). Forecasting Corn Yield With Machine Learning Ensembles. *Frontiers in Plant Science*, *11*. https://www.frontiersin.org/article/10.3389/fpls.2020.01120

Smolders, A. J. P., Lucassen, E. C. H. E. T., Bobbink, R., Roelofs, J. G. M., & Lamers, L. P. M. (2010). How nitrate leaching from agricultural lands provokes phosphate eutrophication in groundwater fed wetlands: The sulphur bridge. *Biogeochemistry*, *98*(1), 1–7. https://doi.org/10.1007/s10533-009-9387-8

Spijker, J., Fraters, D., & Vrijhoef, A. (2021). A machine learning based modelling framework to predict nitrate leaching from agricultural soils across the Netherlands. *Environmental Research Communications*, *3*(4), 045002. https://doi.org/10.1088/2515-7620/abf15f

Zhong, J., Zhang, X., Gui, K., Wang, Y., Che, H., Shen, X., Zhang, L., Zhang, Y., Sun, J., & Zhang, W. (2021). Robust prediction of hourly PM2.5 from meteorological data using LightGBM. *National Science Review*, *8*(10), nwaa307. https://doi.org/10.1093/nsr/nwaa307

APPENDIX: VARIABLES USED IN THE STUDY AND THEIR DESCRIPTION

Table 2: Variables of the dataset with description

Name of the variable	Description and unit
Annual_Q_L_L_year	Annual discharge rate in Liters
AvgTemp	Annual average temperature in degree Celsius
BD	Bulk Density in g/cm^3
Clay	%
Croplands	Amount of cropland of the size of the watershed in %
corn_to_soybean	Ratio of corn plating area to soybean plating area
ID	ID number of watershed
Ksat	micrometers per second
Maize_Yields	bushels/acre
Manure_Maize	kg N per ha
Rain	Annual rain in mm
Radiation	MJ/m^2
Root_Depth	cm
Sand	%
Size	Hectors
Soil_Profile	%
Soy_Yields	bushels/acre
Silt	%
SOM	%

Table 2 continued

Name of the variable	Description and unit
Tile_Drainage	%
YEAR	Year of the datapoint
Yield	Nitrate yield in kg NO_3-N/ha
FWNC	Flow weighted nitrate content in mg NO_3 - N/l
Loads	Nitrate load in kg