Predicting NH4 Levels for Corn Crop in Wisconsin and Michigan

Julia Sochava

Inspirit AI
E-mail: jesochava@gmail.com

Abstract

Ammonium (NH4), an organic matter that accumulates in the top portion of soil, can pose a serious risk to biodiversity. Using machine learning to construct regression models, NH4 levels can be predicted and therefore mitigated. In this paper we used linear, ridge, and lasso regressions. Through the evaluation of crop farming factors that contribute to the NH4 levels, it was concluded that NO3 and N2O have the most direct correlation to NH4. These factors yielded the best accuracy for regression models with the best performing model being a multiple feature linear regression which resulted in 60% accuracy. While certain measures did improve the model's performance, outliers continuously worsened the results.

1. Introduction

NH4, or ammonia, has a significant impact on biodiversity, with certain species particularly falling victim to its pollution. Although nitrogen is the wider known cause of biodiversity loss, ammonia still plays an important role in these changes. While most species only feel consequences when subject to high levels of NH4, lichen and mosses can feel an impact even when low levels of NH4 are present. Biodiversity loss is the main impact of excess NH4, however, it also damages the environment through soil acidification and air pollution. It is important to take action against the impact of NH4 on biodiversity because scientists claim that if its increasing emission continue it could not only significantly damage the environment, but also cost the government approximately $2.50 per kg of ammonia in damage.

Using other factors that correlate to the presence of NH4, a supervised learning model can be constructed to address this issue. Through the use of regressions such as linear, Ridge, and Lasso, along with correlating numerical metrics such as precipitation, N2O, and NO3, the NH4 levels produced by certain crops, in this case corn, can be predicted. Testing the utilization of multiple numerical inputs, to result in the output of future NH4 content, can allow farmers to mitigate potential peaks of NH4. The main sources of NH4 in a farm setting are man made fertilizer and animal manure. Simple tactics such as temporarily switching from urea based fertilizers to ammonium nitrate and washing down animal collection points soon after use, can help lower NH4 levels. Due to the fact that these efforts may be expensive, NH4 prediction can limit their practice to a certain time period when NH4 levels are predicted to be high, instead of full time practice which some farmers may not be able to afford.

2. Materials and Methods

2.1 Dataset

"we got the dataset from this study / university". The crop production dataset used for this research contained string values identifying the experiment type and crop type, and numerical values such as date, N2O flux, fertilization rate, cumulative precipitation, daily average temperature, NH4 content, and NO3 content. During the data preprocessing this larger dataset was split up based on the experiment type resulting in three separate datasets: MCSE, BCSE, and Arlington. While MCSE and BCSE both contain data on the crops from Michigan, Arlington contains data on crops from Wisonsin. From there data for just the corn crop was isolated, narrowing down the scope of the experiment. For each dataset all null values were replaced with the average value for the data in that specific column. For example if there was a null value in the NO3 column for the MCSE dataset, the average of all NO3 values in the MCSE dataset would take its place. After the data was processed, the usage of the data was to be determined. Initial analysis consisted of visualizing each of the metrics over time for all three data sets(figures 1,2,3).

Figure 1
BCSE data

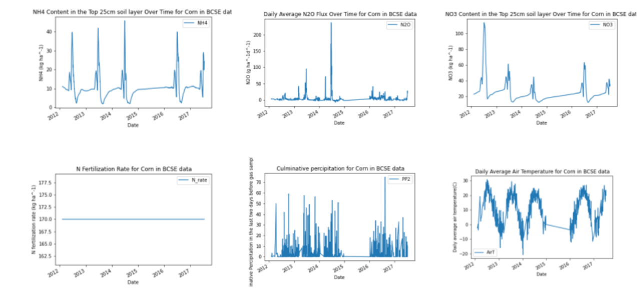


Figure 2
MCSE data

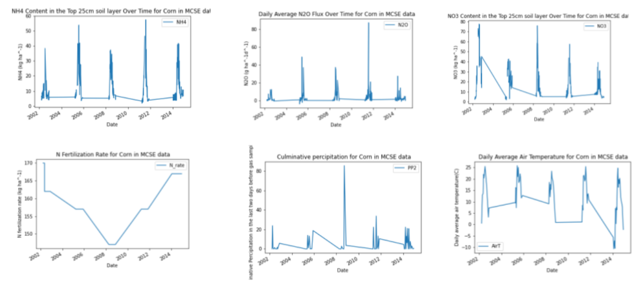
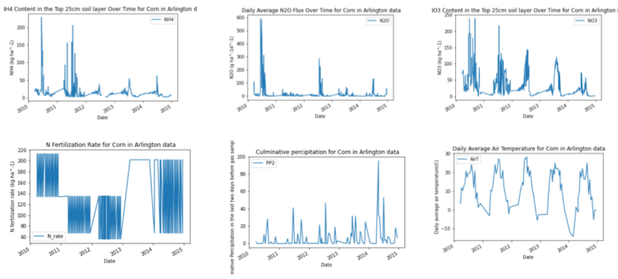


Figure 3
Arlington data



After observing the general trends such as strong correlation between N2O, NO3, and NH4, along with a slight correlation between precipitation and NH4 amongst all data sets, feature importance analysis was used to confirm these correlations on the Ridge model(figures 1,2,3).

Figure 1
BCSE data

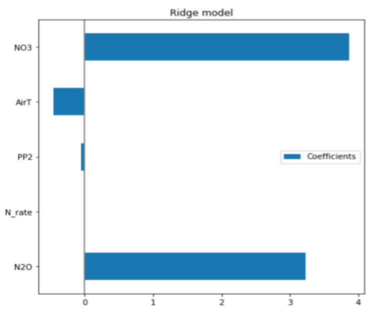


Figure 2
MCSE data

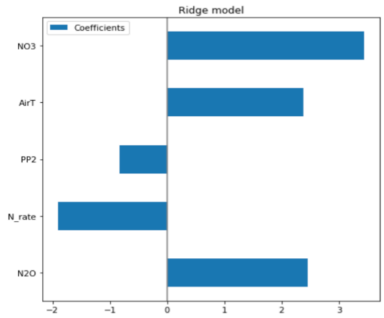
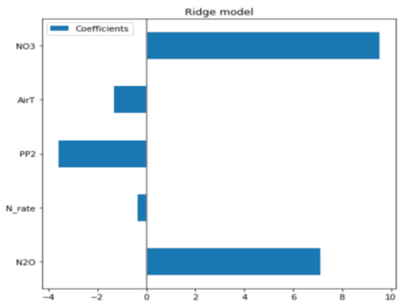


Figure 3
Arlington data



The feature importance analysis showed that NO3 is the most important feature amongst all data sets while fertilization rate and precipitation rate alternate being the least important feature amongst the datasets. Generally N2O is the second most important metric, but there is an exception for this observation for the MCSE data where air temperature is the second most important feature. After this analysis, NO3, N2O and NO3, and NO3 and precipitation, were the chosen inputs for the three regression models.

2.2 Models

With these three different datasets, three types of regressions were tested: linear, ridge, and lasso. Linear regressions are used to predict the value of a variable from the value of another variable. This is the simplest type of regression, compared to the other two. Lasso regressions are similar to linear regressions, except they use shrinkage. Shrinkage is where data values are shrunk towards a central point, like a mean. Ridge regressions are the most unique of the three. Ridge regressions analyze data that is multicollinear, meaning, there is a near linear relationship among multiple variables.

Each of these models were tested with three different sets of inputs: NO3 on its own, N2O and NO3, and NO3 and precipitation. Two thirds of the data was assigned for training and one third for testing. The models were scored using the metrics mean absolute error, mean squared error, and r squared. Mean absolute error takes the absolute error and sums it over all samples, mean squared error which takes the average squared error between predicted and true values, and r squared which takes the square root of the average of the squared difference between the predicted and true values.

3. Results

The model metrics in figures 1, 2, and 3 utilize mean absolute error, mean squared error, and r squared to evaluate the three different models with three different inputs each.

Figure 1
BCSE data

	Linear			Ridge			Lasso		
	NO_3	PP_2 & NO_3	NO_3 & NO_2	NO_3	PP_2 & NO_3	NO_3 & NO_2	NO_3	PP_2 & NO_3	NO_3 & NO_2
MAE	4.247668319012461	4.385531372358979	4.432900261239335	3.80818816560248 14	3.94581730269582 03	4.014001609452263	4.060106686055081	3.92380832224980 75	4.015610306830419
MSE	41.39628504812839	43.113415339905 79	41.2082111411671 55	35.776899977205 29	37.37173214023651	35.851736538726 69	50.4936488688576 26	37.2793823415418 05	36.0570736019360 9
R2	0.40060785467134 53	0.37574488911294 646	0.40333104643705 02	0.39057321061825 945	0.36340670247087 636	0.38929843819679 71	0.22451661177205 606	0.36497979687968 274	0.38580070901253 27

Figure 2
MCSE data

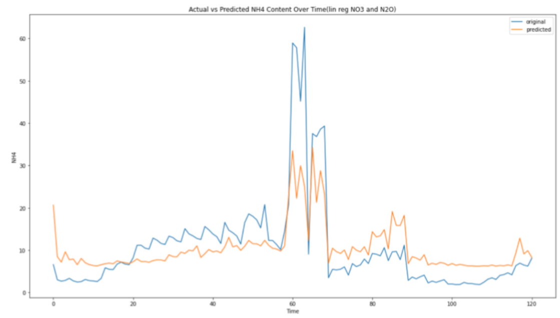
	Linear			Ridge			Lasso		
	NO_3	PP_2 & NO_3	NO_3 & NO_2	NO_3	PP_2 & NO_3	NO_3 & NO_2	NO_3	PP_2 & NO_3	NO_3 & NO_2
MAE	5.723436273414893	5.704445050780953	5.617732392985875	5.619411515783022	5.602562911131387	5.441115102347055	7.167511370636922	5.61500187242517	5.452848240856091
MSE	79.00229244872064	78.27575127712855	76.72139446474279	69.0314573902426	68.44372617604478	66.59692778499797	105.845330262992 62	68.674582964216 21	66.651681689112 69
R2	0.12638283480302 026	0.13441701734445 932	0.15160528809017 815	0.19415851186943 756	0.20101941578448 224	0.22257809091580 927	0.23934391815636 813	0.19832449980330 225	0.22193891901880 97

Figure 3
Arlington data

	Linear			Ridge			Lasso		
	NO_3	PP_2 & NO_3	NO_3 & NO_2	NO_3	PP_2 & NO_3	NO_3 & NO_2	NO_3	PP_2 & NO_3	NO_3 & NO_2
MAE	5.211582874349267	6.699889372886755	4.916597586612884	7.212859532351411	8.5826256033647	6.55535092863323 05	7.167511370636922	5.61500187242517	6.554511656235995
MSE	64.21147254956647	98.115787129422 24	51.9120068895263	110.494524213788 28	148.591337870554 87	85.667185368487 11	105.845330262992 62	68.674582964216 21	85.670372807742 21
R2	0.48176760528144 236	0.20813559781662 372	0.58103306812908 28	0.11724503311124 135	-0.18711531159758343	0.31559383669462 227	0.23934391815636 813	0.19832449980330 225	0.31556837183241 204

The best overall accuracy came from the multiple feature linear regression model from the Arlington dataset that had NO3 and N2O as its inputs. This model had an r squared value of 0.5810330681290828 and closely aligned predicted and true values(figure 4).

Figure 4
Arlington data



In the MCSE dataset the most accurate model was the Lasso regression with just NO3, with an r squared of 0.23934391815636813. This best performing model is an outlier out of the three models because similar to the Arlington dataset, the BCSE dataset's best performing model was the multiple feature linear regression with NO3 and N2O. This model had an r squared of 0.4033310464370502. While the pattern of the predicted values for these models relatively aligns with the pattern of the true values, the outliers significantly distort the accuracy (figures 5,6).

Figure 5
MCSE data

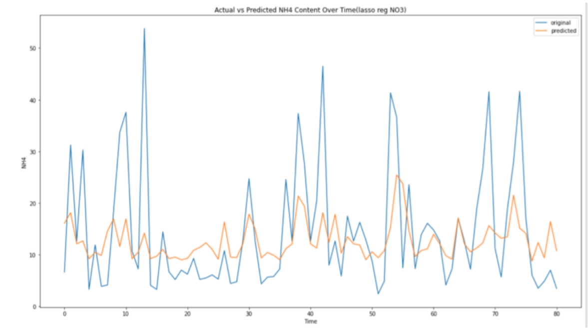
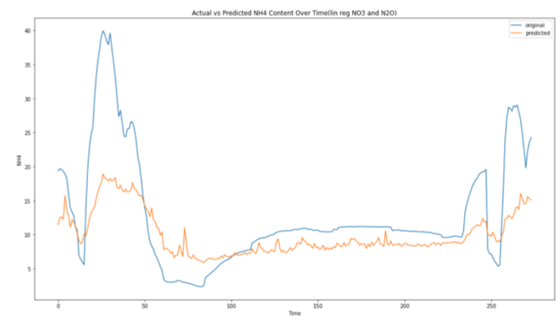
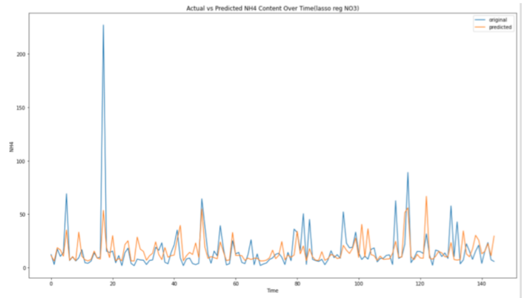


Figure 6
BCSE data



Throughout all of the models, the low accuracy is consistently attributed to the outliers in the data. Many of the worst performing models have closely aligned graphs of the predicted vs actual data. The main issue is the magnitude of some of the peaks which the model does not pick up on, throwing off the predicted values. For example, the lasso regression model that had NO3 as its input with Arlington data had a low r squared of 0.23934391815636813, but the graph of actual vs predicted data aligns very closely as seen in figure 7.

Figure 7
Arlington data



4. Conclusion

Overall, multiple linear regressions with N2O and NO3 as the input features performed the best. However, it is difficult to predict NH4 to prevent dangerously high peaks becasue often the extreme peaks are the outliers which the model does not pick up on. Although general patterns may be accurate, the specific purpose for which this experiment was designed to use them for is not fulfilled. Future research should divert more attention to studying the outliers of the data. Hypertuning regression models to put a heavier weight on the outliers would be a reasonable next step. This research has demonstrated that predicting ammonia levels in order to foresee dangerous peaks may be more difficult that initially imagined because many peaks do not align with the general trends of the rest of the data and are hard to predict.

Acknowledgements

Thank you to Inspirit AI for providing me with the opporutnituty to do this research and Barbie Duckworth for teaching, supporting, and helping me along the way.

References

[1] "About Linear Regression." *IBM*, https://www.ibm.com/topics/linear-regression#:~:text=Resources-,What%20is%20linear%20regression%3F,is%20called%20the%20independent%20variable.

[2] "Agricultural Ammonia Emissions Carry Steep Costs." *RAND Corporation*, https://www.rand.org/randeurope/research/projects/impact-of-ammonia-emissions-on-biodiversity.html.

[3] *Best Management Practices for Reducing Ammonia Emissions - 1.631*. https://extension.colostate.edu/topic-areas/agriculture/best-management-practices-for-reducing-ammonia-emissions-1-631/.

[4] "Ecological Effects of Ammonia." *Minnesota Department of Agriculture*, https://www.mda.state.mn.us/ecological-effects-ammonia.

[5] *An Evidence Synthesis - Royal Society*. https://royalsociety.org/~/media/policy/projects/evidence-synthesis/Ammonia/Ammonia-report.pdf.

[6] Garre, Alberto, et al. "Application of Machine Learning to Support Production Planning of a Food Industry in the Context of Waste Generation under Uncertainty." *Operations Research Perspectives*, Elsevier, 22 Feb. 2020, https://www.sciencedirect.com/science/article/pii/S2214716019301988.

[7] *Machine Learning Improves Predictions of Agricultural Nitrous Oxide ...* https://iopscience.iop.org/article/10.1088/1748-9326/abd2f3.

[8] "Reduce Ammonia Emissions on Your Farm." – *CFE Online*, https://www.cfeonline.org.uk/environmental-management/reduce-ammonia-emissions-on-your-farm/.

[9] Saha, Debasish, et al. "Data from: Machine Learning Improves Predictions of Agricultural Nitrous Oxide (N2O) Emissions from Intensively Managed Cropping Systems." *Dryad Data -- Machine Learning Improves Predictions of Agricultural Nitrous Oxide (N2O) Emissions from Intensively Managed Cropping Systems*, Dryad, https://datadryad.org/stash/dataset/doi:10.5061/dryad.bnzs7h493.

[10] Stephanie. "Lasso Regression: Simple Definition." *Statistics How To*, 27 Apr. 2021, https://www.statisticshowto.com/lasso-regression/.

[11] Team, Great Learning. "What Is Ridge Regression?" *GreatLearning Blog: Free Resources What Matters to Shape Your Career!*, 22 Mar. 2022, https://www.mygreatlearning.com/blog/what-is-ridge-regression/#:~:text=Ridge%20regression%20is%20a%20model,away%20from%20the%20actual%20values.