SML310 Final Project Report

Demetra Yancopoulos

Due: 10 December 2021

A Comparison of Several Multivariate Linear and Polynomial Regressions for Modelling Wheat Yield in Minnesota and Iowa

I. Motivation

In this report, I will examine the utility of multivariate linear and polynomial regressions as a means of predicting wheat yield in two midwestern states: Minnesota, and Iowa. Wheat is an essential component of food and feed products. In the United States, wheat is the main food grain, ranking third in planted acreage and production among field crops [6]. As population increases, and average incomes rise globally, demand for food and feed crops like wheat will continue to rise. By 2050, the global population is estimated to reach about 10 billion people, and total food demand is expected to increase by greater than 50% [5]. So, for the world to prepare to meet future food demand, we must be able to accurately predict future crop yields. By parametrizing crop yields in terms of meteorological parameters, we can use climate projections to predict food production in various future scenarios.

II. Introduction

Wheat yield is dependent on several inputs: sunlight, water, nutrients, and heat. The proper amount of each of these factors facilitates plant growth. While this system may seem simple at a first glance, it is deceptively complex. Plants can be incredibly sensitive to variations in these parameters: for instance, sunlight is necessary for plants to photosynthesize, but too much sunlight can cause plants to dry out, diminishing yields. Plants also require different conditions at different stages in their life cycle: for example, winter wheat, planted in the fall, needs to mature to the point of growing shoots before cold weather sets in, or else it will be stunted by low winter temperatures [2]. So, to accurately predict wheat yield, it is important to

consider the magnitude of factors such as sunlight, water, nutrients, and temperature as a function of time.

For my final project, I have used historical data on meteorological parameters and wheat yields in Minnesota and Iowa to train and compare several multivariate regression models of wheat yield. I implemented a linear regression, as well as a second degree polynomial regression. I tested several methods for feature selection: a manual feature selection involving a p-value significance and a variance inflation factor, a variance threshold, a recursive feature elimination, and a sequential feature elimination. I determined the variance threshold to be the most effective method of feature selection. I also found that, while the polynomial regression tended to perform better than the linear regression, linear regression was competitive with polynomial regression when feature selection parameters were properly tuned.

III. Data collection

Accurate and high resolution historical observations of various meteorological parameters is provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) in the ERA5 dataset [3]. The current release of the ERA5 data ranges from 1979 to 2021, and is available on an hourly basis. From the ERA5 dataset, I retrieved temperature, precipitation, mean surface downward UV radiation flux, and relative humidity. I accessed this data through the Princeton tiger cpu (/tigress/wenchang/data/era5). This data was preprocessed by a research scholar, Wenchang Yang, in The Vecchi Research Group in the Department of Geosciences at Princeton University to generate the following datasets: daily mean 2m temperature, daily maximum 2m temperature, daily minimum 2m temperature, monthly mean total precipitation rate, daily mean surface downward uv radiation flux, monthly mean 2m

relative humidity. For a sample plot of some monthly metrics of 2m temperature data from ERA5, see Appendix A, Figure 1.

Detailed data on crop yield with high spatial and temporal resolution was a bit more difficult to find. I was hoping to find a seasonal or monthly record of crop yields, but most accessible data logged only annual production. Through gro-intelligence, I was able to find a dataset of annual wheat yield in the contiguous United States, amalgamated by district from USDA NASS Crop data [7]. For a sample plot of the wheat yield data, see Appendix A, Figure 2.

For a full description of all datasets in their original form, see Appendix A, Table 1.

IV. Data preprocessing

Surprisingly, the most intensive part of my project was preprocessing the data. This was because of the differing formats between the yield data and the ERA5 data. First, I dealt with preprocessing the wheat yield data. Originally, I had hoped to use the entire wheat yield dataset, which included data across the entire contiguous US. Unfortunately, data availability limited my ability to do so: the locations for each data point in the set are described by county names. I was not able to access the latitude and longitude coordinates of the county names through gro-intelligence, as it requires a costly subscription to the site. As a result, I had to manually input latitude and longitude coordinates for each location in the dataset. In order to make this task manageable, I decided to limit my analysis to only two states: Iowa and Minnesota. I retrieved every data point within Iowa and Minnesota and extracted a list of unique county names across all data points. I then found the latitude and longitude of the center of each county using google maps, and paired these to the list. I added columns for latitude and longitude to the yield data, and iterated through every point in Iowa and Minnesota to add the proper latitude and longitude coordinates.

Next, I needed to extract the ERA5 parameters at all the same locations and times as the wheat yield data. ERA5 data is retrieved by using latitude and longitude coordinates rounded to the nearest 0.25 degree, and a date. So, for every data point in the wheat yield set, I extracted the latitude and longitude, and rounded them to the nearest 0.25 degree. I also extracted the year. I then retrieved the value of the parameter from the ERA5 dataset for each month in the year. As I retrieved the ERA5 data, I performed some operations on the parameters to get features that I believed would have an influence on crop yield. For each iteration, I added a row to a dataframe containing the latitude, longitude, year and feature value for each month in that year. I ultimately stored these dataframes as excel spreadsheets for later use.

From the ERA5 parameters in Appendix A, Table 1, I generated several features, all on a monthly basis. Because crops grow over an entire season, I felt that metrics based only on daily or weekly information would not be very representative of the conditions that the wheat is exposed to over its life cycle, and thus would not be effective predictors of yield. On the other hand, metrics based on temporal scales larger than one month may mask important weather events by averaging over meteorological parameters. So, the goldilocks of timescales seemed to me to be one month. From the ERA5 data, I originally generated 8 different parameters. Among these are mean 2m temperature - the daily 2m temperature averaged over the course of each month, and maximum 2m temperature - the maximum daily 2m temperature occurring during each month. I next decided to generate metrics for the anomalies of each one of these seven parameters. I hoped that by including anomalies as features, rather than just the absolute data, I would be able to account for regional farming practices that cope with local weather conditions: for instance, a region that receives very little rainfall is likely to have developed more extensive systems for irrigation, whereas a region that receives excess rainfall may have better

infrastructure for drainage. Since farmers in each location are equipped to manage crops in their own unique region and climate, I believe that absolute magnitudes of meteorological parameters may in some cases be less indicative of yield than the deviation of those meteorological parameters from normal. So, I calculated the expected value of each of the 8 features for every month at each location by averaging all monthly values of that feature at that location. For each data point, I then subtracted the expected value from the observed value to get the feature anomaly. For a complete description of all features extracted from the ERA5 data, see Appendix A, Table 2.

V. Exploratory data analysis on generated features

To get an idea of how each of the features I generated from the ERA5 data may impact crop yields, I created plots of wheat yield versus each feature. For each row in Appendix A, Table 2, I generated a pairplot displaying that parameter for each month of the year - so, each pairplot displays 12 unique features. All pairplots are attached in Appendix B. Here I will discuss my observations of all the features as they relate to wheat yield:

Mean 2m temperature:

Almost all of the mean monthly temperatures seem to have the same relationship to the yields; I expect that there is multicollinearity across mean temperature for several months. The relationship between mean monthly temperatures and yield is not very distinct for most of the months of the year. It seems relatively random except for the highest yields, which all coincide with relatively high mean monthly temperatures. The month with the most significant relationship to yields seems to be October: as mean temperature in October increases, yield increases.

Maximum 2m temperature:

Similar to mean monthly temperature, higher maximum monthly temperatures tend to be associated with higher yields. Again, I suspect that there is multicollinearity across many of these features. By inspection, January maximum monthly temperature seems to have the most distinct relationship with yields. April, May, October, November, and December also show a similar relationship.

Minimum 2m temperature:

I am not convinced that any of these features have a significant relationship to the yield. There does seem to be a slight positive correlation between minimum monthly temperature and yields, but these plots are even messier than those for mean and maximum monthly temperature. April and October could potentially be useful in predicting yields.

Mean total precipitation rate:

There is not a consistent trend across all months between monthly mean total precipitation and yield. It is interesting to note that in several months (July, September, November) there is a much larger range of yields at lower monthly mean total precipitation than at higher monthly mean total precipitation. It is hard to tell whether this is a sample bias (ie. fewer samples at higher monthly precip), or a legitimate physical phenomenon. Physically, it could mean that:

- When total precipitation is low either: (i) precipitation is not the main control on yields (ie. other parameters contributing to yield are able to induce larger fluctuations in yield) or (ii) crops are more sensitive to the amount of precipitation.
- When total precipitation is high either : (i) precipitation becomes the main control on yields (ie. other parameters contributing to yields are not able to

induce large fluctuations in yield), or (ii) crops are less sensitive to the amount of precipitation

Mean msdwuvrf:

In late spring to summer (April - July) there is a very subtle curvature in the plots. This suggests that there is an intermediary value of UV flux that is optimal for plant growth. Although this isn't too surprising (too little sunlight will limit photosynthesis, too much might cause a plant to dry out), it's super exciting to actually see! In December, there appears to be a positive correlation between mean UV flux and yield.

Maximum msdwuvrf:

Monthly max msdwuvrf displays behavior reminiscent of that in the monthly mean total precipitation rate: in certain months, there is a noticeable change in the range of yields as the max msdwuvrf changes. For instance, in March and April, as monthly maximum msdwuvrf increases, the range of yields decreases. In May, June, July, September, October, as maximum msdwuvrf increases, the range of yields increases. Again, this could be due to a sampling bias, or to some physical phenomenon.

Minimum msdwuvrf:

Monthly minimum msdwuvrf also displays behavior reminiscent of that in the monthly mean total precipitation rate and monthly maximum msdwuvrf. In January, March, September, November, December, as monthly maximum msdwuvrf increases, the range of yields decreases.

Mean relative humidity:

It is very difficult to distinguish a relationship between yield and monthly mean relative humidity based on the pairplot. It seems possible that there is a second

degree relationship between yield and mean relative humidity for some months (April - July), but I don't feel certain about it.

Mean 2m temperature anomaly:

It is difficult to identify a clear relationship between yield and mean monthly temperature anomaly for most of the months. The two parameters seem fairly independent. There may be a slight negative correlation between yield and mean temperature anomaly for June. There may be a slight positive correlation between yield and mean temperature anomaly for January and October.

Maximum 2m temperature anomaly:

There appears to be a slight positive correlation between yield and maximum temperature anomaly for April, October, and November. There appears to be a slight curvature in the plots of yield and maximum monthly temperature anomaly for June, July, and August. Physically, this suggests that in summer months there is an optimal intermediary temperature to maximize crop yield. When I first saw this, I thought it may be indicative of multicollinearity between the mean msdwuvrf and the maximum temperature anomaly. However, the apparent second degree relationship in the plots here occurs in the summer months, whereas for the monthly mean msdwuvrf it occurs in late spring.

Minimum 2m temperature anomaly:

Overall, it seems that the minimum temperature anomaly for each month does not have a substantial effect on yield. Yield and minimum temperature anomaly for June appear to have a second degree relationship. There appears to be a positive correlation between yield and minimum temperature anomaly for October.

Mean total precipitation rate anomaly:

The mean total precipitation rate anomaly exhibits similar behavior to the mean total precipitation rate: the range of yields tends to decrease with increasing precipitation rate anomaly (see February, April, June, July, September, October, November, December). However, there is no obvious relationship between the magnitude of the yield and the magnitude of the anomaly.

Mean msdwuvrf anomaly:

I cannot identify a clear relationship between the magnitude of the yield and the monthly mean msdwuvrf anomaly in most of the months. It is possible that mean msdwuvrf for May, June, July, and October has a second degree relationship with yield. However, it is pretty messy, and I don't feel especially convinced about it.

Maximum msdwuvrf anomaly:

Like for mean total precipitation rate, there appears to be a relationship between the variability in yield and this feature. For instance, in May, as maximum msdwuvrf anomaly increases, the range of yields increases. However, there is not any distinct trend between the yield and the maximum msdwuvrf anomaly for any of the months.

Minimum msdwuvrf anomaly:

The range of yields appears to be dependent on the magnitude of the maximum msdwuvrf anomaly for the months of January, November, and December.

Mean relative humidity anomaly:

Like for mean relative humidity, it is difficult to distinguish a relationship between the yield and the mean relative humidity anomaly here. There may be a second degree relationship between yields and mean relative humidity anomaly for the summer months, but it is not so clear.

From inspection of each of the potential features, I narrowed my list of 192 features down to 62. I used only these 62 features while training my models. For a complete list of these 62 features, see Appendix B, Table 1.

VI. Initializing the models

Machine learning models have proven successful for predicting crop yields. Everything from neural networks to stepwise linear regression has been explored [1]. So, I decided it would be appropriate to explore the use of regression models for predicting wheat yield. In my exploratory data analysis, I saw some features that appeared to have a linear relationship to yield, like mean 2m temperature in October, and others which seemed to have a second degree polynomial relationship to yield, like mean surface downward UV radiation flux in late spring and summer months. As such, it seemed reasonable to implement and compare a linear regression and a degree 2 polynomial regression.

Given the large number of features I was left with after my exploratory data analysis, I also found myself interested in how different feature selection methods would impact the accuracy of the model. While using a lot of features may improve performance metrics of the model such as mean squared error, using too many features comes at a cost: not only can it be difficult to collect all the proper data for model inputs, but it can require substantial memory to store the data and significant computational power to process it. So, I decided to test four unique feature selection methods. The most crude feature selection method I implemented was a variance threshold. A variance threshold removes those features which have a target variable variance below some threshold. The reasoning is sound in the sense that features which are not associated with large changes in the target variable are excluded. However, the method does not

consider any relationship between feature and target variables, and can therefore disregard features who have a high correlation to the target variable but affect it on a smaller scale.

The most intensive feature selection method I performed was a manual recursive selection process. Although there are modules in scikit-learn which execute recursive feature selection, I figured it would be good practice to develop my own implementation. For this feature selection process, I initialized a regression model on the training data, and retrieved the feature with the highest p-value. If the highest p-value was above some threshold, I removed the feature, and generated a new regression with the remaining features. I continued doing this until all remaining p-values were below the set threshold. Since the p-value represents the statistical significance of the relationship between the feature and the target, removing features with a high p-value filters out those features with statistically insignificant relationships with the target variable. Afterwards, I executed another recursive selection based on variance inflation factor, dropping the feature with the highest variance inflation factor until all were below some threshold. The variance inflation factor of a given feature is the ratio of the total model variance to the variance of a model with only that feature. Multicollinear features exhibit high variance inflation factors, since the features are giving feedback to one another in real life, amplifying their influence on the target variable. This inflation makes it more difficult to isolate the relationship between the different features and the target variable. Thus, removing multicollinear features from the model can improve the quality of the regression.

The other two feature selection methods I implemented were provided by ski-kit learn. I used recursive feature elimination with a linear regression estimator, and used a backwards sequential feature selector also with a linear regression estimator.

For a complete description of the 6 models I implemented, see Appendix C, Table 1.

VII. Model training

To generate my final models, I split my data into a training, validation and test set. I used 25 percent of the original data for testing, and 15 percent of the remaining training data for validation. For each of the 6 models, I executed 3 to 6 runs with varying input parameters. In each case, I used the training set to perform the feature selection and finalize the regression. I then assessed the performance of the model on the validation set using the coefficient of determination and mean squared error, and adjusted the input parameters to try to maximize the coefficient of determination and to minimize the mean squared error as well as the number of features used. For a complete table of model inputs and performance metrics, see Appendix C, Table 1.

In general, as the number of features increased, the mean squared error decreased, and the coefficient of determination increased (see Appendix C, Figure 1). While it may be tempting to include the highest possible number of features in the model to minimize error, the improvements made per feature added plateaued after about 50 total features, rendering it less worthwhile to include added features given the costs of data collection, storage, and processing.

The polynomial regression models had more final features at comparable MSE values than the linear regression models. This makes sense given that the number of features grows in a polynomial regression model due to incorporation of higher order terms. So, for the same dataset, generation of polynomial features creates more inputs for the model. Given this, the cost of using more features in a polynomial model is not as damaging in terms of data collection, but can still slow computation. So, I find that having a high number of features in the polynomial models is less deterring than having the same number of features in a linear regression would be.

For model E, the sequential feature selection followed by polynomial regression, the model seemed to be overfitting the training set. For all runs, the MSE on the training set was between 0.01 and 0.05 lower than the MSE on the validation set. Furthermore, the R2 score for the training set was notably higher than the R2 score for the validation set, indicating an overfit model. In contrast, the polynomial regression model F did not exhibit as large of a disparity between the MSE and the R2 score on the training set and validation set. The overfitting in model E is likely due to the high number of features: the number of final features in model E ranged between 36 and 325, while that for model F did not exceed 60. This shows that the overfitting caused by the increased number of features with polynomial regressions can be tempered by an adequate feature selection method.

In terms of the feature selection methods themselves, my manual recursive selection was the most intensive and least rewarding. For 26 features, the MSE on the validation set was 0.29. Similarly, the variance threshold on the linear regression did not perform very well, returning MSE on the validation set of 0.31 for 29 features. In contrast, for the recursive feature elimination in model C, only 20 features returned a MSE of 0.27, a clear improvement upon the selection in models A and B. The backwards sequential feature selection in model D performed even better than the recursive feature elimination on the validation set, returning MSE 0.03 lower than that for the recursive feature elimination with 40 features.

VIII. Model testing

Based on the performance of the trained models on the validation set for varying inputs, I chose finalized input parameters for each model A-F. This returned 6 models for me to use with the test set. For details on the 6 models and their performance metrics, see Appendix C, Figure 1. Models B1 and F3 returned the best performance metrics. This is particularly interesting given

the fact that both of these models used a variance threshold feature selection, which performed worse during validation. While the simplicity of model B1 is appealing, I am skeptical that it would perform as well on other data. I would therefore be more inclined to rely on model F3 in practice. However, the risk of overfitting given a second order polynomial regression should be noted. As such, I would be most likely to choose model C3 or D5, which had comparable performance metrics to F3 for testing, to predict wheat yield in Iowa and Minnesota.

Overall, it seems that model performance is more sensitive to feature selection than to the order of the regression for prediction of wheat yields.

IX. Improving the model

The major challenge in predicting wheat yields does not come from the complexity of yield response to any single variable. It is not especially difficult to predict, in an isolated system, how plant growth may be influenced by perturbation of any single meteorological parameter. The difficulty in predicting wheat yields ultimately comes from the massive number of inputs which can affect yield, and the sensitivity of yields to local conditions. It follows that better data is the most important means of improving crop yield models.

For many meteorological parameters, data exists at high resolution spatial and temporal scales. For instance, the ERA5 data provides hourly estimates for 0.25 degree grids. However, certain meteorological parameters are not accessible at high resolutions and accuracy: soil moisture varies severely at small spatial scales, and can thus be difficult to incorporate into a large scale model. Improving resolution of other meteorological parameters will thus be important for refining predictions of crop yield in the future.

Yield data can be improved by increasing temporal resolution. Most large scale yield data exists only annually. Monthly or seasonal historical yield data could improve models by enabling

us to isolate the specific crops affected by weather and climate events within certain parts of the year. For instance, the wheat yield dataset contained information on various types of wheat, some of which were planted in the fall, and others in the spring. As a result, variations in yield of one type of wheat may be hidden due to different total annual yield trends, making it difficult to parametrize yields in terms of meteorological parameters.

If I had more time, I would have liked to investigate some more complex features derived from the ERA5 data. For example, while I identified several metrics for temperature within each month, temperature related events defined not by there magnitude but by their duration may be more significant: for example, if a day with high temperature is bounded by other days of high temperature, it is likely to have a more severe influence on yields than a high temperature day that stands alone. Identifying something like a heatwave as a feature could therefore be very useful for predicting yields.

X. Conclusions

Regression models are a fairly effective means of predicting wheat yield in Minnesota and Iowa. Increasing the order of the regression does not create substantial improvements in the model performance, and risks overfitting. Therefore, using linear regression to predict wheat yields is preferred. The more important aspect of an effective yield model is feature selection: choosing the right features can substantially reduce the number of features required to effectively predict yield. Given the performance of all 6 final models on the test set, model C3, a linear regression with recursive feature elimination, and model D5, a linear regression with backwards sequential feature selection, seem to be the most appropriate models for predicting wheat yield.