Quant Challenge

Qualifier task

Team Members of ProgramRs

Eperjesi-Kovács Ádám

Péter Milán

CORVINUS UNIVERSITY OF BUDAPEST

Structure

Executive summary

Our new project as a Risk Analyst team was to support an American credit institution - focusing on agricultural investments - in their loan risk assessment by predicting the yield of agricultural crops in the state of Minnesota, with respect to the effects of climate change. After analysing the provided data, we aggregated the weather data to better match the crop data and then assigned the historic weather information to the crop data. Then we cleaned the data from missing values and outliers. After preparing the data, we created several predicting models where our strategy was to create models with increasing levels of complexity. Finally, we evaluated our models and selected a transformed linear multivariate OLS regression to use for making our final prediction.

Description of the Challenge

We are a team of Risk Analysts supporting an American credit institution that is active in the market of loans given to agricultural businesses. Our new project is to assess the climate risk of the bank's investments in agriculture, specifically its loans to farmers in the state of Minnesota. Farmers are expected to repay their loans from the potential profit they achieve from their businesses. A significant drop in productivity endangers the existence of the farmers' business, therefore, leads to a higher chance for loan defaults,

and thus a larger expected loss for the bank. Specifically, the bank is worried about misestimating future crop yields on land in Minnesota due to the impact of climate change on agriculture. Therefore, the bank wants us to build a model that predicts the potential loss of productivity on plots of land that were purchased with loans from the bank.
We are given a database of historical crop yields, weather data, as well as future scenarios of climate change, to be used for forecasting. Our task is to infer from these data how different circumstances influenced the productivity of farmland in Minnesota in the past and build an engine that can predict future crop yields for a given future climate trajectory. A few example climate trajectories and random forecasts are enclosed in the database. The data sets contain data about 3 different crops: corn, oat, and soybean. We start with corn, and then we turn to oats and soybean.

Workflow

For our work to be more clear and more transparent, we summarised the steps we took while solving the task. By outlining each step of the process, we can give the reader a better understanding of how we arrived at our conclusions and help the reader follow along with our thought process. The eight main steps are the following:

1. •Understanding the data
2. •Programming environment and preparations
3. •Matching the county with the nearest station
4. •Aggregating the weather data
   - yearly mean of temperatures, and sum of precipitation
5. •Filtering agricultural production
   - firstly to corn, then soybean and oat
6. •Binding the agriculture data with the weather data
   - including year, county, crop, yield, temperatures, precipitation
7. •Conducting OLS regression
   - respond variable: yield, explanatory variables: temperatures, precipitation
8. •Interpreting our model

Minnesota

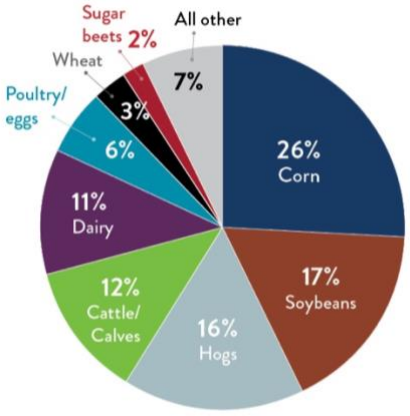
Minnesota is a state in the Upper Midwestern region of the United States. It is the 12th largest U.S. state in area and the 22nd most populous, with over 5.75 million residents. Minnesota is home to western prairies, now given over to intensive agriculture; deciduous forests in the southeast, now partially cleared, farmed, and settled; and the less populated North Woods, used for mining, forestry, and recreation.

Minnesota's leading food products sector sprouted in the state because the industry is rooted in its agricultural history. From farm to table, Minnesota is a food production and agriculture powerhouse. Agricultural production and processing industries generate over $112 billion annually in total economic impact and support more than 431,000 jobs.

Its agricultural economy

- Generated about $17 billion in agricultural sales in 2020.
    - The highest-valued commodities included **corn**, **soybeans**, and hogs.
- Exports about $7.1 billion worth of goods annually.
- Ranks 5th in crops ($8.85 billion) and 5th in total agricultural production ($16.7 billion).
- Ranks 4th compared to other states in terms of total agricultural exports.

Agricultural Products Produced in Minnesota



Source: Minnesota Department of Agriculture

Data

One way to classify the provided databases is by dividing them into two groups: agricultural and weather data. The first and main file in the agricultural data is *minnesota_county_yearly_agricultural_production.csv* which includes years from 1950 to 2022, counties in the state of Minnesota, the produced commodity and its crop subcategory, the acres harvested, production measured in bu. (bushel), and the quotient of the last two: yield measured in bu./acre. The second, geospatial database file is *minnesota_county_location.csv* which contains all the counties in Minnesota, their

capital, and their latitude and longitude. The other group that involves weather data has several files. The one that lists stations capable of collecting weather information is called *Minnesota Station location list.csv*. This is also a geospatial database with the individual code of the stations and their exact locations. There is a folder (*minnesota_daily*) that contains the files of all stations. They provide the measured weather data separately and includes the following variables: a daily date with various domain, the average, minimum, and maximum air temperature °C, and the daily precipitation total in mm. Finally, there is a folder (*prediction_targets_daily*) with similar data but entirely different names, used for prediction at the end of our task.

Preparations

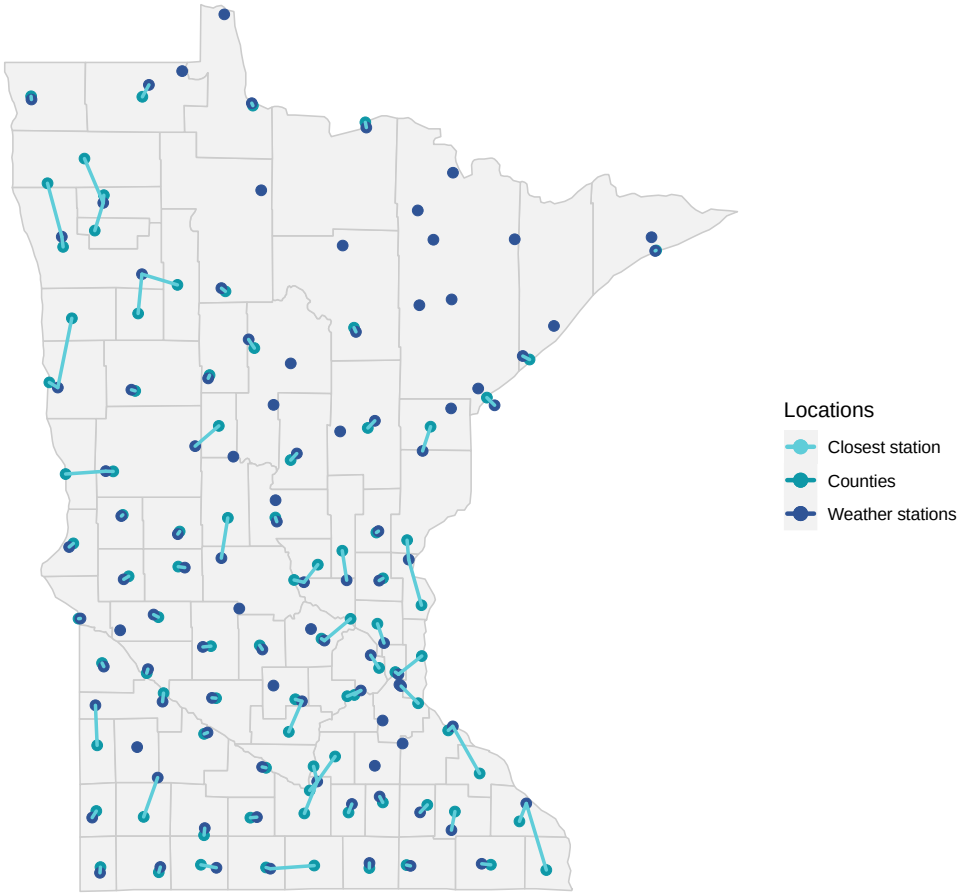
We work extensively with the R programming language for this data analysis and statistical modelling task. R is an open-source software that enables us to analyse large datasets, build predictive models, and visualize complex data patterns efficiently and accurately. With its extensive range of packages and libraries, R allows us to perform advanced statistical techniques and machine learning algorithms. Its reproducible workflow also allows us to document our analyses and share our code including our comments. We marked the different sections in our code with Roman numerals for clarity.

In the **I. Section** (Pre-Settings) the whole coding environment is cleared, and the working directory is set. Individual setup will be required for this. Then in the **II. Section** (Libraries) the packages and libraries are loaded in. After that, the main database – *minnesota_county_yearly_agricultural_production.csv* – is read in the **III. Section**, with small modifications such as renaming its columns.

Locations

In the **IV. Section**, we work with the two geospatial databases: *Minnesota Station location list.csv* and *minnesota_county_location.csv*. The database modifications included renaming the counties so they can be paired and removing insignificant columns that do not carry relevant information. Our task with these databases is to look up and match the counties with the nearest weather station. We completed this task by measuring the distances between all locations from both groups and selecting the shortest one. This

resulted that every county has a closest station but not every station was assigned to a county. Furthermore, there are stations assigned to more than one county.



Weather

In the following **V. Section**, we start to work with files containing all the data about the measured weather – temperatures and precipitations. For computing optimising reasons, we read and load only the involved files, i.e., the ones that were allocated to a county. The data is stored in a three-dimensional array where the first dimension is the given station, the second one is the variables (date, average-, minimum-, maximum temperature, and precipitation) and the third one is the observations.

Then the daily observations must be aggregated. It is because its frequency (daily) does not align with the main file's frequency (yearly). This will be done in two ways: in the case of temperatures, we aggregate them by taking average, and in the case of precipitation, we aggregate by taking sum. So, we produce yearly average temperatures and annual precipitation amount.

Another issue to consider is data cleanliness. While working with the precipitation data, we noticed that in some cases it is heavily incomplete. It generated outliers during the aggregation process, so it had to be filtered. To summarize the data, we only included observations from years with more available data than missing values. This way we were able to eliminate too small, aggregated values that would have been biased our results.

Crops

Our last data processing section is the **VI. Section** in which we assemble the final working databases. The original *minnesota_county_yearly_agricultural_production.csv* file is transformed by filtering by commodity. Firstly CORN, then OATS, and SOYBEAN lastly. In the case of corn, we must filter even for the crop variable, because only CORN, GRAIN ones are necessary for us. Then we bind the stations for the database according to the counties we looked up earlier. We omit the observations where stations are not available for any reason. This is needed for the next step, for matching the stations and years with the aggregated weather data. Finally, we exclude observations where yields are missing or have the value N/A. It is not time-series data because the past yields do not affect the future yields – but we handle it by involving years as a variable. Our task is not to predict the future yields from its past values but from another exogenous factor such as temperatures and precipitation. Consequently, we have the final database including the following variables (the significant ones are highlighted):

Year	County	Commodity	Crop	Harvested	Production	Yield	Station	Latitude	Longitude	Average temperature (tavg)	Minimum temperature (tmin)	Maximum temperature (tmax)	Precipitation (prcp)

Crop	Variables	Observations	Cleaned observations
Corn	14	4994	227
Oats		4671	210
Soybean		4755	211

Models

When it comes to model creation, we built linear multivariate OLS regression models. The table below shows the OLS regression equation of the models we evaluated in the **VII. Section** of the script:

ID	Model	Adjusted R^2
M1	$Yield = tavg + tmin + tmax + prcp$	0.6044
M2	$Yield = Year + tavg + tmin + tmax + prcp$	0.6581
M3	$Yield = Year + tavg + tmin + tmax + prcp^2$	0.6577
M4	$Yield = Year + tavg + tmin + tmax + ln(prcp)$	0.6587
M5	$Yield = Year + tavg + tmin + tmax + prcp + Latitude + Longitude$	0.7028
M6	$Ln(Yield) = Year + tavg + tmin + tmax + prcp$	0.6582
M7	$Ln(Yield) = Year + tavg + tmin + tmax + ln(prcp)$	0.6583

M1 was our first and simplest model where we utilized all the available weather information without any specific transformation. And then on, we gradually increased the complexity of our models. Our model-building strategy was to create increasingly complex models so that we would be able to spot overfitting more easily. For the same reason, we were constantly monitoring the adjusted R^2 values of our models.

In addition to the weather variables, we decided to include the *Year* as a feature variable as well since we observed better prediction results with the *Year* being included. There could be two practical explanations for its inclusion. On one hand, including the year could also indirectly contain other pieces of information that could influence production yield, for example, the improvement of GMO and plant breeding technology over the years. On the other hand, it might indirectly capture the effect of climate change as well.

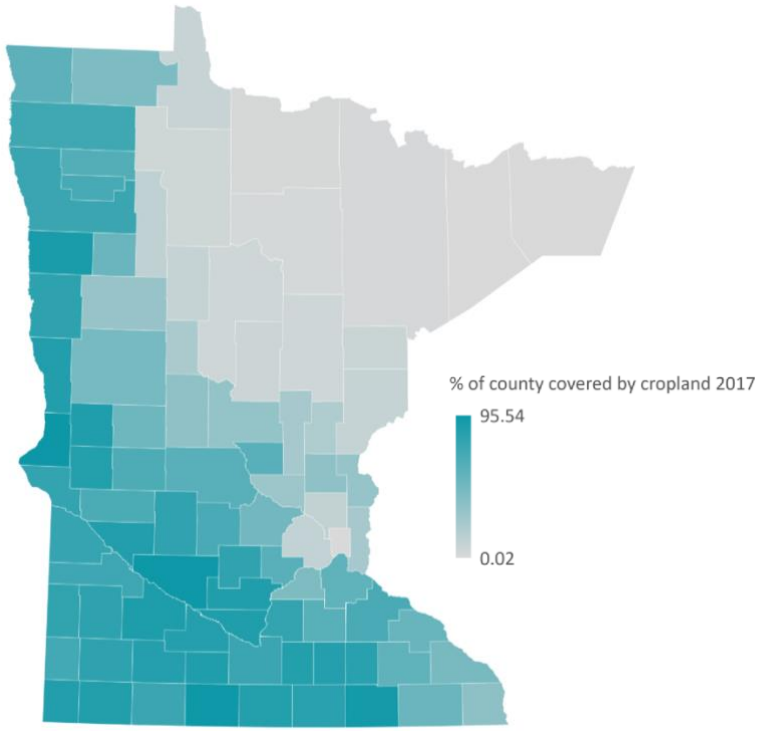
Results

For evaluating our models, we randomly divided our corn data into an 80% work and 20% test data set. Then we performed a $k = 5$-*fold* cross-validation across our work data set, and we got the following RMSE (Root Mean Square Error) values.

Resample	M1	M2	M3	M4	M5	M6	M7
Fold1	28.38314	26.31437	26.31183	26.51518	25.36090	0.231727	0.232796
Fold2	28.38875	25.94965	25.86528	26.13680	24.71830	0.190439	0.191395
Fold3	26.30605	25.43843	25.47930	25.62480	22.31291	0.207183	0.210292
Fold4	25.65210	22.33225	22.20455	22.06907	22.30680	0.194424	0.193693
Fold5	26.21818	25.09792	25.14213	25.11283	23.24929	0.247181	0.247010
RMSE	**27.01465**	**25.06623**	**25.04272**	**25.14166**	**23.6227**	**0.215308**	**0.216139**

We can observe that from the models where the *Yield* was not transformed (M1-M5) the fifth model had the lowest RMSE value, meaning that it had the best-predicting power from the evaluated models. It is worth noting that M6 and M7 cannot be directly compared with the rest of the models since their target variable (*Yield)* was transformed to its natural logarithmic value, therefore the RMSE values are in *ln(Yield)*. In this case, we can only compare these two models relative to each other. It seems that among the logarithmic models, M6 performed slightly better than M7.

Even though model M5 was the best-performing level model, we cannot use it for our final predictions given that the prediction data sets do not contain latitude and longitude values. In the case of Minnesota, the location of the farms matters, because as we already partially discussed it in the Minnesota section of this document, the Southern and the Western parts of the state are more heavily focused on agricultural activity than the rest of the state, thus yields are generally higher in those regions. It can be seen on the attached map. This shows us that in the future it would make the predictions more accurate if the weather predictions included geospatial data as well.

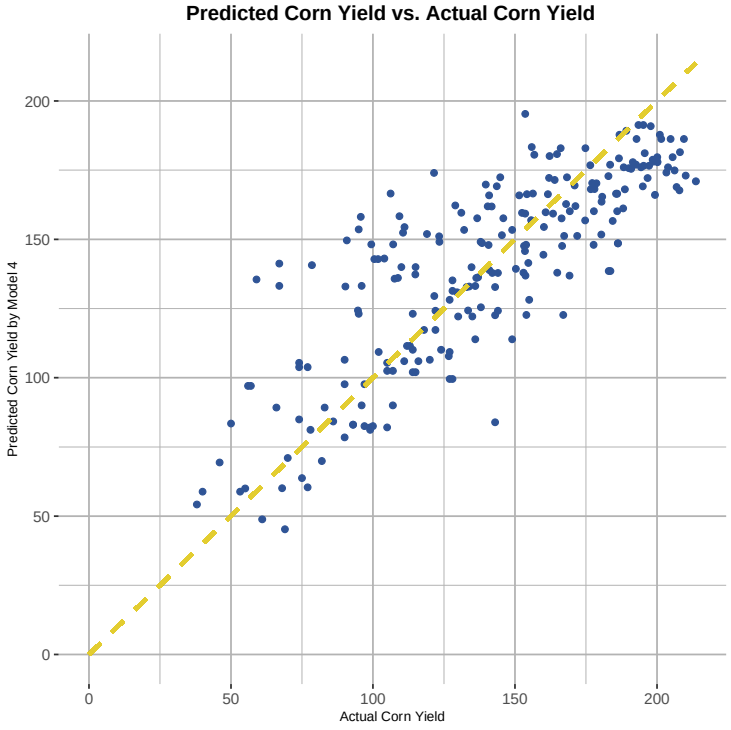


Source: Census of Agriculture, USDA

The next best-performing model was model M3, nevertheless in the end we did not choose it for the actual prediction because, on the test set, it was not the best-performing model. Instead, we selected model M4.

	M1	M2	M3	M4	M5	M6	M7
Test RMSE	26.600259	24.205948	24.224449	24.088656	22.575098	24.496748	24.447635

A chart depicting the actual yield and predicted yield of corn can provide insights into the accuracy of the prediction model, trend analysis, seasonal variations, and model validation. In the case of M4, the predicted yield closely aligns with the actual yield, thus it can indicate that the model is reliable and can be used for decision-making purposes.



Predictions

In the last **VIII. Section**, we can finally train our models on the given datasets. We can make our prediction not only on the data related to corn but even on oats and soybeans. For this, we load all files from the folder *prediction_targets_daily* excluding the ones that contained no data. Then we aggregated the daily weather data in the same way we did in the **V. Section**, then transformed it in the required format. The exported final *csv* file contains the following variables: Target location, Year, Crop, and Predicted yield (BU/acre).

References

Food & Agriculture (2023) Minnesota Department of Employment and Economic Development, Available at: https://mn.gov/deed/joinusmn/key-industries/food-agriculture/ Accessed at: 2023.04.04

Where cropland dominates the landscape: Census of Agriculture, USDA Available at: https://www.mprnews.org/story/2019/04/11/ag-census-2017-minnsota-snapshot Accessed at: 2023.04.06