CSC110 Project Proposal

Yanke Mao, Ziyi Yang, Jiaxu Li, Dakai Hu

November 6, 2020

1. Description

The problem with global food production is serious nowadays. The research reported that it is necessary to increase global food production by 60% by 2050 to satisfy the rising needs of food because of the increasing population and diet changing (OECD and Food and Agriculture Organization of the United Nations, 2012). However, the Fifth Assessment Report (AR5) of the Intergovernmental Panel on Climate Change (IPCC) shows that because of the impact of climate change, from 1960 to 2013, the yields of majors crops trended downwards, and several studies from the report also predicted that there will be large declines in crop yields globally in future, especially in the end of the 21st century (IPCC, 2013). Therefore, it is essential to take action to increase food production.

One of the effective ways to increase food production is to avoid the negative impact on agriculture from climate change. To do this, it is necessary to accurately predict the influences the climate change brings to agriculture, since only with this prediction can the government release suitable mitigation and adaption policy in time, and the farmers can change their planting method to match the climate to reach higher yields. And how to predict crop yields from climate is the question our project is going to address.

As machine-learning is one of the best tools to predict crop yields, we consider using the machine-learning method to perform predictions in our project. With the machine-learning technique developing considerably in recent years, it becomes a reliable and important tool in agriculture. And several machine learning algorithms are applied to predict crop yields for their outstanding accuracy. For instance, the research shows that Artificial Neural Networks (ANN) are commonly used for crop yield prediction (Klompenburg, Kassahun, & Catal, 2020). And Andrew Crane-Droesch, a scientist from Economic Research Service, United States Department of Agriculture, developed semiparametric neural networks (SNN), which is an approach for augmenting parametric statistical models with deep neural networks, to predict the crop yields and the climate change assessment (Crane-Droesch, 2018).

Generally, our project is to construct, train, and test a machine-learning model through the scikit-learn library in python to predict the average yield per hectare for spring wheat in Alberta, with the climate data input. And we are going to obtain the data from websites and compare example predictions with true yields through simple visualization by using the matplotlib library.

2. Datasets

Since we are studying the relationship between climate change and the yield of crops and try to predict the possible production of crops through the model. We have found three datasets for this task: the first one is about the yearly yield of spring wheat in Alberta; the second one is the monthly climate data reported in Alberta; and the last one is about daily climate data reported in Alberta. All datasets are in the *.csv* form. The two datasets about climate can be found on `climate.weather.gc.ca`. The dataset about agriculture can be found on `open.canada.ca`.

- For example, in a case which we collect the data of the production of spring wheat and climate data in Alberta.

    a. The dataset of the monthly climate condition would include 5 columns: time(in terms of year-month, e.g.2007-05), mean temperature(in °C, e.g.3.1), mean max temperature(in °C, e.g.16.3), mean min temperature(in °C, e.g.3.1), and total precipitation(in millimeter, e.g.40.7).

    b. The dataset of the daily climate condition would include 4 columns: date(in the form of year-month-day, e.g.2007-05-01), max temp(in °C, e.g.12.2), min temp(in °C, e.g.-3.4), and mean temp(in °C, e.g. 4.4).

    c. The dataset of agricultural production would include 4 columns: year(in year, e.g.2007), location(e.g.Alberta), value of production(in kilograms per hectare, e.g.200,000), and the name of the crop(e.g. spring wheat).

- To take extreme weather conditions into account, we would first process the daily climate data to count frost days(in days, e.g.4), and summer days(in days, e.g.1). They are defined based on the max and min temperature of the day. Then we will add these two variables to the monthly data and process the monthly data to get to the annual level so that it can match the agricultural data.

- In the sample data, we have given three of the raw datasets and a processed monthly data with extreme weather conditions.

3. Computational Plan

Extreme weather events have a big influence on the production of crops. For instance, a study shows that drastically temperature decrease will cause the production of barley to drop up to 62 percent. Even the least affected species sorghum will still have a chance to lose 21 percent of their normal production. (Paul et al. 2013). And because of this, there are non-linear and threshold-type relationships between crop yields and climate data (Schlenker & Roberts, 2009; Troy et al., 2015). Therefore, in our project, we decide to choose a non-linear regression method in machine-learning to construct our model.

The Random Forest regression, which is a unique Classification and Regression Trees (Breiman et al. 1984), has shown the best fit and least error on the prediction of crop yields based on the climate data among various nonlinear regression models (Konduri, Vandol, S. Ganguly, & R. Ganguly, 2020). Thus, we plan to utilize the Random Forest regression model from the scikit-learn library to perform the prediction. The general process of this method of prediction is divided into two parts, creating a random forest and predict the random forest. When creating a random forest, we need to input the data from the dataset (.csv for example). Then, we need to choose a threshold. Before this threshold is the Training data. After that threshold is Test data. Then, we will calculate the square roots and the average value of inputs and classify the optimal segregation. As a result, our regression forest is created by creating a number of trees. After creating the forest, we are now going to predict by using regression. First, we predict single samples. Then, we can use the predicted single samples to predict the single trees. Finally, the predicted single trees help us to predict the random forest. (Breiman et al. 1984 & Breiman 2004) That is the process of Random Forest Regression, and We plan to use 70% of our data to train our model, and the rest 30% is the data for testing.

In our project, we considered the mean weather indices as predictors, like growing season-averaged maximum and minimum temperature and growing season-averaged precipitation. We also considered the extreme weather indices defined by the CCI/CLIVAR/JCOMM Expert Team on Climate Change Detection and Indices (ETCCDI) (Karl et al., 1999) as other predictors.

After constructing, training, testing the model, and adjusting parameters, our final goal is to predict the production of spring wheat, which is one of the main crops in Alberta, based on the climate data in Alberta through our model.

Lastly, we are going to perform a series of example predictions based on the climate data from the testing dataset, and visualize the prediction results as well as the true production value to perform evaluation and validation of experiment results and its accuracy intuitively and visually. To achieve this, we take advantage of several related libraries in Python: matplotlib and matplotlib's pyplot module.

4. References

1. Andrzej B., Andrzej J. 2010. *Life Time of Correlations and its Applications*. Wydawnictwo Niezależne. pp. 5–21. ISBN 9788391527290.

2. Breiman, L. 2001. *Random forests. Mach. Learn*. 45, 5–32. https://doi.org/10.1023/A:1017934522171

3. Breiman, L., Friedman, J., Olshen, R., and Stone, C. 1984. *Classification and Regression Trees*. Dordrecht: Taylor & Francis.

4. Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.)

5. Crane-Droesch, A. (2018). *Machine learning methods for crop yield prediction and climate change impact assessment in agriculture*. https://doi.org/10.1088/1748-9326/aae159

6. Government of Canada. (n.d.). Station results - historical data. Environment and Natural Resources. Retrieved November 5, 2020, from `https://climate.weather.gc.ca/historical_data/search_historic_data_stations_e.html?searchType=stnProv&timeframe=1&lstProvince=AB&optLimit=yearRange&StartYear=1840&EndYear=2020&Year=2020&Month=11&Day=4&selRowPerPage=25`

7. IPCC (2013). *Summary for Policymakers, Book Section SPM*. Cambridge; New York, NY: Cambridge University Press, 1–30.

8. Karl, T. R., Nicholls, N., and Ghazi, A. (eds.). (1999). *"Clivar/GCOS/WMO workshop on indices and indicators for climate extremes workshop summary," in Weather and Climate Extremes (Dordrecht: Springer), 3–7*. https://doi.org/10.1007/978-94-015-9265-9

9. Klompenburg, T.V., Kassahun, A., & Catal, C. (2020). *Crop yield prediction using machine learning: A systematic literature review*. https://doi.org/10.1016/j.compag.2020.105709

10. Konduri, V. S., Vandal, T. J., Ganguly, S., Ganguly, A. R. (2020). *Data Science for Weather Impacts on Crop Yield*. https://doi.org/10.3389/fsufs.2020.00052

11. OECD and Food and Agriculture Organization of the United Nations (2012). *OECD-FAO Agricultural Outlook 2012*. 286. https://doi.org/10.1787/agr_outlook-2012-en

12. Online documentation and tutorials for scikit-learn library. https://scikit-learn.org/stable/index.html.

13. Paul E., Nicholas Y., Jonathan B. 2013. *How will climate change spatially affect agriculture production in Ethiopia? Case studies of important cereal crops*. Climate Change 119(3-4): 855-873

14. Schlenker, W., and Roberts, M. J. (2009). *Nonlinear temperature effects indicate severe damages to US crop yields under climate change*. https://doi.org/10.1073/pnas.0906865106

15. Statistics Canada. (2020, October 16). Estimated areas, yield, production, average farm price and total farm value of principal field crops, in metric and imperial units. Open Government.
`https://open.canada.ca/data/en/dataset/25b1d384-882d-4aef-949c-68e8c038bf8b`

16. Troy, T., Kipgen, C., and Pal, I. (2015). *The impact of climate extremes and irrigation on US crop yields*. https://doi.org/10.1088/1748-9326/10/5/054013

17. Online documentation and tutorials for matplotlib.
https://matplotlib.org/3.3.2/tutorials/introductory/sampleplots.htmlsphx-glr-tutorials-introductory-sample-plots-py, https://matplotlib.org/3.3.2/tutorials/introductory/pyplot.htmlsphx-glr-tutorials-introductory-pyplot-py