Crop Yield Prediction using Supervised Statistical and Machine Learning Techniques

Prajna Kandarpa
Mechanical and Mechatronics Engineering
University of Waterloo
Waterloo, Canada
Email: spspkand@uwaterloo.ca

Abstract—**Neural Networks and Multiple Adaptive Regressive Splines (MARS), were used to predict crop yield for cereals for the country of India. Historical annual cereal yield data from 1960-2010 for the country of India was used as target data with seasonal mean temperatures, precipitation and arable land size as predictors. The two models were trained with this data and their predictions were compared with the actual yields as well as the performance of more complex crop models.**

I. Introduction

Climate change has brought about increasingly chaotic weather patterns all over the planet making national economic and resource planning difficult for researchers, independent contractors and government organizations. These entities are therefore paying increased attention to the crop forecasts of major food-exporting countries as well as to their own domestic food production. Given the increased volatility of food markets and the rising incidence of climatic extremes affecting food production, food price spikes may increase in prevalence in future years [1]. As such, complex models that can simulate ensemble effects from climate and soil conditions while accounting for geographical variability have assumed paramount importance. Crop growth models perform an abstraction of the dynamic mechanistic of the plants physiological stages by fitting them into a mathematical model [2].

Recent research in this area has sought to build statistical models that perform just as well at forecasting yields as the complex models. Rauff (2015) [3] talk about how crop models can be used to understand the effects of climate change such as elevated nitrogen levels, CO2 levels, temperature and precipitation changes on crop development, growth and yield. For example, sudden onset of warm and humid conditions can lead to plant disease proliferation and cause crop yield loss and financial setbacks for farmers and geographical regions. It is a difficult task to produce a comprehensible, operational representation of a part of reality, which grasps the essential elements and mechanisms of that real world system and even more demanding, when the complex systems encountered in environmental management [3].

A few kinds of crop models have been developed so far, ranging from empirical models to explanatory models. A widely used approach to this prediction problem is to rely on numerical models that emulate the main processes of crop growth and development. Lobell et al (2010)[4] state that these models are typically developed and tested using experimental trials and thus offer the distinct advantage of leveraging decades of research on crop physiology and reproduction, agronomy, and soil science, among other disciplines. Yet these models also require extensive input data on cultivar (plant varieties produced by selective breeding), management, and soil conditions that are unavailable in many parts of the world.

The presence of all the data required for these models still requires calibration, which is difficult due to the high amount of uncertainty in the model parameters. Often, parameter uncertainty is ignored in such complex models and parameter values are picked through subjective deliberation. Iizumi et al. (2009) estimated distributions of parameter values from Markov Chain Monte Carlo techniques with the widths of the distributions reflecting the inability of historical rice, maize yield datasets to effectively constrain parameter values. They found that parameter uncertainties translated to larger uncertainties in projecting yield responses to climate change. They concluded that uncertainty of climate change impact assessment on crop yield may increase if future climate projections are fed to crop models with parameters optimized under current climate conditions [5].

Statistical models, on the other hand, trained with historical crop yield data, and made of relatively simple linear regression ensembles provide a common alternative to the process based models described above. These methods include the following:

1) time series methods - based on time series data from a single point or geographical area.
2) panel methods - these methods rely on variations in both time and feature space.
3) cross sectional methods - based solely on variations in space.

Time series methods are often limited by availability of data but have the advantage of capturing the variability for the chosen location. Panel and cross-sectional methods assume common parameter values across all locations and are prone to errors from omission of features like soil conditions or fertilizer uptake. The main advantages of statistical models are their limited reliance on field calibration data, and their transparent assessment of model uncertainties. For example, if a model does a poor job of representing crop yield responses to climate, this will be reflected in a low coefficient of determination (R^2) between modeled and observed quantities, as well as a large confidence interval around model coefficients and predictions. Although process-based models could in theory be accompanied with similar statistics, in practice they rarely are [4]. A few shortcomings with statistical models include co-linearity problems and assumptions of stationarity of the underlying ecological processes, which is an extremely important problem due to rapidly changing ecologies on the planet as a result of global warming.

Lobell et al.[4] recognized the need for a systematic evaluation of statistical methods in predicting crop yields to climate and recommended that efforts be made to determine the specific conditions under which their predictions are likely to be misleading. This would help in quantifying the common errors that arise from using these simpler if imperfect approaches. Researchers [6] [7] have found that statistical models work best at large spatio-temporal scales like national or regional levels. Other researchers have highlighted the difference between understanding the impact of climate change on crop growth and on crop yields. Predicting crop yields involves understanding predictors that aren't reflected in the physiological crop growth models explaining their inability to reliably predict crop yields. This problem maybe considered as one of dimensionality reduction where multiple predictors need to be aggregated at regional, provincial and national levels and temporal scales like monthly or seasonal climatic variable changes.

A. Objective

The main goal of this report is to evaluate the accuracy of context unaware statistical models at accounting for the uncertainty found in the task of crop yield prediction. Statistical models that use Neural Networks and Multiple Adaptive Regressive Splines are to be built using historical annual crop yield, seasonal temperature and precipitation levels and assess their accuracy at predicting yields using measures such as error bounds, confidence intervals on predictions and compare these measures of variability with similar measures obtained by crop growth models. Many researchers [8] [1] [2] have found that statistical models can explain variability in crop yields due to climate change better than crop growth models do. The important features that explain this variability were found to be seasonal variations in weather, precipitation, fertilizer intake and crop breeding effects. This study aims to make similar inferences about the importance of climatic and geophysical factors on crop yields based on the results obtained from the Neural Network and MARS models.

II. Data and Methods

A. Datasets

Data was obtained from the World Bank's data on socioeconomic indicators for all countries in the world. This data includes historical agricultural indicators like crop yields, fertilizers used, agriculture related carbon dioxide, nitrogen dioxide emissions and extensive climate data going back to 1901.

The crop yield data obtained was for annual cereal production in India from 1961 to 2013. Cereals are a plant family consisting primarily of wheat, oats and corn.

B. Methods

All data was cleaned and was preprocessed to remove

III. Results and Discussion

IV. Summary and Conclusions

Acknowledgment

References

[1] T. Iizumi, H. Sakuma, M. Yokozawa, J.-J. Luo, A. J. Challinor, M. E. Brown, G. Sakurai, and T. Yamagata, "Prediction of seasonal climate-induced variations in global food production," *Nature climate change*, vol. 3, no. 10, pp. 904–908, 2013.

[2] a. Gonzalez-Sanchez, "Predictive ability of machine learning methods for massive crop yield prediction," *Spanish Journal of Agricultural Research*, vol. 12, no. 2, pp. 313–328, 2014. [Online]. Available: http://revistas.inia.es/index.php/sjar/article/view/4439

[3] K. O. Rauff and R. Bello, "A Review of Crop Growth Simulation Models as Tools for Agricultural Meteorology," no. September, pp. 1098–1105, 2015.

[4] D. B. Lobell and M. B. Burke, "On the use of statistical models to predict crop yield responses to climate change," *Agricultural and Forest Meteorology*, vol. 150, no. 11, pp. 1443–1452, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.agrformet.2010.07.008

[5] T. Iizumi, M. Yokozawa, and M. Nishimori, "Parameter estimation and uncertainty analysis of a large-scale crop model for paddy rice: Application of a Bayesian approach," *Agricultural and Forest Meteorology*, vol. 149, no. 2, pp. 333–348, 2009.

[6] "A review on statistical models for identifying climate contributions to crop yields," *Journal of Geographical Sciences*, vol. 23, no. 3, pp. 567–576, 2013.

[7] D. B. Lobell and M. Burke, *Climate change and food security: adapting agriculture to a warmer world*. Springer Science & Business Media, 2009, vol. 37.

[8] S. Wolfram and B. L. David, "Robust negative impacts of climate change on African agriculture," *Environmental Research Letters*, vol. 5, no. 1, p. 14010, 2010. [Online]. Available: http://stacks.iop.org/1748-9326/5/i=1/a=014010