Crop yield prediction using Convolutional Neural Networks

The Mighty Pythons

I. Introduction

Crop modelling is of great relevance for the global economy, to ensure food security and to minimise environmental impact. Crop modelling is a hot topic, with a great variety of models being employed, from regression [1], to random forests [3], neural networks [2] and Gaussian processes [4].

II. Methods

A. Data exploration and pre-processing

From exploration of the yield in each county (Figure 1), we can see a general increasing trend over the years. We can also note a great variation in yield, especially in more recent years. This suggests that as crop yield increases, it is also more susceptible to variation, which can be due to various factors including weather conditions, soil health and labour factors.

While the strong correlation with the minimum yearly temperature was expected, Figure 2 shows an interesting relationship of the yield with the EVI. While we expected a higher EVI to produce a better yield, this is only true for the EVI measured between June and August. However the lower the greenness between the months between the end of March and the beginning May, the higher the yield for the year. This figure also confirmed the selection of only months between April and November, as suggested by [1]. The EVI is also higher in the summer months, and zero during deep winter, reflecting the absence of grown crops during the colder months.

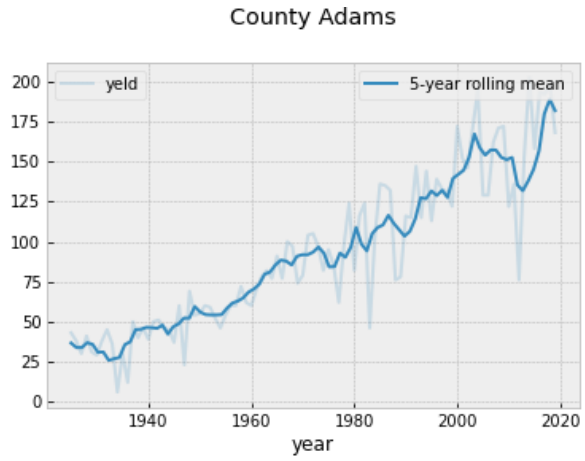


Fig. 1: Yield per year in the county Adams.

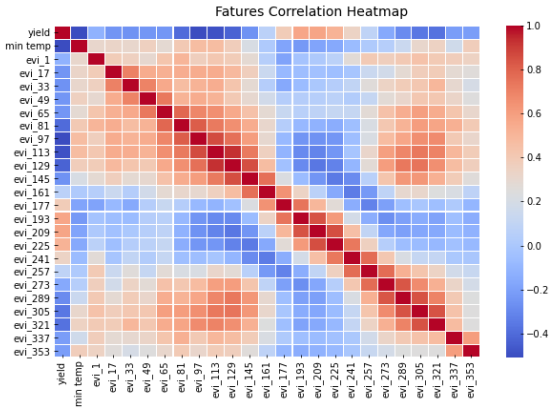


Fig. 2: Pair-wise correlation between the features used.

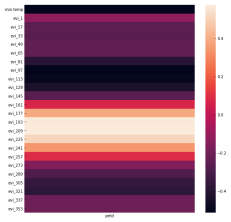


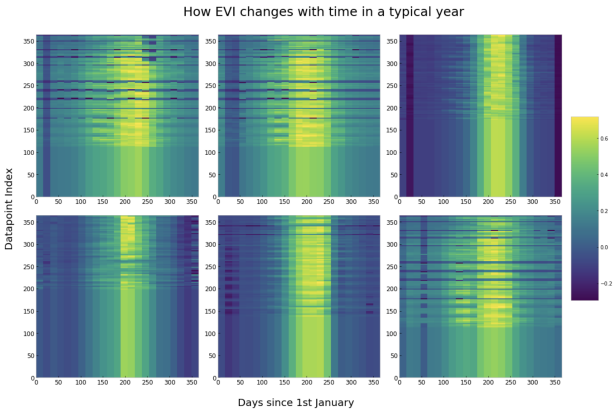
Fig. 3: Correlation between yield, yearly minimum temperature and EVI at different times of year.

We corrected anomalies in the EVI data–which are often due to clouds covering the satellites and therefore not representative of the crops [1]–by removing values that were sufficiently far away from the mean. For a given row y_j, overall matrix mean μ_{evi}, and row mean value , an anomaly was flagged when $e^{-(\bar{y_j}-\mu_{evi})^2} < 1-\epsilon$. Here, ϵ was empirically determined to be 0.04.

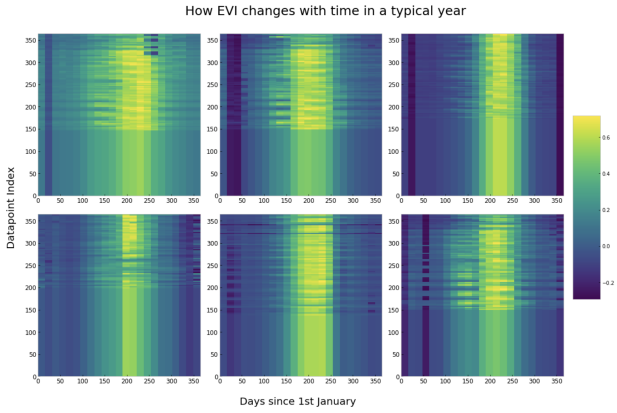
Given missing data at various locations across the years, we estimated the missing values in the EVI matrices by linear interpolation. Interpolation also ensures that anomalies are replaced with sensible values.

B. Method selection

We decided to utilize a deep learning, regression based approach for yield prediction. Furthermore, we also decided to not utilise temporal prediction methods such as LSTMs and RNNs. This is because we have EVI data for years 2001-2019 for most counties (amounting to 20 data inputs for the



(a) EVI data with anomalies



(b) EVI data after removal of anomalies

networks), which intuitively seemed too less to use for county-wise time series predictions. Furthermore, the EVI features for a given county for a given year were in the form of a matrix, which is convenient for input into a CNN. Lastly, it was noticed that data for some years was missing, which doesn't make a time series based approach ideal. Therefore, our main approach was the following:

First, we need to construct appropriate features we could feed into the model. A CNN architecture based on [] was constructed.

- First we performed the pre-processing steps in part A, gathering EVI data for each county for each year in a matrix of size (150,23). This data was cleaned by removing anomalies and padded with mean values of each column.
- We initialized our CNN with the architecture shown in Figure 4.
- We trained the CNN for 25 epochs with an exponentially decreasing learning rate. The convergence plots can be seen in Figure 5.

III. Results

Following are the main findings of this investigation

- The CNN model managed to produce a minimum root mean squared error (RMSE) of $\bar{\varepsilon} \approx 17.5$. The average yield value across the test set was noted as $\bar{y} = 150$. When translated relative to this average, the model made an average error of 11%.
- A few patterns were noted in the EVI data. Firstly, the presence of anomalies threatened to bias our model, but we added anomaly detection using a squared exponential kernel which resolved this problem. The EVI data after removal of anomalies is seen to be symmetric and is highest around the middle (the time corresponding to summer). This pattern suggests that the most significant yield in a given year may be taking place around the lighter band of values.

	Actual Yield	Predicted Yield	County	Year
0	194.7	176.44	CASS	2017
1	130.0	113.48	CLAY	2006
2	163.0	152.96	CLINTON	2008
3	138.0	146.02	COOK	2009
4	120.0	147.44	JASPER	2005
5	150.0	144.56	JASPER	2007
6	129.0	96.20	JOHNSON	2005
7	127.3	106.74	JOHNSON	2011
8	175.2	143.20	KANE	2011
9	129.0	135.00	LAKE	2019
10	145.0	138.83	LIVINGSTON	2005
11	184.0	200.40	LIVINGSTON	2009
12	188.0	171.86	LOGAN	2004
13	153.0	141.72	MACOUPIN	2005
14	142.0	137.74	MARSHALL	2005
15	181.0	172.57	MONTGOMERY	2008
16	109.0	123.56	RICHLAND	2005
17	193.3	172.95	SCOTT	2016
18	165.0	152.53	VERMILION	2003

Fig. 5: A sample table showing some of the yield predictions made by the CNN.

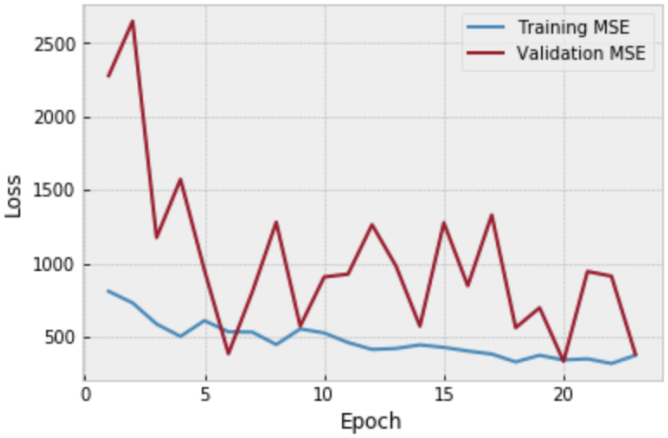


Fig. 6: The training and validation convergence plot of the CNN

IV. Conclusion and Next Steps

Crop estimation is extremely relevant for human survival and environmental impact, but there is still not a definite solution for modelling, given the high heterogeneity and

complexity of the data used for estimation. In this report we have proposed an CNN-based approach trained on the EVI data to estimate the yearly yield for each county. Our model reached good prediction performance, with an average error of 11%.

Some possible ways to extend this exploration include exploring other CNN architectures, adding a Gaussian Process layer using the final features of the network, and adding an extra column of cleaned temperature, latitude, and longitude data to our EVI feature matrix.

REFERENCES

[1] Mateo-Sanchis et al. *Synergistic integration of optical and microwave satellite data for crop yield estimation*, Remote Sensing of Environment, 234, 2019, 111460.

[2] Khaki, Saeed and Wang, Lizhi, *Crop Yield Prediction Using Deep Neural Networks*, Frontiers in Plant Science, 2019.

[3] Li'ai Wang and Xudong Zhou and Xinkai Zhu and Zhaodi Dong and Wenshan Guo, *Estimation of biomass in wheat using random forest regression algorithm and remote sensing data*, The Crop Journal, 2016.

[4] L. Martínez-Ferrer and M. Piles and G. Camps-Valls, *Crop Yield Estimation and Interpretability With Gaussian Processes*, 2020.