Convolutional Neural Network Based Algorithm for Crops Yield Prediction

AI Hack 2021: Crops Yield Prediction Challenge

Karim Alaa El-Din, Susan Chen, Anson Poon, Lorenzo Versini

***Abstract*—We investigate crop yield as a function of EVI index and temperature collected over a range of 19 years for Illinois counties. We propose a model for up-sampling the spatial resolution of the data provided and we outline some of the challenges that this approach presented. Then, we illustrate a second algorithm which predicts the crop yield for each county given the EVI index and the temperature averaged in space across the relevant locations. By comparing the test data with the predictions, we see that there is a good alignment between the two. Finally, we present some remarks regarding the possible applications and future developments of our analysis.**

I. Introduction

CROP harvesting optimisation is of major interest due to the implications it has on population sustainability and the environment. The global population is expected to increase by 2 billion in the next 30 years [1]. In this report we present some tools we developed as part of the AI Hackathon 2021 to characterise the crop distribution as a function of environmental conditions. This can be helpful in the organisation and distribution of resources such as work force and machines.

II. A First Look at the Data

The data provided includes two measures which could affect the crop yield, as well as the total yield per county in Illinois. It contained information about the EVI index and the temperature at multiple locations across Illinois State across an heterogeneous time range. Unfortunately, neither the times nor the place at which measurements were collected matched, resulting in a difference in temporal and spatial resolution. To work with the date-set as a whole, we decided to perform a linear interpolation step in which the temperature data were overlapped on the EVI data both in the time and space domain, in the order mentioned.

Note that the total yield was specified per county, however, only 96 out of the 103 counties in Illinois had a yield provided by the data-set. The yields for the left over unnamed counties were summed and attached to a label of "*OTHER (COMBINED) COUNTIES*". This was taken note of for analyses in later sections of this report.

Prior to work in section III, we decided to explore a few trends in the data-set in order to gain initial insight, presenting them below:

[1] https://www.un.org/en/sections/issues-depth/population/

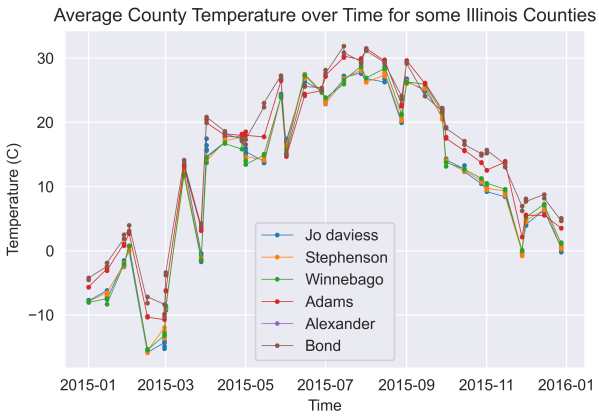


Figure. 1. This plot shows the temperature over the duration of 2015 for the six specified Illinois counties.

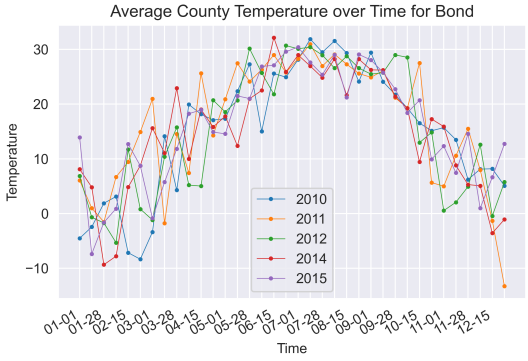


Figure. 2. This plot shows, for the county "Bond", the temperature over the duration of a year, for the above specified years.

III. Method

We now describe the algorithm that we used to return the partial yield for each location in a county. We started with the assumption that EVI and temperature readings were enough to make predictions on the partial yield for that particular location.

Our aim was to define the mapping

$$f : (EVI, T, \text{time}) \to \text{partial yield}, \tag{1}$$

where T is the temperature, which is a function of time, for every location specified in our EVI data-set. The sum of all

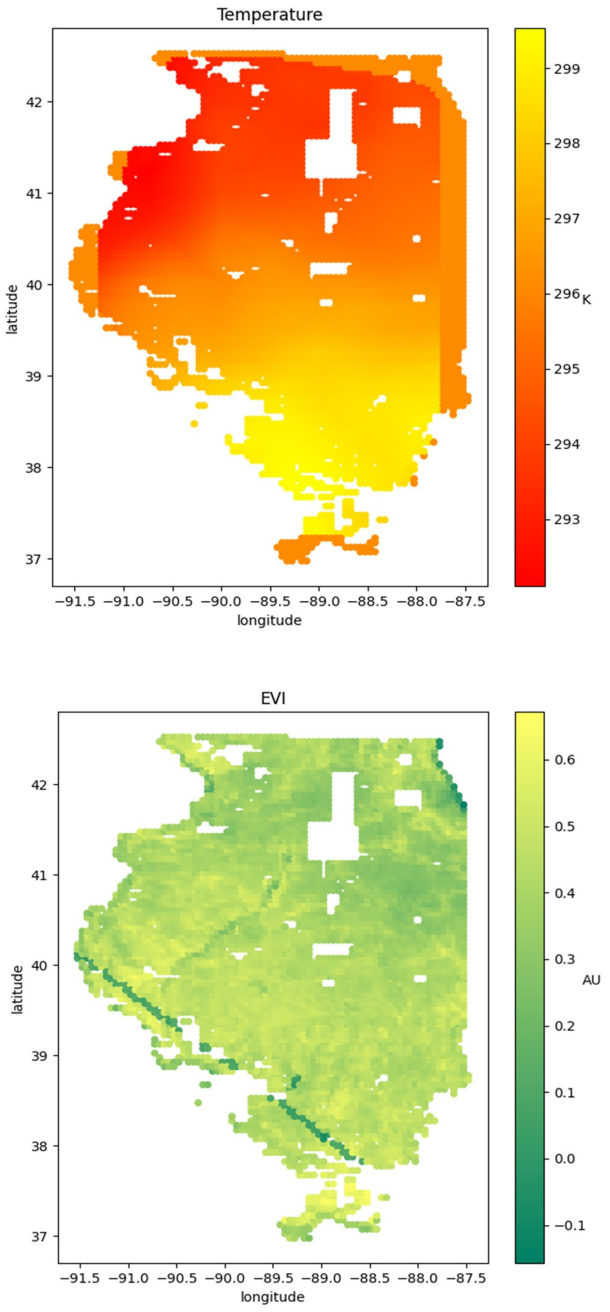


Figure. 3. This plot shows the temperature and EVI throughout Illinois on 1st January 2001.

the partial yields in a county should lead to the total yield in that county in a year, which is our observable. This is shown in fig. 4, where f is the function we are interested to learn. Before creating this map however, we decided to follow the following preprocessing pipeline:

1) interpolate across time (linear) to align T with the EVI data
2) interpolate across space (linear) to align T and EVI across all dimensions
3) group by year and county to match with the yield data

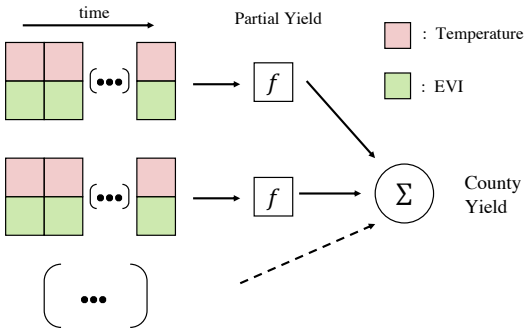


Figure. 4. A schematic of the mapping described in eq. (1).

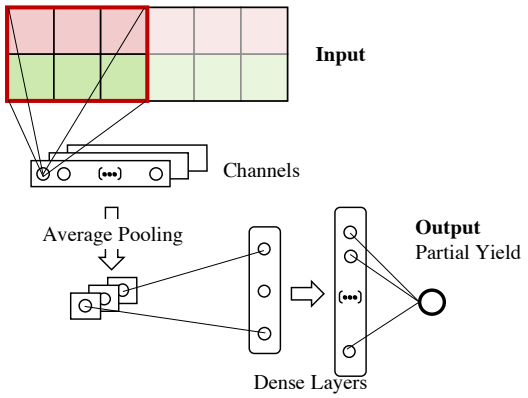


Figure. 5. A schematic of our ANN algorithm: First the input is passed through a CNN to detect relevant features. Then, we perform an average pooling, hence ending up with a number of neurons equal to the number of channels. Two dense layers are then used to perform the last calculations to obtain the partial yield. The *relu* activation function was used at every layer but the output layer.

4) filter where data was available for inputs (EVI + T) and outputs (Yield)
5) filter other cases where data was not available

This left us with 1370 events, each representing a county and year combination. We split these into training + validation and test set (containing 1000 and 370 events respectively).

We then normalised each of these groups (to avoid data leakage from test to training set) and averaged over in-county location to get time series data for EVI and Temperature. We also included the standard deviation in those time series within a country as input features. This left us with inputs of the shape 23x4 and outputs of shape 1, between which we had to build a map.

In order to build the map, we developed an artificial neural network algorithm (ANN) which employed convolutional layers (CNN) to retain the time information in the data-set and some dense units for processing the information collected. A schematic of our algorithm can be seen in fig. 5.

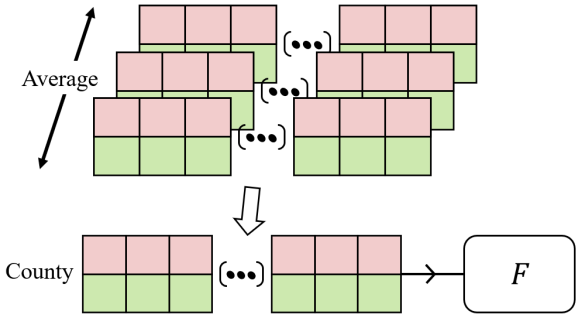


Figure. 6. A schematic of the new algorithm: We take the average over all locations in each county. This is then passed through a NN model consisting of a convolutional layer followed by a series of dense layers with the $relu$ activation function (other than the output layer).

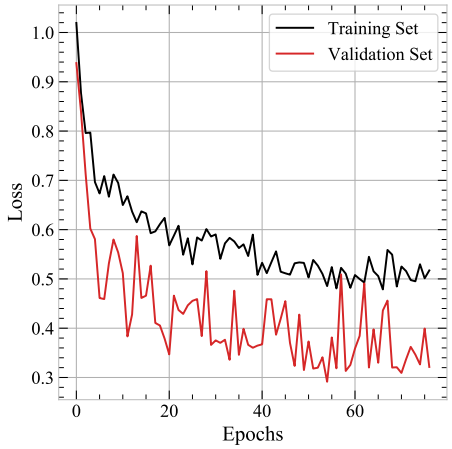
IV. Results and Analysis

We trained the model described in section III on the data-set. One difficulty we faced was the fact that the locations in each county were not fixed, i.e. we had a variable number of entries in the sum showed in fig. 4. This made the process of training over many counties at the same time problematic. Hence we decided to create a customised training routine during which the data was fed into the model one county at a time. This approach made the generalisation of the model more difficult and at the time we wrote this report, an obvious solution to the problem was not found.

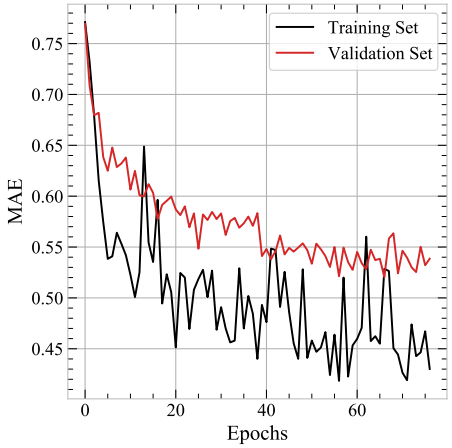
Indeed, training the algorithm led us to values in the loss function (chosen to be the Mean Squared Error) which were on the order of the standard deviation on the yield measurements themselves. Hence, improvements are required before our model could have any practical application, but once applied it would allow *upsampling*, i.e. we would be able to estimate the crop yield with a resolution higher than the available data, hence allowing more precise logistic decisions.

We implemented a model (fig. 6) based on taking the average over single locations in each county for a given time. Dropout was used as a regularisation technique in order to reduce over-fitting. The algorithm itself is similar to the one in Fig. 5, however more dense layers were added, in order to increase the complexity of the model.

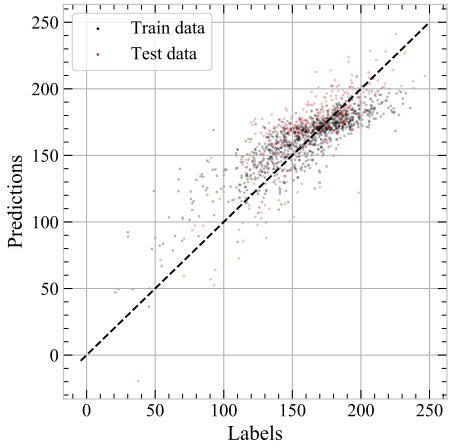
We show the performance of this algorithm in fig. 7. As we can see from fig. 7 (a), the validation loss (*Mean Squared Error*, MSE) is lower than the training loss. However, we note that the training set outperforms the validation set when considering the *Mean Absolute Error* (MAE). This is expected, as we used dropout layers which introduce noise in the training procfig. 7 (c) shows a plot of `label vs prediction` for the test and the train data. As we can see, there is a good alignment between the two.



(a)



(b)



(c)

Figure. 7. Shows the performance of the algorithm predicting yield for every county as a function of the EVI index and temperature measured at regular intervals along the year and averaged spatially across the county. As we can see, the validation data outperforms the training ones in the loss function in (a), which is due to the use of dropout layers. On the other hand (b) shows that the validation set still struggles to reach the performance of the training set when considering the MAE. Finally, (c) compares predictions and true labels for test and training data, showing a good alignment.

V. Conclusions

We suggested an approach for spatial up-sampling of the crop yield data-set which can be useful for better resource management. Moreover, the second algorithm we implemented can be useful for the purpose of simulating the yield as a function of EVI and temperature. Hence, by coupling it with existing climate models, it allows us to estimate the cultivated surface required (as specified by the EVI) in order to meet yearly crop standards. However, more work has to be done to better characterise the uncertainties in our prediction, which can be critical in the food industry. Larger amounts of data with finer spatial resolutions would allow this algorithm to implement stronger predictions.