A Short Literature Survey

Satellite Imagery (Remote Sensing Data), has been widely used for predicting crop yield [1]. This dataset is collected using the sensors mounted on satellites or planes, which detect the energy (electromagnetic waves) reflected or diffracted from surface of the earth [2].

Remote sensing data has a lot of energy bands to offer, but mainly only few of them have been used for crop yield prediction, in the form of features like Normalised Difference Vegetation Index (NDVI), and Normalised Difference Water Index (NDWI) [1]. Yet, there are some people who have tried generating relevant features using the bands which are typically ignored, and they have been successful with improving results with that.

In case of this dataset, most people rarely explore the high-order moments of the features. They either go for mean i.e. 1st moment (like in [4]) or do sampling (like in [5]). Paper [1] has proposed a way of dimensionality reduction by transforming the raw images to a histogram of pixel counts, by discretising the pixel values to bins. They state that this technique of dimensionality reduction helped them train NNs even when the training data is scarce.

One publication also provides with an architecture for managing big data in the agricultural area [3]. It talks about MapReduce structure for Weather Data, ARMA Model etc. Most likely we are not going to need that.

Other datasets used for this problem are Crop Yield Dataset (obviously), Climate Datasets like precipitation, max-min temperature, vapour pressure, Soil properties etc [6]. Based on these datasets people have used algorithms like Regression models, SVM, Random Forest, Nearest Neighbour, Deep Neural Networks. Paper [6] compared these algorithms concluding that NNs give them the best results.

Paper [5] is an example of using Deep Learning on remote sensing data. They have 3 hidden layers in their Deep Neural Net with 256 neutrons in each. Will check more details about their process. On the other hand, Paper [1] uses CNNs and LTSMs with only remote sensing data as the input. Moreover, to incorporate the spatial effect based on the soil properties, paper [1] has used a Gaussian Process Model on top of their Deep NN architecture. They have shown that CNN + GP Model works best amongst all algorithms. Also, progressively, as more and more images for a particular year are available (for more months starting from January) the predictions become better.

What additional things we can do …

- Indian Context: (all the collected papers are in the context of USA or some other countries). Didn't see any great paper for the Indian context. Most of the publications (predicting crop yields for regions in India) that I came across, are using SVM, Regression techniques, Decision Trees.
- Need to go through Microsoft India's work for crop yield production.
- Identify the exact pixels which refer to the green cover in the Satellite Images. Paper [1] mentions that there is a standard dataset available at DAAC website (link) indicating the pixels

referring to the crop production worldwide. They apparently provide some pixel masks to remove the extra pixels from the data. We need to check how much accurate it is for regions in India. Also need to verify if those masks are compatible with the images from that website only or they will work on our dataset too.
- If we find a way to identify exactly which pixels pertain to crop production for any given image, then we can improve the accuracy of those masks, since the crop regions are dynamically changing.
- Architecture of the Deep Neural Networks can be improved to suit our needs better. Also, in addition to the remote sensing data, try including the climate datasets or soil quality datasets as an input.
- If the method of collecting data in paper [1] (through google earth) is more feasible than ours then we need to adapt their method while making any improvements possible specific to our needs.
- After reaching our goal, push for assisting farmers, govt using our predictions. All these publications state they have done better than their competitors but there is no article or public mention of their work being used practically to assist the farmers. If there are some genuine problems in rolling out that work to next stage, then identify those problems and try solving them.

I would say this publication from Stanford University is the most recent (2017) and the best one I came across. Almost all other publications in this field are not from the top tier conferences or journals. (This is my observation).

References:

[1]: “Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data”. Jiaxuan You and Xiaocheng Li and Melvin Low and David Lobell and Stefano Ermon. Department of Computer Science, Stanford University. AAAI 2017.

[2]: https://oceanservice.noaa.gov/facts/remotesensing.html

[3]: “Prediction of crop yield using big data”. Wu Fan, Chen Chong, Guo Xiaoling, Yu Hua. 8th International Symposium on Computational Intelligence and Design, 2015.

[4]: “An assessment of pre-and within- season remotely sensed variables for forecasting corn and soybean yields in the United States”. Johnson, D. M. 2014.

[5]: “Estimating crop yields with deep learning and remotely sensed data. In 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS).” Kuwata, K., and Shibasaki, R. 2015.

[6]: “Machine Learning Approaches to Corn Yield Estimation Using Satellite Images and Climate Data: A Case of Iowa State” Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography. 2016