Crop classification and yield estimation using multispectral satellite imagery

Raul Harnasch
Computer Science Department
University of Massachusetts Lowell
Lowell, USA
rharnasc@cs.uml.edu

***Abstract*—Remote sensing has proved fruitful in the mapping, monitoring and assessment of vegetation health as well as soil conditions. Analysis techniques in this area have become invaluable to the agricultural industry by providing precise near-realtime awareness in many facets of crop production – from irrigation planning to pesticide treatments – as remote sensing data, such as multispectral imagery, has been made more available. Past studies have performed crop yield estimations utilizing both physical measurements to take into account variables such as organic carbon content, as well as various indices and surface measurements derived from multispectral data sources, but have not made use of the full gambit of sensors currently available from orbital satellites. This paper proposes to use a combination of random forrest classification and support vector regression in order to identify crops from their associated spectra and further estimate crop yield.**

***Index Terms*—computer vision, random forrest, support vector machine, regression analysis, machine learning, multispectral analysis, crop yield prediction, agriculture, yield estimation, crop classification, corn, soybeans**

I. Introduction

Crop yield estimation plays an important role vertically across agricultural sectors of different scales, from the local level in planing irrigation schedules to food management of national and international economies year-to-year. [1] Remote sensing has since fit into each of these sectors through the development of a number of indices to measure vegetation health and stress, such as normalized difference vegetation index (NDVI), vegetation condition index (VCI), normalized difference water index (NDWI), and thermal condition index (TCI) – each with the aim to measure different facets of agricultural production, which can potentially be used in combination to estimate crop production at a scale of 35,000 km^2 with current satellite aperture sizes.

Though crop yield estimation is based on various methods such as statistical models, trend or regression analysis, field surveys, and crop growth simulation models [2], the underlying datasets to support these activities are a combination of physical measurements and historical data, only some of which derived from remote sensing. With greater access to multispectral imagery in the visible, shortwave, thermal and near infrared bands, combined with harvest data collected at the local level across 750 acres in central Illinois, this paper describes a method of crop yield estimation with regression analysis via support vector machine after being classified by a random forrest.

II. Background

Many methods have been proposed for crop yield estimation vary with the datasets employed in their analyses and have had moderate success. More traditional approaches leveraging statistical methods[1] diverge between utilizing ”physically-measured”[2] datasets (e.g. rainfall, pH levels) [3] and remote sensing [4]; or data derivable from remote sensor information such as evapotranspiration. [5] And although remote sensing data can be sparse, analysis using an autoregressive (AR) state-space model yielded results suggesting that ”physically-measured” features proved to be less helpful to crop yield estimation in both bivariate and multivariate analysis than features obtained from remote sensors. [6]

With the rise and expanding accessibility of machine learning more automated means of estimation and classification have been studied. In the past, neural networks have been applied to the area of crop prediction and estimation, but have also relied on physical measurements. [7] [8] More recently, computer vision has focused on small-scale detection of crop stressors and health indicators for individual plants [9], green houses [10] and vineyards through the use of UAVs [11], suggesting that the technology is reaching significant levels to produce data at a granularity sufficient to support yield prediction activities at a larger scale.

III. Approach

A. Context

As mentioned earlier, other approaches have utilized extensive spatiotemporal data encompassing several years to build models, often with recurring measurements throughout the growth period. [2] The goal of this effort, however, was to see if collected crop spectra could be used – standalone – as input for predicting yield, and as such focused on satellite imagery captured days before the harvest began.

[1] Least-square regression, exponential-linear, etc.

[2] This paper makes a distinction between data gathered by remote sensing (passive collection) and measurements requiring physical contact or observation.

Fig. 1. Satellite RGB composite of central Illinois, capturing area of analysis

While this effectively reduced the number of images available to process, this limitation was overcome by treating each pixel as individual images for training and testing purposes.

NDVI and NDWI were calculated separately and inserted into the feature vectors, given the values captured from each band given the following equations:

$$\begin{gathered}NDVI = \frac{NIR - Red}{NIR + Red}\\[1em] NDWI_{Gao} = \frac{NIR - SWIR}{NIR + SWIR}\\[1em] NDWI_{McFeeters} = \frac{Green - NIR}{Green + NIR}\end{gathered}\tag{1}$$

All models were built using default models provided by the scikit-learn python library.

B. Crop Classification

The method that yielded the best results for crop classification, using the provided features, turned out to be a random forrest classifier, using the `RandomForrestClassifier` method containing 10 estimators. Training and test subsets were then randomly allocated and divided from the overall dataset using the `train_test_split` method and passed to the K-Fold cross validator (consisting of 10 folds) for training and evaluation.

C. Yield Estimation

Support Vector Regression (SVR) was used to attempt to predict yields with the given multispectral measurements. There was a preoccupation with attempting to find a 'best-fit' model that which produces a linear relationship between the features and yield, so a considerable amount of time was taken using trial end error to identify a model that would prove promising.

Unfortunately no regression technique emerged that yielded remotely usable results for the task at hand. Further details are provided in the Evaluation section.

IV. Dataset

A. Contents and Features

The availability of agricultural yield data was by far the largest limiting factor in the study. Without an open source dataset, the effort focused on building and curating a dataset capable of supporting the task at hand – an activity described in detail in the *Data Curation* subsection.

The dataset is comprised of approximately 6,600 samples of 10 features consisting of multispectral band measurements obtained from Landsat 8 imagery (*See Table I*). For each record, two targets are provided with which to train – yield and crop classification:

```
In [2]: farm = load_data()

In [3]: farm.target_crop
Out[3]: array([1, 0, 0, ..., 0, 1, 0])

In [4]: farm.target_yield
Out[4]: array([ 4033., 20310.,
22225., ..., 26855., 6188., 29274.])

In [5]: farm.target_names
Out[5]: ['CORN', 'SOYBEANS']
```

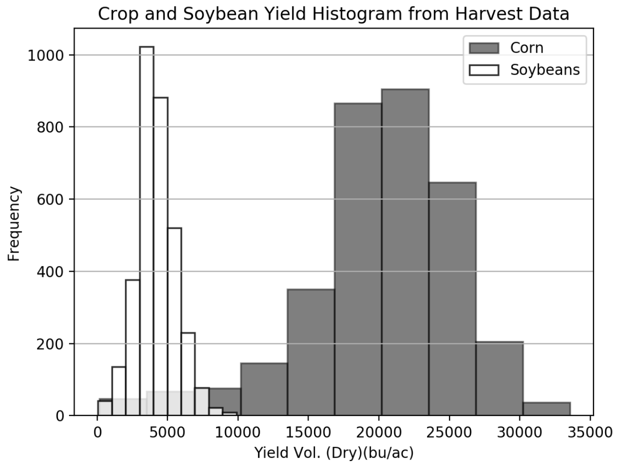

Fig. 2. Combined 2016-2017 harvest yield histogram

TABLE I
RECORD EXAMPLE

Feature	Value
Green	8407
NDVI	0.4495
NDWI (Gao)	0.3661
NDWI (McFeeters)	0.4008
Near IR	19653
Red	7464
Shortwave IR 1	13473
Shortwave IR 2	9119
Thermal IR	27287
Thermal IR 2	24934

where the targets are one-to-one mapped to each record in order by index. `target_crop` consisting of an array of labels whose values corresponds to the list of `target_names`, and `target_yield` consisting of yield measurements in bushel/acres.

B. Data Curation

The most difficult step was to obtain historical yield data from harvesting activities in the agriculture industry, which came in proprietary formats requiring third party software[3] to parse and export into csv files. Harvest data described crop yield information accounting for 750 acres of farm land in central Illinois, covering two years of harvest activities.[4]

Based on harvest times, satellite imagery was identified and collected from the Landsat 8 satellite via the USGS website, where particular focus was placed on the Red, Green, Near Infrared (NIR), Shortwave Infrared (SWIR) and Thermal Infrared (TIRS) bands (*See Table II*).

Using the Geospatial Data Abstraction Library (GDAL), lat/long measurements were correlated to pixels within the satellite images and the values within each band were extracted and concatenated into the feature set. Additionally, for each record, the measured Red, Green, SWIR2 and NIR values were used to calculate the NDVI [12] and NDWI [13] [14] as additional features.

The resultant dataset consisted of 10 features, with correlated crop types and yield sizes on a pixel by pixel level. However, the number of measurements per pixel were not uniform, as the size and shape of the 750 acres of farmland being modeled were (obviously) not uniform 900m^2 plots. Using the number of samples per pixel, the dataset was further filtered to include only the records that were within three standard deviations from the median.

V. EVALUATION

A. Classification

Random forrest classification performed at 89.7±0.2% accuracy.

TABLE II
LANDSAT 8 BAND DESIGNATIONS [15]

Band	Designation	Wavelength (μm)	Res (m)
3	Green	0.0.533 - 0.590	30
4	Red	0.636 - 0.673	30
5	Near Infrared (NIR)	0.851 - 0.879	30
6	Shortwave Infrared (SWIR) 1	1.566 - 1.651	30
7	Shortwave Infrared (SWIR) 2	2.107 - 2.294	30
10	Thermal Infrared (TIRS) 1	10.60 - 11.19	30[a]
11	Thermal Infrared (TIRS) 2	11.50 - 12.51	30[a]

[a] *TIRS bands are acquired at 100 meter resolution, but are resampled to 30m in delivered data product.*

B. Yield Estimation

Methods consisted of permutations of performing standardization and principal component analysis (PCA) prior to being fed into linear regression models. SVR with radius bias function (RBF), linear, and polynomial kernels were also tested which yielded 0.02 r2 scores. Optimization was attempted upon the hyperparameters, for $0 <= C <= 5$ and $0 <= \gamma <= 100$, with no considerable improvements.

VI. CONCLUSION

The use of current features extracted from multispectral sources indicates the probability there is no hard relationship between the gathered data and the yield. While this seemingly reenforces past methods utilizing datasets with wider temporal spread and/or physical parameters, not all factors have been considered in the course of the analysis described in this paper.

More interestingly, the potential use for crop classification using random forests does appear to be promising and could use additional attention.

¯_(ツ)_/¯

VII. ACKNOWLEDGEMENTS

Sincere thanks are due to Brent Lage, Dr. Jeremy Meiners and Dr. Chad Meiners for making this paper possible, and to Dr. Dennis Ross and Jacob Lowey for the guidance and direction.

REFERENCES

[1] M. J. HAYES and W. L. DECKER, "Using noaa avhrr data to estimate maize production in the united states corn belt," *International Journal of Remote Sensing*, vol. 17, no. 16, pp. 3189–3200, 1996. [Online]. Available: https://doi.org/10.1080/01431169608949138

[2] A. K. Prasad, L. Chai, R. P. Singh, and M. Kafatos, "Crop yield estimation model for iowa using remote sensing and surface parameters," *International Journal of Applied Earth Observation and Geoinformation*, vol. 8, no. 1, pp. 26 – 33, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0303243405000553

[3] D. Jones, "A statistical inquiry into crop-weather dependence," *Agricultural Meteorology*, vol. 26, no. 2, pp. 91 – 104, 1982. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0002157182900358

[4] A. S. Oroda, "Application of remote sensing to early warning for food security and environmental monitoring in the horn of africa."

[5] W. Kaicun, W. Pucai, L. Zhanqing, C. M., and S. Michael, "A simple method to estimate actual evapotranspiration from a combination of net radiation, vegetation index, and temperature," *Journal of Geophysical Research: Atmospheres*, vol. 112, no. D15. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2006JD008351

[3] ASF View by Chase IH was used to read and export harvest data collected from sensors onboard the farm equipment.

[4] 2016-2017

[6] O. Wendroth, H. I. Reuter, and K. Kersebaum, "Predicting yield of barley across a landscape: a state-space modeling approach," *Journal of Hydrology*, vol. 272, no. 1, pp. 250 – 263, 2003, soil Hydrological Properties and Processes and their Variability in Space and Time. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S002216940200269X

[7] S. S. Dahikar and D. S. V. Rode, "Agricultural crop yield prediction using artificial neural network approach," 2014.

[8] A. Kehagias, H. Panagiotou, N. Maslaris, V. Petridis, L. Petrou, and V. Spais, "Predictive modular neural networks methods for prediction of sugar beet crop yield," *IFAC Proceedings Volumes*, vol. 31, no. 12, pp. 41 – 45, 1998, iFAC Workshop on Control Applications in Post-Harvest and Processing Technology (CAEA'98), Athens, Greece, 15-17 June 1998. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1474667017360391

[9] A. Elvanidi, N. Katsoulas, K. Ferentinos, T. Bartzanas, and C. Kittas, "Hyperspectral machine vision as a tool for water stress severity assessment in soilless tomato crop." *Biosystems Engineering*, vol. 165, pp. 25 – 35, 2018.

[10] D. Story and M. Kacira, "Design and implementation of a computer vision-guided greenhouse crop diagnostics system." *Machine Vision and Applications*, vol. 26, no. 4, pp. 495 – 506, 2015.

[11] T. Poblete, S. Ortega-Faras, and D. Ryu, "Automatic coregistration algorithm to remove canopy shaded pixels in uav-borne thermal images to improve the estimation of crop water stress index of a drip-irrigated cabernet sauvignon vineyard." *Sensors (Basel, Switzerland)*, vol. 18, no. 2, 2018.

[12] J. W. Rouse, Jr., R. H. Haas, J. A. Schell, and D. W. Deering, "Monitoring Vegetation Systems in the Great Plains with Erts," *NASA Special Publication*, vol. 351, p. 309, 1974.

[13] B.-C. Gao, "Normalized difference water index for remote sensing of vegetation liquid water from space," pp. 2480 – 2480 – 12, 1995. [Online]. Available: https://doi.org/10.1117/12.210877

[14] S. K. McFEETERS, "The use of the normalized difference water index (ndwi) in the delineation of open water features," *International Journal of Remote Sensing*, vol. 17, no. 7, pp. 1425–1432, 1996. [Online]. Available: https://doi.org/10.1080/01431169608948714

[15] J. A. Barsi, K. Lee, G. Kvaran, B. L. Markham, and J. A. Pedelty, "The spectral response of the landsat-8 operational land imager," *Remote Sensing*, vol. 6, no. 10, pp. 10 232–10 251, 2014. [Online]. Available: http://www.mdpi.com/2072-4292/6/10/10232