Machine Learning with Satellite and Survey Data for Index Insurance

Andrew Hobbs and Aboli Khairnar

July 29, 2022

Abstract

Smallholder farmers in developing countries have always been adversely affected by year-to-year variation in weather patterns. Low rainfall, high temperatures, floods, and other disasters can wreak havoc on their livelihoods. Crop insurance has the potential to partially solve this problem, but traditional indemnity-based insurance is generally too costly to administer for smallholder agriculture. Index insurance, which provides payouts based on regional satellite, weather, or crop cut data offers a potential low-cost solution. However, developing accurate indices requires ground-truth data, which itself is costly to collect. This paper explores a new solution to this problem by combining existing household survey data from the World Bank's Living Standards Measurement Survey (LSMS) with satellite data to develop a hypothetical index for maize production in Uganda. We show that by combining remotely sensed data and machine learning techniques, we can construct an accurate crop production index. We compare regularized regression, neural networks, and random forests, and are able to obtain reasonably good yield predictions with neural networks and random forests. This method is a promising new approach for developing accurate index insurance products at low cost with large potential benefits for smallholder farmers.

1 Introduction

Smallholder farmer livelihoods depend on crop production and livestock health, and poor crop yields or animal deaths can lead households to consume less or lower quality food, remove children from school, or forego critical medical

visits. Effective risk management tools therefore have the potential to be enormously beneficial for smallholder households, but designing tools that are both effective and low enough cost to be economically viable has been a challenge. Index insurance is an approach to reducing risk for farmers whose potential is currently growing rapidly due to recent improvements in satellite data and machine learning methods. This paper proposes and tests a new approach to leveraging satellite data for better index insurance by linking it to household surveys that record historic crop yields. To our knowledge, it is the first paper to test the viability of this approach.

Index insurance differs from traditional indemnity insurance in that rather than paying individual farmers for the losses they actually experience, it pays farmers based on typical losses in a region. This has three advantages. First, it is much less costly to administer because individual claims do not need to be processed and investigated. Second, it mitigates the problem of moral hazard, which might lead farmers to put less effort into maximizing their yields if they know insurance payments will make up for shortfalls. Third, it eliminates the problem of adverse selection, which might lead farmers with unusually risky or low quality plots of land to opt into coverage.

Because payouts to farmers are based on regional averages, index insurance is only useful when yield shortfalls are sufficiently correlated at the regional level and when the index itself is accurate enough to map closely to farmer outcomes. As discussed by Carter et al. (2017) and others, when insurance indices do not track farmer yields closely enough their value to farmers is often very low; in many cases too low to justify paying the premium. It is therefore critical that even when indices are calculated at the regional level that they correlate well with farmer yields.

In recent years, the launch of new satellites has led to a dramatic increase in the frequency and resolution of satellite data. Higher resolution data means better yield estimates, which potentially means better index insurance. However, we are only beginning to explore the potential of these new data in this context, in large part because while there is a massive amount of satellite imagery, it is costly and difficult to obtain georeferenced crop yield estimates (labels) to link to satellite data. For a detailed summary of recent advances in satellite data and their potential impact for index insurance, see Benami et al. (2021). This paper proposes linking satellite data to existing household surveys for that reason: these surveys have been conducted for years, many are freely available, and they provide a set of 'labels' in the form of household level reports of crop

yields.

A number of existing studies have explored crop yield prediction by applying machine learning to satellite data (Bose et al., 2016; Gandhi et al., 2016; You et al., 2017; Cai et al., 2019; Mann et al., 2019; Kaneko et al., 2019; Qin et al., 2017; Kamilaris and Prenafeta-Boldú, 2018; Jiang et al., 2020; Chlingaryan et al., 2018). A specific application of machine learning and satellite data to index insurance is explored by Hobbs and Svetlichnaya (2020), who use forage quality data crowdsourced by pastoralists on the ground in lieu of the survey data we use here. Existing studies more broadly differ from this paper in that they tend to use use more precisely georeferenced survey data and are not focused on the index insurance context. Indeed, using data that is only georeferenced to the parish (or regional) level would not be useful for the many applications in which tracking individual farmer yields is the goal. However, in the index insurance case, it has the potential to be adequate since indices themselves are calculated at the regional level even when more detailed data exists. This paper is the first that we are aware of to test the viability of using existing survey datasets without exact georeferencing to construct an insurance index.

The remainder of this paper is organized as follows: Section 1 requires the data sources used in our analysis, Section 2 describes the methods used to predict yields with satellite data, Section 3 describes our models' performance, and Section 4 discusses the usefulness of our model and potential avenues for future work.

1.1 Data

Limited data quality and availability is an important barrier to developing better index insurance products using machine learning methods. Large, accurately labeled datasets of precisely georeferenced crop yields would be ideal, but collecting such data is expensive. Unlike in other domains, humans cannot accurately label aerial images of fields with accurate yields simply by looking at them. Perhaps the most accurate approach is crop-cutting, which means going to a specific plot, cutting down the crops, and recording the quantity produced for that plot. However, since productivity varies from farm to farm and even within farms by plot, putting together an ideal dataset would require doing this sort of harvesting and careful recording for every plot. While crop cuts are in some cases available, they generally correspond to a particular point in a particular

field, and the datasets that contain them are generally not very large.

In light of these challenges, this paper seeks to develop a reasonably accurate yield prediction model with the less-than-ideal data available from household surveys. These data are based on farmer responses, which have been shown to be prone to measurement error (Gollin and Udry, 2021). Thus, a major goal of this study is to test whether we can obtain a useful insurance index despite that measurement error.

An additional challenge associated with attaining quality data is the fact that new harvests only come into existence once a year. Thus, even if we were able to finance and begin an effort to attain better data today (which would be useful), it would take years before we could use the dataset for insurance indices. Farmers are facing droughts today, so finding a workable solution with data we can use now has large potential benefits.

1.1.1 Survey Data

Since approximately-georeferenced household survey data is relatively widely available and goes back quite far in time, it may be possible to link these data to satellite images to create accurate-enough models for identifying poor crop yields. In this study, we use World Bank's Living Standard Measurement Survey Integrated Surveys on Agriculture (LSMS-ISA) from Uganda. We selected Uganda because it has a large number of years available and a relatively large number of households over time.

We obtain crop yield estimates for between 579 and 1260 households per year for the years 2005, 2009, 2010, 2011, 2013, 2017, and 2018. In total, our sample contains 6,447 observations, which is low by machine learning standards but relatively high by household survey standards. We restrict our sample to maize because it is one of the primary staple crops in Uganda and is grown throughout the country; future studies testing this method with other crops would be useful. In the survey, maize production is reported at the plot level in kilograms along with the plot size in acres and the share of the plot used for maize. The total production figure is multiplied by the maize share of the plot and divided by the plot size to obtain maize produced per acre.

Importantly, the survey data we are using does not contain precise geographical locations; the most precise geographical information we have is the name of the parish in which the household is located. However, as shown in Figure 1, parishes in Uganda are quite small, so our goal is to obtain relatively accurate

predictions using this approximate geographic information. To that end, we link the survey data to satellite data aggregated by parish. Since many household survey datasets contain only this level of geographic specificity to protect respondent privacy, this same limitation will be present in most publicly available household survey datasets. Conducting our analysis this way is a good test of the viability of using household surveys for this purpose going forward.

1.1.2 Satellite Data

We link our survey data to an array of satellite images and satellite-derived indicators, combining vegetation indices with data on precipitation, temperature, and other weather variables. We collected all these data using Google Earth Engine.

For both weather data and other images, we follow a similar two-step process. For each parish, we focus on pixels designated as cropland by the Global Food Security Analysis-Support Data (GFSAD) (Teluguntla et al., 2015). We then aggregate those pixels across the parish and link them to households in the survey in the same parish and year. The GFSAD data and parish boundaries are shown in Figure 1; we aggregate other data across the pixels colored green within each parish outlined on the map.

We focus on two vegetation indices. The first, the Normalized Differenced Vegetation Index (NDVI) is among the most widely used for assessing vegetation health from space. We obtain these data from a LANDSAT 8-day composite made from Tier 1 orthorectified scenes. The second, the Green Chlorophyll Index (GCI) is obtained from MODIS Terra Surface Reflectance Daily Data. GCI has been shown to outperform NDVI in predicting crop yields in some settings, which is why we also include it (Burke and Lobell, 2017). We have data for each of these indicators at the pixel level (30m pixels for NDVI and 250m pixels for GCI), and we aggregate them by taking the monthly maximum value for each pixel and averaging those at the parish level so that we can merge it with the satellite data. Taking the pixel-level maximum before aggregation is a commonly used practice in estimating crop yields because it captures pixel when vegetation appears to be greatest (see, for example Lewis et al. (1998)). This is particularly important in harvest months, since after harvest the pixel NDVI is likely to be quite low even if crop yields were high.

We also collect weather data from two sources. Data on temperature and precipitation come from the European Centre for Medium-Range Weather Fore-

casts (ECMWF) ERA5 Dataset. We obtain data on root moisture and transpiration from NASA's Global Land Data Assimilation System (GLDAS) dataset, which combines satellite and ground-level sources. We aggregate these data by taking pixel means by month and then aggregating across the parish boundaries.

Figure 3 depicts annual trends in NDVI, GCI, and precipitation for the three parishes in which we have the largest number of yield observations. A clear annual trend is apparent in the precipitation figure, while the others appear quite noisy. However, despite their noisy, GCI and NDVI are among the most important features in generating accurate crop yield predictions.

Figure 2 provides an overview of the steps we took to aggregate the data, as well as how it is used in each of the two models to be described in detail in the next section. To summarize, by aggregating the satellite data at to the parish level, we are able to link it to the survey data to obtain regional yield predictions. We evaluate those predictions using individual-level survey data since it is the correlation between the index and individual level outcomes that matter most for insurance quality.

2 Method

Because the satellite data we use for this survey are collected at the parish level and our survey data generally includes multiple households per parish, 100% accuracy (or an R^2 of 1) is not possible given the data. However, this data situation reflects the real-world estimation goal of insurance indices: we want to estimate a regional index that is a good-enough approximation of individual yields.

Throughout our analysis, we start by splitting the data into two parts, randomly selecting 70% of the observations to form a training set and 30% to form a test set.

We begin our analysis by creating a benchmark based on an 'area yield' insurance contract. Area yield contracts are based on average regional yields rather than satellite or weather data, and have traditionally been administered via crop cuts across a region. Collecting these data is costly, but is generally considered to be more accurate. Constructing a theoretical area yield therefore provides a reasonable upper bound for what we might hope our satellite data will be able to achieve in terms of prediction.

After creating a benchmark, we test two different methods of predicting

individual yields from satellite data. First, we attempt to directly train an array of machine learning models to predict individual yields. Second, we try a two-step method in which we train models to predict parish averages and then test their ability to predict individual yields. As we will discuss in the Results section to follow, the second method is much more effective, likely because averaging yields at the parish level reduces noise and eliminates the problem of asking the model to predict multiple outcomes from a single set of inputs.

2.1 Benchmark

We construct our benchmark by first generating regional average yields. Let y_{it} represent the yield for farmer i year t and y_{Pt} the parish-year average1 for all farmers in parish P. Using our training set, we then estimate a simple linear regression

$$y_{it} = \alpha + \beta y_{Pt} + \epsilon_{it} \tag{1}$$

where α is a constant, β is the coefficient that relates farmer average yields to individual yields, and ϵ_{it} is a disturbance term. We evaluate the resulting model's predictive power by measuring the R^2 of predictions from the resulting model on the test set. This provides a reasonable upper bound for how well our satellite data based models might do, since this model uses the original data as an input. In other words, it provides a measure of how much of the variation in individual yields is explainable using regional data.

2.2 Predicting Individual Yields

In the next part of our analysis, we train an array of machine learning models. Specifically, we train a linear regression, an elastic net, a simple neural network, and a random forest to predict individual yields based on satellite data aggregated to the parish level. Each month's values for each variable is included as a feature, so there are 12 NDVI features, 12 GCI features, 12 precipitation features, 12 temperature features, 12 root moisture features and 12 transpiration features. In total, this means the models have 72 features to use to predict crop yields. For the elastic net and random forest, we conduct a grid search across possible hyperparameters. For the neural network, we test only a single simple architecture with four hidden layers, each of which contain 200-400 neurons. Thus, while the results show the random forest outperforms the neural network,

it is likely that with additional experimentation these results could be improved upon.

As discussed above, this is potentially challenging for the models, since the satellite data is the same for all households in the same parish-year. Presumably for this reason, we will show in the next section that this does not work particularly well; the two-stage approach we attempt next is much more successful.

2.3 Two-stage Prediction

Since satellite data are aggregated to the parish level, it is intuitive that they might be best suited to predicting parish-level average yields. Unfortunately, that is not our end goal. However, we can use a model trained to predict parish-level averages and test its performance on individual yields. We train the same set of models using the same approach as for the individual yield predictions and achieve much better results using this method. Specifically, we aggregate the training set by calculating parish-year mean yields per hectare and train the models on that dataset. We then test the model on individual yields in the test set.

3 Results

We evaluate goodness-of-fit using R^2 of our model's predictions on the test set. This metric has the advantage of simplicity and ease of interpretation; it can be thought of as the share of the variance in the outcome variable explained by the model. However, for more precise evaluation of index insurance products more sophisticated measures such as the average farm-level R^2 proposed by Stigler and Lobell (2020) or the nonlinear measures that take into account risk aversion proposed by Carter and Chiu (2018) ought to also be considered. Because this study seeks only to provide proof-of-concept and compare different models, we focus for now on traditional R^2 measures.

Our results are summarized in Figure 5. In summary, our two-stage approach which trains a model to predict parish averages and then tests that model's fit against individual yields is much more effective than our initial attempt to directly train a model on individual yields. Further, random forests and neural networks significantly outperform linear models.

3.1 Area Yield Benchmark

As described above, since we know our aggregated satellite data are can be linked to our surveys only at the parish level while our survey in many cases contains multiple observations in a given parish, it is not possible for our models to achieve a perfect fit. The best they could reasonably do is to equal an area-yield index, which is calculated using regional averages.[1] For that reason, as described above, we use the R^2 constructed from regional averages as the benchmark against which to compare other models. The area yield index achieves an R^2 of 0.30, meaning 30% of the variance in individual yields can be explained by regional averages. This indicates that there is significant variation that is idiosyncratic to individual farms, and it also means that building a high-quality index insurance product may be difficult: insurance is less valuable when it covers a smaller share of total risk.

3.2 Individual Yield Prediction

Predicting individual yields directly from satellite data worked quite poorly. An ordinary least squares regression achieved a dismal R^2 of just 0.04, our elastic net and neural net achieved just 0.05, and a random forest just 0.06. This is much worse than our benchmark, and suggests that attempting to predict individual yields with aggregate data is a poor strategy. The two-step strategy described next was much more effective.

3.3 Two-Step Yield Prediction

Our two-step yield prediction used a model trained to predict parish averages and evaluated its performance in predicting individual yields. The results were much more promising. While ordinary least squares achieved predictive accuracy of just 0.07, that figure still exceeded the best outcome from our initial approach. Our elastic net did no better. However, significant improvements were seen in the random forest and neural network, which both achieved predictive accuracy of 0.21 and 0.19, respectively, which compares quite favorably to our benchmark model's 0.30.

[1] Area yield indices are form of index insurance product that use regional averages as the insurance index. Insurance products based area yield indices are generally regarded as the most accurate products, but they are much more costly to implement than satellite data based contracts due to the cost of collecting data in the field.

The R^2 of 0.21 achieved by our best model is quite favorable given the challenges associated with predicting smallholder yields and the fact that we are attempting to predict individual yields using parish-level data. For comparison, Burke and Lobell (2017) achieve a maximum R^2 of 0.4 for predicting smallholder yields using data with exact farm locations, meaning they are able to focus on the precise pixels where the farm is located. The fact that we can reach an R^2 of 0.21 using only very imprecise location data is promising given the challenges associated with predicting smallholder yields.

Another way of evaluating our model performance is to measure its accuracy at predicting parish-level averages in the test set. In a sense, this is a 'fairer' challenge for the algorithm, since the data it uses to predict is at the parish level. Figure 4 shows the predictive accuracy of our two-step model at the parish level as well as the household level. Interestingly, the random forest outperforms the neural net at the parish level, but performs slightly worse at the household level. Again, since the benefits of insurance are at the household level, the household level measures are more important for determining which model would yield the most beneficial insurance product.

4 Discussion

The analysis above showed that satellite data aggregated to the parish level can predict regional and individual yields relatively well. Using simple methods we were able to predict 20% of variation compared to 30% for our benchmark 'area-yield' model based on actual regional averages. In other words, our analysis suggests that about 30% of variation in individual yields is due to factors such as weather that are common throughout a region. The remaining 70% is likely due to farm-specific factors which might include difference in soil, fertilizer usage, planting distances, seeds used, or myriad other factors. The 30% figure therefore represents an approximate upper bound for how well a model based on regional satellite data could do, and our model approaches that upper bound.

As discussed above, satellite data is much more cheaply available than farmer-reported yields. It is also available nearly instantly at any given time, meaning insurance programs based on it can react much more quickly in the event of droughts. Our benchmark model approximates the accuracy that would be achieved by a yield-based insurance program since it is based on yield data just as those programs generally are. Given the expense of obtaining farm-level data

and the fact that our simple models approached its level of predictive accuracy, we think this analysis suggests that models based on satellite and survey data could potentially be used to create low cost insurance products that would help farmers meaningfully manage risk. Of course, as reflected by the R^2 from our benchmark model, farmers face many individual-level risks that cannot be managed with regionally calculated insurance indices, but low cost coverage that can reduce the impact of regional shocks can still be beneficial.

This paper represents an initial attempt at using satellite and survey data to develop insurances where more precisely georeferenced farmer data are not available. The models we trained in this analysis were very simple; additional work using the same data could surely improve on these results and push our accuracy even closer to the area-yield benchmark. There are three areas in particular where improvement is likely possible. First, more careful feature engineering would likely yield improvements: focusing on just harvest months rather than all twelve months of the year, and including information on parish locations might allow the algorithm to learn that harvest dates vary throughout the country, for example. Second, our neural network was very elementary; experimenting with different architectures could surely build on our current results. Third, experimenting with gradient boosting and other methods could probably yield improvements.

Another promising avenue for expansion is to increase the size of our dataset by including survey data from other countries, especially nearby ones with similar farming systems. In particular, LSMS-ISA data are also available for Tanzania, and various household survey datasets are available for Kenya and Ethiopia. Combining several of these datasets would likely improve results, and could also ideally produce a model that would be accurate across East Africa. This sort of model could then be used as a starting point for future index insurance programs in the region.

Future work also ought to test the quality of these indices from a farmer welfare perspective as described by Carter and Chiu (2018). While this study has shown that machine learning methods and satellite data can get relatively close to an area-yield benchmark, it may be that even that benchmark does not explain enough of the variation in individual yields to produce an insurance product whose benefit justifies its cost. In that case, smaller regions than Ugandan parishes may be necessary to make a quality insurance product. Better data and methods cannot overcome problems caused by high within-region yield variation, meaning our benchmark R^2 of 0.3 is an upper bound for the

potential performance of parish-level contracts. This is a common challenge for all insurance indices and does not mean that the model is not useful. Rather, it reflects the fact that many risks affect farmers at the individual level and index insurance does not cover those risks.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

AH developed the methodology, identified appropriate datasets, oversaw the data analysis for this paper, and wrote the final report. AK conducted data collection and analysis.

References

Benami, E., Jin, Z., Carter, M. R., Ghosh, A., Hijmans, R. J., Hobbs, A., Kenduiywo, B., and Lobell, D. B. (2021). Uniting remote sensing, crop modeling, and economics for agricultural risk management. *Nature Reviews Earth and Environment*, 21:2:140–159.

Bose, P., Kasabov, N. K., Bruzzone, L., and Hartono, R. N. (2016). Spiking neural networks for crop yield estimation based on spatiotemporal analysis of image time series. *IEEE Transactions on Geoscience and Remote Sensing*, 54(11):6563–6573.

Burke, M. and Lobell, D. B. (2017). Satellite-based assessment of yield variation and its determinants in smallholder african systems. *Proceedings of the National Academy of Sciences*, 114(9):2189–2194.

Cai, Y., Guan, K., Lobell, D. B., Potgieter, A. B., Wang, S., Peng, J., Xu, T., Asseng, S., Zhang, Y., You, L., et al. (2019). Integrating satellite and climate data to predict wheat yield in australia using machine learning approaches. *Agricultural and Forest Meteorology*, 274:144–159.

Carter, M., de Janvry, A., Sadoulet, E., and Sarris, A. (2017). Index insurance for developing country agriculture: a reassessment. *Annual Review of Resource Economics*, 9:421–438.

Carter, M. R. and Chiu, T. (2018). Quality standards for agricultural index insurance: An agenda for action. Url: https://microinsurancenetwork.org/sites/default/files/SoM$_2$018$_W EB_f inal.pdf$.

Chlingaryan, A., Sukkarieh, S., and Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, 151:61–69.

Gandhi, N., Armstrong, L. J., Petkar, O., and Tripathy, A. K. (2016). Rice crop yield prediction in india using support vector machines. In *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 1–5. IEEE.

Gollin, D. and Udry, C. (2021). Heterogeneity, measurement error, and misallocation: Evidence from african agriculture. *Journal of Political Economy*, 129(1):1–80.

Hobbs, A. and Svetlichnaya, S. (2020). Satellite-based prediction of forage conditions for livestock in northern kenya. In *Workshop on Computer Vision for Agriculture (CV4A)*.

Jiang, H., Hu, H., Zhong, R., Xu, J., Xu, J., Huang, J., Wang, S., Ying, Y., and Lin, T. (2020). A deep learning approach to conflating heterogeneous geospatial data for corn yield estimation: A case study of the us corn belt at the county level. *Global Change Biology*, 26(3):1754–1766.

Kamilaris, A. and Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147:70–90.

Kaneko, A., Kennedy, T., Mei, L., Sintek, C., Burke, M., Ermon, S., and Lobell, D. (2019). Deep learning for crop yield prediction in africa.

Lewis, J., Rowland, J., and Nadeau, A. (1998). Estimating maize production in kenya using ndvi: some statistical considerations. *International Journal of Remote Sensing*, 19(13):2609–2617.

Mann, M. L., Warner, J. M., and Malik, A. S. (2019). Predicting high-magnitude, low-frequency crop losses using machine learning: An application to cereal crops in ethiopia. *Climatic Change*, 154(1-2):211–227.

Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, G., and Cottrell, G. (2017). A dual-stage attention-based recurrent neural network for time series prediction. *arXiv preprint arXiv:1704.02971*.

Stigler, M. and Lobell, D. (2020). Suitability of index insurance: new insights from satellite data. Technical report, Stanford University.

Teluguntla, P., Thenkabail, P. S., Xiong, J., Gumma, M. K., Giri, C., Milesi, C., Ozdogan, M., Congalton, R., Tilton, J., Sankey, T. T., et al. (2015). Global cropland area database (gcad) derived from remote sensing in support of food security in the twenty-first century: current achievements and future possibilities.

You, J., Li, X., Low, M., Lobell, D., and Ermon, S. (2017). Deep gaussian process for crop yield prediction based on remote sensing data. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Figures

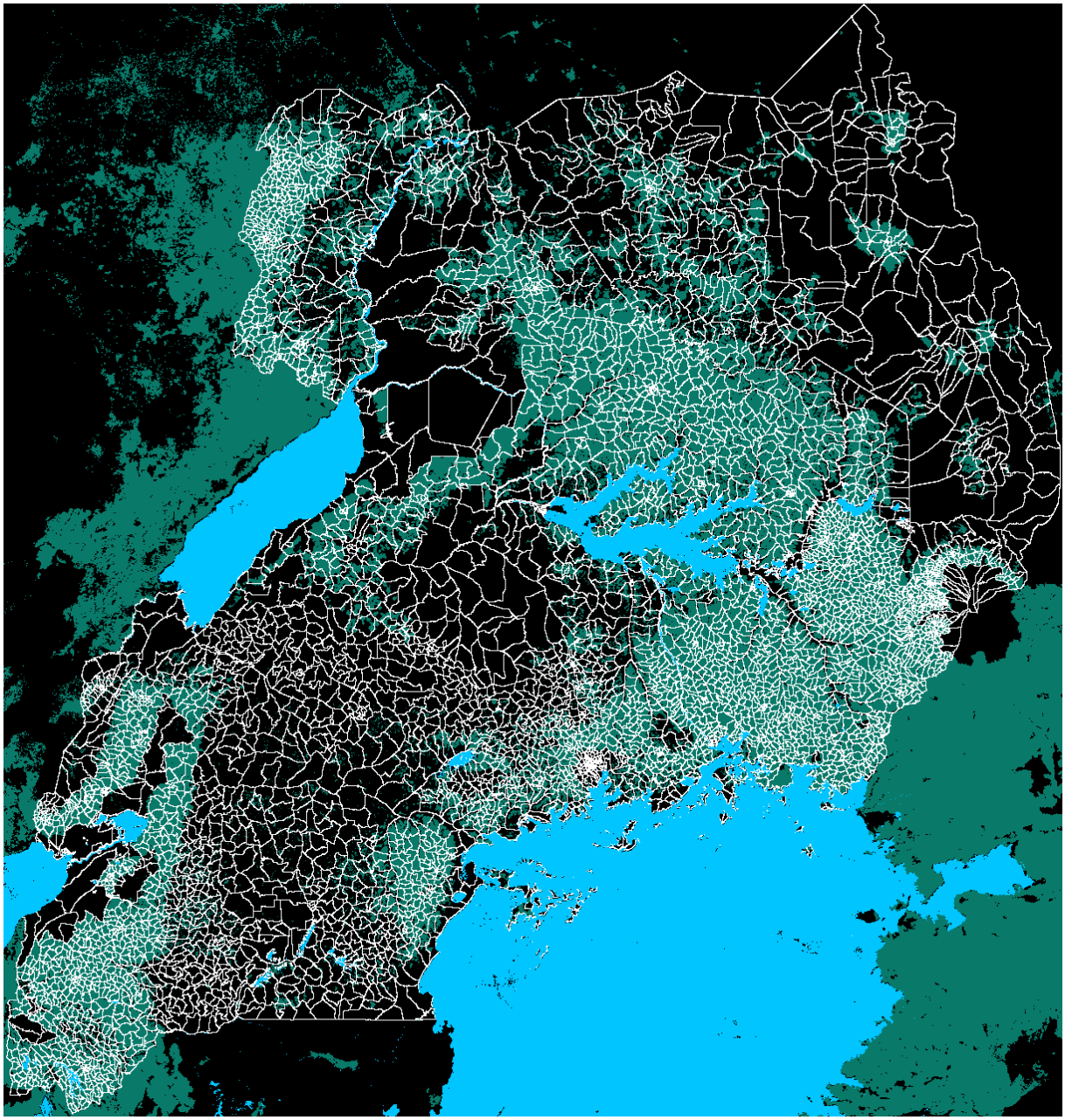

Figure 1: Cropland area in Uganda (green) with parish boundaries (white). All data in cropland pixels were aggregated at the parish level.

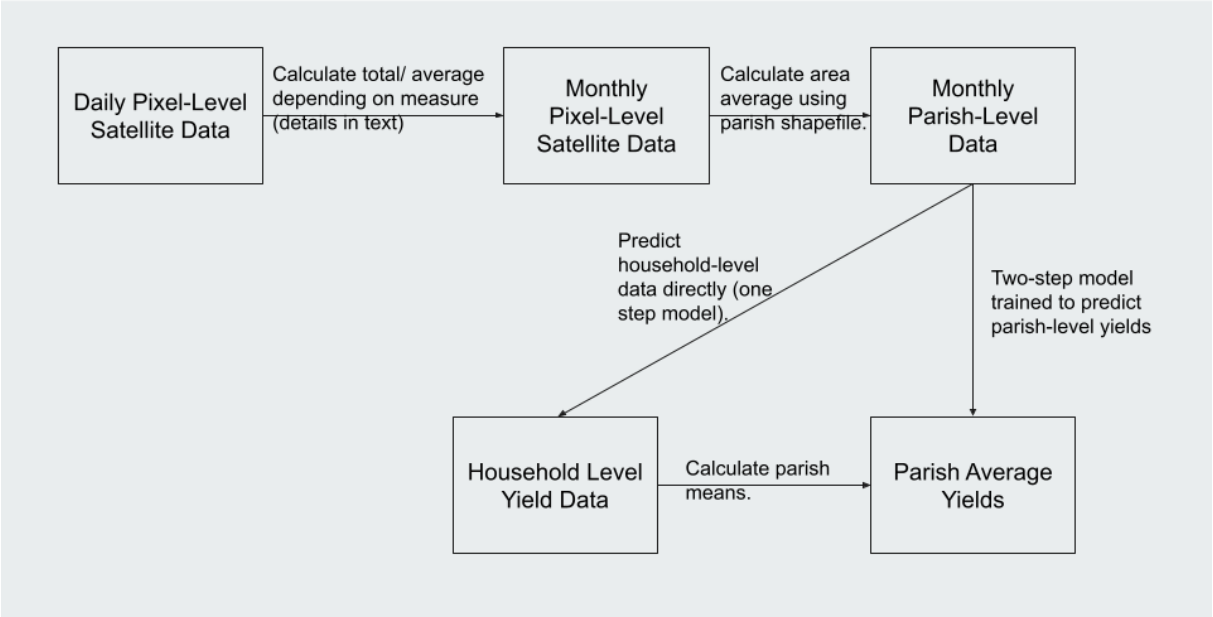

Figure 2: Overview of data aggregation and model training process.

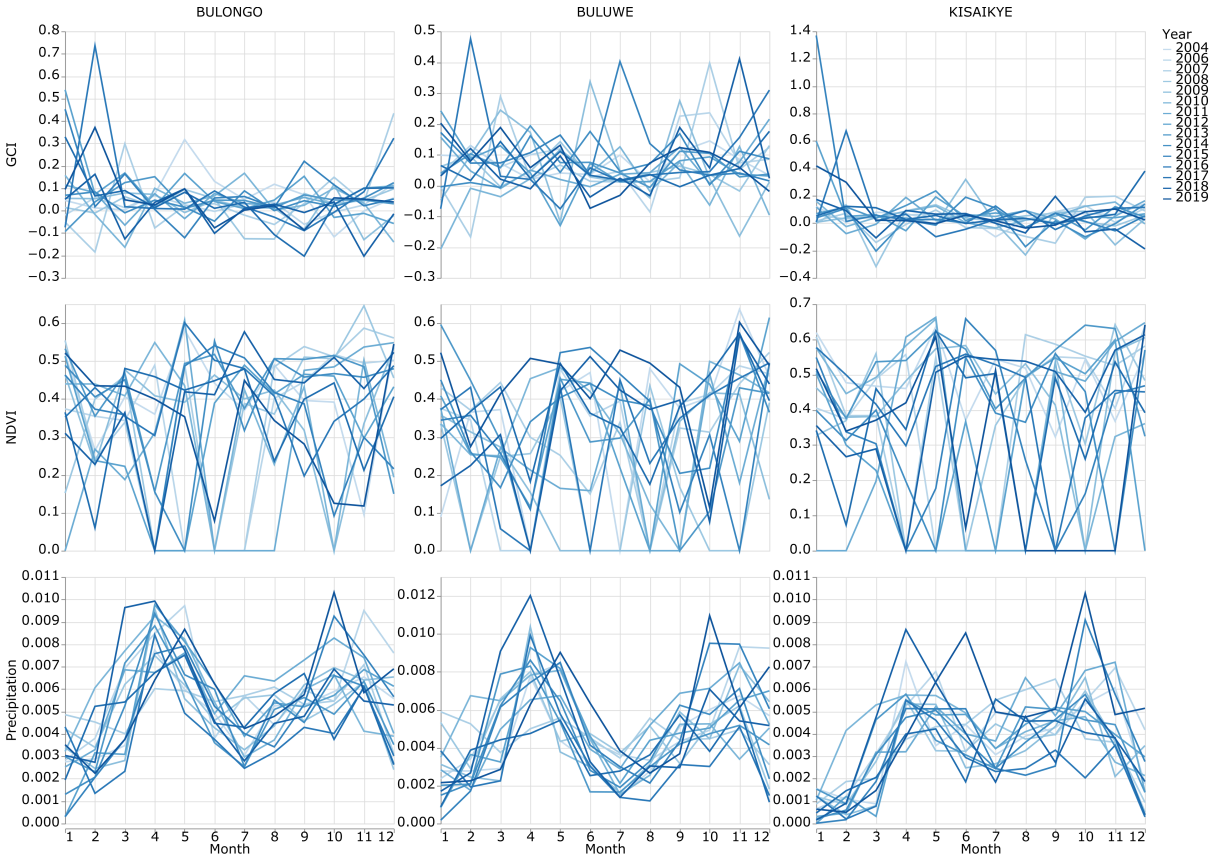

Figure 3: Annual trends in NDVI, GCI, and precipitation in the three parishes with the largest numbers of surveyed households.

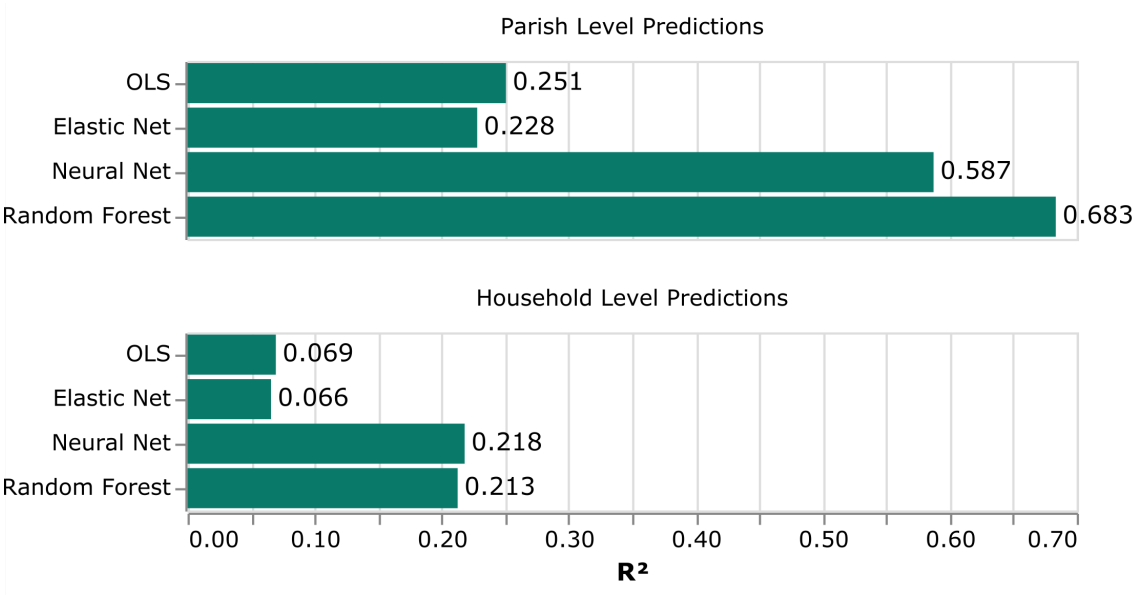

Figure 4: Two-stage models are trained to predict parish averages, and the most accurate model achieves an R^2 of 0.68 in predicting those averages. Predictions are less accurate at the individual level, as we expect, but still compare favorably to our area yield benchmark.

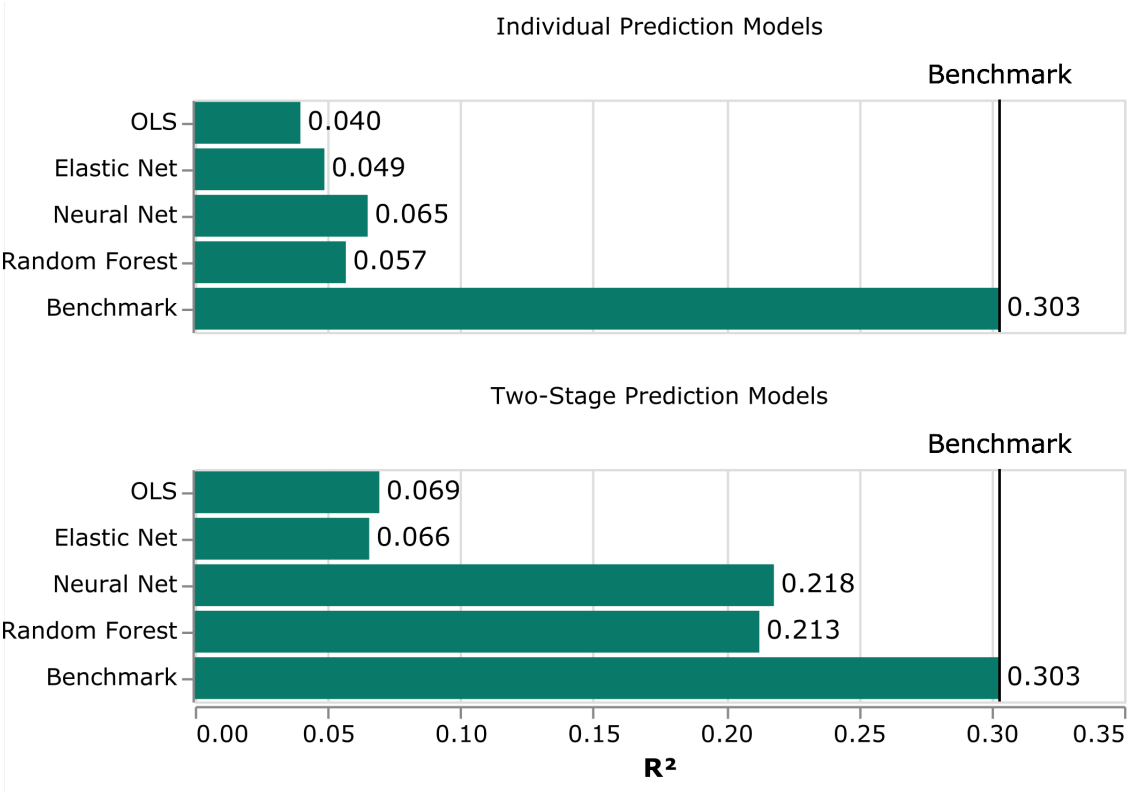

Figure 5: Two-stage prediction models, which were trained on parish average data but then used to predicting individual yields, significantly outperformed models directly trained to predict individual yields.