Using Satellite Data to Investigate the Relationship between Field Size and Agricultural Productivity in Eastern Uttar Pradesh, India

Gautam Mathur, Derek Van Berkel, Meha Jain

Abstract

This study investigates the relationship between field size and agricultural productivity in smallholder farming systems, focusing on three districts in Eastern Uttar Pradesh, India. We employ a Mask R-CNN image segmentation model and high-resolution satellite imagery to predict field boundaries, and estimate crop yield using maximum NDVI from Sentinel-2 imagery during the winter wheat growing season. Our findings reveal a slight positive relationship between field size and productivity across 7 of 8 zones in the study area when controlling for village-level effects. This is contrary to the consensus of an inverse productivity-size relationship in smallholder farming systems. This study provides a framework to investigate the relationship between productivity and field size at a large and dense scale, without the use of self-reported yield data. Limitations include the reliance on maximum NDVI as a yield proxy and the spatial resolution of Sentinel-2 imagery. Future research should explore alternative yield metrics and higher-resolution satellite data to refine our understanding of the productivity-size relationship in smallholder farming systems.

1. Introduction

Optimal agricultural productivity is essential to food security and development in regions with smallholder farming systems. The relationship between agricultural productivity and size has been heavily debated in the literature. Exploring and determining this relationship is important because trends in field sizes can be dynamic. In India, for example, data from agricultural censuses have shown that average sizes of agricultural holdings have decreased, from 2.28 hectares in 1970-71 (I.J Naidu 1971) to 1.08 hectares in 2015-16 (DEPARTMENT OF AGRICULTURE 2020). Additionally, agricultural and economic policy can influence how agricultural land is allocated, and therefore what the size distribution of agricultural plots in a region would be (Zeng et. al. 2018). To understand the implications of the current and possible future trends on food production, exploring the dynamics of the productivity size relationship is necessary.

The theory of an inverse productivity size relationship is largely assumed by agricultural scientists, popularized in the South Asian context by developmental economist Amartya Sen in 1962. The theory states that smaller farms have higher productivity. There have been multiple reasonings behind this hypothesis. Namely, labor intensity due to family labor, efficient land use and management due to limited land, and a greater incentive structure for laborers because of a direct ownership of land (Sen 1962). Numerous studies across various regions and timeframes have consistently demonstrated a negative relationship between farm size and agricultural productivity. In South Asia, the relationship has been found to be negative across three years from surveys in Bangladesh (Gautam and Ahmed 2019), and has also been proven in large studies in India (Gaurav and Mishra 2015). Outside South Asia, this relationship has been found

in numerous developing nations, such as Zambia (Kimhi 2006), Rwanda (Ali and Deininger 2015), and Brazil (Helfand and Taylor 2021).

The lesser adopted hypothesis of a positive relationship comes from the theory of economies of scale. According to this theory, because fixed costs stay constant for all fields, the per-unit cost of cultivating larger farms may be lesser. This may lead to a larger amount of capital for owners of larger farms, who can then invest more resources into their crop, leading to higher yield. This can be seen in areas where land consolidation has caused an increase in technical efficiency in agricultural production. (Zeng et. al. 2018).

While many studies have been conducted to assess this theory in multiple parts of the world, there have been important limitations to these assessments. Firstly, they have not been spatially representative, often occurring on a small number of fields relative to the total agricultural extent of study due to the use of costly household surveys as their data source. A second issue, which is also related to household surveys, is yield reporting bias. Self-reporting surveys have been known to be inaccurate (Carletto et al. 2015), for example due to errors such as heaping (the tendency to round up to multiples of 5 or 10), which can lead to large inaccuracies for small plots (Desiere and Jolliffe 2018). Studies have analyzed self-reported yields from household surveys, and have found that they are insufficient as inputs for remote sensing and economics- based modeling(Paliwal and Jain 2020, Abay 2020). In addition, self-report inaccuracies are associated with yields in smaller fields being over-reported, and yields in larger fields

underreported (Desiere and Jolliffe 2018). Such biases can lead to incorrect inferences about the relationship between field size and yield.

Remote sensing approaches may resolve both the issue of inaccurate self-reported yields, and small sample size. Satellite data are available globally, and different sensors have been used to accurately map field size and crop yield (Zhao et al. 2020, Paliwal et al. 2023). This is particularly important in smallholder systems, where household surveys would only be able to capture a fraction of all farmers and fields. For smallholder farming systems, remote sensing allows the digitizing and cartography of millions of small fields to be done remotely, which saves time and labor, and allows for a much larger coverage.

In this study, we use satellite data to determine the relationship between productivity and field size in three districts in Eastern Uttar Pradesh, India. India is an ideal region to investigate this relationship. With a fast-rising population and existing food scarcity, it is important to understand how the fragmentation, and possible consolidation of farms in India will affect the productivity of agriculture in the nation. While our results are specific to India, the methods applied can be used to more broadly understand the relationship between field size and yield in other agricultural systems across the globe.

2. Study Area

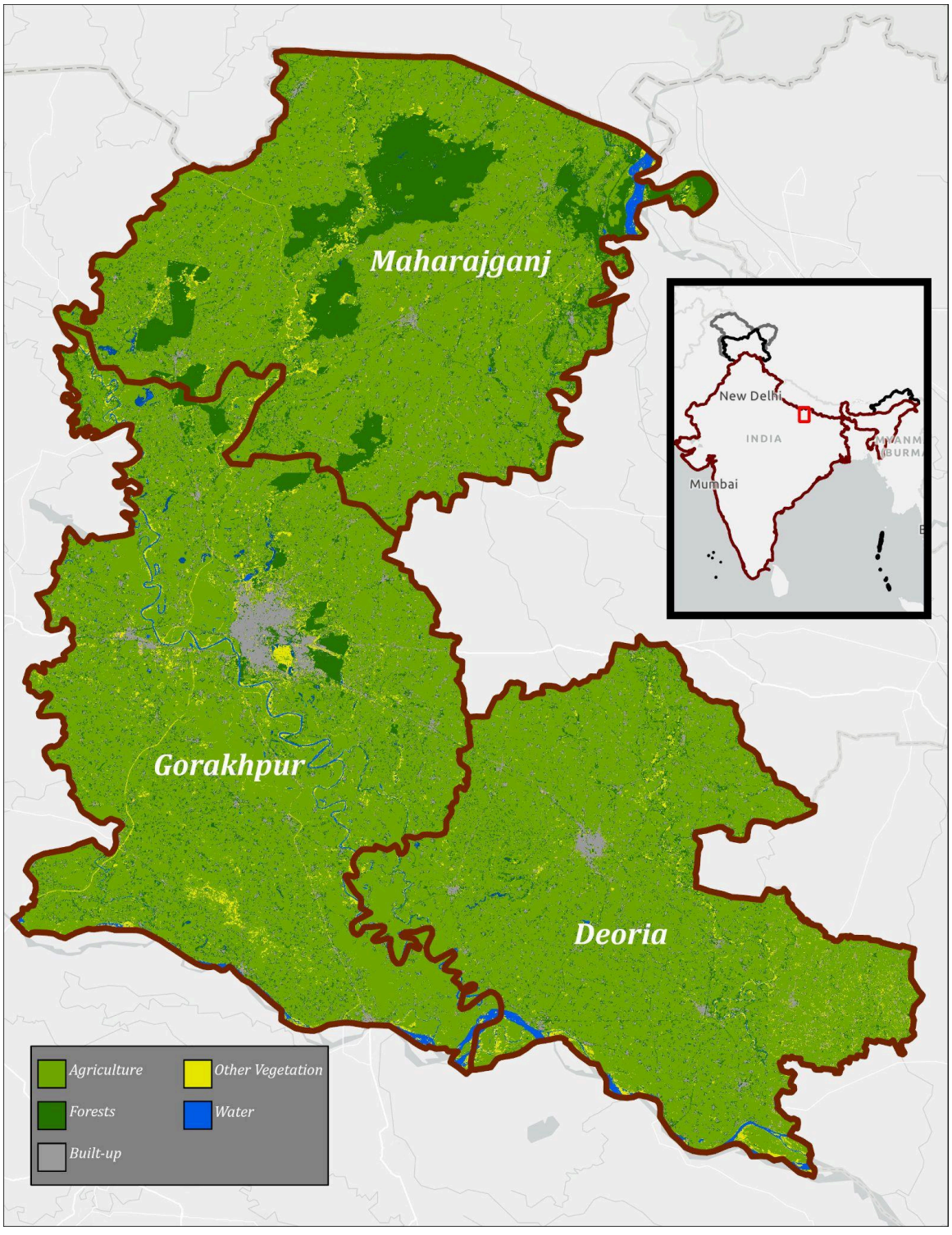

Figure 1. Map of the study area depicting land cover, with an inset map showing the position of the area within India. Map Source: ESRI Topo World. Land Cover Source: ESA WorldCover

The three districts in Uttar Pradesh that make up the region of interest (ROI) are Maharajganj, Deiora, and Gorakhpur (Figure 1). These districts are located in the central-eastern region of the Indo-Gangetic plain. They make up a total area of 8,882 square kilometers. Agriculture covers 74.9% of the total land cover in the districts (European Space Agency 2021). In Uttar Pradesh, 65% of the population's employment is in the agricultural industry. The agriculture in this region, as in the rest of India, is primarily a smallholder system. The average agricultural holding size in Uttar Pradesh as of 2016 is 0.73 hectares (DEPARTMENT OF AGRICULTURE 2020). There are 2 main farming systems in this region- Rabi (winter season), and Kharif (monsoon season). Most crops in this region follow a rice-wheat system, with rice grown in the Kharif season, and wheat grown in the monsoon season.

3. Methods

We detail the steps needed to predict field boundaries (Section 3.1), predict crop yield (Section 3.2), compile the dataset for analysis (Section 3.3), and conduct statistical analyses to identify the relationship between field size and crop yield (Section 3.4).

3.1) Predicting Field Boundaries

We predicted field boundaries using Mask R-CNN and high-resolution WorldView and Quickbird imagery (Mei et al. 2022). Mei et. al. (2022) used Digital Globe images to train, validate, and test a model with these goals in Bihar, India, and the model was able to draw field boundary polygons with a precision of 0.73 and an F1 score of 0.7. This model was then tested in a new area in Uttar Pradesh without any new training data, and a similar level of accuracy was achieved, with a precision of 0.79, and an F1 score of 0.75 (Mei et al. 2022). This model creates the ability to digitize a large number of

smallholder farms with a great reduction in time and effort, and makes large-scale field-level analyses more feasible.

3.1.1) High-resolution Satellite Data

Worldview and Quickbird level 1B imagery were acquired from the DigitalGlobe collection for the three districts in Eastern Uttar Pradesh, India. These images were available as panchromatic images (0.5 m ground sampling distance), and multispectral images with NIR, red, green, and blue bands (2 m ground sampling distance), with panchromatic counterparts.

Our choice of imagery for field boundary segmentation was driven by multiple factors. Initial evaluations found that image segmentation predicted field boundaries with highest accuracy during active agriculture months, which range from July to March. We also prioritized multispectral imagery over panchromatic due to findings in a previous study on their efficacy (Mei et. al. 2022). Maximum coverage of the area was also prioritized. We chose a total of 63 from the months of October, March, and February in 2018, with the majority of the images being from February (Figure 2). All but one of the images were multispectral. One panchromatic image was chosen due to unavailability of multispectral imagery from desired months in the corresponding area. The images in total covered 93.6 % of the region of interest.

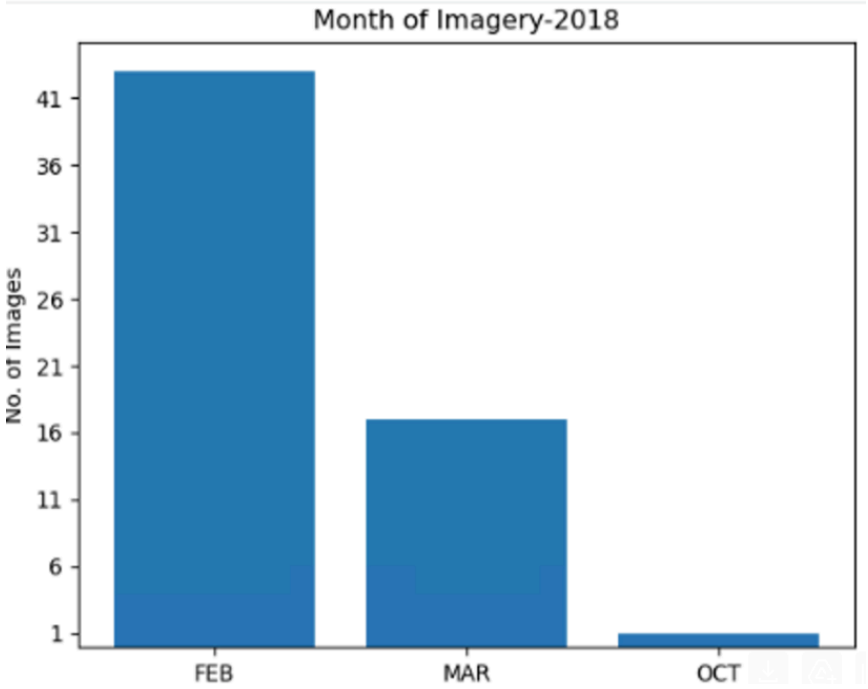

Fig 2: Number of Images in Selected Worldview and Quickbird Image Collection Per Month

3.1.2) Preprocessing Images

The images acquired had minimal cloud cover, were radiometrically corrected, and sensor corrected.. Geometric correction was performed on these images for further improved accuracy. A Digital Elevation Model (DEM) for the region was acquired from the Shuttle Radar Topography mission (NASA (National Aeronautics and Space Administration) 2000). This DEM had a spatial resolution of 3 arc seconds, and was modified to have a 90 m ground Sampling Distance. Using the “Geometric” function from the ArcGIS library of raster functions (Esri Inc. 2023), the DEM was used to produce orthorectified versions of all selected images, projected in UTM Zone 44N. For multispectral imagery, the same function was applied to the panchromatic counterparts of the images.

The panchromatic images had a spatial resolution of 0.5 m, whereas the multispectral images had a spatial resolution of 2 m. To increase the spatial resolution of the multispectral images, they were pan-sharpened using their panchromatic counterparts. This was done in Python using GDAL's pansharpen.py function (GDAL/OGR contributors 2024), and resulted in multispectral images with the spatial resolution of 0.5 m. To input into the model, Python's GDAL library was used to split all images into 512 X 512 image chips, which were also rescaled to have an 8-bit radiometric resolution (0-255 pixel values).

3.1.3) Inputting images into the Mask R-CNN Model

The image chips were input into a Mask R-CNN model trained for field boundary detection (Mei et. al., 2022). This is an instance segmentation model that was trained using hand-digitized field boundaries to segment high-resolution satellite imagery to isolate individual agricultural fields. The output of this model is a raster that is then converted to shapefiles representing detected agricultural fields. The training site for this model was 200 km southeast from our study region. In this study, this model was extrapolated to our region of interest (63 images across 3 districts in Uttar Pradesh, India).

3.1.4) Post-Processing Polygons

The field detection method resulted in data gaps at the edges of each 512 X 512 image chip. To reduce inaccuracies from this phenomenon, fishnet shapefiles for every image were created that reflected the borders of the 512 X 512 image chips. Using Python's Geopandas and Shapely libraries, the maximum and minimum x and y coordinates for

each polygon were calculated. We selected any of these coordinates that were within 2 meters of the corresponding fishnet border, and were likely affected by data gaps at the edge of the respective image chip. We then modified these coordinates to instead have the coordinates of the border of the image chip. Geopandas' "unary_union" and "explode" tools (Jordahl et al. 2020) were then used to merge the polygons together to a single multipolygon, and separate all non-intersecting sections into separate polygons (Figure 3).

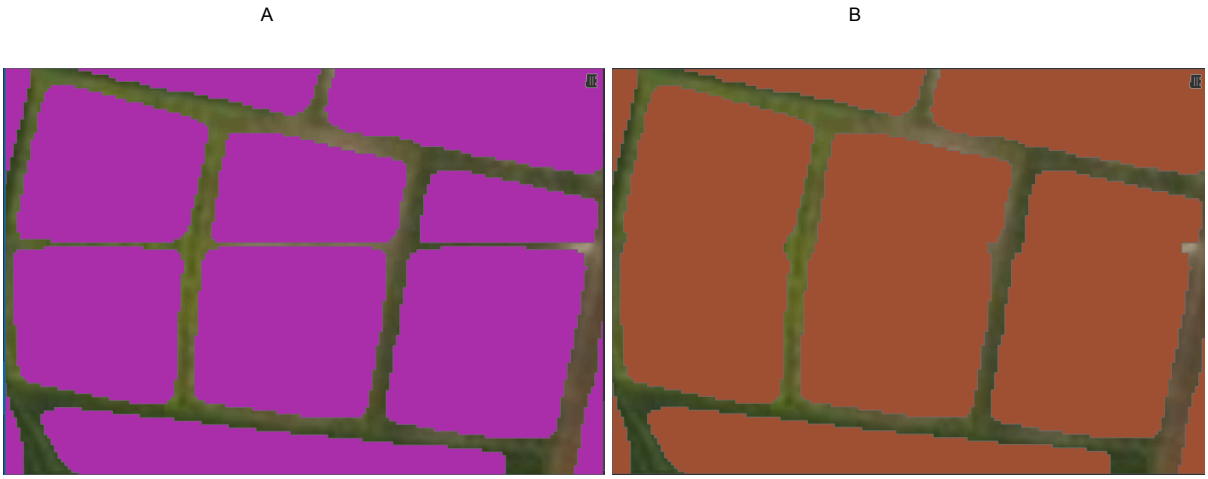

Fig 3: An example of affected polygons A) before and B) after applying the merging function

We visually inspected the resulting polygons and found that the function worked best in dense agricultural areas, and in areas with the most accurate field boundary predictions.

3.2) Predicting Crop Yield

Crop yield for the winter wheat crop was estimated using maximum NDVI of Sentinel-2 Level 2A imagery during the 2019-2020 Rabi (Winter) growing season (October - April). Previous studies have shown that maximum NDVI during the growing season is a good proxy for the yield for grain crops. (Johnson 2016, Liu et al. 2020, Wuepper et al. 2023). We calculated maximum NDVI for our field polygons using Google Earth Engine's

Python API (Gorelick et al. 2017), and GEE's "Quality Mosaic" and "ReduceRegions" functions. This resulted in a raster of maximum NDVI values for the region of interest (ROI), between the months of October and May, 2019-2020 and mean NDVI values for the selected field boundary polygons.

3.3) Compilation and Organization of Data

3.3.1) Shapefile Creation

To manage the heavy data processing workload of a large number of fields, The ROI was split into 8 random zones by merging boundaries of villages within the region of interest (Figure 4). In cases where there were multiple overlapping tiles for the same location, we selected tiles with better prediction accuracy or those from similar months. Next, shapefiles were filtered to only include polygons that 1) had their centroid within the corresponding zone boundary, and 2) did not intersect with polygons that were imported before. ArcGIS Pro (Esri Inc. 2023) was used to visualize the accuracy of predictions for each image, and to create a single merged shapefile for each zone.

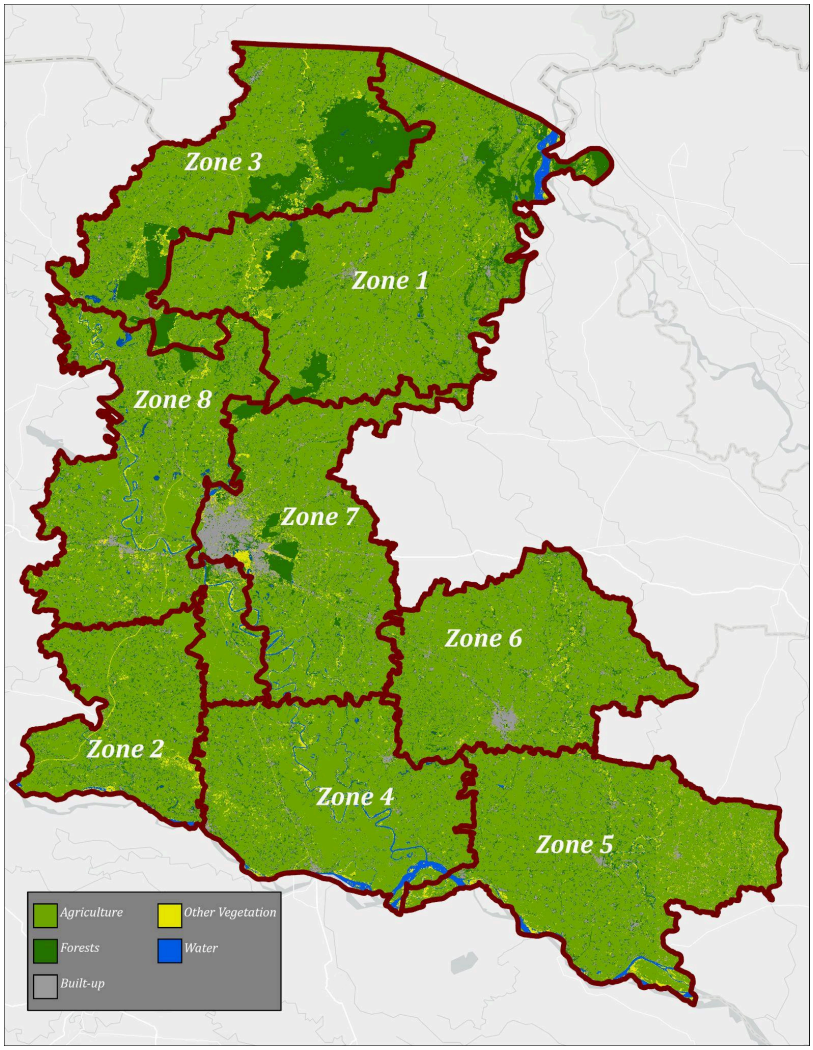

Fig 4: Division of region of interest into 8 zones, shown with land cover (Source: ESA WorldCover)

3.3.2.) Filtering Polygons

Once the shapefiles were created, polygons were further filtered to remove some inaccurately predicted polygons. We removed fields smaller than 85 m^2 as a visual inspection of subsets of the data showed that fields smaller than this size were mostly inaccurately predicted polygons. We also removed polygons that occurred in non-agricultural land cover, specifically forests and water bodies, using land cover

classes from WORLDCOVER (European Space Agency 2021), a 10 m resolution global land cover dataset. Additionally, areas near large rivers with inaccurate predictions were hand digitized and these polygons were also removed. Polygons that were not correctly merged during the gap processing of small image tiles were removed. Based on visual inspection, we identified these as polygons that had one or two vertices greater than 12 m in length. Finally, given the large size of our dataset, we selected only 10% of all fields from each village for our analysis.

3.4) Statistical Analysis

To test theories of field size and yet we fit linear regressions model of field area size and maximum NDVI. One regression was run for each zone, and the regressions were run using village as a fixed effect. This controlled for any factors that varied across villages, such as soil type, irrigation access, and farmer wealth. All analysis used the 'PLM' package (Croissant and Millo 2008) in the R Project Software (R Core Team 2021)

4. Results

The area coefficient of the fixed effects regression was examined. For all zones but one, the coefficient was a small positive number that was statistically significant. For zone 3, the coefficient was a small negative number that was statistically significant. Our results therefore showed a positive productivity size relationship for 7 out of 8 zones, and a negative productivity size relationship for 1 out of 8 zones.

Zone	Coefficient	Sign	P value
1	2.0566E-6	+	1.191E-7
2	6.4251E-6	+	<2.2E-16
3	-7.0536E-6	-	<2.2E-16
4	8.9821E-6	+	<2.2E-16
5	9.0275E-6	+	<2.2E-16
6	5.0061E-6	+	5.745E-16
7	7.3144E-6	+	<2.2E-16
8	7.3307E-6	+	<2.2E-16

Table 1: Coefficient values, signs, and p Values for the "Area" variable in fixed effect regression for each zone

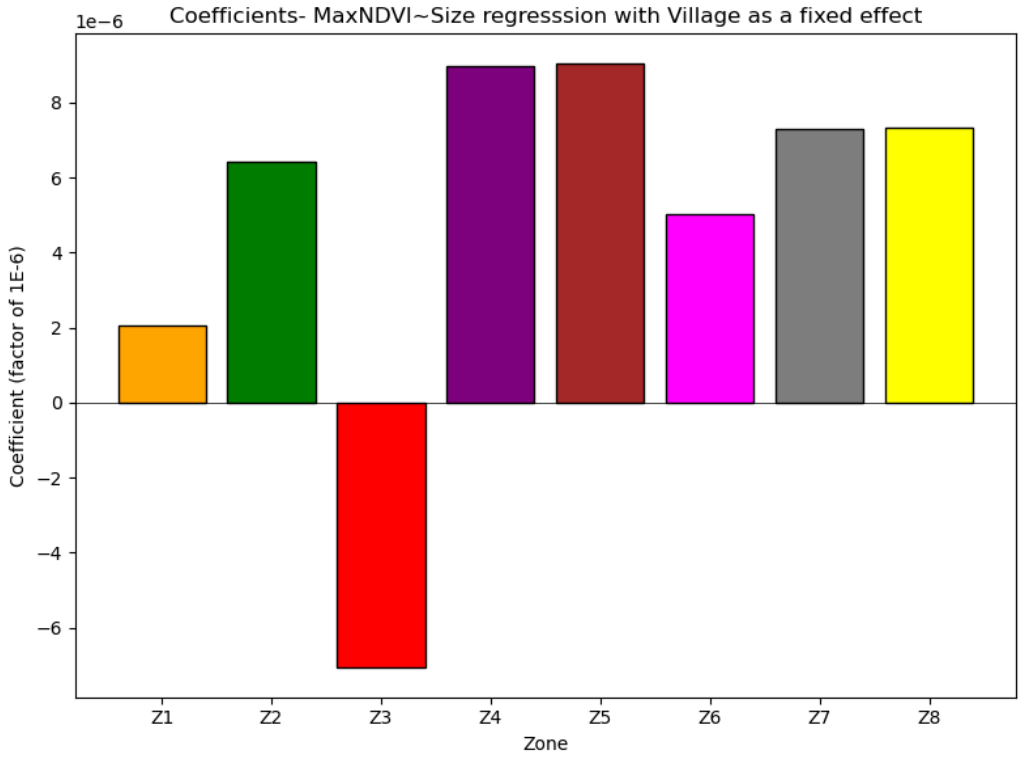

Fig 5: Bar Chart Depicting Coefficient Values for "Area" Coefficient Per Zone

5. Discussion

By finding a positive relationship between field size and productivity in 7 out of 8 zones, our findings are contrary to the popular theory of an inverse productivity-size relationship, and demonstrate the need to reconsider this hypothesis in the context of smallholder farming systems. While the magnitude of these coefficients is small, the significant p values depict that in these zones, an increase in field size adds to productivity, along with the multitude of factors that determine the productivity of an agricultural field.

While zones 1, 2, and 4 through 8 showed a significant positive relationship, zone 3 showed a significant negative relationship. The most prominent differences in zone 3 compared to the rest of the zones were that 1) the profile of land cover in the zone was visibly different from the rest of the zones. Zone 3 had much more forested area, with 25.6% of the zone covered in dense vegetation. This may signify differences in climatic and environmental variables that affect yield, and a larger number of selected polygons in the area close to dense vegetation. 2) Zone 3 also had the lowest average area compared to all the zones, with an average of 610 square meters, whereas the mean of area for the rest of the zones was 890 square meters.

The positive relationship in the majority of zones may be due to economies of scale, and may not be reflected in studies with self-reported yields, due to the previously discussed human errors that come with collecting self-reported data.

One of the limitations of this study is the estimation of yield. Though studies have shown a high correlation between maximum NDVI and field size, there are other metrics that have fared better than maximum NDVI in yield prediction, and/or can improve the accuracy when modeled along with maximum NDVI. The first example of this are crop phenology-based characteristics such as sow date, green-up rate, and senescence rate(Bolton and Friedl 2013). In addition, other vegetation indices that use the red-edge band, such as the Chlorophyll Index (CI), have been shown to more accurately capture yield than other vegetation indices (Zhao et al. 2020). We did not incorporate either of these in our yield metrics as sow date, green up rate, and senescence rate require analyzing time-series data, which is more computationally expensive than the quality mosaic method used in this study. Furthermore, we did not use vegetation indices based on the red-edge band in this study as it has a spatial resolution of 20 m, which is coarse relative to the size of fields in this region.

Yield measurements in our study were also limited by the spatial resolution of Sentinel-2 imagery. The area of one Sentinel-2 pixel is 100 square meters. A 0.09 hectare field, which is close to the average field size in most zones, would be covered by only 9 pixels. This may result in mixed pixels, where pixels on the edges of the field are drawing information from neighboring fields. Using higher resolution imagery, such as Planet imagery that has a pixel area size of 9 square meters, could reduce this problem. Such imagery is not available on an open-source platform, such as Google Earth Engine, and was therefore not used for this study.

Another limitation to consider is the discrepancy in the years between different imagery sources. Due to image availability constraints, the Quickbird and Worldview images used for predicting field boundaries were from 2018, whereas the sentinel-2 images used for estimating yield were from the 2019-20 season. This study assumes minimal to no change in the spatial characteristics of fields within the period of one season.

Lastly, the performance of the field boundary prediction model, reflected by an F1 score of 0.75 in our region of interest, suggests the potential for misinterpretations of field boundaries. While we applied methods to remove and repair fields misinterpreted by the model, some inaccuracies may still exist in the dataset.

6. Conclusion

In conclusion, we found that for 7 out of 8 zones in the region of interest, there was a slightly positive, significant relationship between area and productivity (represented by maximum NDVI). At this level, these results are contrary to the productivity area inverse theory. In further studies, this relationship should be closer explored by controlling for factors that influence yield. In addition, yield estimates can be improved by using alternate satellite imagery with higher spatial resolution, and considering red-edge bands that are more predictive of yield. Future work should examine whether the detected relationships remain after these improvements.

WORKS CITED

Abay, K. A. 2020. Measurement errors in agricultural data and their implications on marginal returns to modern agricultural inputs. Agricultural Economics (United Kingdom) 51:323–341.

Ali, D. A., and K. Deininger. 2015. Is There a Farm SizeProductivity Relationship in African Agriculture? Evidence from Rwanda. Land Economics 91:317–343.

Bolton, D. K., and M. A. Friedl. 2013. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. Agricultural and Forest Meteorology 173:74–84.

Carletto, C., D. Jolliffe, and R. Banerjee. 2015. From Tragedy to Renaissance: Improving Agricultural Data for Better Policies. Journal of Development Studies 51:133–148.

Croissant, Y., and G. Millo. 2008. Panel Data Econometrics in R: The plm Package. Journal of Statistical Software 27:1–43.

DEPARTMENT OF AGRICULTURE, C. & F. W. M. O. A. & F. W. G. O. I. 2020. All India Report on Agriculture Census 2015-16.

Desiere, S., and D. Jolliffe. 2018. Land productivity and plot size: Is measurement error driving the inverse relationship? Journal of Development Economics 130:84–98.

Esri Inc. 2023. ArcGIS Pro . Esri Inc.

European Space Agency. 2021. ESA WORLDCOVER 2021.

Gao, X., A. R. Huete, W. Ni, and T. Miura. (n.d.). Optical-Biophysical Relationships of Vegetation Spectra without Background Contamination.

Gaurav, S., and S. Mishra. 2015. Farm Size and Returns to Cultivation in India: Revisiting an Old Debate. Oxford Development Studies 43:165–193.

Gautam, M., and M. Ahmed. 2019. Too small to be beautiful? The farm size and productivity relationship in Bangladesh. Food Policy 84:165–175.

GDAL/OGR contributors. 2024. GDAL/OGR Geospatial Data Abstraction software Library.

Gorelick, N., M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment.

Helfand, S. M., and M. P. H. Taylor. 2021. The inverse relationship between farm size and productivity: Refocusing the debate. Food Policy 99.

I.J Naidu, A. S. to the Govt. of I. M. of A. 1971. All India Report on Agricultural Census 1970-71.

Johnson, D. M. 2016. A comprehensive assessment of the correlations between field crop yields and commonly used MODIS products. International Journal of Applied Earth Observation and Geoinformation 52:65–81.

Kimhi, A. 2006. Plot size and maize productivity in Zambia: Is there an inverse relationship? Agricultural Economics 35:1–9.

Liu, J., T. Huffman, B. Qian, J. Shang, Q. Li, T. Dong, A. Davidson, and Q. Jing. 2020. Crop Yield Estimation in the Canadian Prairies Using Terra/MODIS-Derived Crop Metrics. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13:2685–2697.

Mei, W., H. Wang, D. Fouhey, W. Zhou, I. Hinks, J. M. Gray, D. Van Berkel, and M. Jain. 2022. Using Deep Learning and Very-High-Resolution Imagery to Map Smallholder Field Boundaries. Remote Sensing 14.

NASA (National Aeronautics and Space Administration). 2000. Shuttle Radar Topography Mission (SRTM).

Paliwal, A., Balwinder-Singh, S. Poonia, and M. Jain. 2023. Using microsatellite data to estimate the persistence of field-level yield gaps and their drivers in smallholder systems. Scientific Reports 13.

Paliwal, A., and M. Jain. 2020. The Accuracy of Self-Reported Crop Yield Estimates and Their Ability to Train Remote Sensing Algorithms. Frontiers in Sustainable Food Systems 4.

R Core Team. 2021. R: A Language and Environment for Statistical Computing. Vienna, Austria.

Sen, A. 1962. An Aspect of Indian Agriculture. Economic Weekly Annual Number 14.

Wuepper, D., H. Wang, W. Schlenker, M. Jain, and R. Finger. 2023. NBER WORKING PAPER SERIES INSTITUTIONS AND GLOBAL CROP YIELDS.

Zhao, Y., A. B. Potgieter, M. Zhang, B. Wu, and G. L. Hammer. 2020. Predicting wheat yield at the field scale by combining high-resolution Sentinel-2 satellite imagery and crop modelling. Remote Sensing 12.