Identifying Potential Positive Deviants (PDs) Across Rice Producing Areas in Indonesia:

An Application of Big Data Analytics and Approaches

June 2020

PULSE LAB JAKARTA

MANCHESTER 1824
The University of Manchester

DATA POWERED POSITIVE DEVIANCE

GIZ DATA LAB

Authors

Basma Albanna
Data Science Fellow, Pulse Lab Jakarta, United Nations Global Pulse
Doctoral Researcher at the Centre for Digital Development, Global Development Institute, University of Manchester

Dharani Dhar Burra
Data Scientist, Agriculture and Food Systems, Pulse Lab Jakarta, United Nations Global Pulse

Michael Dyer
Geospatial Information Systems Officer, Pulse Lab Jakarta, United Nations Global Pulse

Authors' Contributions

BA developed the initial research proposal and the main conceptual idea. BA and DB jointly designed the method and the analytical approach with input from MD. MD forged key partnerships to obtain the administrative data and the Earth Observation data that was used in the analysis. MD and DB scoped, cleaned, preprocessed and prepared Earth Observation and related GIS data for the identification of outliers, as well as prepared the maps that are included in this report. BA sampled and preprocessed the census data and then combined the data set with the Earth Observation data to create the homologous environments and identify the outliers. Inputs on those steps were provided by DB and MD. BA implemented the outlier validation using bivariate analysis; MD implemented the outlier validation using the Google Time Scale tool; and DB implemented the outlier validation using time series Earth Observation data. The results were discussed and interpreted jointly by BA, DB and MD. BA led the writing of this report in collaboration with DB and MD. All authors jointly identified the limitations of the developed method and provided recommendations for future work.

Acknowledgments

The authors acknowledge that Pulse Lab Jakarta, through its partnerships and resources, provided the facilities and opportunity to undertake this research. Special recognition to members of Pulse Lab Jakarta's team, as without their efforts and contributions, this research would not be possible: in particular, Awan Diga Aristo and Angga Gumilar for engaging with the Ministry of National Development Planning (Bappenas) and obtaining access to relevant data sets; Hendrick for structuring the data and providing rich technical insights; and Faizal Tharmin for forging partnerships to gain access to the geospatial administrative data and providing valuable insights. We would like to thank the United Nations Global Pulse for introducing Basma to PLJ through its Data Fellows programme that connects doctoral researchers with relevant expertise to UN entities. Thanks also to Benny Isanto (World Food Programme - Indonesia) for sharing his knowledge on different remote sensing methods. We are grateful for the comments and suggestions offered by Dwayne Carruthers and Utami Diah Kusumawati who reviewed an earlier draft of the report. Special thanks to Sriganesh Lokanathan, data innovation and policy lead at Pulse Lab Jakarta, for his intellectual guidance and mentorship throughout the project; and Richard Heeks, Director of the Centre for Digital Development, University of Manchester, for overseeing Basma's work throughout her fellowship. We would like to thank GIZ Data Lab for their financial support for this project, which is part of the collaborative Data Powered Positive Deviance initiative. We are grateful for the generous support from the Government of Australia and its continued support of Pulse Lab Jakarta.

Table of Contents

Introduction

Public and private organizations working in agriculture development, depend largely on field surveys to identify plot-level management and household-level social, economic, and demographic drivers that determine agricultural productivity. In developing countries, where smallholder agriculture predominates, new and efficient ways of data collection are needed to measure agricultural performance. Current data collection methods fail to capture the complexity of production systems and the varied nature of households that manage these systems, across both spatial and temporal dimensions. For example, traditional field surveys such as the national agricultural census, do not capture certain contributing drivers of productivity, for instance biophysical conditions (e.g. temperature, precipitation, etc.). Nonetheless, earth observation (EO) data has made it possible to map and monitor proxies of croplands and their biophysical environments, which when combined with field surveys and big data analytics, can be used to better characterise the complexity of those production systems. For instance, previous research (Tucker and Sellers, 1998; Mkhabela et al. 2011; Bolton & Friedl 2013; Johnson 2014) used the well-established relationship between net primary production and satellite derived measurements of plant phenology such as Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI); which are both measures that are used to quantify vegetation greenness. Those studies demonstrate that the EVI measure is a valid proxy for crop (annuals) yields. And when combined with field surveys, an EVI measure provides an opportunity to better understand the determinants of agricultural productivity, both in terms of controllable factors (i.e. factors farmers can control) and uncontrollable factors (e.g. biophysical variables) captured from EO data. Improved understanding can lead to the identification of over and underperformers (i.e. households/villages with significantly higher and lower agriculture productivity), with relatively higher confidence, from which interventions can be designed to target underperformers using insights gleaned from practices and strategies of the overperformers. Those individuals or groups that are overperformers are referred to as "positive deviants" (PDs), and adopting their practises and strategies on a wider basis is referred to as the "positive deviance" (PD) approach (Sternin, 2002).

Identifying the overperformers within similar contextual environments was first introduced in 1976 in the case of child stunting to identify dietary practices developed by mothers in low-income families who had well-nourished children (Wishik & Van Der Vynckt, 1976). But it wasn't until the early 2000s that PD was promoted as an effective asset based approach for social development after its successful application in rehabilitating an estimated 50,000 malnourished children in 250 communities in Vietnam (Sternin, 2002). The PD approach starts by discovering the over performing individuals or communities; following that their underlying practices and strategies are determined; based on which interventions are designed to scale those successful practices from the PDs to the under performers. The underlying assumption is that PDs implement unusual practices and strategies that could provide novel insights to solve complex problems, which conventional solutions fail to solve (Cinner, 2016). This type of positive deviance analysis has been also applied in the agricultural domain (Noble et al., 2005; Pant, 2009; Steinke, 2019) and studies show that by analysing what defines a PD within an agricultural community -- with "similar" biophysical, socio-economic, and demographic conditions -- certain drivers can be repeated or introduced and specific constraints can be removed. Drivers could be external agents, innovative technologies and practices and should require a minimum level of human, social, financial, physical, or natural capital. For instance, in the Brazilian state of Parana, some agricultural communities adopted no-till as a better cultivation method and after demonstrating an increase in productivity, income, and sustainability, the practice was adopted across the state (de Vries, 2005).

To the best of our knowledge and according to a recent systematic review on the combined used of PD and big data for development (Albanna & Heeks, 2019), previous PD studies focusing on agricultural development, have not combined EO data with administrative data (e.g. agricultural census) to identify PDs and to understand possible drivers of their agricultural performance. In this study, we propose, and trial a method, which combines those two kinds of data, to identify and validate villages (surveyed in the 2013 Indonesian Agricultural Census) that perform substantially better, in terms of agricultural productivity, than their peers despite having similar socio-economic, biophysical and environmental conditions. We used univariate and multivariate outlier detection techniques for PD identification and used administrative data to understand possible drivers of performance. To validate and denoise the identified PDs, we used the Google earth time scale tool and EO time series data, which further helped us in filtering false PDs from true PDs. Although the scope of this study doesn't fully answer the question "why are some villages faring better than other similar villages" which would require

extensive ground surveys, it paves the way for this type of inquiry through providing a spatial targeting method that could significantly reduce the time and cost needed to identify potential PDs in agriculture.

Data

We relied on two types of data: 1) official administrative datasets i.e. the Agricultural Census Data (2013), and the Village Potential Census data (2014); and 2) EO data that was used to identify homologous environments (i.e. groups of villages that belong to the same biophysical environment). Non-controllable factors, i.e. day temperature and precipitation, that determine agricultural productivity were sourced from Land Surface Temperature and Emissivity data products at 1 km^2 spatial resolution, and monthly temporal resolution, captured since 2004, from NASA's MODIS (Moderate Resolution Imaging Spectroradiometer; Land Surface Temperature and Emissivity (MOD11)), satellite, and CHIRPS (Climate Hazards Group InfraRed Precipitation with Station data), at 0.05 degree arc seconds spatial, and monthly temporal resolution, captured since 1960. In addition, we used earth observation corrected Enhanced Vegetation Index (EVI; derived from MODIS;MODIS/Terra Vegetation Indices Monthly L3 Global 1km SIN Grid (MOD13)), captured at 250 m spatial resolution, and at 16 day time intervals, captured since 2004, as a performance measure, which has been extensively used as a proxy for agriculture performance (Son et al. 2014).

When the research was being conducted, the best available data was relatively outdated (2013 and 2014) but since this work focuses on method development, and subsequent identification of PDs, it was still practical to use this data. In the agriculture census, every house involved in agricultural activity, across Indonesia, is taken as the observational unit. The census contains more than 700 variables related to agriculture production (e.g. type of crops and irrigation systems) per household spanning 18 million agricultural households in Indonesia for each cropping season of the calendar year 2012/2013 for each household. It is important to note that information is not georeferenced and is collected at the household level and not at the plot level. In order to merge the census data with the EO data, we used the unique village geocode to aggregate all the household data to their respective villages. With rice being a major staple food among Indonesians (Hartini et al., 2005), it is cultivated on significantly larger areas compared to other crops. Therefore, we focused this analysis on rice producing households in the census. In addition, it is relatively easier to estimate productivity of annual crops (such as rice), using proxies derived from EO, as annual crops exhibit "strong" seasonality and clearly distinct temporal features from other land use types. Variables directly related to rice farming practices were selected and aggregated for every village. Specifically, agricultural census data for the third season of the cropping calendar (i.e. between January to April 2013) was used for the analysis. The agriculture census does not capture yields, due to its differential design and end use, therefore to circumvent this issue, we used EVI, aggregated to the village as the performance measure.

In parallel, to include socio-economic and demographic information about the households, we used the 2014 village potential survey (PODES). The survey is collected by a different directorate (from the directorate that conducts the Agricultural Census) in the Ministry of National Development Planning (BAPPENAS) by interviewing the head of each village in Indonesia. The 2014 PODES data had more than 800 variables relating to village characteristics such as water resources, public services and facilities, market assets, etc. Similar to the agricultural census data, the PODES data is not geotagged. We selected variables of potential relevance (selected variables can be found in the Appendix) from the agricultural census and PODES, and aggregated the agricultural census data to the village level (a mode function was used for categorical variables and a proportion or average function was used for numerical variables), and joined both datasets with the unique geocode for the village. Notably, selection of variables is a context specific activity, and was dependent on the research question.

Since yields were not captured in the census data, we relied on a commonly used EO metric - the corrected Enhanced Vegetation Index (EVI) - as a proxy for agricultural performance, because it is a well-known measure of plant greenness or leaf area index (Son et al., 2014) and was developed to optimize vegetation signals in regions with high biomass and has less saturation from when compared to the NDVI (Huete et al. 2002; Qiu et al 2013). To derive the EVI, we used MODIS satellite imagery, which has global coverage and a 250 m spatial resolution, with a 16-day revisiting interval. Across the agricultural season between January and April 2013, a global EVI 250m pixel resolution raster layer, was clipped to the extent of Indonesia and was extracted in the

Mercator projection, for each time point. Across the raster brick, the Maximum value for each pixel, across all time-points, was extracted. Since we aggregated the agriculture census and PODES data to the village, the Maximum EVI value for each pixel was also aggregated to the village.

EVI values are sensitive to crop types and the obtained Maximum EVI values could reflect other crop types that are grown in the village. To control for this source of error, we extracted the average Maximum EVI values for the rice growing areas within village boundaries with a rice crop mask provided by the Indonesian Ministry of Forestry. For the purpose of aggregation, Maximum EVI values for each pixel were averaged across all pixels that belong to a village. Village boundary data in a shapefile format were provided by the Indonesian Bureau of National Statistics.

Monthly rainfall data for the season between January 2013 and April 2013, was obtained as raster layers (0.05 arc seconds) from CHIRPS. This is an open source globally gridded dataset, containing 35 years of rainfall data, produced and maintained by the University of California San Diego and USAID. CHIRPS is a hybrid data product, that grids global weather station data, and interpolates the weather station data, with satellite-based precipitation estimates, obtained from NASA's Global Precipitation Missions (GPM) and NOAA's CPC merged analysis of precipitation (CMAP). Although this smart data-fusion approach removes systematic bias associated with the weather station data, the CHIRPS dataset still suffers from issues such as underestimated precipitation measures in complex orography. For this analysis, monthly temperature data for the selected cropping season was downloaded as individual raster files. For each month, the rasters were merged with Indonesia's official administrative boundary shapefile, and values of pixels belonging to each village were averaged separately for each month, to obtain monthly average rainfall (in millimeters) for each village.

The monthly temperature data from Land Surface Temperature and Emissivity data product of MODIS, at 1km spatial resolution, was downloaded as raster files, separately for each month within the selected cropping cycle. To obtain average monthly temperature values, the raster files were merged with the administrative boundary shapefile, and the pixel values averaged across, and extracted at the village level. The temperature data from the MODIS sensor was obtained in digital numbers (DN), which were then converted to temperature in degrees centigrade, by multiplying the DN value with 0.02 (i.e. scale factor), to obtain temperature in Kelvin, and then subtracting with 273.15 to obtain temperature values in degree centigrade.

In summary, the following datasets were used for the analysis:

- Average monthly rainfall data (in millimeters for the cropping season between January to April 2013) from Climate Hazards Group InfraRed Precipitation with station data (CHIRPS),

- Average monthly temperature data (in degree centigrade, for the cropping season between January to April 2013) from the Land Surface Temperature and Emissivity data product of MODIS

- Averaged Maximum EVI for each village obtained from bi-weekly MODIS EVI data produced every 16 days for the cropping season between January to April 2013

- Land use 2014 rice crop mask data intersected with the village boundaries in order to extract EVI of the rice areas in each village

- 2013 Indonesia Agricultural Census Data capturing agriculture production data for more than 18 million households that are involved in agriculture for the seasons between years 2012 and 2013

- 2014 Indonesia Village Potential Data for 82,000 villages

- 2014 Administrative boundaries of villages data (bureau of statistics)

The following datasets were used for subsequent validation of the results:

- Monthly aggregates of precipitation (in millimeters) from January 2001, until December 2015, for Indonesia at 5 square kilometer (0.5 arc seconds) resolution, from Climate Hazards Group InfraRed Precipitation with station data (CHIRPS)

- Monthly aggregates of day time temperatures (in degree centigrade) from January 2001, until December 2015, for Indonesia at 4 square kilometer resolution, from the MOD11 data product of MODIS

- Monthly aggregates of EVI from January 2001, until December 2015, for Indonesia at 1 square kilometer resolution, from the MOD13 data product of MODIS

Figure 1 presents the rice growing area in Indonesia in the year 2014. It is important to note that the rice areas might be slightly different in 2014 than what is stated in the 2013 census data. We used the rice mask to reduce the error of capturing EVI values from non-rice areas instead of using if we captured EVIs for the entire village. Seen in Figure 1, this rice mask covers both types of rice, the wetland and the dryland, without differentiating between the two. Therefore our sample included villages growing both types of rice.

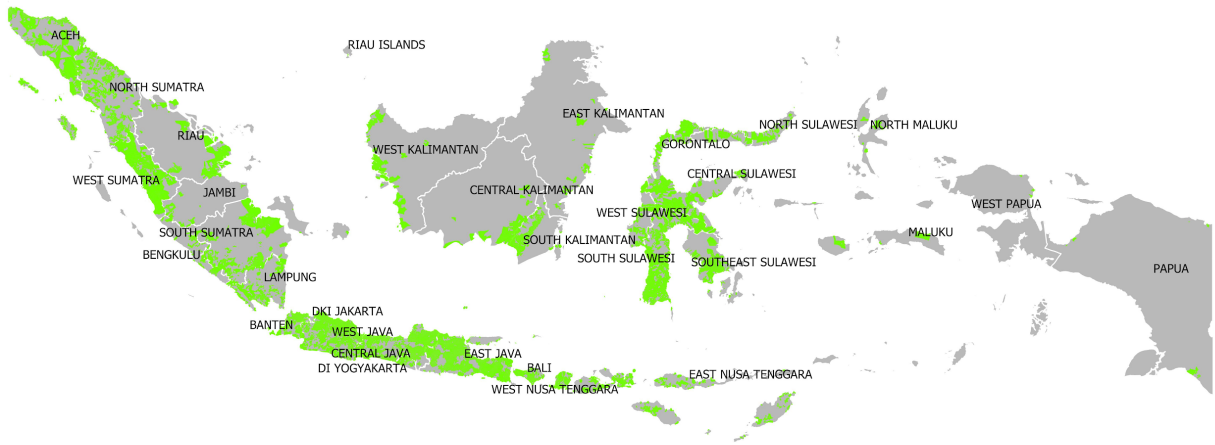

Figure 1: Rice Crop Mask in Indonesia - 2014

Study Sample

Our data sample included villages in Indonesia having at least one household growing any type of rice (i.e. dryland or wetland rice) in the third season (i.e. January to April 2013). Combined with average monthly rainfall and temperature, and average Maximum EVI during the cropping season between January to April 2013. According to the agricultural census, there are 41,664 villages growing rice in Indonesia. However, we were able to extract EVI values for only 18,978 villages. To prevent cross-signal issues from other crops, we used a rice mask layer for each village, to extract average Maximum EVI values. However, the rice mask was recorded for the year 2014 and had limited metadata about which cropping season and rice variation it represented. The significant reduction in the number of villages in the census, for which EVI values could be obtained, can be attributed to this temporal mismatch between the rice crop mask layer and the agricultural census data. It is also possible that the rice mask for 2014 represented a different cropping season than the cropping cycle selected for the analysis, or there could be a temporal shift in the amount of rice production in 2014, compared to 2013. The total of 18,978 villages was further reduced to 17,517 villages, due to missing temperature and/or rainfall data or not being captured in the PODES census data. Our final data sample of 17,517 villages constituted a total of

4,051,416 households growing rice, wherein each village had data from the agricultural census, PODES, average monthly rainfall and temperature data, and an average Maximum EVI, all for the cropping cycle between January and April 2013.

Methodology

The primary objective of this study was to develop a method that combines EO data with existing, varied administrative data, to identify rice villages that are performing significantly better than rice villages having similar conditions. To test the viability of the proposed method it is important to validate the identified PDs by checking whether they are true PDs or not. Figure 2 provides a summary of the approaches used for potential PD identification, which are explained in further details in the sections to follow. In this study, all the data analysis was done using the statistical software R v3.4.1, and QGIS v3.8 and ArcGIS v10.7.1 for spatial processing.

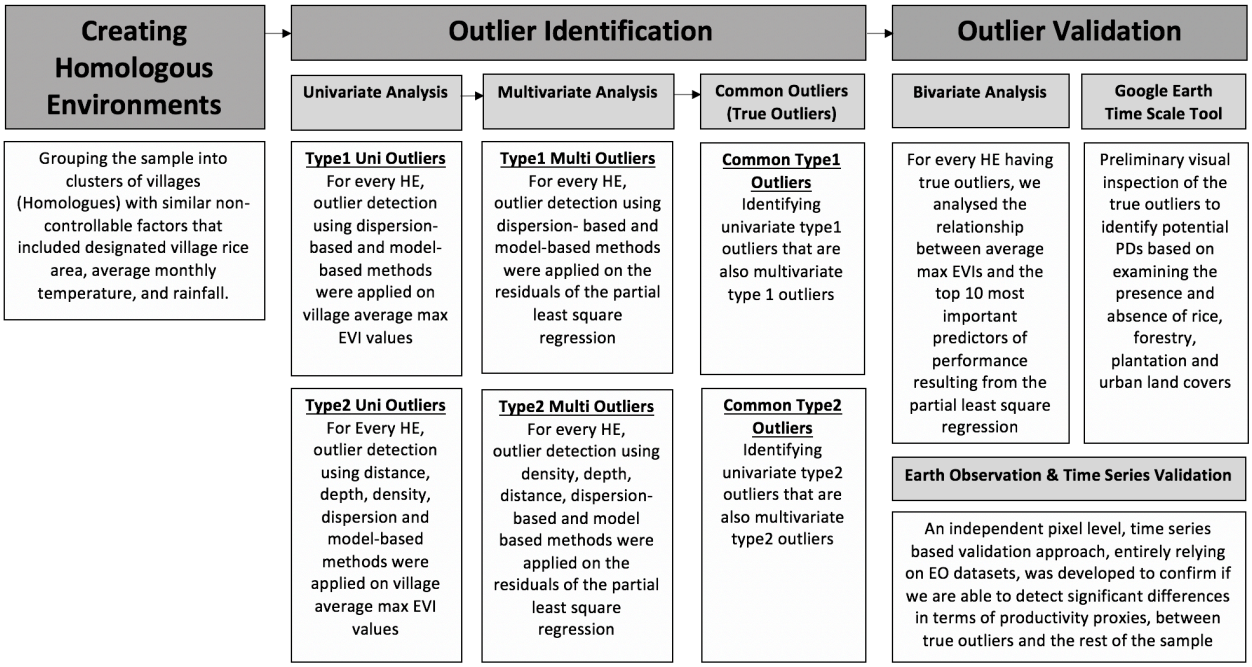

Figure 2: A summary of the approaches used for potential PD Identification and Validation

Creating Homologous Environments

The first step of this methodology was to create homologous environments (HEs) i.e. grouping the 17,517 villages, into clusters of villages with similar non-controllable factors, that included designated village rice area, average monthly temperature (in Degree Celsius) and rainfall (in millimetres). The total rice area was clustered using the *densityMclust* function of the mclust R Package into 5 clusters, such that each cluster would have groups of villages with similar total rice areas. The number of clusters (k=7) was determined automatically based on the score of the Bayesian Information Criterion (BIC). The number of villages and the mean total rice area in every cluster is shown in Table 1.

Area Clusters	Number of Villages	Mean of Village total rice area (meter square)
Cluster 1	815	4,231
Cluster 2	1,881	49,495
Cluster 3	6,158	2,811,223
Cluster 4	6,459	1,108,841
Cluster 5	2,204	3,922,707

Table 1: Village clusters based on rice area

Next, when clustering villages based on biophysical environments (i.e. based on average temperature and rainfall), we performed hierarchical clustering with principal components using the *HCPC* function in the factominer R Package. This function performs an agglomerative hierarchical clustering on results from a factor analysis. The first step was to perform a principal component analysis (PCA) on a dataset containing 8 columns of normalized. PCA looks for a few linear combinations called "Principal components" (PCs) that can be used to summarize the data without losing too much information (Maitra & Yan 2008). Ordered quantile normalization was done using the *orderNorm* function in the bestNormalize R package We chose to retain the three principal components that were able to explain 80% of the variance using the "elbow" method applied on a scree plot. The output of the PCA was then passed to the *HCPC* function which clustered the villages into three biophysical clusters containing 7135, 5348, and 5034 villages respectively. Based on the area clusters and biophysical environment clusters, we developed 15 HEs, wherein each homologous environment consisted of villages having similar rice area and biophysical conditions as shown in table 2.

Homologue	Biophysical Cluster	Area Cluster	Number of Villages
11	1	1	259
12	1	2	667
13	1	3	2,532
14	1	4	2,943
15	1	5	734
21	2	1	163
22	2	2	436
23	2	3	1,361
24	2	4	2,438
25	2	5	950
31	3	1	393
32	3	2	778
33	3	3	2,265
34	3	4	1,078
35	3	5	520

Table 2: Number of villages in each homologue

Outlier Identification

Before identification of potential PDs in each HE, we used two different, yet complementary approaches, to identify outliers within each HE. The two approaches are as follows:

1. Univariate analysis approach, where outlier villages in each HE are identified based on the performance measure, i.e. average Max EVI value for each village
2. Multivariate analysis, where outlier villages are identified in a relative sense using multiple control variables (the selected controllable factors from Agricultural census and PODES), that could determine average Max EVI values for each village

The final set of outlier villages would be those that consistently appear in both types of analysis. The reason we perform two types of analysis was to ensure that outliers are identified with higher confidence, after controlling for multiple factors, and are not merely identified by chance. The universe of factors, used as controls in this study, are derived from a comprehensive list of factors, that are collected repeatedly, across time by governments. If a village continues to be an outlier, with (multivariate analysis) and without the use (univariate analysis) of these controllable factors, it would suggest, that practices in the outlier villages are different from the rest of the villages, within the same HE, and are not completely captured by the current universe of factors, which further increases the chances of identifying potential PDs, from this list of outliers, in the later stages of the analysis.

Univariate Analysis

In this analysis, for every HE, the average Max EVI values, for all the villages, were used to identify outliers. As expected, the average Max EVI values of villages within an HE didn't follow a normal distribution, further indicating the presence of complex, non-uniform production systems within HE. Therefore, a method that identifies observations as outliers, solely based on the assumption of normal distribution cannot be used. Instead we used an outlier detection approach, that uses multiple measures, in addition to the distribution of the performance measure, to identify outliers. We used the *OutlierDetection* function in the OutlierDetection R Package, which identifies outliers using a combination of different methods. Since the focus is to identify over-performing villages, we term them as positive outliers. Two types of outliers were identified using this function:

- **Type1:** Outliers are identified using the default *OutlierDetection* function which finds outlier observations using dispersion based and model based methods for outlier detection. The total number of positive outliers, across all HEs, identified using this method is 144 villages.
- **Type2:** Outliers are identified using the *OutlierDetection* function but after adding depth, density and distance methods to the dispersion-based and model-based methods for outlier identification. It can be considered as a narrower filter for outlier detection, hence, it resulted in a smaller number of outliers. The total number of outliers identified using this method is 29 villages.

Table 3 presents the distribution of Type1 and Type2 outliers across the 15 homologues. The above approach was unable to identify outliers in certain HEs (e.g. "11", "21", "22" and "31"). Alternatively, this approach yielded only Type1 outliers and not Type2, in certain HEs (e.g. 12 and 35).

Homologue	Number of Villages	Type1 outliers	Type2 outliers
11	259	0	0
12	667	4	0
13	2,532	27	2
14	2,943	22	4
15	734	4	2
21	163	0	0
22	436	0	0
23	1,361	1	1
24	2,438	20	1
25	950	15	5
31	393	0	0
32	778	2	2
33	2,265	30	7
34	1,078	11	5
35	520	7	0
Total	**17,517**	**144**	**29**

Table 3: Number of PDs in each homologue

Multivariate Analysis

In the previous approach, we used only the average Max EVI value, to identify positive outliers, but we didn't consider other drivers of agricultural performance that could have influenced those values. Those drivers include but are not limited to the village income, crop ecosystems, other plantation farming activities and possible environmental stresses. In this section, we present results from the multivariate analysis that was applied to identify positive outlier villages having an observed Max EVI value that is significantly higher than the predicted Max EVI, which was modelled based on possible drivers of performance. Variable selection and dimensionality reduction is a crucial step in multivariate analysis, especially when you have a large number of possible explanatory/predictive variables (our data sample had 75 variables). Additionally, if the independent variables are highly correlated, they increase the variance in the regression estimates, and this requires special methods of regression that could overcome this problem of multicollinearity (Kleinbaum et al. 1988). Among those methods, is the PCA and Partial Least Square (PLS). In principal component regression, the PCs are used to predict the dependent variable, which in our case is the average Max EVI.

One drawback of doing regressions using PCA is that the selection of the principal components doesn't give much importance to how each independent variable may be related to the dependent variable, as it is an unsupervised dimensionality reduction technique. Since we are trying to capture possible drivers of EVI, it is crucial to reduce the dimensionality of the data by identifying PCs that not only summarize the independent/predictor variables, but that are also related to the dependent/outcome variable. PLS allows us to achieve this balance by using a dimensionality reduction technique that is supervised by the outcome variable (Maitra & Yan 2008). In comparison to PCA, PLS regression achieved lower RMSE, higher R-square scores and higher percentage of explained variance in the outcome variable. The r function *train* in the caret R Package was used to compute the PLS regression by invoking the pls R Package. The numeric variables were scaled using the *scale* function in the base Package to make them comparable with the categorical variables which were

transformed into dummy variables in the PLS regression. Cross validation was used to identify the optimal number of PCs to be incorporated in the model. The optimal number of components is the one that achieves the lowest cross validation error (RMSE). For each of the 15 HEs, PLS regression was applied, and the optimum number of PCs were used to model the predictor variable. The PC residuals (i.e. the difference between the observed value and the fitted value of the outcome variable predicted by the PLS principle components) were used for outlier analysis using the *OutlierDetection* function in the OutlierDetection Package. In a similar way as the univariate analysis, two types of PDs were identified in the multivariate analysis. The first type used the default methods and the second type used a combination of all methods (i.e. density, depth, dispersion and distance) for outlier identification. While the univariate and multivariate outlier analyses detected 144 and 539 Type1 outlier villages respectively, only 32 Type1 outlier villages were common to both sets of analyses. Similarly, while the univariate and multivariate outlier analyses detected 29 and 48 Type2 outlier villages respectively, only 6 Type2 outlier villages were common to both sets of analyses.

Table 4 summarizes the number of components used in modelling the outcome variable in the PLS regression which was applied for each HE separately. It also presents the cumulative proportion of variance explained, the RMSE and r square scores, the positive outliers identified by multivariate analysis and the common outliers which were also identified using the univariate analysis for each of the two types of outlier detection. In total, out of the 15 HEs, there were nine HEs that had common Type1 outliers and three HEs that had common Type2 outliers. Common here refers to outliers identified by both univariate and multivariate approaches. These common outliers, specifically for Type1 and Type2, are now referred to as True Outliers. It is also interesting to see that for few HEs, all outliers identified using the univariate analysis remained as outliers in the multivariate analysis too. For example, in homologue 23, there was a univariate Type1 outlier village that is also a multivariate Type1 outlier village.

HE ID	Max EVI % of explained variance	RMSE	R square	Type1 Outliers			Type 2 Outliers		
				Uni	Multi	Common (True Outliers)	Uni	Multi	Common (True Outliers)
11	27.3%	0.82	0.07	0	1	0	0	1	0
12	41.4%	0.81	0.35	4	12	0	0	1	0
13	39.3%	0.80	0.32	27	136	5	2	5	0
14	33.5%	0.91	0.25	22	79	4	4	8	2
15	44.7%	0.73	0.45	4	17	2	2	2	1
21	79.9%	0.69	0.61	0	5	0	0	1	0
22	74.9%	0.62	0.63	0	14	0	0	3	0
23	49.1%	0.69	0.54	1	20	1	1	3	0
24	31.6%	0.83	0.28	20	69	0	1	3	0
25	34.4.%	0.90	0.22	15	23	1	5	1	0
31	41.8%	0.80	0.23	0	8	0	0	1	0
32	43.7%	0.72	0.44	2	29	1	2	5	0
33	30.2%	0.85	0.23	30	80	13	7	8	3
34	35.1%	0.80	0.29	11	35	3	5	4	0
35	47.7%	0.97	0.12	7	11	2	0	2	0
Total number of PDs				**144**	**539**	**32**	**29**	**48**	**6**

Table 4: PLS regression Statistics

The **"Max EVI % of variance explained"** column in Table 4, suggests that despite controlling for various factors, there are several controllable and uncontrollable factors, which are not captured by administrative data, that contribute to the variance of observed performance between villages within a HE. The *varImp* function in the caret R package was used to identify the most important predictor variables (i.e. controllable factors) in the model produced by the *train* function. For PLS regression, the variable importance measure is based on the weighted sum of the absolute regression coefficients. The weights are a function of the reduction of the sums of squares, across the number of PLS components and are computed separately for each outcome. Therefore, the contribution of the coefficients is weighted proportionally based on its ability to reduce the sums of squares. The top 10 important variables in the nine HEs containing common outliers (outliers that were identified by both the multivariate and univariate analysis) were analysed again, to identify variables that were common across HEs and variables that were specific to them. In total 35 variables collectively appeared in the top 10 list in each of the nine homologues. Table 5 provides a summary of those variables and how they are ordered across the different homologues.

Important Variables		Homologue								
Name	Code	13	14	15	23	25	32	33	34	35
Doing plantation farming	r2042	1	1	-	2	-	3	1	1	7
Age of main farmer	r216	2	2	-	3	-	4	9	6	5
% of households growing dryland rice in Season3	r301bk4	6	-	6	-	9	1	4	-	9
% of rain fed irrigation	r901a3k2	-	4	3	-	10	6	8	3	-
% of households growing wetland rice in Season3	r301ak4	7	-	10	-	8	5	5	-	8
Flood Events in 2013	R601B_K7	10	-	-	4	1	7	-	-	10
Number of families without electricity	R501B	8	7	-	-	-	-	3	5	3
% of households growing dryland rice in Season2	r301ak2	-	-	5	9	-	8	-	-	1
Number of markets without buildings	R1205	-	-	-	1	-	2	-	2	4
% of households growing wetland rice in Season2	r301ak3	5	8	4	-	-	9	-	-	-
Cooking Fuel used is "LPG"	R5032	-	9	-	-	-	-	-	-	2
Number of active saving and loan cooperatives	R1212C	9	-	-	-	-	-	2	4	-
Main cooking fuel used is firewood	R5034	-	6	-	-	-	-	-	-	-
Village Revenue	R1501A_K3	-	-	1	-	4	-	-	-	-
Main source of income for the majority of the population is plantation	R404B14	3	-	-	-	-	-	6	8	-
Number of female migrant workers	R403B2	-	-	-	-	3	-	-	-	6
Number of landslides	R601A_K7	-	-	-	-	-	-	10	-	-

% of technical irrigation	r901a1k2	-	5	-	-	-	-	-	-	-
Main type of household business is plantation	r214204	4	-	-	-	-	-	7	-	-
Proportion of simple irrigation	r901a2k2	-	-	2	-	-	-	-	-	-
Doing Horticulture Activities	r2032	-	10	-	-	-	-	-	-	-
Majority of wetland rice is managed with revenue sharing	r301ak82	-	-	-	-	5	10	-	-	-
Drainage through river/irrigation channel/lake/sea	R5064	-	-	-	10	-	-	-	-	-
Water source for bathing is well	R507B4	-	-	8	6	-	-	-	-	-
Burning of fields before farming	R5132	-	-	-	-	-	-	-	7	-
Number of male migrant workers	R403B1	-	-	9	-	2	-	-	-	-
Doing Aquaculture Activities	r2082	-	3	-	-	-	-	-	-	-
Water source for bathing is drilling well or pump	R507B3	-	-	-	8	-	-	-	-	-
Drainage through sewage system	R5062	-	-	-	-	7	-	-	-	-
Road surface type from production centre to the main village road is land	R404B23	-	-	-	7	-	-	-	-	-
Village area that borders by the sea	R307A2	-	-	-	-	-	-	-	9	-
Pollution Incidents	R512A_K22	-	-	-	-	6	-	-	-	-
The existence of settlements	R511A2	-	-	-	5	-	-	-	-	-
Water source of bathing is river/lake or pond	R507B6	-	-	7	-	-	-	-	-	-
Utilization of the sea for public transportation	R307B1E2	-	-	-	-	-	-	-	10	-

Table 5: PLS Important Variables

As shown in table 5, doing plantation farming along with rice was identified as a key predictor of average Max EVI values. Villages with the majority of households doing plantation farming were associated with better average Max EVI values. It ranked first in four out of the nine analysed HEs, and on average, it is the most highly ranked variable. The age of the main farmer came second, it appeared as one of the top 10 variables in 7 out of the 9 HEs. As the average age of the main farmer in a village increases, the average Max EVI value decreases. Other top predictors included the proportion of rice households with rain fed irrigation (as it increases, the average Max EVI values increases), the number of flood events (as it increases, the average Max EVI values decreases) and the existence of families without electricity (as it increases, the average Max EVI values increases). There were also predictors that were specific to certain homologues, like aquaculture household activities and horticulture farming in HE “14”, the existence of settlements in HE “23”, burning of the field in preparation of the agricultural land in HE “34” and pollution incidents in HE “24”.

Outlier Validation

The previous steps, i.e. construction of HEs and identification of true outliers, rely on several assumptions, and are performed at a higher aggregation level. For instance, the EO data sources used in the previous steps, provide data at a very high spatial resolution (e.g .precipitation data from CHIRPS is provided at 5 square kilometer resolution), or at a higher temporal resolution (e.g. EVI values are generated once every 16 days). In order to merge these EO data sources with administrative data (such as the agricultural census), that come at a different spatial and temporal resolution, the EO data is aggregated to the smallest administrative unit, i.e. the village level, at which the agricultural census is often collected. Infact, since the agriculture census is collected at the household level, we also aggregate every variable to the village level. Additionally, since administrative data is collected at large scales, systematic errors especially during data collection and (pre) processing, can occur. Therefore, it is necessary to validate whether the identified true outliers are potential PDs. Validation of true outliers needs to happen in terms of errors resulting due to: 1) aggregation of the various data sources used; 2) systematic errors during data collection/processing of the administrative data and 3) False positives resulting from the use of varied statistical approaches such as clustering (to identify HEs), outlier detection and PLS regression. To specifically address potential errors arising from the above mentioned sources, three outlier validation approaches were conducted:

(a) Bivariate Analysis: For every HE containing common true outliers, bivariate analysis was conducted to understand the relationship between each of the top 10 variables (resulting from the PLS regression) and the average Max EVI. The bivariate analysis explained in the next section, describes the contribution of each variable, in explaining the observed variance (which also includes the additional variance), on the average Max EVI, and how this differs across HEs. Validation here was done through 1) identifying if there are variables that are common across true outliers that are known (based on the literature) to have a positive impact on agriculture productivity and 2) if growing rice is the main driver of average Max EVI for true outliers, and it is not plantation what is driving their EVIs.

(b) Google Earth Time Scale Tool: A preliminary visual inspection to assess coherence between the results of PLS/bivariate analysis, performed to identify the true outliers, and from generalized assessments, resulting from manual labelling and examination of a subset of factors, identified in the PLS/bivariate analysis, using the time scale tool in Google Earth Pro.

(c) Earth Observation and Time Series Analysis: An independent pixel level, time series based validation approach, entirely relying on EO datasets, was developed to confirm if we are able to detect significant differences in terms of productivity proxies, between outliers and the rest of the sample. To build, and test this method, HE21 and 22, that had the highest EVI variance explanation in the PLS were used.

Bivariate Analysis

HE "13": This HE contains 2532 villages out of which five villages are true (common) outliers. As shown in figure 3, plantation farming was one of the key predictors of average Max EVI and four out of the five PD villages did plantation farming along with rice farming, three of which had plantation farming as the main type of household business and two had plantation farming as the main source of income. The average age of the main farmer was also identified as one of the top predictors of Max EVI. As the average age increases, the Max EVI decreases. The majority of true outlier villages had an average main farmer age around the thirties. The figure also shows that as the proportion of households growing dryland rice in season three increases, the average Max EVI increases, but this is not the case with the proportion of wetland rice, which affects Max average EVI negatively. The true outliers were divided between both groups, two villages had the majority of their households growing wetland, and two had the majority of households growing dryland rice. And the wetland rice true outliers didn't grow rice in season two. Additionally, the existence of active saving and loan cooperatives was inversely proportional to average Max EVI values, as true outliers were villages having zero cooperatives. Figure 3 also shows that as the number of flood events increases, average Max EVI decreases and true outlier villages had 0 to 3 floods in the year 2013.

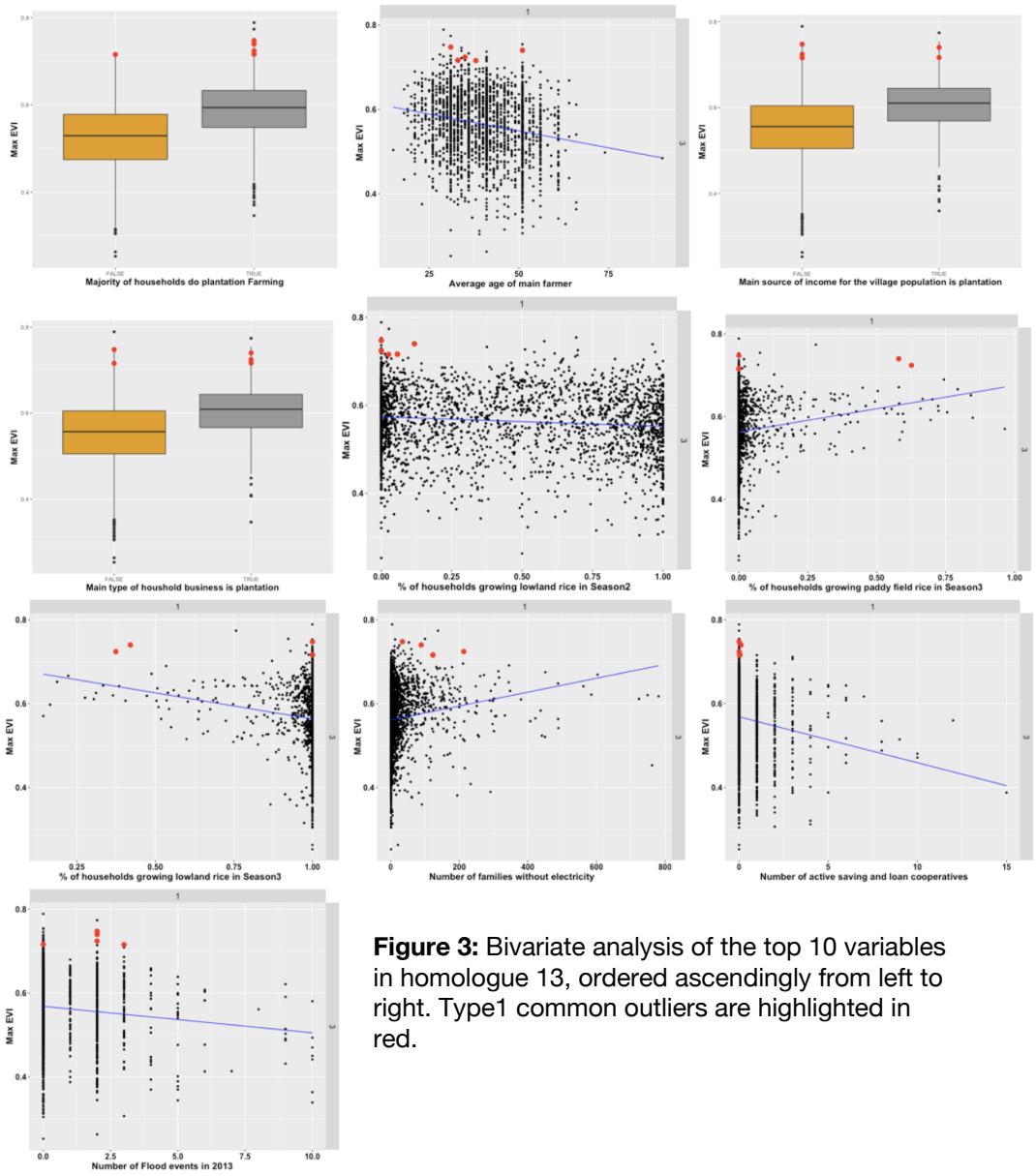

Figure 3: Bivariate analysis of the top 10 variables in homologue 13, ordered ascendingly from left to right. Type1 common outliers are highlighted in red.

HE "14": contains 2943 villages out of which four villages are common PDs. As shown in figure 4, doing plantation farming was also one of the key predictors of Max EVI, but three of the four identified true outliers didn't do plantation farming, despite its association with higher average Max EVI values. The average age of the main farmer is also one of the most important predictors, and true outlier villages had an average main farmer age ranging from 25 to 50 years old. In this homologue, doing aquaculture household activities and horticulture farming were identified among the top predictors of the outcome variable and they were associated with higher average Max EVI values, however none of the true outliers did aquaculture household activities and one outlier did horticulture farming activities. Figure 4 also shows that rain fed and technical irrigation were identified as top predictors of average Max EVI, the higher the proportion of the former the better the EVI values while the latter showed a slight decrease in Max EVI values when there is an increase in technical irrigation. As the number of families without electricity increases (i.e. rural areas) the EVI values increases, however true outliers were located at the lower end with zero to few families without electricity. Similar to homologue "13", as the proportion of households growing wetland rice in season 2 increases, Max EVI values decreases and true outliers were evenly found at the two ends of the spectrum.

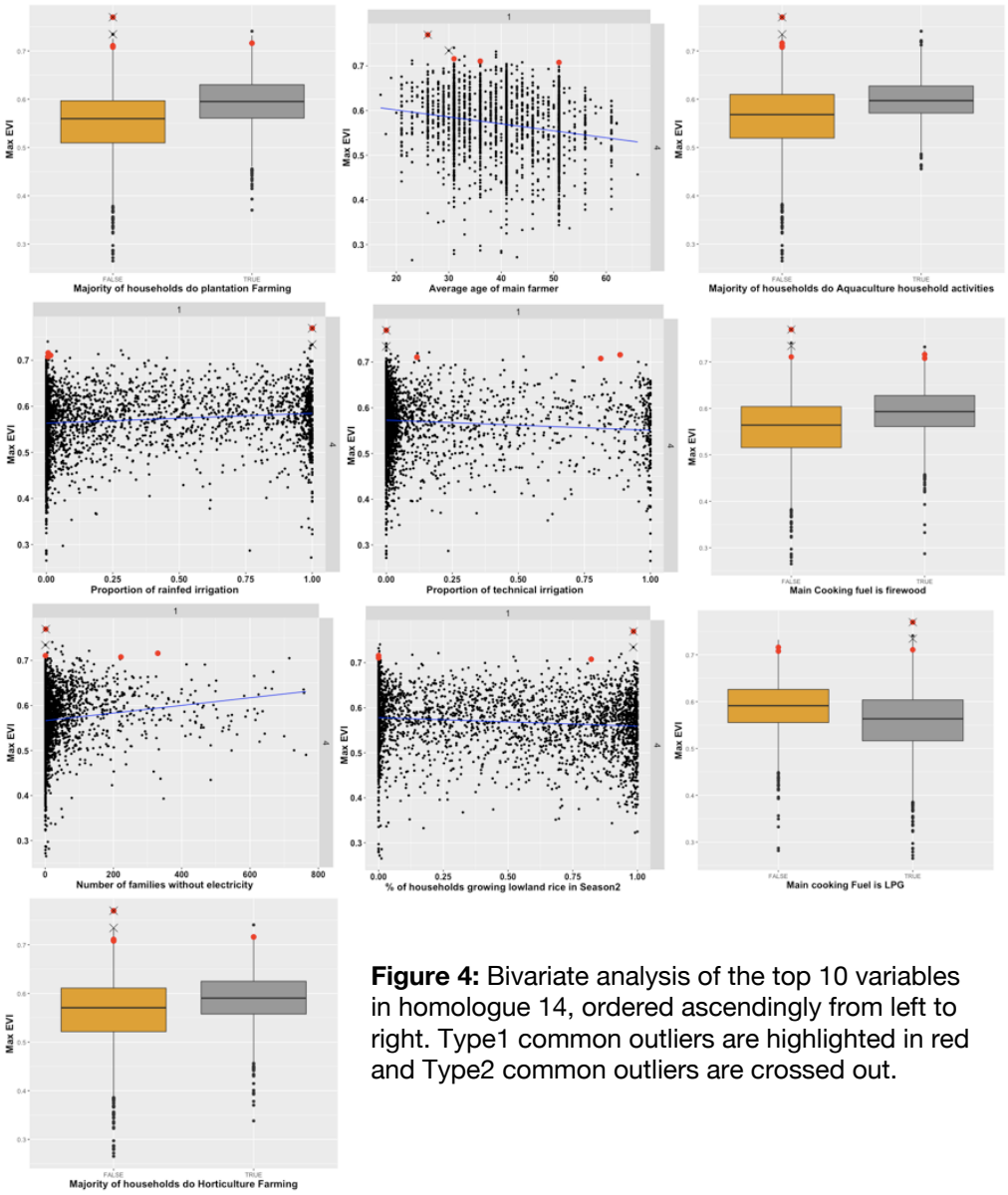

Figure 4: Bivariate analysis of the top 10 variables in homologue 14, ordered ascendingly from left to right. Type1 common outliers are highlighted in red and Type2 common outliers are crossed out.

HE "15": contains 734 villages out of which 2 villages are common true outliers. As shown in figure 4, the most important predictor is village revenue, as it increases Max EVI values increases. One outlier was at the lower end (revenue for this village wasn't provided) and the other outlier was at the higher end. Figure 4 also shows that rain fed and simple irrigation were identified as top predictors of Max EVI, the higher the proportion of households with them the better the EVI values. One outlier village had all households with rain fed irrigation and the other outlier had almost half the households with simple irrigation and the other half with rain fed irrigation. None of the outliers grew dryland rice in season three although it's directly proportional with Max EVI and despite the fact that both PDs had respectively around 0.25% and 0.95% of their households growing dryland rice in season two. However, both PDs had households growing wetland rice as well with similar percentages in season two. The main water sources for bathing in the village was also identified as an important predictor. Rivers, lakes or ponds were associated with lower Max EVIs and using wells for bathing were associated with higher Max EVIs. However, true outliers were divided evenly across both sources. Figure 4 also shows that as the number of male migrant workers increases, EVI values increase. However, both outliers had a very small number of migrant workers. Almost all households in outlier villages grew wetland rice in season three despite the association with lower Max EVI values.

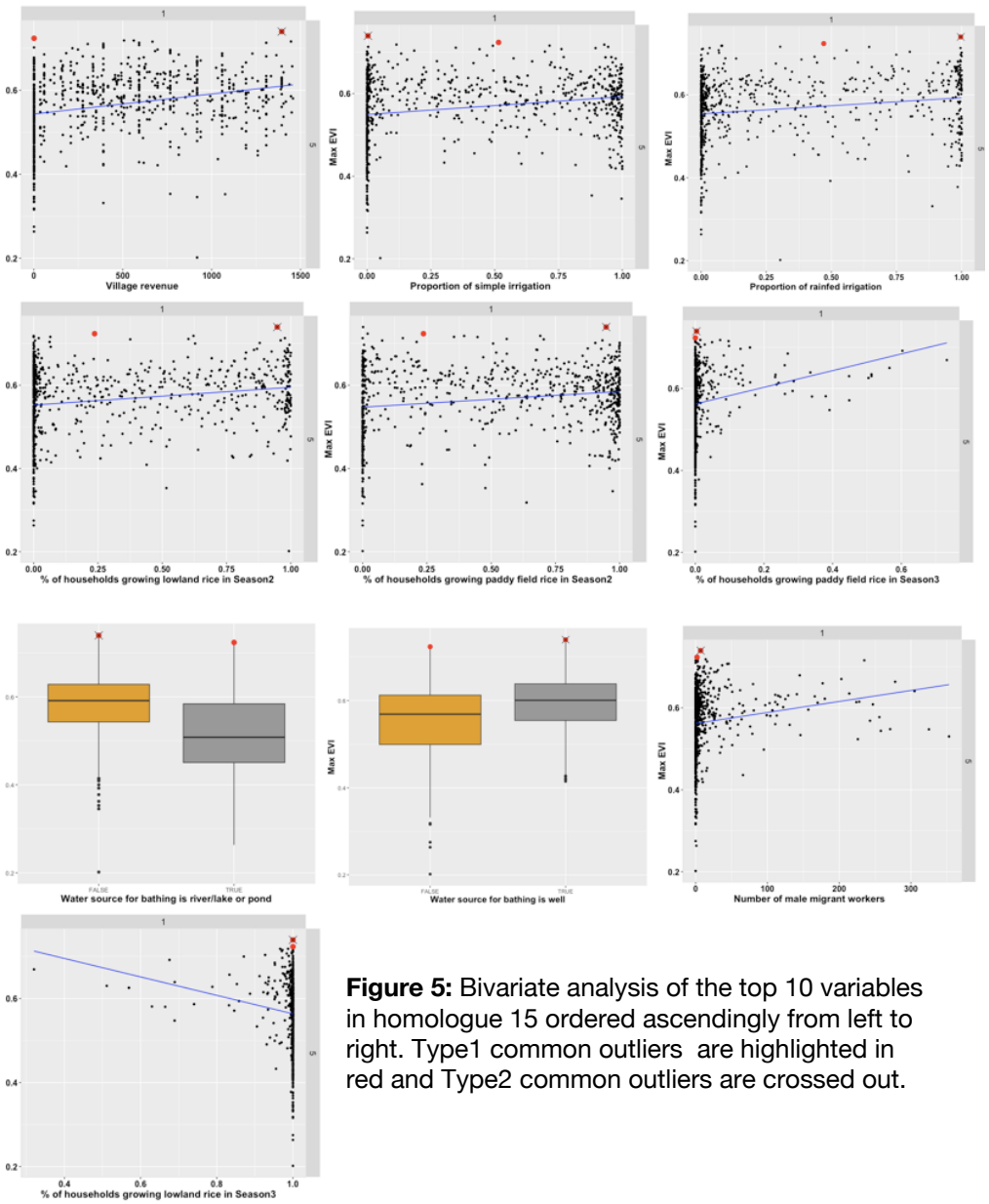

Figure 5: Bivariate analysis of the top 10 variables in homologue 15 ordered ascendingly from left to right. Type1 common outliers are highlighted in red and Type2 common outliers are crossed out.

HE "23": contains 1361 villages out of which only one village is a true outlier. As shown in figure 6, the most important predictor of Max EVI is the number of markets without buildings. As the number of markets without buildings increases, the Max EVI values decreases and the outlier village had no such markets. Doing plantation farming appeared again as an important predictor and the outlier village did plantation farming. Average age of the main farmer appeared again as an important predictor and the outlier village average age of the main farmer was in the forties. Figure 6 also shows that the number of flood events have a negative impact on the outcome variable and the outlier village had no flood events in the year 2013. The results also show the existence of settlements and the use of water pumps for bathing is inversely proportional to Max EVIs (the outlier village had non) and using wells as the primary source of water for bathing is directly proportional to Max EVIs (the outlier village used it). The existence of a land road surface from the production centre to the main village road was associated with better EVI values and the outlier village had a land road surface. Drainage through irrigation channels, lakes or in the seas was associated with higher Max EVI values and the outlier village didn't have such drainage systems. Finally, the outlier village had almost zero households growing dryland rice in season two.

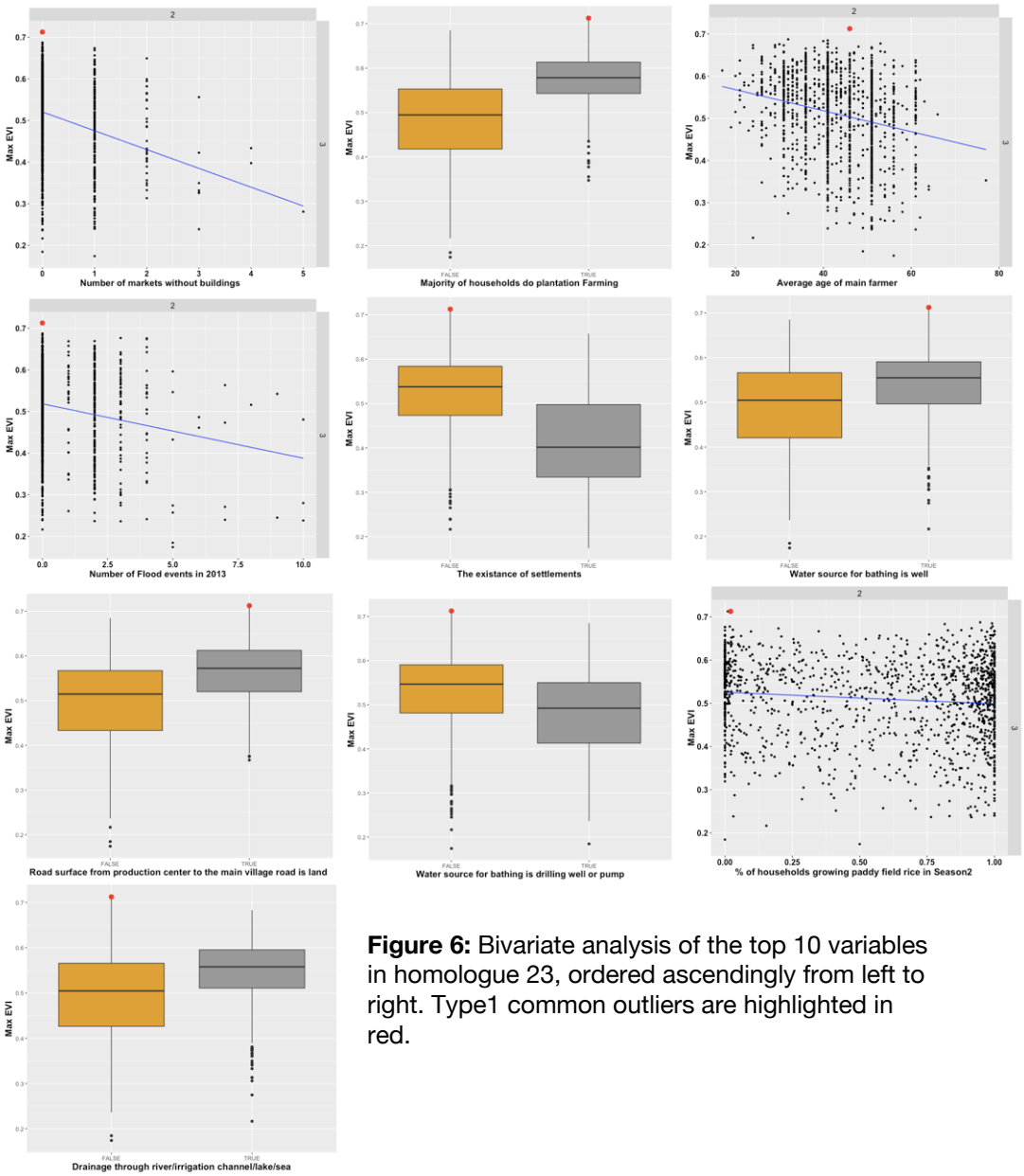

Figure 6: Bivariate analysis of the top 10 variables in homologue 23, ordered ascendingly from left to right. Type1 common outliers are highlighted in red.

HE "25": contains 950 villages out of which only one village is a true outlier. As shown in figure 7, the number of flood events was the most important predictor and the outlier village had no flood events in the year 2013. The number of male migrant workers came second with the outlier village having none, however, in contrast to homologue "15", as the number of male migrant workers increases, Max EVI increases. The same was true for female migrant workers and the outlier village had few of them. Village revenue came fourth, with the outlier village having a relatively low revenue. Figure 7 also shows that revenue sharing in managing wetland rice had an inversely proportional relationship with Max EVIs and the majority of rice households in the outlier village didn't use this type of land management. Pollution incidents also appeared as one of the important predictors of Max EVI and the outlier village didn't experience any. Drainage through sewage systems showed slightly lower Max EVI values than drainage through other systems. In season three, we can also see the majority of villages had a large proportion of households growing wetland rice and a very small proportion of households growing dryland rice. The outlier village had almost 90% of the households growing wetland rice and less than 20% of the households growing dryland rice in season three.

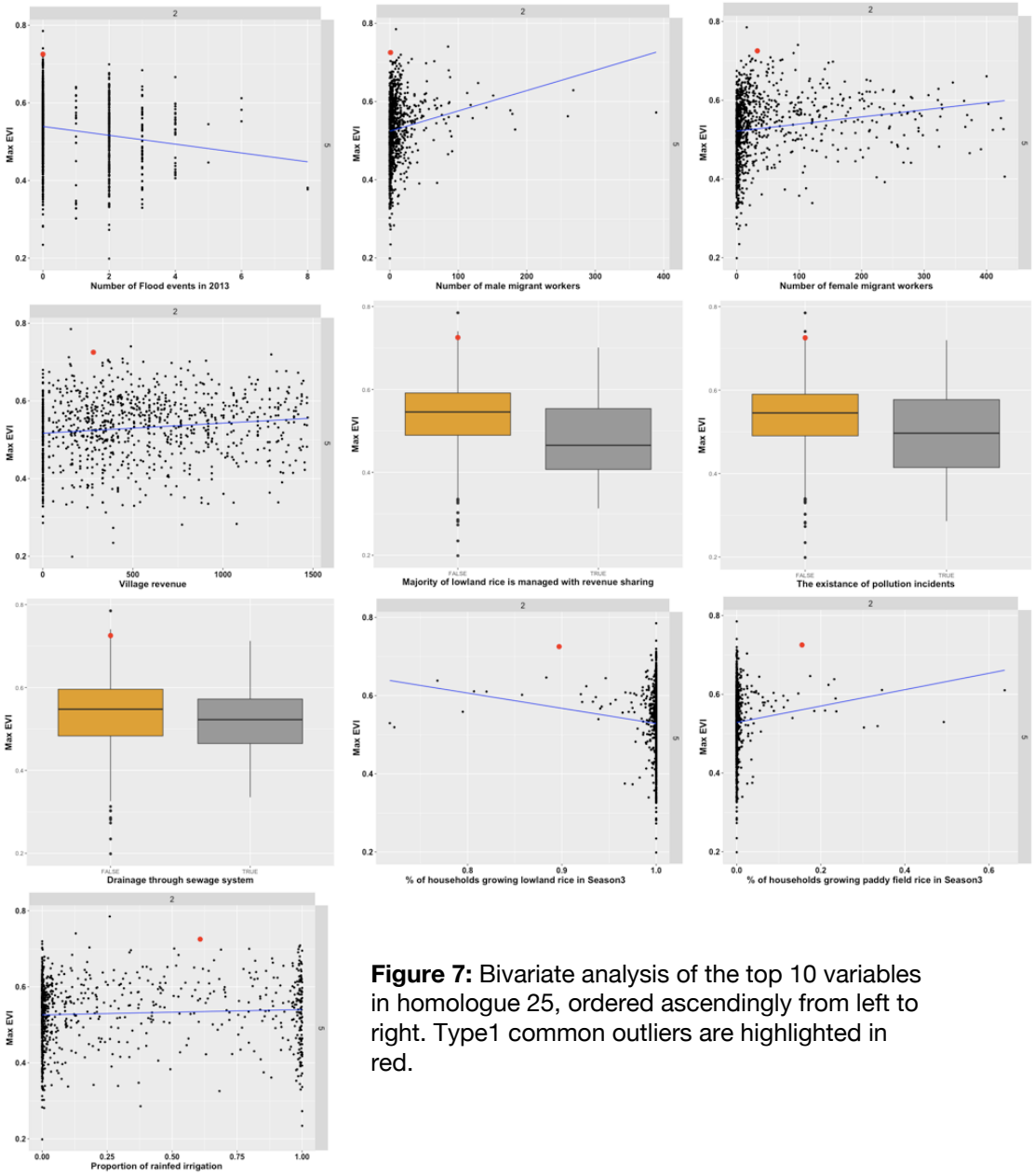

Figure 7: Bivariate analysis of the top 10 variables in homologue 25, ordered ascendingly from left to right. Type1 common outliers are highlighted in red.

HE "32": contains 778 villages out of which only one village is a common true outlier. As shown in figure 8, the proportion of households growing dryland rice in season three is the most important predictor. However, the outlier village didn't have households growing dryland rice in season three despite having almost 70% of the households growing dryland rice in season two. On the other hand, almost all rice growing households in the outlier village grew wetland rice in season 3 and 90% of the rice households grew wetland rice in season two. The average age of the main farmer appeared again as an important predictor, with the outlier village having an average age around 45. Opposite to previous homologues, the number of flood events here was directly proportional to Max EVI values. Figure 8 also shows that rain fed irrigation is one of the most important predictors of Max EVI, however none of the rice growing households in the outlier village used rain fed irrigation. Similar to the previous homologue, revenue sharing in managing wetland rice had an inversely proportional relationship with Max EVIs and the majority of rice households in the outlier village didn't use this type of land management.

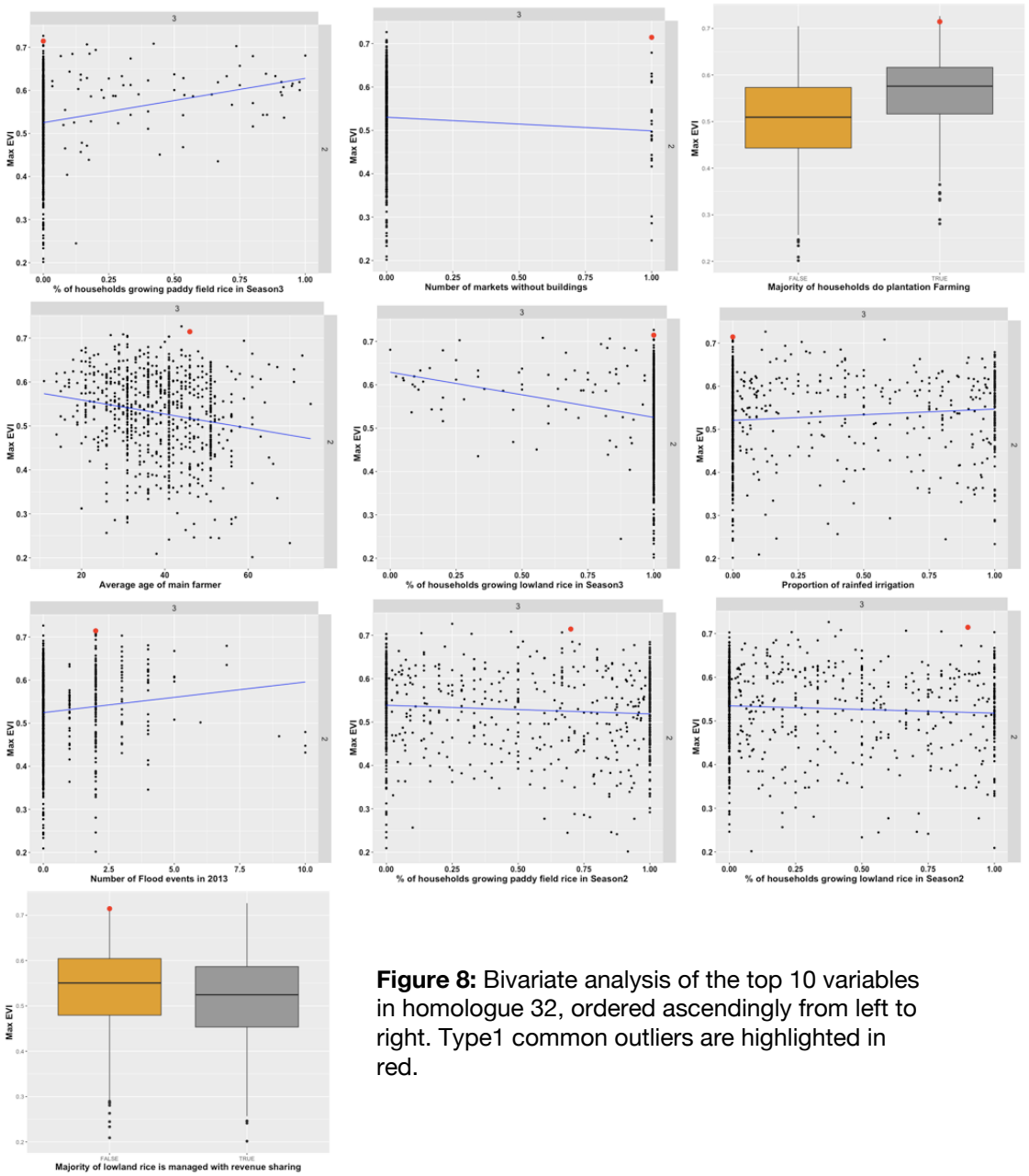

Figure 8: Bivariate analysis of the top 10 variables in homologue 32, ordered ascendingly from left to right. Type1 common outliers are highlighted in red.

HE "33": contains 2265 villages out of which 13 villages are true outliers. As shown in figure 9, practicing plantation farming is again the most important predictor of Max EVI, however two outlier villages did not undertake plantation farming. Active saving and loan cooperatives appeared again with the majority of outliers having one or zero cooperatives. The number of families without electricity appeared also as one of the most important predictors. However, the majority of outliers had a very low number of families without electricity. In season three, the majority of villages - including outliers - had a very high proportion of rice households growing wetland rice and a very low proportion of rice households growing dryland rice. Having plantation farming as the main source of income and the main type of household business for the majority of households in the village, was identified among the most important predictors of Max EVI. Figure 9 also shows that outliers had varying proportions of households depending on rain fed irrigation in growing rice. The average age of the main farmer appeared again as an important predictor with outlier villages ranging from 30 to 50 years old. Finally, the number of landslides seemed to be positively correlated with Max EVI values and outliers had either one or zero landslides in the year 2013.

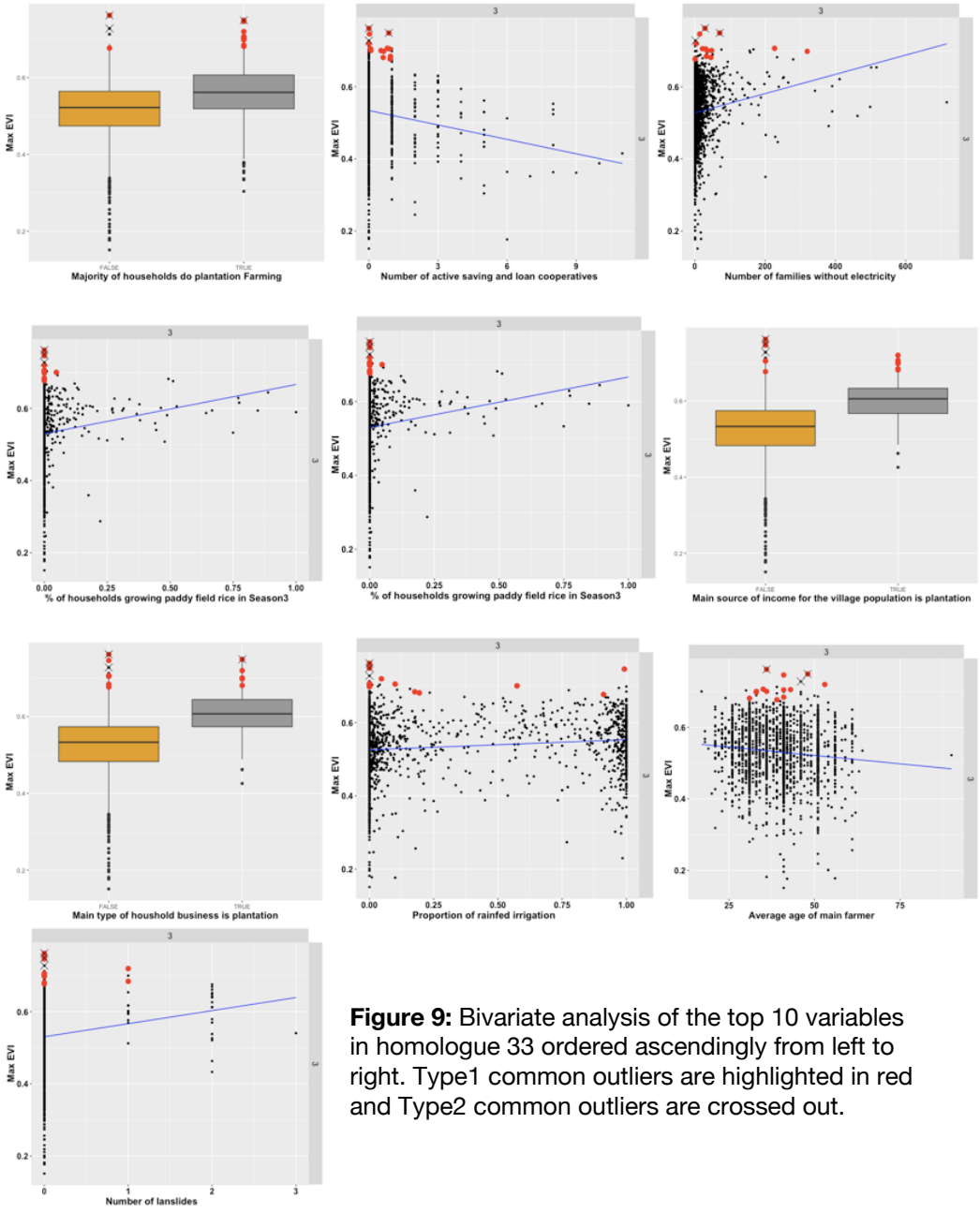

Figure 9: Bivariate analysis of the top 10 variables in homologue 33 ordered ascendingly from left to right. Type1 common outliers are highlighted in red and Type2 common outliers are crossed out.

HE "34": contains 1078 villages out of which three villages are common outliers. As shown in figure 10, doing plantation farming is also the most important predictor of Max EVI, however, one of the three outlier villages did not do plantation farming. Markets without buildings and active saving and loan cooperatives appeared again as an important predictor, however, all three outlier villages didn't have any of those markets and cooperatives. The proportion of rice households with rain fed irrigation in outlier villages is very low. Similarly, there were very few families without electricity in outlier villages. Average age of the main farmer in outlier villages ranged from 25 to 50 years old. In figure 10 burning fields to prepare the agricultural land appeared as an important predictor for the first time in this homologue and it is associated with better Max EVI values, two out of the three outlier villages did this practice. Other predictors that also appeared in this homologue and they might be related to its geographical location is having an area in the village that borders with the sea and the utilization of the sea for public transportation, the former was associated with lower Max EVI values and the latter was associated with higher Max EVI values.

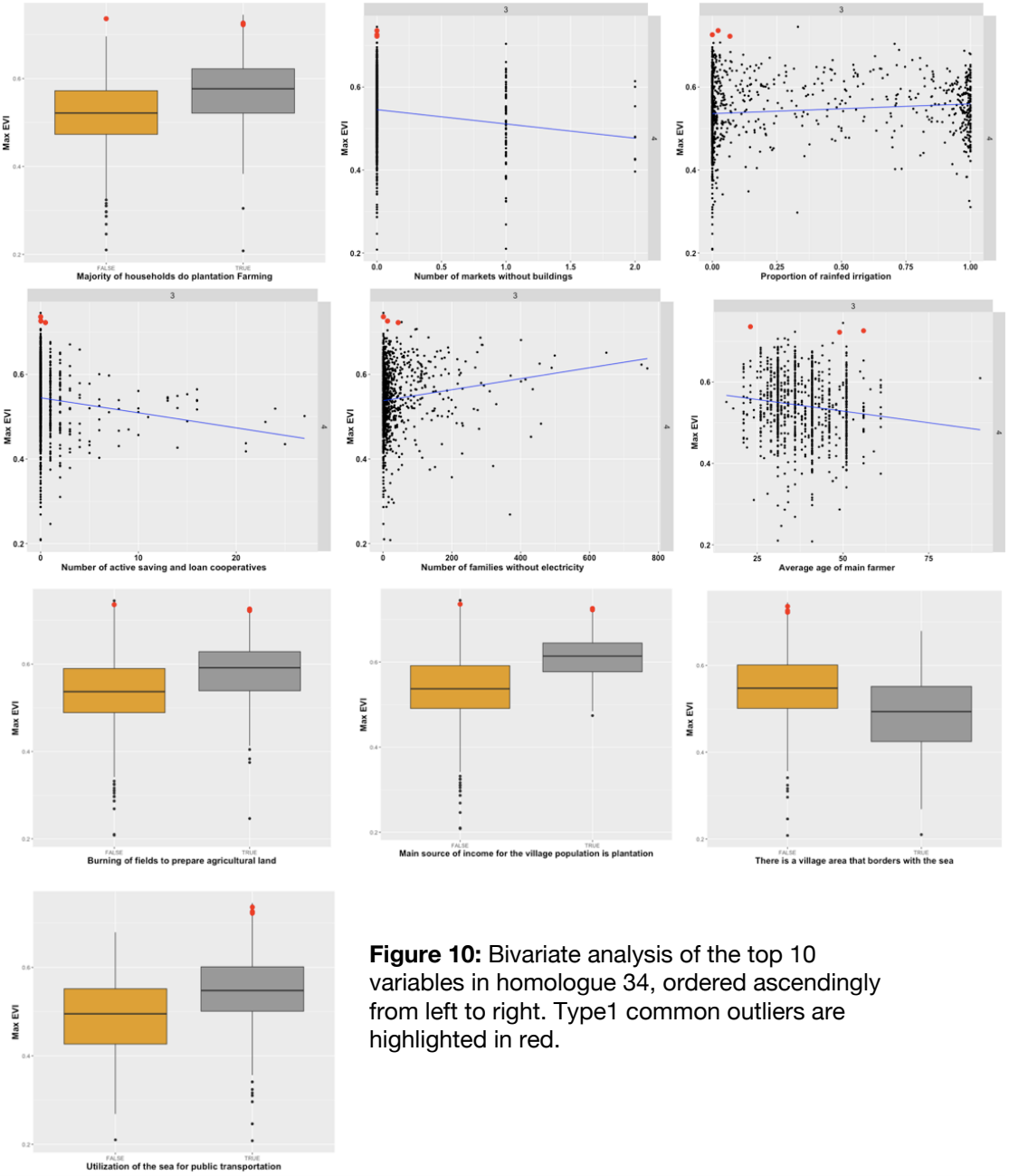

Figure 10: Bivariate analysis of the top 10 variables in homologue 34, ordered ascendingly from left to right. Type1 common outliers are highlighted in red.

HE "35": contains 520 villages, of which two villages are true outliers. As shown in figure 11, the proportion of rice households growing dryland rice in season two is the most important predictor of Max EVI with both outliers having 50% and 90% respectively. The use of LPG as the main cooking fuel also appeared as one of the top predictors and it was true for one of the outlier villages. Markets without buildings and families without electricity appeared again as important predictors, however, the two outlier villages had none. The average age of the main farmer for both the outlier villages ranged from 30 to 40. Figure 11 also shows that the number of female migrant workers was negatively correlated with Max EVIs, however the outlier villages had a very low number of those migrants. Doing plantation farming appeared again as a top predictor, however, the majority of rice households in outlier villages did not do plantation farming. In season three, the majority of villages - including outliers - had a very high proportion of rice households growing wetland rice and a very low proportion of rice households growing dryland rice. The number of flood events also appeared as an important predictor that is negatively correlated with Max EVI and outlier villages didn't have any flood events.

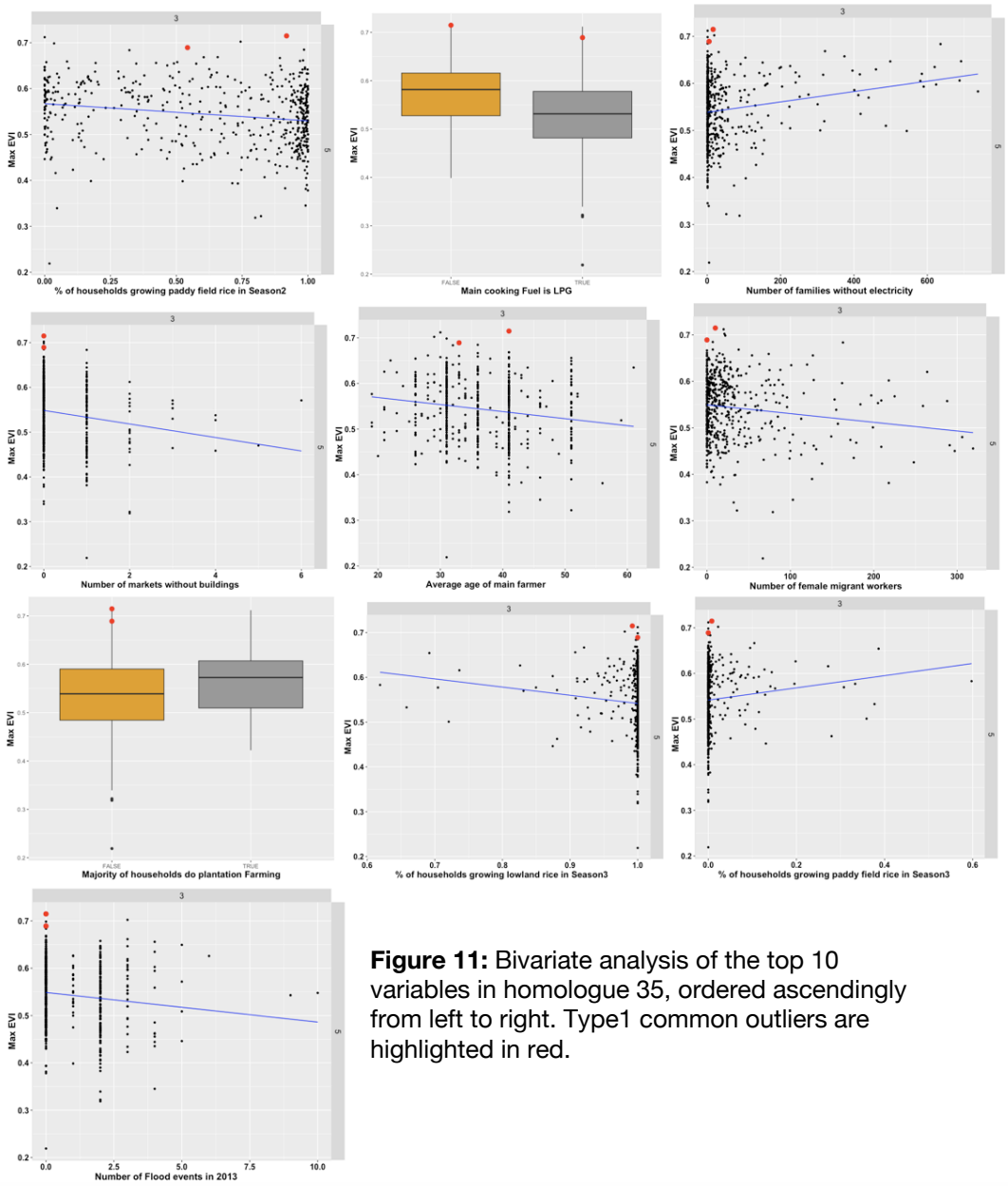

Figure 11: Bivariate analysis of the top 10 variables in homologue 35, ordered ascendingly from left to right. Type1 common outliers are highlighted in red.

Discussion

- In some of the HEs, it appears the absence of rice cultivation in the second season is associated with higher average Max EVI values. This is supported by evidence from a study by Lantican et al. (1999) which suggests that farmers who grew non-rice crops before the rice season had better weed control, hence, improved rice productivity.
- We found predictors that were specific to certain homologues: for e.g. aquaculture household activities and horticulture farming in HE "14"; the existence of settlements in HE "23"; burning of the field in preparation of the agricultural land in HE "34"; and pollution incidents in HE "24". These results, further add credence to the fact that production systems are complex, and highly varied, and the use of standardized administrative and EO data, with a PD approach, can help identify location specific constraints, and opportunities to improve agricultural performance.
- The results suggest that absent the use of rice type (i.e. dryland and wetland rice) as a control variable, there were no issues observed with identifying true PD rice producing areas across both types. Because, as shown in HE "13" of the bivariate analysis, out of the four outlier villages identified, two had the majority of households growing dryland rice and two had the majority of households growing wetland rice and still both were identified as outlier villages in the same homologue.
- The age of the farmer seemed to be negatively correlated with average Max EVIs. This supports the findings of Osanyinlusi & Adenegan (2016) who conducted a study that examined the factors affecting rice farmers' productivity of 160 randomly selected farmers in Nigeria. They offered evidence that farming experience was negatively significant to farmers' productivity. True outliers had an average age around the 30s.
- The bivariate analysis also shows that flooding events affect Max EVIs negatively. This is supported by the findings of Osanyinlusi & Adenegan (2016) which identified flooding as one of the constraints limiting rice production. Our true outliers had a small number of flooding events ranging from 0 to 3 floods in the year 2013.
- Type1 PDs didn't necessarily have the same values across the most important variables, on the other hand, type 2 PDs had clearly more conformity i.e. whenever they existed in a homologue they had similar values across the various predictors of performance. This indicates that type 2 outlier detection doesn't only result in a smaller number of outliers, but also very similar outlier villages. This is evident in HE "14" & "33" of the bivariate analysis.
- Despite the positive relationship between the number of families without electricity and the average Max EVI values, the true outliers always appeared at the lower end with zero to few families without electricity.
- Other interesting findings from the bivariate analysis include the burning of rice fields to prepare the agricultural land and the existence of wells as bathing sources, which are both positively correlated with average Max EVI values. While we can not explain the relationship between agricultural productivity and having wells as bathing sources, existing literature (Mandal et al. 2004) has shown that rice straw burning returns a considerable amount of plant nutrients to the soil in rice-based crop production systems.
- Despite the fact that "doing plantation farming" was listed as the top predictor of average Max EVI in a number of HEs, this doesn't necessarily mean that plantation farming is the main source of income for the village or the main type of business for the majority of households in this village. They could be villages that are growing other forms of plantation along with rice. For example in HE "33" we were able to identify 13 true outliers, from figure 9 we can see that 3 out of the 13 villages don't do plantation farming and 10 do plantation farming. Out of those 10 villages, only 4 had plantations as the main source of income and the main type of household business.

Google Time Scale Tool

We conducted a rapid check to investigate potential PDs and differences from true outliers using Google earth time scale tool, which enables us to view imagery obtained as near as possible to the selected cropping cycle (i.e. January to April 2013). All the true outliers were inspected in addition to 28 negative outliers. The presence and absence of rice, forestry/plantation and urban land cover was marked for each selected village, as well as basic notes describing the land cover. We did not quantify the land cover percentage but observed optical trends that became apparent. Inspecting each village using this method cannot conclusively determine whether a true outlier is a positive deviant. However, it can help determine if further investigation should take place for particular villages, to determine if certain land covers are potentially causing errors and where there are inaccuracies in the rice mask. Few of the trends we were able to identify among true outliers are presented below:

- True outliers often had mixed land use including forestry and agricultural plantation around and within the rice area. The influence of mixed vegetation on the EVI value is currently unclear, however different vegetation covers will have different levels of 'greenness' (Heute et al. 2002; Megue et al. 2019). Figure 12 presents a village, which is a Type1 and Type2 outlier in HE "33". The white boundary is the rice mask and within the boundary there are rice fields. There is a small amount of urban land cover within the rice mask on the left. On the right, there appears to be mixed vegetation cover. This current analysis cannot determine if this village is a true PD but we suggest the village should be investigated further.

Figure 12: Tulleng village in East Nusa Tenggara, HE 33. The imagery shown was obtained on the 9th of September 2012.

- Several Type1 outliers appeared to have monocultures of rice within the village boundary. Figure 13 presents two of such villages. These villages should be of interest for future analysis, as the influence of other land covers is potentially minimal. Weru is both Type1 and Type2 outlier, the rice boundary nearly covers the entire village, with other land covers being absent, within and around the rice mask.

Figure 13: Weru, Banten and Kubangkampil, Pandeglang, Java, HE 15. The imagery was collected on the 10th of November 2014.

- Many Type1 and Type2 outliers had accurate rice masks, with minimal urban and forestry land cover within the dedicated rice mask. Figure 14 presents two of such villages which had an accurate rice mask, with minimal mixed land-use within. There may be vegetation planted on the sides of the rice fields but it isn't detectable in this imagery.

Figure 14: Padang Subur, South Suliwesi, HE 25 and Bonne-Bonne Village, West Suliwesi, HE 33. The imagery was collected on the

- Four true outliers did not appear to have any rice within the rice mask. Figure 15 presents two of those villages who appeared to have no rice growing in the dedicated rice area and they had forest cover instead.

Figure 15: Soga, Soppeng Regency, South Sulawesi, HE 32 and Tada Selatan, Central Sulawesi, HE 33. Imagery was collected on the 4th of August 2010 and

- On the other hand, a large number of negative outliers had low average Max EVI values because within the rice mask there was large amounts of urban land cover, with minimal rice production. Figure 16 shows two such villages.

Figure 16: Cibaduyut Kidul, West Java & Wumialo Gorontalo both in HE 33. Imagery was collected on the 9th of September 2012 and on the 8th of December 2014 respectively.

- Based on the Google time scale tool, out of the 32 identified true outliers, rice was present in 29 villages and was absent in the remaining three villages. The remaining three villages which had plantation/forestation cover that may have contributed largely to the high Max EVI values, for the cropping season (Megue et al. 2019). This does not necessarily mean that those villages do not grow rice, but more accurate rice mapping data and a better proxy for productivity are needed before any conclusions can be drawn about rice productivity. The remaining 29 villages either had rice only or rice mixed with forestry, plantation or natural vegetation.

Earth Observation and Time Series Analysis

Hypothesis:

For this approach, we built and tested the following hypothesis:

"For a village to be an outlier, it is necessary for its agriculture production systems to be independent of climatic patterns, which means, despite fluctuations in climatic patterns, productivity (measured as the average max Enhanced Vegetation Index (EVI), wherein peak EVI value for each pixel is averaged across all pixels belonging to a village, for the January to April 2013 season) of outlier village should remain consistent. Alternatively, we assume that agricultural productivity is tightly linked to climatic patterns among non-outlier villages, and their productivity is significantly associated with seasonality. The basis for this assumption is that, outlier villages have adopted approaches and practices, and have established production systems, that delink climatic patterns with productivity, whereas the non-outlier approaches have not, at least for the target season, i.e. January to April 2013."

Previous studies have shown that Enhanced Vegetation Index (EVI), works well as a proxy for agriculture productivity, especially for annual crop systems in the tropics. It is also a well-known fact that agriculture productivity is linked to climatic factors such as temperature and precipitation. Therefore, based on the stated hypothesis, and the relationship between EVI and agricultural productivity, and agricultural productivity and biophysical factors, the proposed validation step consists of the following logic:

(a) The triangular relationship can be leveraged to construct a per-pixel relationship between EVI, temperature and precipitation, and the same can be modelled across time (from 2001, until 2012) for both outlier and non-outlier villages.

(b) From 2012, for the same set of pixels, the model can be used to predict the EVI, given the temperature and precipitation values, since 2012.

(c) Based on the stated hypothesis, the assumption is that in 2013, for pixels belonging to outlier villages, the observed EVI values are significantly higher than the predicted ones, and this observation is reversed in the case of pixels of non-outlier villages (the rest of the sample in the HE), wherein the predicted EVI values are either equal or below the observed values.

Method

Using the stated hypothesis, and the logic, this additional validation step consisted of the following workflow:

A. Based on the PLS regression results, we identified homologous environments (HE) that had the highest explanatory power for EVI. HE 21 and HE 22 were selected as both the environments explained ~ 75% of the observed variation in EVI. In addition, both the selected HE's revealed outliers based on only the multivariate approach (under both Type1 and Type2 outlier detection methods), and not based on the univariate approach. If differences between the observed and predicted EVI are identified using the proposed time series based approach, those differences can be attributed to the factors identified using the PLS approach.
B. Pixels belonging to outlier villages within these two HEs were grouped into one class (total sample size equalling to ~ 650 pixels of 1 square kilometre each), while pixels belonging to non-outlier villages in the same HEs were grouped into another class (total sample size equalling to ~ 4500 pixels of 1 square kilometre each). **** All pixels are standardized to 1 square kilometre, see explanation in step e.*** The fact that outliers in HE 21 and 22, belonged to only multivariate based analysis (under both Type1 and Type2 outlier detection methods), suggests that outliers, and the rest of the villages in HE 21 and 22 (non-outliers), are two independent, distinct samples, with different production practices and constraints, ~ 75% of which can be attributed to factors identified to be important differentiators between the two samples using the PLS approach. Therefore, comparisons were made within the pixels of outliers, and non-outliers, across time, rather than between outliers and non-outliers across time.

C. Hence, the following steps were conducted separately for both the classes, and comparisons were performed across time within each class.
D. For each pixel, monthly data since January 2001, until December 2016, for temperature, precipitation and EVI was obtained. The temperature data was obtained from MODIS's Land Surface Temperature and Emissivity sensor (MOD11C3), that provides global monthly day time land temperature, at ~ 5 square kilometre resolution. Similarly, for the same time period (i.e. between 2001 January and December 2016), monthly per pixel EVI values were obtained from MODIS terra sensor (MODIS VI), at 1 square kilometre spatial resolution. Monthly average precipitation data for the same set of pixels and for the same time period, were obtained from the CHIRPS (Climate Hazards Group Infrared precipitation with Station Data) database at ~ 5 square kilometre resolution.
E. In order to bring temperature and precipitation data to the same spatial resolution as the EVI data, temperature and precipitation raster stacks (stacked across time) were resampled using the resample algorithm of Raster Package in R statistical environment.
F. Spatial and temporal gaps across the three datasets were assessed using the Amelia Package in R, and a simple moving average approach from ImputeTS Package in R statistical environment was used to fill data gaps across time.
G. For each class (i.e. outliers and non-outliers) separately, monthly data from January 2001, until December 2012, was used for model building and validation, wherein EVI was used as the response variable, while precipitation and temperature were used as predictors. Separately for each class, monthly data for temperature and precipitation since January 2013, was provided to the model, in order to obtain predicted EVI values.
H. For the modelling approach, we relied on the grid search capabilities, and a deep learning model (a feed-forward multilayer perceptron) provided by the in R statistical environment.
I. For both the models for each class, hyper-parameter tuning was conducted using a random-discrete based grid search approach, with rectifier and Maxout, with and without dropout as activation functions. The grid search approach also included three different combinations of hidden nodes, and a range of values for lasso and ridge regularization (to prevent overfitting). Lastly, the grid search was restricted to testing for a total of 20 models, with 5 folds and 100 epochs.
J. The best performing deep learning model was selected using RMSE and MAE validation metrics.
K. The best model for both the classes, was used to perform predictions. For the predictions, the best performing model was provided with monthly temperature and precipitation data from 2013 January onwards, until December 2016.
L. Package CAST in R statistical computing environment was used to build spatiotemporal cross validation folds, to obtain training and validation datasets for model building purposes. Two folds across space and time were derived from the training data separately for the two classes. Fold one was used to build the model, while the other fold was used as the validation dataset.
M. For the modelling approach, temperature, precipitation data was normalized to scale between 0 and 1, using the normalize function. In addition, the EVI data was also rescaled to 0 and 1. Data rescaling was done using the Normalize function in the BBMISC Package in the R statistical environment.
N. Scatter plots for comparing the predicted EVI values and scaled and observed EVI values from the validation dataset were constructed using the ggplot2 plotting Package in R statistical environment.
O. A new variable called difference (diff) was constructed, which quantifies the difference between Observed (and scaled), and predicted EVI values, for each pixel, separately for both outlier and non-outlier villages. Histograms for the difference values were constructed using base R functions, and density plots for the same were constructed using the ggplot2 Package in R statistical environment.
P. For true outlier and non-outlier villages separately, total number of positive and negative difference values were counted. Positive difference values reflect that the observed is higher than the predicted, while the negative difference values correspond to pixels, wherein the predicted values are higher than the observed.
Q. Count data for positive and negative values was subjected to chi-square analysis, in R statistical environment. The null hypothesis in this case referred to an equal number of positive and negative values. Chi square test was done separately for true outlier and non-outlier villages.

R. Since the classification of outlier and non-outlier villages was performed with village as the observational unit, we performed additional analysis, to reflect pixel level differences in observed and predicted EVI values, to the observational unit.
S. For the village level analysis, EVI raster layer, in which pixel values were masked using spatial polygon shapefiles of either outlier or non-outlier villages, wherein each polygon belongs to a village, was used. Each pixel was converted into a polygon using the Raster Package in the R statistical environment. Following this, a cell number was assigned to each polygon.
T. The cell number in the above step, corresponds to the cell number of pixels for which difference values were calculated.
U. Original values for cells in the EVI raster layer were replaced with the corresponding differenced values of the corresponding cells.
V. The difference values in the EVI raster layer were again extracted, this time into the corresponding village, using the spatial polygon shapefiles of either outlier or non-outlier villages.
W. For each village, the total number of positive and negative difference values were counted, and the counts were subtracted (referred to as diff2 in the figures)

Results and Discussion:

The model building and validation employed in this approach, relies on the established knowledge regarding the relationships between (a) biophysical covariates, i.e. Temperature and Precipitation, and EVI, (b) relationship between EVI and agricultural productivity, and (c) biophysical covariates and agricultural productivity. Therefore although two different models, one for outlier village pixels, and the other for non-outlier village pixels were developed, the strategy was to select a model, which best reflects this existing knowledge of relationship. Therefore, the model building process for both the classes involved the same hyperparameter ranges and search strategy. Table 6 presents model parameters for outliers and non-outliers, which best describes the relationship between biophysical covariates and EVI. Interestingly model validation metrics, for both with the validation dataset, and the prediction dataset, showed that the model for pixels belonging to outlier villages had a better fit than the pixels of non-outlier villages (See RMSE and MSE values in Table 6). The number of pixels in the non-outlier villages were significantly much more than those in the true outlier villages. Therefore pixels in non-outlier villages could encompass larger variation in biophysical covariates and EVI, compared to those in the outlier villages, hence rendering it difficult to find a model that best describes the relationship in comparison to pixels in outlier villages, which is reflected in the model validation metrics. The focus of this modelling strategy is to perform predictions on data from the year 2013, and not to describe the relationship between biophysical covariates and EVI, however it is interesting to note that EVI is significantly influenced by precipitation in pixels among non-outlier villages, while EVI is influenced by temperature among pixels in outlier villages.

Best Model Specification (obtained after grid search)	**Non-Outlier villages**	**Outlier villages**
Activation	Rectifier	Max out with dropout
Hidden neurons	20, 15	20, 15
l1 (lasso optimization)	1.0E-5	0.001
l2 (ridge optimization)	1.0E-5	0.001

Mean squared error (MSE) on validation data (Fold 2)	0.012	0.0005
Root mean squared error (RMSE) on training on validation data (Fold 2)	0.111	0.02
Top important variable	Precipitation	Temperature
Mean squared error (MSE) on prediction data	0.015	0.002
Root mean squared error (RMSE) on prediction data	0.126	0.05

Table 6: Selected parameters and validation metrics for the best model obtained after employing the same search strategies, across identical ranges of hyper-parameter values for pixels belonging to both the classes, i.e. outlier and non-outlier villages

Prediction results from the selected models for both classes, also reflect the model validation metrics presented in Table 6, wherein the slope of the relationship between observed and predicted scaled EVI values for pixels belonging to outlier villages, differed significantly to those from the non-outlier villages. Relatively steeper slope, in the case of pixels belonging to outlier villages, suggests that in general, the observed EVI values for the period January to April 2013, are significantly higher than the predicted values, unlike those among pixels of non-outlier villages (Figure 17). This further suggests that production systems belonging to pixels in outlier villages are performing better than expected (i.e. predicted, and are not entirely dependent on climatic conditions).

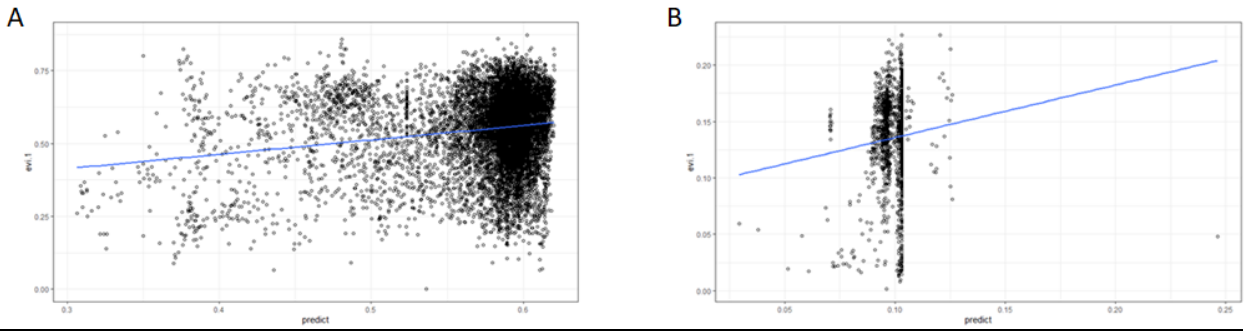

Figure 17: Scatter plot presenting prediction results from the best model, using monthly values of biophysical covariates starting from January to April 2013 (i.e. the season for which data from the Agriculture census was used to identify true and non-outlier villages). X axis represents the predicted EVI values (scaled), while the Y-axis represents the observed EVI values (also scaled) for each pixel, between January and April 2013. The blue line in each scatter plot represents a linear model fit. Subset A represents the predictions for pixels belonging to the non-outlier villages, while B represents the same for pixels belonging to outlier villages.

The observation that pixels in outlier villages perform better than expected, is further evidenced by the fact, that the distribution of the difference values for pixels (i.e. per pixel difference between scaled values of observed and predicted EVI), is skewed to the right (Figure 18 C and D), indicating presence of more number of positive difference values, in comparison to pixels in non-outlier villages, wherein no skew was observed (Figure 18 A and B), indicating relatively equal numbers of positive and negative difference values. More number of positive than

negative difference values, as observed in the case of pixels of outlier villages, suggests that many pixels performed better than expected, during the January to April 2013 season.

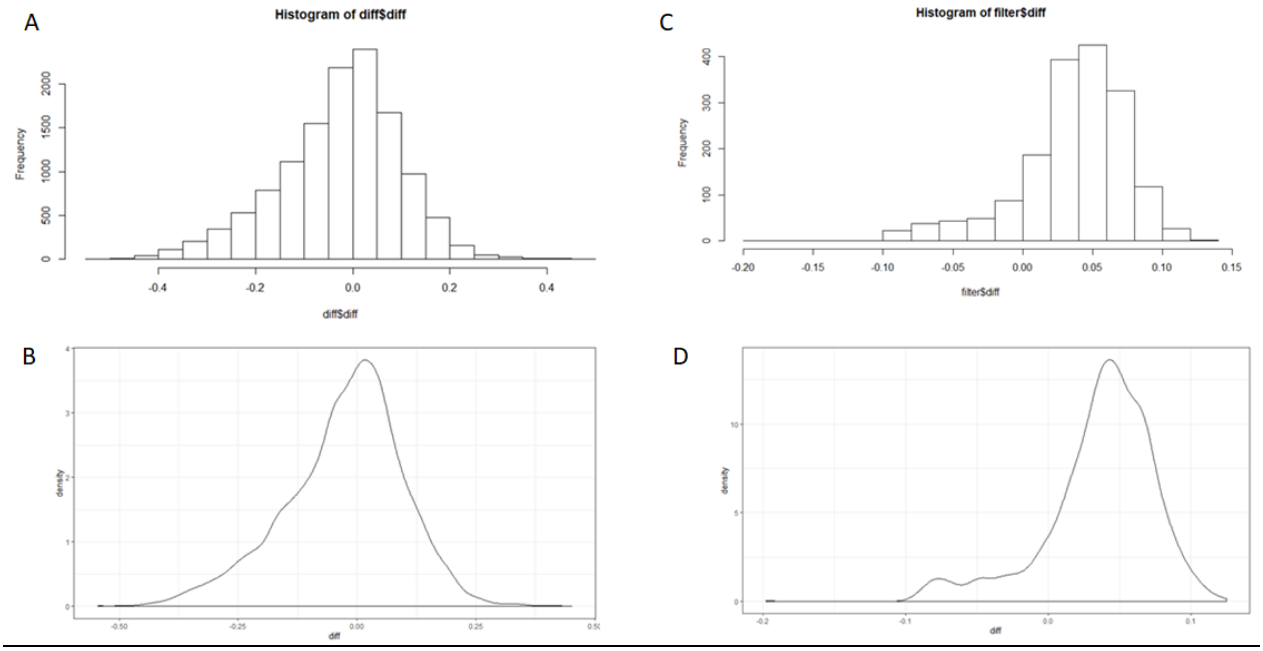

Figure 18: Histograms and density distribution plots representing the distribution of differenced values (difference between observed and predicted scaled EVI values) across all pixels belonging to non-outlier villages in subsets A and B, and outlier villages in subsets C and D.

In order to test if the difference in the number of positive and negative difference values, for either pixels of and non-outlier villages, is not a chance observation, we performed a chi-square test, with the null hypothesis of equal proportion of positive and negative difference values.

Chi square test for pixels in both outlier and non-outlier villages showed that the difference between positive and negative difference values is significant, and the observation is not by chance (Table 7 and Table 8; chi square statistic for non-outlier villages = 48.099, chi square statistic for outlier villages = 512.4256, p-value < .01). However, the number of positive difference values are significantly higher than negative values in outlier pixels, in contrast to that of the pixels in non-outlier pixels, wherein the number of negative difference values are significantly higher than positive values, showing that pixels of outlier villages did indeed perform better than expected during the January 2013 to April 2013 cropping season.

Non-Outlier pixels	**neg**	**pos**	
Null Hypothesis	6324	6324	
Actual	6875	5773	
The chi-square statistic is 48.099			
The p-value is < 0.00001. Significant at p < .01			

Table 7: Chi-square test analysis to assess if the difference between the total number of negative (neg) and positive (pos) difference values among pixels on non-outlier villages are significantly different, and are not observed by mere chance. Actual in the above table refers to the observed number of pixels with positive and negative difference values, while Hypothesis refers to the null hypothesis of equal number of positive and negative difference value pixels.

Outlier pixels	**neg**	**pos**	
Null Hypothesis	860	859	
Actual	240	1477	
The chi-square statistic is 512.4256			
The p-value is < 0.00001. Significant at p < .01			

Table 8: Chi-square test analysis to assess if the difference between the total number of negative (neg) and positive (pos) difference values among pixels of outlier villages are significantly different, and are not observed by mere chance. Actual in the above table refers to the observed number of pixels with positive and negative difference values, while Hypothesis refers to the null hypothesis of equal number of positive and negative difference value pixels.

Since the identification of outliers and non-outliers was performed at the lowest administrative unit (i.e. village) in the agriculture census, the results from the time series validation were rolled up from the pixel level to the village level. Distribution of the difference2 values (i.e. difference between the number of positive and negative difference values) across all villages, belonging to either outlier and non-outlier class revealed that a major proportion of villages in the case of outlier (18 out of 19; Figure 19B) villages obtained positive difference2 values. This is in contrast to non-outlier villages, wherein relatively lower proportion of villages (251 out of 580; Figure 19A) obtained positive difference2 values, further indicating that, aggregation of the pixel results to village level, also shows that indeed villages belonging to true outlier class performed better than expected, during the cropping season between January 2013 to April 2013.

These results also reveal that the per pixel and village level differences observed between outliers and non-outliers, can be attributed to the factors, identified from the PLS regression, responsible for differentiating the two village types.

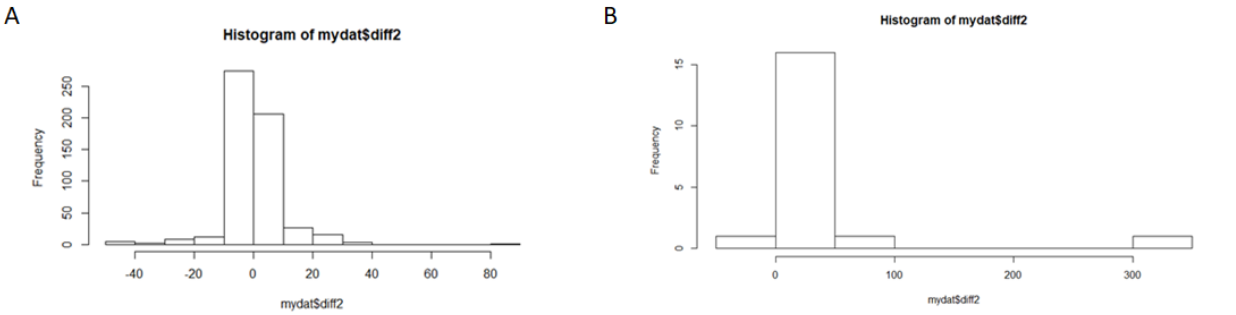

Figure 19: Histogram presenting the distribution of positive and negative difference2 values (i.e. the difference between the number of positive and negative difference values) aggregated at the village level for pixels belonging to (a) Non-Outlier villages (b) True Outlier villages

Histogram analysis of difference2 values showed that Sungai Lumpur village (Geocode: 1602031004), in Sumatra Selatan province, had the highest difference2 value (= 349) among true outlier villages, while Barabai Utara village (Geocode: 6307050006) in Kalimantan Selatan province, was the only village with a negative difference2 value (= -1) among true outlier villages. Interestingly, histogram analysis also revealed several villages with positive difference2 values. Tirto Sari village (Geocode: 1607060026) in Sumatra Selatan, had the highest positive difference2 value (= 88) among non-outlier villages. In contrast Bintaran village (Geocode: 1607091004) in Sumatra Selatan province, had the highest negative difference2 value (= -49), suggesting that, among all the non-outlier villages, this village performed the worst, during the January 2013 to April 2013 cropping season. These results also reveal an interesting observation with respect to province Sumatra Selatan, as both the best and worst performing villages, were identified in the same province.

In addition to validating the identification of true outlier and non-outlier villages, this step can also be used to further narrow down the number of villages for additional ground truthing, to identify true positive deviants. It is important to note that this method needs further development to test which methods (multivariate or univariate - or Type1 and Type2 outlier detection approaches), yield true PDs.

Challenges and Limitations

- **Rice mask errors:** To minimize the influence of other land covers on the extracted EVI values, we used the intersection of the rice mask from the 2014 Indonesian Land Use shapefile and the village boundaries. However, the 2014 Land Use shapefile (provided by the Ministry of Forestry) only indicates the rice area as "Sawah" (Indonesian for "rice field") and does not differentiate between "wetland rice" and "dryland rice." It is unknown what rice variety is grown in the mask and it cannot be directly paired to the census and PODES datasets. Additionally, as discussed in the section earlier on validation using Google Earth, the rice mask boundaries are not accurate, likely because it was created with a visual classification with medium resolution data (Setiawan et al. 2013). When viewing the land cover of villages, rice can be found outside the dedicated rice mask and alternative land covers, such as urban land, industry, and other agricultural systems are also often found within.

- **Dataset integration errors:** We identified a number of potential sources of error when combining the different datasets. The first potential source of error identified, was the false geometry in the village administrative boundary shapefile. To combine EO data with the census data, there is a crucial step of extracting the raster values using the administrative boundary shapefile. It was assumed that correcting the geometries of the shapefiles would reduce the errors and decrease the reduction in the sample size. Whilst correcting geometry improved the results of the extraction, there was still a reduction in the sample size. This could be due to two reasons. First, the extraction method is failing because the spatial resolution of the rasters in comparison to the administrative boundaries are too large. Second, the spatial data have inherent errors coming from the coordinate reference systems, data frames, and extent. Whilst we corrected and pre-

processed the data, there were still errors when combining the data. In the future, different extraction methods such as binary, centroid or interpolation methods should be trialled.

- **EO Data Errors:** When extracting the CHIRPS data into the 78,811 villages there were a total of 1,328 villages with NA values. When displayed spatially, the NAs are mostly present along the coast, on small islands amongst the archipelago and near inland lakes. And when we attempted to extract the temperature raster onto the 78,811 villages, there were 620 villages with NA values. When displayed spatially we found an overlap with the errors previously displayed with the CHIRPS data. There was an additional source of error when we extracted the average EVI Max values for rice growing areas within villages. Across the administrative boundary of a village, there are numerous rice growing areas that return null values and negative values. Of the 36,787 villages, there were 1,194 villages with no value or a negative value from the raster. Figure 17 presents a map showing the spatial location of those errors, with insets displaying randomly selected locations at a higher resolution.

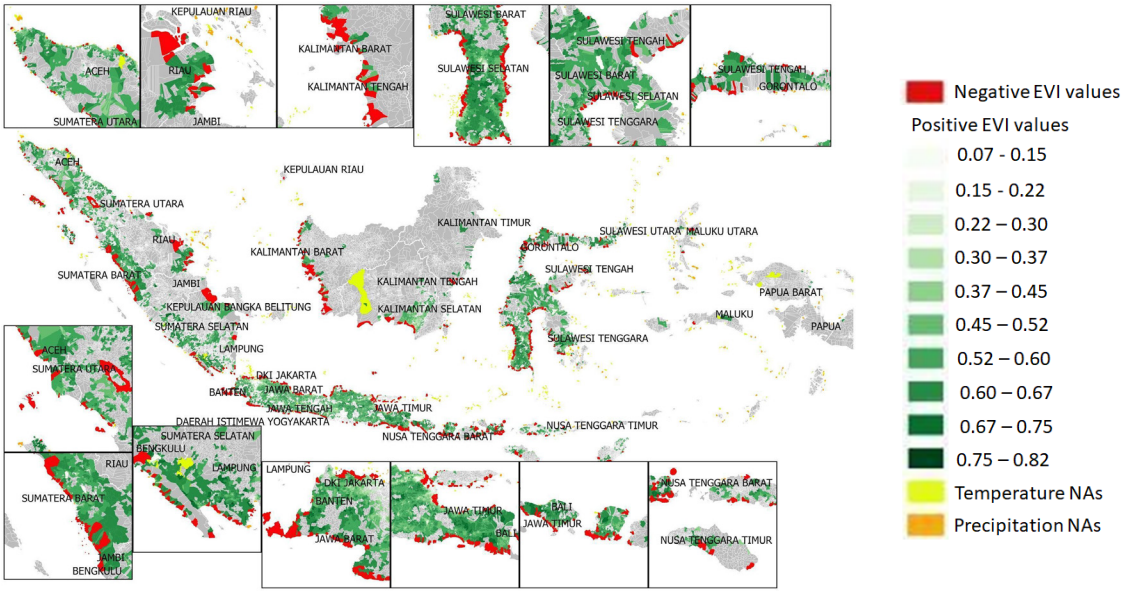

Figure 17: The location of missing EVI, precipitation and temperature data across Indonesia, after extracting the raster values with the village administrative boundaries.

- **Complex Land Cover Errors:** Where the rice land use area is relatively small and surrounded by forest or plantations, there could be a potential source of error resulting from the MODIS imagery, which has a moderate spatial resolution of 250 metres per pixel (Setiawan et al. 2013). There is a possibility that the EVI value extracted from the rice area is influenced by the spectral signature of the surrounding land use (Setiawan et al. 2013). Given the complexity of the landscape, it is possible that other high performing villages have low EVI values due to the spatial proximity of low reflecting land covers, i.e. urban settlements, roads and barren areas. For example, figure 18 presents two villages that are potentially subject to such errors. In the village to the left, the rice mask within this village is mostly rice, with some other vegetation within. However, the rice growing area is surrounded by forestry and native vegetation. It is unknown if the spectral signature of the surrounding land cover influences the EVI value within the rice mask. Whereas in the village to the left, there is a large amount of urban and industrial land cover within and surrounding the rice mask, potentially reducing the EVI value.

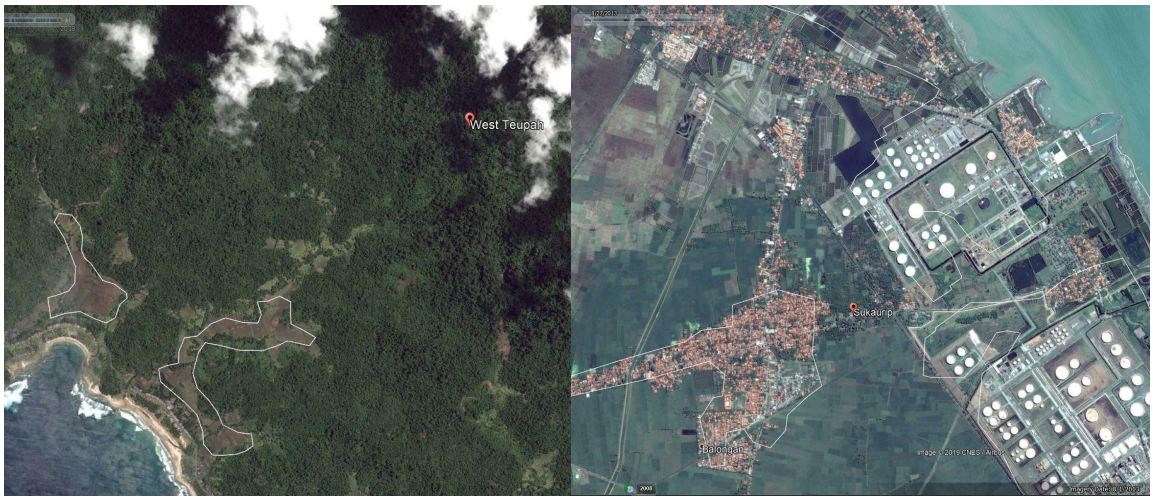

Figure 18: On the left is West Teupah Village, Aceh (HE 13). The imagery shown was obtained the 5th of March 2014. The dedicated rice growing area in this village is mostly rice, with some other vegetation within. On the right, Sukaurip, Indramaya Regency (HE 14). The imagery was collected the 17th of August 2013.

Recommendations for Future Work

- **Performance Measure:** The performance measure used in this study, i.e. average Max EVI, was calculated by extracting the Maximum EVI value for each pixel, within a village boundary, and averaging the Maximum value, across all pixels within this boundary. However, several nuances of rice crop phenology across a season are missed in this performance measure. For instance, using temporal vegetation indices and biophysical covariates data, for pixels within the rice mask, it is possible to predict, and construct a performance measure that captures multiple characteristics of rice crop phenology for the target season, such as initiation of greening (beginning of season), senescence (end of season), length of season etc, all of which can be captured per pixel, and then be aggregated to the rice mask for each village.

- **Studying Negative outliers:** With PLJ's focus on researching and developing innovative tools and approaches to promote inclusive growth that leaves no one behind, future investigations should include the negative outliers as well. Comparing the practices and enablers of positive deviants with those of negative deviants is crucial in designing effective interventions and for identifying the key elements responsible for higher productivity.

- **Rice Area Mapping:** Instead of using rice crop masks that are inaccurately separating areas growing rice from other land covers and land use, alternative methods could be used for rice area mapping. One of the methods is using MODIS multi-temporal satellite imagery to map rice areas. A study by Lee et al. (2012) produced a map of dryland distribution in Indonesia. The algorithm they developed uses time series of various vegetation indices (i.e. EVI and NDVI) to identify the initial period of dryland flooding and transplanting based on the sensitivity of the land surface water index (LSWI).

- **Control Variables:** In this analysis, we only use data from biophysical variables that were open source and readily available. There is a need to control for other drivers of agricultural productivity that farmers have no control over (such as soil) by engaging with agricultural experts, that could suggest such variables.

- **Expert driven validation:** Administrative/government agencies, who in this case are also data collectors, will be the end users of the results obtained from the proposed PD approach. Therefore, it is necessary to validate these results, through focus group meetings and consultations to understand: 1) If indeed the variables identified with the PLS/Bivariate analysis, differ between true outliers from the remaining villages in the same HE; 2) if these are also determinants of higher average Max EVI values among true outliers and 3) in comparison to the rest of the outliers, and in general if agricultural performance among true outliers is better than other villages in the same HE.

- **Complex Land Covers:** Land cover across Indonesia, including agriculture is complex, as 80% of 250 m pixels are not homogenous (Setiawan et al. 2011) and we suspect that there are potential influences from the reflectance of multiple land covers within a MODIS pixel (Setiawan et al. 2013). Obtaining higher spatial resolution data land use data and (Setiawan et al. 2011; Setiawan et al. 2014). In the future, it may be worthwhile to explore Landsat imagery with a 30 metres spatial resolution and other imagery to obtain more accurate agricultural mapping (Mengue et al. 2019).

- **Google Earth Validation:** The open sourced imagery available in the Google Earth Time Scale Tool does not have a consistent time stamp across Indonesia. Some areas have more available imagery and some areas of interest did not have imagery available during the target time period. We consistently tried to inspect imagery with a collection data as close as possible to our target time period to the chance of land cover change but due to the time difference. By inspecting imagery as close as possible, we assume that the land cover, especially the rice and forest cover did not change between the date of collection and our target date.

Conclusion

The positive deviance (PD) approach for development programming relies on identifying and scaling the strategies of positive deviants (PDs) i.e. individuals or communities who use uncommon practices and behaviours that enable them to achieve better outcomes than their peers. Disseminating and analyzing the behaviours and other factors underpinning PDs are demonstrably effective in delivering development results. However, conventional PD approaches are time and labour-intensive, and often are not scalable across communities. This is because they rely mainly on primary data collection for the identification of PDs with costs proportional to the sample size. Hence, the samples are usually small, which can make it hard to identify PDs statistically and practically, given their relative rarity. Innovations in digital technologies and platforms that record, mediate or observe individual and community behaviors, led to the proliferation of digital datasets "big data" (e.g. online data, mobile data and earth observation data) that could enable us, in specific domains, to identify and understand PDs in new and/or better ways (Albanna & Heeks, 2019). In agricultural development programming, earth observation (EO) data have the potential to provide deeper insights on behaviours of rural communities at temporal and spatial scales not previously possible using conventional methods.

In this study we presented a stepwise approach for the identification and validation of PD rice villages in Indonesia i.e. villages with significantly higher agricultural productivity in comparison to neighbouring villages with similar socio-economic and environmental conditions. It is a step towards building evidence for the use of big data to facilitate PD related development programming in agriculture. We were able to demonstrate that big data sources (such as EO data), can be combined with administrative data, in order to spatially locate and identify PD communities and some of their underlying practices such as straw burning, demographic variables such as average age, and contextual variables such as type of irrigation; which is a precursor to mainstreaming PD in development programming.

The presented analysis shows that the administrative data was able to explain variance in agricultural performance - captured through the EO derived measure - ranging from 21% to 75% across the 15 pre-determined homologues. This suggests that there are factors affecting performance that are not fully captured using the administrative data and some of those factors could be identified through extensive ground surveys and ethnographic methods. Collecting such data for large samples is difficult. However, in this study we provide a systematic way to identify information rich small samples - characterized in true outliers or PDs - that could be targeted for ground data.

We specifically focus on the use of administrative data, as we see national governments, as primary stakeholders. Through this study, we provide evidence that administrative data, when combined with open source Earth Observation data, can be reused in multiple ways to facilitate targeted development planning.

Although the advantage of big data is in its ability to provide more data, the trade-off however is that more data, also could incorporate more noise. Therefore, in this study, we focused on developing a statistically rigorous approach, with multiple validation steps, in our bid to separate noise from true insights.

The developed methodology can be used to draw other insights from the combination of open source EO with administrative data, that are not directly connected to PD. For instance, we identified 4 homologous environments (HEs) within Aceh province, whereas 3 HEs where in Java, yet the sampling strategy for the two provinces is the same in the Agriculture census, thereby leading to under-representation of diverse conditions and complexities in Aceh, within the census.

This proof-of-concept analysis contributes to the evidence that big data sources and analytics, administrative data, and open source EO data, has the potential to facilitate mainstreaming of PD into development programming, further empowering national and local governments, with methods that can enable targeted bottom-up solution development. However, the roll out of this method would require the use of recent administrative data along with EO data in order to move to the next stage of PD inquiry i.e. ground surveys and ethnographic methods targeting the true outliers or PDs to understand their underlying behaviours.

References

Albanna, B. and Heeks, R., 2019. Positive deviance, big data, and development: A systematic literature review. *The Electronic Journal of Information Systems in Developing Countries*, *85*(1), p.e12063.

Bolton, Douglas K., and Mark A. Friedl. "Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics." *Agricultural and Forest Meteorology* 173 (2013): 74-84.

Cinner, J.E., Huchery, C., MacNeil, M.A., Graham, N.A., McClanahan, T.R., Maina, J., Maire, E., Kittinger, J.N., Hicks, C.C., Mora, C. and Allison, E.H., 2016. Bright spots among the world's coral reefs. *Nature*, *535*(7612), p.416

de Vries, F.P. ed., 2005. *Bright spots demonstrate community successes in African agriculture* (Vol. 102). IWMI.
Johnson, D.M., 2014. An assessment of pre-and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. *Remote Sensing of Environment*, *141*, pp.116-128.

Hartini, T.N.S., Padmawati, R.S., Lindholm, L., Surjono, A. and Winkvist, A., 2005. The importance of eating rice: changing food habits among pregnant Indonesian women during the economic crisis. *Social science & medicine*, *61*(1), pp.199-210.

Heute, A., Didan, T., Miura, T., Rodriguez, E.P., Gao, X., and Ferreira, L.G., 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices, *Remote Sensing of the Environment, 83,* pp 195 - 213.

Kleinbaum, D.G., Kupper, L.L., Muller, K.E. and Nizam, A., 1988. *Applied regression analysis and other multivariable methods* (Vol. 601). Belmont, CA: Duxbury Press.

Lee, N., Monica, A. and Daratista, I., 2012. Mapping Indonesian dryland using multiple-temporal satellite imagery. *African J. Agricul. Res.*, *7*(28), pp.4038-4044.

Lantican, M.A., Lampayan, R.M., Bhuiyan, S.I. and Yadav, M.K., 1999. Determinants of improving productivity of dry-seeded rice in rainfed wetlands. *Experimental Agriculture*, *35*(2), pp.127-140.

Maitra, S. and Yan, J., 2008. Principle component analysis and partial least squares: Two dimension reduction techniques for regression. *Applying Multivariate Statistical Models*, *79*, pp.79-90.

Mandal, K.G., Misra, A.K., Hati, K.M., Bandyopadhyay, K.K., Ghosh, P.K. and Mohanty, M., 2004. Rice residue-management options and effects on soil properties and crop productivity. *Journal of Food Agriculture and Environment*, 2, pp.224-231.

Mengue, V.P., Fontana, D.C., da Silva, T.S., Zanotta, D. and Scotta, F.C., 2019. Methodology for classification of land use and vegetation cover using MODIS-EVI data, *Revista Brasileria de Engenharia Agricola e Ambeintal, 23*(11), pp. 812-818.

Mkhabela, M.S., Bullock, P., Raj, S., Wang, S. and Yang, Y., 2011. Crop yield forecasting on the Canadian Prairies using MODIS NDVI data. *Agricultural and Forest Meteorology*, *151*(3), pp.385-393.

Noble, A. et al. in Bright spots demonstrate community successes in African agriculture (ed. F. W. T. Penning de Vries) 7 (International Water Management Institute, 2005)

Osanyinlusi, O.I. and Adenegan, K.O., 2016. The determinants of rice farmers' productivity in Ekiti State, Nigeria.

Pant, L.P. and Hambly Odame, H., 2009. The promise of positive deviants: bridging divides between scientific research and local practices in smallholder agriculture. *Knowledge management for development journal*, *5*(2), pp.160-172.

Qiu, B., Zeng. C., Tang, Z., and Chen, C. 2013. Characterizing spatiotemporal non-stationarity in vegetation dynamics in China using MODIS EVI dataset. Environmental Monitoring and Assessment. 185. pp. 9019-9035.

Setiawan, Y., Yoshino, K., Philpot, W.D. 2011. Characterizing temporal vegetation dynamics of land use in regional scale of Java Island, Indonesia. *Journal of Land Use Science, 8.,* pp. 1- 30.

Setiawan., Y., Yoshino., K. and Prasetyo., L.B., 2013. Characterizing the dynamics change of vegetation cover on tropical forestlands using 250m multi-temporal MODIS EVI. *International Journal of Applied Earth Observation and Geoinformation, 26,* pp. 132-144.

Son, N.T., Chen, C.F., Chen, C.R., Minh, V.Q. and Trung, N.H., 2014. A comparative analysis of multitemporal MODIS EVI and NDVI data for large-scale rice yield estimation. *Agricultural and forest meteorology*, *197*, pp.52-64.

Sternin, J. (2002) 'Positive deviance: a new paradigm for addressing today's problems today', *The Journal of Corporate Citizenship*. Greenleaf Publishing, pp. 57–63

Steinke, J., Mgimiloko, M.G., Graef, F., Hammond, J., van Wijk, M.T. and van Etten, J., 2019. Prioritizing options for multi-objective agricultural development through the Positive Deviance approach. *PloS one*, *14*(2), p.e0212926.

Tucker, C.J. and Sellers, P.J., 1986. Satellite remote sensing of primary production. *International journal of remote sensing*, *7*(11), pp.1395-1416.

Wishik, S. M., & Van Der Vynckt, S. (1976). The use of nutritional "positive deviants" to identify approaches for modification of dietary practices. *American Journal of Public Health*, 66(1), 38–42. https://doi.org/10.2105/AJPH.66.1.38

Acronyms

BAPPENAS Ministry of National Development Planning of the Republic of Indonesia
CHIRPS Climate Hazards Group InfraRed Precipitation with Station data
CMAP CPC merged analysis of precipitation
EO Earth Observation
EVI Enhanced Vegetation Index
GPM Global Precipitation Measures
HE Homologues Environment
MODIS Moderate Resolution Imaging Spectroradiometer
MSE Means Squared Error
NDVI Normalized Difference Vegetation Index
PCs Principal Components
PCA Principal Component Analysis
PLS Partial Least Square
PD Positive Deviance
PDs Positive Deviants
PODIS Village Potential Survey
RMSE Root Mean Squared Error

Appendix

Agricultural Census

Variable Code	Variable Name	Variable Values	Variable Labels
prop	Province		
kab	District		
kec	Subdistrict		
desa	Village		
Idruta	Household number		
r103	Age of head of household		
r104	Gender of Head of Household		
r201	Rice Farming business	1	Yes
		0	No
r213l	No. of Male Household managing agribusiness	Numeric	
r213p	No. of Female Household managing agribusiness	Numeric	
r214	Main types of household business	201	Rice Farming
		202	Other Crops
		203	Horticulture
		204	Plantation
		205	Livestock
		206	Fish farming
		207	Catching Fish

		208	AquaCulture
		209	Wild Animals
		2010	Agricultural Service
r217	Sex of the main famer of the main business household	1	Male
		2	Female
r301ak1_2	Code of the paddy rice plant	1101	wetland rice
r301ak2 wetland	Season1	numeric	area of rice m square
r301ak3	Season2	numeric	area of rice m square
r301ak4	Season3	numeric	area of rice m square
r301ak5	Sum	numeric	area sum
r301ak6	Main harvesting method	1	Harvested young
		2	Harvested other forms
		3	Harvested yourself
		4	Released
		5	Allowed
		6	Not harvested
r301ak7	The yields are sold/exchanged for wetland rice	1	Yes
		2	No
r301ak8	Management status	1	Manager Owned
		2	Managed with revenue sharing
		3	Managing and you are receiving a wage
r301bk1_2	Code of the rice field crops	1102	dryland
r301bk2 Rice Fields	Season1	numeric	area of rice m square
r301bk3	Season2	numeric	area of rice m square
r3o1bk4	Season3	numeric	area of rice m square
r301bk5	Sum	numeric	area sum
r301bk6	Main harvesting method	1	Harvested young
		2	Harvested other forms
		3	Harvested yourself
		4	Released
		5	Allowed
		6	Not harvested
r301bk7	The yields are sold/exchanged for field rice	1	Yes
		2	No
r301bk8	Management status	1	Manager Owned
		2	Managed with revenue

			sharing
		3	Managing and you are receiving a wage
r302	Types of rice plants that have highest production value	1101	wetland rice
		1102	Rice fields
r306a	Household members doing agricultural services other than farm labor	1	Yes
		2	No
r306b1	Household members doing rice products	1	Yes
		2	No
r901a1k2	Area of irrigated rice fields	numeric	area
	Location of irrigated rice fields	1	in the village
		2	outside the village within the subdistrict
		3	outside the sub district within the district
		4	outside the district
r901a2k2	Area of simple irrigation	numeric	area
r901a3k2	Area of rainfed	numeric	area
r901a4k2	Area of tidal swamp	numeric	area
r901a5k2	Area of wetland (swampy swamp)	numeric	area
r901a6k2	Agricultural land that is rice	numeric	area
r901b8k2	Agricultural land that is not rice	numeric	area
r902k2	Non-agricultural land	numeric	area
r903k2	Total land (Agricultural and non-agricultural)	numeric	area

PODIS

Variable Code	Variable Name	Variable Values	Variable Labels
R101	Province Code		
R102	Regency Code		
Q103	District Code		
Q104	Village Code		
R101N	Province Name		
R102N	Regency Name		
R103N	District Name		
R104N	Village Name		

R301	Government Status	1	Village
		2	Village
		3	UPT/SPT
		4	Other
		5	Natagara
R302	Consultative Body	1	Yes
		0	No
R303	Village Boundaries Lawful Map	1	Yes
		2	No
R304a	Existence of a local Environmental Unit	1	Yes
		2	No
R305B	Topography	1	Slope/Peak
		2	Valley
		3	Plain
R306	Village Head Office Location	1	Yes in the village
		2	Yes outside the village
		3	No Office
R307a	Village direct access to the ocean	1	Yes
		2	No
R307B1A	Fishing	1	Yes
		2	No
R307B1B	Utilization of fishing for aquaculture	3	Yes
		4	No
R307B1C	Utilization of fishing for salt ponds	5	Yes
		6	No
R307B1D	Utilization of Ocean for tourism	7	Yes
		8	No
R307B1E	Utilization of Ocean for public transportation	1	Yes
		2	No
R307B2	Existence of Mangroves	1	Yes
		2	No
R308A	Where is the village located	1	In the forest
		2	At the edge of the forest
		3	Outside the forest

R308B	Forest Function	1	Conservation
		2	Production
R4031	Residents working abroad	1	Yes
		2	No
		3	I Don't Know
R403B1	Male Migrant workers	numeric	
R403B2	Female Migrant workers	numeric	
R404A	Main source of Income	1	Agriculture
		2	Mining and excavation
		3	Manufacturing
		4	Trading
		5	Transportation
		6	Service
R404B1	Main commodity	1	Rice
		2	Other Crops
		3	Horticulture
		4	Plantation
		5	Animal Husbandry
		6	Capture fisheries
		7	Aquaculture
		8	Forestry
R404B2	Road Surface type from village to agricultural area	1	Concrete
		2	Hardened
		3	Land
		4	Other
R501A1	Families with PLN Electricity	numeric	
R501A2	Families without PLN Electricity	numeric	
R501B	Families without Electricity	numeric	
R503	Cooking fuel used by household	1	City gas
		2	LPG
		3	Kerosene
		4	Firewood
		5	Other
R504	Access to bathrooms	1	Independently
		2	Together
		3	Public toilets
		4	No toilets

R506	Drainage system	1	Infiltration hole
		2	Drainage Sewage system
		3	River or Ocean
		4	In a hole
		5	other
R507B	Source of drinking water	1	Bottle of water
		2	Plumbing with ammeter
		3	Plumbing without ammeter
		4	Drilling well
		5	well
		6	spring
		7	River or Lake
		8	Rainwater
		9	Other
R508AK2	Existence of river	1	yes
		2	No
R508AK3	Existence of Irrigation Channels	1	yes
		2	No
R508AK4	Existence of lake or reservoir	1	yes
		2	No
R508B3K2	Use rivers for irrigation of agricultural land	1	yes
		2	No
R508B3K3	Use of irrigation channels for irrigation	1	yes
		2	No
R508B3K4	Use of lakes and reservoirs for irrigation	1	yes
		2	No
R511A	Existing of Slums	1	yes
		2	No
R512AK2	Pollution Incident (water)	1	yes
		2	No
R512BK2	Pollution Incident (Soil)	1	yes
		2	No
R512CK2	Pollution Incident (Air)	1	yes
		2	No
R513	Burning fields for agricultural purposes	1	yes
		2	No

R601AK7	How many landslides in 2013	numeric	
R601BK7	How many flood in 2013	numeric	
R601CK7	How many Flash Flood in2013	numeric	
R601DK7	How many earthquakes in 2013	numeric	
R601JK7	How many drought events in 2013	numeric	
R701AK2	Number of levels of education	numeric	
R702A	Functional Literacy activities	1	Yes
		2	No
R704AK2	Hospital Facilities	1	Yes
		2	No
R709(All)*	Health Epidemics	1	Yes
		2	No
R710	Number of residents suffering from bad nutrition	numeric	
R807A	Habit of mutual cooperation of residents	1	Yes
		2	No
R1001B2	Roads can be accessed by cars or larger	1	All year long
		2	all year except certain times
		3	all year except wet season
		4	Not passable
R1103A_K4	Conversion from rice to non-rice agriculture	1	Yes
		2	No
R1103C_K2	Conversion from non-rice to rice agriculture	1	Yes
		2	No
R1205	Number of markets without buildings	numeric	
R1212B	Number of small industry cooperatives	numeric	
R1212C	Number of savings and loans cooperatives	numeric	
R1213A	The existence of stalls that sell agricultural production facilities owned KUD	1	Yes
		2	No
R1214C	Small Business credit facility	1	Yes

	received by residents		
		2	No
R1401A4_K2	Programmes community development, irrigation, markets, agriculture	1	Yes
		2	No
R1401A4_K3	Source of programme intervention	1	PNPM
		2	Non PNPM
		3	PNPM and Non PNPM
R1401A4_K4	Programme implementers	1	poor population
		2	non-resident
		4	farmer
		8	business group
		16	other
R1401A4_K5	Direct beneficiaries	1	poor population
		2	non-resident
		4	farmer
		8	business group
		16	other
R1401B1_K2	Programmes for capacity building, lending for agriculture	1	Yes
		2	No
R1401B1_K3	source of the programme	1	PNPM
		2	Non PNPM
		3	PNPM and Non PNPM
R1401B1_K4	Implementers	1	poor population
		2	non-resident
		4	farmer
		8	business group
		16	other
R1401B1_K5	Direct beneficiaries	1	poor population
		2	non-resident
		4	farmer
		8	business group
		16	other
R1501A_K3	Revenue	numeric	
R1501B_K2	Village fund allocation value	numeric	
R1503C	The existence of village market assets	5	Yes

		6	Nothing
R1601A_K	Gender of the head	1	Yes
		2	Nothing
R1601A_K5	Education of the head	1	Never attended school
		2	Not finished
		3	Graduated from elementary school / equivalent
		4	Junior high school / equivalent
		5	High school / equivalent
		6	Academy / DIII
		7	Diploma IV / S1
		8	S2
		9	S3
KCR803B2K2	Number of markets specifically fruit and vegetables	numeric	
KCR803B3K2	Number of special markets for rice	numeric	
KCR803B3K3	Special types of rice market buildings	1	Permanent
		2	Semi permanent
		4	No Building
KBR801A	Disaster management efforts	1	Yes
		2	Nothing