Assessing rural productive capabilities and identifying potential products by municipality

Final report submitted to Bancóldex and National Planning Department as part of Datlas 2.0 project

Ricardo Hausmann
Principal Investigator
Director of the Center for International Development
Harvard University
Cambridge, Massachusetts

Authors:[1]
Sid Ravinutala
Andres Gomez-Lievano
Eduardo Lora

CENTER FOR INTERNATIONAL DEVELOPMENT

GROWTH LAB

cid.harvard.edu

Center for International Development at Harvard University

[1]We acknowledge many useful comments made by Bancoldex and DNP technical staff at the Seminar held in Bogotá, July 12-13, 2017. We also thank Juan Camilo Medellín for compiling and processing some of the databases used here. We also received useful comments and suggestions from Frank Neffke, Dario Diodato, Ljubica Nedelkoska, Michele Coscia and Matte Hartog.

Contents

List of Tables

List of Figures

1 Introduction

How can the productive capabilities of each municipality be unleashed taking into consideration the resources available to them? A first pass at this ambitious question begins by understanding the set of outputs a municipality is capable of producing. We answer this by discovering relationships between agricultural inputs and outputs and ask a relatively simpler question: how similar are agricultural outputs in terms of the inputs they use? Answering this question is made difficult by the fact that most UPAs cultivate just one or two crops. This may be a rational response to economies of scale. Given a plot of land and inputs, it may be easier to cultivate one crop on the entire land than plant a number of them with each requiring a different care regimen[2]. It may be that the inputs available only allow for a few types of crops.

In this paper, we use the rural census data from Colombia to build an agricultural product space capturing the similarities between outputs. We test the predictive power of the product space and use this to answer the question above. In section 2, we discuss the various sources of data and how they are merged, cleaned, and transformed before processing. In section 3, we look at some high level features of the dataset and how inputs, outputs, and land use are related. In section 4, we explore the mechanics of diversification. We construct similarity and density matrices and show that they do indeed predict what a municipality produces. Finally, in section 5 we use Machine Learning algorithms and the density matrices to predict municipalities that are best suited to produce a given output. Further, we identify "missing" municipalities-output pairs i.e. municipalities that should be producing a given output at high yield but currently are not. Finally, we summarize our findings and suggest areas for further work.

In this report we will be making extensive use of concepts described in more detail in the companion report "How Industry-Related Capabilities Affect Export Possibilities", especially with respect to Machine Learning techniques.

[2]We acknowledge that there are also reasons to not produce just a few crops. Demand may not be elastic; increased production of one crop may reduce its market price. Second, multiple crops allow the farmer to hedge against crop specific pests and diseases.

2 Data

The main source of data for this paper is the (2013-14) Agricultural Census, which DANE[3] has recently released for public use. After some basic processing of the relevant variables for our purposes, the farm[4] level data for the rural census dataset in provided as a series of *base* Stata files labelled A to H. Table 2 details all the variables in the dataset.

- Content of each *base* Stata file:
  - A and B have inputs.
  - C has non-farm activities (e.g., "sugar refinery").
  - D has raw agricultural outputs.
  - E has cattle.
  - F and G have fishing outputs[5].
  - H has some demographics such as poverty.
- The bases A and B and C have 2,913,163 rows, which correspond to *all* farms. D has only the subset of those farms (958,530) that produce raw agricultural outputs and forestry, called agricultural production units, UPAs, in the Census. The remaining farms being the non-agricultural producing units, or UPNAs. We include all farms, UPAs and UPNAs, in the dataset.
- Dataset has 483 different agricultural product (excluding fish), and around 400 possible input variables. Some of these will be described in more detail in the sections that follow.

To create our agri-product space we make use of files A through E. We create a dataset where each row is a farm and columns are inputs and outputs. We also remove approximately 100,000 observations, where output is produced solely for self-consumption or barter (see fig. 2 below for more detail).

In addition to this farm level data, municipality level characteristics such as soil type, governance and public expenditure, land conflict indicators, distance to other municipalities, and population numbers are available from sources other than the Census. We use these as controls to test the robustness of our models in section 4.5.

2.1 Variables in the microdata

Here we will present a rough overview of some of the variables of the microdata at the level of farms. The goal of this subsection will simply be to provide a sense of what the Data consists of, and a rough idea of how the typical farms in Colombia look like.

First, we present the types of columns of interest in our analysis:

[3] Departamento Administrativo Nacional de Estadística

[4] We use the term 'farms' to include both UPAs and UPNAs

[5] F has information on fisheries, while G on fishing activities

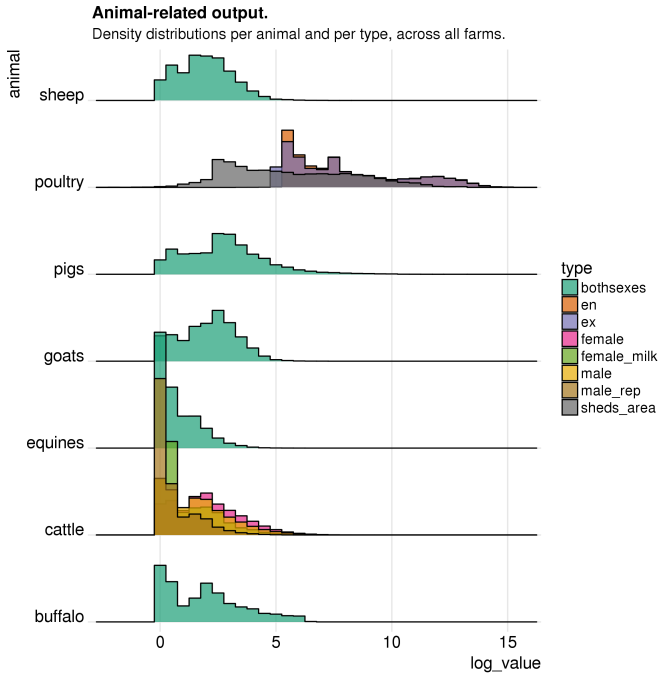

Figure 1: The x-axis in the density plots is the logarithm of the counts of "heads".

Animals: Variables for farm animals are individually coded by the type of animals, and by some of their characteristics, such as sex and age: buffalo, cattle, equines, goats, pigs, poultry and sheep. These consist of a total of 28 different columns. For pigs, buffalo, equines, goats, and sheep, we collapse and count "both sexes". For cattle, however, we separate by female and male, and also whether they are used for milk production or reproduction[6]. Poultry also has several types, but we collapse between it into whether it is for domestic consumption (coded using "en"), or for exports (coded using "ex"). Figure 1 shows the distribution of counts and output related to animals across farms.

Non-agricultural activities: There are 68 columns that code for non-agricultural activities. These columns are recognized because they start with "agric_nonagric" and "nonagric_nonagric", which is what identifies UPAs from UPNAs. Each come in 34 different types. Table 1 lists these 34 types, and how they are coded in the data. These are binary variables, and farms (both UPAs and UPNAs) report whether they have, or not, non-agricultural activities. The table shows the number of farms (per 1,000 farms) that reported "yes" to each of the 34 types for "agric_nonagric" (UPAs) and "nonagric_nonagric" (UPNAs) across non-agricultural activities. As is observed, not many farms report non-agricultural activities.

Destination-types: One of the important characteristics of farms is that their production has different uses and is destined for different purposes. These consist of eleven different types of destinations. As mentioned, we care about the production whose destination is not self-consumption. Hence, we create

[6]Note that the two sets are a not not mutually exclusive. E.g. A cattle can be female and milk producing. Further analysis may wish to include just male and female cattle.

Table 1: Non-agricultural activities, and the number of UPAs and UPNAs that have them (per 1,000 farms).

Type	Meaning	UPAs	UPNAs
comm1	Comercio de productos alimenticios y bebidas alcoholicas	1.43	2.50
comm2	Comercio de productos diferentes a alimentos y bebidas alcoholicas	0.91	1.37
minmfg1	Fabricacion de productos de plÃąstico, metalurgicos, sustancias y productos quimicos	0.12	0.25
minmfg2	Petroleo	0.04	0.34
minmfg3	Mineria con titulos	0.54	0.79
minmfg4	Mineria sin titulos	1.01	0.53
minmfg5	Gas, generacion y transmision de energia	0.07	0.18
rawpro1	Extraccion de aceite	0.33	0.09
rawpro10	Produccion de alimentos para consumo humano	1.07	0.28
rawpro11	Elaboracion de alimentos preparados para animales	0.30	0.04
rawpro12	Destilacion de bebidas alcoholicas o fermentadas	0.29	0.02
rawpro13	Obtencion de biocombustibles	0.06	0.10
rawpro14	Elaboracion de artesanias	5.15	1.39
rawpro15	Acerrado, impregnacion de madera	0.34	0.20
rawpro16	Fabricacion de pulpas	0.02	0.02
rawpro17	Fabricacion de muebles	0.52	0.16
rawpro2	Fabricacion de azucar	0.09	0.02
rawpro3	Molineria de arroz	0.06	0.04
rawpro4	Desmote de algondon	0.05	0.01
rawpro5	Elaboracion de panela y mieles	3.75	0.36
rawpro6	Transformacion de productos de la flora	0.32	0.07
rawpro7	Fabricacion de productos de caucho	0.04	0.01
rawpro8	Sacrificios de animales	0.44	0.16
rawpro9	Procesamiento de leche	0.61	0.17
serv1	Actividades de apoyo a la actividad agricola	21.10	1.42
serv10	Actividades ambientales	0.36	0.65
serv2	Actividades de apoyo a la ganaderia	9.11	0.68
serv3	Actividades de apoyo a la silvicultura	0.85	0.20
serv4	Servicios turisticos, de alojamiento, hospedaje y otros	0.66	1.31
serv5	Servicios educativos	0.75	3.95
serv6	Servicios de salud	0.09	0.39
serv7	Servicios religiosos	0.25	0.72
serv8	Servicios recreativos	0.16	0.36
serv9	Servicios de seguridad nacional	0.05	0.12

a new columns that separates explicitly those. Figure 2 presents these differences, and the number of farms corresponding to the different destinations.

Inputs: In the dataset, several columns start with the code name "farms_" and "famach". These are the columns which we identify as the inputs of production in farms. The former includes management practices, types of irrigation techniques, natural resources used, among others. The columns of "famach" are about machinery, and the quantities of them per age. All together they consist of 410 different input columns.

Outputs: Output variables are recognized because these are agricultural products whose production is measured in tons. Hence, in the dataset, there is a single column for each product and starts with "tons_". Many of these are sown and harvested. Accordingly, there are the corresponding column variables "harv_" and "sown_". In total there are 483 output products.

Lastly, to get a sense of what are the variables that generate the most important differences between farms, we run a Principal Component Analysis (PCA) over a sample of farms (100,000), and we include all columns for inputs, outputs, non-agricultural activities, and animals, described above. For PCA to be meaningful, we centered and scaled each column, including those that consist of binary values, so that all have mean 0 and standard deviation 1 (since we work with a sample of farms, it is possible that some columns only have a single value, such as 0, and when this occurs we drop those columns from the analysis). PCA returns, for each farm, a new representation in terms of a number of "scores" in some "principal components". If we restrict our analysis to just 2 "components", PCA returns 2 scores per farm. And a "component" here is a vector representation of the inputs, outputs, etc., with weights in each of the elements such that the vector resembles most of farms. In this sense, the first component is the vector that is the one

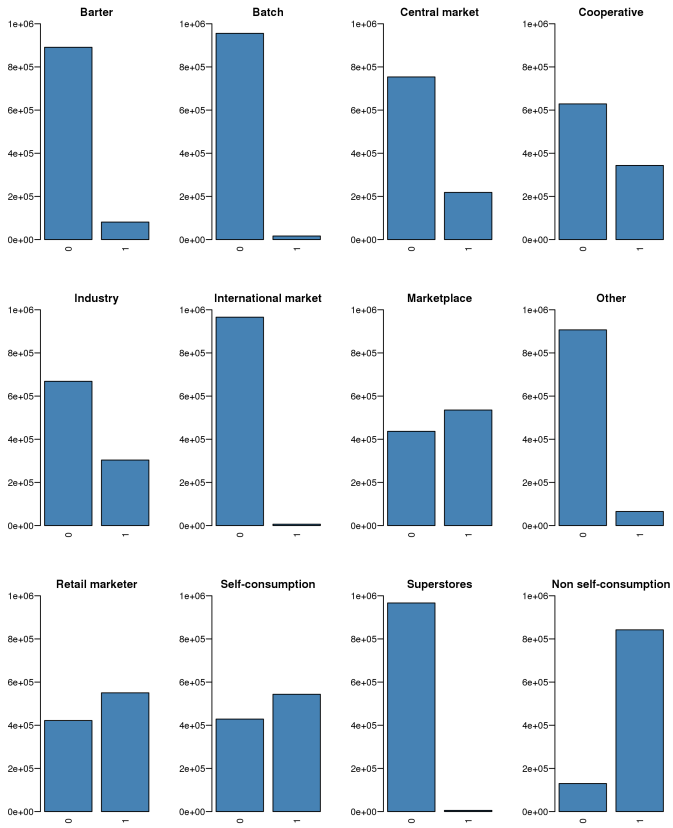

Figure 2: Count of the number of farms per production destination type.

closest to all farms, and therefore explains most of the variance (i.e., most of the differences in the vectors that represent each farm). The second component is another vector, orthogonal to the first component, that explains the most variance unexplained by the first. Hence, PCA represents each farms with 2 scores, which serve to tell us that farms (each of which is a vector of inputs, outputs, etc.) can be seen, on average, as the linear combination of these two "components" (which are fixed for all farms after PCA has been applied), each component weighted by the farm's corresponding scores.

The algorithm tells us the following results when we run it. First, that the first component explains approximately 20% of the variance across farms. And second, that this allow us to identify the dimensions that differentiate farms the most. Hence, we check the scores in the 1st principal component across farms, and identify those farms that have the highest value and those with the lowest. And then, we observe in what elements of the first component they differ the most on. A quick glance tells us that what differentiates farms the most is whether they own rather sophisticated, and uncommon assets, such as airplanes ("Avioneta"), boilers ("Calderas") or incinerator furnaces ("Hornos incineradores"). PCA itself is useful to get a sense of multidimensional data, and we will encounter PCA again in what follows (we will remind the reader again about the characteristics of this methodology). But, as a descriptive tool, PCA analysis here does not tell us if having or lacking any of these inputs increases production or not. This is precisely the question that we

will try to answer in the following sections.

2.2 Why UPA and UPNAs

As we noted earlier, the dataset constructed includes both UPAs and UPNAs even though we know before hand than UPNAs do not produce any agricultural produce. The main reason for this is to learn what kind of farms (based on available inputs, co-occuring products) do not produce any output. Put another way, the UPNAs provide a contrast to allow us to learn the features of a farm producing at high yield. As we will see, when doing predictions, we use data collapsed at the municipality level. Since there is no land allocated to crop output in UPNAs, only UPAs are inluded in this collapsed dataset.

In summary, in order to be able to predict the output of municipalities (coming just from UPAs), we also need to learn the features of farms *not* producing any output (i.e. UPNAs).

2.3 Creating outcome indices

We use three outcome variables.

- Yield: A measure of how productive a farm or municipality is at a given crop.
- Land usage: A measure of how intensely a crop is grown in a municipality, given the availability of soil that is best suited to it.
- Non-agricultural activities: A measure of intensity of non-agricultural activities in a municipality.

In this section, we elaborate how these outcome measures are constructed. These indices are constructed at different levels as outlined in the respective sections.

2.3.1 Creating yield indices

Our first variable of interest is yield. We create this by dividing the output by land harvested for each product at each farm. The 'yield' for animals is constructed slightly differently. Instead of ouput per harvested hectare, we look at the number of heads per hectare of pasture land.

$$y_{ip} = \frac{x_{ip}}{h_{ip}} \tag{1}$$

where:

x_{ip} : the output in tonnes (heads) of crop (animal) p at farm i

h_{ip} : the land harvested (pastures available) in hectares for crop (animal) p at farm i

This metric is also created at the municipality level.

$$y_{mp}^{muni} = \frac{x_{mp}^{muni}}{hi_{mp}^{muni}} \tag{2}$$

where:

x_{mp}^{muni} : the output in tonnes (heads) of crop (animal) p at municipality m

h_{mp}^{muni} : the land harvested (pastured available) in hectares for crop (animal) p at municipality m

Since we cannot directly compare yield between products, we instead create matrix R^{yield} of *yield indices* where elements r_{ip}^{yield} are the ratio of yield of product p at farm i and the average yield of that crop over all farms in Colombia.

$$r_{ip}^{yield} = y_{ip} \cdot \left(\frac{\sum_{n=1}^{N} h_{np}}{\sum_{n=1}^{N} x_{np}} \right) \tag{3}$$

Next, we discretize this as follows to produce matrix O^{yield} with element o_{ip}^{yield} to create a binary variable for outputs:

$$o_{ip}^{yield} = \begin{cases} 1 & \text{if } r_{ip}^{yield} > 1 \\ 0 & \text{if } r_{ip}^{yield} \leq 1 \end{cases} \tag{4}$$

Note that this assigns a 1 to farms where yield is above average, and 0 where it is not. Similarly, we create matrix O^{muni} at the municipality level.

2.3.2 Creating land usage indices

The methodology for creating the land usage variable is analagous to the previous section but the metric is constructed at the municipality level. We create a land usage metric as follows:

$$l_{mp} = \frac{h_{mp}}{s_{mp}} \tag{5}$$

where:

h_{mp} : the land harvested for crop p at municipality m

s_{mp} : the amount of suitable soil available for crop p at municipality m

To determine the amount of suitable land for crop we previously classified (almost) all products as either permanent or transitory and by climate (cold, mild, warm or any combination) based on the predom-

inant soil vocation of the municipalities where they are currently harvested (multi-soil municipalities were not taken into account for this classification). Information of soil vocations by municipality come from ICA and Instituto Geográfico Agustín Codazzi and was released in 2002).

Similar to the yield index, the land harvested index is constructed as a relative measure of land use in Colombia. Specifically, the land harvested index is given by:

$$r_{mp}^{land} = l_{mp} \cdot \left(\frac{\sum_{n=1}^{N} s_{np}}{\sum_{n=1}^{N} h_{np}} \right) \tag{6}$$

Next, we discretize this as follows to produce matrix O^{land} with element o_{mp}^{land} to create a binary variable for outputs:

$$o_{mp}^{land} = \begin{cases} 1 & \text{if } r_{mp}^{land} > 1.0 \\ 0 & \text{if } r_{mp}^{land} \leq 1.0 \end{cases} \tag{7}$$

2.3.3 Creating non-agri indices

We also create indices for intensity of non-agricultural activities at the vereda level as follows. Note that this includes both UPAs and UPNAs.

$$r_{vp}^{non\_ag} = \frac{n_{vp}}{z_p} \tag{8}$$

where:

$$n_{vp} = \frac{\text{\# of farms with non-agri activity } p \text{ in vereda } v}{\text{\# of farms in vereda } v}$$
$$z_p = \frac{\text{\# of farms with non-agri activity } p \text{ in Colombia}}{\text{\# of farms in Colombia}}$$

Next, discretize these indices as follows:

$$o_{vp}^{non\_ag} = \begin{cases} 1 & \text{if } r_{vp}^{non\_ag} > 1 \\ 0 & \text{if } r_{vp}^{non\_ag} \leq 1 \end{cases} \tag{9}$$

2.4 Creating indices for inputs

Most inputs are provided as dummies which means no transformations are necessary. A few inputs are not; these are usually counts of inputs. The inputs that are not dummies are:

- variables ending in '_w': Count of workers working in various activities
- variables ending in '_n': Count of machinery of a given type
- variables ending in 'less5n': Count of machinery of a given type less than 5 years old
- variables ending in '5moren': Count of machinery of a given type more than 5 years old
- variables starting with 'work': Count of the number of employees from within and outside of the household

We convert these to 'per hectare of harvested land' measures and then into binary variables as we did for outputs in equations 1, 3, and 4[7].

2.4.1 Soil type

A critical input that is not available in the Census data is 'soil vocation'. As mentioned in Section 2.3.2 above this data from the ICA and Instituto Geográfico Agustín Codazzi (2001) classifies the land in a municipality by crop type (permanent or transitory crops) and by climate – if it is cold, mild, or warm. We convert these shares to indexed measures as above and include them as inputs. Note that these are municipality level characteristics. In the absense of farm level soil vocation data, we assume all farms in the municipality have the same distribution of soil types as the municipality.

We call the final matrix containing all inputs and soil type data expressed as either dummies or binary variables, the I matrix[8].

2.5 The M matrix

Now we merge the input, I, and output matrices, O^{yield} to form matrix M with a row for each farm with ones and zeroes for inputs and outputs as shown in Figure 3.

2.6 Additional controls

Our hypothesis is that the ability to produce an output can be predicted by the availability of inputs and the presence of similar outputs. We add additional municipality level characteristics to improve the robustness of our models. In section 4.5, the municipality level variables are used as controls in the density regression.

[7] We divide inputs by the total area harvested. Note that we may be overestimating the total area harvested as some crops might be grown on the same piece of land but in different seasons. However, in the absence of a 'total area' variable, this is a reasonable approximation.

[8] Note that soil type is used as an input and is also used to construct the intensity of land use outcome variable. There are two major reasons this is not an issue. First, these input aand output matrices are condensed to similarities in Section 4, where the soil type of the current product is not taken into account. Second, these measures are at the farm level while predictions are at the municipality level. As a sanity check, we repeated the excercise omitting soil as an input and observed no significant change in our results.

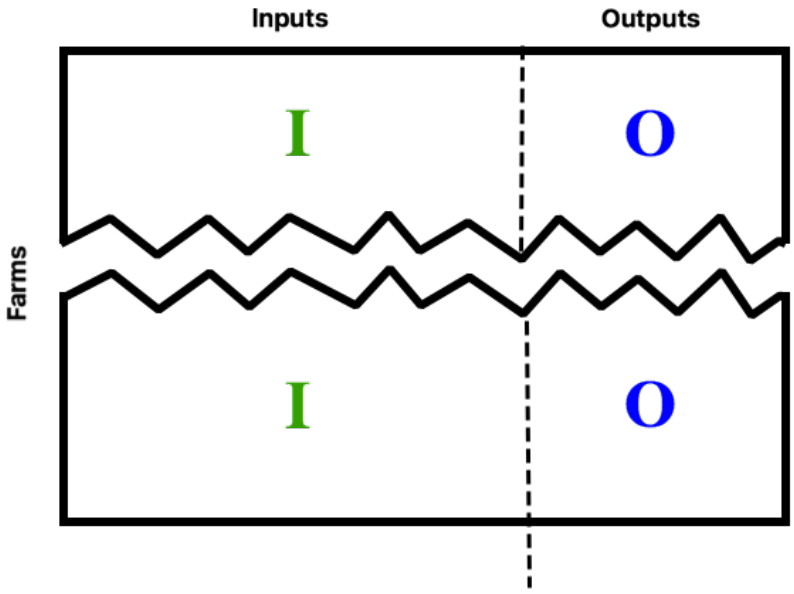

Figure 3: Matrix M with sub-matrices I and O.

In contrast, in section 5, these variables are added as additional predictors to the model. Indicators for governance, public expenditure, and land conflicts, which have been compiled by CEDE (Universidad de los Andes) from a large number of sources, most of them official.[9]. These use of these controls is noted as 'Muni-level Controls' in the regressions that follow.

2.6.1 Soil vocation data

Soil data was included in the I matrix (and therefore the M matrix). We also use these as controls to see if yields are directly influenced by the soil types available.

2.6.2 Adjusted yield index

The *adjusted yield index* is intended to capture the productivity of the municipality excluding the product of interest. It is calculated as the harvested land weighted average of the yield indices of all other crops in the municipality. Specifically, adjusted yield index, a_{mp} for a municipality m and product p, it is created as follows:

[9]The CEDE Municipality Panel can be accessed at `https://datoscede.uniandes.edu.co/microdatos-detalle.php/263/2/156/`.

$$a_{mp} = \sum_{i \neq p} w_{mi,p} \cdot r_{mi}^{yield} \tag{10}$$

where:

$$w_{mi,p} = \frac{h_{mi}}{\sum_{j \neq p} h_{mj}}$$

Note that when calculating the adjusted yield index for a crop, we consider the average yield of all the other crops in the municipality i.e. we exclude the crop of interest. This because we wish to create an index of the average yield of everything else that is grown in the region.

2.6.3 Non-agricultural activities

A number of farms (both UPAs and UPNAs) undertake non-agricultural activities, such as providing security services, milling rice, or making fruit juice. A problem with using these as outputs it that they do not depend on inputs directly but rather may be the result of high density of a related output in the region. For example, a farm with no inputs available may choose to make fruit juice if neighbouring farms are producing fruit.

Using non-agricultural activies as an input would give a substantial but unfair leg-up to our algorithms. It would be trivial to predict that one can find a rice farm near rice mills. Further, the direction of causality is not obvious. A farm may choose to produce rice if there is a rice mill in the vicinity. Equally possible is that a farm might choose to setup a rice mill since there are rice farms around. In order to separate these two effects, we use the vector of industrial employment in a municipality (from the PILA dataset) to predict the presence of non-agricultural activities. There are two underlying assumptions here. First, industrial activity drives non-agricultural activity - if there is a juice bottling plant we would expect a higher concentration of juice makers. Second, industry activity is not strongly correlated with agricultural activity. The validation of these assumptions requires deeper analysis and is left for future work.

In order to predict non-agricultural activities, we train and tune Support Vector Machine models using a radial basis function kernel. The AUC and ROC[10] for the models are shown in Figure 4. We use this trained model to generate a vector of non-agricultural activities for each municipality. Note that the industry vector is not a strong predictor. The predicted non-agricultural activity measure may indeed be noisy but would be largely free from the effect of nearby agricultural activity.

2.6.4 Population density

'Population density' may be a proxy for a few key predictors of agricultural activity, such as access to markets, access to public goods such as roads, and even access to knowledge. We constructed two metrics

[10] See section 5 for definitions of these metrics and the algorithm.

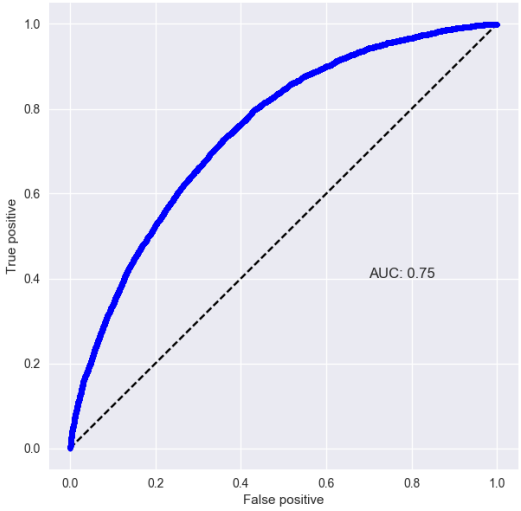

Figure 4: Predicting non-agricultural activities using industry employment in municipality.

for population density using the municipality populations and the distances betweem them[11].

$$
\begin{aligned}
pd_m^{linear} &= \sum_{i \in M} pop_i \cdot (1 - dist_{mi}) \\
pd_m^{gravity} &= \frac{\sum_{i \in M} pop_i \cdot pop_m}{dist_{mi}}
\end{aligned}
\tag{11}
$$

where:

$dist_{mi}$: distance between municipalities m and i normalized to (0,1)

pop_i : population of municipality i

[11] When distances between municipalities is not available, it's population is not used in the construction. This only applies to distances between municipalities far from each other and therefore does not effect the population density substantially. Future work may look to use better distance data to correct for it.

3 Descriptive Statistics

3.1 Inputs and outputs

Figure 5 shows the relationship between inputs and outputs at the municipality level. We created input and output indices as in section 2 but at the municipality level, instead of the farm level. Note that this results in a matrix of dummies where each row is a municipality. We take a sum of all inputs and outputs for each municipality. Since, these sums are log-normal in distribution, we log and rescale to produce the *input_mod* and *output_mod* variables. These variables allow us to explore the functional form of the relationship between inputs and outputs. The Pearson R coefficient for *output_log* and *input_log* is the highest suggesting a power-law relationship; a relative change in inputs is correlated with a relative change in outputs:

$$\ln(y_i) = \alpha + \beta \cdot \ln(x_i) \tag{12}$$

Table 4 shows the result of regressing output on inputs. We might have expected the relationship to be exponential i.e. a level increase in inputs leads to a percentae increase in outputs. This is the functional form observed when exploring the relationship between industries and product export. This might still be true here. Inputs are no completely independant. For example, soil protection practices are split into multiple column dummies, each for a different type of soil protection practice. Therefore these multiple columns actually code the same variable. Note that the Pearson R coefficient for *output_log* and *input_mod* is very close to *ouput_log* and *input_log*. It is probably that if the columns were coded differently, each representing an independent 'input', we would indeed observe an exponential relationship.

3.2 Yield index and the land use index

In section 2 we created two indices - yield and land use. Yield indices specify how productive the output is in a municipality whereas land use indices represent the intensity with which a crop is grown i.e. how much of the available suitable land in the municipality is dedicated to the crop. How similar are these measures? If we believe farmers to be rational agents who dedicate largest section of the land to the highest yielding crops, these measure are exactly the same. On the other hand, if we believe yield has little effect on the decision of what to grow is, these measure will be completely uncorrelated. We regress land usage indices on yield indices as follows:

$$r_{mp}^{land} = \beta_0 + \beta_1 \cdot r_{mp}^{yield} + \alpha_m + \gamma_p \tag{13}$$

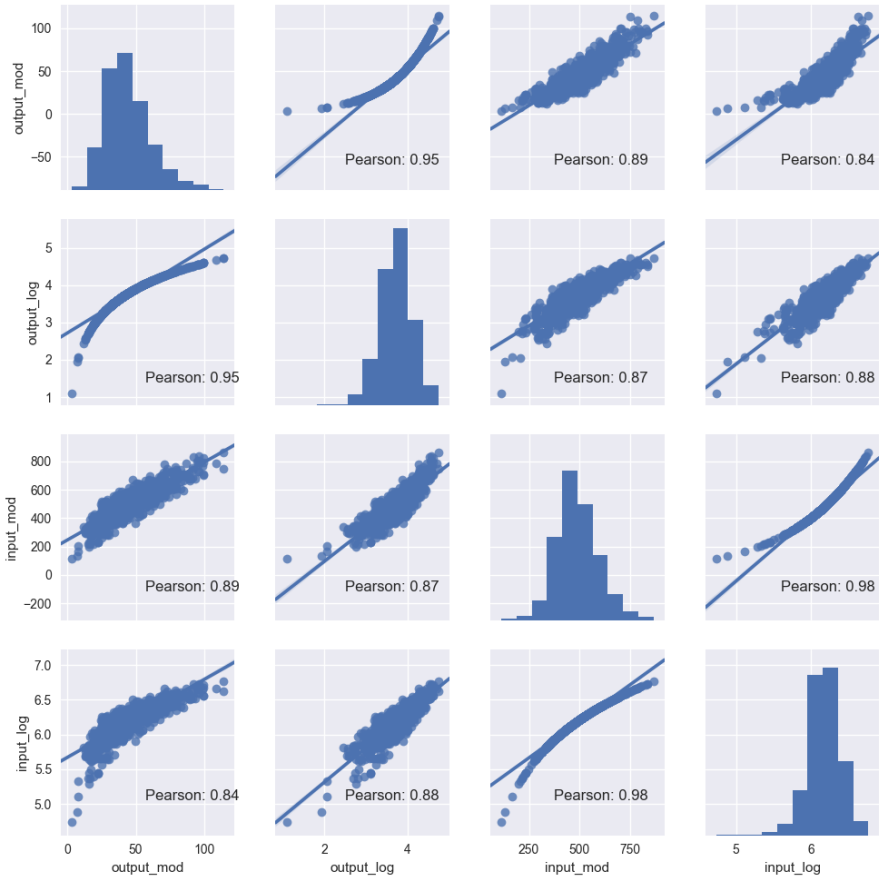

Figure 5: Distributions of input and output summary variables at the municipality level key

where:

$$r_{mp}^{land} : \text{land usage index}$$
$$r_{mp}^{yield} : \text{yield index}$$
$$\alpha_m : \text{municipality fixed effects}$$
$$\gamma_p : \text{product fixed effects}$$

Table 5 shows the results. Note that the coefficient on yield is positive and significant in all specifications. A one unit increase in yield index correlates with a 0.85 increase in land usage. Further, the R^2 is very high even without the use of fixed effects. A large portion of the variation in land usage is explained by yield. This leads us to conclude that yield and land usage are indeed closely tied. It may be that farmers select crops that provide high yield. Another possibility is that farmers have found ways to extract high yield for crops on which they focus. It is difficult to make any causal claims with just one year of data. Regardless, we can conclude that land allocation to crops is fairly optimized in municipalities.

3.3 Distribution of output yield indices

In section 2, we discretize the output yield indices by testing if they are above or below one. One concern may be that if the yield indices, r_{ip}^{yield} are in fact clustered close to one i.e. all farms are equally productive, we may be arbitrarily choosing half of the points. Figure 6 shows the distribution of the log of yield index for a sample of about 40 different outputs. Note that the threshold of one corresponds to a threshold of zero in log terms. We observe that although observations are distributed around zero in most cases (as should be expected), they do not bunch strongly at that level, and many have very wide distributions.

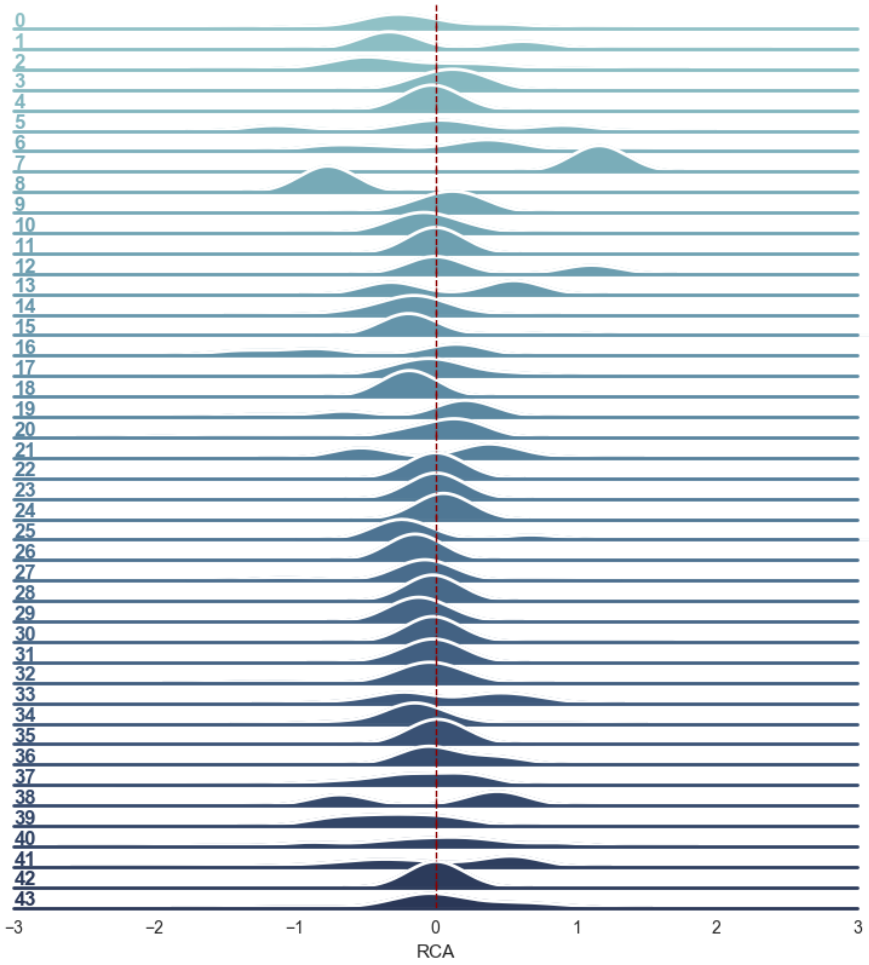

Figure 6: Distribution of log of yield index of a few output variables.

4 Similarity Matrices

In the previous section, we created a farm level dataset , M, with almost a million farms and with inputs and outputs as ones and zeroes. To understand relationships between inputs and outputs, we look at how often they co-occur at the farm level. The underlying assumption is that items that co-occur more frequently than expected are connected, or are *similar*, and the strength of their connection is proportional to how often they co-occur. In the next section, we construct density matrices from these similarity measures to test if our underlying assumption is valid.

Looking at outputs specifically, we can say that two outputs are similar if:

- They co-occur more frequently than expected: We assume that outputs that are geographically co-located are similar since they share the same conditions - from weather to skill-level of farmer to access to market. Note that this takes into account a number of unobservable characteristics but is still a narrow definition of similarity. Two farmers might be similar in all relevant characteristics and inputs but may have chosen to produce different crops due to historical reasons (family always grew corn) or lack of knowledge (was not aware that she could grow corn).
- The vectors of inputs with which they co-occur are similar: Outputs that require similar vectors of observable (and recorded) inputs are similar. This is a less strict definition and can show similarities between products that are not necessarily grown together at the farm level.

To create these two similarity matrices, we must first create a matrix of co-occurrences, convert these to indexed values to adjust for ubiquity of certain inputs and outputs, and use this matrix to generate the similarity matrices. The following sections detail these steps.

4.1 Creating a matrix of co-occurrences

Since M is a matrix of ones and zeroes, we can create a co-occurrence matrix C with elements c_{ij} as follows:

$$C = M^T M \tag{14}$$

c_{ij} is the number of times i and j co-occur at a farm. Note that i and j can be either input or output. Figure 7 shows this matrix. The top left quadrant of the C matrix is the co-occurrence of inputs with other inputs, where we observe the greatest density. The rest of the matrix appears to have relatively very small numbers since very few outputs are created at a farm[12].

[12]This is also an artifact of using presence dummies for inputs while using discrete yield indices, which is a lot more sparse, for output. The next few sub-sections attempt to correct for this difference in ubiquity.

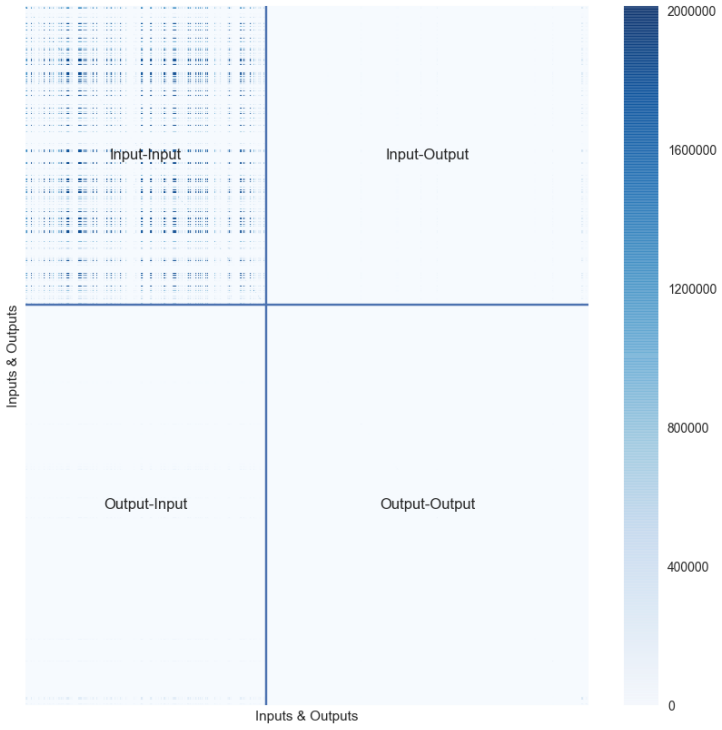

Figure 7: The C matrix - a matrix of count of co-occurrences

4.2 From co-occurrences to index measures

Some items occur more frequently than others. An input, say water, occurring with an output is not significant if it is highly ubiquitous and co-occurs with almost all outputs. To adjust for this, we generate a Q matrix of indexed values. Specifically, we do the following operations on the C matrix:

- Calculate indexed values: C is a square matrix. To adjust for the ubiquity of certain inputs and outputs, we convert the values to indices as follows:

$$c'_{ij} = \frac{c_{ij}}{c_{ii} \cdot c_{jj}} \cdot N \tag{15}$$

where N is the total number of farms. Note that c_{ii} and c_{jj} are diagonal terms that provide the occurrence of i and j in the data. How do we interpret c'_{ij}? If i and j were independant, then the probability of i and j co-occurring would be $Pr(i) \times Pr(j)$ which is $\frac{c_{ii}}{N} \times \frac{c_{jj}}{N}$. But the data gives the probability of i and j co-occurring or $Pr(i,j)$ as $\frac{c_{ij}}{N}$. The ratio of these gives us $\frac{Pr(i,j)}{Pr(i) \times Pr(j)}$ or c'_{ij} as per equation 4.2. Therefore c'_{ij} is a measure of association between i and j, taking into account their

probability of occurence. If c'_{ij} is greater than 1 then i and j are positively associated and co-occur more frequently than we would expect if they were independant. Similarly, if c'_{ij} is less than 1 then i and j are negatively associated.

- Drop items that never occur: Some outputs are not produced by any farms or produced by just one farm. We drop these.

- Take 'modlog' of the indexed value: This c'_{ij} has a long tail and has a log-normal distribution as seen in Figure 8. To adjust for this, we convert these to log terms. There are two major problems with taking logs. First, an output (or input) only co-occurs with very few other outputs (or inputs). Therefore C would have a number of zeroes, for which log is undefined. Second, log of values less than 1 would be negative which maybe undesirable when dealing with matrices.

  In order to correct for these, we do 'modlog' transformation as follows:

$$modlog(c'_{ij}) = \begin{cases} 0 & \text{if } c'_{ij} = 0 \\ \frac{log(c'_{ij}) + m}{m} & \text{if } c'_{ij} > 0 \end{cases} \tag{16}$$

  where:

$$m = \frac{log(c'^{min}_{ij})}{c'^{min}_{ij} - 1}$$
$$c'^{min}_{ij} = min(c'_{ij}) \qquad \text{for } c'_{ij} > 0$$

  More details on this transformation can be found in the companion report, "How Industry-Related Capabilities Affect Export Possibilities".

  Plot B in Figure 8 shows the distribution of the Q matrix with element q_{ij} resulting from taking the modlog of c'_{ij}. Note that, though not perfect, it has more of a normal distribution.

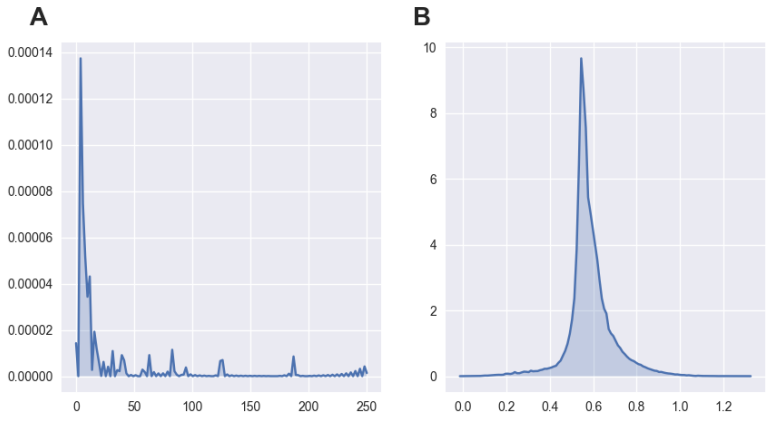

Figure 8: Taking 'modlog' - Distribution of (A) $c'_{ij}i$ (B) q_{ij}

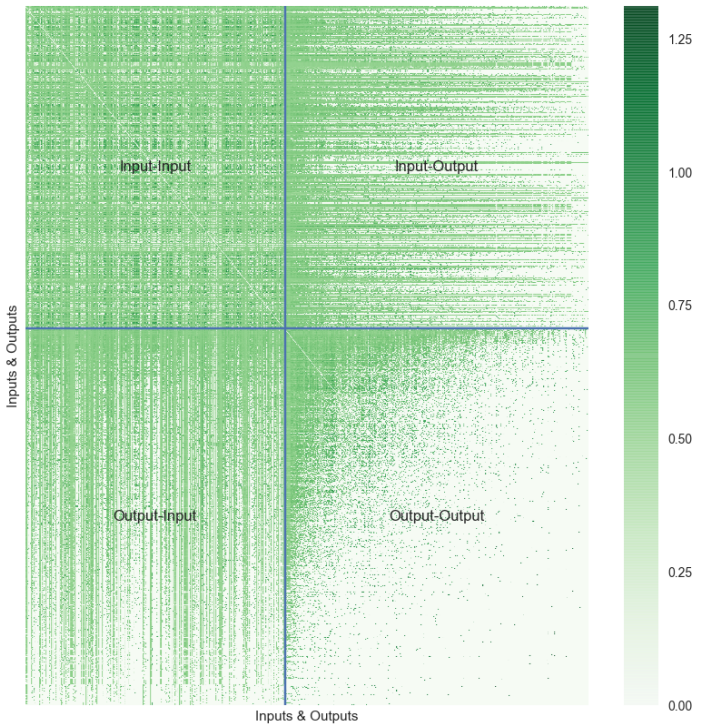

Figure 9: The Q matrix - Index transform of C matrix. Matrix is sorted by row and columns sums.

The resulting Q matrix is shown in Figure 9.

Comparing Q to C, we note that though the matrix continues to be sparse, a lot more significant associations between outputs and outputs, and outputs and inputs are now identified.

4.3 The output similarity matrices

From the bottom right quadrant of matrix Q, we can directly extract a relationship between outputs. We call this matrix the OO matrix. Since two outputs rarely occur together, (1) this may be a noisy signal and (2) it may miss relationships between outputs that do not co-occur in the data.

We can generate another relationship between outputs by considering the similarity between two row-vectors of the OI matrix in the bottom left quadrant of the Q matrix. It is implied that similar outputs would require similar inputs. We construct this similarity matrix using pairwise Pearson correlations[13] between the rows of the OI matrix. We call this the OO' matrix with elements oo'_{ij}:

[13] We leave exploration of other measures of similarity for future work. We may be able to use higher order kernels to construct better measures.

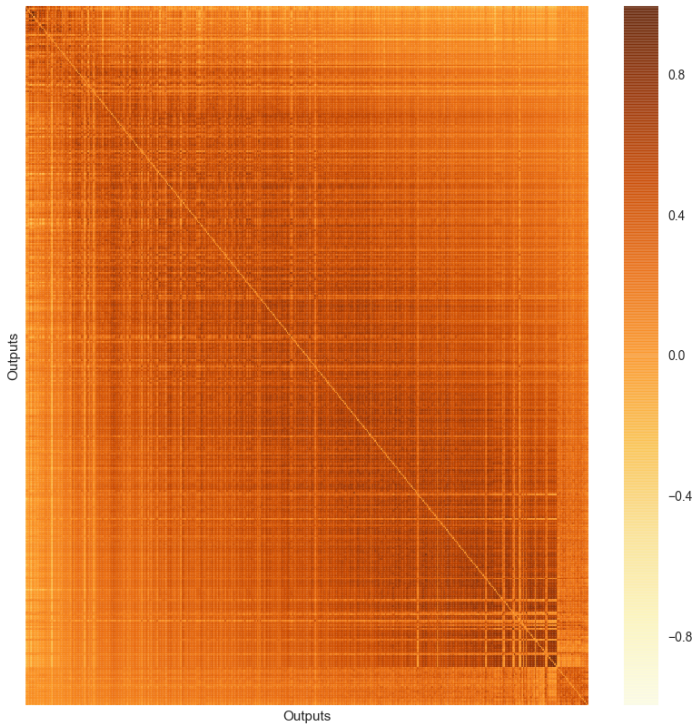

Figure 10: The OO' matrix - from correlations of rows of OI matrix.

$$oo'_{ij} = \begin{cases} corr(\vec{oi}_i, \vec{oi}_j) & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \tag{17}$$

where $\vec{oi}_i$ is row i of matrix OI[14]. The OO' matrix is shown in Figure 10. How similar are OO' and OO? In section 4.5, we conduct a more detailed analysis showing that they explain different components of the variation in outputs. The Pearson correlation coefficient of these matrices is -0.02, suggesting that they carry very different signals.

We now have two product spaces; a direct one (OO), based on product co-occurence, and an indirect one (OO'), based on similarity of inputs with which they co-occur.

The inputs can be broken into two groups: machinery and labour inputs, and other inputs. The first group includes inputs such labour, and various types of equipment and machinery and can be seen as rival goods since their use in cultivating an output would mean that other production processes are unable to use

[14]It is important to set the diagonal to zero so that the density measure we construct in section 4.5 does not take presence of item itself into account.

it. The second group includes inputs such as water source, soil protection practices, or waste management practices. Use of these for the production of one output does not preclude them for being used elsewhere. Table 8 shows how the various types of inputs were classified[15].

We construct two additional matrices, OO'_r and OO'_{nr} which are constructed as in equation 17 but using just machinery and labour inputs, and other inputs respectively.

4.4 Clusters of output

Can we discover groups of similar outputs based on the OI matrix? Since OO' is based on the indices matrix Q, it takes into account the ubiquity of the output itself. Therefore, before we can cluster outputs, we construct an alternative matrix, P, that looks at the vector of inputs occurring *with one unit of output*:

$$p_{ij} = c_{ij}/c_{ii} \tag{18}$$

Note that the diagonal for this matrix is the number of farms where these inputs and outputs co-occur[16]. By dividing a row by the value of the diagonal, we get the probability of co-occurrence i.e. $P(j|i) = p_{ij}$, can be interpreted as probability of j occurring when i occurs. Therefore each output/input is now represented as a vector of probabilities of outputs/inputs co-occurring. Figure 11 shows the P matrix. The OI matrix in the bottom left quadrant of P matrix gives an estimate of the conditional probability of requiring input j when producing output i.

We use the OI_p matrix in the bottom left quadrant of P to cluster output based on Euclidean distance[17] between probability vectors of inputs.

4.4.1 The clustering algorithm

We use K-means and Affinity Propagation (AP) to create the clusters. For K-means it is necessary to specify the number of desired clusters. We algorithmically determine these by maximizing the silhouette score.

4.4.2 Principal Component Analysis (PCA)

Clustering at high dimensions leads to unexpected behaviour due to '*the curse of dimensionality*'. One way around this is to take the top n principal components of the data. Figure 12 shows the PCA components and

[15]These groups are meant to represent rival and non-rival goods. Though empirical results in the next section appear to support this grouping, it can be improved upon by analyzing each input separately. For example, we may find that 'water' and 'credit' are infact rival goods and should belong with machinery and labour. Similarly, it can be argued that capital and labour goods are not fully utilized and are hence not rival goods. Though we have not done so in this report, the exercise could easily be repeated with the inputs split differently.

[16]We use 'co-occur' loosely here. Since we use indexed values, we really mean the number of farms where input usage or output yield is higher than expected.

[17]We leave exploration of other distance metrics for future work. Differing critical inputs like soil should make two outputs very 'far' from each other. This logic is not captured through Euclidean distance.

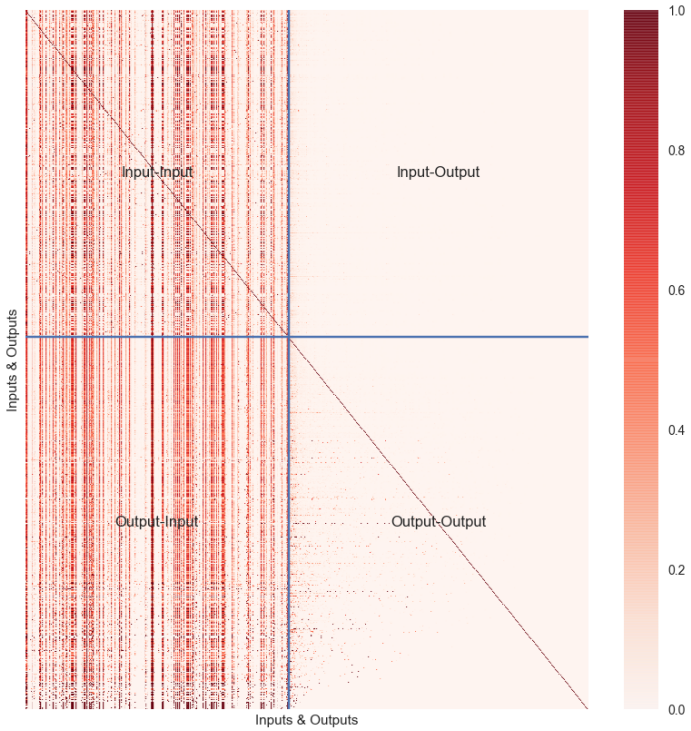

Figure 11: The P matrix - normalizing rows by the diagonal.

the variation explained by the first 40[18]. We chose n to be 20 which explains around 80% of the variation.

4.4.3 Setting preferences manually

We need to specify the preference vector for affinity propagation. We would like the larger or more common outputs to have a high preference to be exemplars. Therefore we use ubiquity[19] of outputs to create the preference vector.

4.4.4 The clusters

The clusters produced can be seen in Figures 13 and 14. These plots use the first two principal components as axes.

[18]The maximum components is equal to the total number of inputs. Having greater than 40 components adds very little to the explained variation, and these components have been omitted to make the figure easier to read

[19]Actually, square root of ubiquity. Ubiquity has a very large tail. Taking a square root reduces the skew.

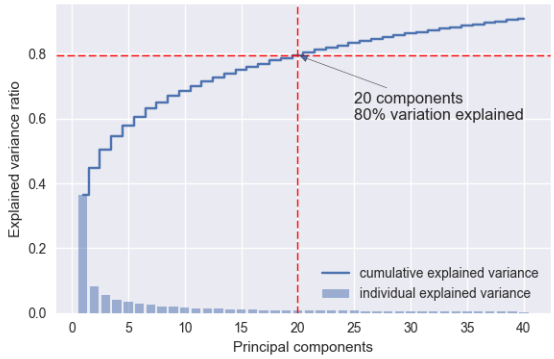

Figure 12: PCA components and explained variation

4.4.5 Discussion

The K-means clustering with silhouette score creates 2 clusters. These are quite coarse and no clear categories are apparent.

The Affinity Propogation algorithm on the other hand creates 4 clusters as shown in Figure 13. Again, the classification into clusters does not neatly delineate products. The following list shows some of the categories of crops in each cluster:

- Cluster 1: root vegetables (potatoes, carrots, onions), fruits (plum, passion fruit, coconut), coffee (6 varieties)
- Cluster 2: coffee (20 varieties), herbs (clover, cowpea, parsley)
- Cluster 3: coffee (8 varieties), grains(rice, maize, sorgum, soy), avocado , banana, mango.
- Cluster 4: coffee (6 varieties), flowers (azalea, snapdragon, miniclavel, dahlia, fressia)

This type of clustering is rarely perfect[20]. They are sensitive to the algorithm, the choice of metric to determine number of clusters, and other initialization parameters. A number of unsupervised learning techniques can be used to improve this clustering and is left for future work.

4.5 Density matrices

In section 4.3, we constructed similarity matrices OO, OO', OO'_r, and OO'_{nr}. Next, we explore the predictive power of these matrices. Specifically, we wish to know if they are individually and jointly predictive of output.

[20]This can be seen by the fact that though most coffee varieties are in cluster 2, some end up in other clusters as well.

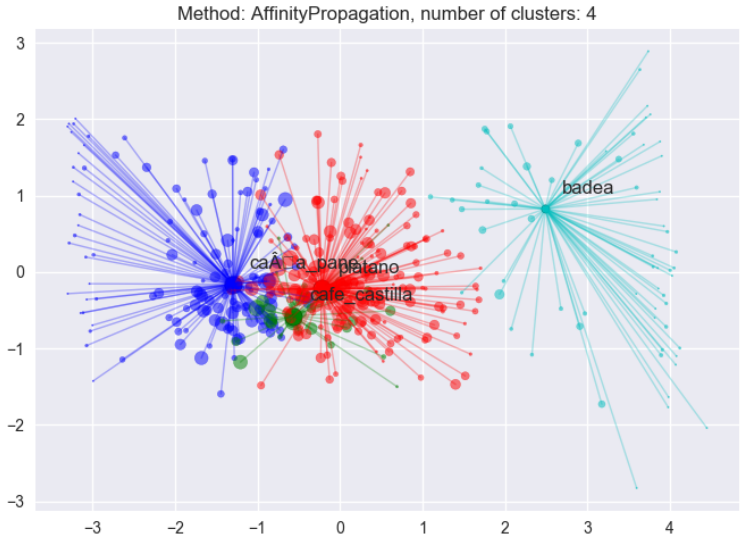

Figure 13: Clustering Outputs using OI_p using Affinity Propagation

Farms tend to specialize and produce just one product. This means that the density for the outputs would be very low for most farms. Therefore, we test our predictions at the municipality level where the output space is more vibrant. In addition, this allows us to control for a number of municipality level characteristics in order to test the robustness of our coefficients. We construct four density measures as follows:

$$\begin{aligned} D &= (O_m) \cdot OO \\ D' &= (O_m) \cdot OO' \\ D'_r &= (O_m) \cdot OO'_r \\ D'_{nr} &= (O_m) \cdot OO'_{nr} \end{aligned} \tag{19}$$

Where O_m is the matrix of output yield indices at the municipality level. Note that the D, D', D'_r, and D'_{nr} are all M×P matrices, where M is the number of municipalities and P is the number of outputs. The diagonals for OO', OO, OO'_r, and OO'_{nr} are zero, therefore element d_{ij} in matrix D is a measure of density around output j for municipality i but excludes the presence of output j itself. Density around an output j would be higher for municipalities producing outputs that tend to co-occur with j within farms. Similarly, d'_{ij}, would be larger for municipalities that produce outputs that require similar inputs to those required by output j. d_{ij} and d'_{ij} can be interpreted as matrices of *implied comparative advantage* i.e. the higher the number, the better the municipality i would be at producing j. If our assumptions in section 4 are indeed valid, we would expect D and D' to be predictive of the output. Specifically, we would expect β_1 and β_2 to

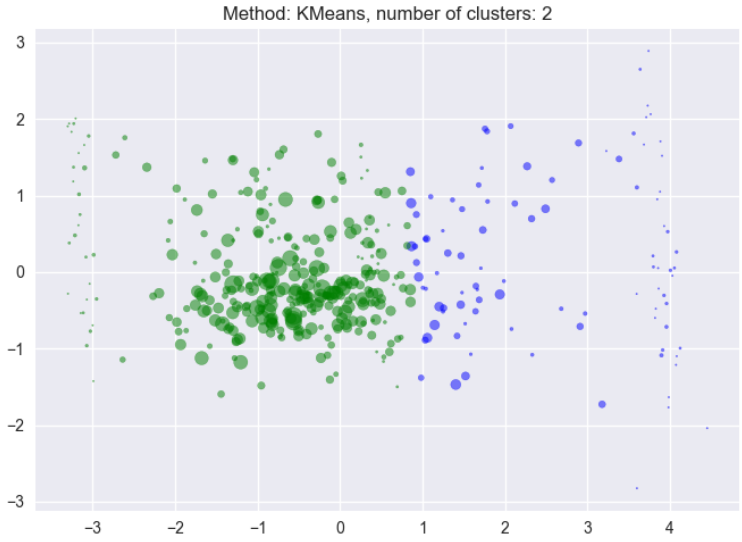

Figure 14: Clustering Outputs using OI_p using K-means

be positive and significant in the following regression:

$$y_{i,p} = \beta_0 + \beta_1 \cdot d_{i,p} + \beta_2 \cdot d'_{i,p} + \gamma_i + \delta_p + \varepsilon_{i,p} \tag{20}$$

where $d_{i,p}$ and $d'_{i,p}$ are elements (i,p) for D and D' respectively. γ_i is the municipality fixed effect and δ_p is output fixed effect.

Futher, we would also expect D'_r and D'_{nr} to be predictive of output but with opposite effects. A municipality i and product j with high density, $d'_{r,ij}$ would imply that there are a number of outputs in i that use similar *machinery and labour* inputs which tend to be *rival* goods. Therefore, we would expect this to have negative effect on the ability of i to produce j. Specifically, we would expect the α_2 to be negative but α_3 to be positive in the following specification:

$$y_{i,p} = \alpha_0 + \alpha_1 \cdot d_{i,p} + \alpha_2 \cdot d_{r,ip} + \alpha_3 \cdot d'_{nr,ip} + \gamma_i + \delta_p + \varepsilon_{ip} \tag{21}$$

We run various specifications, including some municipality level characteristics, that test the predictive power of D', D, D'_r, D'_{nr}. In all the specifications reported, we cluster standard errors within products and within municipalities. Further, we look at how predicitive our density measures are for agricultural crops and animals separately. Tables 12 and 13 show these results.

4.5.1 Densities and land usage

We also explore if the second outcome variable, land usage, can be predicted using these matrices. We run specifications analagous to 20 and 21:

$$l_{i,p} = \beta_0 + \beta_1 \cdot d_{i,p} + \beta_2 \cdot d'_{i,p} + \gamma_i + \delta_p + \varepsilon_{i,p} \tag{22}$$

and

$$l_{i,p} = \alpha_0 + \alpha_1 \cdot d_{ip} + \alpha_2 \cdot d_{r,ip} + \alpha_3 \cdot d'_{nr,ip} + \gamma_i + \delta_p + \varepsilon_{ip} \tag{23}$$

Again, we would expect the coefficient to be positive for all densities except D'_r since a similar logic applies. A municipality i is less able to grow j if $d'_{r,ij}$ is high since there would be competition for machinery and labour inputs.

4.5.2 Presence and high yield

Finally, we see if the density measures explain two factors: (1) if a crop is produced or not at all in the municipality (eq. 24), and, given that the crop is produced, if it is produced at high yield (eq. 25).

$$y^{exist}_{i,p} = \alpha_0 + \alpha_1 \cdot d_{i,p} + \alpha_2 \cdot d_{r,ip} + \alpha_3 \cdot d'_{nr,ip} + \gamma_i + \delta_p + \varepsilon_{ip} \tag{24}$$

$$y^{high}_{i,p} = \alpha_0 + \alpha_1 \cdot d_{i,p} + \alpha_2 \cdot d_{r,ip} + \alpha_3 \cdot d'_{nr,ip} + \gamma_i + \delta_p + \varepsilon_{ip} \tag{25}$$

The results for these are shown in Table 15 and 16 respectively.

4.6 Interpreting the results

We consider the two outcome variables separately.

4.6.1 Yield

First, we consider the results in table 12. D is significant and positive under all specifications. Further, the coefficient is stable in the presence of additional controls and fixed effects. This provides strong evidence for our first assumption that crops that grown on the same farm are indeed similar. Note that D', which captures the availability of inputs, is a significant and *negative* predictor of output once municipality and product fixed effects are included (specification 4). Most of its predictive power can be explained by the product and municipality fixed effects. Since D' was constructed using all inputs, it may well be that the

opposing effects of machinery and labour, and other goods are conflated. On the other hand, D is significant and positive under all specifications. Futher, the coefficient is stable in the presence of additional controls and fixed effects. This provides strong evidence for our first assumption that crops that grown on the same farm are indeed similar.

We also note that D'_r and D'_{nr} are indeed predictors of yield. Their coefficients are significant in most specifications with the sign as expected, D'_r is negative while D'_{nr} is positive in all specifications except[21] (6). Also note, that the magnitude of the coefficient for D'_{nr} in almost all specifications is greater than D'_r. From this we infer that, the *diversification effect* of having similar outputs that use the similar inputs is greater than the *specialization effect* of having similar outputs that use similar machinery and labour.

Adjusted yield index, which is an indicator of how productive the other crops in the municipality are, has a negative and significant coefficient in all specifications. This would suggest that if a municipality already grows crops at high yield, it is less likely to diversify into other crops[22]. Since soil is a scarce and possibly fully utilized resource, this is what we would expect. If the crops currently utilizing the available soil provide high yields and economies of scale exist, there would be little reason to produce a different crop.

Finally, we see if these results hold true when controlling for the output being a crop or an animal. Table 13 shows these regressions. Overall the results are less stable across specification. Note that in specificatoin 3, where we run a simple OLS for just animals, the coefficient on D is negative. Specification 7 repeats this using fixed effects for municipality and product and the coefficient on D is again negative. This would suggest that the presense of outputs that animals co-occur with has a negative effect of yield: crops and animals compete. It may be that they require very different skill sets to produce at high yield and therefore specialization is natural and maybe necessary. Other specifications include both crops and animals but include a dummy and interaction terms. Since crops account of a majority of the observations, the results are mainly driven them and are similar to regressions in previous sections.

4.6.2 Land usage

We consider the second outcome of interest, land usage index, as per specifications 22 and 23. Table 14 shows these results. Note that the sign and significance for D is the same as with *yield* and the same conclusions can be drawn[23]. For D'_{nr} and D'_r the results are less robust. Under certain specifications, the coefficients are not significant[24] though their signs are in the correct direction.

[21] In all the regressions, we cluster standard errors at both product and municipality level to provide the strongest possible specifications.

[22] More accurately, less likely to produce other crops or produce them at high yield.

[23] *yield* and *land usage* are closely linked. The correlation between them is 0.78. It may not be surprising that in a place where a crop provides high yield, a larger portion of the soil is dedicated to its cultivation. Identifying where this is not true may be an interesting extension for futher study.

[24] The coefficients are significant if we do not cluster standard errors at both municipality or product level.

4.6.3 Presence and high yields

A slight variation on looking at "high yield" as we did in 4.6.1 is to look at *presense* of an output. Here we test if the densities can explain what is grown in a municipality. We classify ‘grown’ as the presense of one or more farms producing an output. We do not differentiate if it is produced at high yield or low yield. Table 15 shows these results. Note that these are very similar to the results we found in section 4.6.1. This prompts a follow-up question; of the crops that are produces, can the density metrics explain high yields. Table 16 shows these results. A few of the conclusions remain unchanged. D, $D_n r$ are both positive and significant predictors of high yield. The coefficient on adjusted yield is still negative and significant. The major difference is that under some specification, D_r is not significant. This presents an interesting hypothesis - competition for inputs may deter entry into new products but it does not reduce the likelihood of producing at high yield. Once the crop is already being grown, access to machinery and labour inputs has already been negotiated and causes little negative effect on yield.

4.7 Conclusion

In this section, our goal was to understand the mechanism of diversification. We constructed multiple measures of similarities between outputs. First, based on co-occurence of outputs at the farm level. Second, based on the similarity of machinery and labour inputs used. Third, based on the similarity of other, non-machinery and labour, inputs used. We discovered that municipalities do indeed diversify into similar goods though the definition of ‘similar’ is important. The presence of outputs that compete for machinery and labour puts a downward pressure on diversification though only when product is not currently being produced. We also discovered that if a municipality is already growing other crops at high yield, it is less likely to diversify. Controlling for all these forces, presense of outputs that use similar non-machinery inputs increases the likelihood of diversification. These forces differ in magnitude with the net effect that diversification can be observed at the municipality level.

These results suggest that we should be able to predict which output a municipality should produce given its densities. In the next section, we use Machine Learning algorithms and methodologies to predict these. The flexibility of models and the robustness of predictions, offered by these algorithms allow us to accurately predict and therefore recommend municipality-output pairs.

5 Machine Learning Methods

In the previous section, we constructed multiple measures of similarity. One, OO, was based on co-occurence of agricultural outputs at the farm level and the other, OO', was based on correlation between the vector of co-occuring inputs. We also constructed two additional similarity matrices, OO'_r and OO'_{nr} using correlation between vector of co-occuring *machinery and labour inputs* and *other non-machinery and labour* inputs measures respectively. While OO' was not found to be correlated with our outcome measures, all other similarity matrices were.

After having explored the mechanisms of diversification, we switch to the task of recommendation or prediction. Using the density measure we have created, we predict the presense of outputs in municipalities and use our models to recommend outputs that are currently "missing".

We utilize various Machine Learning (ML) algorithms to predict the two outcome measures. We utilize the densities created in the previous section in addition to the input indices and other disaggregated measure as predictors. Most of our predictions are done at the municipality level - partly due to computational considerations, and partly to allow us to use predicted non-agricultural outputs and soil vocation data as additional inputs.

5.1 Why machine learning?

Traditional regression methods such as generalized least squares or maximum likelihood estimators fail us in two major ways.

- Over-fitting: For each output, we have around 1000 municipalities or observations and around 318 possible inputs and another 268 non-agricultural activities. Since the ratio of observations to explanatory variables is low, traditional methods such as ordinary least squares could lead to "over-fitting"; they would model the given data very closely but fail for any out of sample data.
- Functional form: Traditional regression methods require some prior knowledge of the functional form of the relationship between agricultural inputs and outputs. Are some inputs substitutes while others are complements? Are certain inputs critical for high yield of a crop? We have no such prior though we suspect that some interaction effects between inputs exist. Exploring this infinite space of possible interaction effects is not trivial.

ML gets around the problem of over-fitting by using regularizers[25]. These penalize complexity and converge on simpler models that are less likely to overfit. Other techniques like cross-validation and grid-search allow us to tune models that accurately describe the data in general and not just the dataset in hand.

[25] This adds a regularization parameter that needs to be tuned. Techniques, such as k-fold cross-validation, exist for this purpose.

Some machine learning algorithms like decision trees[26] and support vector machines[27] do not require a functional form to be specified.

In general, tuned ML have been shown to perform better than traditional econometric techniques at prediction[28]. The downside is that ML models can be difficult to unpack or interpret. The coefficients returned cannot be seen as correlations in the traditional econometric sense. This is made even worse by ensemble methods. Therefore ML algorithms provide strong predictive power but weak explanatory power.

5.2 Machine learning techniques

As part of this project, we explored a number of algorithms. In this report, we present the results from only the best performing algorithms. Some of the algorithms used in this paper are:‘

- LASSO: LASSO adds a *regularizer* term to the standard least squared models.

$$\min_{\beta} \left\{ \frac{1}{N} \|y - X\beta\|^2 + \lambda \|\beta\| \right\} \tag{26}$$

  where $\lambda\|\beta\|$ is the L1 regularizer which selects only the subset of the independent variables (X) that have high explanatory power. The model is useful if we believe the true model to be sparse i.e only a few of the independent variables explain the dependant variable.

- Ridge: Instead if we believe the true model to be dense i.e almost all independent variables explain the dependant variables, we may choose the Ridge L2 regularizer instead:

$$\min_{\beta} \left\{ \frac{1}{N} \|y - X\beta\|^2 + \lambda \beta^2 \right\} \tag{27}$$

- Random Forest: We can also train a decision tree to predict probability of exports. But deep trees i.e. with a lot of independent variables, tend to overfit the training set. Random Forests (RF) overcomes this with bootstrapping. It creates a number of decision trees, each trained on a bootstrapped random sample, and averages their predictions.

- Support Vector Machines: Given a training set, $(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n)$, where y_i can be -1 or 1, linear Support Vector Machines (SVM) find the hyperplane that maximally divides the groups of points for which $y_i = 1$ and from those with $y_i = -1$. Figure 15 shows an example of one such hyperplane. By utilizing higher order and non-polynomial kernels, we are able to fit a hyperplane in a tranformed (usually higher order) feature space.

[26] It can be argued that decision trees impose their own functional form. This is true but the hierarchical nature allows for a lot more flexibility. The use of ensemble methods like random forests provide even greater flexibility.

[27] A choice of a kernel can be seen as choosing a functional form, but this can also be selected using grid-search or similar parameter search techniques.

[28] *Prediction* is accurately identifying $\hat{y}$ whereas *estimation* is accurately estimating the $\hat{\beta}$. ML algorithms sacrifice un-biasedness for low variance.

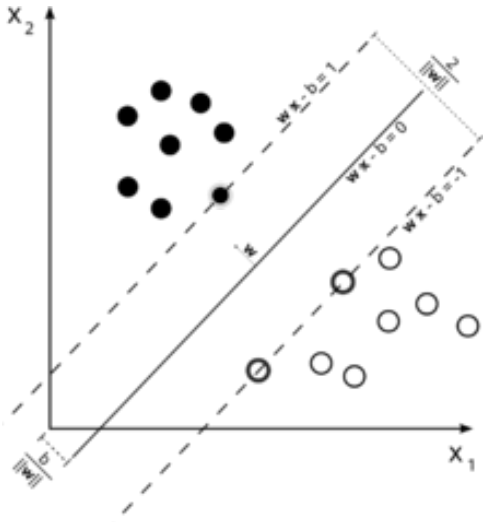

Figure 15: An example of a linear hyperplane for SVM. Courtesy: wikimedia/Public

- Gradient Boosted Trees: Gradient Boosted Trees (GBT), like RF, is another ensemble method that combines multiple decision trees. But unlike Random Forests, GBT builds the model in stages by fitting a tress to the *pseudo-residuals*. XGBoost is a popular implementation of GBT that we utilize in this paper.

5.3 Defining metrics

As we noted in section 2, most municipalities produce only a handful of products with high yield. Therefore, our dataset contains significantly more cases of municipalities not producing an output than producing it. If we use a simple measure such as percentage correct, an algorithm that classifies all output as not being produced will appear do very well. As an example, if only 5 out of 100 municipalities produce product A, an algorithm that simply says that no municipality will produce product A would have an accuracy of 95%. Instead, we need to look at the *types* of errors made by the algorithm to understand performance.

A classification algorithm can make two types of errors:

- False positive (FP): Predicts that an output should be produced when it isn't.
- False negative (FN): Predicts that an output will not be produced when it is.

For completeness, we also define the following:

- True positive (TP): Correctly predicts that an output is produced.
- True negative (TN): Correctly predicts that an output is not produced.

One way to express the results from the classification is as a **confusion matrix** as shown in Figure 16. Note that this allows you to see types of errors made and calculate metrics such as sensitivity and specificity, or precision and recall.

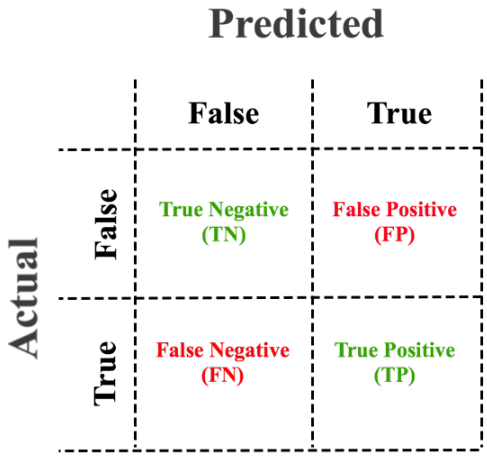

Figure 16: A confusion matrix

There is often a trade-off between these errors. We may be able to increase the true positive rate at the expense of a higher number of false positives. This relationship is captured in the **receiver operating characteristic (ROC) curve**. At each point of the curve, a new confusion matrix can be created. Another metric for the performance of the classifier is the **area under the curve (AUC)** of the ROC. A perfect classifier that makes no errors would have an AUC of 1 while a classifier that randomly guesses would have an AUC of 0.5.

5.4 Results

We predict three outcome measures and train multiple models based on different algorithms.

5.4.1 Predicting yield

The specification for the predicting yield is:

$$\hat{o}_{mp} = \hat{f}((d)_{mp}, (d')_{mp}, (d'_r)_{mp}, (d'_{nr})_{mp}, I_m, a_{mp}, \hat{X}_m, S_m, \alpha_m, \gamma_p) \tag{28}$$

where:

$$I_m : \text{Vector of inputs in municipality } m$$
$$a_{mp} : \text{Adjusted yield index}$$
$$\hat{X}_m : \text{Predicted non-agric activities}$$
$$S_m : \text{Soil vocation indexed data}$$
$$\alpha_m : \text{Municipality dummies}$$
$$\gamma_p : \text{Product dummies}$$

And the d variables are the four densities we constructed in the previous section.

We consider three classes of outcomes when predicting yield. Therefor o_{mp} can take three possible values:

- Class 2 - High yield: Output is produced with a high yield index at the municipality.
- Class 1 - Low yield: Output is produced with a low yield index at the municipality.
- Class 0 - Not produced: Output is not produced at all at the municipality.

All the results presented are 'one versus the rest' i.e. they show performance of the model predicting the outcome of interest vs. the other two outcomes. We add an additional case, 'High vs. low' where we only consider municipality-crop pairs where the crop is actually produced and predict if the crop will be produced at high yield or low yield. The ROC/AUC for these is shown in Figure 17.

Note that the model does remarkably well when predicting if a crop will be grown or not, with an AUC of over 0.9, but finds it difficult to differentiate between low-yield and high-yield. Therefore we can identify products a place should be producing with high accuracy but, given that it is produced, fail to predict if it will be produced at high yield or low yield accurately. Do we fail to differentiate between low-yield and high-yield for all products or just a few? Figure 18 shows the results from training the model for each product. Though this yields a lower performance than the model that includes all products, it gives us an understanding of the performance by product.

Figure 18 shows the products where average AUC from 10 simulations is greater than 0.85. 88 products meet this threshold and we can have high confidence when predicting high yield for these products.

5.4.2 Predicting land usage

The setup for predicting land usage is very similar to the previous section. The specification is as follows:

$$\hat{o}_{mp}^{soil} = \hat{f}((d)_{mp}, (d')_{mp}, (d'_r)_{mp}, (d'_{nr})_{mp}, i_m, a_{mp}, \hat{X}_m, S_m, \alpha_m, \gamma_p) \tag{29}$$

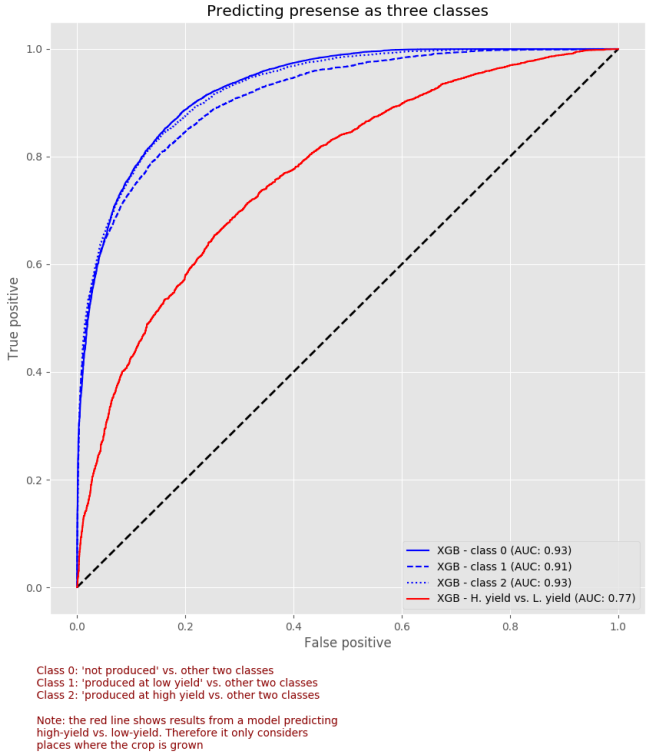

Figure 17: Predicting three classes of yield

where:

I_m : Vector of inputs in municipality m

a_{mp} : Adjusted yield index

$\hat{X}_m$: Predicted non-agric activities

S_m : Soil vocation indexed data

α_m : Municipality dummies

γ_p : Product dummies

Recall that o_{mp}^{soil} is the intensity with which product p uses the appropriate soil type in municipality m. As before, we predict the three classes as 'one versus the rest' and an additional prediction where the model predicts high versus low land use for the crop in municipalities where it is grown. Figure 19 shows the ROC/AUC for the four predictions. It is interesting to note that though the performance in the 'one versus the rest' predictions is poorer that when predicting yield, it is higher when differentiating between high-intensity and low-intensity. It may not be a surprising result that predicting land use is a more difficult problem than predicting yield. Crop choice may be influenced by factors outside of our dataset, such as identity ("I'm a wheat farmer!"), knowledge ("I know mangoes but not coffee"), and competition ("Jack is

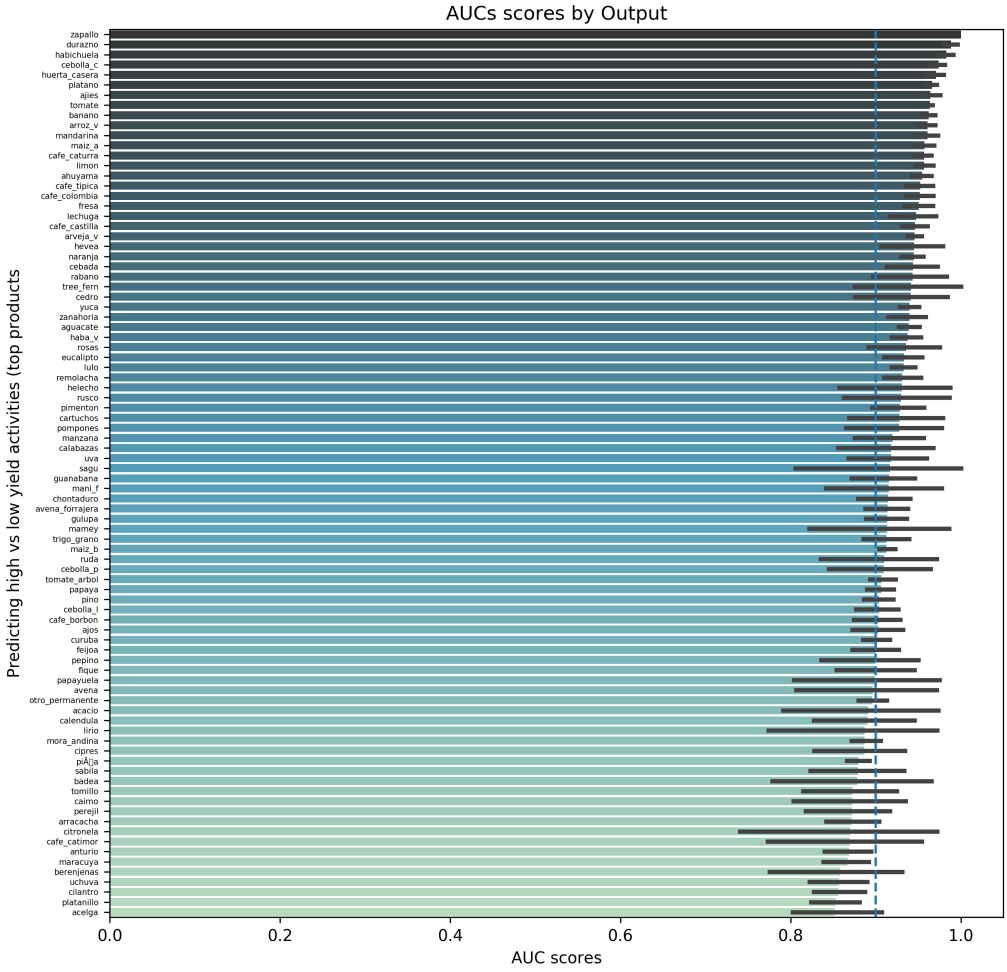

Figure 18: Predicting yield by products

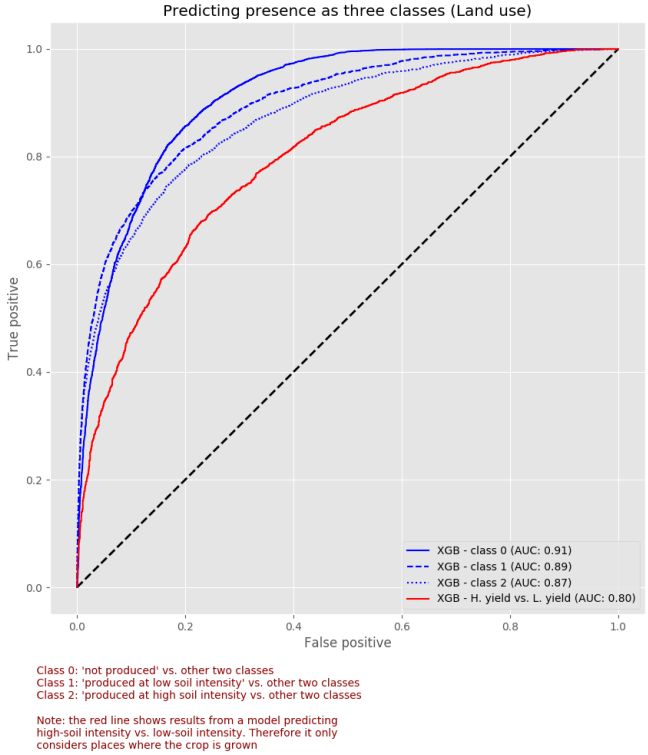

Figure 19: Predicting three classes of land use

already growing oranges").

As in the previous section, we train models for each crop. Given that a crop is grown, we look at how well we are able to predict if the intensity of land use for it will be high or low. The results are show in **??**.

5.4.3 Predicting non-agricultural activities

The final specification is ifor predicting the intensity of non-agricultural activities.

The specification for the prediction now includes output as features. As mentioned earlier, the non-agricultural activities present may be determined by the agricultural activites in the area. For example, in an area with a number of rice farms, we would expect to see a rice mill[29]. Since this might be particularly true at the vereda level, we construct the metric and conduct the prediction at this level. In addition, the dataset also includes industry employment RCAs (constructed using PILA data) and the vector of inputs available.

$$\hat{o}_{vp} = \hat{f}((d)_{vp}, (d')_{vp}, (d'_r)_{vp}, (d'_{nr})_{vp}, I_v, N_m, O_v, \alpha_m, \gamma_p) \tag{30}$$

[29]In the previous section, we did not use non-agricultural activities in the density regression as we expected them to have high explanatory power. Here, instead of seeking a deeper understanding of the mechanisms, we are solely concerned with prediction and hence are able to use outputs

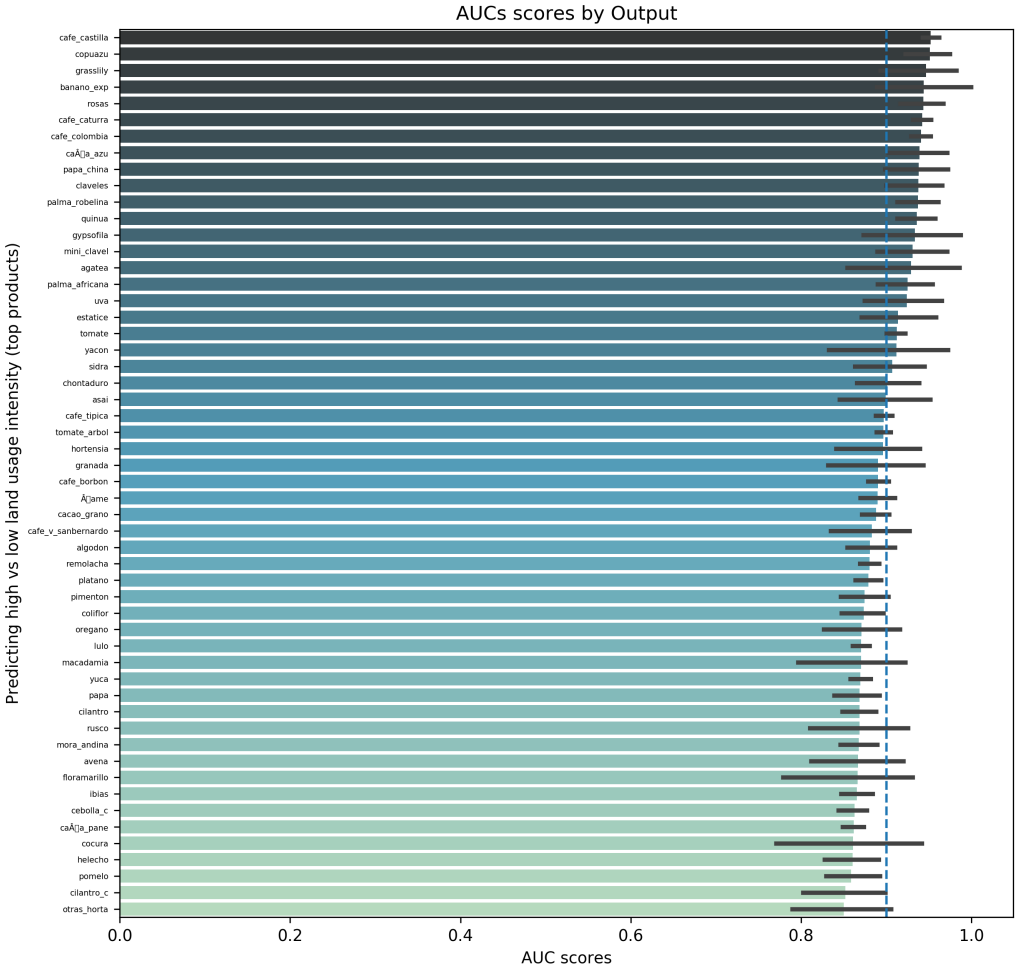

Figure 20: Predicting land usage intensity by products

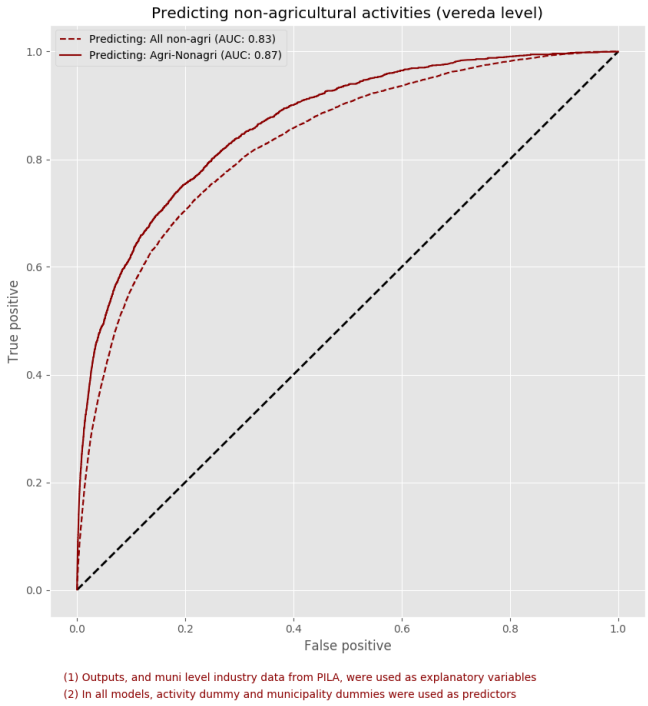

Figure 21: Predicting non-agricultural output

where:

I_v : Vector of inputs in municipality m

N_m : Vector of industry employment in municipality

O_v : Vector if output indices in the vereda

α_m : Municipality dummies

γ_p : Product dummies

The results are shown in Figure 21. We create a model for just non-agricultural activities associated with raw processing of agricultural outputs and repeat it for all non-agricultural activites. As with predicting high versus low yield, though the performance overall is not extermely high, there is a large variation in the prediction performance by the non-agricultural activity. Figure 22 shows this variation.

5.5 Identifying missing products

We were able to train models with very high AUCs (at least for a few products). Using these, we are now able to identify municipality-output pairs that are "missing" i.e. our model predicts that municipality should be producing the output at high yield but it does not. This can be considered in two ways. We can look at

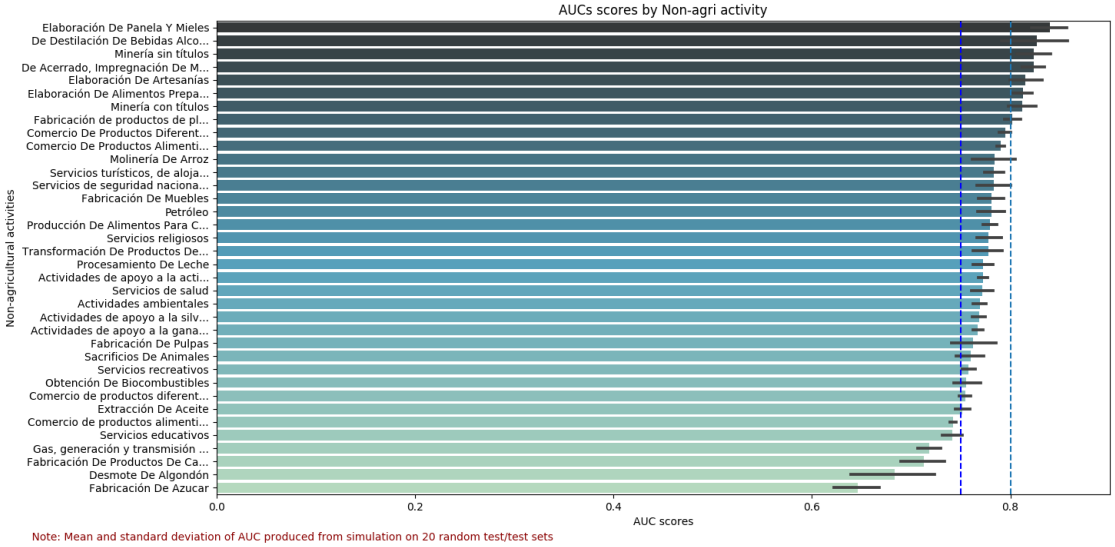

Figure 22: Predicting non-agricultural output by product

the top products a municipality should be producing, or identify the municipalities best suited to produce a product.

A similar "missing" analysis can be done for non-agricultural activites. These can be explored at:

- : *Missing products*: `https://goo.gl/dyxbVk`
- : *Missing non-agricultural activities*: `https://goo.gl/MPyzu7`

Note that some of the predictions of 'Missing products' includes product-municipality combinations where the appropriate climate is not present. Therefore, we refine the results further by filtering the predictions to show only those where the appropriate climate is present[30]. These results can be found at `https://goo.gl/tEZgK7`.

These predictions are not infallable. Even though the model produces probabilities, they cannot be directly interpreted as the model's confidence in its prediction. These visualizations should instead be used to prompt a deeper investigation into why they might be missing. More specifically, why they do not match the pattern found in other municipalities in Colombia. There may be perfectly good reasons, like the model does not contain a relevant variable, and future efforts could aim to correct these. This may also reveal some binding constraints that if loosened will allow for growth in these municipalities and products.

[30]Climate information is available in the soil vocation dataset.

5.6 A new product space

In the previous section, we created three measures of similarity between products. Densities created with these similarities were shown to be predictive of what crops a municipality might diversity into. We also show that allowing a flexible functional form using machine learning, we can combine these densities in non-linear ways to improve our predictive power. Therefore an accurate product space must be based on a non-linear combination of all the three types of similarities. Unfortunately, as also discussed in section 5.1, machine learning does not (easily) allow us to 'unpack' the coefficients to identify factors that make two products similar.

We attempted to cluster the products using their co-occurence but found little coherence. Another way, we may be able to create a product space may be to look at feature importance in a Random Forest model. Products that have similar vectors of feature importance would be similar. An even more sophisticated way might be to create embeddings in a neural network and generate similarity as the distance between two embedding vectors. These require a new line of investigation and we leave this for future work.

5.7 Conclusion

In this section, We use a number of ML techniques to predict three main outcomes: yield index of outputs, land usage at the municipality level, and non-agricultural activities at the vereda level. We show that there is heterogeneity in prediction performance by product and even when the model does not perform exceptionally well, it is able to do so for some of the products. We used these models to identify products and non-agricultural activities "missing" in municipalities and present them online to be explored.

Throughout this document, we suggest areas for future work and enhancements. One addition area of work may be feature creation. Feature creation is a critical part of machine learning, where context-specific knowledge can be used to create new features. This has been shown to improve performance of models significantly. We do some feature creation when we create adjusted yield and density measures but leave more detailed feature creation for may be possible. In addition, we leave exploration of other frameworks, and ensemble methods that combine various algorithms including those in this paper to create predictors to future work.

6 Tables

Bases	Type of data	Description
Base A	Water	Source, water protection practices, water use problems
	Soil	Soil protection practices
	Energy	Type of energy source
	Technical Assist	Type of technical assistance
	Credit	What credit was requested for
	Workers	Number of works
	Crop management	Use of fertilizers etc.
	Other	Waste management and natural resources
	Irrigation	Type of irrigation practice
	Plaque control	Type of plaque controls practices
Base B	Equipment (Y/N)	Dummies for 49 types of equipment
	Equipment (Count)	Number of each type of equipment
	Equipment new (Count)	Number of each type of equipment less than 5 yrs old
	Equipment old (Count)	Number of each type of equipment more than 5 yrs old
Base C	Non-agri production (Y/N)	E.g. Deforestation, Milk processing, sugar refinery
	Non-agri services (Y/N)	E.g. Education, health, religious services, mining
	Workers in non-agri prod (Count)	Count of workers in each activity
	Workers in non-agri services (Count)	Count of workers in each activity
Base D	Tons of production	For 484 products
	Land sown	For 484 products
	Land harvested	For 484 products
Base E	Count of animals	Cattle, pig, poultry etc.
Base F	Farmed fish harvest (count)	Number of harvests of 38 types of fish
	Farmed fish (count)	Count of fish harvested - 38 types
	Farmed fish (average weight)	Avergae weight of a fish harvested - 38 types
	Farmed fish (total weight)	Total weight of type of fish harvested - 38 types
Base G	Salt water fish caught (total weight)	92 types
	Fresh water fish caught	50 types
Base H	Poverty headcount	Number of male/female/students in poor households
	Quality of life indicators	better vs. worse life than 5 years ago

Table 2: Description of variables in dataset

row_total_rca	No.	%	%
0	269209	28.1	28.1
1	386262	40.3	68.4
2	200558	20.9	89.3
3	72339	7.5	96.9
4	22008	2.3	99.1
5	5837	0.6	99.8
6	1515	0.2	99.9
7	430	0.0	100.0
8	132	0.0	100.0
9	80	0.0	100.0
10	44	0.0	100.0
11	23	0.0	100.0
12	16	0.0	100.0
13	13	0.0	100.0
14	10	0.0	100.0
15	9	0.0	100.0
16	9	0.0	100.0
17	5	0.0	100.0
18	5	0.0	100.0
19	7	0.0	100.0
20	4	0.0	100.0
21	3	0.0	100.0
22	3	0.0	100.0
23	1	0.0	100.0
24	1	0.0	100.0
25	2	0.0	100.0
26	2	0.0	100.0
30	1	0.0	100.0
35	1	0.0	100.0
37	1	0.0	100.0
Total	958530	100.0	

Table 3: Distribution of farms by # of products with yield index > 1

	(1) Total outputs	(2) Total outputs	(3) Log total outputs	(4) Log total outputs	(5) Log total outputs	(6) Log total outputs
Total inputs	0.144***		0.00334***			0.00328***
	(64.39)		(60.20)			(62.42)
Log total inputs		63.72***		1.565***		
		(52.80)		(62.70)		
Log average farmsize					0.0690***	0.0469***
					(8.09)	(11.59)
Constant	-26.16***	-348.8***	2.085***	-5.939***	3.678***	2.087***
	(-23.45)	(-46.84)	(75.44)	(-38.56)	(299.13)	(79.89)
Observations	1122	1122	1122	1122	1122	1122
R^2	0.787	0.713	0.764	0.778	0.055	0.789

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: Regressing total outputs produced over total inputs used at the municipality level

	(1) Land Use	(2) Land Use	(3) Land Use	(4) Land Use	(5) Land Use	(6) Land Use	(7) Land Use
Yield	0.977***	0.971***	0.870***	0.851***	0.862***	0.863***	0.858***
	(448.11)	(76.43)	(27.52)	(27.35)	(26.63)	(27.36)	(26.22)
Product FE	No	No	Yes	Yes	Yes	Yes	Yes
Munic FE	No	Yes	No	Yes	No	No	No
Non-ag Controls	Yes	No	No	No	No	Yes	Yes
Soil-voc Controls	Yes	No	No	No	Yes	No	Yes
Observations	109417	131331	131331	131331	109417	131331	109417
R^2	0.617	0.629	0.682	0.703	0.686	0.689	0.691

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5: Regressing land usage indices over yield indices

	(1) row_total_rca	(2) row_total_rca	(3) row_total_rca	(4) row_total_rca	(5) row_total_rca
row_total_harv	0.00126***	0.00121***	0.000804*	0.00152**	0.00132**
	(7.82)	(7.94)	(2.57)	(2.91)	(2.91)
workers_perm_per_area			0.00360***	0.00509***	0.00702***
			(7.44)	(11.40)	(15.72)
workers_permanent_total				-0.0000135**	-0.0000118**
				(-2.87)	(-2.87)
Constant	1.178***	0.363***	1.209***	1.202***	0.345***
	(783.93)	(13.60)	(635.51)	(341.04)	(10.34)
Munic Fixed Effects	No	Yes	No	No	Yes
Observations	958530	958530	668504	668504	668504
R^2	0.014	0.324	0.002	0.003	0.334

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6: Regressing total discrete yield index over total land harvested

	(1) log_rca_total	(2) log_rca_total	(3) log_rca_total	(4) log_rca_total	(5) log_rca_total
log_harv_total	0.0570***	0.0678***	0.0691***	0.0474***	0.0548***
	(194.41)	(196.30)	(171.69)	(31.48)	(37.72)
workers_perm_per_area			-0.0000349	-0.000934***	-0.000985***
			(-0.21)	(-4.02)	(-4.17)
log_workers_total				0.00896***	0.0135***
				(6.33)	(10.01)
Constant	0.357***	0.142***	0.162***	0.370***	0.156***
	(677.07)	(6.24)	(5.18)	(429.29)	(4.99)
Munic Fixed Effects	No	Yes	Yes	No	Yes
Observations	689321	689321	484390	483640	483640
R^2	0.049	0.281	0.289	0.045	0.289

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7: Regressing log of total discrete yield index over log of total land harvested

Rival goods	Non-rival goods
Boiler	Water source
Tillage Equipment - Sowing & Harvesting	Water protection practices
Outboard motor	Water issues
Motor pump	Soil protection practices
Aerator	Energy sources
Chainsaw	Technical assistance
Incinerator Oven	Credit
Winche	Soil management practices
CableVia	Use of natural fertilizer
Tractoelevadores	Applied for and received/not received loan
Cooling Tank	Did not apply for credit
Cold room	Natural forest or moor with exploitation
Weighing machine	Handling of natural waste
Seeder	Use of natural wastes to fertilize
Deburring machine	Handle debris from plastic, glass, PVC, etc.
Motor Cleaner	
Pasture Mincer	
Combine With Less Than 100 Hp	
Combine With Over 100 Hp	
Real Estate Management	
Manuring machine	
Tractor With Less Than 100 Hp	
Tractor With Over 100 Hp	
Motorized Fumigator	
Automatic Watering System	
Automatic feeder	
Automated Production Equipment	
Ahoyadora	
Laser Leveler	
Water Pump	
Thermal Plant	
Power plant	
Solar panel	
Prenez Detector	
Insulation Equipment	
Climate Record Equipment	
Environment Controller	
Vehicle	
Motorboat Lango-Canoa	
Wooden Sailing Ship	
Wooden Boat To Stick Or Rowing	
Motor Boats	
Fiberglass Boat Sailing Or Rod Or Rowing	
Glass Fiber Vessel With Motor	
Helicopter	
Plane	
Motorbike	
Manual Traction Equipment	
Other machinery N.C.P.	
Soil type	

Table 8: Two groups of inputs - Machinery and labour, and ‘other’ inputs

	(1) Yield	(2) Yield	(3) Yield	(4) Yield	(5) Yield	(6) Yield	(7) Yield	(8) Yield	(9) Yield	(10) Yield
Density	0.0143***	0.00752***	0.00749***	0.0139***	0.00929***	0.00922***	0.0116***	0.0116***	0.00976***	0.0106***
	(152.36)	(66.99)	(67.88)	(16.29)	(6.91)	(10.13)	(10.80)	(10.80)	(12.30)	(12.78)
Density prime	-0.00223***			0.00484						
	(-20.63)			(1.71)						
Density prime rival		-0.0129***	-0.0113***		-0.0146***	-0.00223	-0.00266*	-0.00266*	-0.00289*	-0.00387**
		(-73.85)	(-63.23)		(-5.29)	(-1.51)	(-2.05)	(-2.05)	(-2.12)	(-3.21)
Density prime non-rival		0.0163***	0.0138***		0.0141***	0.00392	0.0101***	0.0101***	0.00435*	0.00580**
		(70.61)	(59.07)		(4.06)	(1.70)	(3.60)	(3.60)	(2.03)	(2.73)
Adj. Yield			-0.0155***	-0.0288**	-0.0405***	-0.0126***	-0.0280*	-0.0280*	-0.00946***	-0.0100***
			(-16.14)	(-2.64)	(-3.59)	(-4.74)	(-2.57)	(-2.57)	(-5.68)	(-7.04)
Constant	-0.0378***	-0.0403***	-0.0247***							
	(-70.36)	(-75.76)	(-24.13)							
Product FE	No	No	No	Yes	No	Yes	Yes	Yes	Yes	Yes
Munic FE	No	No	No	Yes	Yes	No	Yes	Yes	No	No
Non-ag Controls	No	No	No	No	No	No	No	Yes	No	Yes
Muni-level Controls	No	No	No	No	No	No	No	No	Yes	Yes
Observations	483582	483582	463386	463386	463386	463386	463386	463386	454713	454713
R^2	0.131	0.152	0.144	0.273	0.155	0.265	0.275	0.275	0.270	0.272

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 9: Regressing yield outcome using densities

	(1) Yield	(2) Yield	(3) Yield	(4) Yield	(5) Yield	(6) Yield	(7) Yield	(8) Yield
Density	0.00756***	0.00757***	-0.0108***	0.00780***	0.00848***	0.00894***	-0.0872*	0.0116***
	(67.71)	(67.76)	(-6.18)	(70.27)	(5.82)	(6.16)	(-2.35)	(10.80)
Density prime rival	-0.0104***	-0.0103***	-0.0165***	-0.0102***	-0.0116***	-0.0119***	0.00410	-0.00266*
	(-59.61)	(-59.00)	(-8.97)	(-58.43)	(-3.76)	(-3.81)	(0.15)	(-2.05)
Density prime non-rival	0.0133***	0.0130***	0.0411***	0.0127***	0.0145***	0.0137***	-0.0773	0.0101***
	(57.49)	(56.34)	(16.79)	(55.17)	(4.18)	(3.93)	(-2.10)	(3.60)
Is Animal	0.260***	0.204***			0.248**	0.199**		
	(58.26)	(19.76)			(3.30)	(3.30)		
Dp rival.Is Animal		-0.00138				-0.00553		
		(-0.75)				(-0.29)		
Dp non-rival.Is Animal		0.00530***				0.0334		
		(4.22)				(1.40)		
D.Is Animal						-0.0225		
						(-1.34)		
Adj. Yield								-0.0280*
								(-2.57)
Constant	-0.0400***	-0.0391***	0.183***	-0.0393***				
	(-75.70)	(-75.06)	(17.55)	(-75.44)				
Product FE	No	No	No	No	No	No	Yes	Yes
Munic FE	No	No	No	No	Yes	Yes	Yes	Yes
R2	0.179	0.179	0.0556	0.139	0.186	0.187	0.523	0.275
Observations	483582	483582	12342	471240	483582	483582	12342	463386
Animals only			Yes				Yes	
Crops Only								Yes

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 10: Regressing yield outcome using densities with animals

	(1) Land Use	(2) Land Use	(3) Land Use	(4) Land Use	(5) Land Use	(6) Land Use	(7) Land Use	(8) Land Use	(9) Land Use	(10) Land Use
Density	0.0264***	0.0140***	0.0140***	0.0261***	0.0203***	0.0135***	0.0218***	0.0218***	0.0141***	0.0162***
	(102.19)	(47.89)	(48.14)	(16.94)	(5.99)	(6.97)	(11.20)	(11.20)	(7.85)	(8.96)
Density prime	-0.00471***			0.00614						
	(-13.90)			(1.15)						
Density prime rival		-0.0317***	-0.0316***		-0.0329***	-0.000404	-0.00892*	-0.00892	-0.00231	-0.00608
		(-67.88)	(-67.70)		(-5.91)	(-0.11)	(-2.03)	(-0.00)	(-0.70)	(-1.94)
Density prime non-rival		0.0349***	0.0349***		0.0389***	0.00459	0.0191***	0.0191	0.00634	0.00954*
		(65.96)	(66.16)		(6.14)	(0.91)	(3.39)	(0.00)	(1.35)	(2.07)
Adj. Yield			-0.0377***	-0.00110	0.0103	-0.0290**	-0.00179	-0.00179	-0.0318***	-0.0216***
			(-11.38)	(-0.10)	(0.40)	(-3.26)	(-0.16)	(-0.00)	(-7.69)	(-6.16)
Constant	-0.0969***	-0.101***	-0.0665***							
	(-52.81)	(-56.73)	(-18.83)							
Product FE	No	No	No	Yes	No	Yes	Yes	Yes	Yes	Yes
Munic FE	No	No	No	Yes	Yes	No	Yes	Yes	No	No
Non-ag Controls	No	No	No	No	No	No	No	Yes	No	Yes
Muni-level Controls	No	No	No	No	No	No	No	No	Yes	Yes
Observations	131331	131331	131331	131331	131331	131331	131331	131331	131033	131033
R^2	0.157	0.207	0.207	0.411	0.265	0.379	0.413	0.413	0.382	0.390

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 11: Regressing land usage outcome over densities

	(1) Yield	(2) Yield	(3) Yield	(4) Yield	(5) Yield	(6) Yield	(7) Yield	(8) Yield	(9) Yield
Density	0.0195***	0.0120***	0.0120***	0.0133***	0.0132***	0.0143***	0.0143***	0.0137***	0.0141***
	(169.18)	(83.48)	(84.09)	(8.03)	(11.82)	(10.62)	(10.62)	(13.31)	(12.96)
Density prime	-0.00362***								
	(-26.64)								
Density prime rival		-0.0133***	-0.0126***	-0.0140***	-0.00151	0.00247	0.00247	-0.00177	-0.00196
		(-68.33)	(-63.82)	(-4.77)	(-1.13)	(1.57)	(0.00)	(-1.47)	(-1.88)
Density prime non-rival		0.0161***	0.0145***	0.0154***	0.00199	0.00971***	0.00971	0.00195	0.00325
		(59.99)	(53.43)	(3.83)	(0.88)	(3.45)	(0.13)	(0.94)	(1.58)
Adj. Yield			-0.0188***	-0.0672***	-0.0117***	-0.0424**	-0.0424***	-0.00810***	-0.0106***
			(-15.61)	(-3.94)	(-3.55)	(-2.87)	(-4.51)	(-3.87)	(-5.98)
Constant	-0.0401***	-0.0433***	-0.0258***						
	(-65.37)	(-71.64)	(-20.49)						
Product FE	No	No	No	No	Yes	Yes	Yes	Yes	Yes
Munic FE	No	No	No	Yes	No	Yes	Yes	No	No
Non-ag Controls	No	No	No	No	No	No	Yes	No	Yes
Muni-level Controls	No	No	No	No	No	No	No	Yes	Yes
Observations	374063	374063	353867	353867	353867	353867	353867	345442	345442
R^2	0.193	0.210	0.210	0.224	0.353	0.368	0.368	0.358	0.361

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 12: Regressing yield outcome using densities (considering only products that exist in muni)

	(1) Yield	(2) Yield	(3) Yield	(4) Yield	(5) Yield	(6) Yield	(7) Yield	(8) Yield
Density	0.0119***	0.0119***	-0.0108***	0.0123***	0.0121***	0.0121***	-0.0872*	0.0143***
	(82.77)	(82.75)	(-6.18)	(85.85)	(6.70)	(6.71)	(-2.35)	(10.62)
Density prime rival	-0.0111***	-0.0111***	-0.0165***	-0.0109***	-0.0101**	-0.0101**	0.00410	0.00247
	(-56.90)	(-56.92)	(-8.97)	(-56.06)	(-2.90)	(-2.84)	(0.15)	(1.57)
Density prime non-rival	0.0134***	0.0134***	0.0411***	0.0129***	0.0157***	0.0157***	-0.0773	0.00971***
	(49.90)	(49.73)	(16.79)	(47.97)	(3.88)	(3.83)	(-2.10)	(3.45)
Is Animal	0.202***	0.204***			0.185*	0.168**		
	(44.71)	(19.68)			(2.45)	(2.93)		
Dp rival.Is Animal		0.000542				-0.000851		
		(0.29)				(-0.04)		
Dp non-rival.Is Animal		-0.000442				0.00184		
		(-0.35)				(0.16)		
Adj. Yield								-0.0424**
								(-2.87)
Constant	-0.0434***	-0.0434***	0.183***	-0.0436***				
	(-72.18)	(-73.22)	(17.55)	(-73.51)				
Product FE	No	No	No	No	No	No	Yes	Yes
Munic FE	No	No	No	No	Yes	Yes	Yes	Yes
R2	0.226	0.226	0.0556	0.200	0.236	0.236	0.523	0.368
Observations	374063	374063	12342	361721	374063	374063	12342	353867
Animals only			Yes				Yes	
Crops Only								Yes

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 13: Regressing yield outcome using densities with animals (considering only products that exist in muni)

	(1) Land Use	(2) Land Use	(3) Land Use	(4) Land Use	(5) Land Use	(6) Land Use	(7) Land Use	(8) Land Use	(9) Land Use
Density	0.00170***	0.00334***	0.00332***	-0.00338*	0.00576***	0.0147***	0.0147***	0.00563***	0.00637***
	(4.41)	(7.75)	(7.69)	(-2.13)	(4.93)	(11.37)	(11.37)	(6.50)	(7.16)
Density prime	0.00392***								
	(7.34)								
Density prime rival		0.00383***	0.00381***	0.000658	-0.00509*	0.0000967	0.0000967	-0.00461**	-0.00313*
		(9.59)	(9.56)	(0.57)	(-2.37)	(0.06)	(0.00)	(-3.15)	(-2.47)
Density prime non-rival		-0.00139**	-0.00140**	-0.00526**	0.00227	0.00572**	0.00572	0.00195	0.00125
		(-2.67)	(-2.71)	(-2.62)	(1.03)	(2.98)	(0.00)	(1.19)	(0.81)
Adj. Yield			0.00862	-0.00841	0.00679	0.0338*	0.0338	0.0111	0.0233***
			(1.59)	(-0.39)	(0.48)	(1.98)	(0.01)	(1.96)	(4.15)
Constant	0.956***	0.959***	0.951***						
	(240.35)	(240.14)	(151.06)						
Product FE	No	No	No	No	Yes	Yes	Yes	Yes	Yes
Munic FE	No	No	No	Yes	No	Yes	Yes	No	No
Non-ag Controls	No	No	No	No	No	No	Yes	No	Yes
Muni-level Controls	No	No	No	No	No	No	No	Yes	Yes
Observations	21812	21812	21812	21810	21804	21802	21802	21754	21754
R^2	0.020	0.021	0.022	0.356	0.138	0.471	0.471	0.144	0.249

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 14: Regressing land usage outcome over densities (considering only products that exist in muni)

	(1) Grown	(2) Grown	(3) Grown	(4) Grown	(5) Grown	(6) Grown	(7) Grown	(8) Grown
Density	0.0137***	0.0137***	0.00999***	0.0165***	0.0133***	0.0133***	0.00998***	0.0165***
	(96.06)	(96.17)	(9.62)	(6.56)	(10.93)	(10.92)	(9.65)	(6.57)
Density prime rival	-0.0154***	-0.0154***	-0.000981	-0.0198***	-0.00351*	-0.00351*	-0.00114	-0.0198***
	(-75.59)	(-75.51)	(-0.58)	(-4.89)	(-2.01)	(-2.01)	(-0.67)	(-4.89)
Density prime non-rival	0.0170***	0.0172***	0.00466	0.0194***	0.0145***	0.0145***	0.00507	0.0194***
	(61.00)	(61.67)	(1.79)	(3.69)	(4.61)	(4.61)	(1.94)	(3.69)
Avg yield		-0.0292***				-0.0118	-0.0256***	0.0117
		(-24.55)				(-1.56)	(-5.58)	(0.75)
Constant	-0.0286***	-0.00172						
	(-51.12)	(-1.40)						
Product FE	No	No	Yes	No	Yes	Yes	Yes	No
Munic FE	No	No	No	Yes	Yes	Yes	No	Yes
Observations	472362	472362	472362	472362	472362	472362	472362	472362
R^2	0.172	0.173	0.389	0.201	0.404	0.404	0.390	0.201

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 15: Regressing existence of crop production by densities

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Grown - High yield	Grown - High yield	Grown - High yield	Grown - High yield	Grown - High yield	Grown - High yield	Grown - High yield	Grown - High yield
Density	0.00621***	0.00621***	0.00657***	0.00687***	0.00801***	0.00797***	0.00656***	0.00690***
	(57.06)	(57.08)	(7.52)	(5.44)	(8.04)	(8.01)	(7.52)	(5.46)
Density prime rival	-0.00886***	-0.00885***	-0.00289	-0.00895***	-0.000667	-0.000664	-0.00292	-0.00891***
	(-52.39)	(-52.36)	(-1.84)	(-4.03)	(-0.47)	(-0.47)	(-1.86)	(-4.02)
Density prime non-rival	0.0112***	0.0113***	0.00482*	0.0127***	0.0103***	0.0102***	0.00489*	0.0127***
	(48.77)	(48.92)	(2.01)	(3.67)	(3.61)	(3.59)	(2.03)	(3.67)
Avg yield		-0.00688***				-0.0327**	-0.00405*	-0.0420***
		(-8.10)				(-3.15)	(-2.18)	(-4.03)
Constant	-0.0253***	-0.0190***						
	(-57.03)	(-21.08)						
Product FE	No	No	Yes	No	Yes	Yes	Yes	No
Munic FE	No	No	No	Yes	Yes	Yes	No	Yes
Observations	472362	472362	472362	472362	472362	472362	472362	472362
R^2	0.107	0.107	0.237	0.115	0.245	0.245	0.237	0.116

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 16: Regressing high-yield vs. low-yield of crop production by densities