Growth of Firms and Migration in Mexico: Twin Processes

David Mayer-Foulkes[1]

Contents

[1] This project received invaluable help and guidance from Amir Ziai, mentor in springboard.com’s Data Science Intensive course.

1. Looking at the data

The search for economic opportunity drives both the creation of new firms and migration in Mexico. As economic development proceeds, people go after firms for employment and to purchase goods, and firms go after firms and people, seeking labor, inputs, and customers.

These twin processes can be examined using firm data from the National Statistical Directory of Economic Units (DENUE) for 2012 and 2016, and Census population data for 2010 and 2015. I aggregate this data to the municipal level so as to make the local economies our unit of observation.

Migration

Municipalities include city districts ("Delegaciones") and can thus number millions of people. On the other hand in 2015 the smallest one had about 87 people.

Define municipal rank as the proportion of people that live in smaller municipalities. The long history of migration that began with the shift from agriculture to industry can be seen simply by plotting municipal population against municipal rank (Figure 1).

Migration still continues, except that today workers seek both rural and urban employment. A plot of population growth against population (Figure 2) shows that population grew less than average in smaller municipalities. Also, population growth was somewhat slower in the very populated municipalities. If we subdivide municipalities into the four population intervals [0, 0.03), [0.03, 0.27), [0.27, 0.63), [0.63, 1.00], mean population growth increases across the first, second and third intervals. However, it then decreases to the last interval. (These

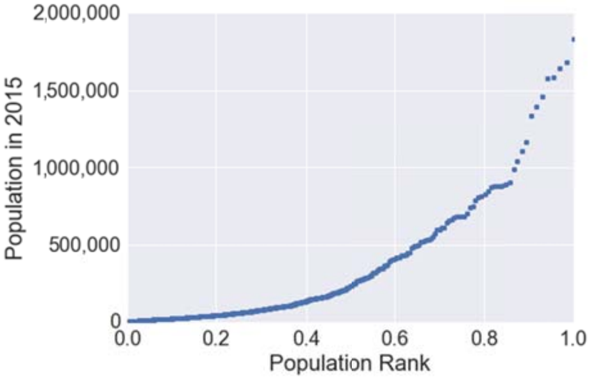


Figure 1. Municipal population versus municipal rank (proportion of people living in smaller municipalities)

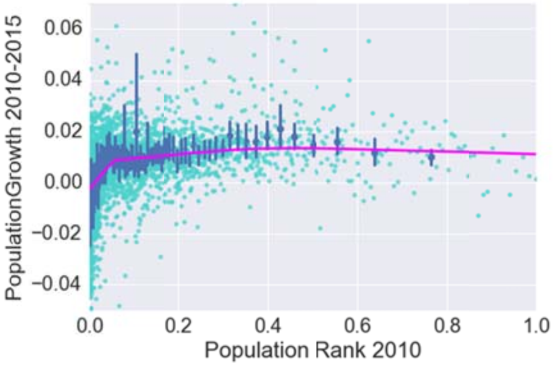
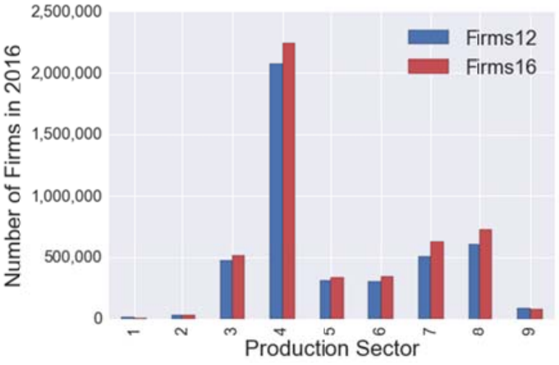


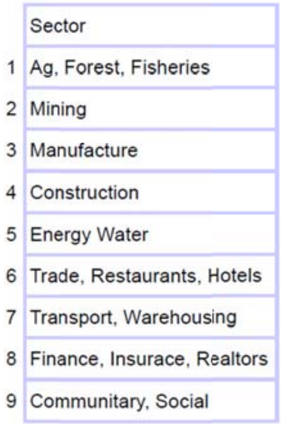
Figure 2. Municipal population growth 2010-2015 versus population. In light blue, the scatterplot (some outliers not shown). In dark blue, 95% confidence intervals for mean municipal population growth for groups of 25 municipalities. In magenta, locally weighted linear regression.

comparisons are significant at better than 1% confidence using a means comparison test). By the way, the number of municipalities in each of these four intervals is 946, 1164, 295 and 51. This means that the municipalities in the lowest group from which there is migration are quite small, holding 3% of the population. On the other hand the 51 largest are quite large, holding 37% of the population.

1.1. Firm Growth

The DENUE data classifies firms into 9 main production sectors (the 1 digit level). The sector with the most firms was construction (Figure 3). This was followed by finance, insurance and realtors; transport and warehousing; and manufacturing. Trade, restaurants and hotels overtook energy and water during the period. These were followed by communitary and social, mining, and agriculture, forestry and fishing, which all decreased in numbers during the period. However, using a means comparison test, the only significant differences in firm numbers at the 1% level were the decrease in agriculture, forestry and fishing, and the increase in transport and warehousing. The increases in finance, insurance, and realtors, and trade, restaurants and hotels were significant at the 4.7% and 5.6% levels.

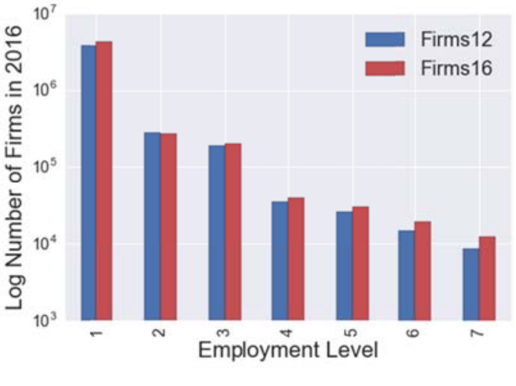


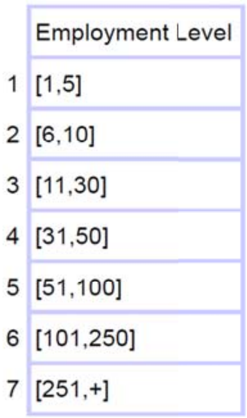


	Sector
1	Ag, Forest, Fisheries
2	Mining
3	Manufacture
4	Construction
5	Energy Water
6	Trade, Restaurants, Hotels
7	Transport, Warehousing
8	Finance, Insurace, Realtors
9	Communitary, Social

Figure 3. Firm growth by sectors, 2012-2016.

We can also examine the growth in firm numbers by employment levels (Figure 4). The number of firms increased in every employment level except for [6, 10]. However, using a means comparison test, the only significant differences in firm numbers at the 1% level were for firms with 51 employees or higher, employment levels 5, 6, and 7. The increases in employment at levels [1, 5] and [31, 50] were significant with a confidence of 2%.





	Employment Level
1	[1,5]
2	[6,10]
3	[11,30]
4	[31,50]
5	[51,100]
6	[101,250]
7	[251,+]

Figure 4. Firm growth by employment ranges, 2012-2016.

1.2. Interaction of Firms and Population Numbers

There are several general questions on how firms numbers relate to population numbers. First, is there some "law" relating these quantities? Second, when the economy grows, how do the numbers of firms in different sectors and employment levels grow? Do numbers of firms grow proportionally, or is there a "migration" from small firms to large firms? That is, is development achieved with larger firms rather than more firms? Monetary data on production is not readily available so we work with numbers and sizes of firms instead.

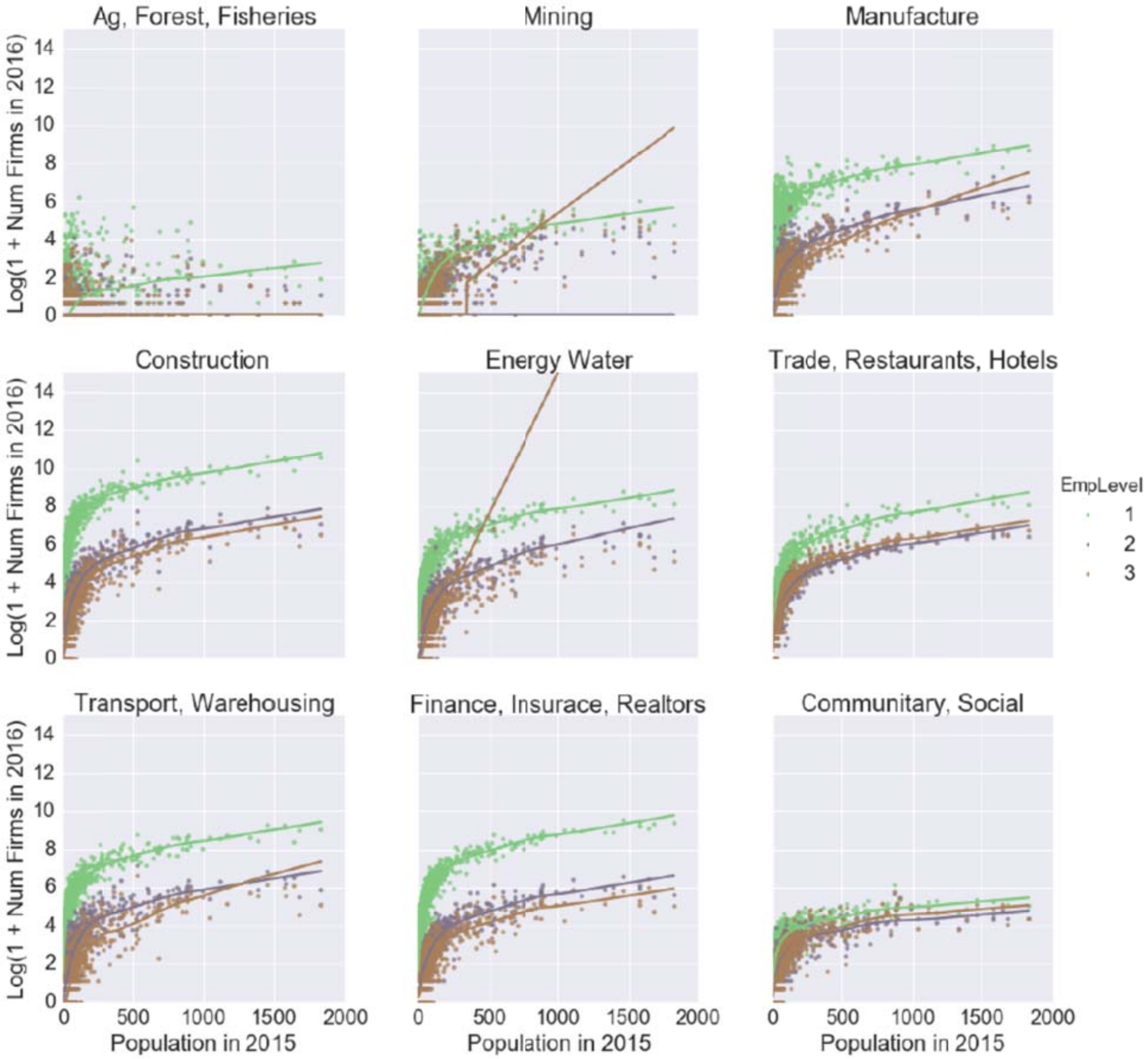


Figure 5. log(1 + Number of Firms) against Municipal Population (in thousands). Each panel represents a sector of production. Each panel shows a scatterplot for employment levels [1, 5], [6, 10] and [11, 30], together with a are locally weighted regression plot. log(1 + Number of Firms) is used to avoid log(0).

Figures 5 and 6 plot, for each production sector, a scatterplot log(1 + Number of Firms) against municipal population. 1 is added to the number of firms before taking the logarithm to avoid the occurrence of log(0) when there are no firms of a certain type. Figure 5 concentrates on the three lower employment levels, [1, 5], [6, 10] and [11, 30], and Figure 6 on larger firms with employment levels [31, 50], [51, 100] and [101, 250], [251, +]. Both figures show that after a threshold, the number of firms grows approximately exponentially as compared to the population of Mexican municipalities, with clear differences across employment levels. In fact, the exponential coefficient tends to be larger for smaller firms. This is verified to a 5% confidence level in several of the plots. Perhaps larger firms in fact eschew high population areas.

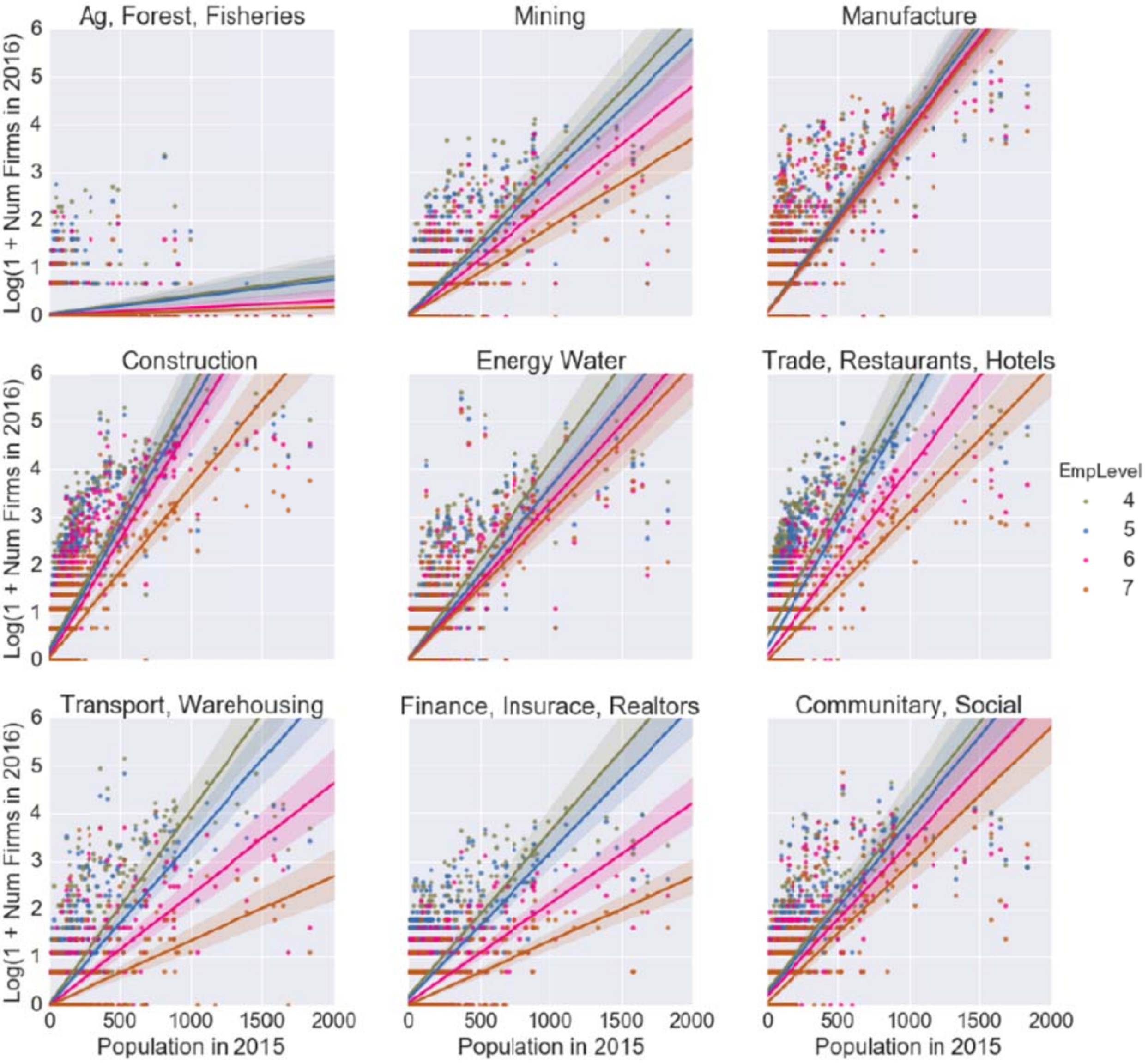


Figure 6. log(1 + Number of Firms) against Municipal Population (in thousands). Each panel represents a sector of production. Each panel shows a scatterplot for employment levels [31, 50], [51, 100], [101, 250], [251, +], together with a linear regression plot. log(1 + Number of Firms) is used to avoid meaningless log(0).

Two particular qualitatively exceptional behaviors are noticeable in the figures. First, agriculture, forestry and fishing behave quite differently from the other production sectors. Second, employment levels [6, 10] and [11, 30] behave similarly rather than distinctly. In manufacture also behavior is similar across employment levels [31, 50], [51, 100], [101, 250], and [251, +].

1.3. Firm Growth and Population Numbers

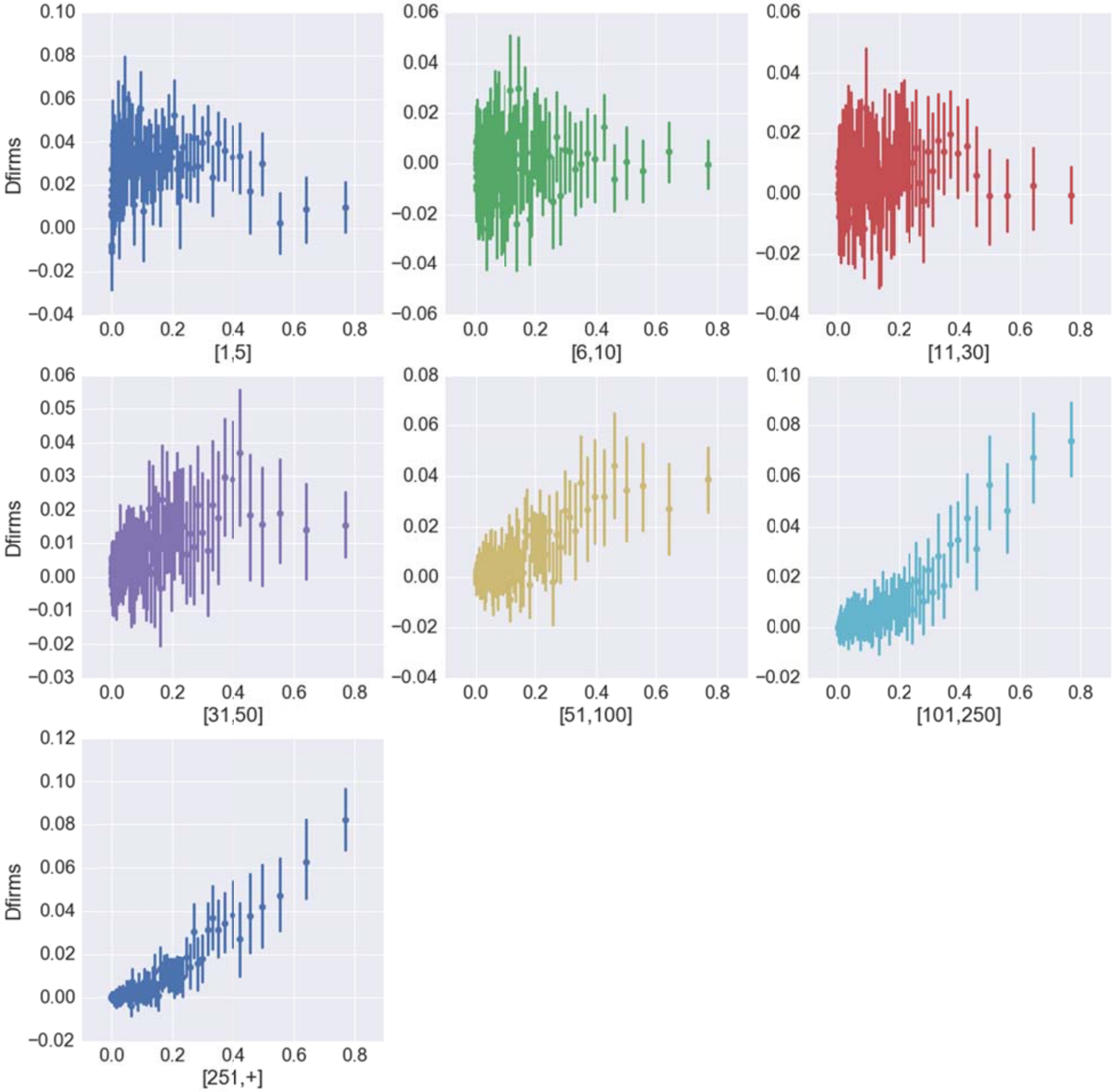


Figure 7. Growth in firm numbers by employment level (2012-2016). The graph shows a binned scatterplot of growth against population: growth means for groups of 25 municipalities, and a 95% confidence interval.

Growth in firm numbers varies across municipalities according to their population rank, just as population growth does. We first consider firms by employment levels. Overall, the mean growth in firm numbers follows an inverse U curve for smaller employment levels. As higher employment levels are reached, the maximum of the inverse U curve occurs for higher values of the population ranking until finally only the increasing section remains. At the same time, there is very much variability in the firm growth data.

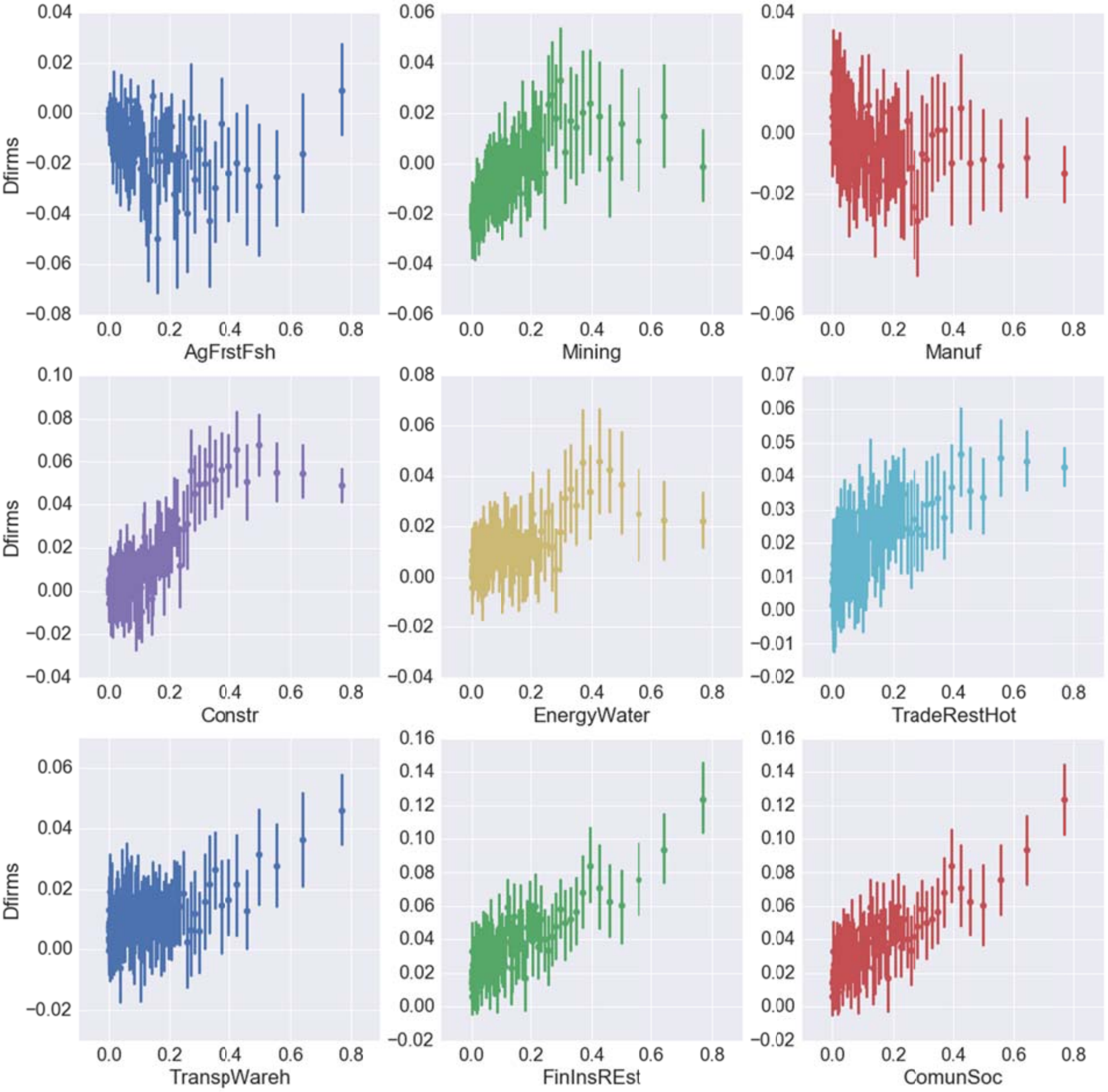


Figure 8. Growth in firm numbers by production sector (2012-2016). The graph shows a binned scatterplot of growth against population: growth means for groups of 25 municipalities, and a 95% confidence interval.

When we consider firms by production sectors (Figure 8), similar inverse U curve patterns are found for mining, construction, energy and water, and then just the increasing section for trade, restaurants and hotels, transport and warehousing, finance, insurance, and realtors, and community and social. One difference is that small municipalities may remain close to zero growth, the inverse U curve only appearing at a population rank of 0.1 or 0.2. Manufacturing shows the inverse U curve pattern, but has an additional region of new firms at low municipal populations. Agriculture, forestry and fishing is also atypical, displaying growth at both low and high municipal populations, with negative firm number growth displayed for a considerable number of intermediate municipal rankings.

Interplay between Firm Number and Population Growth

The next step in considering the interaction between firm number growth and population growth is considering the phase space for the dynamics between these two variables. This is a two dimensional plot with population along the x axis and firm numbers along the y axis, which displays arrows starting from the 2012 value of these variables and pointing in the direction of change of the two variables. While we could map every municipality onto the phase diagram in this way (for a particular class of firms), a 2,456 arrow plot would not really work. Instead we subdivide the population-firm number rectangle [0, 1] x [0, 1] into a 10 x 10 grid, and plot the averages of the municipal arrows. This is a binned phase diagram, similar to a binned scatterplot. The population rank gives us a 0 to 1 population measure. For firm numbers we use the variable log(+firm number)/max[log(1+firm number)], where the maximum is taken over both years 2012, 2016, We therefore also have a 0 to 1 measure for firm numbers, with the vertical dimension in log firm numbers representing a rate of change. Both dimensions are normalized to a yearly rate of change. For visualization purposes, the arrows are multiplied by 8 in length. They therefore represent change extrapolated to an 8 year period.

Figure 9 shows the result, for each combination of production sector and employment level. Each is plotted as a subpanel of the figure. These arrow plots represent the combined firm and population dynamics. They vary quite considerably across the different subplots. In particular agriculture, forestry and fishing display a considerable number of downward arrows. Many of the displays instead concentrate on what can be described as a parabolic trajectory in which the number of firms rises quite fast as the population rises from minimum levels. These “normal” trajectories tend to move towards the right with the number of firms rising exponentially.

It is noteworthy, though, that at low firm sizes, other than in the agriculture, forestry and fishing sector, the number of firms rises faster for smaller employment levels than for larger employment levels, then often keeps to the parabolic trajectory. On the other hand the higher employment levels display growth spurts across municipal population sizes so long as they are above the smallest. The community-social sector loses lots of [1, 5] level firms.

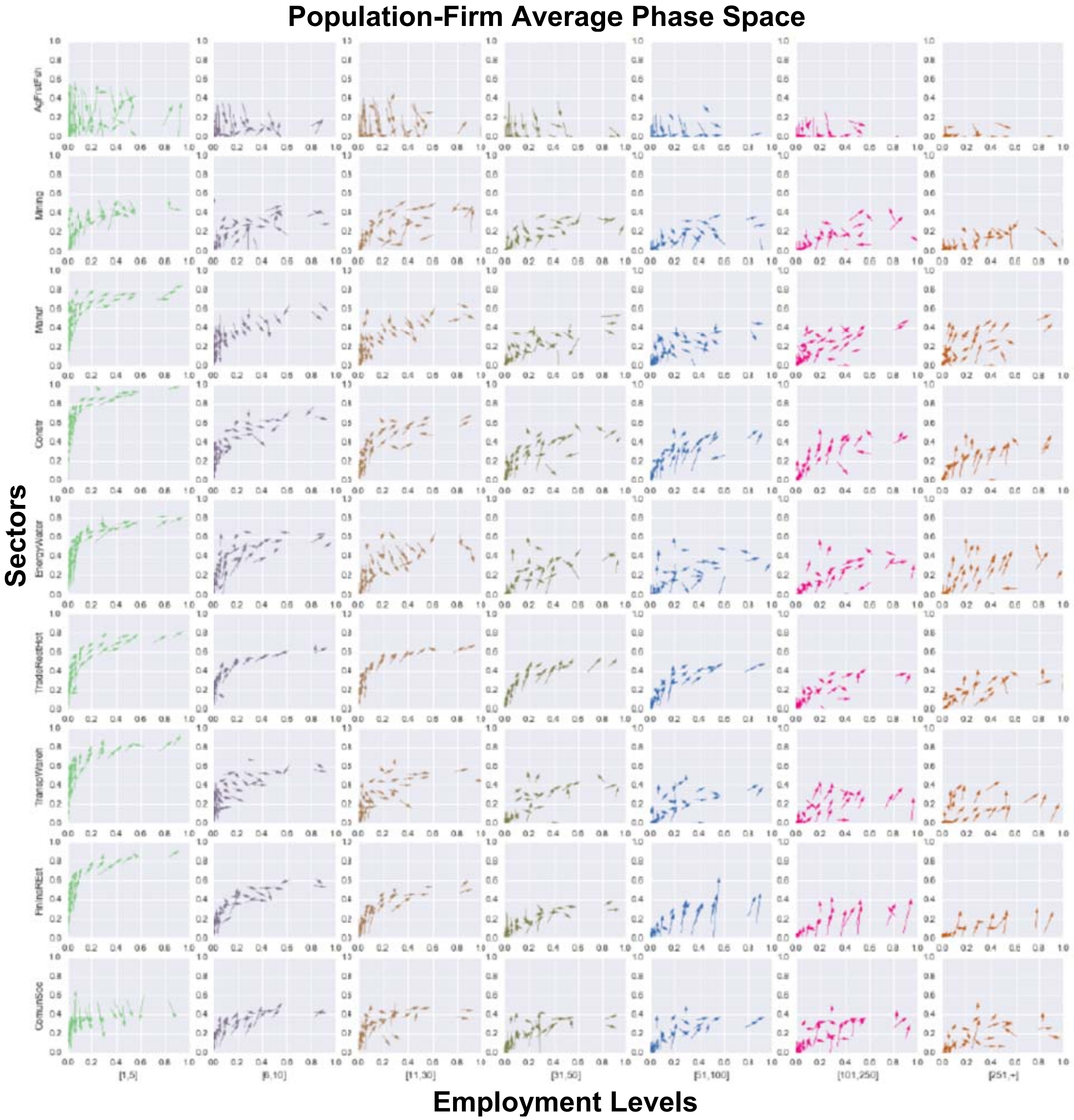


Figure 9. Each panel's horizontal and vertical axes are population 2010 and log(1 + NumFirms2012). Each arrow indicates average municipal rate of change in these variables (to 2015 and 2016) for bins forming a 10×10 grid subdividing each subplot. For a clearer view magnify the subplots using PDF capabilities.

1.4. Complexity of Firm Change and Migration

While the graphs uncover some regularity in the patterns of firm and population growth, in fact they also show that the data is complex. Whenever we used a stronger lens we found again a diversity of phenomena. And this is precisely the definition of complexity. We are using highly aggregated data. Production sectors are in themselves diverse. Also municipal population characteristics and infrastructures vary immensely.

For example, when we consider small scale firms in the [1, 5] range, the first subplot in Figure 7 is quite similar to population Figure 2. In fact, this size firm is associated with the livelihood for many people, and therefore with population growth. However, when we include the scatter plot (Figure 10), this shows that there is a lot of additional variation. This is consistent with the idea that there are many external factors that play a role in particular instances of municipal evolution. Something similar occurs with each of the subplots in Figure 7. However, the municipal density moves towards higher population rankings, consistently with the idea that the observation that the maximum of the inverted U curve moves to the right.

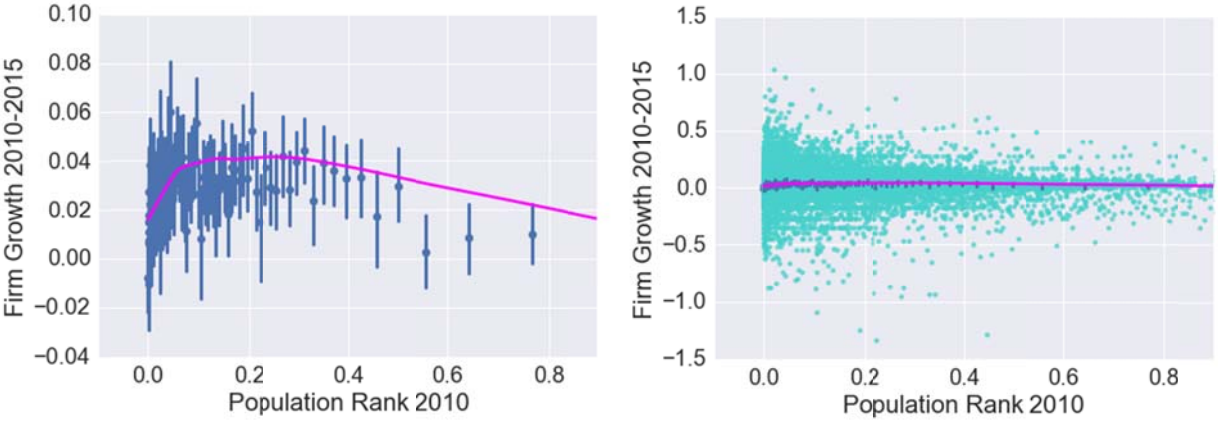


Figure 10. Both panels show growth in firm numbers (2012-2016) for employment level [1, 5], in different scales. In dark blue, both graphs shows a binned scatterplot of growth against population: growth means for groups of 25 municipalities, and a 95% confidence interval. They also show, in magenta, the results of a locally weighted linear regression. Finally, the panel on the right shows the scatterplot in light blue.

On the other hand, the shape of the municipal scatterplot does not change as much across the production sectors considered in Figure 8.

Now let us expand Figures 7 and 8 to consider all possible combinations of production sector and employment level. The results confirm that an inverted U pattern is often present. However, this certainly does not describe many other features that appear at this level of detail, and that are lost in the averages taken before.

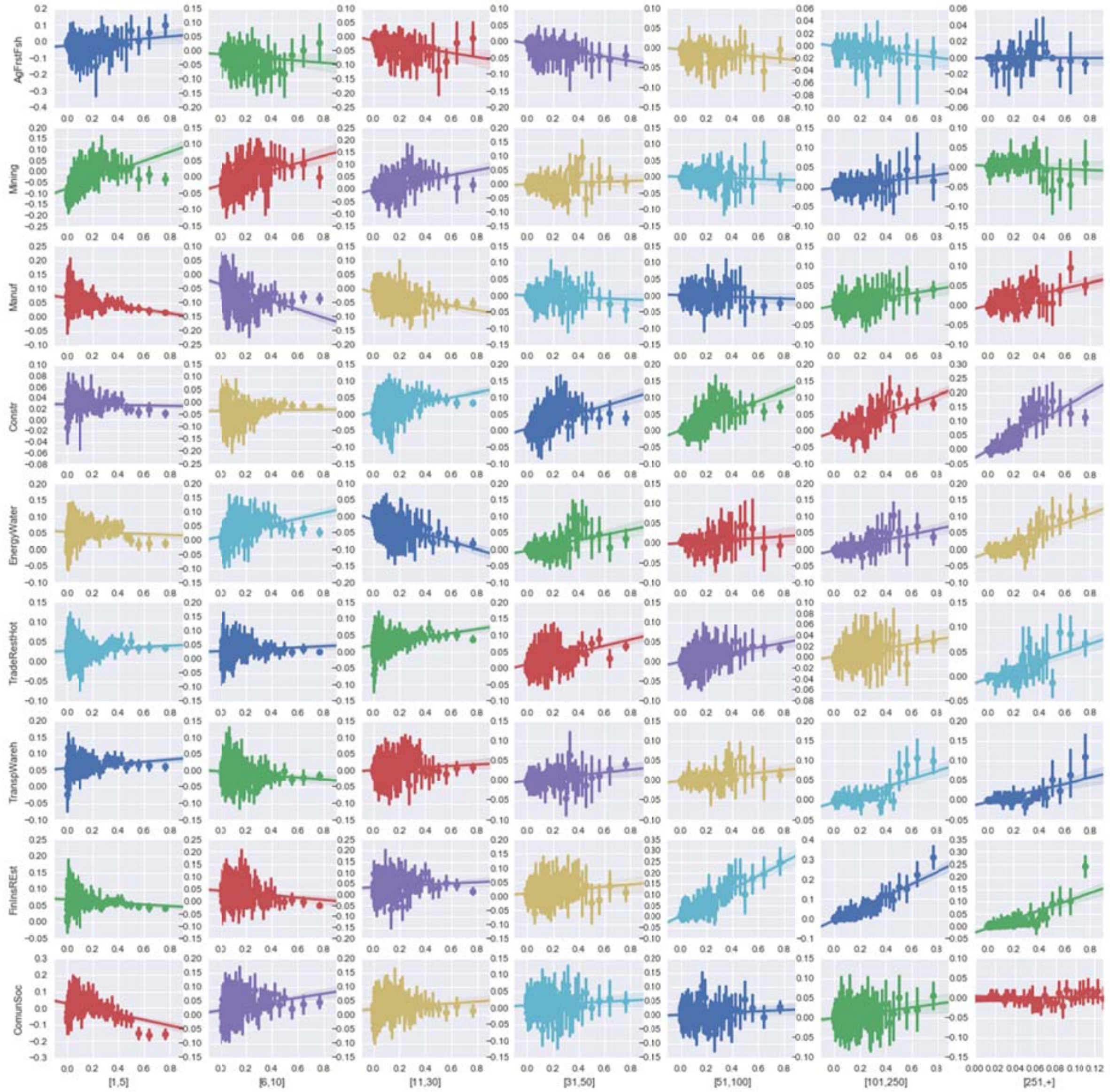


Figure 11. Growth in firm numbers (2012-2016) by production sector and employment level. Each panel shows a binned scatterplot of growth against population: growth means for groups of 25 municipalities, with a 95% confidence interval. A line obtained by linear regression is shown for reference.

Concluding, methods that will approach the data in detail, such as machine learning, will be very useful to approximate the considerable underlying complexity of firm growth and migration.

2. Preparing the data for a supervised machine learning application

The qualitative observation of the data has uncovered both complexity and interaction in the evolution of firm and population numbers at the municipal level in Mexico. Is there anything we can say from a bird's eye viewpoint about the aggregate?

The detailed firm information provided by DENUE indirectly portrays municipalities in quite a detailed way. We could certainly use this information to seek to model the behavior of particular production sectors in particular employment categories, in a way that could be useful to entrepreneurs or for policy purposes.

However, here we conduct a first approach tailored at eliciting information on the firm and population growth process at the aggregate level. An understanding of the process as a whole can help to inform the basic perspectives from which policy is made.

The DENUE information does not provide a way for weighting the different types of firms across production sectors and employment categories (such as production or employment) to construct a single indicator of firm numbers. For this reason we turn to a principal component analysis, which also serves the purpose of dimensional reduction.

2.1. Principal components for firm numbers

We take for each municipality the 63 indicators $log(1 + N_{SE}), 1 \leq S \leq 9, 1 \leq E \leq 7$ of firm numbers, where N_{SE} is the number of firms in production sectors *S* and employment categories *E*. To this number is added 1 so that the logarithm is not zero when the number of firms is zero. This way we have a logarithmic indicator differences of which are essentially rates of growth. These variables are scaled to have mean 0 and standard deviation 1. For the year 2012, the first four components control for 64.4%, 7.97%, 5.6% and 1.8% of the variance. Therefore the first three of these principal components are included as *features* in the machine learning evaluation. These are the principal components of firm growth. Specifically, we refer to the first component as *Firm Growth*. A scatterplot of the first two components is shown in Figure 12. The shape of the figure indicates a process of transition which is also supported by other figures below. For the 63 corresponding number of firm variables for 2016 we use the same 2012 scaling to construct corresponding features for 2016. The rate of change *dx* of the first component gives us a rate of development that we use as *label* in our evaluation. Figure 13 shows a binned scatterplot for *dx* along the *x* axis, together with a locally weighted linear regression.

2.2. Principal components for population numbers

A similar process is conducted for the population and migration related variables log population, proportion of people born in the same state, proportion of people living in the

US, and the CONAPO marginalization index, all for 2010. The first two principal components, xpop and ypop are kept, accounting for 93.4% and 3.88% of the variance.

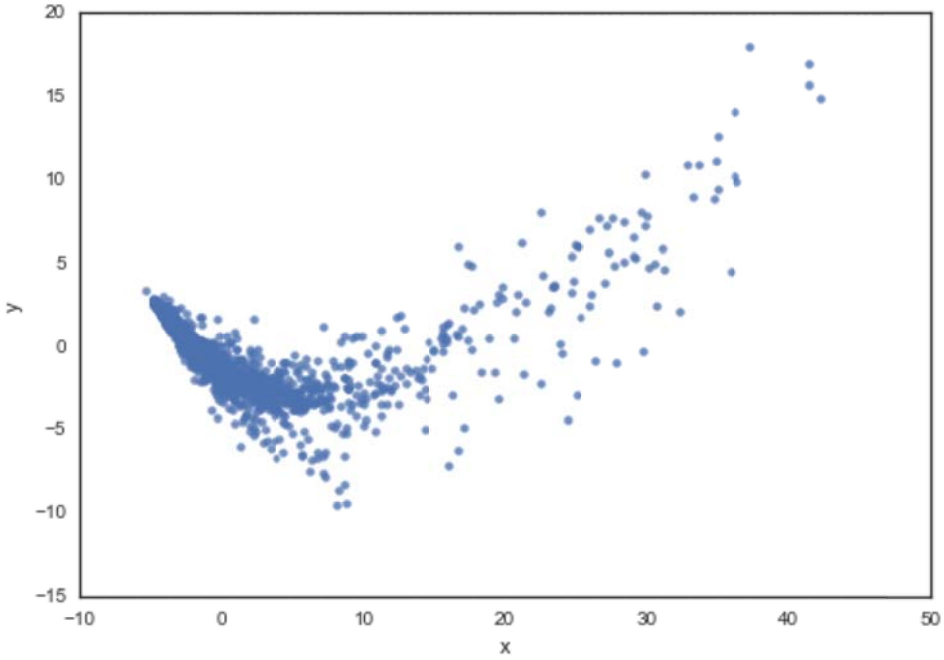


Figure 12. First two principal components *x* and *y* of 63 log firm number indicators.

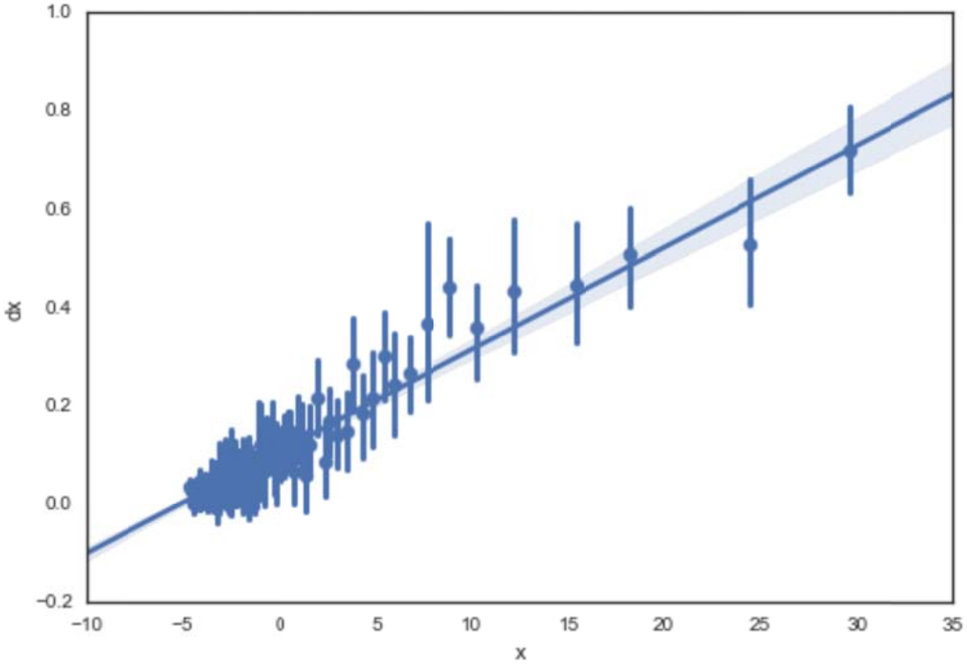


Figure 13. Binned scatterplot for mean municipal population of *dx* versus *x* (see text) for groups of 25 municipalities. 95% confidence intervals shown.

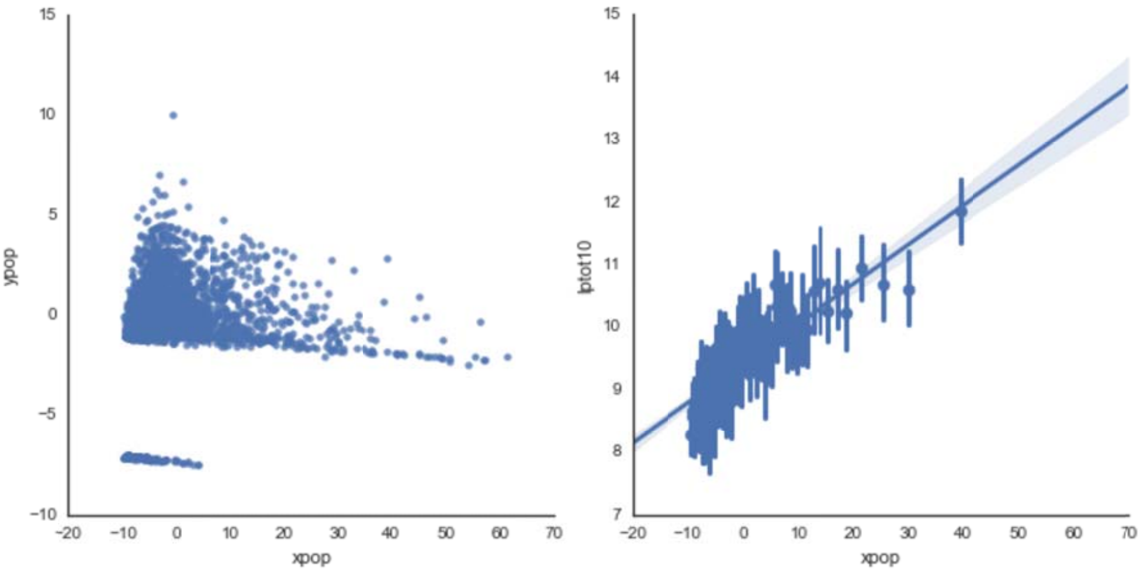


Figure 14. Scatterplots of ypop versus xpop and log population versus xpop.

The relation between the two principal components of population-migration and of the first principal component versus log population are shown in scatterplots in Figure 14. The relation between the principal component variables *x*, *y*, *xpop*; log population *lptot10*, and population rank *ptot10cum* for 2010, are shown in Figure 15 in terms of matrices of scatterplots and histograms. There are two panels, the first with points colored according to the employment category having the maximum number of firms, in each municipality, scored in standard deviations from the mean, the second according to the production sector holding the analogous maximum score. What is evident in these plots is that the selected variables carry a lot of information of the firm and population development process. They are thus excellent features for the analysis.

Note in particular how directly the population rank *ptot10cum* maps to the Firm Growth variable *x*. The same holds for log population *lptot10* and *x*. Correspondingly, the transition shape observed in Figure 12 between *x* and *y* is also observed between *lptot10* and *y*.

The principal components *xpop* and *ypop* are extrapolated to 2015 in the same way as before, using the 2010 principal component transformation, this time applied to the 2015 variables. However, two of these variables, the proportion of people born in the same state, and the proportion of people living in the US are unavailable for 2015. Thus we use instead the 2010 indicators. Both of these variables are proportions of the population, so only a relatively small error is introduced.

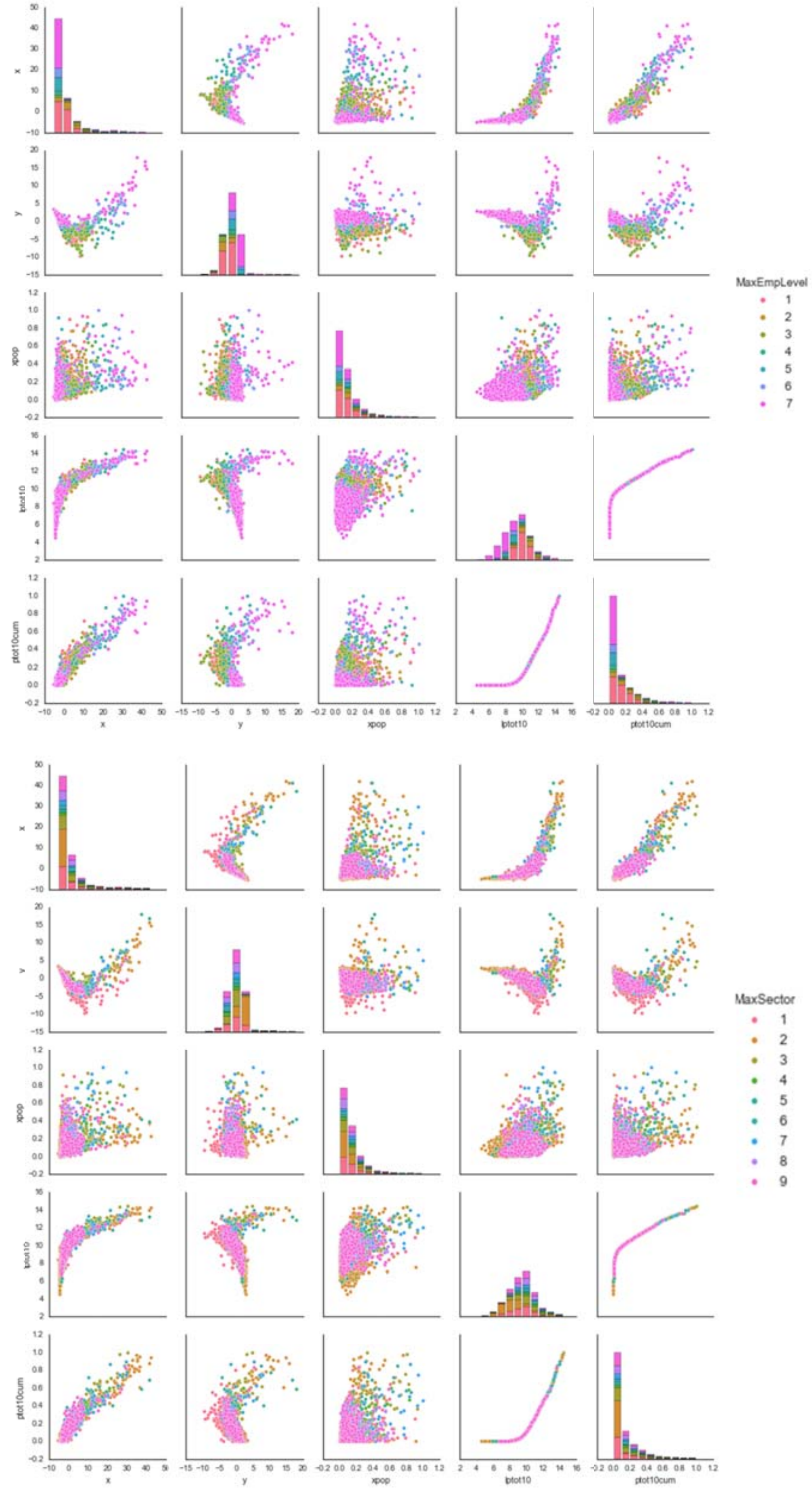


Figure 15. Matrix of scatterplots and histograms for the variables *x*, *y*, *xpop*, log population *lptot10*, and population rank *ptot10cum* for 2010.

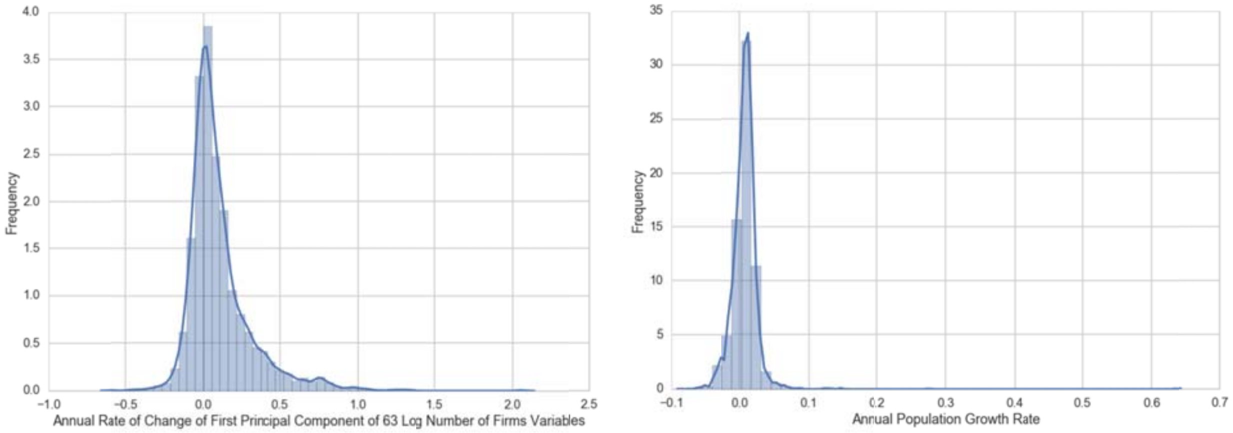


Figure 16. Histograms for the municipal firm and population growth indicators, Firm Growth *x* ≥ 5% and population growth *dptot* ≥ 1%

2.3. Labels

Two parallel analyses are conducted, one on firm growth and the other on population growth. For firm growth our label will be based on the variable *dx* already mentioned. For population growth we use just that – population growth *dptot*. Because we are interested in obtaining qualitative information on the firm and population growth process, we use the Random Forest Classifier and the Random Forest Regression for which feature importance can be retrieved.

The distribution of the firm and population growth variables is shown in Figure 16. We can now

Population Growth	0	1
PC Firm Growth		
0	854	392
1	555	654

Figure 17a. Cross Tabulation of municipal Firm Growth *dx* ≥ 0.05 and Population Growth *dptot* ≥ 0.01.

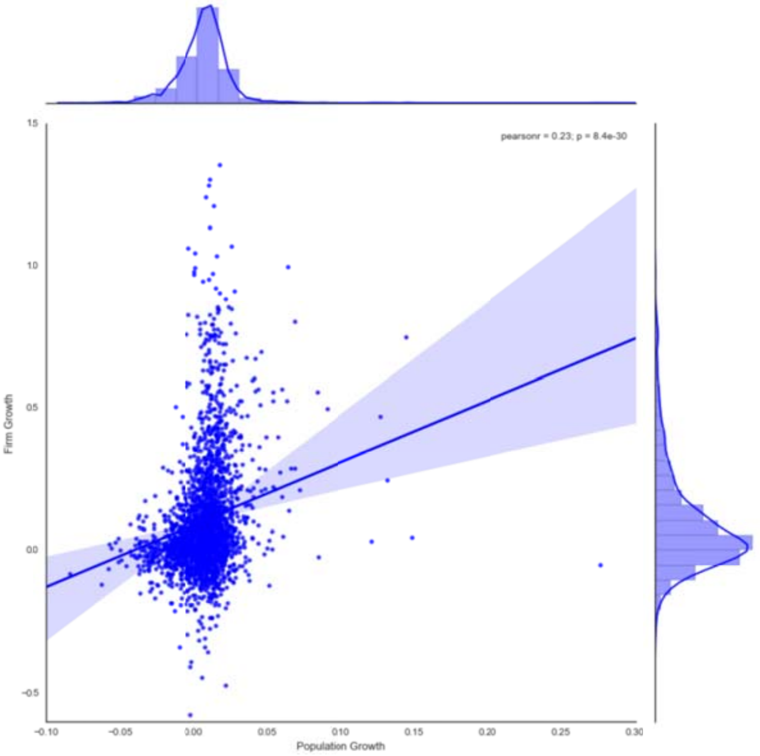


Figure 17b. Scatterplot and histograms of municipal Firm Growth and Population Growth.

define two categorical 1 and 0 indicators for firm and population growth, according to whether $dx \geq 0.05$ (1,209 versus 1,246 municipalities) and $dptot \geq 0.01$ (1,046 versus 1,409 municipalities). These two will be the labels modeled by the Random Forest Classifier. They correspond to a qualitative inquiry into the determinants of overall healthy firm and population growth. Their cross tabulation is shown in Figure 17a.

The continuous variables *dx* and *dptot* will be used for the quantitative analysis provided by the Random Forest Regressor. The scatterplot of these two variables is shown in Figure 17b. They are mapped in Figure 18. Recall that population growth can be both positive and negative and implicitly includes migration

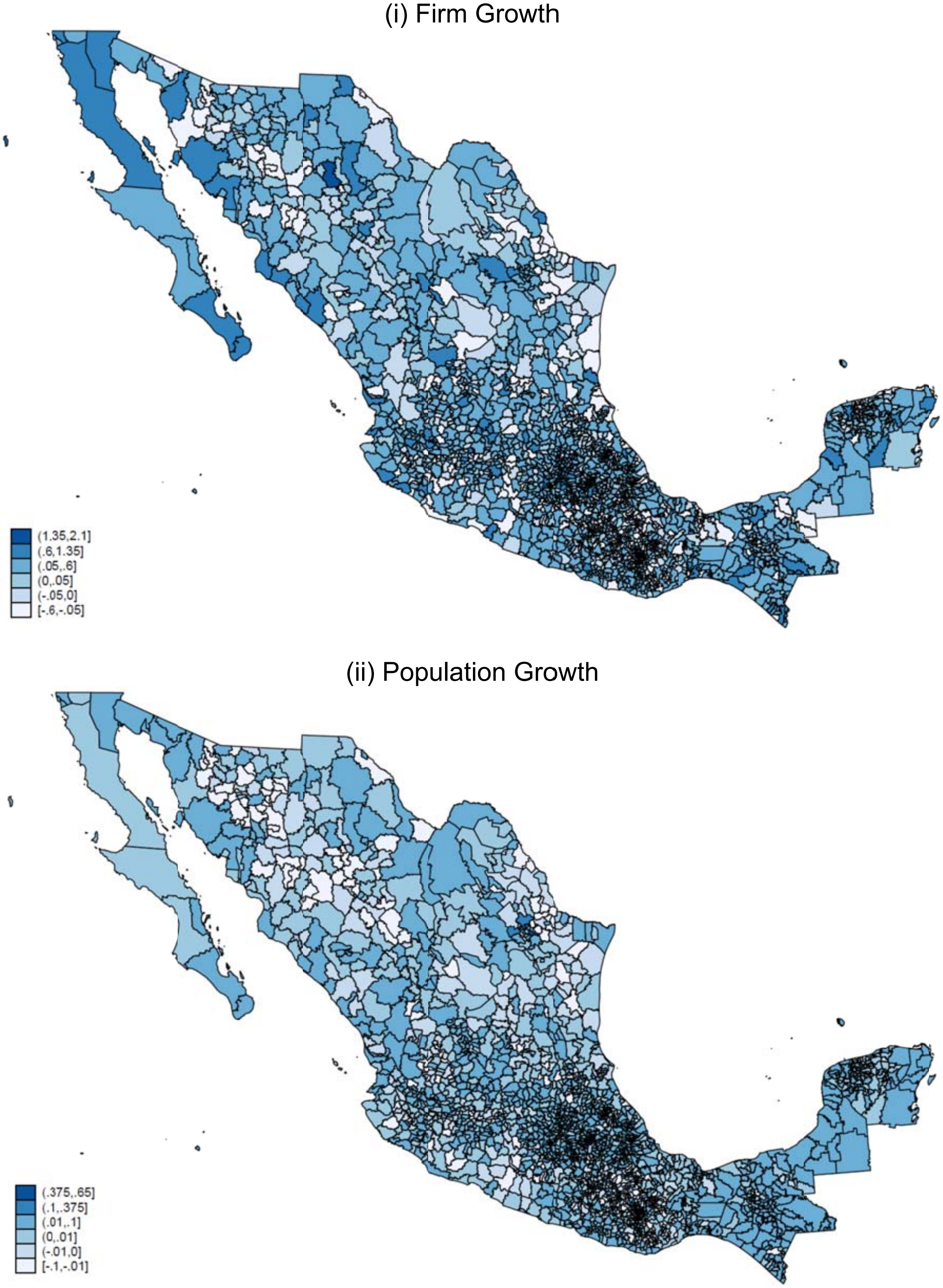
2.4. Features

In examining the features to be used, consider the following. Numbers of firms are observed at the municipal level. To use cross-validation for the learning algorithms, samples will be reduced to around 400 observations (out of the 2456 municipalities). This puts a limit to the number of features we should employ. For this initial analysis we keep to a one digit classification of the production sectors. Log total firm numbers (variable named *SE*) are also subcategorized according to their employment level category (variables *E1*, *E2, … E7*) and their production sector (*S1*, *S2, … S9*), providing a total of 17 features. Now, firms in different municipalities rely on their state context for both inputs and markets, and therefore we add to these features the corresponding 17 indicators constructed analogously at the state level (*SEent, E1ent*, … *S9ent*). To these local features we add log population, *lptot10*, migration (proportion born in state, *nacent10*, proportion living in the US, *viveu10*), marginalization, *im10*, and log population also at the state level, *lptot10ent* (5 variables). Finally, we include the three PC firm and population growth variables *x*, *y*, *z*, *xpop*, *ypop*, and the population ranking variable *ptot10cum* mentioned in the introduction (5 variables). This makes for a total of 44 *local* variables.

However, economic competition does not only occur at a local level. In effect any given municipality competes with all other municipalities in the country for firm and population growth, which are interlinked. Therefore we construct indicators based on rankings of the principal components *x*, *y*, *z*, *xpop*, and *ypop*, and on the population ranking variable *ptot10cum*. I construct seven indicators for each of these six variables *v*, named *vmi* for *i* = *–3*, *–2*, *–1*, and *vi* for *i* = *0*, *1*, *2*, *3*. The definition is as follows. Let *wmi* (alternatively *wi*) be the proportion of municipalities whose *v* value is lower than $v - 0.05 \times i$ (alternatively with a plus sign). Define *v0* = *w0*, and for the remaining *i*, *vi* = *wi – w0*, *vmi* = *w0 – wmi*. These variables define the proportion of municipalities whose ranks lie in successive value intervals above or below *v*. These define *competition corridors* between municipalities regarding feature *v*. All in all this adds 42 *nonlocal* features.

Armed with our 84 features for 2456 municipalities, we now proceed to apply the Random Forest Classifier for a qualitative analysis of the indicators of healthy firm and population growth ($x \geq 0.05$, $dptot \geq 0.01$) and the Random Forest Regressor for a quantitative analysis of *x* and *dptot*.

Figure 18. Maps of Mexican municipalities showing Firm and Population Growth.



(i) Firm Growth

(ii) Population Growth

3. Application of Random Forest Classifier and Regressor

3.1. Specifying the parameters

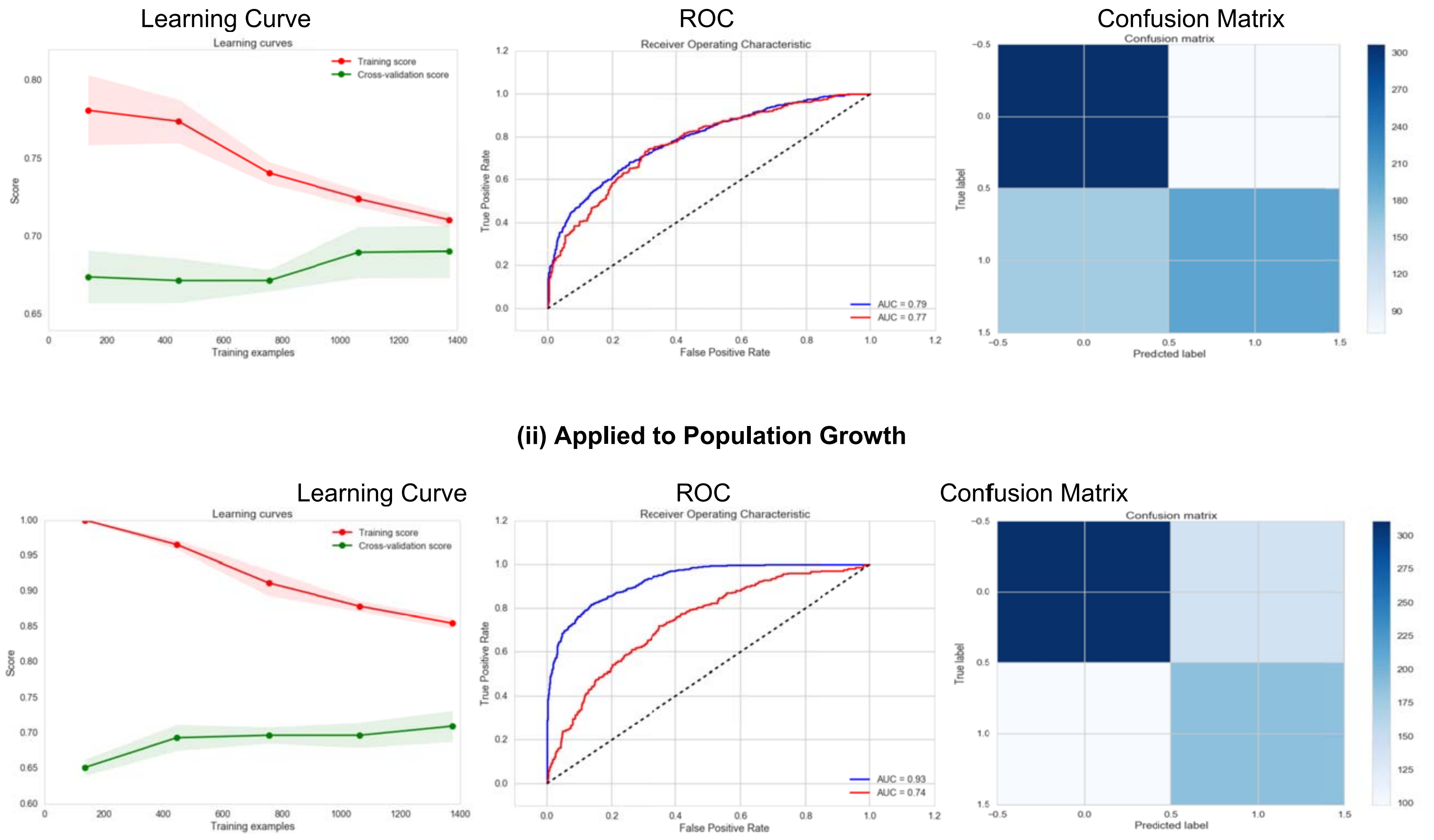
The first step in applying the Random Forest (RF) Classifier and Regressor is selecting the maximum depth of the decision trees and the number of trees. To do this we performed a grid search in these parameters. Now, economic growth in general and municipal firm and population growth in particular are quite noisy indicators. One way of stating this is that the predictable part of these indicators, from a several-year-perspective, only represents a certain portion of these indicators. Therefore measures such as accuracy or R^2 are somewhat week, since they are considerably affected by the unpredictable component of growth. This means that the results of any single grid search are somewhat random. This favored selecting as large a number of trees as was practical. 2,000 was too time consuming on a laptop (particularly for the RF Regressor) so 1,000 was selected. In the case of the RF Classifier, all of the grid searches that were observed for Firm Growth favored a maximum decision tree length of 3, that was selected as minimum. In the case of Population Growth, on the other hand, a higher maximum depth tended to be selected. In this case a ceiling was set at 7, which already models quite a bit of complexity. In the case of the RF Regressor, the grid searches that were observed favored increasing the maximum decision tree length up to depths of 11 that were offered to the algorithm, for both growth indicators. However, computing times were also impractical, so I settled for maximum of 5.

	Random Forest Classifier		Random Forest Regressor	
Maximum Decision Tree Depth	Firm Growth	Population Growth	Firm Growth	Population Growth
3	0.6926	0.6927	0.4123	0.2182
4	0.6810	0.7002	0.4160	0.2385
5	0.6810	0.7025	0.4179	0.2422
6	0.6822	0.7049		
7	0.6851	0.7089		
Accuracy or R^2	0.6879	0.6757	0.4962	0.2486
Test Score	0.7313	0.9362	0.7199	0.5453
Final Score	0.7055	0.8147	0.6065	0.6424

Table 1. Maximum decision tree depth grid search for the four Random Forest applications.

In each of these grid searches a test sample with 30% of the municipalities was first selected at random and set apart. Then cross-validation was applied to the selected training set (with 2012 data), on k-folds with k=5. After the maximum depth was selected, the RF was trained on the full training set, yielding an accuracy or R^2 (for the RF Classifier or Regressor; see Table 1), and then applied to the test sample, yielding the corresponding test score. At this stage further evaluation metrics were applied to each RF application. The learning curve was estimated, and in the case of the RF Classifier also the Receiver Operating Characteristic (ROC, which shows results were better than random), and the confusion matrix, see below. Finally, the RF was trained on the full feature dataset for 2012, yielding a final score. At this final stage, feature importance statistics were collected, which yield the main results of our analysis.

Figure 20. Evaluation Metrics for the Random Forest Classifier



(i) Applied to Firm Growth

Learning Curve

ROC

Confusion Matrix

(ii) Applied to Population Growth

Learning Curve

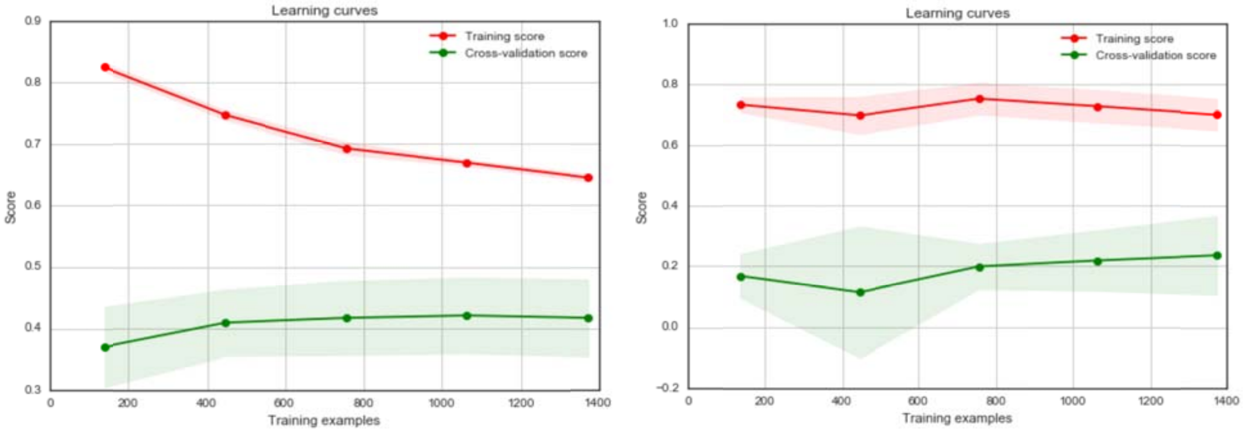
ROC

Confusion Matrix

Figure 20 shows for each application of the RF Classifier the Learning Curve, the ROC curve and the Confusion Matrix. Overall, the learning curve took a smaller sample for lower maximum decision tree depths. Correspondingly, the Area Under the Curve for the test results tended to be closer to the training results for a lower maximum curve. Finally, the confusion matrix was reasonably diagonal in both cases. However, in the case of Firm Growth, false predictions tended to be false negatives, while in the case of Population Growth they tended to be false positives.

Turning to the RF Regressor, Figure 21 shows its learning curves. Again these were slower for the population growth case. Note that municipal "population growth" refers at the same time to fertility, mortality and migration. This includes tendencies for the population to decrease as well as to increase (recall Figure 2), and implies that the population process may be more complex than the process of firm growth, or at least that we have included less relevant features about it, which is consistent with its analysis calling for a higher decision tree depth and a larger number of samples.

Figure 21. Learning Curves for the Random Forest Regressor


(i) Applied to Firm Growth

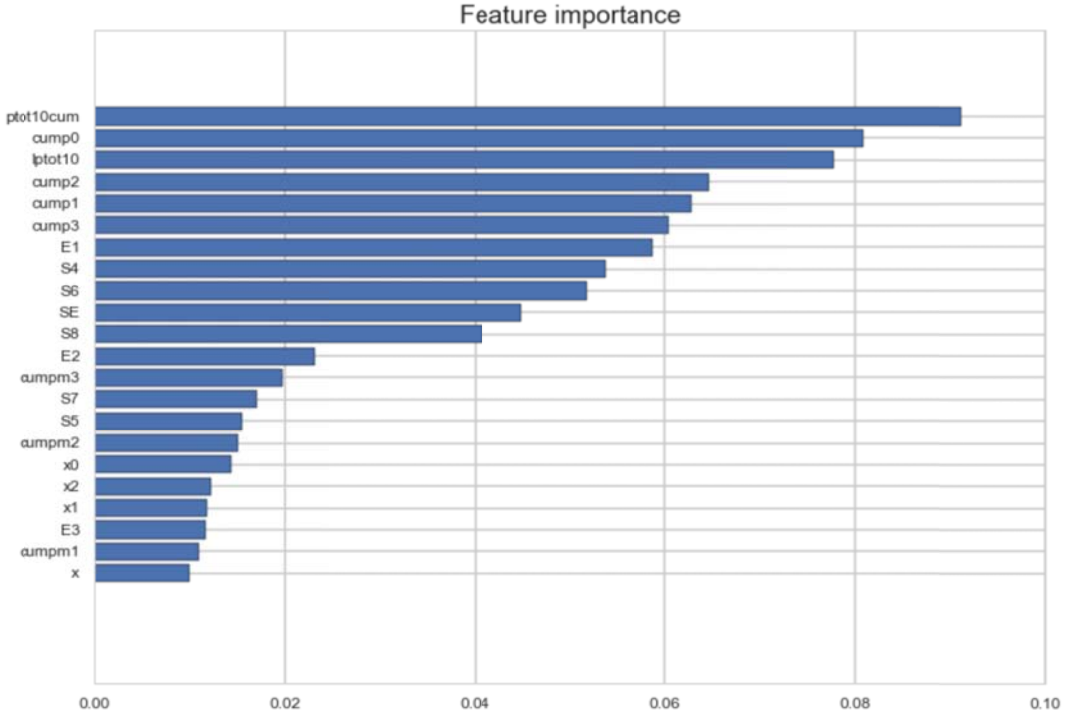
(ii) Applied to Population Growth

3.2. Feature importance for Firm Growth and Population Growth

The RF Classifier and the RF Regressor yield complex estimates for Firm and Population Growth, as functions of the features. Information on the importance of individual features is provided by the percentage of times that they intervene in the decision trees. Figure 22 provides graphs for these percentages, when they are higher than 1%. Figures 23 and 24 provide a synthesis of these results for Firm and Population Growth, discussed below.

Figure 22. Feature Importance

(i) Random Forest Classifier applied to Firm Growth



(i) Random Forest Classifier applied to PC Firm Population

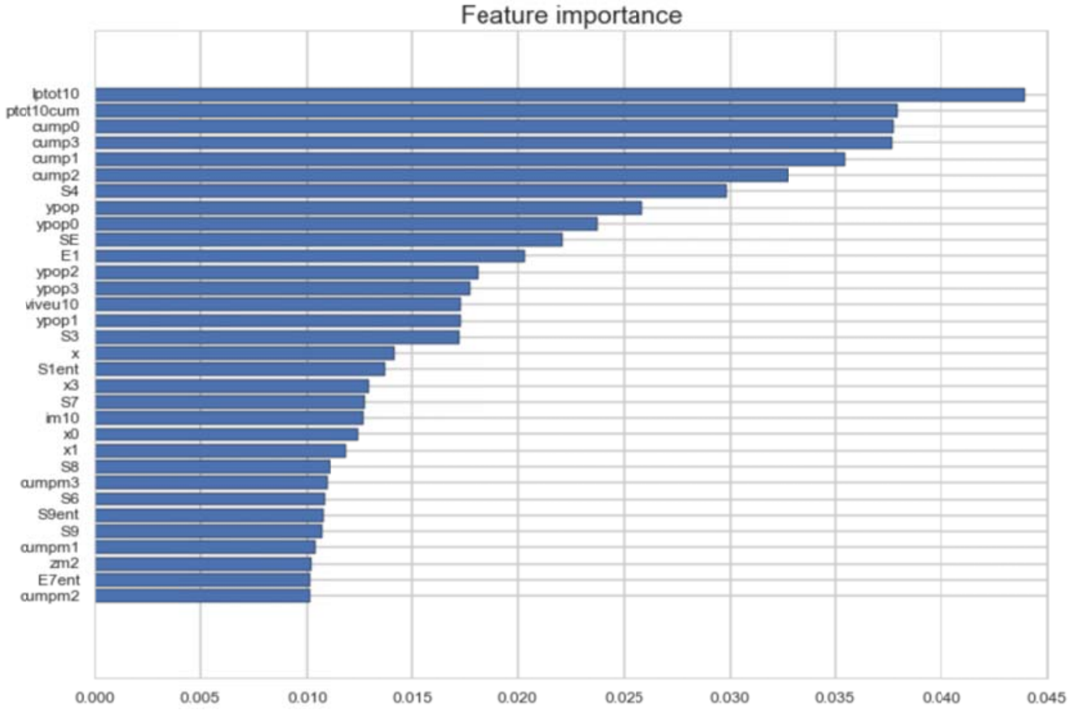
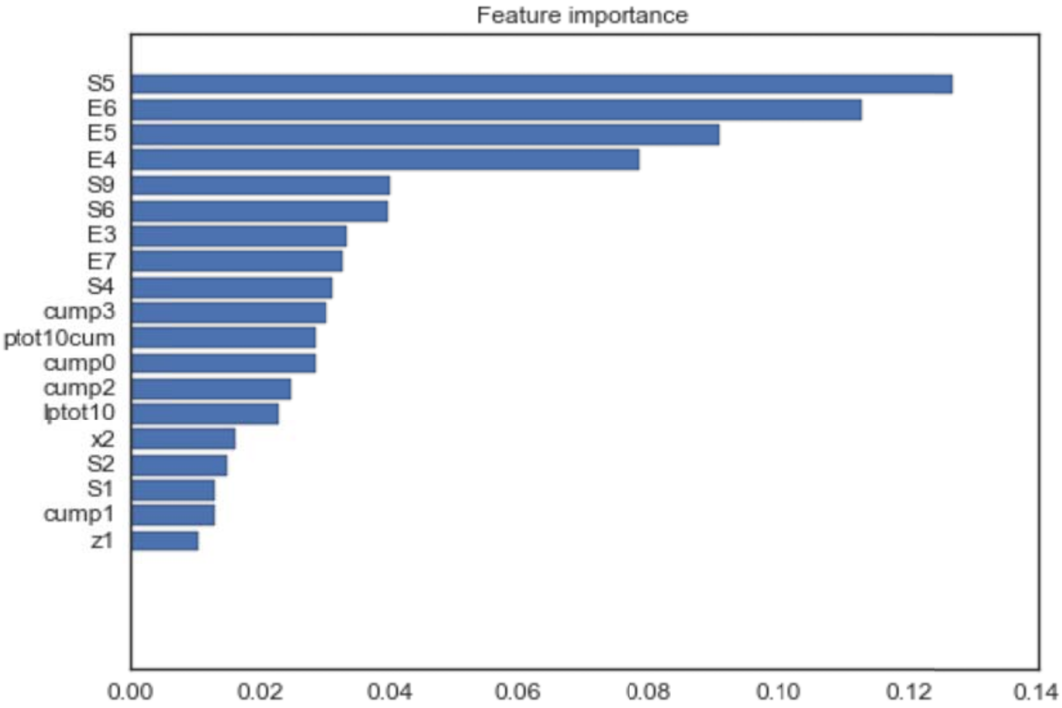
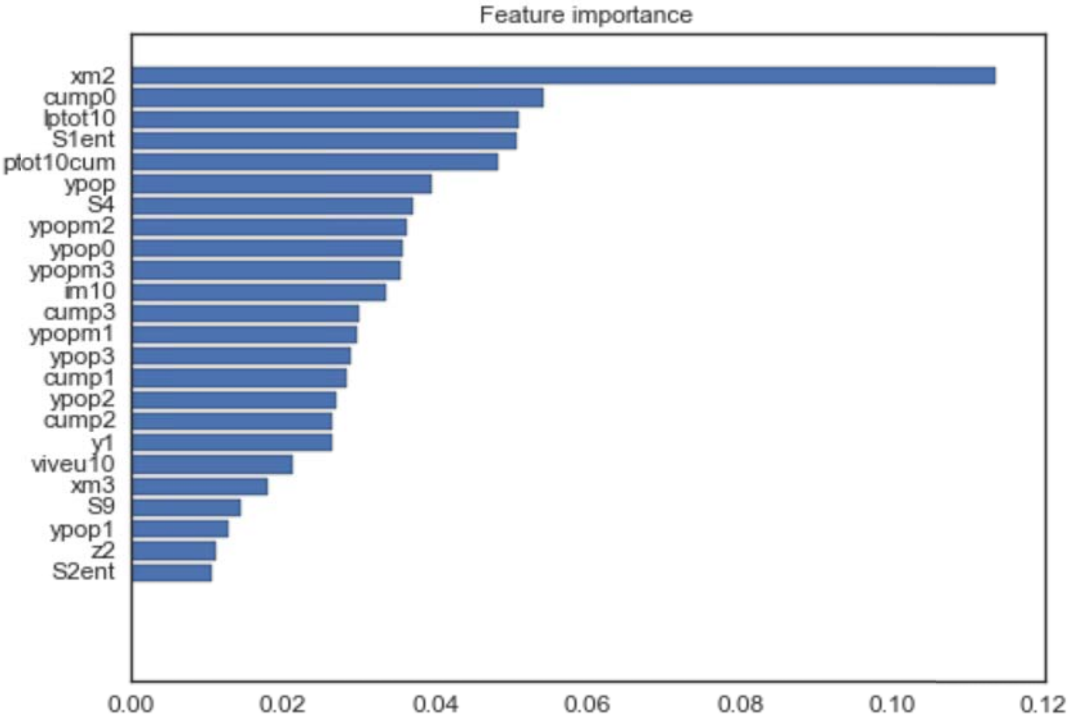


Figure 22. Feature Importance (Continued)

(iii) Random Forest Regressor applied to Firm Growth



(iv) Random Forest Regressor applied to Population Growth



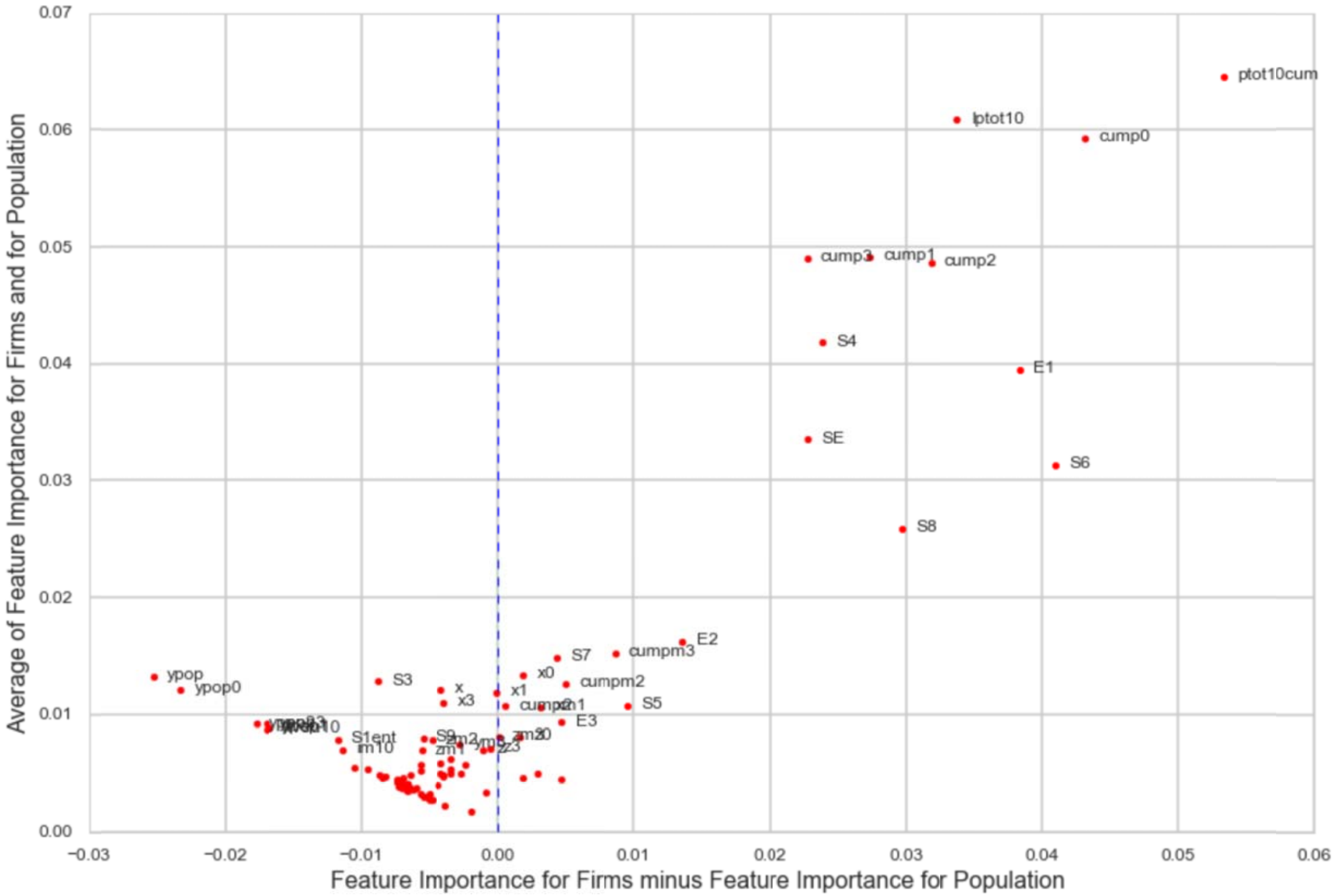


Figure 23. Scatter Plot of Average versus Difference in Feature Importances obtained by Random Forest Classifier for Firm Growth and Population Growth. To the right of the blue dashed line features are more important for Firm Growth than for Population Growth.

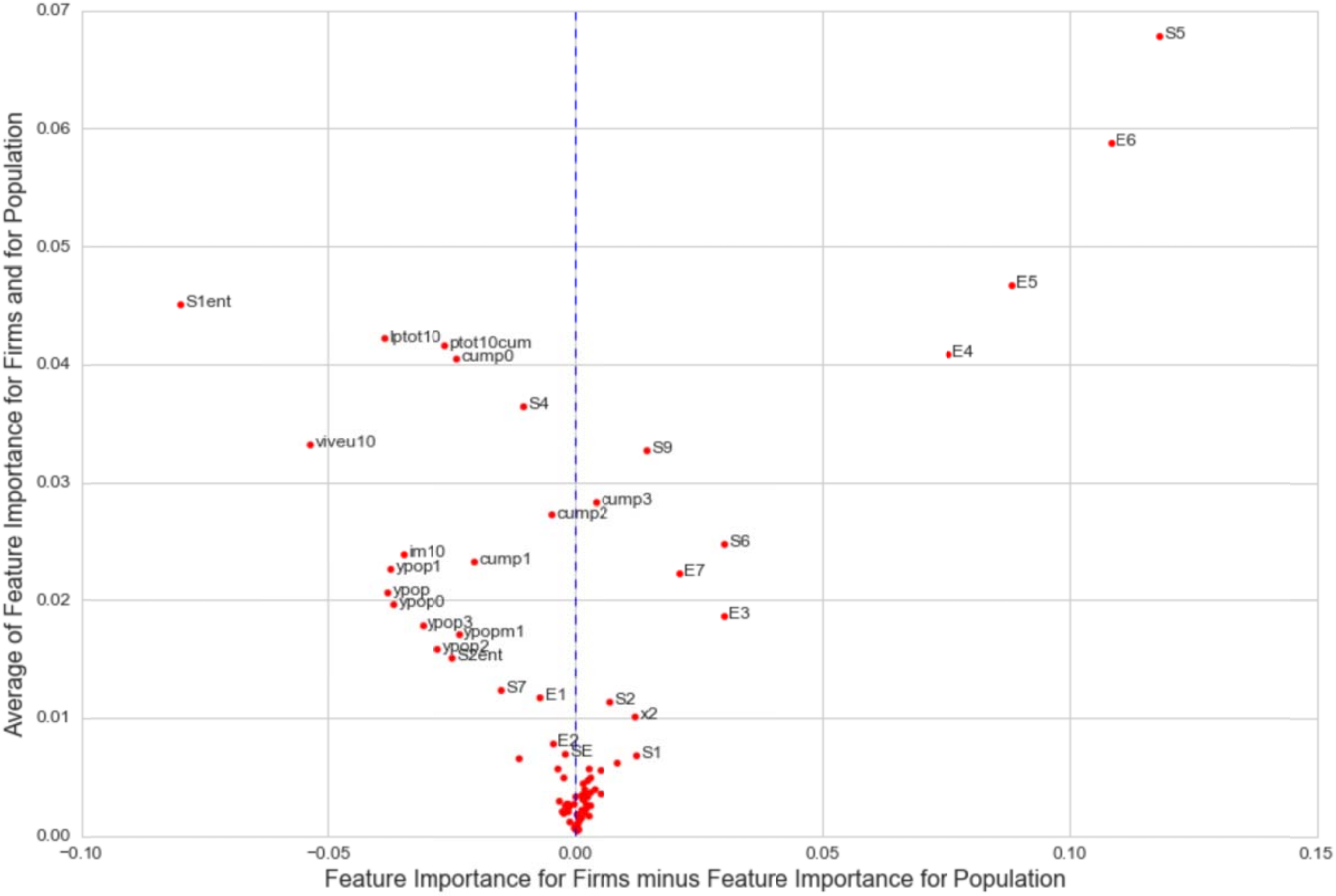


Figure 24. Scatter Plot of Average versus Difference in Feature Importances obtained by Random Forest Regressor for Firm Growth and Population Growth. To the right of the blue dashed line features are more important for Firm Growth than for Population Growth.

3.3. The combined process of firm and population growth

When the feature importances are combined in a single two dimensional plot, qualitative aspects of the combined dynamics of population and firm growth emerge. We already commented how Figure 12 reveals a process of transition.

A moment's thought shows that when there is migration some municipalities must operate as sources while others operate as sinks, even if these functions change location through time.

Moreover, as Figure 25 shows, the first principal component of firm numbers *x* is highly and almost linearly correlated with the Cumulative Population Rank. In addition, there is clearly a logic to the productive sectors and employment levels that are dominant at different levels of these variables.

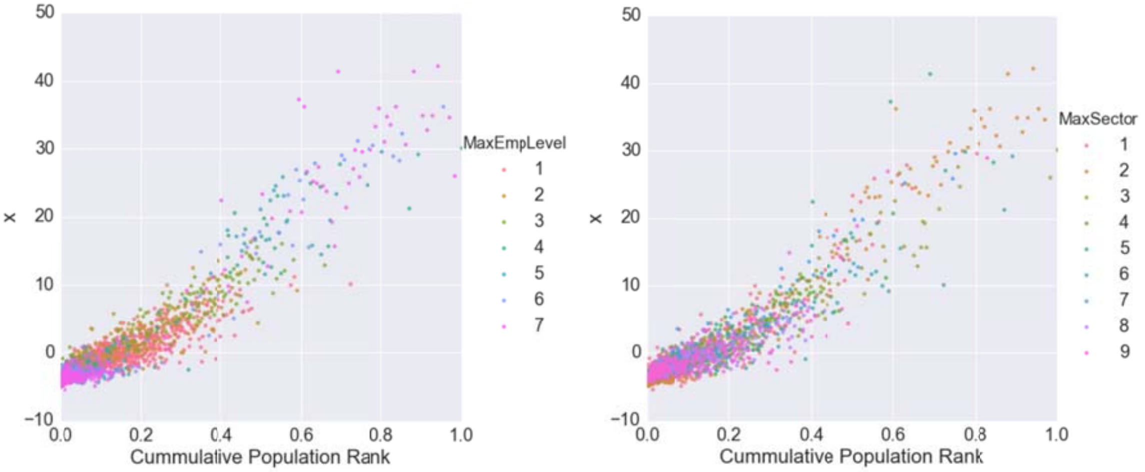


Figure 25. Scatterplots of the first principal component of firm numbers *x* with the Cumulative Population Rank. MaxEmpLevel is defined as the employment category that has highest value expressed in a normalized scale with mean 0 and standard deviation 1. Similarly MaxSector regarding production sectors.

Figures 23 and 24 scatterplot the average feature importance obtained by Firm and Population Growth on the y axis, against the difference in these importances on the x axis. Features that are more important for Firm Growth are on the right of the dashed blue line, while those that are more important for Population Growth are on the left. Figures 23 shows the result for the RF Classifier and Figure 24 for the RF Regressor. Both plots give evidence of a transition. There are few features that are both important and equally important for both processes. Figure 23 shows results for "healthy growth." In this case we are only concerned with this qualitative question. Five population indicators emerge as the indicators most important on average, and they are more important for firms. These include population, log population, cumulative population, and the percentages of municipalities in rings of cumulative population 5, 10 and 15 percent *above* the municipality's ranking. Then follows the construction production sector, number of small firms, total number of firms (these are in fact similar, because there are fewer large firms), then follow trade restaurants and hotels and finance, insurance and realtors, Other sectors can be observed in the figure.

On the left, affecting the population dynamics, are indicators related with the second component of population *ypop*. This component has higher marginalization, lower population, a higher proportion of people born in the state, and a higher proportion of people living in the US, consistently with being a migration source (Table 2). A look at the data shows municipalities where *ypop* is high occur in many Mexican states. It also features agricultural production at the state level, and the community and social sector, which is related to rural ejidos.

Concept	**Variable**	**xpop**	**ypop**
Marginalization	im10	-0.5957	0.0067
log population	lptot10	0.5182	-0.2555
Proportion born in the state	nacent10	-0.5771	0.1107
Live in the US	viveu10	0.2085	0.9604

Table 2. Coefficients of principal components of population

Turning to Figure 24, once we look at Firm and Population Growth quantitatively, the population features we had mentioned shift to the left, and instead the energy and water sectors becomes prominent, as well as employment levels [101,250], [51,100], [31,50], in that order. Agriculture at the state level and percentage having migrated to the US become prominent for Population Growth (or change), together with marginalization and other indicators mentioned before. The figures thus unequivocally detect both interlinkage and qualitative differences between Firm and Population Growth.

A more careful look at the dynamics could help to determine the types of production sectors that serve to promote firm growth over the development process intimated in Figure 25.

3.4. Prediction of 2016-2020 Growth

Finally, the trained RF Classifier and Regressor was applied to 2016 data, obtaining predictions for 2016-2020 municipal Firm Growth and Population Growth.

Population and marginalization were available for 2015, but not the proportion of population born in the state or living in the US, so for these two variables we used the 2010 values.

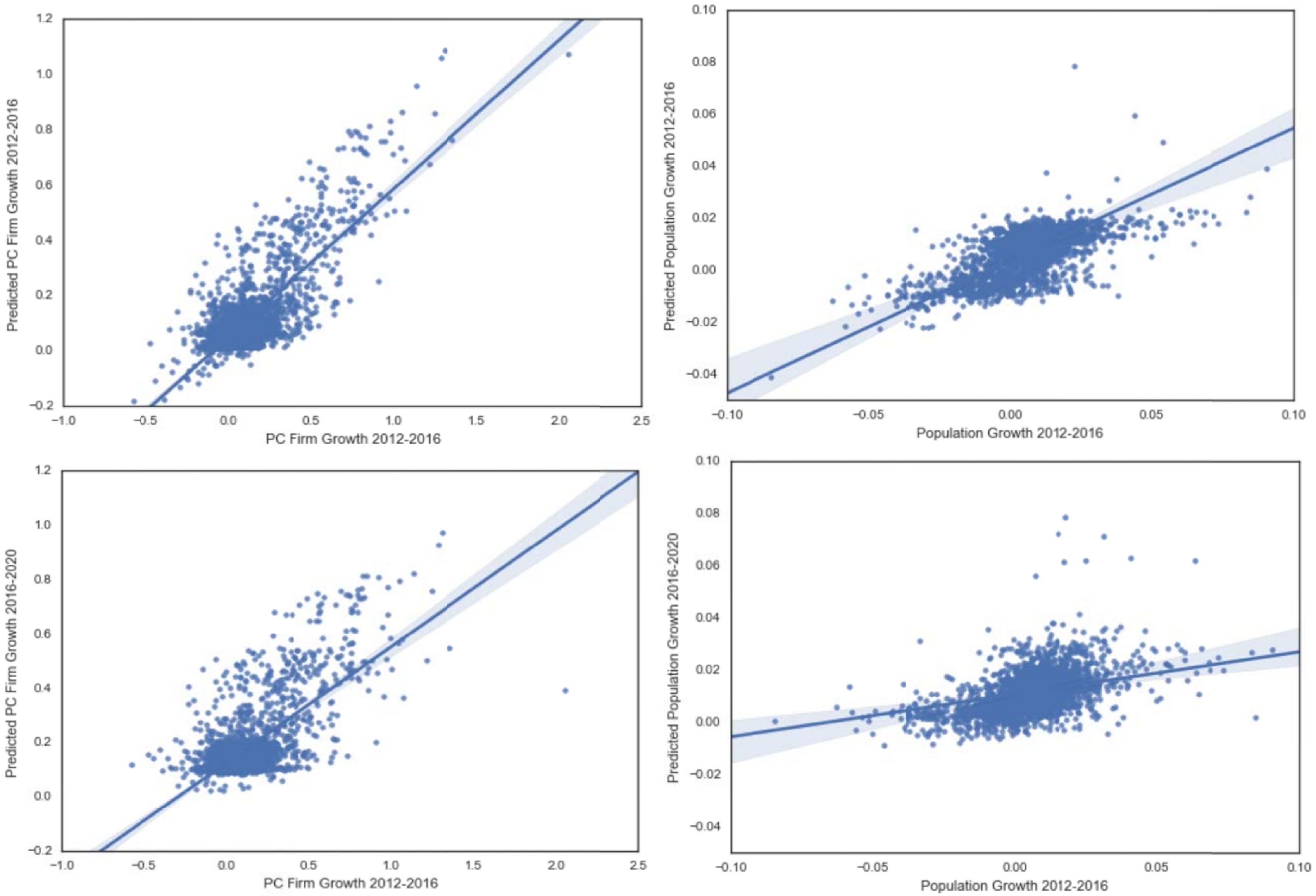


Figure 26. Scatter Plot of predicted municipal Firm and Population Growth for 2012-2016 and 2016-2020 vs actual growth over the period 2012-2016 (some outliers not shown).

Figure 26 shows a scatter plot of predicted municipal Firm and Population Growth for 2012-2016 and 2016-2020 (by the RF Regressor) vs actual growth over the period 2012-2016vs actual growth over the period 2012-2016. Recall that sometimes municipalities grow exceptionally when some state or national project is situated in them. Therefore we do not expect extreme outlier growth behavior to continue, something that is confirmed. We exclude some of these outliers from the graphs mainly to get a more detailed view of the overall process in the main cloud of municipalities. In both Firm and Population growth a trending process of continuing growth is present in significant parts of the cloud.

We can also see the results of the RF Classifier in Table 3.

firmgrowthclasspred1620	**0**	**1**
firmgrowthclasspred		
0	1063	366
1	0	1026

popgrowthclasspred1620	**0**	**1**
popgrowthclasspred		
0	1136	274
1	58	987

Table 3. Firm and Population Growth Prediction in 2012-2016 versus 2016-2020. This shows the behavior of the predictable rather than unexpected component of growth. Once municipalities are growing, they are mostly expected to keep growing.

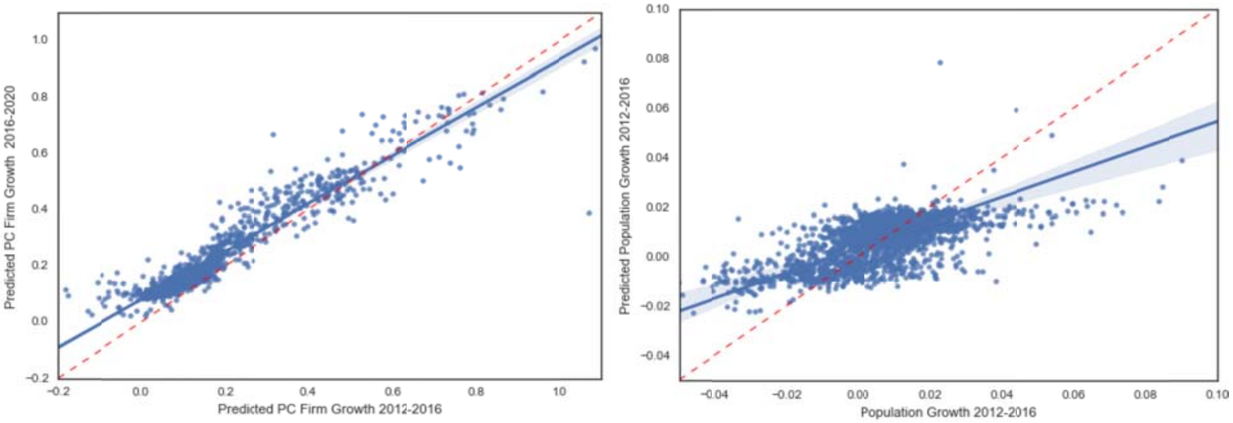


Figure 27. Scatterplot of predicted growth in 2012-2016 vs predicted growth for 2016-2020, according to the RF Regressor. A dashed 45° line is plotted in red for comparison.

We can also compare predicted growths for 2012-2016 and 2016-2020 for the RF Regressor (Figure 27). The predictable rate of growth is very persistent in the case of firms. Comparison of the OLS approximation to the 45° line shows that it has a significantly smaller slope. This implies that currently, according to the estimates the predictable rate of Firm Growth converges in the long-run to a value of about 60%, clearly a transitional rate of growth. Similarly the Population Growth currently converges to about 1.5% per year.

4. Discussion

4.1. Implications of the results

Application of the machine learning techniques, in particular the Random Forest Classifier and Regressor, have clearly made a difference in understanding the interaction between Firm Growth and Population Growth/Migration.

In the introductory section, it was very hard to find a clear interpretation of how these processes interact. The first approach was in fact to try to understand Firm Growth on its own. The population data was brought in mainly as a complement. As we conducted the data wrangling, slowly the interrelationships emerged, until we actually gave the Capstone Project the title “Growth of Firms and Migration in Mexico: Twin Processes.”

This first understanding was emphasizing parallel growth in firms and population, reflecting that people go after firms for employment and to purchase goods, and firms go after firms and people, seeking labor, inputs, and customers.

However, the results also underline the dual phenomenon that occurs in municipalites from which people migrate. Feature importance can be plotted to show that this constitutes a distinct process. This also implies that when the RF estimators are applied to population, their application requires more complexity (deeper decision trees), with a slower process of learning.

Thus the stylized facts that emerge for the process of Firm Growth and Population Growth/Migration are the following.

1) Overall, large firm numbers occur in municipalities with large populations
2) These municipalities also tend to grow more rapidly.
3) Nevertheless, there is a tendency for middle municipalities to grow a little faster.
4) On the other hand there is a set of municipalities which tend to be sources of migration, for which firm growth is quite different. Here instead of complementarity between the population process and firm growth, we have the opposite.

A further exploration of these complementary and dual processes needs to consider more data on the population process, such as fertility and mortality. We also have not included other data such as local government and local geography.

4.2. Policy Recommendations

In this first approach I have concentrated on the bird's eye viewpoint on firm growth in Mexico. Below I note that in fact the program is ready to use to consider the growth rates of other, more detailed indicators.

Policies for economic growth in Mexico should not only concentrate on firms, so to speak, for instance through economic policies on competition, technology, finance, education, and so on. Mexico's development continues to occur through a transition that includes migration tending to concentrate the population. Migration is an integral component of the current stage of Mexico's development.

With hindsight, it is no surprise that the construction sector is so prominent (see Figure 3). Indeed many migrant workers first become construction workers when they move.

Migration is in fact a very costly process in which people have to find work, build homes, find schools and so on. This can constitute a bottleneck for development. Whole cities have to be built and require services that are provided by the state. Up to now government responds when the problems are already there. Planning urbanization in a flexible format that nevertheless provides for the necessary services and facilitates the necessary investments can be a way to streamline development in Mexico, and to ease what tends to be a painful process of migration marked by unemployment, homelessness and poverty.

Since much of what firms decide and do is taken care of by the private sector, I concentrate here on the role of the public sector in Mexico. From this bird's eye view there emerges one main policy recommendation:

An integral part of development policy in Mexico must be facilitating and planning ahead for the movement of the population and urbanization, in a flexible but effective format.

A second insight gained from the data wrangling and from the analysis is that the process of firm and population growth is marked by complexity. This means that different detail and process emerges at different levels of scale and specificity. It is enough to recall Figures 5, 6, 7, 8, 9, and 11, and also the many unlabeled dots in Figures 23 and 24, which express feature importance in specific municipalities. This means that there must be a place for coordinated mid-level and local policies.

In addition, the importance of features *cump0*, *cump1*, *cump2*, *cump3* in Figures 23 and 24, as well as *ypopm1*, *ypop0*, *ypop1*, *ypop3* in Figure 24, underlines the importance of competition between municipalities. It can therefore be ventured that:

A competitive context for public services can be defined across municipalities that can help to make both public service and the growth process more efficient.

4.3. Further applications of the project

Migration only marks one aspect of the transition. Firm growth is marked by a series of stages in municipalities at different levels of development that need to be attended and can also be studied with the program I have written. All that needs to be done is to change the labels. A set of 63 growth rates has already been constructed from the data, corresponding to each production sector in each given employment category. The growth rates of the two other principal components of Firm Numbers have also been included.

Thus my Capstone Project provides a program that can be used for a whole set of analyses. The features and labels can be extended to allow the examination of firm growth for specific sectors at any level of the six digit classification. While at this initial stage only the one digit classification of production sectors has been included in the estimations, now that the framework has been constructed, it would not be difficult first, two predict the growth rate of any production sector based on the current set of features, and second to include further details of the production sectors as features for estimating the expected growth of specific production sectors.