International Journal of Economics and Finance; Vol. 10, No. 10; 2018
ISSN 1916-971X E-ISSN 1916-9728
Published by Canadian Center of Science and Education

Efficient Urbanization for Mexican Development

David Mayer-Foulkes[1]

[1] Centro de Investigación y Docencia Económicas, Ciudad de México, México

Correspondence: David Mayer-Foulkes, Centro de Investigación y Docencia Económicas, Carretera México-Toluca 3655, Del. Alvaro Obregón, 01210 México CDMX, México. Tel: 52-55-5727-9800. E-mail: david.mayer@cide.edu

Received: August 9, 2018 Accepted: September 5, 2018 Online Published: September 10, 2018

doi:10.5539/ijef.v10n10p1 URL: https://doi.org/10.5539/ijef.v10n10p1

Abstract

By applying Data Science techniques we find strong evidence that urbanization plays a key role in the process of development in Mexico, This process necessarily involves government action and therefore must be the subject of policy. We suggest that there are ways of streamlining the government's role in providing the public goods of urbanization that can combine with and stimulate the competitive economic context.

We apply Data Science techniques including visualization of the full universe of the object of study, and application of the Random Forest Classifier and Regressor machine learning algorithms, to municipal firm number growth obtained from Mexico's full Directory of Economic Units for 2012 and 2016. These are aggregated at the municipal level by employment scales and one-digit production sectors, and combined with municipal demographic census data. Our visualization exercises also show that the dynamics of firm and population numbers is complex, such as in a changing fractal.

Keywords: machine learning, social policy, economic growth, combined micro and macro analysis, urbanization

1. Introduction

1.1 Using Machine Learning to Discern the Salient Processes of Economic Development

The question motivating this research is, what can we say about economic development in the Mexican Economy taking a fresh look using a big data approach? Can we take a bird's eye macroeconomic approach on economic growth which at the same time is cognizant of detailed microeconomic data on the economy across the country, to distinguish countrywide economic processes linked to growth?

The quintessential components of the economy are firms and their workers. Mexico has a full Directory of Economic Units that is available for the recent years 2012 and 2016, with information on firm location, activity, and employment scale. Of course, firms operate in a context. The smallest context that includes a rounded universe of economic, social and governance information in Mexico is the municipality. If we aggregate information on firms at the municipal level by classification and employment scale, the result is a summary of demand and supply, itself reflecting income distribution and population characteristics.

1.2 The Methodological Challenge

There is a double methodological challenge here. From the theoretical point of view, a series of economic processes interact and occur jointly when this range of scales is included. These include migration, demographic change, human capital accumulation, technological change, and institutional change, among others (Durlauf & Aghion, 2005). But our objective is not to check one particular theory. Instead it is to detect at least one ongoing, aggregate, significant aspect of development in Mexico as a whole. Can we discover in this fresh big data exploration, using information on firms and workers, some central characteristic of the development process in Mexico which plays an important role and deserves policy attention? And then, can our model lend credible insights?

1.3 The Incipient Use of Machine Learning in Economics

Before continuing, let us review current applications of machine learning in Economics, which offer tremendous new opportunities. Machine learning takes advantage of great amounts of data to use highly complex functions to approximate highly complex data patterns. At the same time, it has at its disposal visualization techniques that allow a close look at the data as a step towards bringing out its main features, understanding, and testing them

(Varian, 2014; Mullainathan & Spiess, 2017).

The literature on machine learning applications to economics is taking off rapidly. Initial articles introduced machine learning and its methodology in a general way and explained what applications there may be in economics (Green & Richards, 2016). Interest has grown rapidly, and now one can find books not only on the application of machine learning to economics but on the economic impact of the application of artificial intelligence in production. Agrawal, Gans, and Goldfarb (2018) introduce a book in preparation on "The Economics of Artificial Intelligence" reporting contributions to a conference with the same name.

At this initial stage in the development of the statistics of machine learning, some authors have stressed prediction as one of the strengths of machine learning for economics (Mullainathan & Spiess, 2017).

Machine learning prediction is now sought and applied in diverse areas of economics. Chakraborty and Joseph (2017) explore the application of machine learning at central banks, using diverse methods and comparing them with a range of standard methods. They find machine learning generally outperforms traditional modeling in prediction tasks, while open questions remain regarding causal inference. Papadimitriou et al. (2014) use Machine Learning to forecast recessions, based on a variety of short (treasury bills) and long term interest rates (bonds) for the period from 1976:Q3 to 2011:Q4 in conjunction with the real GDP. Gogas, Papadimitriou and Karagkiozis (2018) use machine learning predictions, showing these improve on linear regressions to test theoretical models.

The use in machine learning of functions with very general forms has also led to other applications. Gründler and Krieger (2018) use machine learning to develop complex indices such as a democracy index. Mareckova and Pohlmeier (2017) uses machine learning to construct new personality indices with long-term predictive power on the impact of early childhood circumstances on long-term earnings and unemployment.

As the following examples show, machine learning lends itself to many kinds of applications and we are likely to see more of them in the future. Samii, Paler, and Daly (2016) apply causal inference techniques in a machine learning context. Milgrom and Tadelis (2018) uses machine learning to analyze and improve market performance. Aufenanger (2017) uses machine learning to assist and improve experimental design. Andini et al. (2017) use machine learning to improve targeting in social policy. Renner and Scheidegger (2017) use machine learning to perform complex numerical calculations that can arise in Dynamic optimization.

1.4 Methodological Outline of Our Application

In our case, we use machine learning to gain a qualitative perspective on economic growth in Mexico, which can inform public policy. As we mentioned, from the econometric point of view, we are using techniques that invite the use of many variables but are somewhat statistically underdeveloped (e.g. Van der Laan & Rose, 2011). On the other hand, we draw from the strengths of this new methodological approach born in the new experiences of Data Science, a new relative of Econometrics. We take as our methodology its standard steps. We first visualize the data, trying to find the main relationships that characterize it. Next, we choose the main one or two "Labels," variables that we seek to predict or understand. Then we choose the main "Features," variables that can serve to distinctly describe the behavior of our "Labels." Finally, we apply an appropriate algorithm to furnish the model.

When we visualize the data, we are not looking at a sample. We are looking at the full set of firms in the full set of municipalities, and the actual growth in their numbers from 2012 to 2016. This means that we are looking, as far as this is possible, at facts. Adding municipal information to our data, we examine in this way that firm growth (short throughout for growth in firm numbers), migration, and specifically the interaction of firm and municipal population numbers, firm growth and population numbers, and firm numbers and population growth. The question is, are there approximate laws governing these relationships? We do find that, across employment scales, firm numbers are approximately exponential in population. We also find that the principal components of firm numbers (aggregated through employment scales and one-digit production sectors) are related to population numbers and the population ranking across municipalities. Finally, examining phase diagrams firm change and population change display complexity, in that looking closer uncovers further, unsuspected dynamic variety across production sectors and employment scales, as in a changing fractal.

Finally we focus on predicting firm and population growth. (Recall population growth is local and therefore also reflects migration.) This calls for a supervised machine learning application. We rely on Random Forests, machine learning techniques used as a Classifier and as a Regressor, since these deliver feature importance. The machine learning algorithms use as features firm numbers and demographic data, the features which have been found to simplify visualization, and features that represent competition between municipalities. The result is that population and firm growth share common main features driving them – municipal population and numbers of

similarly populated municipalities amongst them. The trained algorithms are used to predict 2016-2020 growth, displaying some of the dynamic features of current growth.

We find strong evidence in our visualization of the data for the almost fractal complexity of the economic process. Then, when we review the results of the machine learning algorithms, we find that population growth and firm number growth share many of the same predictors, especially population growth (reflecting migration) variables. This means that migration continues to coincide with and therefore to play a very important role in economic growth. In some ways this is not a very surprising result. Anyone visiting the Mexican countryside knows that there are migrant workers everywhere from all parts of the country seeking work. This means that on the margin workers relocate to where economic growth occurs. On the other hand, academic exploration of economic growth in Mexico has focused much more on other determinants of economic growth, such as human capital, technological change, institutions, and convergence. How often is the rural to urban migratory process first pointed out by Harris and Todaro (1970) referred to as not having fully played out in Mexico?

The conclusion that urbanization plays an important role in the process of development in Mexico is significant. The reason is that urbanization necessarily involves public action and therefore must be the subject of policy. We are not arguing that there is a one-way causality from urbanization to growth. It is clear that causality can run both ways and also that locally this process can exhaust itself. However, it is also clear that when urbanization and economic growth happen together, the efficiency of the process depends on the efficiency of both its public and private components. Finally, it is necessary to observe that some of the underlying economies of scale might make migration and inevitable component of the economic process.

We suggest below that there are ways of streamlining the government's role in providing the public goods of urbanization that combine with the competitive context.

The rest of the paper follows the methodology we have described, looking at the data, preparing it for an application of the Random Forest Classifier and Regressor, and discussing the results.

2. Preliminary Descriptive Analysis: Visualization of the Data

The search for economic opportunity drives both the creation of new firms and municipal population change 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.

2.1 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 a municipality's rank as the proportion of people in Mexico that live in smaller municipalities (e.g. Rowland, 2001). The long history of migration and the concentration of population that began with the shift from agriculture to industry can be observed 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 rank (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 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, 1,164, 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.

2.2 Firm Numbers Growth

The DENUE data classifies firms into nine main production sectors (at the 1 digit level). The sector with the most firms was construction (see 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 2012–2016. These were followed by communitarian 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.

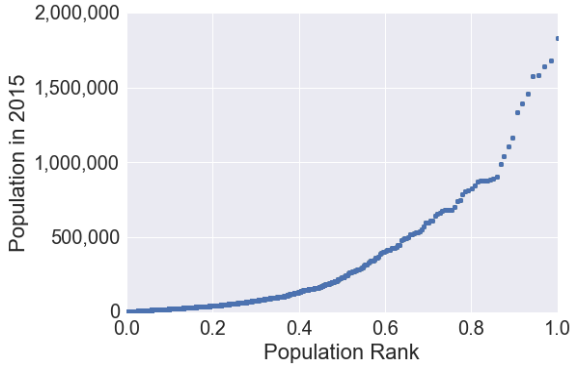


Figure 1. Municipal population versus municipal rank (proportion of people living in smaller municipalities)

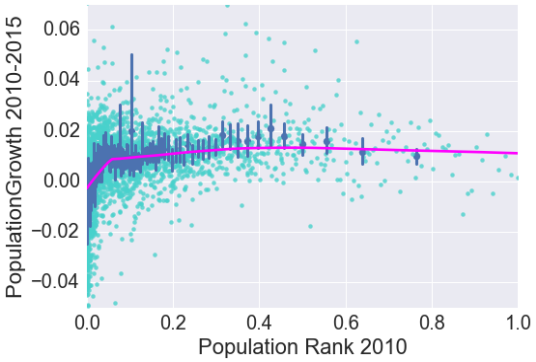


Figure 2. Municipal population growth 2010-2015 versus population rank

Note. 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.

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

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

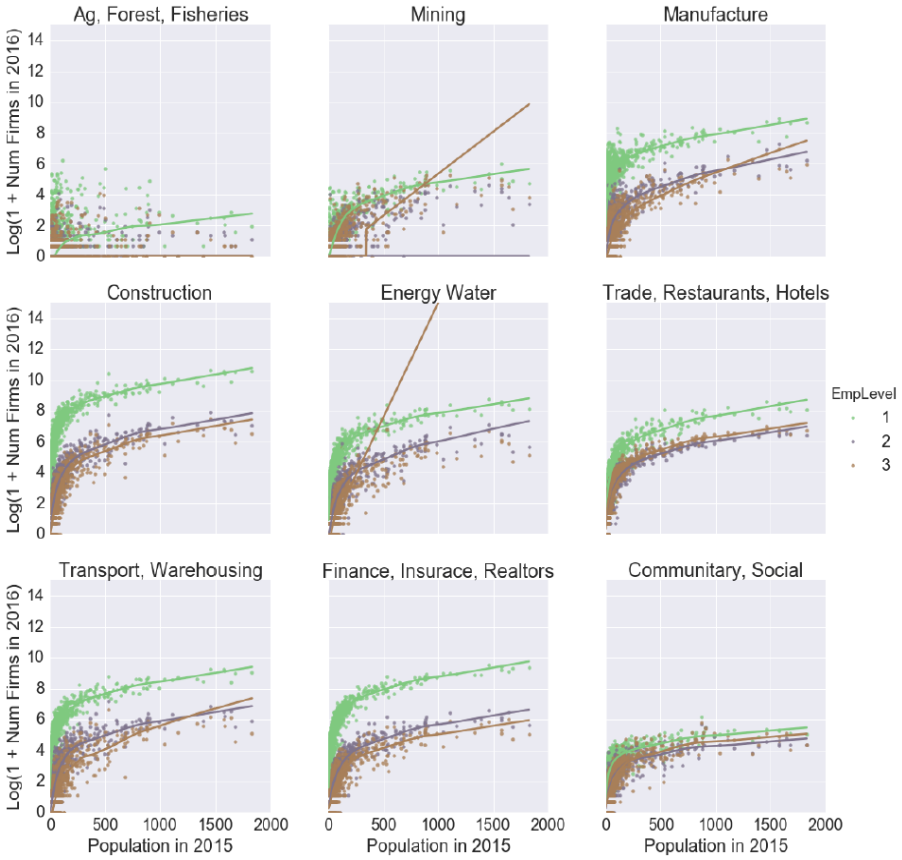


Figure 5. log(1 + Number of Firms) against Municipal Population (in thousands)

Note. The 1 is added to avoid log(0). Each subplot represents a sector of production, and shows a scatterplot for employment levels [1, 5], [6, 10] and [11, 30], as well as a locally weighted regression plot.

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%.

2.3 Interaction of Firm and Population Numbers

There are several general questions on how firm numbers relate to population numbers (Krugman, 1991; Durlauf & Aghion, 2005). First, is there some "law" relating these quantities? Second, when the population moves and 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.

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 no firms are present of some given 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. This is verified to a 5% confidence level in several of the plots. In fact, the exponential coefficient tends to be larger for smaller firms. Perhaps larger firms in fact eschew high population areas.

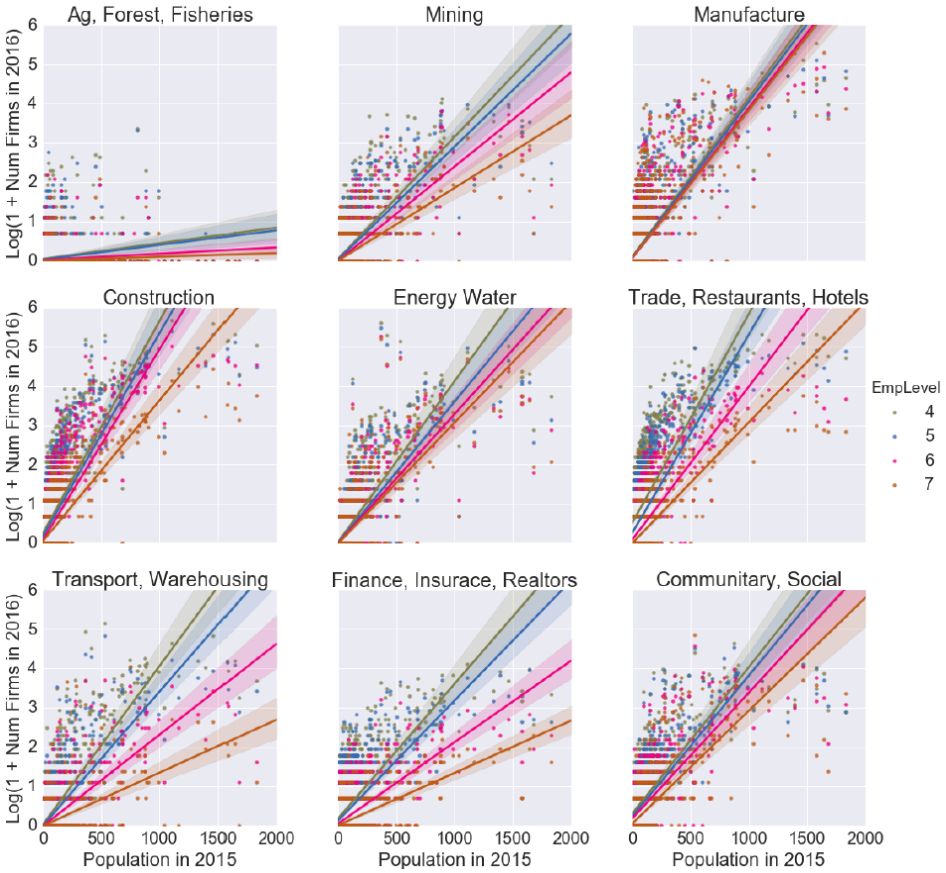


Figure 6. log(1 + Number of Firms) against municipal population (in thousands)

Note. The 1 is added to avoid log(0). Each subplot represents a sector of production, and shows a scatterplot for employment levels [31, 50], [51, 100], [101, 250], [251, +], together with a linear regression plot. There is not enough data for a locally weighted linear regression.

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

2.4 Firm Number and Population Growth

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

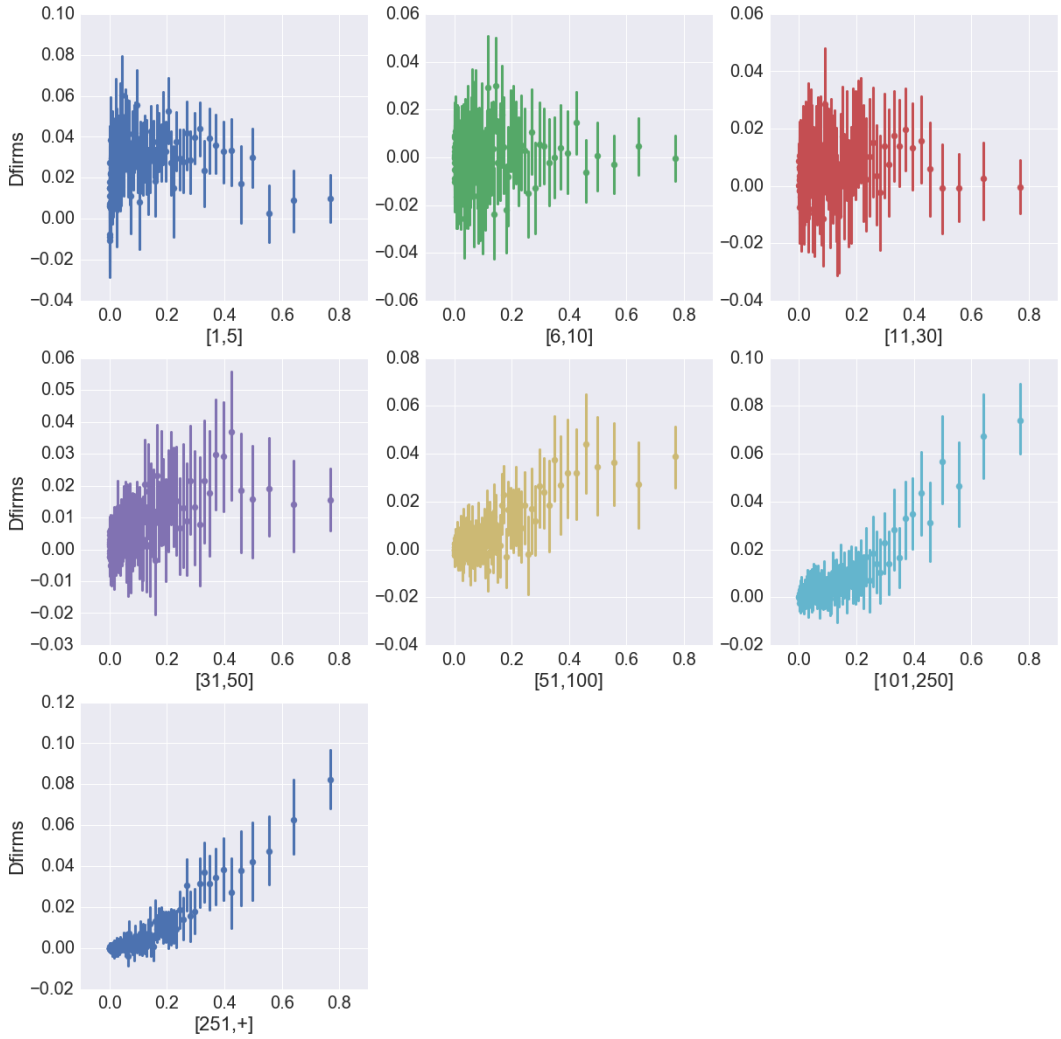


Figure 7. Growth in firm numbers by employment level (2012-2016)

Note. The graph shows a binned scatterplot of growth against population rank: growth means for groups of 25 municipalities, with 95% confidence intervals.

Now consider firms by production sectors (Figure 8). Similar inverted 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 inverted U curve only appearing at a population rank of 0.1 or even 0.2. Manufacturing shows the inverted 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.

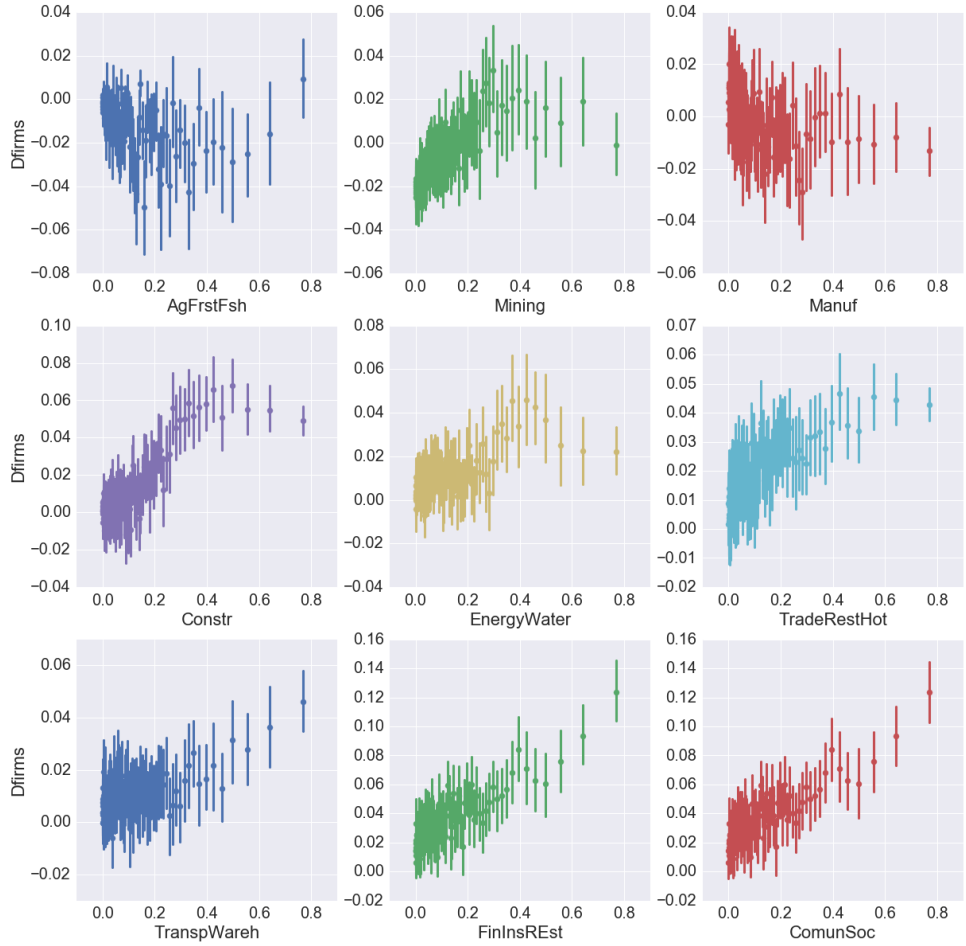


Figure 8. Growth in firm numbers by production sector (2012-2016)

Note. The graph shows a binned scatterplot of growth against population: growth means for groups of 25 municipalities, with 95% confidence intervals.

2.5 Interplay between Firm Numbers and Population Growth

Are any regularities apparent when we consider firm number growth and population growth? (Recall that population growth is the combination of migration, fertility and mortality.) One way to examine this is considering the phase space for the dynamics between these two variables. This is a two dimensional plot with population rank along the x axis and firm numbers along the y axis, which displays arrows representing the changes in these variables as a vector. Municipal rank for 2010 runs on the [0, 1] interval. Log (1 + Number of Firms) is scaled to [0, 1] for each of 2012, 2016. Municipal population change is the difference in population ranks between 2010 and 2015. Firm number growth is the rate of change of the two normalized Log (1 + Number of Firms) 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 therefore a binned phase diagram, similar to a binned scatterplot.

Figure 9. Each panel's horizontal and vertical axes are municipal population rank 2010 and log(1 + NumFirms2012), normalized to the [0, 1] range

Note. 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 in a PDF.

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 paradigmatic or 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.

2.6 Complexity of Firm Change and Population Change

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 use a stronger lens we again find a diversity of phenomena. 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. This size firm is closely linked with the livelihood of many people (indeed it can represent a disguised form of unemployment), and therefore with population growth. However, when we also view the scatter plot (Figure 10), this shows 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 development. Something similar occurs with each of the subplots in Figure 7 ranging across employment levels. However, the municipal density grows faster at higher population rankings, consistently with the idea that the observation that the maximum of the inverted U curve moves to the right.

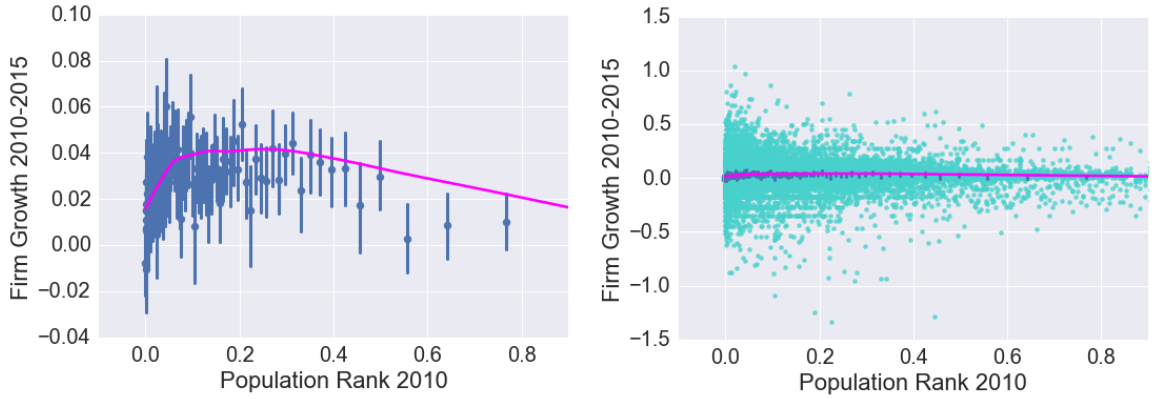


Figure 10. Both panels show growth in firm numbers (2012-2016) for employment level [1, 5], in different scales

Note. 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 scatterplots does not change as much across production sectors (Figure 8).

Now let us expand Figures 7 and 8 to consider all possible combinations of production sector and employment level. The results (Figure 11) 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, that are lost in the average measures considered before.

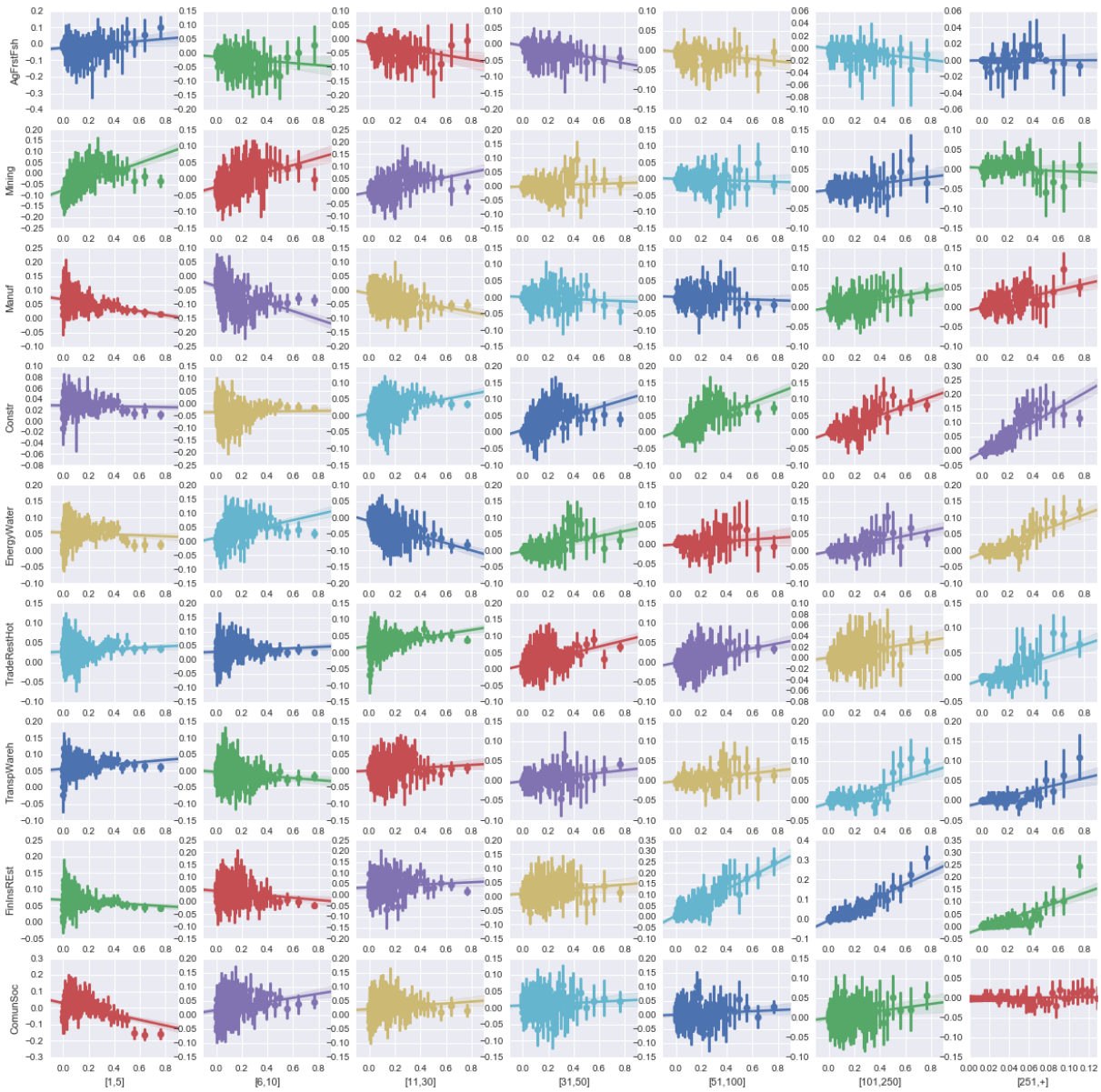


Figure 11. Growth in firm numbers (2012-2016) by production sector and employment level

Note. Each panel shows a binned scatterplot of growth against normalized 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.

3. Method: Preparing Data and Parameters for the Supervised Machine Learning Application

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 process?

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 means 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.

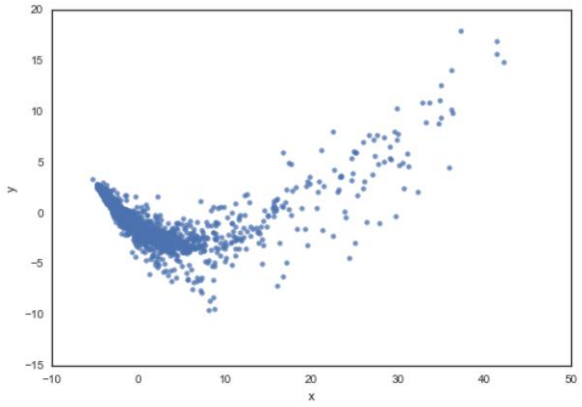


Figure 12. First two principal components x and y of 63 log firm number indicators

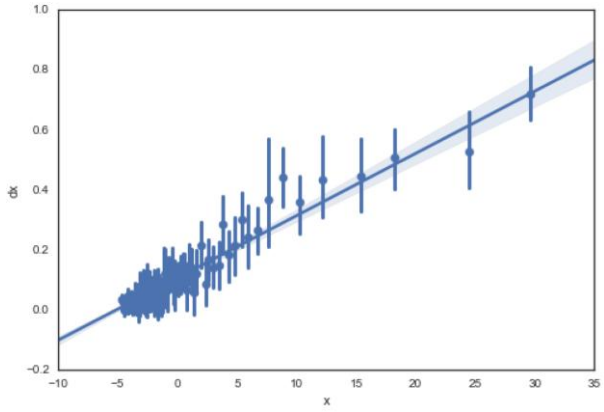


Figure 13. Binned scatterplot for mean municipal population of dx versus x (see text) for groups of 25 municipalities

Note. 95% confidence intervals shown.

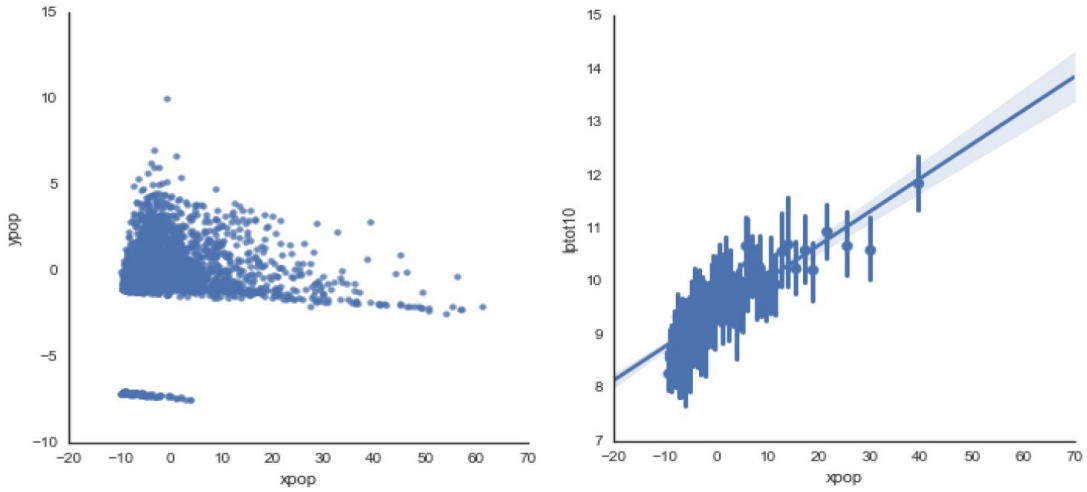


Figure 14. Scatterplots of ypop versus xpop and log population versus xpop

Principal components for population numbers

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. 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. For these we use instead the 2010 indicators. Since each is a proportion, only a small error is introduced.

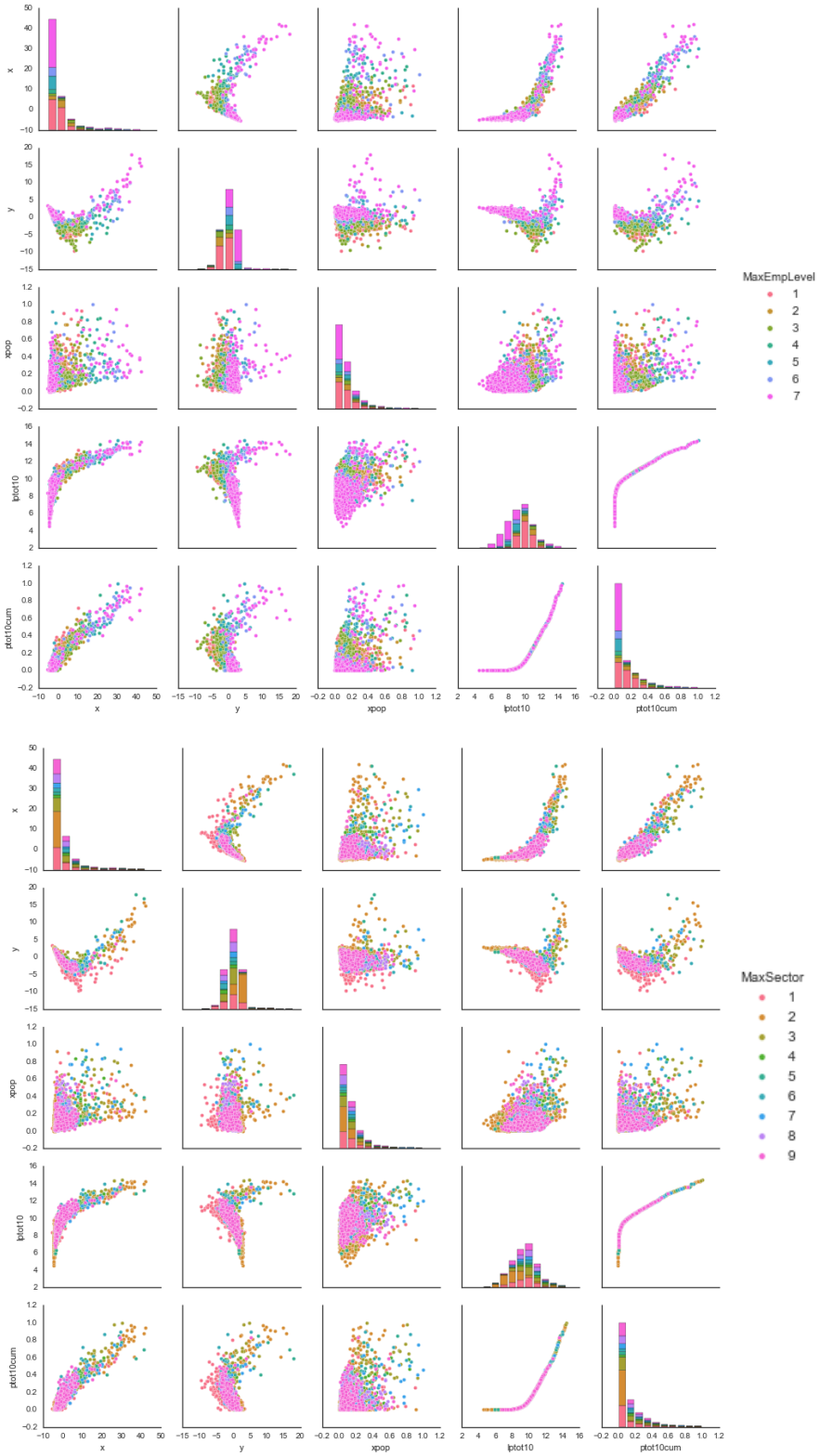


Figure 15. Matrix of scatterplots and histograms for the variables x, y, xpop, log population lptot10, and population rank ptot10cum for 2010

3.1 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 first 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, for simplicity of language I 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 will also be supported by other figures below. For the 63 corresponding number of firm variables for 2016, we use the 2012 scaling parameters to construct corresponding features for 2016, and define the first three components using the parameters of the 2012 decomposition. 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. This is remarkably linear. The principal component analysis has successfully removed the data's bias towards small firms that follows simply from the number of data points.

A similar analysis 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.

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.

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

Table 1. Cross tabulation of municipal firm growth dx ≥ 0.05 and population growth dptot ≥ 0.01

	Population Growth:	Yes	No
First PC Firm Growth	Yes	654	555
	No	392	854

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 17. They are shown in maps of Mexico in Figure 18. Recall that population growth can be both positive and negative and implicitly includes migration.

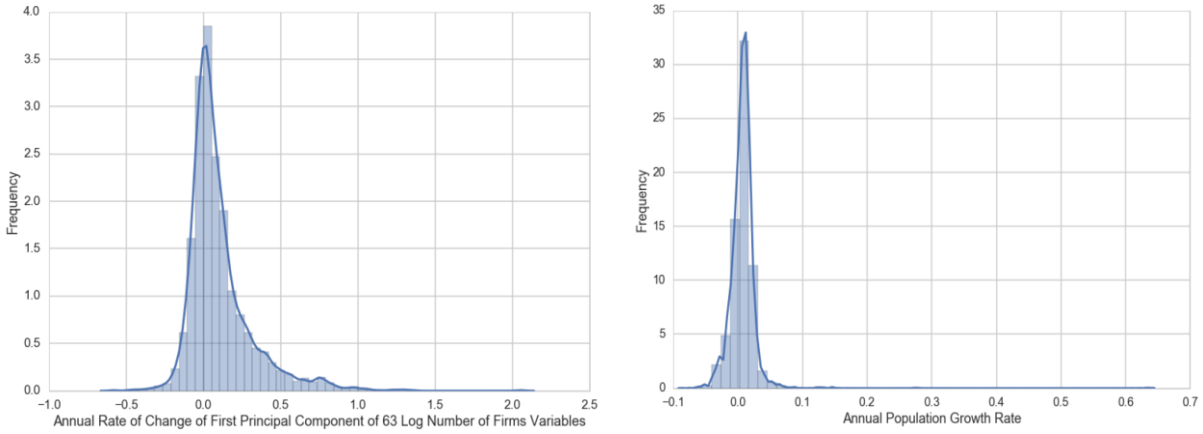


Figure 16. Histograms for the municipal firm and population growth indicators, Firm Growth x ≥ 5% and population growth dptot ≥ 1%

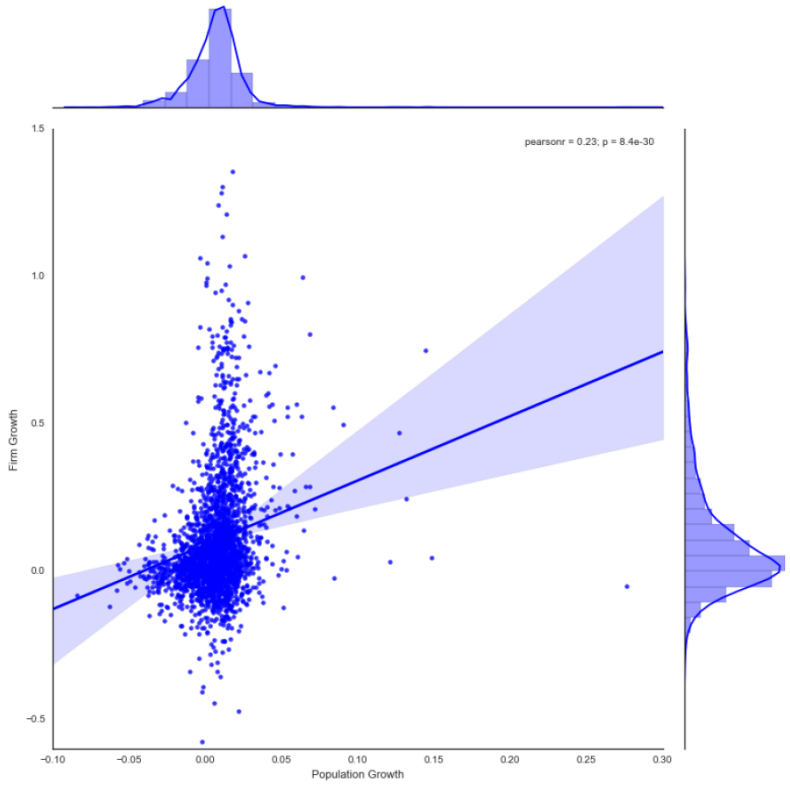


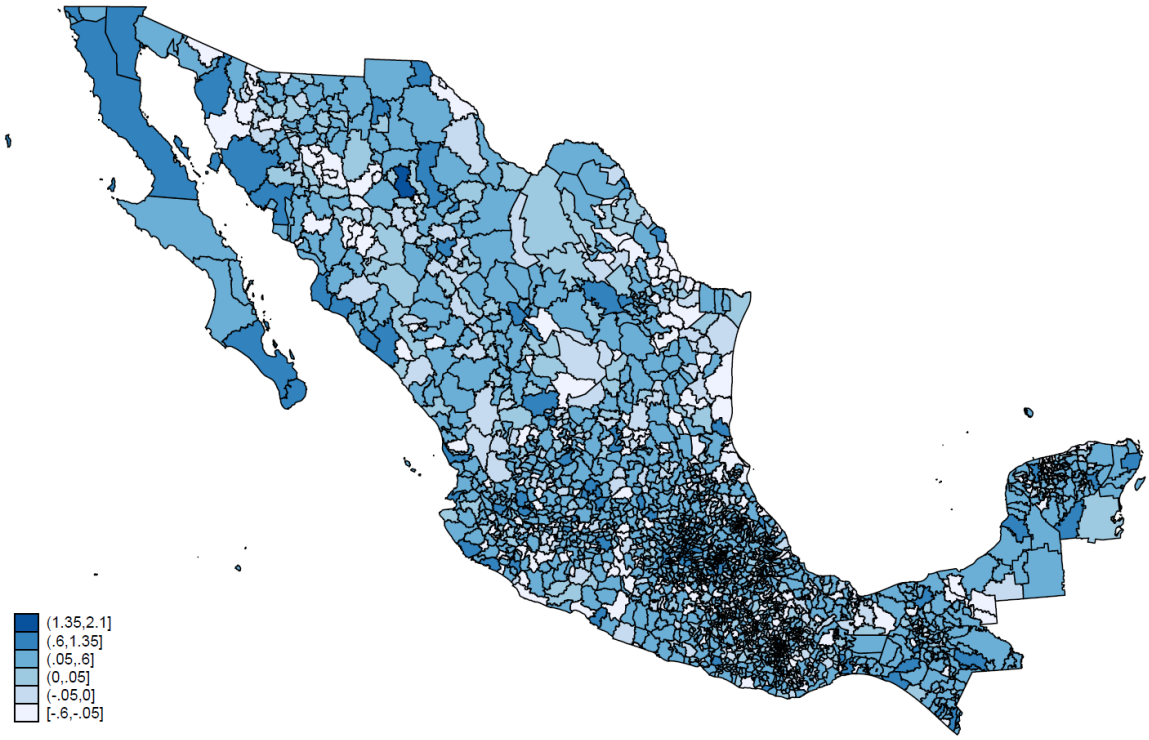
Figure 17. Scatterplot and histograms of municipal firm growth and population growth

3.3 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.

(i) Firm Growth



(ii) Population Growth

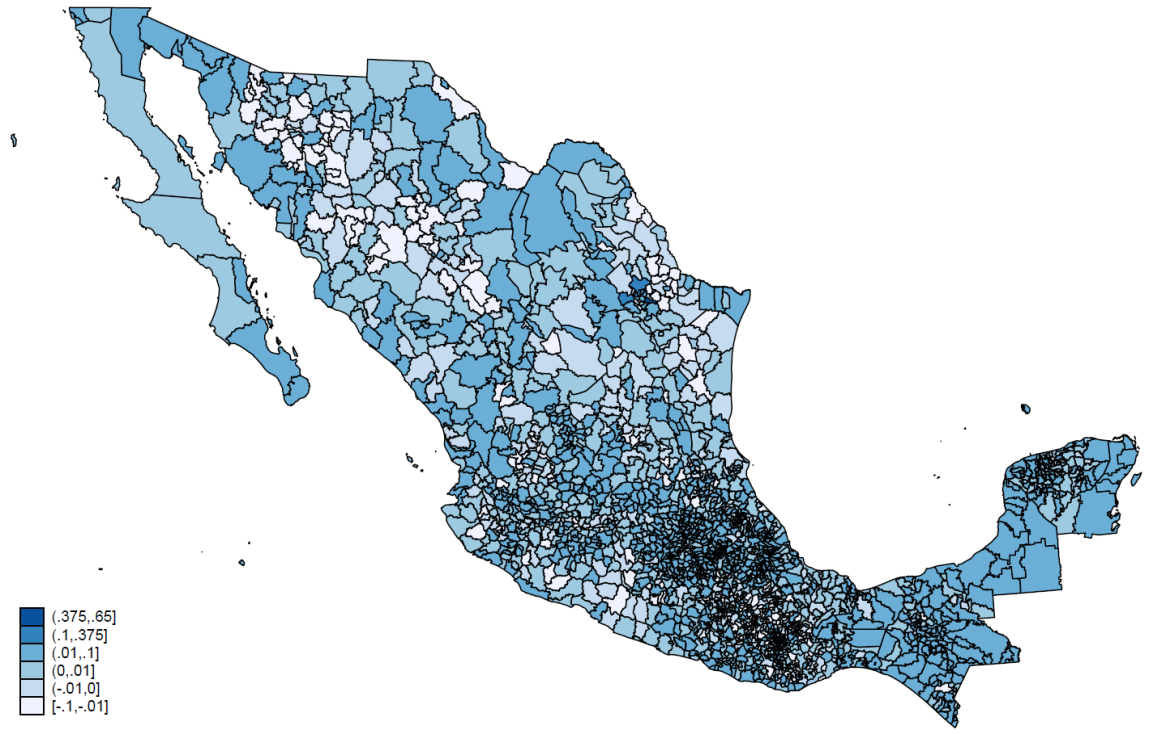


Figure 18. Maps of Mexican municipalities showing firm and population growth

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 × 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.

3.5 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 (Liaw & Wiener, 2002). 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 R2 are somewhat weak, at least compared to deterministic processes, since they are considerably affected by unpredictable components 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.

Table 2. Maximum decision tree depth grid search for the four random forest applications

	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

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 R2 (for the RF Classifier or Regressor; see Table 2, column maximae highlighted in yellow), and then applied to the test sample, yielding the corresponding test score. In fact the 70% accuracy obtained for Firm Growth is quite high. 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.

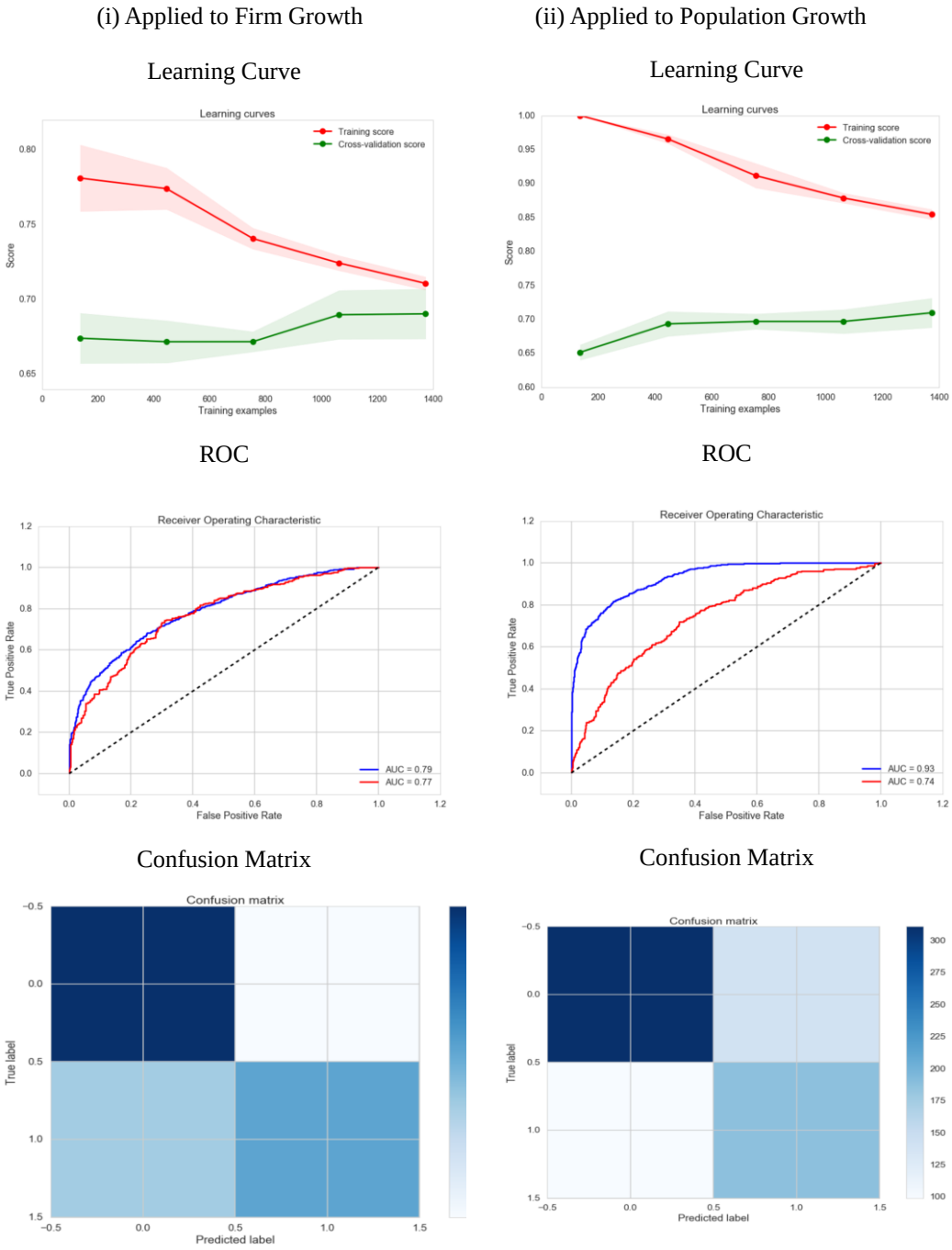
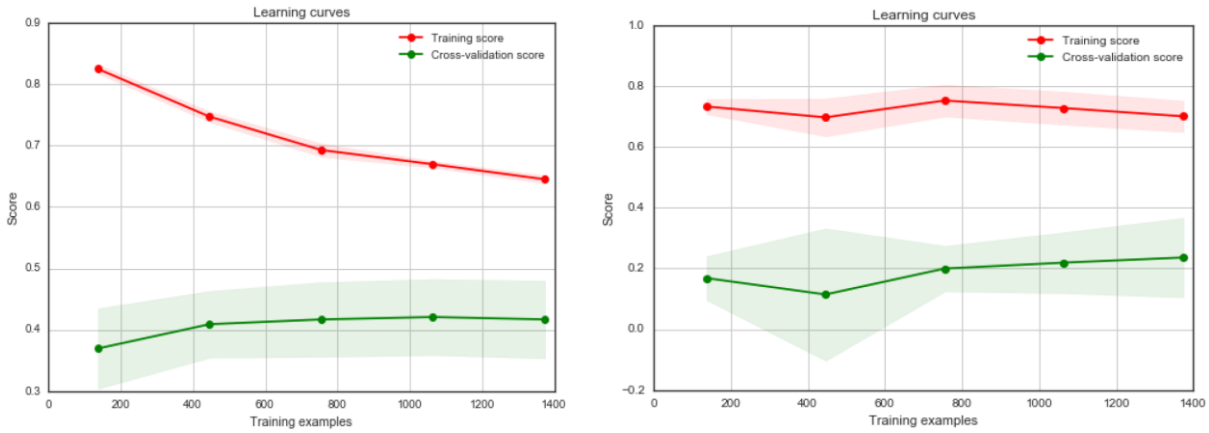


Figure 20. Evaluation metrics for the random forest classifier

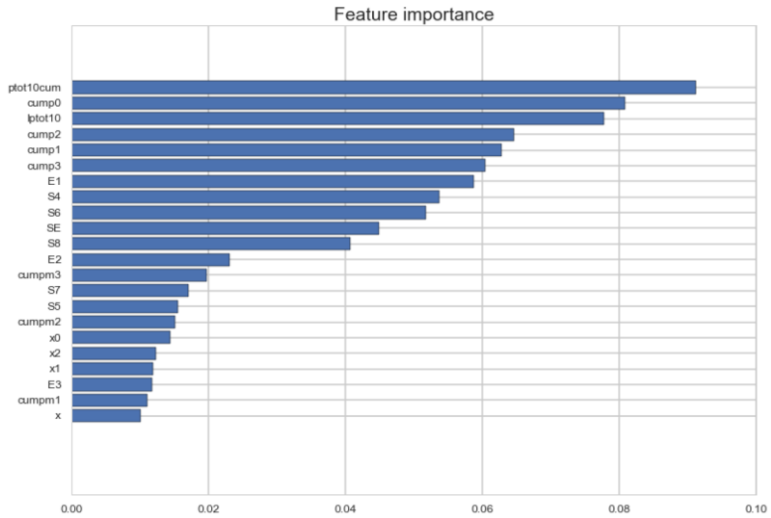
Figure 20 shows for each application of the RF Classifier the Learning Curve, the ROC curve and the Confusion Matrix (Liaw & Wiener, 2002). Overall, the learning curve required a smaller sample for lower maximum decision tree depths. Correspondingly, the Area Under the Curve for the test results tended 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.



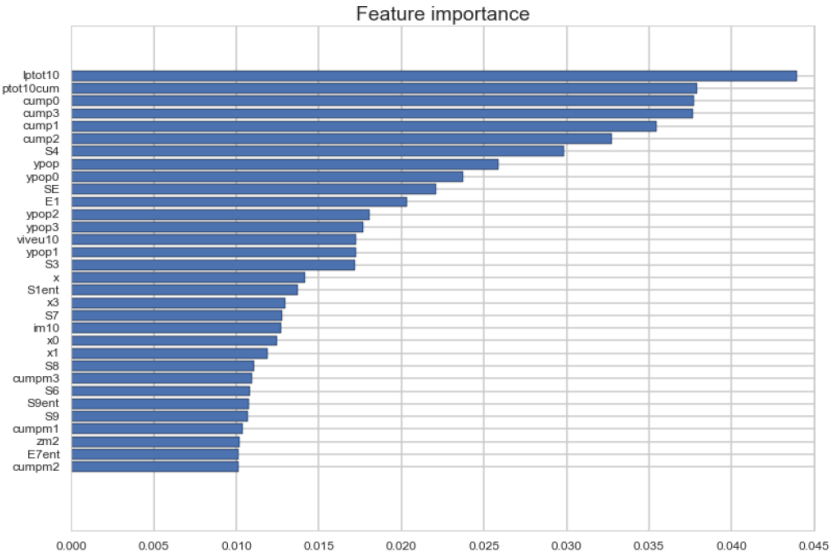
(i) Applied to Firm Growth (ii) Applied to Population Growth

Figure 21. Learning curves for the random forest regressor

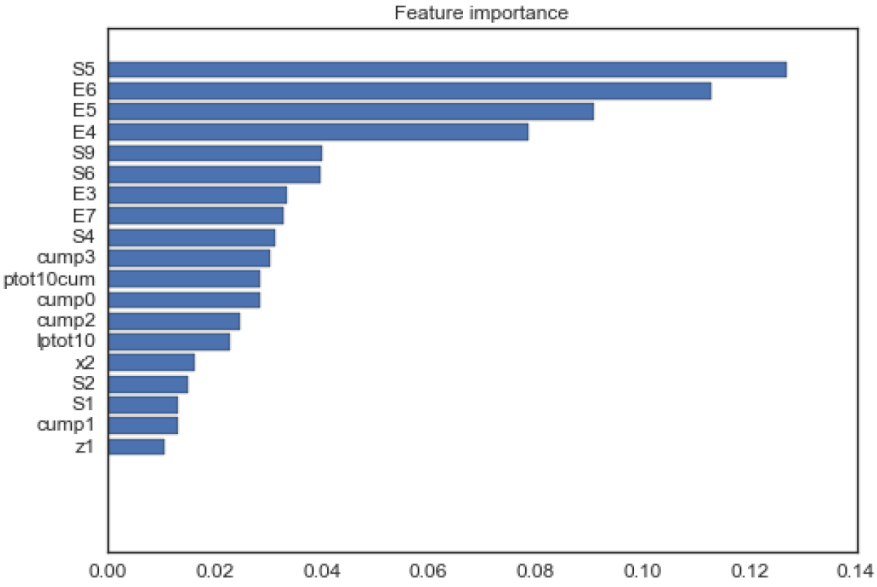
(i) Random Forest Classifier applied to Firm Growth



(ii) Random Forest Classifier applied to Population Growth



(iii) Random Forest Regressor applied to Firm Growth



(iv) Random Forest Regressor applied to Population Growth

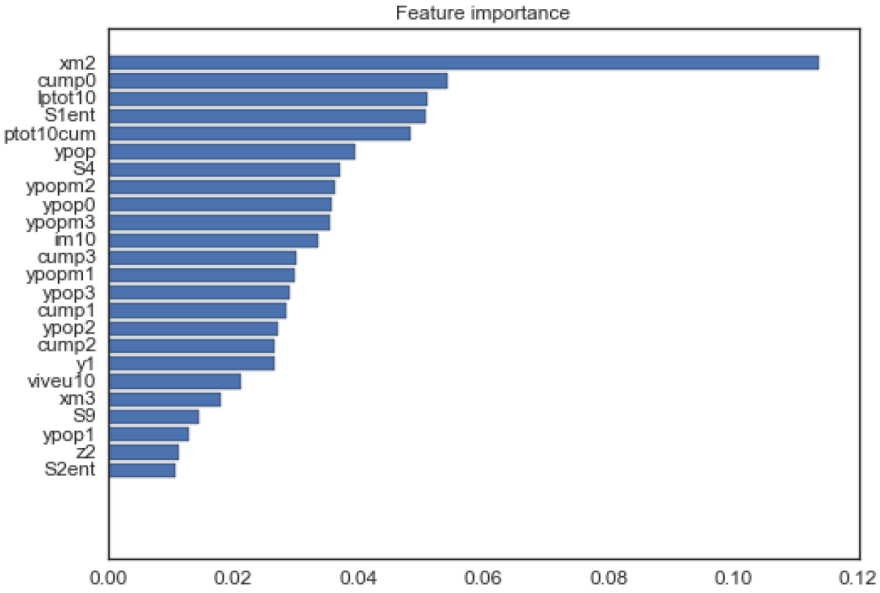


Figure 22. Feature importance

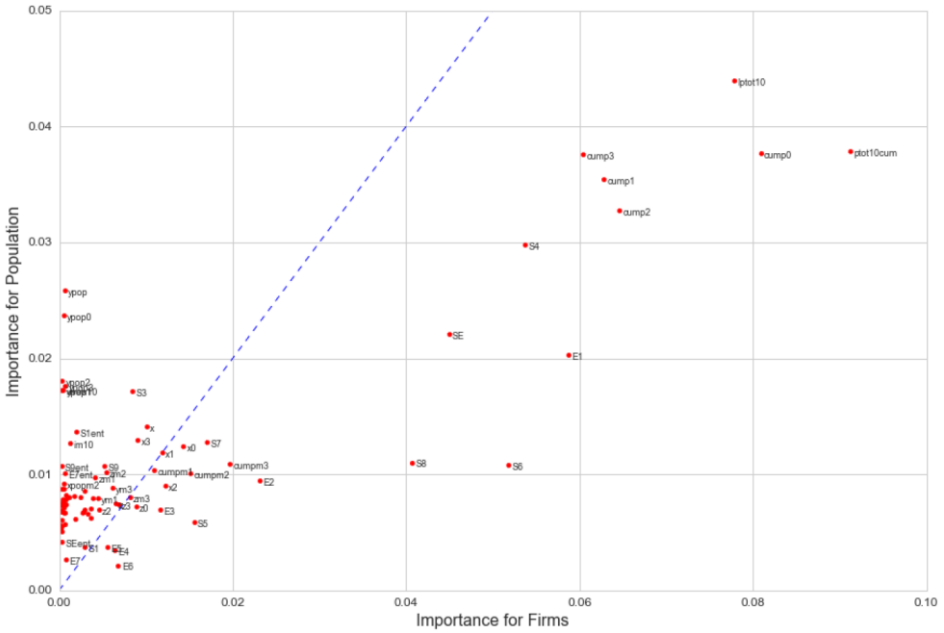


Figure 23. Scatter plot of feature importances obtained by random forest classifier for firm and population growth

Note. The blue dashed 45° line indicates whether the features are more important for Firm or for Population Growth.

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.

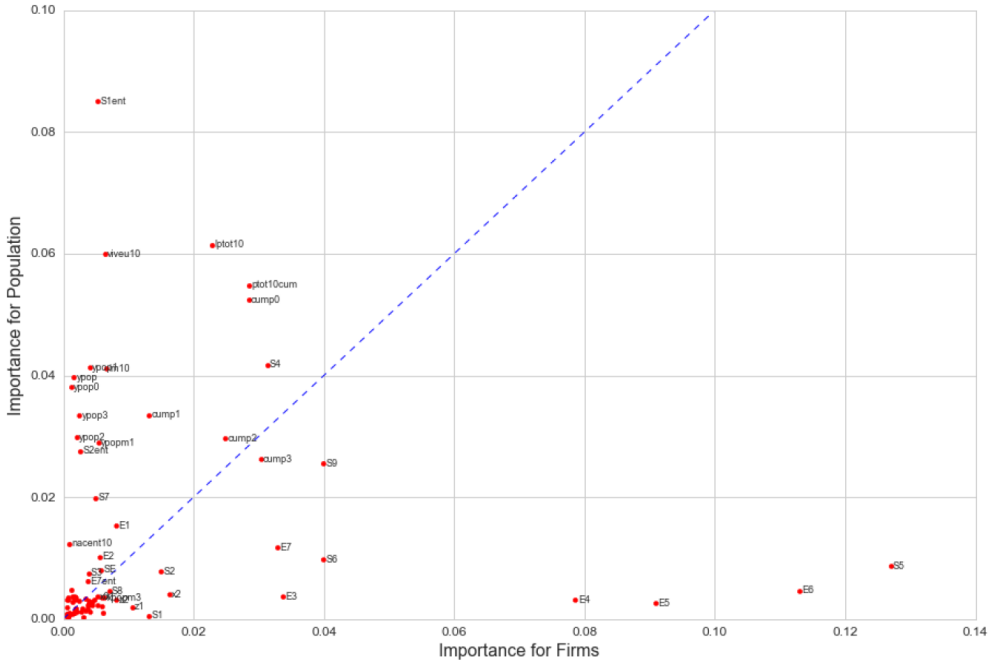


Figure 24. Scatter plot of feature importances obtained by random forest regressor for firm and population growth

Note. The blue dashed 45° line indicates whether the features are more important for firm or for population growth.

4. Results: Application of Random Forest Classifier and Regressor

4.1 Feature Importance for Growth of Firm Numbers and Population

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 significantly 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.

4.2 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 roles 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 sequential logic to the productive sectors and employment levels that are dominant at different levels of these variables.

The scatterplots of the feature importances obtained for Firm and Population Growth by the RF Classifier and Regressor are shown in Figures 23 and 24. 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. Both plots give evidence of a transition because of the U-shaped boundary of the features. Figure 23 shows results for "healthy growth". In this case we are only concerned with a qualitative question. Six population indicators emerge as the most important features for both processes, and they are more important for firms. These include population, log population, cumulative population and its rank, and the percentages of municipalities in rings of cumulative population 5, 10 and 15 percent above the municipality's ranking, a correlate of the competitive environment between municipalities, and of the appropriateness of this municipal size for economic growth. Then follow specific firm indicators reflecting the local economic environment such as the construction production sector, number of small firms, total number of firms (these are in fact similar, because there are fewer large firms), trade restaurants and hotels and finance, insurance and realtors, that may indicate specific local growth processes. Other sectors can be observed in the figure.

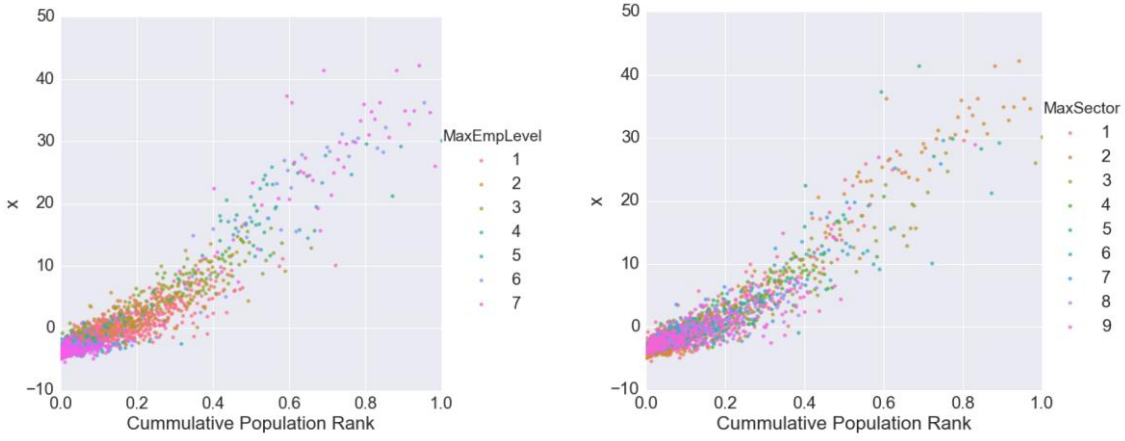


Figure 25. Scatterplots of the first principal component of firm numbers x with the cumulative population rank. maxemplevel is defined as the employment category that has highest value expressed in a normalized scale with mean 0 and standard deviation 1. Similarly MaxSector regarding production sectors.

As we mentioned in the introduction, what is important here is not causality as such, which may be mutual, but the fact that the processes of growth in firm numbers and population coincide and take place together. If there were a steady-state in the population distribution, predictive indicators would follow quite a different pattern. For example, population growth could follow firm growth and firm growth could follow sectoral patterns. There

would be little reason for there to be competition across municipalities according to population patterns, nor for population indicators per se to guide firm growth.

On the left of Figure 23, affecting the population dynamics, are indicators related with the second principal component of population ypop. This population 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 3). A look at the data shows municipalities where ypop is high occur in many Mexican states. Important features for population change include agricultural production at the state level, and the community and social sector, which is related to rural ejidos.

Table 3. Coefficients of principal components of population

Concept	Variable	xpop	ypop
Marginalization	im10	-0.5957	0.0067
log Population	lptop10	0.5182	-2555
Proportion born in the state	nacent10	-0.5771	0.1107
Live in the US	viveu10	0.2085	0.9604

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 for Firm Growth, as well as employment levels [101, 250], [51, 100] and [31, 50], in that order. Agriculture at the state level and percentage having migrated to the US become prominent for Population Growth (positive or negative), together with marginalization and other indicators mentioned before. The two figures thus detect both interlinkage and qualitative differences between Firm and Population Growth.

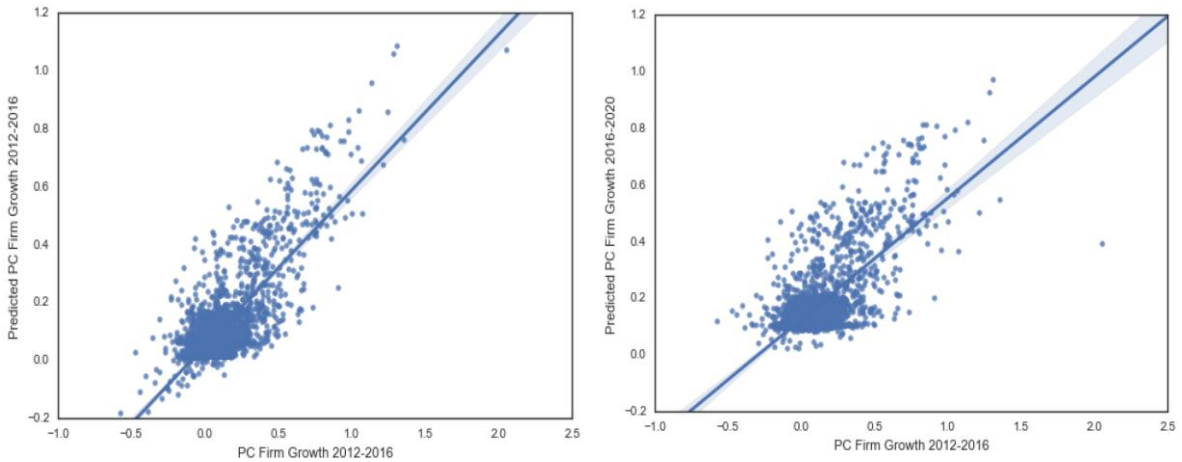


Figure 26a. Scatter plot of predicted municipal firm growth for 2012-2016 and 2016-2020 vs actual growth over the period 2012-2016 (some outliers not shown)

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.

4.3 Prediction of 2016-2020 Growth

Finally, the trained RF Classifier and Regressor were 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, which are expressed as percentages, we used the 2010 values.

Figures 26a and 26b 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 sections of the cloud.

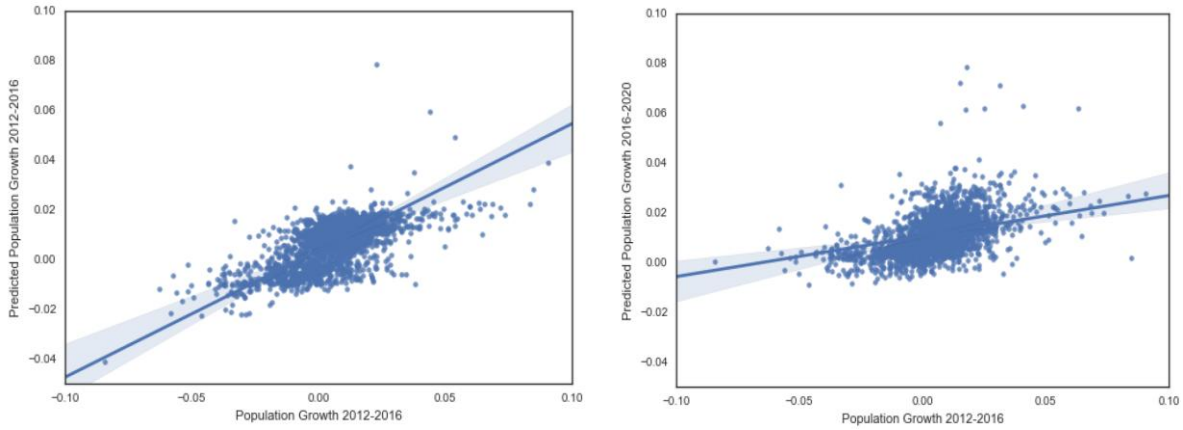


Figure 26b. Scatter plot of predicted municipal population growth for 2012-2016 and 2016-2020 vs actual growth over the period 2012-2016 (some outliers not shown)

We can also see the results of the RF Classifier in Table 4.

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. The figure shows the locally weighted linear regression, which predicts low but positive Firm Growth where growth has been negative and a tendency for Firm Growth to converge in the long-run to a value of about 75%, clearly a transitional rate of growth. Predictable Population Growth is more varied and currently converges to about 1.5% per year. The convergence rate is slower above the equilibrium level than below it.

Table 4. Firm and population growth prediction in 2012-2016 versus 2016-2020

Predicted Firm Growth Class				Predicted Population Growth Class			
	2016-2020:	Yes	No		2016-2020:	Yes	No
2012-2016	Yes	1,026	0	2012-2016	Yes	987	58
	No	366	1,063		No	274	1,136

Note. This shows the behavior of the predictable rather than unexpected component of growth. Once municipalities are growing, they are mostly expected to keep growing.

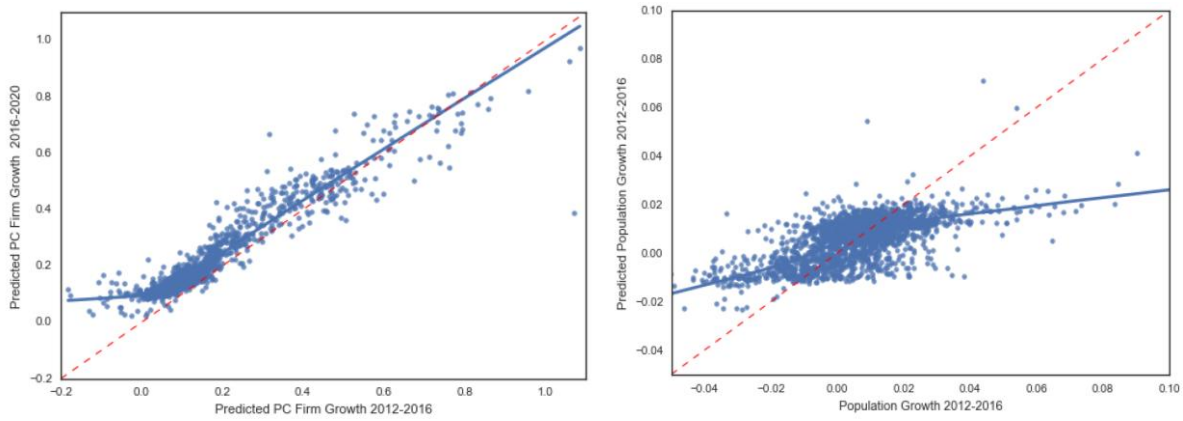


Figure 27. Scatterplot of predicted growth in 2012-2016 vs predicted growth for 2016-2020, according to the RF Regressor. A dashed 45° line is plotted in red for comparison.

5. Discussion

5.1 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 examined the data, slowly the interrelationships emerged, until we actually came to think of firm number growth and migration in Mexico as 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 Random Forest 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, proportionally higher firm numbers and sizes occur in municipalities with larger populations.
2) These municipalities also tend to grow more rapidly.
3) Nevertheless, there is an inverted U-curve, with a tendency for firm numbers to grow even faster in middle municipalities.
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 education, fertility and mortality. We also have not included other data such as local government and local geography.

5.2 Policy Recommendations

This machine learning approach has concentrated on the bird's eye viewpoint on firm growth in Mexico. Below I note that in fact the same machine learning program can be used 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. It is likely that migration follows economies of scale which give it a degree of inevitability. Only the provision of productive infrastructure that accelerates convergence between the poorer states (such as Oaxaca, Chiapas, and Guerrero) and the rest of the country, e.g. roads, communication, internet, technological and institutional know-how, and so on, can reduce this.

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 basic infrastructure 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, as provider of public goods. 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 visualizing the data and from the analysis is that the process of firm and population growth is marked by complexity. This means that qualitatively different detail and process emerge 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 a place must be made 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 rewarding excellence in public services can be defined across municipalities that can help to make both public service and the growth process more efficient.

5.3 Further Applications

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 machine learning 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 this research is based on a program that can be applied 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, to 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.

Acknowledgments

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

References

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Introduction to The Economics of Artificial Intelligence: An Agenda* (Preliminary draft). Retrieved from http://www.nber.org/chapters/c14005.pdf

Andini, M., Ciani, E., De Blasio, G., D'Ignazio, A., & Salvestrini, V. (2017). Targeting policy-compliers with machine learning: an application to a tax rebate programme in Italy. *Temi di discussion. Economic working papers, 1158*. Bank of Italy, Economic Research and International Relations Area. https://doi.org/10.2139/ssrn.3084031

Aufenanger, T. (2017). Machine learning to improve experimental design. *FAU Discussion Papers in Economics, 16/2017*. Friedrich-Alexander University Erlangen-Nuremberg, Institute for Economics.

Chakraborty, C., & Joseph, A. (2017). Machine learning at central banks. *Bank of England working papers, 674*. https://doi.org/10.2139/ssrn.3031796

Durlauf, S. N., & Aghion, P. (2005). *Handbook of Economic Growth*.

Gogas, P., Papadimitriou, T., & Karagkiozis, D. (2018). The Fama 3 and Fama 5 factor models under a machine learning framework. *Working Paper series 18-05*. Rimini Centre for Economic Analysis.

Green, G., & Richards, T. (2016). *Interpreting Results of Demand Estimation from Machine Learning Models.* 2016 Annual Meeting, July 31-August 2, 2016, Boston, Massachusetts, Agricultural and Applied Economics Association.

Gründler, K., & Krieger, T. (2018). Machine Learning Indices, Political Institutions, and Economic Development. *CESifo Working Paper Series 6930*. CESifo Group Munich.

Harris, J. R., & Todaro, M. P. (1970). Migration, unemployment and development: A two-sector analysis. *The American Economic Review*, *60*(1), 126-142.

Krugman, P. (1991). Increasing returns and economic geography. *Journal of Political Economy*, *99*(3), 483-499. https://doi.org/10.1086/261763

Liaw, A., & Wiener, M. (2002). Classification and regression by Random Forest. *R news, 2*(3), 18-22.

Mareckova, J., & Pohlmeier, W. (2017). Noncognitive Skills and Labor Market Outcomes: A Machine Learning Approach. *Annual Conference 2017 (Vienna): Alternative Structures for Money and Banking, Verein für Socialpolitik / German Economic Association.*

Milgrom, P., & Tadelis, S. (2018). How Artificial Intelligence and Machine Learning Can Impact Market Design. NBER Chapters, in: *The Economics of Artificial Intelligence: An Agenda*. National Bureau of Economic Research, Inc. https://doi.org/10.3386/w24282

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives, 31*(2), 87-106. https://doi.org/10.1257/jep.31.2.87

Papadimitriou, T., Gogas, P., Matthaiou, M., & Chrysanthidou, E. (2014). Yield curve and Recession Forecasting in a Machine Learning Framework. *Working Paper series 32_14, Rimini Centre for Economic Analysis*. https://doi.org/10.2139/ssrn.2388719

Renner, P., & Scheidegger, S. (2017). Machine learning for dynamic incentive problems. *Working Papers 203620397, Lancaster University Management School, Economics Department.*

Rowland, A. M. (2001). Population as a determinant of local outcomes under decentralization: Illustrations from small municipalities in Bolivia and Mexico. *World Development, 29*(8), 1373-1389. https://doi.org/10.1016/S0305-750X(01)00045-6

Samii, C., Paler, L., & Daly, S. (2016). Retrospective Causal Inference with Machine Learning Ensembles: An Application to Anti-recidivism Policies in Colombia. *Political Analysis, 24*(04), 434-456. https://doi.org/10.1093/pan/mpw019

Van der Laan, M. J., & Rose, S. (2011). *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4419-9782-1

Varian, H. R. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives, 28*(2), 3-27. https://doi.org/10.1257/jep.28.2.3