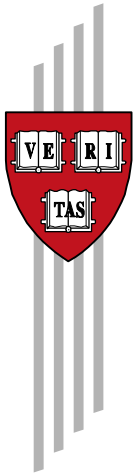
Forecasting Formal Employment in Cities

Forecasting Formal Employment in Cities

Eduardo Lora

CID Research Fellow and Graduate Student
Working Paper No. 114
May 2019



Working Papers

Center for International Development
at Harvard University

Forecasting formal employment in cities

Eduardo Lora[1]
RiSE and Department of Economics, Universidad Eafit and
Center for Internationational Development, Harvard University
May 2019

Abstract

Can "full and productive employment for all" be achieved by 2030 as envisaged by the United Nations Sustainable Development Goals? This paper assesses the issue for the largest 62 Colombian cities using social security administrative records between 2008 and 2015, which show that the larger the city, the higher its formal occupation rate. This is explanined by the fact that formal employment creation is restricted by the availability of the diverse skills needed in complex sectors. Since skill accumulation is a gradual path-dependent process, future formal employment by city can be forecasted using either ordinary least square regression results or machine learning algorithms. The results show that the share of working population in formal employment will increase between 13 and nearly 32 percent points between 2015 and 2030, which is substantial but still insufficient to achieve the goal. Results are broadly consistent across methods for the larger cities, but not the smaller ones. For these, the machine learning method provides nuanced forecasts which may help further explorations into the relation between complexity and formal employment at the city level.

[1] Comments by Mauricio Quiñones are acknowleged.

1. Introduction

United Nations Sustainable Development Goal 8 is "Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all". More specifically, target 8.3 seeks to "[b]y 2030, achieve full and productive employment and decent work for all women and men, including for young people and persons with disabilities, and equal pay for work of equal value". This paper assesses how achievable this target is for Colombia, based on a novel theory of formal employment creation in cities and two complementary forecasting methods: standard regressions and machine learning.

Cities are necessary for economic growth to take place through a process of diversification and innovation that leads to productive employment and decent work for larger shares of the population. However, urbanization is not a sufficient condition for industrialization and productive employment: the expected relation between urbanization, industrialization and employment quality is absent in many parts of the world (Gollin, Jedwab and Vollrath, 2016). Urbanization patterns, and not just urbanization rates or macroeconomic factors (such as natural resource abundance) may shed light on the role of cities in economic growth and formal employment creation as suggested by two strong stylized facts (O'Clery et al 2018): (1) formal occupation rates are more variable across cities within countries than across developing countries (Figure 1), and (2) larger cities create proportionally more formal employment (Figure 2).

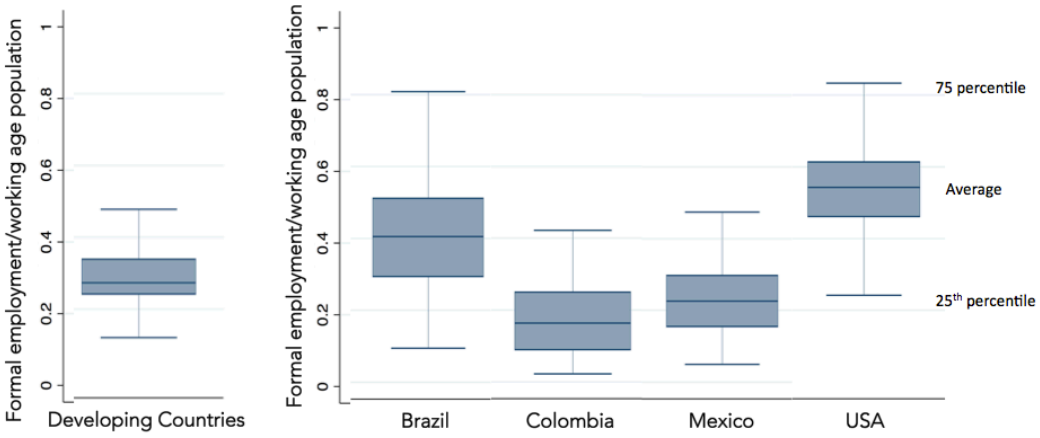


Figure 1. Box plots for the distribution of formal occupation rates in a set of 56 developing countries (left plot) and cities in Brazil, Colombia, Mexico and the US. We observe a larger variance in formality rates across cities within countries than across countries, suggesting that the study of the determinants of formality across cities is a relevant area of study in connection with Sustainable Development Goal 8 ("full and productive employment and decent work for all"). Source: O'Clery et al (2018).

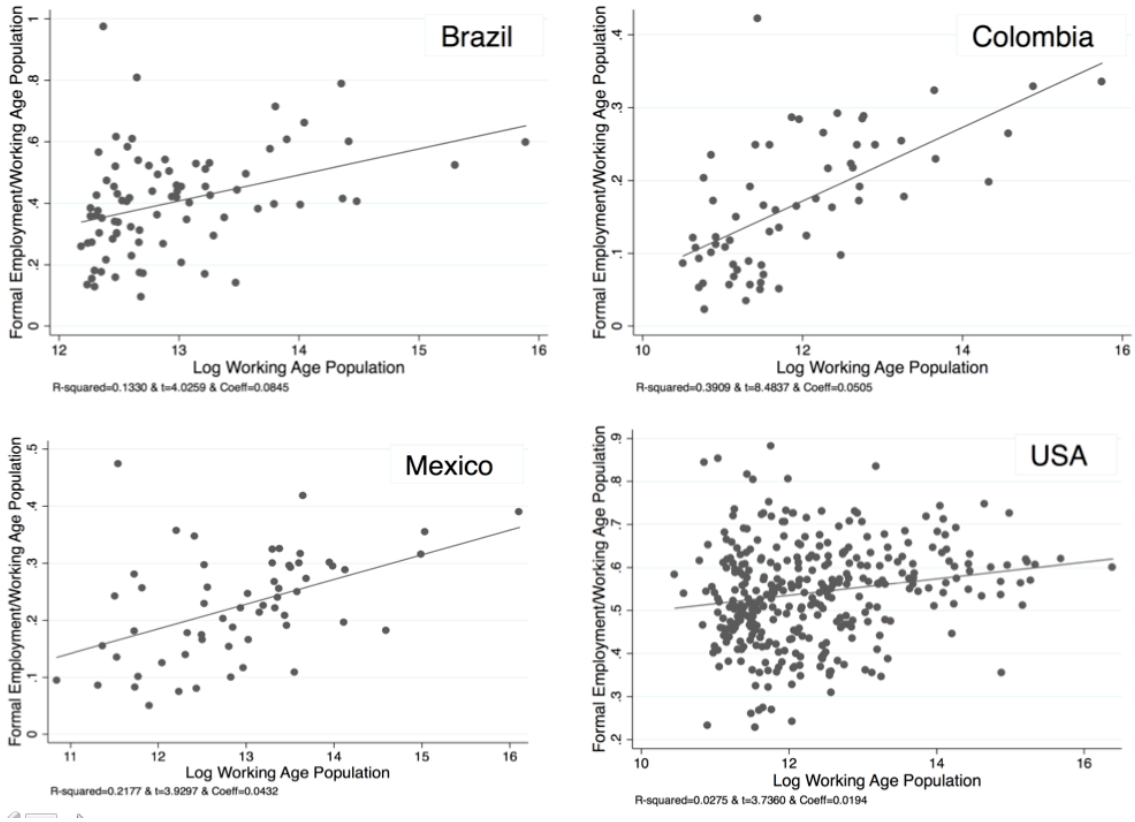


Figure 2. Formality rates increase with working age population for cities across Brazil, Colombia, Mexico and the US: larger cities have disproportionately more workers in the formal sector than smaller cities, a pattern that is statistically significant in the four countries as shown at the bottom of the figures. Source: O'Clery et al (2018).

2. Theoretical framework

One of the central issues in economic development theory is the reason for the size and persistence of informal labor in developing economies. Since formal firms have access to capital and technology that make them more productive than small or family businesses, what explains that large chunks of the labor force are not occupied in the formal sector where labor conditions are better than in the informal sector? Economic theory has provided several explanations. In dualistic models of informality, the self-employed and their family businesses are fundamentally different from formal firms in the type of human capital they use –mainly uneducated and unproductive entrepreneurs and managers–, and in what they produce –mainly low-quality products for low-income customers. The formal and informal sectors coexist because they are different (Lewis 1954, Harris and Todaro 1970, Rauch 1991). An alternative view is that of De Soto (1989, 2000), who considers that informal firms are an untapped reservoir of productive resources held back by government regulations. Relatedly, Levy (2008) sees informal

businesses as entrenched firms that survive in spite of their low productivity by avoiding taxes and regulations. Lastly, in labor search models that take into account the costs and benefits of labor regulations, informal employment is not the consequence of exclusion, but the result of labor market frictions between heterogeneous workers and firms (Albrecht, Navarro and Vroman 2009; Bosch and Maloney 2010; Ulyssea 2010; Meghir, Narita and Robin 2015).

While empirical evidence has been provided in support to each of these explanations of informality, none of them recognizes the two stylized facts mentioned in the introduction, namely that formal occupation rates across cities (within a given country) have a larger variance than across countries, and that formal occupation rates are directly and significantly associated with city size. In other words, none of the mainstream theories can explain the role of cities in formal employment creation. Furthermore, some of the main variables put forward by those theories to explain the presence of informality –such as social security regimes and labor hiring and firing legislation—have little or no variance across cities within each country.

In view of these shortcomings, this paper adopts the theoretical framework developed by O'Clery et al (2019), which differs from previous theories in a number of ways. First, it focuses on cities rather than countries, because cities are the actual locations where workers and their employers interact. Second, it emphasizes skill diversity –which is central in urban economics—rather than skill levels, educational attainment or managerial capabilities. Third, it assumes that firms evolve by tinkering with skills because many feasible technologies cannot be known in advance, but need to be discovered. Formal employment creation in cities results from this evolutionary process. In larger cities, firms have better access to the diverse skills they need to produce more sophisticated goods.

The main components of the model can be summarized as follows (for the complete model see O'Clery et al 2018):

City size and skill diversity are taken as exogenous. Each firm is located in a city, but sells to the whole national market under perfect competition. The output of firm r, which belongs to industry j at time t, is given by a CES production function whose only production factors are types of labor differentiated by skill, where skills k are hard to substitute for one another:

$$y_t^r = A_t^r \left[\sum_{k \epsilon j_t(r)} l_t^r(k)^\rho\right]^{\frac{1}{\rho}} \qquad (1)$$

Optimal formal labor demand of each skill is given by the solution to the cost minimization problem facing the firm, from which wage by skill is obtained:

$$W_t^j = (\overline{w}_t)^{\frac{\rho}{\rho - 1}} C_t^j \qquad (2)$$

where $\overline{w}_t$is the survival (or minimum) wage of workers with general skills (which are supplied in excess of what firms demand), and where C_t^j reflects the diversity of skills that are needed in industry *j*, and can therefore be interpreted as a measure of the complexity of the industry,

$$C_t^j = \sum_{k\,\epsilon\,j} \theta_t(k)^{\frac{\rho}{\rho-1}} \qquad (3)$$

Since $\theta_t(k) \geq 1$ for every skill, industry complexity is larger in industries which combine a larger subset of sophisticated skills.

Finally, firms transition from less to more sophisticated industries following a probabilistic rule such that the conditional expected value of a firm´s complexity one period ahead is:

$$E_t\left(C_{t+1}^j \mid j_t(r) = j\right) = p\, C_t^j + (1-p) \underbrace{\left[\sum_{j' \in J} \beta_t(j,j')\, C_t^{j'}\right]}_{\text{Complexity Potential}} \qquad (4)$$

Due to the definition of $\beta_t(j,j')$, the complexity potential of a given firm depends on (i) the industry to which the firm currently belongs, (ii) the distance between such industry and those industries to which the firm could migrate, (iii) the relative abundance in the local labor market of those new skills which the firm must hire in order to carry out an industry transition, and (iv) the size of labor force in city.

Total formal employment at the city level one period ahead (F_{t+1}) is thus obtained by aggregating labor demand across skills and firms:

$$F_{t+1} = \sum_r \sum_{k \in j_{t+1}(r)} l_{t+1}^r(k)^* = \sum_r \left[\left(\frac{y_{t+1}^r}{A_{t+1}^r}\right)\left(C_{t+1}^j\right)^{\frac{-1}{\rho}} \sum_{k \in j_{t+1}(r)} \left(\theta_{t+1}(k)\right)^{\frac{1}{\rho-1}}\right] \qquad (5)$$

Aggregate formal employment in a city depends on current complexity of all its firms, which in turn depends on past complexity potential (equation 4). Therefore, formal employment in period *t+1* is a function of complexity potential in period *t*:

$$F_{t+1} = f\underbrace{\left[\sum_{j' \in J} \beta_t(j,j')\, C_t^{j'}\right]}_{\text{Complexity Potential}} \qquad (6)$$

Notice that complexity potential is a weighted average of the *industry complexity* of the *missing sectors* with weights given by the *skill similarity* between those sectors and the ones already present.

In order to operationalize equation (6), data are needed on industry complexity, missing sectors, and skill similarity between all pairs of industries. Since skills are tacit knowledge and therefore unobservable, industry complexity and complexity

potential must be computed indirectly. To that end, O'Clery et al (2019) make use of the methodologies developed by Hidalgo and Hausmann (2009) and Neffke and Henning (2013). In essence, *industry complexity* is a measure of the range of skills needed in an industry, which is obtained from the number of industries present in the cities that have the industry (ie those industries that have revealed comparative advantage greater than 1 in city, based on formal employment shares) and the number of cities that have the industry (ie those cities where the industry has revealed comparative advantage greater than 1). *Skill similarity* between a pair of industries is measured by the relative intensity of the labor flows between the two industries, and *missing industries* in city are those with revealed comparative advantage lower than 1 (Appendix 1 provides further details on computation methods).

3. Data and empirical definitions

Like in O'Clery et al (2019), I use data for Colombian cities larger than 50,000 inhabitants. My definition of cities rests on the methodology proposed by Duranton (2015) to define metropolitan areas. It consists of adding iteratively a municipality to a metropolitan area if there is a share of workers, above a given threshold, that commute from the municipality to the metropolitan area. Assuming a 10 percent threshold, the methodology generates 19 metropolitan areas that consist of two or more municipalities (comprising a total of 115 municipalities). Since another 43 individual municipalities have populations above 50,000 inhabitants, a total of 62 cities is obtained.

The main data source for the 62 cities is the social security administrative data collected by the Health and Social Security Ministry, known as PILA (Planilla Integrada de Liquidación Laboral). PILA contains information by worker and firm on days of work, sector of activity and municipality.[2] To aggregate these data, I count the share of the year t that each worker effectively contributed to the social security system through firms per city c per industry j ($emp_{c,j}$). This is the *formal employment* for a given sector (or for the aggregate of all sectors within a city). Sectors are defined at the 4-digit industry level of the International Standard Industrial Classification (ISIC, revision 3.0).

The *formal employment rate* in city c in year t ($F_{c,t}$) is defined as formal employment divided by the city-wide population 15 years old or older ($pop_{c,t}$, estimated by DANE):

$$F_{c,t} = emp_{c,t} \,/\, pop_{c,t} \qquad (7)$$

[2] The datasets have information on age and gender, which we do not use. Unfortunately, it provides no information on education, which prevents us from testing our model predictions vis-à-vis the findings of previous works discussed in the introduction.

The (simple) average formal occupation rate in cities was only 20.3 percent of working age population in 2015, with a relatively large standard deviation (between 11.1 percent points). Important changes in urban formal occupation rates occurred between 2008 and 2015: the *aggregate f*ormal occupation rate for the 62 cities went up from 21.1 percent to 31.2 percent, with a (simple) average increase across cities of 8.1 percent points and a standard deviation of 5.4 points. Formal occupation was facilitated by a rate of GDP growth of 4.1 percent, and probably also by the elimination in May of 2013 of payroll taxes representing 5 percent of the wage bill (Kugler, Kugler and Herrera-Prada 2017).

Since the formal employment rate is a variable bounded between 0 and 1, and the aim is to assess how fast it approaches 1, it is transformed to its logistic form, time-differentiated and expressed in annual terms:

$$y_{c,t-i} = \frac{1}{t-i}\left(\frac{e^{F_{c,t}}}{1+e^{F_{c,t}}} - \frac{e^{F_{c,t-i}}}{1+e^{F_{c,t-i}}}\right) \qquad (8)$$

where $y_{c,t-i}$ will be the dependend variable and the subscript *i* is the *year-interval* or number of years for the time-differentiation (which may take values between 1 and 7, given that the data cover an 8-year span). For intuition's sake, I will refer to the dependent variable as the "annual speed towards full employment", or "speed", for short.

The independent variables (at time *t-i*) will be *complexity potential,* $CP_{c,t-i}$ as explained above, the (log of) working age population, $lpop_{c,t-i}$, the *logistic of formal occupation rate*, $\frac{e^{F_{c,t-i}}}{1+e^{F_{c,t-i}}}$, a dummy for the *oil-producing cities* (those with more than one oil well per 10,000 inhabitants: Acacías, Arauca, Barrancabermeja, Neiva and Yopal) and a synthetic measure of the *exogenous sectoral shocks* by city *c* (following McGuire and Bartik 1991, the so-called *Bartik shock* measure for city *c* at time *t* is a weighted average of the rates of change between *t-i* and *t* of formal employment by sector at the national level, excluding city c, with weights equal to the employment share of each sector in city *c* in year *t-i*).[3]

Two forecasting methods will be used in a complementary way. The first one will be based on ordinary least square regressions for all the possible time frequencies of the yearly data between 2008 and 2015. After discussing the lack of consistency of some of the coefficcients, two regressions are chosen to forecast the dependent variable by city and compute the formality rates by city in 2030. The second method, further explained in section 5, will be a machine learning technique known as "random forest", by which a set of alternative results are predicted based on combinations of explanatory variables presumedly associated with the results (in an

[3] In O'Clery et al (2019) the measure of complexity potential depends on working age population, while here I am taking the latter as a separate explanatory variable. In this way, the relation between both variables can be explored in the machine learning exercises.

unknown non-linear fashion). The two methods are complementary because, while OLS provides light on the possible influence of each individual variable, its predictions can only be reliable if the coefficients can be consistently estimated and the relation between the dependent and the independent variables (or combinations thereof) is linear and known in advance. These limitations do not apply to machine learning techniques, which are intended to produce reliable predictions using probabilistic methods that make efficient use of all the data that may be relevant.

4. Regression-based forecasts

Table 1 is a summary of the regressions. Only the 7-year (ie full 2008-2015) and 1-year interval regressions are presented (see Appendix 2 for all the intervals). In the upper panel the 62 observations correspond to the number of cities because there is only one period. In the two other panels, the number of observations is 434, since there are 7 one-year periods (434=62 x 7).

Table 1. Regressions of speed towards full formal employment on complexity potential and other controls

(Pooled ordinary least squares for different intervals, with year dummies)

Full 7-year period	**Coefficient**	**Standard error**	**t statistic**	**P>|t|**
Complexity potential at t-7 (log)	0.003043	0.0007914	3.85	0
Working age population at t-7 (log)	-0.0006131	0.0003166	-1.94	0.058
Formality rate at t-7 (logistic)	0.1132962	0.046996	2.41	0.019
Oil producing city	0.0037701	0.0007497	5.03	0
Bartik shock between t-7 and t	-0.0419715	0.0237082	-1.77	0.082
Constant	-0.0388139	0.0235932	-1.65	0.106
Number of obs = 62				
Adj R-squared = 0.5891				

1-year intervals (full specification)	**Coefficient**	**Standard error**	**t statistic**	**P>|t|**
Complexity potential at t-1 (log)	0.0033963	0.0006686	5.08	0
Working age population at t-1 (log)	-0.0006598	0.0002322	-2.84	0.005
Formality rate at t-1 (logistic)	-0.0272684	0.0122967	-2.22	0.027
Oil producing city	0.0016853	0.0005864	2.87	0.004
Bartik shock between t-1 and t	0.2048173	0.0303162	6.76	0
Constant	0.0329708	0.0071898	4.59	0
Year dummies	F(6, 422) =		5.841	0
Number of obs = 434				
Adjusted R-squared = 0.5020				

1-year intervals (simplified specification)	**Coefficient**	**Standard error**	**t statistic**	**P>|t|**
Complexity potential at t-1 (log)	0.0030968	0.0003987	7.77	0
Oil producing city	0.0036794	0.0005224	7.04	0
Constant	0.0118205	0.0011629	10.16	0
Year dummies	F(6, 422) =		36.571	0
Number of obs = 434				
Adjusted R-squared = 0.4331				

The interpretation of the coefficients is not straighforward because of the way the dependent variable is defined. However, it is clear that although all the explanatory variables are significantly associated with the speed towards full employment, some change sign between the 7-year and the 1-year full specification (upper and middle panels). This suggests that their relation with the dependent variable is not adequately captured: there may be important interactions between the explanatory

variables or dynamic issues that are ignored in the specification adopted. Since both the number of cities and the number of periods are small, not much can be done to overcome these problems with standard econometrics. As we will see, machine learning techniques are able to deal with these limitations.

The lower panel shows a simplified version of the 1-year interval regression, which only includes the explanatory variables that are significantly and consistently directly or inversely associated with the dependent variable. Those are just complexity potential and the dummy variable for oil-producing cities.

I use the coefficients of the middle- and lower-panel regressions to forecast formal employment in 2030, with the following additional assumptions and methods:

- Complexity potential by city is assumed constant at the 2015 values
- Working age population by city is projected at the same growth rate observed between 2008 and 2015
- Formality rate at *t-1* (logistic) by city is calculated recursively with the forecast of the dependent variable for the previous year
- Oil producing city dummy is kept unchanged throughout the forecast period
- Bartik shock by city is assumed constant at the mean of the yearly data for 2008-2015.

The results appear in Figures 3-5 (and Appendices 2 and 3). A brief summary is in order. Figure 3 shows that formality rates will increase througout the whole sample of cities and forecast options: all cities will advance towards the full-formal employment target. However, it is unclear whether formality rates will tend to converge. In the full specification, formality rates tend to converge (because all increase by about the same), but in the simplified specification they tend to diverge (increases are proportional to the initial values). Also, with the full specification, formal employment rates in many cities will be above 0.6, and even 0.8 in 2030, suggesting that "full and productive employment and decent work for all women and men" may be within reach. But in the simplified specification, only a handful of cities will get that high.

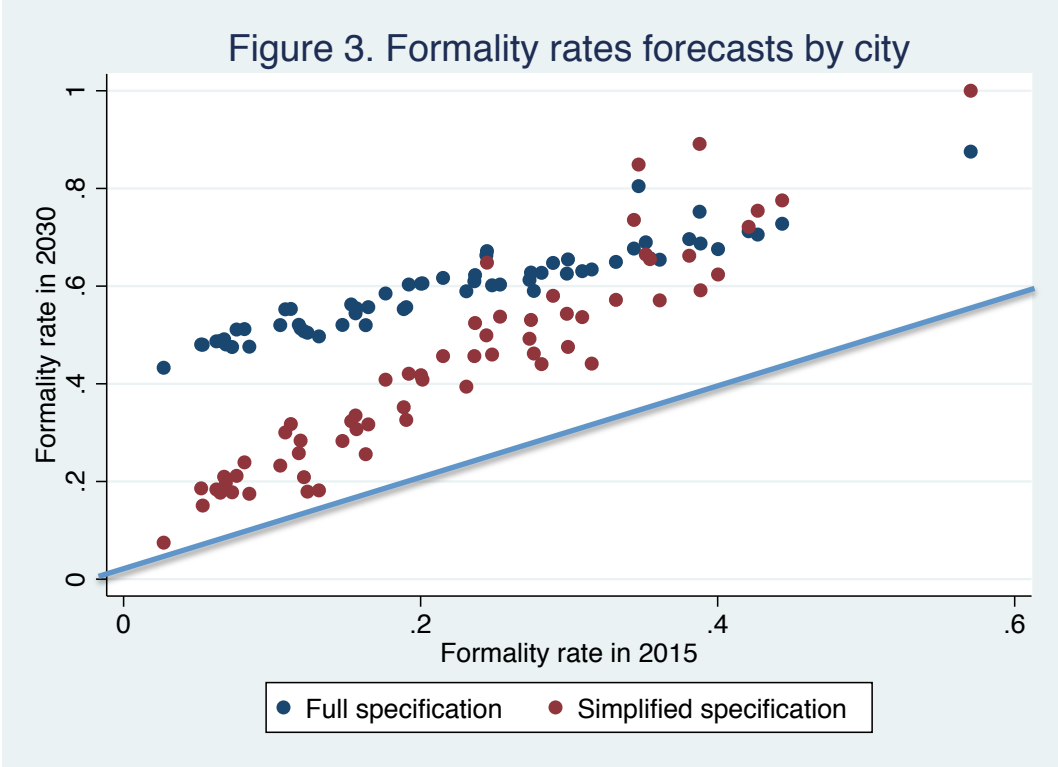


Figure 3. Regression-based forecasts of formal employment by city show that all cities would move towards full employment. In the simplified specification, formal employment rates tend to diverge.

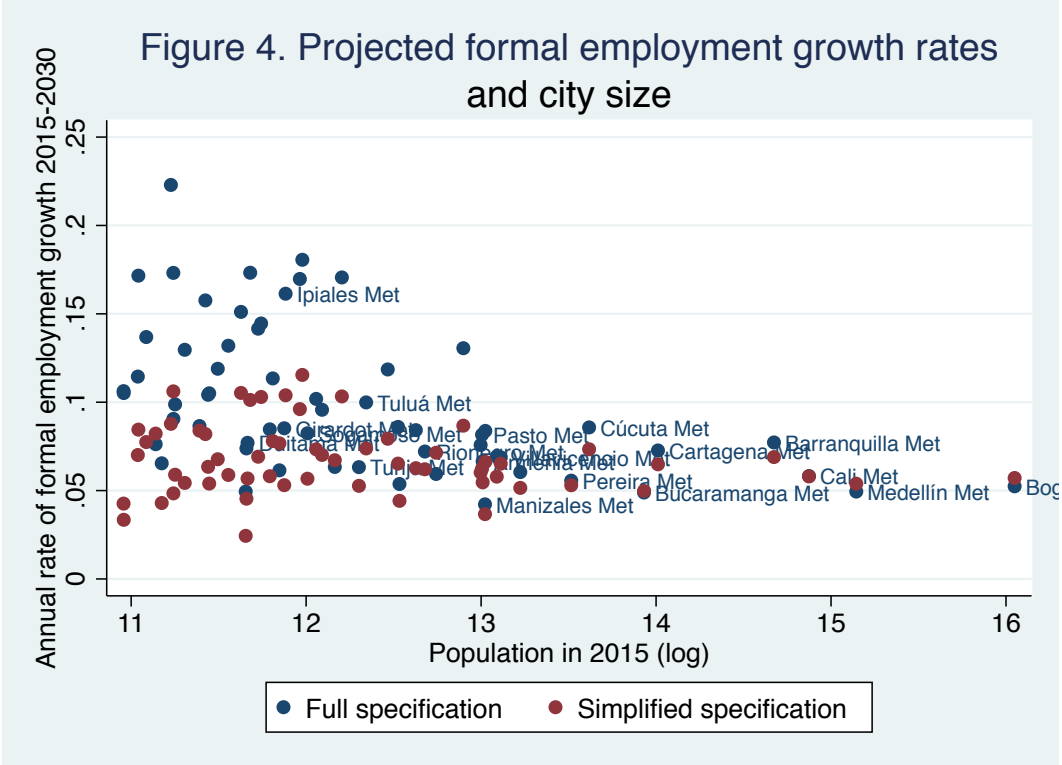


Figure 4. Regression-based forecasts of formal employment growth rates by city show substantially more dispersion in growth rates and between the two specifications among the smaller cities.

Figure 4 makes clear that the differences between the two forecasts are strongly related to city size: while for the smaller cities the rates of employment growth can differ by more than 10 percent points, for the largest cities the differences are negligible (the figure shows the names of the multi-municipality cities only, most of which are also the largest cities).

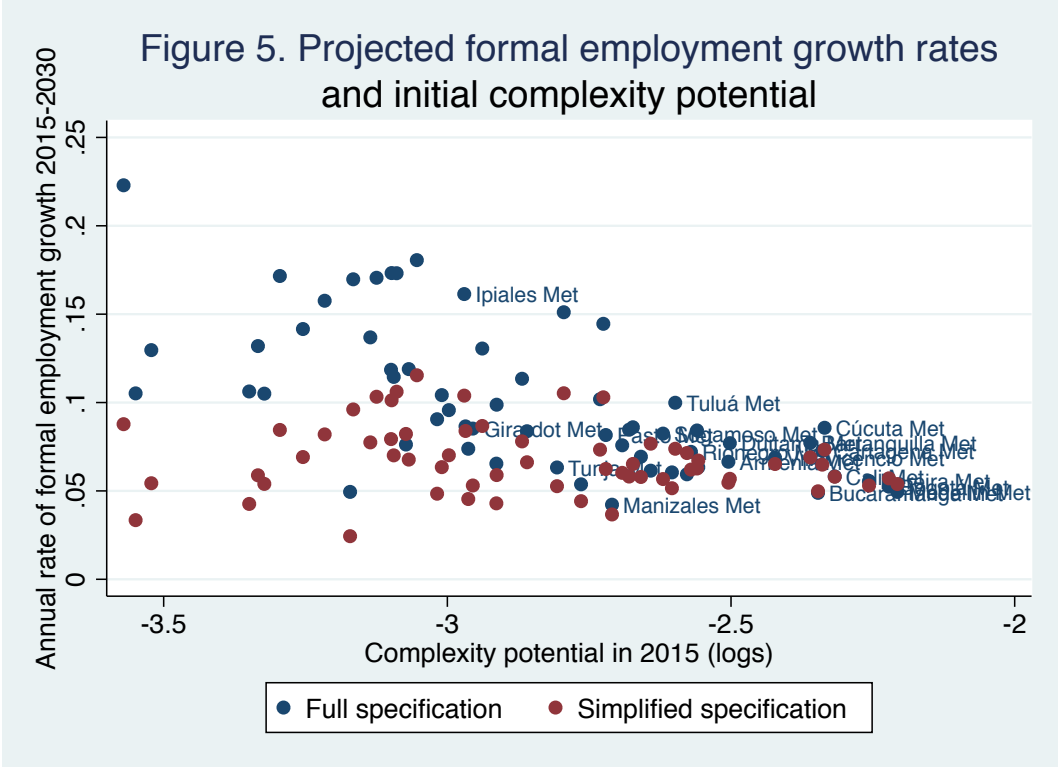


Figure 5. Regression-based forecasts of formal employment growth rates by city show high dispersion among cities whose initial complexity potential is low.

Although the theoretical framework emphasizes the importance of complexity potential, it may not be the unique factor influencing the forecasts, as suggested by Figure 3: with the full specification, that includes other variables, many of the low-complexity cities show high formal employment rates, which is not apparent in the simplified specification. In the latter, the fastest growing cities have medium levels of initial complexity potential.

To conclude the presentation of the regression-based forecasts, Table 3 shows the aggregates of the most relevant results. In 2015, the formal employment rate in the urban areas was 34 percent of the population in working age, and the average across cities 22 percent. Remember that our definition of formal employment takes into account the actual number of weeks of work of every employee. From this basis, the formal employment rate will probably reach between 63 and 66 percent in

2030, and the simple average will be between 43 and 59 percent, depending on the regression specification on which the forecasts are based. While formal employment in the 62 cities grew 8 percent per year between 2008 and 2015 (or 10.5% on average), it will probably slow down to a rate of growth of about 6 percent in the future (or between 7 and 10 percent on average). This is due to the fact that the largest cities will see more modest rates of formal employment growth. These results suggest that the choice of specification does not make much of a difference for the (weighted) aggregate of the 62 cities, this is certainly not the case for the simple averages or for the individual cities, as we have seen. That is where the machine learning techniques may be more adequate.

Table 3. Regression-based forecasts for the aggregate of the 62 cities

			Projected (2030)	
		Current	Full specification	Simplified specification
Formal employment rate	Weighted average	34.3%	66.1%	62.5%
	Simple average	22.0%	59.0%	43.0%
Formal employment growth rate	Weighted average	7.7%	6.3%	5.9%
	Simple average	10.5%	10.0%	6.8%

5. Machine learning forecasts

Machine learning is a type or artifical intelligence used to predict outcomes from input data without explicitly specifying the relation between the outcomes and the input data. The algorithms used in machine learning are able to discover the patterns in the data that best fit the outcomes, without any theory or model that relates the outcomes and the inputs.

I will use the machine learning technique known as “random forest”, which is typically applied to predicting categories of an outcome using random subsets of the data to randomly constructed decision trees. A decision tree is simply a step by step process to decide a category something belongs to.

It should be noted that there are two types of randomness in random forests. One is the random selection of the data in each subset and the other is the random branching or splitting of the inputs in the subset. The two types of randomness are

ways to prevent overfitting and determine how reliable the predictions are (for an intuitive introduction to random forests see Hartshorn, 2016).

Several decisions must be made to apply the random forest technique. Basically:

- Outcome categories must be defined. In our case, the outcome is the dependent variable defined in equation (9) and the categories will be its quartiles. Since I use the 434 observations of the 1-year intervals (as in the middle and lower panel regressions in Table 2), each quartile contains 108 or 109 observations. The program's objective will be to predict the category to which each observation belongs.
- Input data must be selected: I will use the same set of explanatory variables in the "full specification" (listed in the middle panel of Table 2). Since I want to make predictions of the outcome categories for 2030, I also include the input data for that year (the same used in the regression-based forecasts).
- Input data categories: although it is not strictly necessary to "discretize" the input data, it improves the reliability of the results when the number of observations is small, as in our case. I have constructed deciles of each variable for the 434 observations between 2008 and 2015, except the dummy for oil producing cities. I then applied the categorization criteria to the 62 observations of the 2030 input data.
- Number of trees, or simulations: 1000.
- Other: although many features of the program may be modified, I have used the default options in the Stata program for random forests.

The prediction scores are summarized in Table 4. The "success rate" for the whole sample is 78 percent, meaning by that the percent of outcomes predicted in the correct outcome category (listed in the first column). The success rates of each of the categories range between 86 percent for category 1 (slowest speed of formal employment change) and 72 percent for category 4 (fastest). Keep in mind that, since there are four categories, the expected success rate of a completely random prediction would be 25 percent in each category (and therefore in the total as well).

The success rate should not be confused with the probability that the category predicted for an individual outcome is the correct one. Since each of the 434 individual outcomes will enter in many of the simulations (more exactly 63.2 percent of the simulations, see Hartshorn, 2016), the program computes the percent of those cases in which it has made the correct prediction. The last column of Table 4 shows that, on average, that probability is 44 percent (and very similar for each of the categories).

Table 4. Score summary of machine learning predictions

Annual speed towards full employment category	Falsely predicted	Correctly predicted	Total number of cases	Success rate	Mean probability of the correct predictions
1=Less than 0.05 pp	15	94	109	86%	46%
2=Between 0.05 and 0.28 pp	28	80	108	74%	40%
3=Between 0.28 and 0.54 pp	24	85	109	78%	42%
4=More than 0.54 pp	30	78	108	72%	48%
Total	97	337	434	78%	44%

Table 5 presents a summary of prediction scores for a selection of cities (all of them multi-municipality cities). For three of those, random forest predicts correctly the speed category every year between 2008 and 2015. Although the probability of each of those individual events is moderate (again, around 44 percent), the consistency of the prediction suggests, for instance, that it is highly reliable that Barranquilla and Rionegro belong to speed category 3, while Ipiales belongs to speed category 1. At the bottom of the table is Bogotá, with only three correct predictions that it belongs to category 4 (the fastest).

Table 5. Score of past formal employment change predictions by machine learning, selected cities

City	Number of correct predictions 2008-2015 (out of 7)	Median growth group predicted 2008-2015	Mean probability of belonging to growth group 2008-2015
Barranquilla Met	7	3	48%
Rionegro Met	7	3	44%
Ipiales Met	7	1	43%
Villavicencio Met	6	4	47%
Cúcuta Met	6	3	47%
Armenia Met	6	3	41%
Pereira Met	5	4	49%
Tunja Met	5	4	45%
Duitama Met	5	3	45%
Sogamoso Met	5	3	40%
Girardot Met	5	2	39%
Tuluá Met	5	1	38%
Cartagena Met	4	3.5	51%
Manizales Met	4	4	49%

Medellín Met	4	4	48%
Cali Met	4	3.5	44%
Bucaramanga Met	4	4	43%
Bogotá Met	3	4	45%

The objective of the exercise is to forecast the speed category of each city in the future. A summary of the results for the same selection of cities is presented in Table 6.

Table 6. Future formal employment change group predicted by machine learning

(groups of formal employment rate change: 1=Less than 0.05 pp
2=Between 0.05 and 0.28 pp
3=Between 0.28 and 0.54 pp
4=More than 0.54 pp)

City	Growth group predicted	Probability of belonging to group
Manizales Met	4	55%
Pereira Met	4	55%
Tunja Met	4	51%
Medellín Met	4	50%
Bogotá Met	4	48%
Cali Met	4	45%
Bucaramanga Met	4	43%
Villavicencio Met	4	42%
Armenia Met	4	39%
Rionegro Met	4	37%
Cúcuta Met	3	59%
Barranquilla Met	3	51%
Sogamoso Met	3	40%
Tuluá Met	3	38%
Cartagena Met	3	36%
Duitama Met	2	44%
Girardot Met	2	31%
Ipiales Met	1	41%

Most of the large cities belong to the fastest category of formal employment growth in the future, which in many cases differ from the past, as we will see below. The probability of that event is relatively high for some of those cities. Only three of the multi-municipality cities are classified in the slower categories. Appendix 6, which presents the complete list of cities, shows that 18 cities are classified in the slowest category, and in some cases with high probabilities. Most of those are small cities.

How different are these machine learning forecasts from the regression-based ones and the past records of the cities presented in the previous section? Table 7 focuses again in the same selection of cities, and complete results can be seen in Appendix 7. As the last column of the table indicates, in only a handful of the cities (Tunja, Manizales, Villavicencio and Pereira), do the three classifications coincide. This strongly suggests that the cities belong to the fastest group, where they are consistently classified. The machine-learning based forecasts are less optimistic than the ones based on the simplified regression (or the ones based in the full specification regression, which are all category 4 and not included in table), but more optimistic of what a simple extrapolation of the past would suggest.

Table 7. Comparison of regression and machine-learning predictions of future formal employment change

(groups of formal employment rate change: 1=Less than 0.05 pp
2=Between 0.05 and 0.28 pp
3=Between 0.28 and 0.54 pp
4=More than 0.54 pp)

City	2008-2015 median	Regression-based (simplified specification)	Machine-learning based	Number of same categories
Tunja Met	4	4	4	3
Manizales Met	4	4	4	3
Villavicencio Met	4	4	4	3
Pereira Met	4	4	4	3
Medellín Met	3	4	4	1
Rionegro Met	3	4	4	1
Bogotá Met	3	4	4	1
Armenia Met	3	4	4	1
Bucaramanga Met	3	4	4	1
Cali Met	3	4	4	1
Barranquilla Met	3	4	3	1
Cartagena Met	3	4	3	1
Sogamoso Met	3	4	3	1
Pasto Met	3	4	3	1
Cúcuta Met	3	4	3	1
Tuluá Met	1	4	3	0
Girardot Met	3	3	2	1
Pamplona	2	3	2	1
Duitama Met	3	4	2	0
Ipiales Met	1	3	1	1
Averages and percent same	3.0	3.9	3.3	14%

In order to compare the forecasts for 2030 by the different methods, the category predictions by machine learning must be converted into formal employment growth rates and then extrapolated to 2030. To that end, I assume that the value of the dependant variable (speed) in each category exactly corrresponds to the median of the category, which I then use to make the calculations. Figure 6 compare the forecasts by the three methods of the formality rates in 2030. Notice that the machine learning forecasts form four straight lines: each one of them corresponds to a speed category, given than I have used the same speed for all the cities in each category. As already mentioned, the machine learning predictions are less optimistic than the regression-based ones. Furthermore, for the cities classified in category 1 (slowest speed), formality rates will not change, according to the machine-learning forecast. Although most of these cities initially have low formality rates, two of them have initial formality rates about the average (Barrancabermeja and Buga) and one of them starts from a very high formality rate (Yopal).

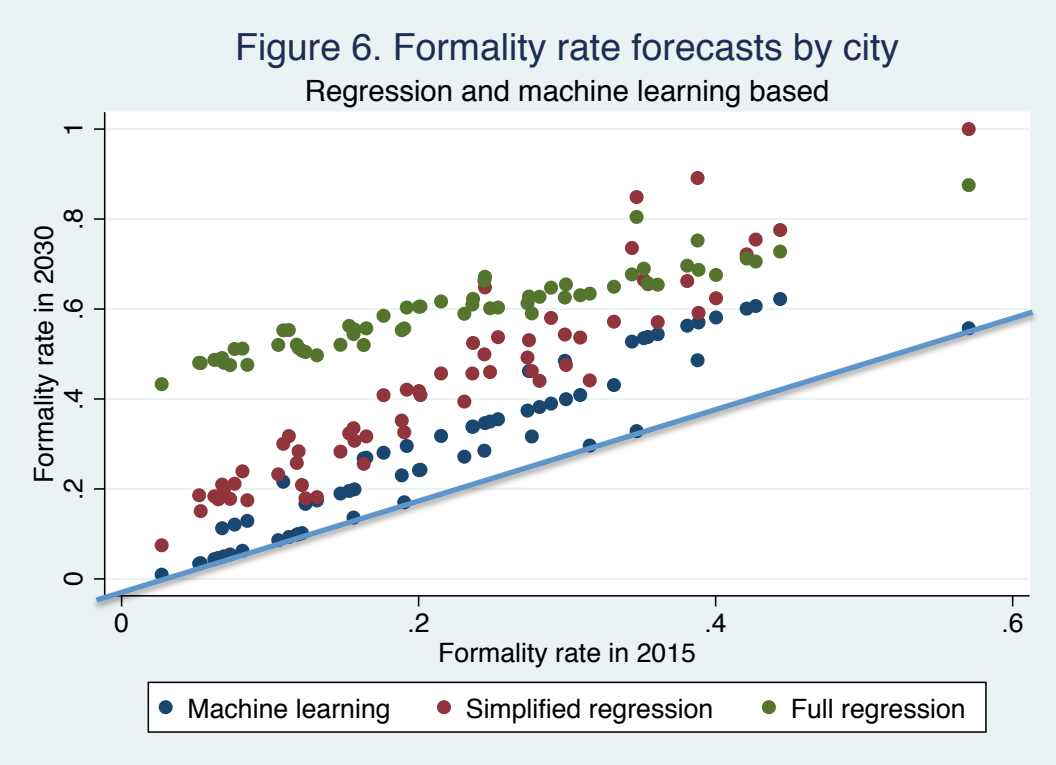


Figure 6. Machine-learning based forecasts of formal employment rates are lower and less differentiated by city than those based on regressions.

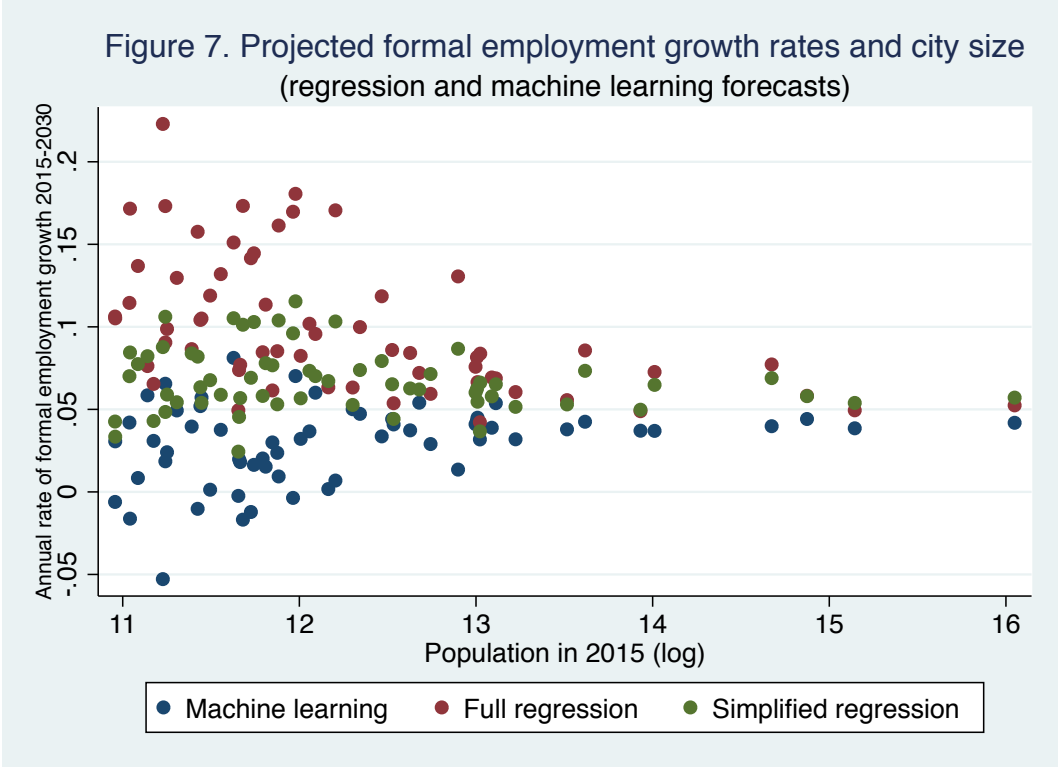


Figure 7. Machine-learning based forecasts of formal employment growth rates are lower than those based on regressions, especially for many of the smaller cities.

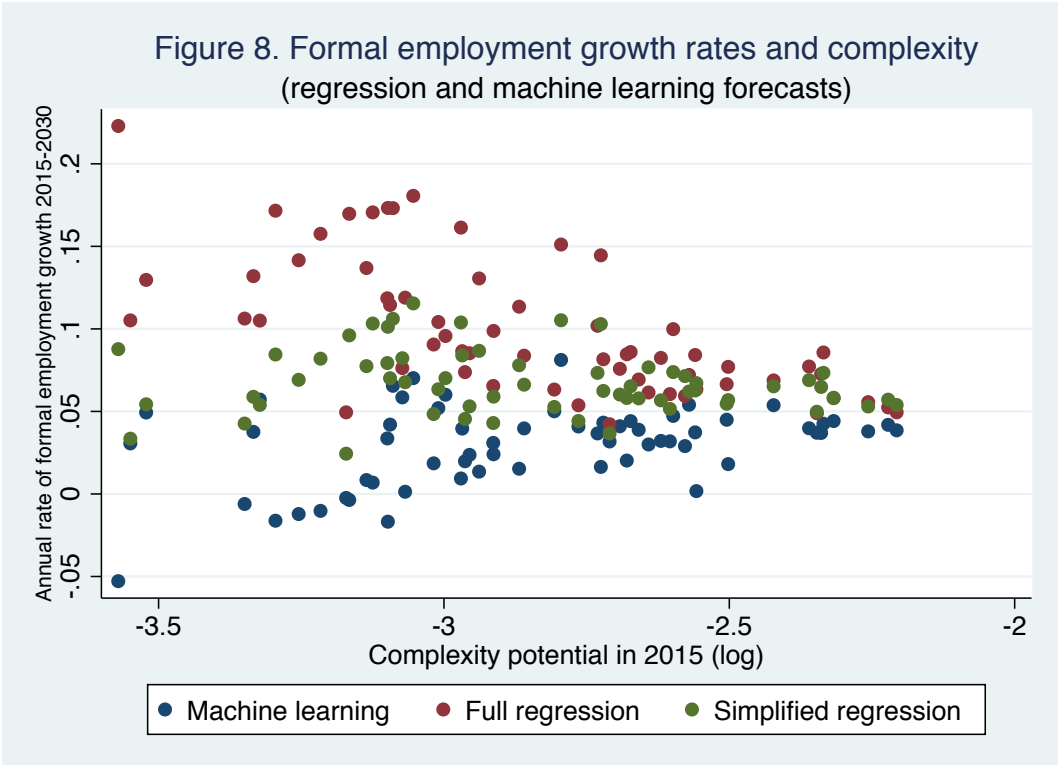


Figure 7. Machine-learning based forecasts of formal employment growth rates are much less disperse than those based on regressions, especially for many of the smaller cities.

Figure 7 shows that the formal employment growth rates of the three methods are similar for the largest cities but tend to diverge for smaller cities. The same pattern holds in relation to initial complexity potential.

Finally, to conclude the presentation of the results, Table 8, compares the aggregates of the 62 cities from the three methods. The formal employment rate for the aggregate, currently 34.3 percent, may reach between 47.9 percent and 66.1 percent, depending on the forecast method (and the simple average between 29.1 and 59.4 percent, starting from 22 percent). While in the period 2008-2015, total formal employment in the 62 cities grew 7.7 percent per annum, it may be expected to grow in the future between 4 and and 6.3 percent (simple average between 2.9 and 10 percent, compared with 10.5 percent in the recent past).

Table 8. Forecasts summary for the aggregate of the 62 cities

			Projected (2030)		
		Current	Regression-based, full specification	Regression based, simplified specification	Machine leaning based
Formal employment rate	Weighted average	34.3%	66.1%	62.5%	47.9%
	Simple average	22.0%	59.4%	43.0%	29.1%
Formal employment growth rate	Weighted average	7.7%	6.3%	5.9%	4.0%
	Simple average	10.5%	10.0%	6.8%	2.9%

6. Discussion

In order to assess these results, it must be recalled that the definition of formal employment used in this paper is *not* the share of the *occupied* that had *some formal employment* or *social security* in the reference period. With the formal employment criterion used by DANE (employees in establishments of more than 5 workers) and a 3-month (rolling) reference period, the formality rate in 2015 in the 23 largest cities and their metropolitan areas was 50.7 percent. With the social security criterion, it was either 64.6 or 46.8 percent, depending on whether social security

affiliation refers to health or pensions. In any of these definitions, there is only one margin through which the formality rate may increase, which is the status (either formal or informal) of the occupied. In my definition, there are four margins, as can be seen in this expression, which is an expansion of equation (7):

$$F_{c,t} = \frac{emp_{c,t}}{pop_{c,t}} = \left(\frac{emp_{c,t}}{workers_{c,t}}\right) * \left(\frac{workers_{c,t}}{occupied_{c,t}}\right) * \left(\frac{occupied_{c,t}}{laborforce_{c,t}}\right) * \left(\frac{laborforce_{c,t}}{pop_{c,t}}\right) \quad (9)$$

$$F_{c,t} = \frac{emp_{c,t}}{pop_{c,t}} = work\ intesity\ rate_{c,t} * official\ formality\ rate_{c,t} * \left(1 - unemployment\ rate_{c,t}\right) * participation\ rate_{c,t} \quad (10)$$

where the work intensity rate is the share of the year *t* that workers on average effectively contribute to the social security system, given my definition of $emp_{c,t}$. My formal employment rate and the official formality rate would move proportionally as long as the three other margins remain unchanged. If so, the official formality rate would go up from a range between 46.8 and 64.6 percent, as we have just seen, to a range between 65.3 percent and 90.2 percent in the machine-learning based forecast. But this conclusion is unwarranted because, although I have not explicitly modelled the three other margins (ie the work intensity, the unemployment and the participation margins) they are implicitly considered in the forecasts and it would not be reasonable to expect substantial increases in the official formality rate without increases in the other rates. As argued before, the official definitions of (in)formality are not adequate to assess the feasibility of the sustainable development goal of "full and productive employment and decent work for all women and men". My definition is much better suited to this end.

Being so, it is abundantly clear from the forecasts that reaching the full employment goal lies much further in the future than 2030. This does not contradict the finding that, most likely, formality rates will increase in most if not all Colombian cities larger than 50,000 inhabitants. Also, it does not deny that the different forecast methods consistently indicate that the formal employment growth rates in the largest cities will be about 5 percent. However, there is much less consistency in the predictions for the mid-size and smaller cities, many of which are not very optimistic.

Given the limitations of the regression-based forecasts, the machine-learning based one should be given serious consideration. The main strength of the latter lies not in its ability to predict aggregates, but in all the nuances it provides with respect to the individual predictions. For some of the smaller cities (such as Carmen de Bolívar and Chiquinquirá), it predicts with confidence that formal employment rates will stagnate at their low initial level, contrary to what the full specification regression would suggest. In other cases (such as Tunja and Popayán), it strongly predicts a fast process of labor formalization, consistent with the still incipient past tendencies, but also with the predictions based on regressions. Yet in others, the predictions not

only differ widely across methods, but those by machine-learning are statistically weak (Fusagasugá, Tulúa).

As argued in the theoretical section and shown in the regression results, complexity potential is the strongest and most consistent predictor of formal employment rate changes in cities. However, the machine-learning method suggests that the relation between the two variables is less straightforward than assumed in the regression-based methods. Further research is needed to understand how the ability of cities to make use of their skill mix in developing new industries may be affected by urban features such as density, availability of transportation means, women's access to work places, etc.

References

Albrecht, James, Lucas Navarro, and Susan Vroman. 2009. "The effects of Labour Market Policies in an Economy with an Informal Sector". *Economic Journal*, 119(539): 1105-29.

Bosch, Mariano, and William F. Maloney. 2010. "Comparative analysis of labor market dynamics using Markov processes: An application to informality". *Labour Economics*, 17(4): 621-31.

De Soto, Hernando. 1989. The Other Path: The Invisible Revolution in the Third World. New York: Harper and Row.

_____. 2000. The Mystery of Capital: Why Capitalism Triumphs in the West and Fails Everywhere Else. New York: Basic Books.

Duranton, Gilles. 2015. Delineating Metropolitan Areas: Measuring Spatial Labour Market Networks Through Commuting Patterns. In: Watanabe T., Uesugi I., Ono A. (eds) The Economics of Interfirm Networks. *Advances in Japanese Business and Economics*, vol 4. Springer, Tokyo.

Gollin, D., Jedwab, R. & Vollrath D., "Urbanization with and without Industrialization", *Journal of Economic Growth* (2016) 21: 35. https://doi-org.ezp-prod1.hul.harvard.edu/10.1007/s10887-015-9121-4

Harris, John R., and Michael P. Todaro. 1970. "Migration, Unemployment, and Development: A Two-Sector Analysis." American Economic Review 60(1): 126–42.

Hartshorn, Scott. *Machine Learning With Random Forests And Decision Trees: A Visual Guide For Beginners*. Kindle Edition, 2016.

Hidalgo, César and Ricardo Hausmann 2009. "The Building Blocks of Economic Complexity", *Proceedings of the National Academy of Sciences*, 106(26): 10570-5. DOI:10.1073/pnas.0900943106.

Kugler, Adriana, Maurice D. Kugler and Luis O. Herrera-Prada. 2017. "Do Payroll Tax Breaks Stimulate Formality? Evidence from Colombia's Reform," *Economia, Journal of the Latin American and Caribbean Economic Association*, Fall 2017: 3-40.

Levy, Santiago. 2008. *Good Intentions, Bad Outcomes: Social Policy, Informality, and Economic Growth in Mexico*. Brookings Institution Press.

Lewis, W. Arthur. 1954. "Economic Development with Unlimited Supplies of Labor." Manchester School of Economic and Social Studies 22(2): 139–91.

McGuire, T. J., Bartik, T. J., 1991. Who benefits From state and local economic development policies? JSTOR.

Meghir, Costas, Renata Narita, and Jean-Marc Robin. 2015. "Wages and Informality in Developing Countries". *American Economic Review*, 105: 1509-46.

Neffke, Frank and Martin Henning. 2013. "Skill Relatedness and Firm Diversification", *Strategic Management Journal*, 34(3): 297-316

O'Clery, N., Chaparro, J.C., Gómez-Liévano, A., *Lora, E. 2019. "Skill Diversity and the Evolution of Formal Employment in Cities", submitted to *Research Policy.*

Rauch, James E. 1991. "Modeling the Informal Sector Formally." *Journal of Development Economics* 35(1): 33–47.

Ulyssea, Gabriel. 2010. "Regulation of entry, labor market institutions and the informal sector". *Journal of Development Economics*, 91(1): 87 -99.

Appendix 1 - Calculation Methods for Industry Complexity

This appendix explains the methods for calculating the industry complexity variable introduced at the end of Section 2. It is adapted from Hidalgo and Hausmann (2009) and Neffke and Henning (2013). The actual calculations used formal employment data of all industries producing either goods or services (ISIC-AC, Rev. 3, at 4 digits, using social security data from PILA).

In the equations below, the sub-index *c* indicates cities and the sub-index *p* indicates industries. While no time sub-index is used here, all calculations are applied for each year separately (2008-2015).

Calculation of Revealed Comparative Advantages

The computation starts with data for employment by industry, city and year, organized in matrix form:

$$X_{cp}$$

From this matrix, the following aggregates are computed:

$$X_c = \sum_p X_{cp}$$

$$X_p = \sum_c X_{cp}$$

$$X = \sum_c \sum_p X_{cp}$$

These metrics are used to calculate the Revealed Comparative Advantage (RCA) for each city/industry combination:

$$RCA_{cp} = \frac{X_{cp}/X_p}{X_c/X}$$

Diversity and Ubiquity Calculations

The RCA matrix is transformed in a binary matrix depending on whether a particular value is larger than 1 or not:

$$M_{cp} = \begin{cases} 1 & RCA_{cp} \geq 1 \\ 0 & RCA_{cp} < 1 \end{cases}$$

This matrix indicates the industries that are relatively large in each city. This matrix is then used to compute the Diversity indicator at the city level, and the Ubiquity indicator at the industry level –that is, the count of the number of industries with relatively large employment for each city, and the count of the cities that have a given industry with a relatively high intensity:

$$k_{c,0} = \sum_{p} M_{cp} \qquad k_{p,0} = \sum_{c} M_{cp}$$

Industry Economic Complexity

The complexity of an industry can be measured by its ubiquity weighed by the diversity of the localities that have revealed comparative advantage in such industry. Extending this exercise *ad infinitum*, correcting diversity with ubiquity and vice-versa with consecutive iterations, is called the *method of reflections*. It can be expressed as follows:

$$k_{c,n} = \frac{1}{k_{c,0}} \sum_{p} M_{cp} \frac{1}{k_{p,0}} \sum_{c'} M_{c'p} k_{c',n-2}$$

$$= \sum_{c'} k_{c',n-2} \sum_{p} \frac{M_{c'p} M_{cp}}{k_{c,0}\, k_{p,0}}$$

$$= \sum_{c'} k_{c',n-2}\, \widetilde{M}^{c}_{c,c'}$$

Where:

$$\widetilde{M}^{c}_{c,c'} \equiv \sum_{p} \frac{M_{c'p} M_{cp}}{k_{c,0}\, k_{p,0}}$$

Using vector notation, the calculation method can be written in a compact manner as:

$$\vec{\boldsymbol{k}}_{\boldsymbol{n}} = \widetilde{\boldsymbol{M}}^{\boldsymbol{c}} \times \vec{\boldsymbol{k}}_{\boldsymbol{n-2}}$$

when $n \to \infty$, the following expression obtains:

$$\widetilde{\boldsymbol{M}}^{\boldsymbol{C}} \times \vec{\boldsymbol{k}} = \boldsymbol{\lambda}\vec{\boldsymbol{k}}$$

Where $\vec{\boldsymbol{k}}$ is an eigenvector of $\widetilde{\boldsymbol{M}}^{\boldsymbol{C}}$.

The second largest eigenvector of $\widetilde{\boldsymbol{M}}^{P}$ is taken as the Industry Complexity Index. The Index is calculated on employment levels per industry/city combination, including only industries with at least 50 formal employees in an average month, and only cities with at least 10 industries with 50 or more formal employees.

Appendix 2. Regressions of speed towards full formal employment on complexity potential and other controls

(Pooled ordinary least squares for different intervals, with year dummies)

Full 7-year period	**Coefficient**	**Standard error**	**t statistic**	**P>|t|**
Complexity potential at t-7 (log)	0.003043	0.0007914	3.85	0
Working age population at t-7 (log)	-0.0006131	0.0003166	-1.94	0.058
Formality rate at t-7 (logistic)	0.1132962	0.046996	2.41	0.019
Oil producing city	0.0037701	0.0007497	5.03	0
Bartik shock between t-7 and t	-0.0419715	0.0237082	-1.77	0.082
Constant	-0.0388139	0.0235932	-1.65	0.106

Number of obs = 62

Adj R-squared = 0.5891

6-year intervals	**Coefficient**	**Standard error**	**t statistic**	**P>|t|**
Complexity potential at t-6 (log)	0.0030322	0.0005672	5.35	0
Working age population at t-6 (log)	-0.0005583	0.0002257	-2.47	0.015
Formality rate at t-6 (logistic)	0.0777203	0.026448	2.94	0.004
Oil producing city	0.0034717	0.0005683	6.11	0
Bartik shock between t-6 and t	-0.0258881	0.0154644	-1.67	0.097
Constant	-0.0221075	0.0132719	-1.67	0.098
Year dummies	F(1, 117) =		8.784	0.004

Number of observations = 124

Adjusted R-squared = 0.5776

5-year intervals	**Coefficient**	**Standard error**	**t statistic**	**P>|t|**
Complexity potential at t-5 (log)	0.0029394	0.000478	6.15	0
Working age population at t-5 (log)	-0.0004868	0.0001827	-2.66	0.008
Formality rate at t-5 (logistic)	0.0371817	0.0154663	2.4	0.017
Oil producing city	0.0027807	0.0004671	5.95	0
Bartik shock between t-5 and t	-0.0046998	0.0114487	-0.41	0.682

Constant	-0.0031799	0.007911	-0.4	0.688
Year dummies	F(2, 178) =		2.3	0.103

Number of observations = 186

Adjusted R-squared = 0.5334

4-year intervals	**Coefficient**	**Standard error**	**t statistic**	**P>|t|**
Complexity potential at t-4 (log)	0.0029197	0.0004596	6.35	0
Working age population at t-4 (log)	-0.0004501	0.0001706	-2.64	0.009
Formality rate at t-4 (logistic)	0.0158137	0.0133056	1.19	0.236
Oil producing city	0.0022181	0.0004436	5	0
Bartik shock between t-4 and t	0.0154851	0.0119289	1.3	0.195
Constant	0.0070455	0.0069345	1.02	0.311
Year dummies	F(3, 239) =		6.548	0

Number of observations = 248

Adjusted R-squared = 0.514

3-year intervals	**Coefficient**	**Standard error**	**t statistic**	**P>|t|**
Complexity potential at t-3 (log)	0.0029632	0.0004778	6.2	0
Working age population at t-3 (log)	-0.0005133	0.0001734	-2.96	0.003
Formality rate at t-3 (logistic)	-0.0015121	0.0120469	-0.13	0.9
Oil producing city	0.0018829	0.0004502	4.18	0
Bartik shock between t-3 and t	0.0446801	0.0134568	3.32	0.001
Constant	0.0166122	0.0064531	2.57	0.011
Year dummies	F(4, 300) =		6.922	0

Number of observations = 310

Adjusted R-squared = 0.5149

2-year intervals	**Coefficient**	**Standard error**	**t statistic**	**P>|t|**
Complexity potential at t-2 (log)	0.0032913	0.0005331	6.17	0
Working age population at t-2 (log)	-0.0006558	0.0001903	-3.45	0.001
Formality rate at t-2 (logistic)	-0.0025988	0.0124833	-0.21	0.835
Oil producing city	0.001869	0.0004873	3.84	0

Bartik shock between t-2 and t	0.0717888	0.0178002	4.03	0
Constant	0.0199271	0.0068108	2.93	0.004
Year dummies	F(5, 361) =		8.34	0

Number of observations = 372

Adjusted R-squared = 0.5402

1-year intervals (full specification)	**Coefficient**	**Standard error**	**t statistic**	**P>|t|**
Complexity potential at t-1 (log)	0.0033963	0.0006686	5.08	0
Working age population at t-1 (log)	-0.0006598	0.0002322	-2.84	0.005
Formality rate at t-1 (logistic)	-0.0272684	0.0122967	-2.22	0.027
Oil producing city	0.0016853	0.0005864	2.87	0.004
Bartik shock between t-1 and t	0.2048173	0.0303162	6.76	0
Constant	0.0329708	0.0071898	4.59	0
Year dummies	F(6, 422) =		5.841	0

Number of obs = 434

Adjusted R-squared = 0.5020

1-year intervals (simplified specification)	**Coefficient**	**Standard error**	**t statistic**	**P>|t|**
Complexity potential at t-1 (log)	0.0030968	0.0003987	7.77	0
Oil producing city	0.0036794	0.0005224	7.04	0
Constant	0.0118205	0.0011629	10.16	0
Year dummies	F(6, 422) =		36.571	0

Number of obs = 434

Adjusted R-squared = 0.4331

Appendix 3. Current and projected formality rates

(ordered by mid projection)

City	Current (2015)	Projected (2030)	
		Full specification	Simplified specification
Yopal	57%	88%	100%
Medellín Met	44%	73%	78%
Bogotá Met	43%	71%	75%
Bucaramanga Met	42%	71%	72%
Manizales Met	40%	68%	62%
Tunja Met	39%	69%	59%
Neiva	39%	75%	89%
Villavicencio Met	38%	70%	66%
Popayán	36%	65%	57%
Cali Met	35%	66%	66%
Pereira Met	35%	69%	66%
Barrancabermeja	35%	80%	85%
Acacías	34%	68%	74%
Ibagué	33%	65%	57%
Guadalajara de Buga	32%	63%	44%
Santa Marta	31%	63%	54%
San Andrés	30%	65%	48%
Rionegro Met	30%	63%	54%
Cartagena Met	29%	65%	58%
Apartadó	28%	63%	44%
Valledupar	28%	59%	46%
Armenia Met	27%	63%	53%
Montería	27%	61%	49%
Barranquilla Met	25%	60%	54%
Pasto Met	25%	60%	46%
Arauca	24%	67%	65%
Duitama Met	24%	66%	50%
Cúcuta Met	24%	62%	52%
Sincelejo	24%	61%	46%
Quibdó	23%	59%	39%
Palmira	22%	62%	46%
Florencia	20%	61%	41%
Cartago	20%	60%	42%
Sogamoso Met	19%	60%	42%
Riohacha	19%	56%	33%
Girardot Met	19%	55%	35%

Tuluá Met	18%	58%	41%
Aguachica	16%	56%	32%
Santander de Quilichao	16%	52%	26%
Espinal	16%	55%	31%
Fusagasugá	16%	54%	33%
La Dorada	15%	56%	32%
Granada	15%	52%	28%
Pamplona	13%	50%	18%
Montelíbano	12%	50%	18%
Fundación	12%	51%	21%
Buenaventura	12%	51%	28%
Ocaña	12%	52%	26%
Pitalito	11%	55%	32%
Caucasia	11%	55%	30%
Chiquinquirá	11%	52%	23%
Ciénaga	8%	48%	17%
Ipiales Met	8%	51%	24%
Chigorodó	8%	51%	21%
Magangué	7%	48%	18%
San Andres de Tumaco	7%	48%	20%
Turbo	7%	49%	21%
Cereté	7%	49%	18%
Maicao	6%	49%	18%
Corozal	5%	48%	15%
Lorica	5%	48%	19%
El Carmen de Bolívar	3%	43%	7%
Total 62 cities	34%	66%	63%
Correlation with past	100%	95%	95%

Appendix 4. Past and projected formal employment growth rates

(ordered by mid projection)

City	Past (2008-2015)	Projected (2015-2030)	
		Full specification	Simplified specification
Fusagasugá	22%	11%	8%
Aguachica	21%	10%	6%
Magangué	20%	14%	7%
Acacías	19%	8%	8%
Granada	18%	11%	7%
Yopal	17%	6%	8%
Ocaña	16%	12%	7%
Lorica	16%	17%	10%
Quibdó	16%	7%	5%
Pitalito	15%	14%	10%
Ciénaga	14%	13%	6%
Valledupar	14%	8%	7%
Villavicencio Met	14%	7%	7%
Girardot Met	13%	9%	5%
San Andres de Tumaco	13%	17%	10%
El Carmen de Bolívar	13%	22%	9%
Maicao	12%	17%	10%
Montería	12%	8%	6%
Chiquinquirá	12%	14%	8%
Sincelejo	11%	9%	7%
Pasto Met	11%	8%	6%
Caucasia	11%	15%	11%
Neiva	11%	6%	7%
Pamplona	11%	11%	3%
Rionegro Met	10%	7%	6%
Popayán	10%	5%	4%
Ipiales Met	10%	16%	10%
Arauca	10%	9%	8%
Cartagena Met	10%	7%	6%
Fundación	10%	11%	4%
Chigorodó	10%	17%	11%
Florencia	10%	10%	7%
Bucaramanga Met	10%	5%	5%
Ibagué	10%	6%	5%
Cúcuta Met	9%	9%	7%
Santa Marta	9%	7%	6%

Riohacha	9%	12%	8%
Buenaventura	9%	13%	9%
Armenia Met	9%	7%	5%
Barrancabermeja	9%	6%	7%
Corozal	8%	17%	8%
Cereté	8%	16%	8%
San Andrés	8%	7%	4%
Barranquilla Met	8%	8%	7%
Santander de Quilichao	8%	11%	5%
Tunja Met	8%	6%	5%
Manizales Met	8%	4%	4%
Sogamoso Met	7%	8%	6%
Espinal	7%	9%	5%
Bogotá Met	7%	5%	6%
Pereira Met	7%	6%	5%
Duitama Met	7%	8%	6%
La Dorada	7%	10%	6%
Cartago	7%	8%	6%
Cali Met	7%	6%	6%
Apartadó	6%	10%	7%
Montelíbano	6%	13%	5%
Medellín Met	6%	5%	5%
Turbo	6%	18%	12%
Palmira	5%	8%	6%
Tuluá Met	3%	10%	7%
Guadalajara de Buga	1%	5%	2%
Total 62 cities	8%	6%	6%
Correlation with past	100%	24%	27%

Appendix 5. Score of past formal employment change predictions by machine learning

City	Number of correct predictions 2008-2015 (out of 7)	Median growth group predicted 2008-2015	Mean probability of belonging to growth group 2008-2015
Yopal	7	4	62%
Neiva	7	4	52%
Barranquilla Met	7	3	48%
San Andrés	7	3	45%
Rionegro Met	7	3	44%
Ipiales Met	7	1	43%
Cartago	7	2	41%
Florencia	7	2	39%
Apartadó	7	2	38%
Chigorodó	6	1.5	53%
El Carmen de Bolívar	6	1	52%
Turbo	6	1.5	50%
Villavicencio Met	6	4	47%
Cúcuta Met	6	3	47%
Arauca	6	3	46%
Santander de Quilichao	6	1.5	46%
Chiquinquirá	6	1	46%
Magangué	6	2	45%
Quibdó	6	2	45%
Popayán	6	3.5	45%
Ibagué	6	3.5	44%
Acacías	6	4	43%
Pasto Met	6	3	43%
Montelíbano	6	1.5	42%
Guadalajara de Buga	6	2	41%
Armenia Met	6	3	41%
Valledupar	6	2.5	41%
Pamplona	6	2	40%
Riohacha	6	1	40%
Palmira	6	2	38%
Caucasia	6	2	37%
Barrancabermeja	5	4	51%
Lorica	5	1	49%
Cereté	5	1	49%
Pereira Met	5	4	49%
Maicao	5	1	48%

Tunja Met	5	4	45%
Duitama Met	5	3	45%
San Andres de Tumaco	5	2	45%
La Dorada	5	2	45%
Montería	5	3	42%
Buenaventura	5	1	40%
Pitalito	5	2	40%
Sogamoso Met	5	3	40%
Ocaña	5	2	40%
Girardot Met	5	2	39%
Santa Marta	5	3	39%
Espinal	5	2	38%
Tuluá Met	5	1	38%
Sincelejo	5	2	38%
Fusagasugá	5	3	37%
Corozal	4	1	52%
Cartagena Met	4	3.5	51%
Manizales Met	4	4	49%
Medellín Met	4	4	48%
Ciénaga	4	1.5	48%
Cali Met	4	3.5	44%
Bucaramanga Met	4	4	43%
Granada	4	1.5	40%
Aguachica	4	2	39%
Bogotá Met	3	4	45%
Fundación	3	2	38%
Median	5.5	2	44%

Table 6. Future formal employment change group predicted by machine learning

(groups of formal employment rate change: 1=Less than 0.05 pp
2=Between 0.05 and 0.28 pp
3=Between 0.28 and 0.54 pp
4=More than 0.54 pp)

City	Growth group predicted	Probability of belonging to group
Popayán	4	59%
Manizales Met	4	55%
Pereira Met	4	55%
Tunja Met	4	51%
Medellín Met	4	50%
Acacías	4	48%
Bogotá Met	4	48%
Cali Met	4	45%
Bucaramanga Met	4	43%
Villavicencio Met	4	42%
Armenia Met	4	39%
Rionegro Met	4	37%
Cúcuta Met	3	59%
Arauca	3	59%
Barranquilla Met	3	51%
Montería	3	49%
San Andrés	3	47%
Palmira	3	46%
Santander de Quilichao	3	45%
Aguachica	3	44%
Neiva	3	43%
Santa Marta	3	43%
Pasto Met	3	43%
Sincelejo	3	43%
Ibagué	3	40%
Sogamoso Met	3	40%
Caucasia	3	40%
Apartadó	3	38%
Tuluá Met	3	38%
Cartagena Met	3	36%
Quibdó	2	51%
Chigorodó	2	50%

Espinal	2	46%
Cartago	2	45%
Duitama Met	2	44%
La Dorada	2	44%
Turbo	2	43%
Pamplona	2	42%
Valledupar	2	41%
Montelíbano	2	39%
Ciénaga	2	38%
Granada	2	38%
Florencia	2	37%
Girardot Met	2	31%
El Carmen de Bolívar	1	63%
Cereté	1	58%
Chiquinquirá	1	55%
Maicao	1	53%
Corozal	1	52%
Magangué	1	49%
San Andres de Tumaco	1	47%
Yopal	1	47%
Buenaventura	1	45%
Lorica	1	44%
Ocaña	1	44%
Barrancabermeja	1	43%
Guadalajara de Buga	1	42%
Ipiales Met	1	41%
Riohacha	1	37%
Pitalito	1	35%
Fundación	1	34%
Fusagasugá	1	30%

Appendix 7. Comparison of regression and machine-learning predictions of future formal employment change

(groups of formal employment rate change: 1=Less than 0.05 pp
2=Between 0.05 and 0.28 pp
3=Between 0.28 and 0.54 pp
4=More than 0.54 pp)

				Same predictions?			
City	2008-2015 median	Regression-based (simplified specification)	Machine-learning based	Median 2008-2015 and regression-based	Median 2008-2015 and machine-learning based	Regression-based and machine-learning based	Total (ou of 3)
Aguachica	3	3	3	1	1	1	3
Tunja Met	4	4	4	1	1	1	3
Manizales Met	4	4	4	1	1	1	3
Popayán	4	4	4	1	1	1	3
Villavicencio Met	4	4	4	1	1	1	3
Acacías	4	4	4	1	1	1	3
Pereira Met	4	4	4	1	1	1	3
San Andrés	3	3	3	1	1	1	3
Florencia	2	4	2	0	1	0	1
Santander de Quilichao	2	3	3	0	0	1	1
Lorica	1	3	1	0	1	0	1
Ocaña	3	3	1	1	0	0	1
Sincelejo	3	4	3	0	1	0	1
Medellín Met	3	4	4	0	0	1	1
Apartadó	2	3	3	0	0	1	1
Chigorodó	2	3	2	0	1	0	1
Rionegro Met	3	4	4	0	0	1	1
Turbo	2	3	2	0	1	0	1
Barranquilla Met	3	4	3	0	1	0	1
Bogotá Met	3	4	4	0	0	1	1
Cartagena Met	3	4	3	0	1	0	1
El Carmen de Bolívar	1	3	1	0	1	0	1
Chiquinquirá	1	3	1	0	1	0	1
Sogamoso Met	3	4	3	0	1	0	1
La Dorada	2	3	2	0	1	0	1
Montería	3	4	3	0	1	0	1
Cereté	1	3	1	0	1	0	1
Montelíbano	2	3	2	0	1	0	1
Girardot Met	3	3	2	1	0	0	1
Quibdó	2	3	2	0	1	0	1

Neiva	4	4	3	1	0	0	1
Riohacha	1	3	1	0	1	0	1
Santa Marta	3	4	3	0	1	0	1
Ciénaga	2	3	2	0	1	0	1
Granada	3	3	2	1	0	0	1
Pasto Met	3	4	3	0	1	0	1
Ipiales Met	1	3	1	0	1	0	1
Cúcuta Met	3	4	3	0	1	0	1
Pamplona	2	3	2	0	1	0	1
Armenia Met	3	4	4	0	0	1	1
Bucaramanga Met	3	4	4	0	0	1	1
Corozal	1	3	1	0	1	0	1
Ibagué	3	4	3	0	1	0	1
Espinal	2	3	2	0	1	0	1
Cali Met	3	4	4	0	0	1	1
Cartago	2	4	2	0	1	0	1
Arauca	3	4	3	0	1	0	1
Yopal	4	4	1	1	0	0	1
Duitama Met	3	4	2	0	0	0	0
Caucasia	2	4	3	0	0	0	0
Magangué	2	3	1	0	0	0	0
Valledupar	3	4	2	0	0	0	0
Fusagasugá	3	4	1	0	0	0	0
Pitalito	2	4	1	0	0	0	0
Maicao	2	3	1	0	0	0	0
Fundación	2	3	1	0	0	0	0
San Andres de Tumaco	2	3	1	0	0	0	0
Barrancabermeja	3	4	1	0	0	0	0
Buenaventura	2	3	1	0	0	0	0
Guadalajara de Buga	2	3	1	0	0	0	0
Palmira	2	4	3	0	0	0	0
Tuluá Met	1	4	3	0	0	0	0
Averages and percent same	2.5	3.5	2.4	21%	56%	26%	34%