Forecasting Formal Employment in Cities

Received: September 13th, 2020 –Accepted: November 4th, 2020

Doi: https://doi.org/10.12804/revistas.urosario.edu.co/economia/a.10029

Eduardo Lora*

Abstract

Can "full and productive employment for all" be achieved by 2030 as per UN Development Goal 8? The issue was assessed for 62 Colombian cities using administrative data. Larger cities have higher formal occupation rates because formal employment creation is restricted by the availability of the skills needed in complex sectors, which follows a path-dependent process. Alternative forecasts were produced using OLS and machine learning algorithms. Results: the share of the working population in formal employment will increase between 13 and 32 percentage points, which is insufficient to achieve the goal. Consistency across methods was good for large cities, not for small ones.

Keywords: Employment; cities; skills; forecasts; sustainable development; Colombia.
Classification JEL: J21.

* Senior Fellow, Center for International Development (CID), Harvard University and Associate Researcher, Research in Spatial Economics (RISE), Universidad Eafit. Principal mail address: Eduardo.A.Lora@gmail.com; secondary mail address: Eduardo_Lora@hks.harvard.edu. Address: 5203 Westbard Avenue, Bethesda-MD 20816, USA. Phone: +1-240-893-3142. ORCID:

Comments by Juan Pablo Chauvin, Johanna Fajardo, Ana María Ibáñez, Carmen Pagés, Mauricio Quiñones and participants at seminars held at the Inter-American Development Bank and Universidad Eafit are acknowledged. This research took place thanks to the "Urban PEAK" Program, with the support of UKRI's *Global Challenge, Research Fund* (Grant reference: ES/P011055/1).

To quote this article: Lora, E. (2021). Forecasting Formal Employment in Cities. *Revista de Economía del Rosario, 24*(1), 79-116. https://doi.org/10.12804/revistas.urosario.edu.co/economia/a.10029

ISSN 0123-5362 - ISSNE 2145-454x

Pronósticos de empleo formal urbano

Resumen

¿Se puede lograr el "empleo pleno y productivo para todos" en el 2030, según el Objetivo de Desarrollo 8 de la ONU? El asunto se evaluó para 62 ciudades colombianas, utilizando datos administrativos. Las ciudades más grandes tienen tasas de ocupación formal más altas porque la creación de empleo formal está restringida a la disponibilidad de las habilidades necesarias en sectores complejos, un proceso dependiente de la trayectoria. Se hicieron pronósticos alternativos utilizando OLS y algoritmos de aprendizaje automático. Resultados: el porcentaje de la población en edad de trabajar con un empleo formal aumentará entre 13 y 32 puntos porcentuales, lo que es insuficiente para alcanzar la meta. La coherencia entre los métodos es buena para las grandes ciudades, no para las pequeñas.

Palabras clave: empleo; ciudades; habilidades; pronósticos; desarrollo sostenible; Colombia.
Clasificación JEL: J21.

Previsões para o emprego formal urbano

Resumo

Um "emprego pleno e produtivo para todos" pode ser alcançado até 2030 de acordo com o Objetivo de Desenvolvimento 8 das Nações Unidas? O assunto é avaliado em 62 cidades colombianas com base em dados administrativos. As cidades maiores têm taxas de emprego formal mais altas porque a criação de empregos formais é restringida pela disponibilidade das habilidades necessárias em setores complexos, um processo dependente da trajetória. São realizadas previsões alternativas usando OLS e algoritmos de aprendizado automático. Resultados: o percentual da população em idade ativa com emprego formal aumentará entre 13 e 32 pontos percentuais, o que é insuficiente para atingir a meta. Coerência entre métodos: bom para cidades grandes, não para cidades pequenas.

Palavras-chave: emprego; cidades; habilidades; previsões; desenvolvimento sustentável; Colômbia.
Classificação JEL: J21.

Introduction

United Nations Sustainable Development Goal 8 is to "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 et al., 2016). Urbanization patterns, 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 reliable facts (O'Clery et al., 2020): (i) formal occupation rates are more variable across cities within countries than across developing countries and (ii) larger cities create proportionally more formal employment.

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 the large quantity of labor force 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 & 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 despite their low productivity by avoiding taxes and regulations. Lastly, in labor search models, which 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 et al., 2009; Bosch & Maloney, 2010; Ulyssea, 2010; Meghir et al., 2015).

While empirical evidence has been provided to support each of these explanations of informality, none of them recognizes the two facts mentioned in the introduction: 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, 2020), 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.

Accordingly, in O'Clery et al. (2019), the creation of formal employment between period t and period $t{+}n$ in city c, Δemp_C depends on "complexity potential" CP in period t, which is a measure of the availability of skills of the local labor force needed in the more complex industries not yet present in the city:

$$\Delta emp_C = emp_{C,t+1} - emp_{C,t} = \mathrm{f}\,(CP_{c,t}) \tag{1}$$

$$CP_{c,t} = \frac{1}{\left|M_{c,t}\right|}\sum\nolimits_{i\in M_{c,t}} d_{c,i}C_i \tag{2}$$

Notice that complexity potential is a weighted average of the *industry complexity* C_i of the *missing sectors* at time t $M_{c,\,t}$, with weights given by the *density* $d_{c,\,i}$ of industries that use *skills similar* to those of the missing sectors.

In order to operationalize equation (2), 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 (i.e., those industries that have revealed a comparative advantage greater than 1 in a city, based on formal employment shares) and the number of cities that have the industry (i.e., those cities where the industry has revealed a 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 a city are those with a revealed comparative advantage lower than 1.

Data and Empirical Definitions

Like in O'Clery et al. (2019, 2020), 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 % 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 was obtained.

The main data source for the 62 cities was the social security administrative data collected by the Ministerio de Salud y Seguridad Social (Health and Social Security Ministry), known as PILA (Planilla Integrada de Liquidación de Aportes). PILA contains information of workers and firms on the days worked, the sector of activity, and the municipality.[1] To aggregate these data, I count the share of the year t that each worker effectively contributed to the social

[1] The datasets have information on age and gender, which I did 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.

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):

$$f_{c,\,t} = emp_{c,\,t}/pop_{c,\,t} \tag{3}$$

The (simple) average formal occupation rate in cities was only 20.3 % of the working age population in 2015 with a relatively large standard deviation (11.1 %). Important changes in urban formal occupation rates occurred between 2008 and 2015: the *aggregate* formal occupation rate for the 62 cities went up from 21.1 to 31.2 % with a (simple) average increase across cities of 8.1 % and a standard deviation of 5.4 %. Formal occupation was facilitated by a rate of GDP growth of 4.1 % and probably by the elimination in May of 2013 of payroll taxes and surcharges representing up to 13.5 % of the wage bill of some groups of workers (Kugler et al., 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:

$$y_{c,t-i} = \frac{1}{t-i}\left(\frac{e^{f_{c,t}}}{1+e^{f_{c,t}}} - \frac{e^{f_{c,t-i}}}{1+e^{f_{c,t-i}}}\right) \tag{4}$$

Where $y_{c,\,t\,-\,i}$ will be the dependent 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*) were *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*).[2]

Two forecasting methods were used in a *complementary* way. As I will explain, the two methods look at the issue from different angles, the first being better grounded in theory but quite limited in the choice of specifications, and the second being more flexible operationally but theoretically *ad hoc*. The first method was 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 coefficients, two regressions were 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, was 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 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 independent variables (or combinations thereof) is 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. While the machine learning method is agnostic from a theoretical point of view and does not produce unbiased estimates of the coefficients that relate the dependent and the independent variables, it provides more nuanced forecasts at the city level, as I will discuss further below.

Regression-Based Forecasts

Table 1 is a summary of the regressions. Only the 7-year (i.e., full 2008-2015) and 1-year interval regressions are presented (see Appendix 1 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 seven one-year periods (434=62 x 7).

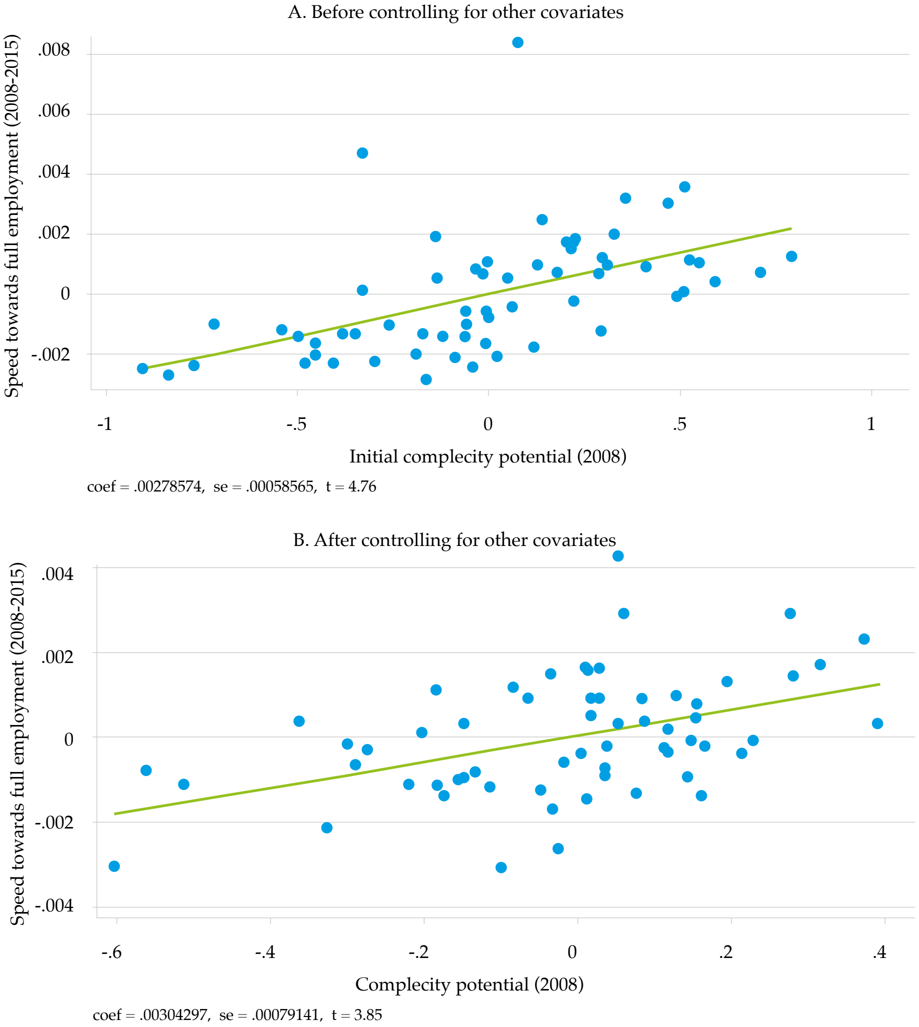
[2] In O'Clery et al. (2020) the measure of complexity potential depends on working age population, while here I am taking the latter as a separate explanatory variable. In this way, the relation between both variables can be explored in the machine learning exercises.

Table 1. Regressions of speed towards full formal employment on complexity potential an 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				
Adj R–squared= 0.5020				
1 – year intervals (simplified specification)	Coefficient	Standard error	t statistic	P > |t|
Complexity potential at t – 1 (log)	0.0038597	0.0007055	5.47	0
Working age population at t – 1 (log)	–0.0002965	0.0002262	–1.31	0.191
Oil producing city	0.0034578	0.0005486	6.3	0
Constant	0.0175317	0.0045106	3.89	0
Year dummies	F(6,422)=		36.605	0
Number of obs= 434				
Adj R–squared= 0.4341				

Source: Own calculations with Ministry of Health's PILA data.

Before explaining the results in detail, it is important to highlight that the main explanatory variable in all the regressions is the initial complexity potential: our measure of the initial availability in the city of the diversity of skills needed in the industries more complex than the industries already present in the city. Figure 1 shows (for the full period 2008-2015) the remarkable relationship between this variable and the speed towards full employment, the dependent variable. The left side panel shows the relation without any controls (apart from a constant term), and the right one shows the relation after controlling for the other explanatory variables, based on the regression of the upper panel.


A. Before controlling for other covariates

coef = .00278574, se = .00058565, t = 4.76

B. After controlling for other covariates

coef = .00304297, se = .00079141, t = 3.85

Figure 1. The relationship between complexity potential and speed towards full employment, before and after controlling for other covariates

The interpretation of the coefficients in Table 1 is not straightforward 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 relationship with the dependent variable is not fully captured: there may be important interactions between the explanatory variables or dynamic issues that are ignored in the specification adopted. Since the number of cities, as well as the number of periods is small, not much can be done to overcome these problems with standard econometrics.

The lower panel shows a simplified version of the 1-year interval regression, which only includes three explanatory variables. While complexity potential and the dummy variable for oil-producing cities are significantly and consistently directly or inversely associated with the dependent variable in other regressions, the working-age population is not. However, it is included because it is the variable that motivates the theoretical model summarized in a previous section.

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

- The complexity potential by city is assumed constant at the 2015 values.
- The working-age population by city is projected at the same growth rate observed between 2008 and 2015.
- The oil producing city dummy is kept unchanged throughout the forecast period.

In order to compute formal employment at the end of the forecast period, the formality rate at *t-1* (logistic) by city must be calculated recursively with the dependent variable's forecast for the previous year. This procedure has no bearing in the results.

Basing a 15-year forecast (i.e., 2015 to 2030) on eight years of observations is dictated by the availability of data and the purpose of the exercise. However, neither on theoretical nor on empirical grounds is the exercise far-fetched. The theory rests on the assumption that formal creation is an *evolutionary* process, which, as such, is reinforced rather than weakened through time. This is corroborated empirically by the fact that the 7-year specification (in the high panel above) has a similar explanatory power than the comparable 1-year specification (middle panel), in spite of the fact that the latter include time dummies. Furthermore, the coefficient of complexity potential (the key variable in the theoretical model) is highly significant and practically identical in both regressions.

The results appear in Figures 2-4 (and Appendices 1 and 2). Figure 2 shows that formality rates will increase throughout 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.

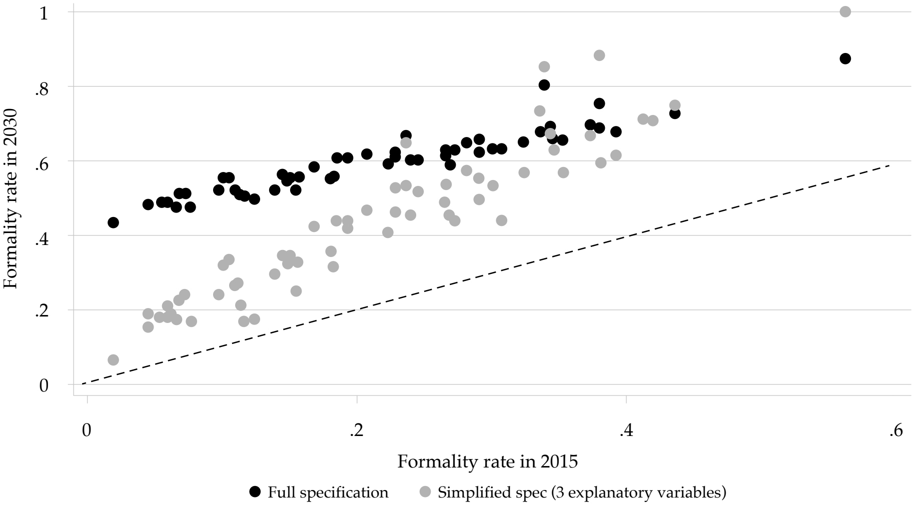


Figure 2. Formality rate forecasts by city

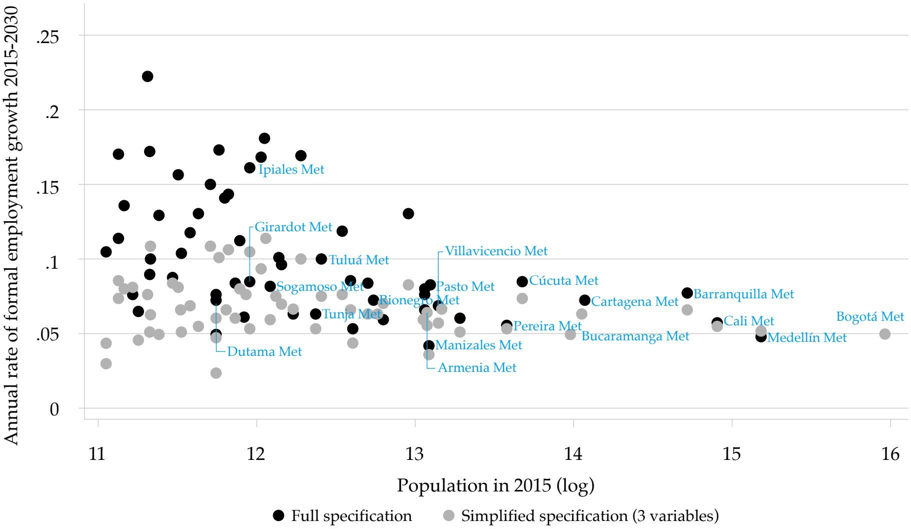


Figure 3. Projected formal employment growth rates and city size

Figure 3 makes clear that the differences between the two forecasts are strongly related to city size (although this variable is included in both regressions, it is not significant in the second one and its effect on the forecasts is extremely small): while for the smaller cities the rates of employment growth can differ by more than 10 %, for the largest cities, the differences are negligible (the figure shows only the names of the multi-municipality cities, most of which are also the largest ones).

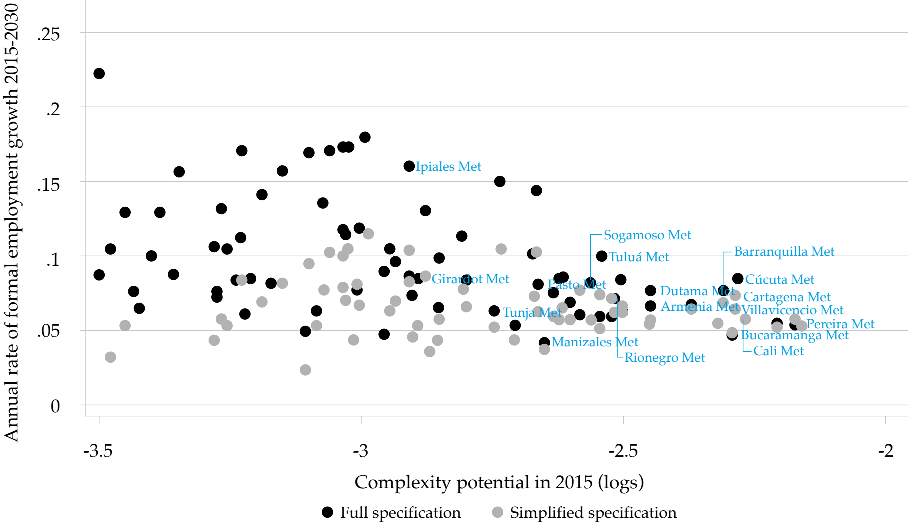


Figure 4. Projected formal employment growth rates and initial complexity potential

Although the theoretical framework emphasizes the importance of complexity potential (and Figure 1 above corroborates it), it may not be the unique factor influencing the forecasts, as suggested by Figure 2, 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 2 shows the aggregates of the most relevant results. In 2015, the formal employment rate in the urban areas was 34 % of the population of working age, and the average across cities, 22 %. 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 % in 2030, and the simple average will be between 43 and 59 %, depending on the regression specification on which the forecasts are based. While formal employment in the 62 cities grew 8 % per year between 2008 and 2015 (or 10.5 % on average), it will probably slow down to a rate of growth of about 6 % in the future (or between 7 and 10 % on average), 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 a substantial difference for the (weighted) aggregate of the 62 cities, but this is certainly not the case for the simple averages or for the individual cities, as we have seen. There is where machine learning techniques may be helpful.

Table 2. Regressions–based forecasts for the aggregate of the 62 cities

		Current	Projected (2030)	
			Full specifications	Simplified specifications
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 %

Source: Own calculations with Ministry of Health's PILA data.

Machine Learning Forecasts

Machine learning is a type of artificial 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.

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:

- The outcome categories must be defined. In this case, the outcome is the dependent variable defined in equation (4), and 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. To check the robustness of the categorization, I categorize the dependent variable in two other ways, using three and six groups, respectively, instead of the original four groups (see further below).
- The 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).
- The 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 is the case here. 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. To check the robustness of the categorization, I "discretize" the relevant variables in two additional ways using five and 15 groups, instead of the original 10 (see further below).
- The number of trees or simulations: 1.000.

- 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 3. The "success rate" for the whole sample was 78 %, meaning 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 % for category 1 (slowest speed of formal employment change) and 72 % 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 % in each category (and therefore in the total as well).

The success rate should not be mistaken 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 % 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 3 shows that, on average, that probability is 44 % (and very similar for each of the categories).

Table 3. 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 prediction
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 %

Source: Own calculations with Ministry of health's PILA data.

Table 4 presents a summary of the 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 %), the consistency of the prediction suggests that it is highly likely 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 4. 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 beloging to growth group 2008-2015
Barranquilla Meet	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 %

Source: Own calculations with Ministry of health's PILA data.

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 5.

Table 5. 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 %

City	Growth group predicted	Probability of belonging to group
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 %
Tulúa Met	3	38 %
Cartagena Met	3	36 %
Duitama Met	2	44 %
Girardot Met	2	31 %
Ipiales Met	1	41 %

Source: Own calculations with Ministry of health's PILA data.

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 5, which presents the complete list of cities, shows that 18 are classified in the slowest category and, in some cases, with high probabilities. Most of those are small cities.

As mentioned, for robustness, I have rerun the machine learning algorithms just explained using two alternative variable categorizations. In the "coarser" categorization, the dependent variable is "discretized" in three categories instead of four, and the explanatory variables are "discretized" in five groups instead of ten. In the "smoother" categorization, the dependent variable is "discretized" in six groups and the dependent variables in 15 groups. With the "coarser" categorization, the "success rate" for the whole sample falls from 78 % to 65 %, and the average probability of the correct predictions falls from 44 % to 36 %. With the "smoother" categorization, the success rate increases to 88 %, while the average probability of the correct predictions is 37 %. Therefore, the original categorization is intermediate between the two alternative ones in terms of the success rate but produces better results than the two alternatives in terms of confidence of the predictions. For this reason, I use it as my preferred categorization; the remainder discussion of the results refers to it.

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 6 focuses again on the same selection of cities used above (complete results can be seen in Appendix 6). As the last column of the table indicates, in only a handful of cities (Tunja, Manizales, Villavicencio, and Pereira), 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 the table but more optimistic than what a simple extrapolation of the past would suggest.

Table 6. Comparison of regression and machine-learning predictions of future formal employment change

(group 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	3
Rionegro Met	3	4	4	3
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

City	2008-2015 median	Regression-based (simplified specification)	Machine-learning based	Number of same categories
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	2	1
Average and percent same	3.0	3.9	3.3	14 %

Source: Own calculations with Ministry of health's PILA data.

While the previous comparisons across methods refer to the speed categories, it is also relevant to compare the forecasts for 2030 of the formal employment rates, to which the sustainable development target of full and decent employment refers. In order to do this, 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 dependent variable (speed) in each category corresponds to the median of the category, which I then use to make the calculations. Although it would be desirable to have a different speed for each city, this is not possible with the results of the machine learning technique, which only provides a classification by speed categories. Similarly, the technique does not provide any basis to establish whether a city may jump from one speed-category to another.

Figure 5 compares the forecasts by the three methods of formality rates in 2030. For this reason, the machine-learning forecasts form four straight lines: each one of them corresponds to a speed 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).

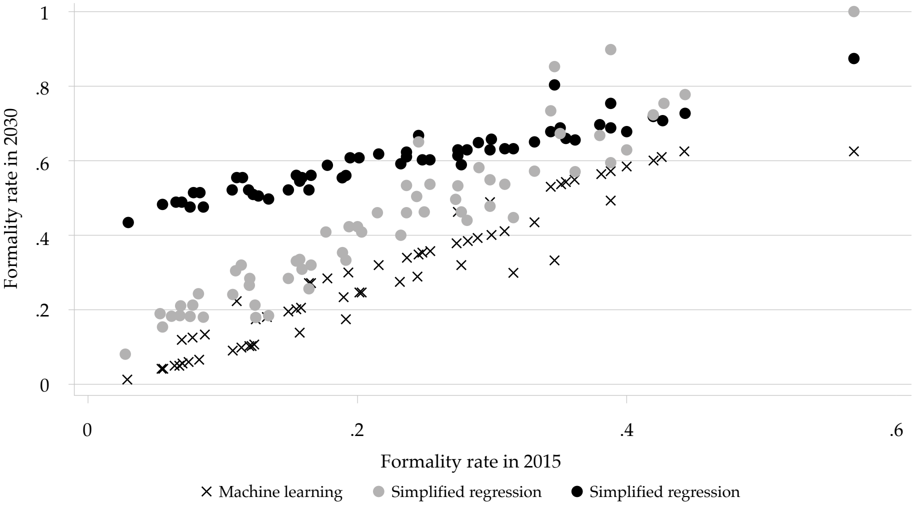


Figure 5. Formality rate forecasts by city (regression and machine-learning-based)

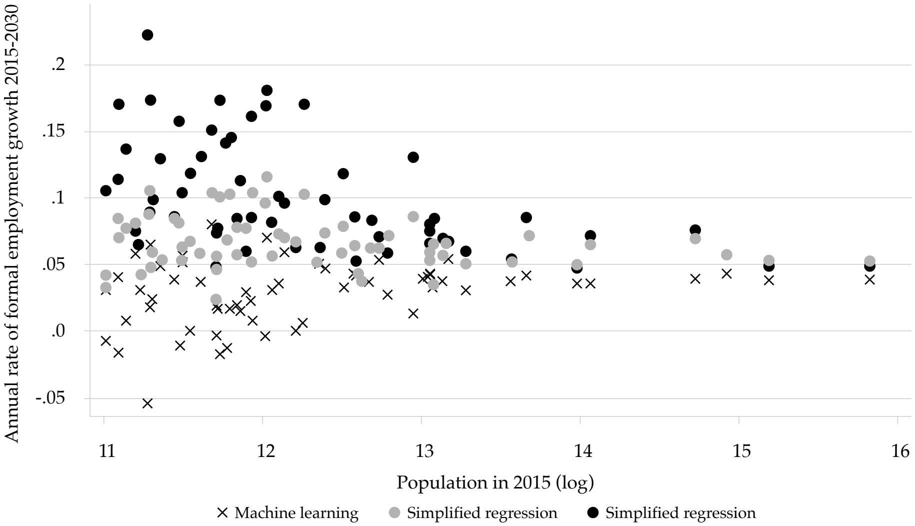


Figure 6. Projected formal employment growth rates and city size (regression and machine-learning-based)

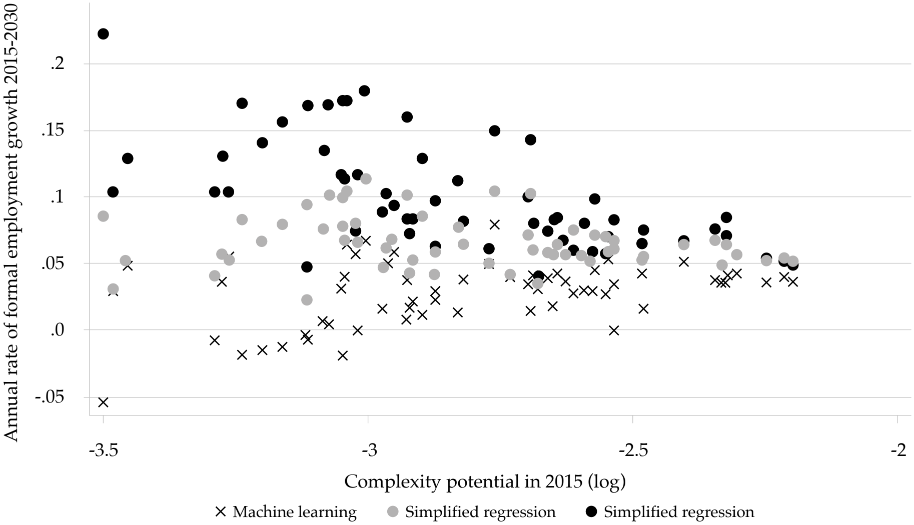


Figure 7. Projected formal employment growth rates and complexity (regression and machine-learning-based)

Figure 6 shows that the formal employment growth rates in 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 (Figure 7).

Finally, to conclude the presentation of the results, Table 7 compares the aggregates of the 62 cities from the three methods. The formal employment rate for the aggregate, currently 34.3 %, may reach between 47.9 % and 66.1 %, depending on the forecast method (and the simple average may reach between 29.1 and 59.4 %, starting from 22 %). While in the period 2008-2015, total formal employment in the 62 cities grew 7.7 % per annum, it may be expected to grow in the future between 4 and 6.3 % (simple average between 2.9 and 10 %, compared with 10.5 % in the recent past).

Table 7. Forecasts summary for the aggregate of the 62 cities

		Current	Projected (2030)		
			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 %

Discussion

In order to assess the 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 five 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 %. With the social security criterion, it was either 64.6 or 46.8 %, 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 (3):

$$f_{c,t}=\frac{emp_{c,t}}{pop_{c,t}}=\left(\frac{emp_{c,t}}{workers_{c,t}}\right)*\left(\frac{workers_{c,t}}{occupied_{c,t}}\right)*\left(\frac{occupied_{c,t}}{laborforce_{c,t}}\right)*\left(\frac{laborforce_{c,t}}{pop_{c,t}}\right) \tag{5}$$

$$\begin{aligned} f_{c,t}=\frac{emp_{c,t}}{pop_{c,t}} &= work\ intensity\ rate_{c,t} * official\ formality\ rate_{c,t} \\ &*\left(1-unemployment\ rate_{c,t}\right) * participation\ rate_{c,t} \end{aligned} \tag{6}$$

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 %, as we have just seen, to a range between 65.3 % and 90.2 % in the machine-learning-based forecast. But this conclusion is unwarranted because, although I have not explicitly modeled the three other margins (i.e., 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 %. However, there is much less consistency in the predictions for the mid-size and smaller cities, many of which are not very optimistic.

Combining regression-based and machine-learning-based forecasts enrich our understanding of cities formal employment prospects. The main strength of the latter lies not in its ability to predict aggregates for which the regression-based method is better suited but in the nuances it provides on the predictions by city. 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 implicitly 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 (however, see O'Clery et al. 2019), women's access to workplaces, etc.

References

Albrecht, J., Navarro, L., & Vroman, S. (2009). The effects of labour market policies in an economy with an informal sector. *The Economic Journal, 119*(539), 1105-1129. https://doi.org/10.1111/j.1468-0297.2009.02268.x

Bosch, M., & Maloney, W.F. (2010). Comparative analysis of labor market dynamics using Markov processes: An application to informality. *Labour Economics, 17*(4), 621-631. https://doi.org/10.1596/1813-9450-4429

De Soto, H. (1989). *The other path: The invisible revolution in the Third World*. Harper and Row.

De Soto, H. (2000). *The mystery of capital: Why capitalism triumphs in the west and fails everywhere else*. Basic Books.

Duranton, G. (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* (pp. 107-133). Springer. https://doi.org/10.1007/978-4-431-55390-8_6

Gollin, D., Jedwab, R., & Vollrath D. (2016). Urbanization with and without industrialization. *Journal of Economic Growth, 21*(1), 35-70. https://doi.org/10.1007/s10887-015-9121-4

Harris, J. R., & Todaro, M. P. (1970). Migration, unemployment, and development: A two-sector analysis. *The American Economic Review, 60*(1), 126-142. https://www.jstor.org/stable/1807860

Hartshorn, S. (2016). *Machine learning with random forests and decision trees: A Visual guide for beginners*. Kindle Edition.

Hidalgo, C., & Hausmann, R. (2009). The building blocks of economic complexity. *Proceedings of the National Academy of Sciences, 106*(26), 10570-10575. https://doi.org/10.1073/pnas.0900943106

Kugler, A., Kugler, M. D., & Herrera-Prada, L. O. (2017). Do payroll tax breaks stimulate formality? Evidence from Colombia's reform. *Economia Journal of the Latin American and Caribbean Economic Association*, (Fall), 3-40. http://dx.doi.org/10.3386/w23308

Levy, S. (2008). *Good Intentions, bad outcomes: Social policy, informality, and economic growth in Mexico*. Brookings Institution Press.

Lewis, W.A. (1954). Economic development with unlimited supplies of labor. *Manchester School of Economic and Social Studies, 22*(2), 139-191. http://faculty.smu.edu/tosang/pdf/Lewis_1954.pdf

McGuire, T. J., & Bartik, T. J. (1991). *Who benefits from state and local economic development policies?* W.E. Upjohn Institute. https://www.jstor.org/stable/j.ctvh4zh1q

Meghir, C., Narita, R., & Robin, J-M. (2015). Wages and informality in developing countries. *The American Economic Review*, *105*, 1509-1546. https://doi.org/10.1257/aer.20121110

Neffke, F., & Henning, M. (2013). Skill relatedness and firm diversification. *Strategic Management Journal*, *34*(3), 297-316. https://doi.org/10.1002/smj.2014

O'Clery, N., Prieto Curiel, R., & Lora, E. (2019). Commuting times and the mobilisation of skills in emergent cities. *Applied Network Science*, *4*(1), 118. https://doi.org/10.1007/s41109-019-0235-z

O'Clery, N., Chaparro, J. C., Gómez-Liévano, A., & Lora, E. (2020). *Skill diversity and the evolution of formal employment in cities, submitted to Research Policy*. Peak Urban.

Rauch, J. E. (1991). Modeling the informal sector formally. *Journal of Development Economics*, *35*(1), 33-47. https://doi.org/10.1016/0304-3878(91)90065-4

Ulyssea, G. (2010). Regulation of entry, labor market institutions and the informal sector. *Journal of Development Economics*, *91*(1), 87-99. https://doi.org/10.1016/j.jdeveco.2009.07.001

Appendix

Appendix 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				
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

Keep going

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 observations = 434				
Adjusted R-squared = 0.4331				

Appendix 2. 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 %

City	Current (2015)	Projected (2030)	
		Full specification	Simplified specification
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 current	100 %	95 %	95 %

Appendix 3. 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 Andrés 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 %

City	Past (2008-2015)	Projected (2015-2030)	
		Full specification	Simplified specification
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 current	100 %	24 %	27 %

Appendix 4. 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 %

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
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 Andrés 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 %

Appendix 5. 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 %

City	Growth group predicted	Probability of belonging to group
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 6. 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	Same predictions?			
				Median 2008-2015 and regression-based	Median 2008-2015 and machine-learning based	Regression-based and machine-learning based	Total (out 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

City	2008-2015 median	Regression-based (simplified specification)	Machine- learning based	Same predictions?			
				Median 2008-2015 and regression-based	Median 2008-2015 and machine-learning based	Regression-based and machine-learning based	Total (out of 3)
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

Keep going

City	2008-2015 median	Regression- based (simplified specification)	Machine- learning based	Same predictions?			
				Median 2008-2015 and regression-based	Median 2008-2015 and machine-learning based	Regression-based and machine-learning based	Total (out of 3)
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 %