Economic Complexity and Technological Relatedness: Findings for American Cities

Carlos Daboín, Marcela Escobari, Gabriel Hernández and José Morales-Arilla

Abstract

Implementing and expanding on the methods introduced in Hausmann & Hidalgo (2009), we find that a city's *Economic Complexity* - a measure of the productive capabilities available in a location - informs that city's future growth prospects. Moreover, we find that the *technological relatedness* between an industry and the activities already present in a city informs the industry's local growth prospects. The out-of-sample predictive value of these findings suggests that the evolution of industries within a city is path-dependent. Consequently, policymakers interested in the local economic development of American cities may obtain valuable insights from these models' predictions.

Keywords: City Growth, Economic Complexity, Technological Relatedness

1. Introduction

Countries diversify their productive structures into increasingly uncommon activities as they develop[1]. Figure 1 shows this pattern by plotting the diversity and average ubiquity of the export baskets of different countries, along with their level of economic development. The salient feature of this visualization is that, on average, richer countries lie on the high-diversity, low-ubiquity end of the graph, while poorest countries tend to be found in the low-diversity, high-ubiquity quadrant.

This international pattern is often mirrored in subnational analyses: Developed cities or regions tend to diversify towards uncommon economic sec-

[1]Described in detail in Hausmann et al (2014).

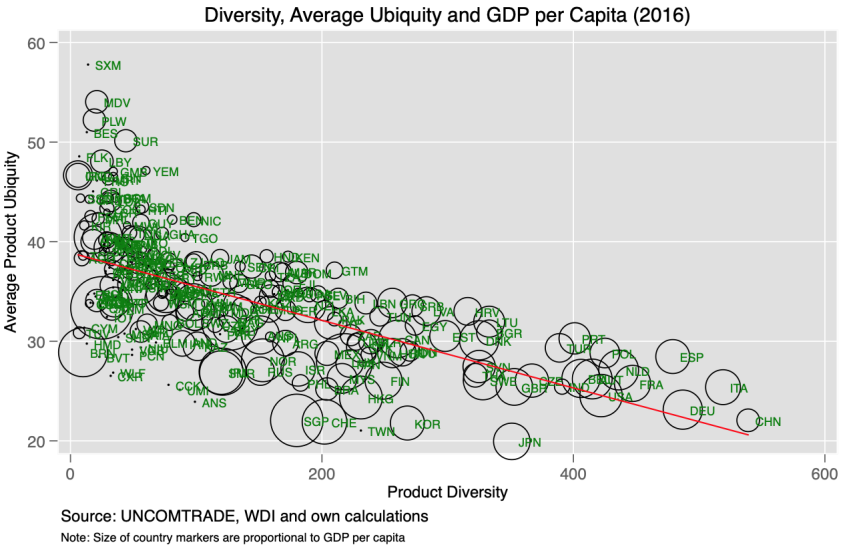

Figure 1: Diversity, Ubiquity and Wealth

tors, while lagging areas remain focused on fewer, prevalent activities. Figure 2 illustrates this pattern within the United States. While variation in income levels is disperse within the US, we still observe richer cities grouped towards the bottom-right of the visualization, while poorer cities group in the upper-left.

These findings contradict Ricardian comparative advantage theory, which suggests that specializing in a narrow set of activities should lead to higher levels of efficiency. The observation that diversity and ubiquity of a society's productive structures associate with levels of wealth is consistent with a view that points to the progressive accumulation of productive capabilities and know-how as the path to economic development.

The intuition behind one of such views, the *Economic Complexity* theory of economic development[2], goes as follows:

- Productive capabilities and tacit know-how, which are not perfectly

[2]Described in detail in Hausmann & Hidalgo (2009).

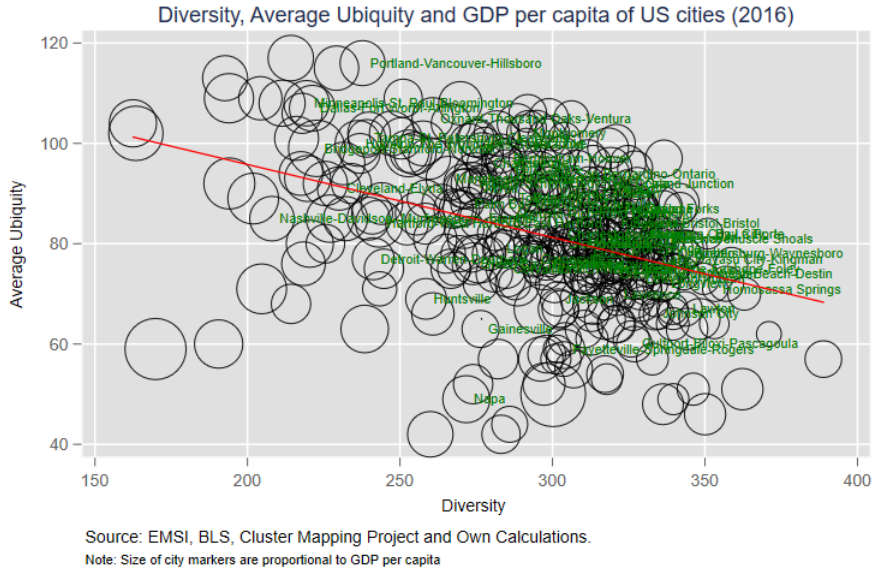

Figure 2: Diversity, Ubiquity and Wealth within the US

observable, are combined in the production of goods and services.

- Regions that lack many capabilities will only be able to assemble a relatively modest number of goods and services, which will also be feasible in many other regions.

- To the contrary, regions that accumulate many capacities will be able to assemble a relatively large number of activities, many of which will only be feasible in the small group of other regions with the necessary capabilities.

- As they expand their stock of productive capabilities, developing regions become able to diversify their productive mix into less ubiquitous activities.

From this perspective, the concept of productive diversity and average ubiquity of a region's productive mix are indicative of its level of economic development. The concept is captured through the Economic Complexity Index (ECI) which measures the unobserved stock of capabilities in an economy. For a city, ECI is derived through an iterative process that interacts the

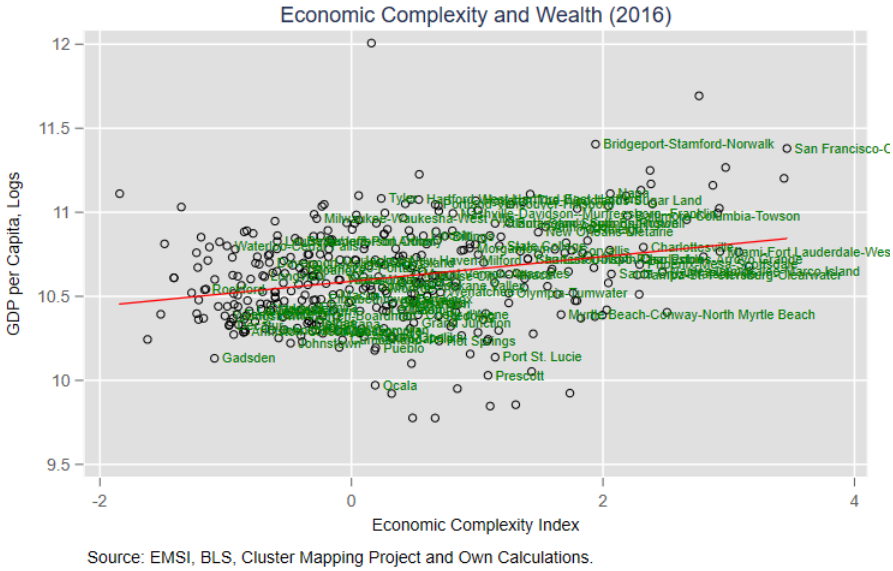

Figure 3: Complexity and wealth among US cities, 2016

measures of industrial diversity and ubiquity to correct for the noise in one of the variables with the average values of the other (Hausmann & Hidalgo, 2009). The result is an index which captures the ability of a city to make many products and services, and thus host many industries, including those that are hard to host in capability-poor cities. Figure 3 shows how the ECI of American cities relates to their GRP per capita in the same year.

Authors in urban economics discussing the issue of local economic growth have emphasized the process of *spatial equilibria*, in which the effects of productivity gains on local earnings growth are partially driven out by labor mobility into growing cities that can accommodate a growing workforce with relative ease[3]. These spatial dynamics suggest that:

1. Assessing the association between complexity and wealth should control for population size, which could capture the effects of past economic gains at the city level.
2. Assessing the effects of complexity on development should consider city

[3]See Glaeser and Gottlieb (2009).

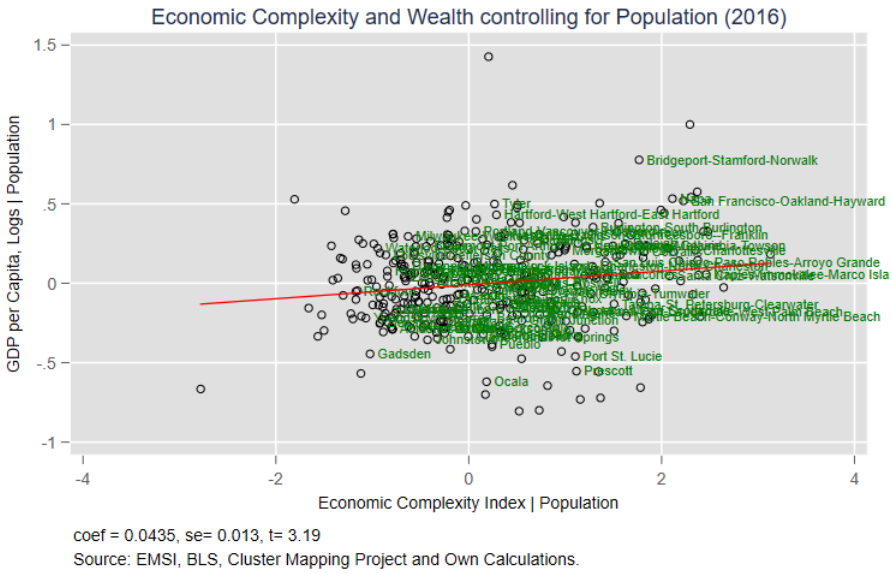

Figure 4: Complexity and wealth controlling for population

population growth as dependent variable of interest.

3. Assessing the effects of complexity on economic growth should consider the heterogeneity between large and small cities and between dense and sparse cities.

Figure 4 shows that even after controlling for population size, there is a positive association between the gross regional product (GRP) per capita of US cities and their level of economic complexity. However, Figures 5 and 6 also underscore how this relationship is contingent on city size with a stronger effect among more populous cities. The reason for this is that while small cities benefit from concentrating their resources towards the accumulation of economies of scale on a single sector, large cities more readily gain wealth by accumulating economies of scale in many different sectors.

Nevertheless, our main question is whether local levels of economic complexity are predictive of the future growth of cities, which as we mentioned, should be measured using population growth. Our analyses suggest that this is indeed the case: a 1 standard deviation increase in ECI associates with a 0.1 standard deviation increase in the logarithm of the population

growth factor. This result is robust to controlling for a number of factors that the economics literature has identified as relevant for local economic growth. Moreover, our most complex model, which fits 5-year growth data between 2001 and 2013, captures over 27% of the out-of-sample variations in the logarithm of city growth ratios between 2013 and 2017. We find that ECI is among the variables that contribute most to this forecasting capacity. Finally, we test whether the association between ECI and future city growth varies along city population and population density dimensions, finding some evidence that city size dampens the association between ECI and city growth.

If the accumulation of productive capabilities is relevant for future economic performance, a key question to address is the process of accumulation of such capacities, which should be reflected in the patterns of diversification of cities. International and subnational evidence shows that this process is path-dependent: the accumulation of new productive capabilities depends on how complementary these are to other capabilities already present in an economy. From a diversification perspective, this suggests that absent industries that share capabilities with the competitive activities of a city are more likely to appear and grow, while present sectors that do not share capabilities with the rest of a city's economy are more likely to shrink and disappear.

The *technological proximity* between pairs of products or industries captures the degree to which they share the same productive capabilities as inputs. This proximity can be estimated in a number of ways. In an international trade context, these proximities have been estimated as the tendency for the exports of different pairs of goods to co-locate in the same countries of origin. These estimates constitute the building blocks of the Product Space[4], a network visualization that shows each product most closely connected to those other products with which it had been historically co-exported.

Similarly, in a subnational setting, it is expected for industries that rely on the same inputs to cluster together in the same cities. The urban economics literature refers to this tendency as Marshallian agglomeration externalities. Measuring this clustering tendency provides an implicit measure of the relatedness between different pairs of industries.

[4]Introduced in Hidalgo et al (2007)

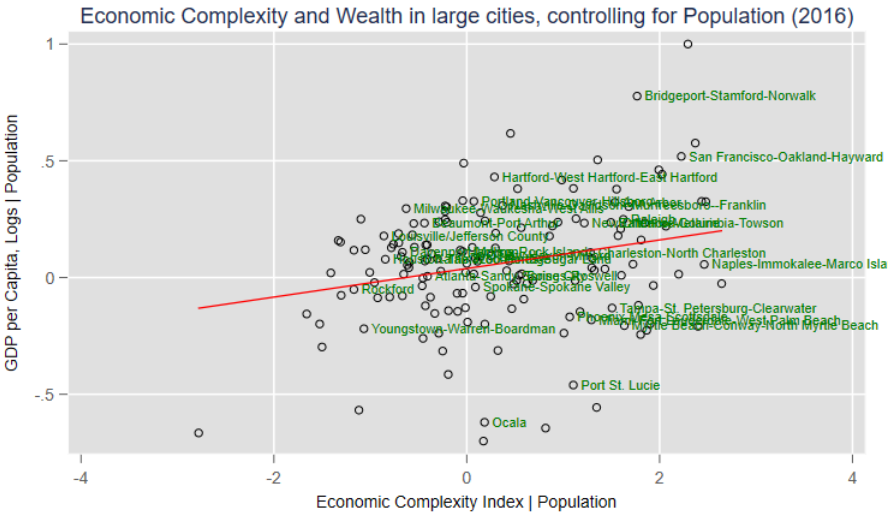

Figure 5: Complexity and wealth in large cities

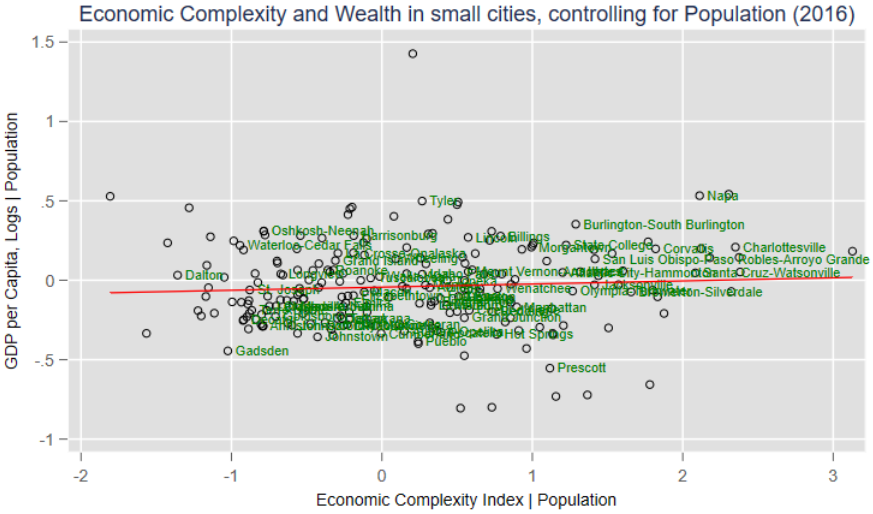

Figure 6: Complexity and wealth in small cities

Separately, one can explicitly observe the tendency for sectors to rely on the same inputs. For instance, if the occupational staffing patterns of industries are observable, one could construct an explicit technological relatedness metric based on the similarity in the occupation demand between every pair of industries.

Building on these technological relatedness metrics, one can calculate the overall relatedness of an industry to the competitive industries present in a city. The path-dependence hypothesis suggests that this technological density of citys economy around an industry should be positively associated with the local prospects of that industry.

We test this hypothesis on density measures based on implicit relatedness captured by the tendency of industries to cluster geographically, and on explicit relatedness measured by the occupation similarity between industry pairs. We provide network visualizations, or industry spaces for these two measures of relatedness. We find the expected results: a 1% increase in the implicit and explicit relatedness densities associate with a 0.07% and 0.01% increase in the growth factor of cities, respectively. Our model, fit with 5-year growth data between 2002 and 2012, captures over 12% of the out-of-sample variations in city-industry growth between 2012 and 2017.

Moreover, we test the value of these density variables in the extensive margin, assessing whether they predict the appearance and disappearance of industries. We find that our models capture an out-of-sample area under the ROC curve of 60% for the appearance of absent industries, and 70% of the variation in the disappearance of present industries.

Given the out-of-sample predictive value of our models at the city and the city-industry levels, we believe that they are informative forecasting tools for policy-makers and analysts interested in local economic development in the US.

The paper continues as follows: Section 2 describes how the relevant complexity metrics are calculated. Section 3 presents our results predicting city population growth as a function of the economic complexity index of a city. Section 4 presents our results predicting city-industry growth and discrete

changes (appearance and disappearance) as a function of implicit and explicit relatedness densities. Section 5 concludes. Section 6 describes the datasets used in our analyses.

2. Calculating complexity and relatedness measures

2.1. Economic Complexity Index

The calculation of complexity metrics for a given year starts with a matrix of employment of all industries in all cities. We'll define this matrix as J_{ci}. From here, we can calculate total employment levels by city across industries and by industry across cities.

$$X_c = \sum_i J_{ci} \tag{1}$$

$$X_i = \sum_c J_{ci} \tag{2}$$

$$X = \sum_i \sum_c J_{ci} \tag{3}$$

We now calculate the Revealed Comparative Advantage (RCA_{ci}) of a city in a given industry as the ratio of the share of a given industry in a city's employment and the national share of the industry. We define the M_{ci} matrix as the condition that a given industry represents a larger share of an city's employment than the national share.

$$RCA_{ci} = \frac{X_{ci}/X_c}{X_i/X} \tag{4}$$

$$M_{ci} = 1[RCA_{ci} \geq 1] \tag{5}$$

From M_{ci} we can now estimate the diversity of a city as the count of industries in which a given city has an RCA_{ci} greater than 1, and the ubiquity of an industry as the number of cities in which the industry is observed with an RCA_{ci} greater than one.

$$\text{Diversity}_c = K_{c0} = \sum_i M_{ci} \tag{6}$$

$$\text{Ubiquity}_i = K_{i0} = \sum_c M_{ci} \tag{7}$$

Following Hausmann & Hidalgo (2009), we can now refine the metric of diversity of a city with the ubiquity of the industries in which the city shows an intensive concentration on. This yields the average ubiquity. This would help improve the diversity metric for places that are not very diverse but concentrate in very unique sectors (as San Jose, which we observe in the figures above showing low diversity but high complexity).

Similarly, we can improve on the ubiquity of an industry by the diversity of the cities that concentrate in it intensively. This would help correct for sectors that are not very ubiquitous but are also not very complex in the way they share productive inputs with other industries (such as extractive sectors of the economy).

This process of refining an industry metric by the average values of the relevant city metrics and vice versa is called the "method of reflections", and if performed *ad infinitum*, it would converge to a metric at the city level and a metric at the industry level. These would be the Economic Complexity Index and the Industry Complexity Index, respectively.

$$\text{Av. Ubq.}_c = K_{c1} = \frac{\sum_i K_{i0} * M_{ci}}{K_{c0}} \rightarrow K_{c2} \rightarrow \cdots \rightarrow K_{c\infty} = ECI_c \tag{8}$$

$$\text{Av. Div.}_i = K_{i1} = \frac{\sum_c K_{c0} * M_{ci}}{K_{i0}} \rightarrow K_{i2} \rightarrow \cdots \rightarrow K_{i\infty} = ICI_i \tag{9}$$

Another way to estimate the ECI, which is mathematically equivalent, would be as follows:

$$K_{c,n} = \frac{1}{k_c, 0} M_{ci} \frac{1}{k_{i,0}} \sum_{c'} M_{c'i} k_{c',n-2} \tag{10}$$

$$K_{c,n} = \sum_{c'} k_{c',n-2} \sum_{i} \frac{M_{c'i} k_{c',n-2}}{k_{c,0} ki, 0} \tag{11}$$

$$K_{c,n} = \sum_{c'} k_{c',n-2} \tilde{M}^C_{c,c'} \tag{12}$$

Where:

$$\tilde{M}^C_{c,c'} = \sum_{i} \frac{M_{c'i} k_{c',n-2}}{k_{c,0} ki, 0}$$

In vector notation:

$$\vec{k}_n = \tilde{M}^C * \vec{k}_{n-2} \tag{13}$$

As $n \to \infty$:

$$\tilde{M}^C * \vec{k}_{n-2} = \lambda \vec{k} \tag{14}$$

Where $\vec{k}$ is an eigenvector of $\tilde{M}^C$. The second largest eigenvector of $\tilde{M}^C$ in the city/industry matrix M_{ci} accounts for the Economic Complexity Index (ECI) at the city level, and the second largest eigenvector of $\tilde{M}^i$ accounts for the Industry Complexity Index (ICI). The ECI of a city is mathematically equivalent to the average of the ICIs of those products (or sectors) in which a location has an RCA larger than 1.

2.2. Proximity and Density

From the M_{ci} we can count the number of cities in which a given pair of products appear with high concentration.

$$\text{co-occurrence}_{i,i} = U_{i,i} = M_{ci}^T * M_{ci} \tag{15}$$

This co-occurrence matrix is by definition symmetric, and its diagonal captures the number of occurrences of each industry. By dividing the co-occurrences between industries i and i' by the maximum between the diagonal position for i and i', we can estimate the minimum conditional probability for a city to be competitive in an industry given that it is competitive in another. This estimate captures the tendency for industries to cluster together in the same cities, and we will refer to it as co-location *implicit* technological proximity between industries.

$$\phi_{i,i'} = \frac{U_{i,i'}}{max(U_{i,i}, U_{i',i'})} \tag{16}$$

We can visualize this co-location proximity as a network, which we call the "industry space". In this case, connections in the industry space are "implicit" technological relatedness measured by the tendency for industries to cluster together. Figure 7 shows the industry space retaining 5% of the strongest proximity connections and the strongest link to every industry. This exercise yields a layout that permits identifying some visible industrial

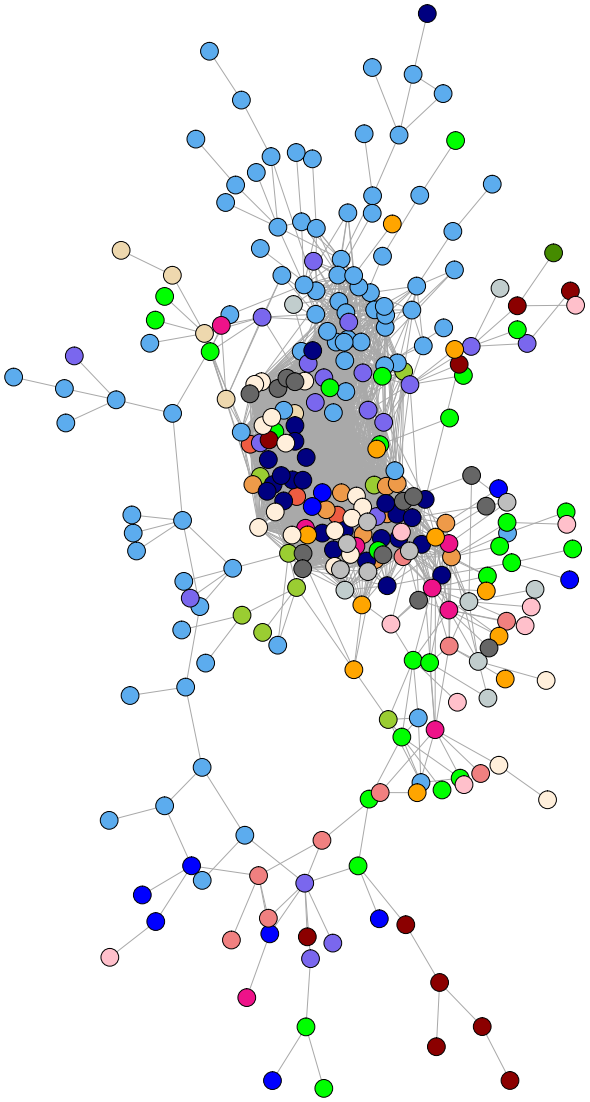

Figure 7: The Co-Location Industry Space. The co-location Industry Space shows the implicit relatedness of one industry to another. The distance between nodes is calculated by the probability of any two industries to appear in the same city.

communities by sector.

A similar exercise could be executed on a matrix on the national staffing patterns of industries by the occupations of their workers, so that we end up with a new proximity matrix between pairs of industries that captures the minimum conditional probability that an industry i demands a given occupation with relative intensity, given that i' also does. This is why we call the occupation similarity *explicit* technological proximity between industries. We will call this matrix $\psi_{i,i'}$. Figure 8 displays the occupation-similarity "industry space" similarly displaying the top 5% of the industry-to-industry proximities and the strongest link to every industry.

From $\phi_{i,i'}$ and $\psi_{i,i'}$ we get the implicit and explicit technological proximities between every pair of industries. These metrics and their resulting structures are expected to differ. To assess this difference, we can measure the *weighted degree centrality*[5] of every industry in both industry spaces:

$$centrality_{i'}^{\text{implicit}} = c_{i'}^{I} = \frac{\sum_i \phi_{i,i'}}{\sum_i \sum_{i'} \phi_{i,i'}} \tag{17}$$

$$centrality_{i'}^{\text{explicit}} = c_{i'}^{X} = \frac{\sum_i \psi_{i,i'}}{\sum_i \sum_{i'} \psi_{i,i'}} \tag{18}$$

Figures 9 and 10 show that the distributions of centrality for the sets of tradable and non-tradable industries. Interestingly, while non-tradable industries are relatively central in the implicit proximity matrix $\phi_{i,i'}$, tradable industries are so in the explicit proximity matrix $\psi i, i'$. We interpret this result as a consequence of the high ubiquity of non-tradable industries and the high occupational diversity of tradable industries. The former makes non-tradable sectors relatively likely to co-locate with many industries, while the latter makes tradable sectors share occupational vectors with many industries.

Following Hausmann et. al (2014), one way to assess the degree to which the industries present in a city are relatively proximate to a given industry i is to add up all proximities to that industry in the set of industries present

[5]These centrality measures capture the sum of all proximities to an industry as a proportion of all the proximities in each network.

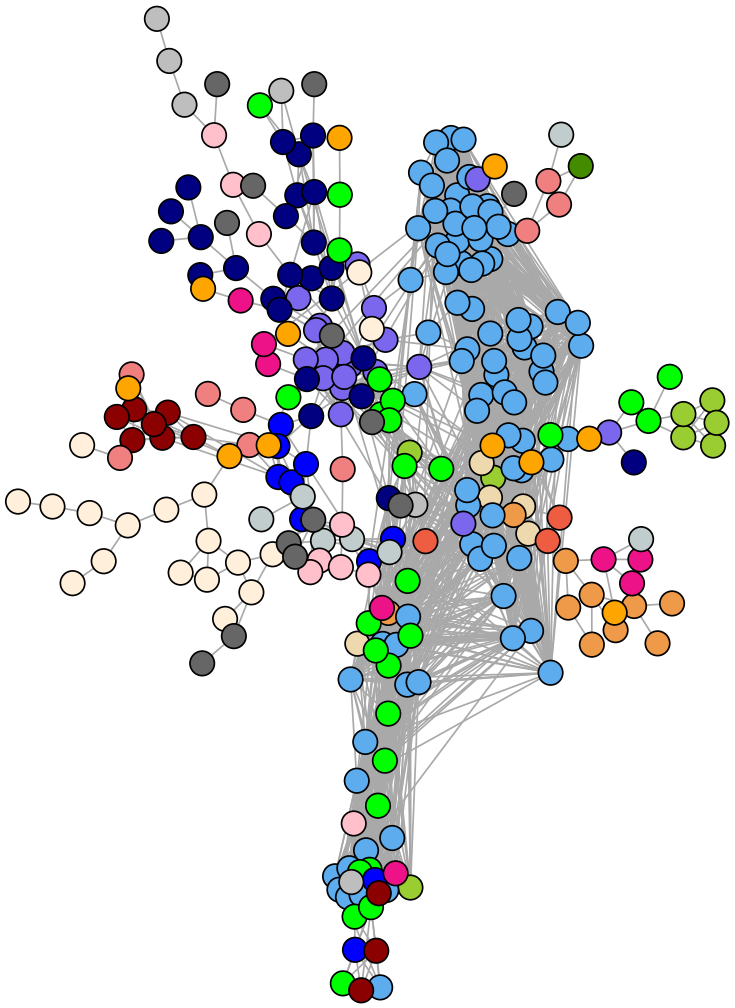

Figure 8: The Occupational Similarity Industry Space. This industry space shows the explicit relatedness of one industry to another. The distance between the nodes is determined by the tendency of different industries to employ the same composition of workers.

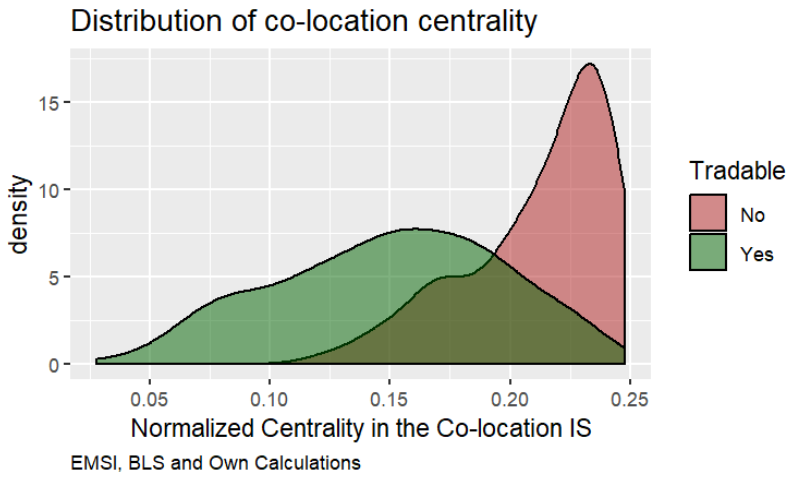

Figure 9: Distribution of co-location centrality by tradability. The co-location centrality is the sum of all the implicit proximities to an industry.

in the city and divide it by the total sum of all proximities. This metric would capture the *density* of the industrial structure in a city around every industry, both present or absent.

$$density_{c,i'}^{\text{implicit}} = d_{c,i'}^{I} = \frac{\sum_i M_{c,i} * \phi_{i,i'}}{\sum_i \phi_{i,i'}} \tag{19}$$

$$density_{c,i'}^{\text{explicit}} = d_{c,i'}^{X} = \frac{\sum_i M_{c,i} * \psi_{i,i'}}{\sum_i \psi_{i,i'}} \tag{20}$$

Finally, based on the density and proximity metrics, we can estimate the share of all densities that are captured by a city's productive structure, weighting by each industry's ICI - a measure we call Strategic Index - (SI)-, and how much the SI of a city would improve by adding a given missing industry - a measure we call Strategic Gain (SG). We capture these metrics based on the implicit co-location proximities.

$$SI_c = \sum_i d_{c,i}(1 - M_{c,i})ICI_i \tag{21}$$

$$SG_{c,i} = \left[\sum_{i'} \frac{\phi_{i,i'}}{\sum_{i''} \phi_{i'',i'}}(1 - M_{c,i'})ICI_{i'}\right] - d_{c,i}ICI_i \tag{22}$$

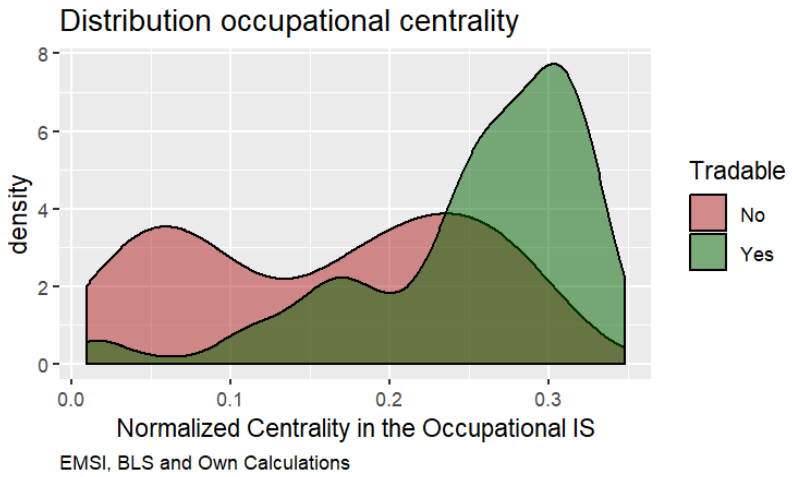

Figure 10: Distribution of occupational centrality by tradability. The occupation centrality is the sum of all the explicit proximities to an industry.

3. City growth as a function of economic complexity of cities

We assess how City population growth associates with lagged economic complexity through the following equation.

$$log\left[\frac{P_c^{t+5}}{P_c^t}\right] = \beta_0 + \beta_1 ECI_c^t + X^t\kappa + \epsilon_c \tag{23}$$

Where P_c^t stands for the population of city c at year t, ECI_c^t is the complexity of city c at time t, X^t is a vector of city covariates at time t, and ϵ_c is a residual uncorrelated with regressors. We take US micropolitan and metropolitan areas for which we can capture all the relevant variables and fit a 5-year growth model between 2006 and 2011 and test the predictive accuracy of its forecasts on city growth between 2011 and 2016. In the appendix we provide estimates restricting for metropolitan areas which allow us to include education controls, as well as longer-term growth regressions for 8-year spans - lack of data availability on some of the most relevant controls prevent us from testing the out-of-sample accuracy of the 8-year growth models.

Table 1 provides estimates of different specifications, iterating on the set of controls being included. It reports standardized "beta" coefficients and robust p-values of the relevant t-tests for statistical significance of coefficients,

and also provide the out-of-sample R^2 for city population growth between 2011 and 2016. In our pooled model with demographic, welfare, productivity and age structure controls (regression 6), we find that 1 standard deviation increase in a city's ECI associates with a statistically significant increase of 0.15 standard deviations in the logarithm of the population growth factor. This model yields an out-of-sample R^2 value of 27%, a significant level of forecasting accuracy. Including baseline population and population density interactions with ECI (regression 8) we find an increase in one order of magnitude in the direct association between ECI and population growth, but a negative and significant association between future growth and the interaction of ECI and city population size. This suggests that the association of ECI with future population growth is weaker for larger cities. This latter model captures 29% of the out of sample variation in city growth between 2011 and 2016 - figure 11 displays the association between estimates of city growth from the model of regression 8 and the observed levels of population growth.

VARIABLES	(1) Population Growth 5 years	(2) Population Growth 5 years	(3) Population Growth 5 years	(4) Population Growth 5 years	(5) Population Growth 5 years	(6) Population Growth 5 years	(7) Population Growth 5 years	(8) Population Growth 5 years
L5 - ECI (Standardized)	0.377	0.286	0.328	0.296	0.270	0.154	0.886	1.293
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.009)	(0.002)	(0.000)
L5 – Log Population		0.292					0.294	0.262
		(0.000)					(0.000)	(0.479)
L5 - (ECI * Log Pop.)							-0.180	-0.836
							(0.667)	(0.030)
L5 – Pop. Density		-0.250					-0.186	-0.061
		(0.000)					(0.002)	(0.307)
L5 – (ECI * Pop. Density)							-0.442	-0.247
							(0.073)	(0.266)
Constant	.	.	.	.	.	.	.	.
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Observations	704	704	704	704	704	704	704	704
R-squared	0.142	0.179	0.158	0.433	0.171	0.452	0.192	0.468
Controls	None	Population size and Density	Welfare / Income	Age Structure	Productivity	2 + 3 + 4 + 5	2+ Demographics Interacting with ECI	2+ 3 + 4+ 5+ Demographics Interacting with ECI
R^2 Out-Sample	0.163	0.183	0.189	0.171	0.209	0.271	0.181	0.293

Table 1: City growth as a function of economic complexity

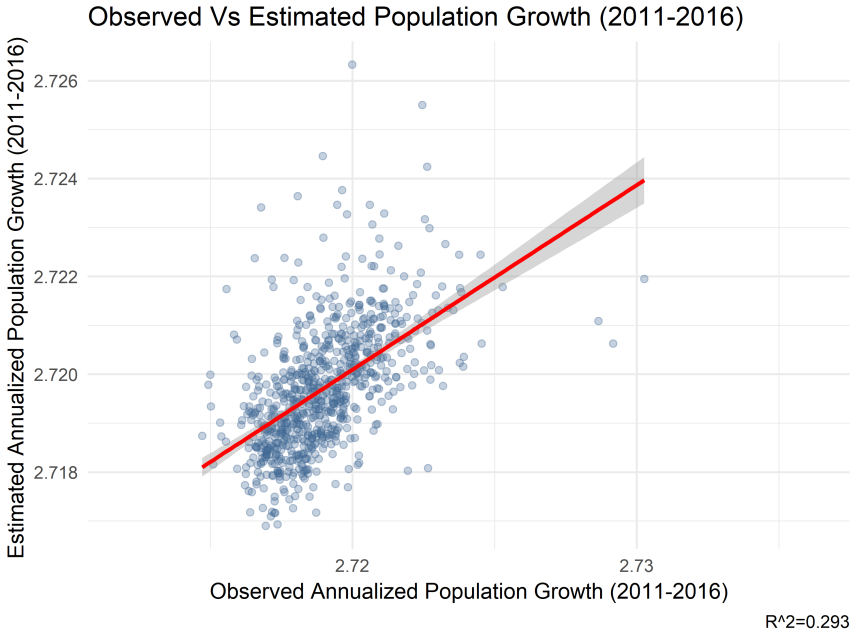

Figure 11: Predicted and actual growth: 2011-2016

Moreover, table 2 shows that if we take model 5 and iteratively remove different groups of variables, we find that removing ECI from the regression worsens it's out-of-sample R^2 by 1.6%, as opposed to removing baseline demographics and welfare variables, which actually improves out-of-sample predictive accuracy. While removing age structure or productivity variables worsens the predictive quality of the model by more than ECI (7.7% and 9.7% respectively), these sets of controls incorporate 8 and 7 variables each, while ECI is one single regressor. This suggests that baseline ECI is a meaningful predictor of future city growth.

VARIABLES	(1) Population Growth 5 years	(2) Population Growth 5 years	(3) Population Growth 5 years	(4) Population Growth 5 years	(5) Population Growth 5 years	(6) Population Growth 5 years
L5 - ECI	0.154		0.149	0.157	0.287	0.265
	(0.009)		(0.009)	(0.006)	(0.000)	(0.000)
Constant	.	.	.	.	.	.
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Observations	704	704	704	704	704	704
R-squared	0.452	0.446	0.450	0.451	0.198	0.442
Controls	Demographics + Welfare + Age Structure + Productivity	Excluding ECI	Excluding Demographics	Excluding Welfare	Excluding Age Structure	Excluding Productivity
Decrease In-Sample	0	-0.00570	-0.00205	-0.00140	-0.254	-0.00957
R^2 Out-Sample	0.271	0.255	0.273	0.276	0.194	0.174
Decrease Out-Sample	0	-0.0165	0.00170	0.00488	-0.0777	-0.0973

Table 2: City growth as a function of economic complexity

Finally, we further explore the heterogeneity of the effects of complexity by city size, either by splitting the sample into large and small cities (above or below the median) or by splitting it into city size quantiles. While regression 8 of table 1 shows a negative interaction between ECI and city size, we find no significant differences in the effect of ECI between small and large cities by interacting ECI with an indicator of whether the city was large at baseline (regression 1). We do find evidence of a negative interaction between ECI and the highest city size quantile (regression 2), but it's significant only at a 10% level of confidence.

VARIABLES	(1) Population Growth 5 years	(2) Population Growth 5 years
L5 - ECI	0.205	0.334
	(0.044)	(0.008)
L5.Q_2		0.102
		(0.060)
L5.Q_3		0.202
		(0.001)
L5.Q_4		0.295
		(0.000)
L5.Q_5		0.369
		(0.000)
L5.eciQ_2		-0.040
		(0.479)
L5.eciQ_3		-0.004
		(0.934)
L5.eciQ_4		-0.013
		(0.795)
L5.eciQ_5		-0.117
		(0.082)
L5.large	0.140	
	(0.002)	
L5.eciLarge	-0.025	
	(0.700)	
Constant	.	.
	(0.000)	(0.000)
Observations	704	704
R-squared	0.461	0.481
Controls	2+ 3 + 4+ 5+ Size quintiles SQ + SQ interactions with ECI	2+ 3 + 4+ 5+ SQ + SQ interactions with ECI
R^2 Out-Sample	0.283	0.286

Table 3: City growth as a function of economic complexity

4. City-industry growth as a function of relatedness densities

4.1. Intensive Margin: Predicting growth of industries

After observing that ECI contributes meaningfully to predicting future city growth, we now want to assess whether the density measures of relatedness between a city's productive structure and a given industry are predictive of that industry's local growth. To test for this, we test for the following specification:

$$log\left[\frac{J_{ci}^{t+5}}{J_{ci}^{t}}\right] = \alpha_0 + \alpha_1 log[J_{ci}^{t}] + \alpha_2 log[dI_{ci}^{t}] + \alpha_3 log[dX_{ci}^{t}] + X^t\rho + \mu_{ci} \quad (24)$$

Where J_{ci}^t is the employment level in city c and industry i at year t, dI_{ci}^t and dX_{ci}^t capture the implicit (co-location based) and explicit (occupation similarity based) densities of the industrial structure of city c around industry i at time t. X^t stands for lagged controls and μ_{ci} is an error term that is uncorrelated with controls. In this setup, α_1 estimates a mean-reversion term, while α_2 and α_3 capture the 5-year elasticity of employment to the implicit and explicit densities.

Table 4 shows the relevant estimates for the density elasticities and their respective heteroskedasticity robust standard errors fit in 5-year growth windows between 2002 and 2012, along with the out-of-sample R^2 of models' predictions of city/industry growth between 2012 and 2017. We find that estimates of α_1 are consistently negatives while estimates of α_2 and α_3 are consistently positive and statistically significant - although the implicit density's coefficients are more robust and higher in size. Importantly, a model that only controls for these variables (model 3) shows an out-of-sample R^2 of 4.7%. Adding industry, city, city-year and industry-year fixed effects in model 4 does not alter the sign or significance of elasticities, but it marginally worsens out-of-sample predictiveness - which is not unexpected of fixed effects models. Adding new controls capturing lagged aggregate size of cities and industries (model 7) improves predictiveness, and so does including their lagged growth rates and the city-industry lagged growth rate to correct for trends (model 8). This does not affect the signs of elasticities, but model 7 does remove statistical significance from the explicit density's elasticity. Adding city and industry fixed effects worsens predictiveness (model 9). Adding current levels of city and industry growth does improve predictiveness (model 10), but we cannot use such a model for forecasting purposes. Hence we pick model 8 as our preferred model for forecasting purposes. Figure 12 shows the observed and out-of-sample predicted factors of city/industry growth between 2012 and 2017.

VARIABLES	(1) Jobs growth	(2) Jobs growth	(3) Jobs growth	(4) Jobs growth	(5) Jobs growth	(6) Jobs growth	(7) Jobs growth	(8) Jobs growth	(9) Jobs growth	(10) Jobs growth
L5.ln_jobs	-0.014***	-0.015***	-0.015***	-0.044***	-0.065***	-0.035***	-0.0384***	-0.027***	-0.033***	-0.027***
	(0.000205)	(0.000212)	(0.000215)	(0.000384)	(0.000407)	(0.000339)	(0.000356)	(0.000494)	(0.000540)	(0.000488)
L5.ln_density_ocu		0.0635***	0.0534***	0.0270***		0.0426***	0.00821***	0.00176	0.0208***	-0.00108
		(0.00122)	(0.00153)	(0.00205)		(0.00122)	(0.00153)	(0.00203)	(0.00270)	(0.00199)
L5.ln_density	0.0577***		0.0194***	0.201***	0.0781***		0.0673***	0.0567***	0.167***	0.0570***
	(0.00139)		(0.00176)	(0.00363)	(0.00134)		(0.00181)	(0.00243)	(0.00486)	(0.00240)
L5.ln_jobs_r					0.0483***	0.0306***	0.0315***	0.0239***		0.0232***
					(0.000394)	(0.000381)	(0.000386)	(0.000530)		(0.000522)
L5.ln_jobs_i					0.0520***	0.0388***	0.0427***	0.0268***		0.0282***
					(0.000442)	(0.000410)	(0.000435)	(0.000606)		(0.000599)
L5.ln_jobs_ratio_5								-0.149***	-0.142***	-0.151***
								(0.00392)	(0.00391)	(0.00390)
L5.ln_jobs_radial_r_5								0.196***		0.122***
								(0.0235)		(0.0232)
L5.ln_jobs_radial_i_5								0.694***		0.135***
								(0.0198)		(0.0211)
ln_jobs_radial_i_5										0.909***
										(0.0133)
ln_jobs_radial_r_5										0.610***
										(0.0272)
Constant	0.122***	0.146***	0.157***	0.469***	-0.774***	-0.607***	-0.618***	-0.425***	0.361***	-0.416***
	(0.00239)	(0.00245)	(0.00264)	(0.00559)	(0.00732)	(0.00765)	(0.00771)	(0.0106)	(0.00757)	(0.0104)
Observations	317,315	317,315	317,315	317,315	318,167	317,315	317,315	153,179	153,179	153,179
R-squared	0.020	0.024	0.024	0.131	0.224	0.066	0.071	0.091	0.147	0.121
R squared out of sample	0.0417	0.0465	0.0470	0.0415	0.0990	0.0957	0.101	0.123	0.0622	0.146
Industry FE				YES					YES	
City FE				YES					YES	
Forecast Model								YES		
Industry-Year FE				YES						
City-Year FE				YES						

Table 4: Growth of industries in cities as a function of densities by co-location and occupation similarities

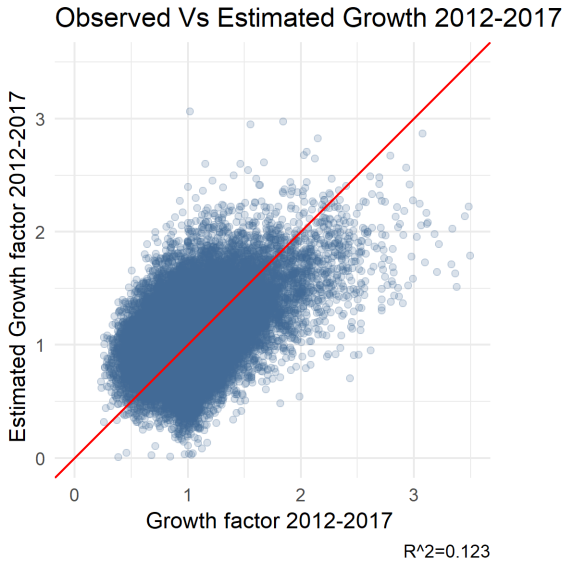

Figure 12: Predicted and actual growth: 2012-2017

4.2. Extensive Margin: Predicting industry appearance and disappearance

Beyond predicting continuous growth factors of industries in cities, a question of interest is that of predicting instances of industrial appearance or disappearance in cities. These are instances of discontinuous success of absent activities or failure of present ones, always measured in terms of employment.

When answering this question it is important to only consider important shifts in industries' performance. In order to do so, we consider as present those industries with an $RCA_{ci} \geq 0.25$ and as absent those industries with an $RCA_{ci} < 0.05$. By doing this, we consider as appearances those cases in which an industry jumps from $RCA_{ci,T-1} < 0.05$ to $RCA_{ci,T} \geq 0.25$, which means increasing their RCA_{ci} by more than 500% during the observation window. We define disappearances as those cases in which the opposite happens.

If density metrics indeed capture the coherence between an industry and the productive structure of a city, one would expect that:

- Higher densities associate with a lower chance of an industry being absent.
- Among absent industries, the chance of appearance is higher for high density industries.
- Higher density associates with a higher chance of an industry being present.
- Among present industries, the chance of disappearance shrinks with higher densities.

We test these hypothesis with the following logit specifications:

$$\text{Logit}[X_{c,i}] = \gamma_0 + \gamma_1 dI^t_{ci} + \gamma_2 dX^t_{ci} + \theta_{ci} \tag{25}$$

Where X_{ci} stands for the presence, absence, appearance or disappearance of industry i in city c. The absence and presence regressions take the density around each industry in 2012 and capture its relationship with industries' absence or presence in that year. In a similar way, the appearance and disappearance regressions take density of 2007 and measure its relationship with

the probability of appearance or disappearance for the period 2007-2012. In each regression we evaluate whether the coefficients for each of the density variables have the expected signs and statistical significance. Later, we assess the out-of-sample predictive accuracy of the models by comparing the predicted probabilities of absence or presence and appearances or disappearances for the period 2012-2017, with what actually happened in that period.

Given that this is a classification exercise, we use the Area Under the Curve (AUC) of the Receiver Operating Characteristics (ROC). This measure captures the capacity of a model to separate between binary events, as higher values suggest a higher sensitivity for correctly classifying positives at marginal increases in the false positive rate. The most relevant benchmark of random classification is captured by the 45-degree line. We report ROC values of each model, and display the relevant ROC.

4.2.1. Absence

Table 5 shows that all specifications capture negative and statistically significant coefficients as expected, and provide similar AUC scores of about 75%. Figure 13 shows how the predictive sensitivity of the three models is virtually indistinguishable along the specificity dimension.

VARIABLES	(1) Industry Absence	(2) Industry Absence	(3) Industry Absence
density	-14.15***		-10.02***
	(0.0705)		(0.0928)
density_ocu		-13.93***	-6.270***
		(0.0770)	(0.0970)
Constant	3.238***	2.576***	3.534***
	(0.0188)	(0.0171)	(0.0197)
Observations	261,345	261,345	261,345
Pseudo R-squared	0.152	0.130	0.165
AUC ROC 2017	0.758	0.740	0.767

Table 5: Logit models of absence on densities in 2012

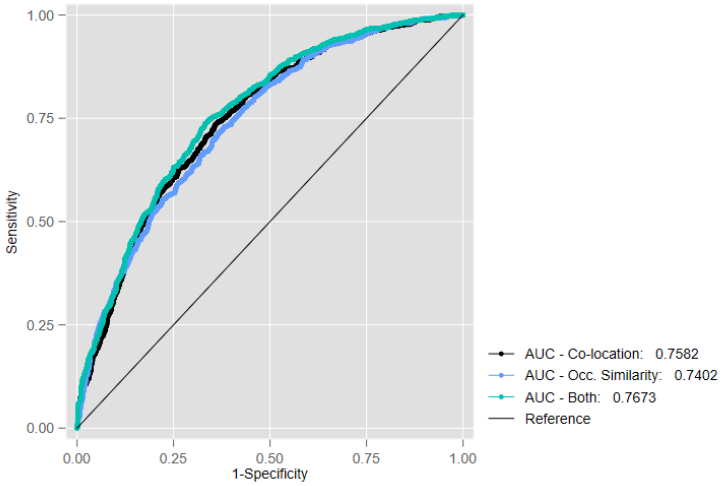

Figure 13: Out-of-sample ROC curve for absence: 2017

4.2.2. Appearance

Table 6 shows that all specifications capture positive and statistically significant coefficients as expected, and provide similar AUC scores of about 60%. Figure 14 shows how the three models yield very similar prediction results.

VARIABLES	(1) Industry Appearance	(2) Industry Appearance	(3) Industry Appearance
L5.density	6.856***		6.203***
	(0.209)		(0.260)
L5.density_ocu		4.300***	1.024***
		(0.182)	(0.241)
Constant	-4.355***	-3.558***	-4.403***
	(0.0546)	(0.0401)	(0.0558)
Observations	97,122	97,122	97,122
Pseudo R-squared	0.0234	0.0113	0.0238
AUC ROC 2017	0.615	0.586	0.616

Table 6: Logit models of appearance on densities. Period: 2007-2012

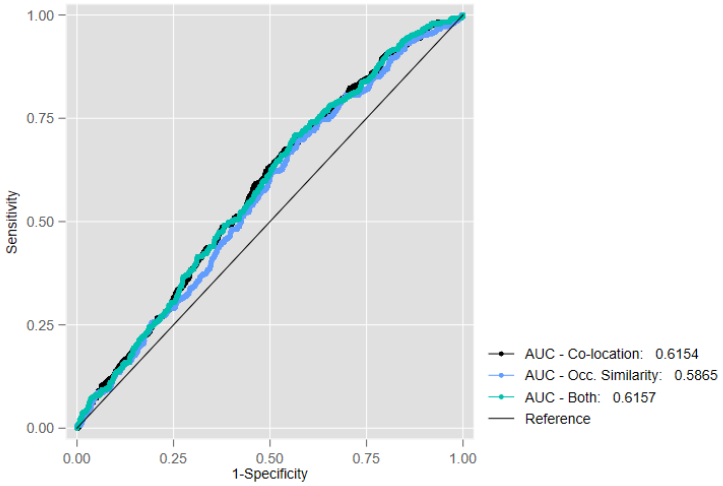

Figure 14: Out-of-sample ROC curve for appearance: 2012-2017

4.2.3. Presence

Table 7 shows that all specifications capture positive and statistically significant coefficients as expected. Model 1 and model 3 capture the highest out-of-sample AUC scores of 77%, while model 2 using only the occupation similarity density has an AUC of 73%. Figure 15 shows the predictive accuracy of model 2 is slightly worse than that of model 1 and model 3.

VARIABLES	(1) Industry Presence	(2) Industry Presence	(3) Industry Presence
density	14.89***		11.92***
	(0.0699)		(0.0910)
density_ocu		13.11***	4.350***
		(0.0717)	(0.0894)
Constant	-3.980***	-2.913***	-4.165***
	(0.0195)	(0.0167)	(0.0201)
Observations	261,345	261,345	261,345
Pseudo R-squared	0.167	0.124	0.174
AUC ROC 2017	0.766	0.734	0.771

Table 7: Logit models of presence on densities in 2012

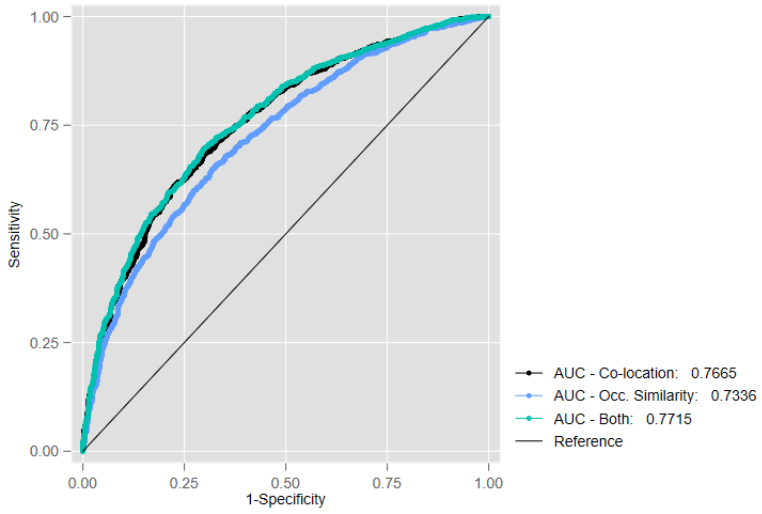

Figure 15: Out-of-sample ROC curve for presence: 2017

4.2.4. Disappearance

Table 8 shows that all specifications capture negative and statistically significant coefficients as expected. Model 3 captures an AUC of 72%, while models 1 and 2 capture an AUC of about 70% each. Figure 16 shows how model 3 is marginally better than models 1 and 2 at predicting future disappearances.

VARIABLES	(1) Industry Disappearance	(2) Industry Disappearance	(3) Industry Disappearance
L5.density	-5.000***		-2.997***
	(0.0917)		(0.124)
L5.density_ocu		-5.034***	-2.951***
		(0.0961)	(0.127)
Constant	-0.256***	-0.490***	-0.130***
	(0.0257)	(0.0223)	(0.0267)
Observations	136,395	136,395	136,395
Pseudo R-squared	0.0634	0.0628	0.0745
AUC ROC 2017	0.704	0.707	0.721

Table 8: Logit models of disappearance on densities. Period: 2007-2012

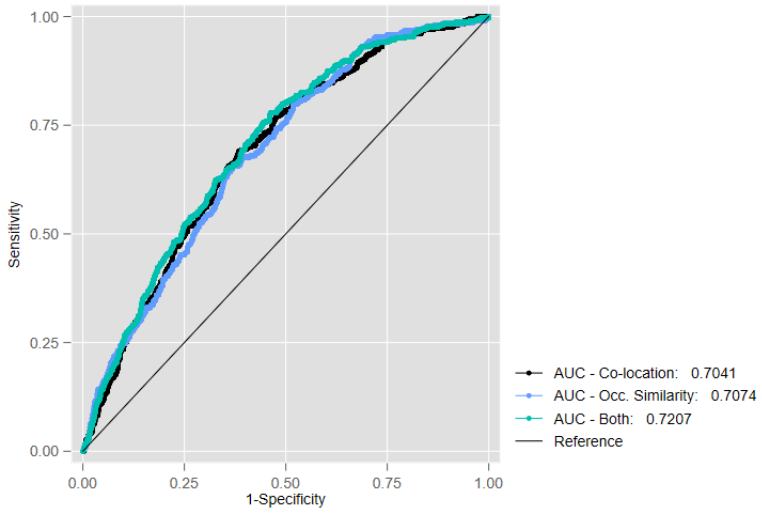

Figure 16: Out-of-sample ROC curve for disappearance: 2012-2017

5. Conclusion

This work is a preliminary step in a research agenda trying to assess the skills and other productive capacities required to host different economic activities, and how cities can best organize to grow faster and provide high quality jobs to their workforce.

This paper discusses the relevance of the economic complexity measure introduced by Hausmann & Hidalgo (2009) in order to forecast the population growth of cities and the local employment growth of industries. We find that this is the case: most empirical specifications confirm the presence of significant coefficients with the hypothesized signs, and show sizeable predictive accuracy for the future performance of cities and industries. Given these results, we consider that these metrics and the models presented above provide valuable forecasting tools that can help analysts and policy-makers make choices. We provide data visualization tools that are readily available for the use of parties interested in the discussion of local economic development of American cities.

In future work, we hope to focus on how tradable sectors develop locally according to the presence of their non-tradable inputs, on how other explicit

measures of technological proximity (input/output linkages, patent/R&D similarity, etc.) help improve predictions on local industrial growth, appearance and disappearance, and how the economic complexity analysis can be expanded to predict growth at the city/industry/occupation level.

6. Description of datasets

6.1. City/Industry data

The main data source for all our analyses is a city/industry jobs panel from EMSI. We aggregate EMSI's data at the Micro/Metropolitan geographic level (what we refer to as cities) and at the 4-digit NAICS industrial classification. While this local labor market data is proprietary and relies on modelling and imputation, we build on it as it provides trustworthy estimates of city/industry cells that are censored in official statistical sources. The industry/occupation data used to build the occupation-similarity explicit proximities was taken from the Occupational Employment Statistics (OES). From these baseline datasets we are able to reproduce all complexity metric calculations described above, and we can take metrics fixed at the city/year level (ECI, SI, etc.) for analyses predicting city growth using other city level data sources. Table 9 shows descriptive statistics for variables used in our models at the city/industry for 2012, which is our baseline year for out-of-sample prediction of city/industry job growth.

Variables	Description	Mean	Sd	Min	Med	Max
density_ocu	Occ. Similarity base density	0.26	0.07	0.00	0.25	0.96
density	Co-location based density	0.30	0.07	0.02	0.31	0.61
cog	Complexity Outlook Index	0.57	0.68	-1.09	0.28	2.63
rca	Revealed Comparative Advantage	1.93	9.57	0.00	0.79	884.67
jobs	Jobs	678	4204	0	78	457136

Table 9: Descriptive Statistics at City-Industry Level, 2012 (part 2)

6.2. City data

As mentioned above, we take a city panel of complexity variables to advance our city population growth predictions. For this purpose, we also include baseline controls on the demographics, productivity, wealth, age structure and education of American cities. Tables 10 and 11 present each of the variables used by group, providing a description, source and descriptive statistics for 2011, which is our baseline year for out-of-sample prediction of the population growth of cities.

	Variables	Description	Source	Mean	Sd	Min	Med	Max	Missing	Via
Complexity	eci	Jobs-based Economic Complexity Index	EMSI	0	1	-1.9	-0.21	3.63	0	EMSI
	coi	Jobs-based Complexity Outlook Gain	EMSI	0	1	-4.7	-0.04	3.43	0	EMSI
	diversity	Number of industries with an employment-measured presence higher than in the average CBSA	EMSI	69.86	15.99	8.00	69.00	120.00	0	EMSI
Demographics	population	Population of CBSA's population from ti to population years	U.S. Census Bureau	319708	1078551	13412	75657	19763868	0	Cluster Mapping Project API
	population_density	Total population per squared miles. Based in CBSA's area reported in the 2010 census	U.S. Census Bureau	154.67	230.93	1.73	88.03	2668.76	0	Cluster Mapping Project API
Productivity	manufacturing_intensity_tf	Manufacturing jobs as percentage of total jobs	U.S.. Bureau of Labor Statistics	0.1256088623	0.083993912	0.0	0.11	0.48	6	Cluster Mapping Project API
	innovation_tf	Utility Patents per 10k Employees	U.S. Patent and Trademark Office	5.18	8.68	0.06	2.65	114.78	96	Cluster Mapping Project API
	avg_firm_size_tf	CBSA's average firm size	U.S. Census Bureau	15.89	3.04	6.42	16.00	35.52	0	Cluster Mapping Project API
	utility_patents_tf	Number of utility patents	U.S. Patent and Trademark Office	117.33	598.50	0.00	6.00	10256.00	0	Cluster Mapping Project API
	patent_count_tf	Number of patents	U.S. Patent and Trademark Office	118.75	586.47	0.11	6.23	9927.58	104	Cluster Mapping Project API
	manufacturing_jobs_tf	Number of CBSA's jobs in manufacturing	U.S. Bureau of Labor Statistics	11809	34229	0	3452	519644	0	Cluster Mapping Project API
	labor_prod	Units of GRP by active worker	EMSI	91073	24736	56254	84922	287865	0	EMSI
Wealth	Median_Income	Median Wages	Brookings Metro	28621	4677	15964	28025	48500	554	Brookings Metro
	real_personal_income	Millions of 2009 chained dollars	Bureau of Economic Analysis	30777	73652	2466	9938	943692	536	Bureau of Economic Analysis
	gdp_pc_cmp	GDP per capita, 2005 real dollars	Moody'seconomy.com	38988	15428	2608	35936	180669	5	Cluster Mapping Project API
	gdp_pc	GDP per capita, 2009 real dollars	Bureau of Economic Analysis	40879	12157	18172	39026	136684	534	Bureau of Economic Analysis
	real_pc_personal_income	Real_Pc_Personal of CBSA's adult population with inco rson high school	Bureau of Economic Analysis	41791	6669	27366	41271	104038	536	Bureau of Economic Analysis
	median_household_income_	Median Household Income	Cluster Mapping Project API	52982	144893	24537	47994	1992016	0	Cluster Mapping Project API
	median_family_income_tf	Median Family Income	Cluster Mapping Project API	117790	179375	27392	61423	2430613	0	Cluster Mapping Project API

Table 10: Descriptive Statistics at City Level, 2011 (part 1)

	Variables	Description	Source	Mean	Sd	Min	Med	Max	Missing	Via
Age structure	popultn_0_t_4_prschl_tf	Share of CBSA's population from 0 to 4 years	U.S. Census Bureau	20743	69966	735	4561	1227891	0	Cluster Mapping Project API
	popltn_5_t_17_schl_g_tf	Popltn of CBSA's population from 5 to 17 years	U.S. Census Bureau	55256	186327	1693	12457	3244484	0	Cluster Mapping Project API
	ppltn_18_t_24_cllg_g_tf	Share of CBSA's population from 18 to 24 years	U.S. Census Bureau	32287	104656	931	7713	1864820	0	Cluster Mapping Project API
	ppltn_25_t_44_yng_dlt_t	Ppltn of CBSA's population from 25 to 44 to years	U.S. Census Bureau	85446	305431	2956	17643	5539202	0	Cluster Mapping Project API
	popultn_45_t_64_ldr_dlt_tf	Popultn of CBSA's population from 45 to 64 to years	U.S. Census Bureau	84443	282253	2611	21109	5267671	0	Cluster Mapping Project API
	popultn_65_nd_ldr_ldr_tf	Popultn of CBSA's population of 65 and older to years	U.S. Census Bureau	41533	133351	1403	11632	2619800	0	Cluster Mapping Project API
Education	share_lessthanhs	Share of CBSA's adult population with less than high school	Brookings Metro	0.13	0.06	0.0	0.12	0.38	554	Brookings Metro
	share_baplus	Share of CBSA's adult population with at least a bachelor degree	Brookings Metro	0.26	0.08	0.1	0.25	0.57	554	Brookings Metro
	share_hs	Share of CBSA's adult population with high school completed	Brookings Metro	0.30	0.06	0.1	0.30	0.48	554	Brookings Metro
	share_somecolaa	Share of CBSA's adult population with some cola high school	Brookings Metro	0.31	0.04	0.2	0.31	0.45	554	Brookings Metro

Table 11: Descriptive Statistics at City Level, 2011 (part 2)

References

- Alabdulkareem, A., Frank, M. R., Sun, L., Alshebli, B., Hidalgo, C., & Rahwan, I. (2018). Unpacking the polarization of workplace skills. Science Advances, 4(7). doi:10.1126/sciadv.aao6030.

- Autor, D. H., Levy, F., & Murnane, R. J. (2003). Skill Demand, Inequality, and Computerization: Connecting the Dots. Technology, Growth, and the Labor Market, 107-129. doi:10.1007/978-1-4615-0325-5_6.

- Autor, D. H. (2015). Why Are There Still So Many Jobs? The History and Future of Workplace Automation. Journal of Economic Perspectives, 29(3), 3-30. doi:10.1257/jep.29.3.3.

- Bahar, D., Stein, E., Wagner, R. A., & Rosenow, S. (2019). The Birth and Growth of New Export Clusters: Which Mechanisms Drive Diversification? World Development Journal. doi:10.2139/ssrn.3035605.

- Beaudry, C., & Schiffauerova, A. (2009). Whos right, Marshall or Jacobs? The localization versus urbanization debate. Research Policy, 38(2), 318-337. doi:10.1016/j.respol.2008.11.010.

- Campante, F. R., & Chor, D. (2017). "Just do your job": Obedience, Routine Tasks, and the Pattern of Specialization. Economic Research Institute for ASEAN and East Asia (ERIA)., DP-, 35th ser.

- Delgado, M., Porter, M., & Stern, S. (2012). Clusters, Convergence, and Economic Performance. US Census Bureau Center for Economic Studies. doi:10.3386/w18250.

- Delgado, M., Porter, M. E., & Stern, S. (2015). Defining clusters of related industries. Journal of Economic Geography, 16(1), 1-38. doi:10.1093/jeg:lbv017.

- Ellison, G., E. L. Glaeser, and W. R. Kerr (2010, jun). What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns. American Economic Review 100 (3), 11951213.

- Hausmann, R., & Rodrik, D. (2002). Economic Development as Self-Discovery. doi:10.3386/w8952.

- Hausmann, R., Hwang, J., & Rodrik, D. (2005). What You Export Matters. Journal of Economic Growth. doi:10.3386/w11905.

- Hausmann, R., & Klinger, B. (2006). Structural Transformation and Patterns of Comparative Advantage in the Product Space. SSRN Electronic Journal. doi:10.2139/ssrn.939646.

- Hausmann, R., Hidalgo, C., Stock, D. P., & Yildirim, M. A. (2014). Implied Comparative Advantage. SSRN Electronic Journal.

- Hausmann, R., Hidalgo, C. A., Bustos, S., Jimenez, J., Coscia, M., & Yildirim, M. (2014). The Atlas of Economic Complexity. Cambridge: MIT Press.

- Hausmann, R., & Hidalgo, C. A. (2009). The building blocks of economic complexity. Proceedings of the national academy of sciences, 106(26), 10570-10575. doi:10.1073/pnas.0900943106.

- Hidalgo, C. A., Balland, P., R. B., Delgado, M., Feldman, M., Frenken, et al. (2018). The Principles of Relatedness. ICCS, (Unifying Themes in Complex Systems), ix, 451-457. Doi: 10.1007/978-3-319-96661-8_46.

- Hidalgo, C. A., Klinger, B., Barabsi, A. L., & Hausmann, R. (2007). The product space conditions the development of nations. Science, 317(5837), 482-487.

- Santos, M. , Morales, J. R., & Hausmann, R. (2017). Panam Beyond the Canal: Using Technological Proximities to Identify Opportunities for Productive Diversification. Center for International Development.

- Scherer, F. (2003). Technology Flows Matrix Estimation Revisited. Economic Systems Research, 15(3), 327-358.