Are Big Cities Important for Economic Growth?*

Matthew A. Turner† and David N. Weil‡

January 2025

Abstract: Cities are often described as engines of economic growth. We assess this statement quantitatively. We focus on two mechanisms: a static agglomeration effect that makes production in bigger cities more efficient, and a dynamic effect whereby urban scale impacts the productivity of invention, which in turn determines the speed of technological progress for the country as a whole. Using estimates of these effects from the literature and MSA-level patent and population data since 1900, we ask how much lower US output would be in 2010 if city size had been limited to one million or one hundred thousand starting in 1900. These effects are small. If city sizes had been limited to one million people since 1900, output in 2010 would have been only 8% lower than its observed value.

JEL: O40, R10
Keywords: Agglomeration economies, Economic growth

*The authors are grateful to Enrico Berkes for generously sharing the CUSP patent database, to Daniele Goffi for research assistance, and to seminar audiences at UCLA and Brown for helpful comments.

†Brown University, Department of Economics, Box B, Brown University, Providence, RI 02912. email: Matthew_Turner@Brown.edu. Also affiliated with PERC, IGC, NBER, PSTC.

‡Brown University, Department of Economics, Box B, Brown University, Providence, RI 02912. email: David_Weil@Brown.edu. Also affiliated with NBER

1. Introduction

Cities are economic dynamos. They are hubs of innovation and breeding grounds for new industries. Highly skilled workers, entrepreneurs, and scientists congregate in cities to take advantage of the efficiencies of thick markets and the externalities associated with agglomeration. In 2010, the 27 cities that constituted the top decile of the MSA size distribution accounted for 47% of US population, 54% of output, and 65% of patents (the data are discussed below). Cities are often referred to as "engines of economic growth."

In this paper we evaluate this idea quantitatively. Our approach follows Fogel's (1964), analysis of the role of railroads in US economic growth. Prior to Fogel's work it was widely noted that late 19th century railroads carried the bulk of inter-regional trade, trade that was in turn a vital driver of growth. The natural conclusion was that railroads were necessary for that growth. Fogel's innovation was to note that, even though railroads were in practice the dominant carrier of freight, in a world without railroads the same freight traffic could still flow, only at higher cost – which he showed by constructing the transportation network that could have existed in such a case. Analogously, we would like to ask how much slower US economic growth would have been, or how much poorer the country would be today, if the large cities in which so much economic activity takes place had not existed. This question cannot be answered by simply observing how much output is produced in cities or how much inventive activity takes place there. Had big cities not existed, much of the benefit of agglomeration would have been lost, but there still would have been skilled workers, entrepreneurs, new ideas waiting to be discovered, and so on.

Our main tools for pursuing this agenda will be explicit estimates of urban scale effects in two specific dimensions. The first is a static agglomeration effect that makes production of goods and services in larger cities more efficient. The second is a similar agglomeration effect that makes production of inventions in large cities more efficient. Because new technologies raise productivity everywhere in the country and technological progress accumulates over time, this latter effect has an impact external to a particular city and works dynamically. We begin with existing estimates of the magnitudes of these effects. Using a simple growth model, we consider counterfactual scenarios where urban scale – specifically, the size of the largest cities – differs from the historically observed path. The gap between income (or growth) in the counterfactual relative to the historical baseline is our measure of the growth effects of cities.

Throughout our analysis, we take both the observed and counterfactual distributions of city sizes as given. This relieves us of the problem of specifying the equilibrium process that determines this distribution, a problem that is itself the subject of a large and distinguished literature (Duranton and Puga, 2023, Fajgelbaum and Gaubert, 2020, Hsieh

and Moretti, 2019, for example). If we were instead to treat realized and counterfactual distributions of city sizes as the consequences of some explicit equilibrium process, our conclusions would be specific to the particular driver of the distribution of city sizes under consideration, such as zoning restrictions or transport costs. This would reduce the generality of our results. By contrast, our approach, which follows the literatures on development and growth accounting (Caselli, 2005, Hulten, 2010), requires remarkably weak assumptions and allows an immediate mapping from data to results.

The rest of this paper is organized as follows. Section 2 describes the data on city-level population, output, and patents that we use, and presents an overview of their contemporary and historical relationship. Section 3 introduces our counterfactual approach to assessing the importance of urban scale and applies it to study the effect of the distribution of city sizes statically on total factor productivity. We specifically consider counterfactual scenarios in which MSAs in the US are limited in population to one million or one hundred thousand individuals. Section 4 then takes the same approach to study technological progress, using data on MSA-level patents to assess the impact of the city size distribution on inventive activity at a point in time. In Section 5, we then cumulate differences in inventive activity between our counterfactual and the baseline of the actual development of the US, to calculate the reduction in TFP due to slower technological progress that results from limitations on city sizes. In Section 6, we combine the static and dynamic effects to calculate the overall impact on output of our counterfactual restriction on city sizes. Section 7 concludes.

2. A First Look at the Data

We investigate how the distribution of city sizes affects aggregate output via two mechanisms that operate at the city level. The first is a static agglomeration effect that leads to increases in city level productivity as city size increases. The second is a similar increase in the productivity of cities at research as city size increases. Because research output improves economy wide productivity, scale effects in city level research productivity increase economy wide productivity, an effect that compounds over time. To set the stage for this investigation, we present data on the cross sectional relationship between city size, as measured by population, city output, and research, as measured by patents.

For our cities, we consider a set of 275 constant boundary MSAs in the continental US defined to the same boundaries as Duranton and Puga (2023), along with a single rural area that aggregates all non-metropolitan counties.[1] We construct decadal population data by combining population data in replication files from Duranton and Puga (2023)

[1]Strictly, our MSAs are the set of all CMSAs, MSAs and NECMAs drawn to 2000 boundaries.

with 1900-1990 county population data from Forstall and NBER (1995). This results in an MSA by decade panel of population stretching from 1900 to 2010.[2]

We measure output using the county level output data from the BEA (USDOC/BEA/RD, 2023), and aggregate counties to MSAs. These data are available beginning in 2000.[3] We rely on the CUSP data (Berkes, 2018), to measure patents. These data report on all patents issued by the US Patent office from 1836 to 2015, along with the year of issue and county of residence for all listed inventors. Using these data, and pro-rating patents with multiple inventors, we construct county-by-year counts of patents. Because MSAs are defined as collections of counties, we can easily aggregate to counts of patents produced in each MSA during each decade, e.g. 1900-1909, from 1900 to 2010.

Figure 1(a) is a histogram of population, output, and patents across cities for the year 2010. Cities are grouped in deciles by population, and we include an eleventh non-MSA category. The figure shows the importance of large cities. San Antonio, with a population of 1.99 million, is the smallest MSA in the top decile. In total, the top decile of cities accounted for 47% of population, 54% of output, and 65% of patents. Large cities have higher per-capita output and patent production than small cities, on average. Non-metropolitan counties accounted for 19% of population, 14% of output and 6% of patents.

Figures 1(b) and 1(c) repeat the analysis of Figure 1(a) for the years 1900 and 1950, although we have no MSA-level output data for these years. The concentration of patenting in the largest decile of cities is less pronounced in 1900 and 1950 than in 2010. In 1900 there is also a significant over-representation of patents in the second largest decile of cities. The under representation of non-MSA areas in patenting is more pronounced in the earlier years.

Figure 2 describes correlations in our data in 2010. Panel (a) describes the relationship between the log of MSA output and log population. The tight linear relationship of the logs implies an elasticity of output per capita with respect to population of 8%. This is slightly smaller than the 13% elasticity reported by Glaeser and Gottlieb (2009) for the same regression using data for 2000 and slightly different MSA definitions. Panel (b) describes the relationship between the log of MSA patents and the log of population. The relationship between patents and population is noticeably nosier than is the relationship between output and population, and also considerably steeper. The elasticity of patents to city population in the raw data is 41% in 2010. Panel (c) describes the relationship between patents and output. Unsurprisingly, this relationship is also positive. More

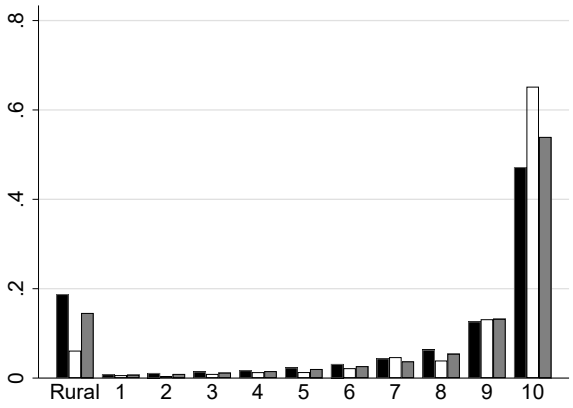
[2]Our sample of MSAs decreases slightly in the early part of our sample. This largely reflects the fact that some states had not yet joined the union and so census data and county boundaries do not exist. For example, Arizona, New Mexico and Oklahoma all joined the US after 1900.

[3]The BEA productivity data does not report for the two counties that make up the Danville, VA MSA, and so the BEA data describes 274 MSAs instead of 275.

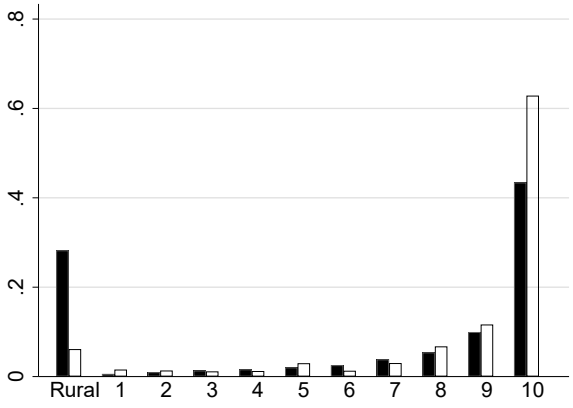
interesting is the fact that many smaller cities appear to be relatively specialized in innovation (e.g., Corvalis, Boise, or Rochester, MN) or output, while larger cities almost always do both. San Francisco is an obvious exception. It is the only large MSA that is relatively specialized in innovation.

Figure 3 describes the relationship between patents and city size over time. Panel (a) repeats panel (b) of figure 2. The relationship between patenting and city size is always positive, but flatter as we go back in time. The slope of the regression lines in 1950 and 1900 are 1.35 and 1.11. versus 1.41 in 2010. Looking at the names of cities that are particularly successful at innovation shows the westward march of the center of innovations, from New England in the 19th century, to the Midwest in the 20th century, to the West in the 21st century.

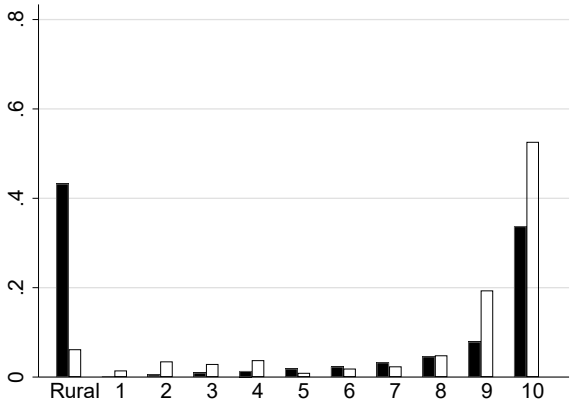
Figure 1: Distribution of population, output and patents by city size in 1900, 1950 and 2010



(a) 2010



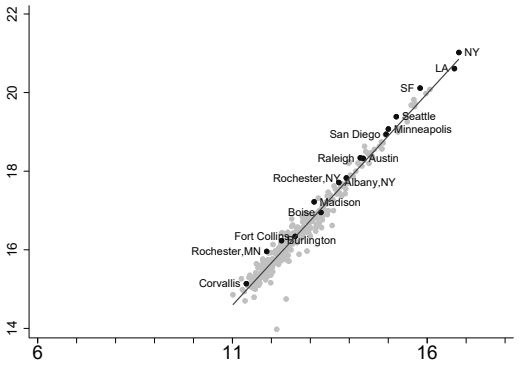
(b) 1950



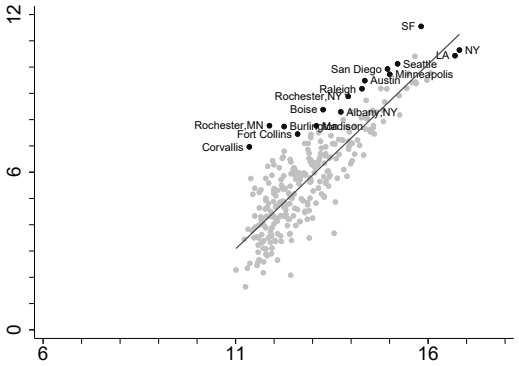
(c) 1900

Note: *(a) Share of population, patents and total output by deciles of city size and rural status for 2010. (b) Share of population and patents by deciles of city size and rural status for 1950. (c) Same as (b) but for 1900. In each panel, black bar is population share and white bar is patent share. In panel (a) the gray bar is output share.*

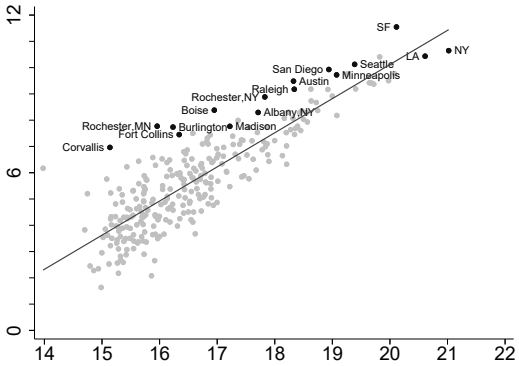
Figure 2: Joint distribution of output, patents and city size in 2010.



(a) Output vs Pop. 2010



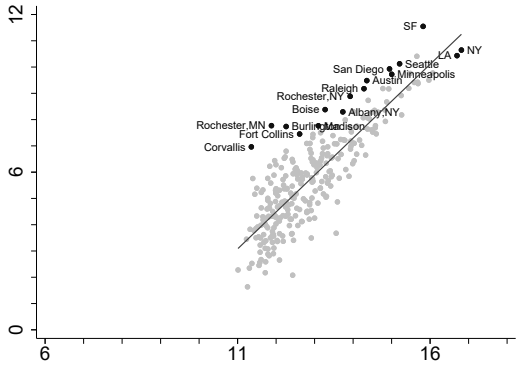
(b) Patents vs Pop. 2010



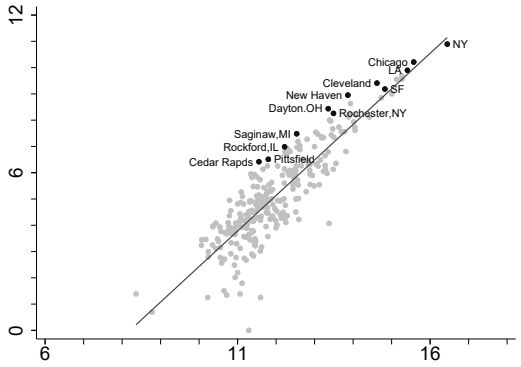
(c) Patents vs Output. 2010

Note: *Each panel shows a scatter plot and OLS regression line. (a)* $\ln(Output)$ *vs.* $\ln(Population)$ *2010;* $\beta = 1.08, s.e. = 0.013$. *(b)* $\ln(Patents)$ *vs.* $\ln(Population)$ *2010;* $\beta = 1.41, s.e. = 0.053$. *(c)* $\ln(Patents)$ *vs.* $\ln(Output)$ *2010;* $\beta = 1.29, s.e. = 0.047$.

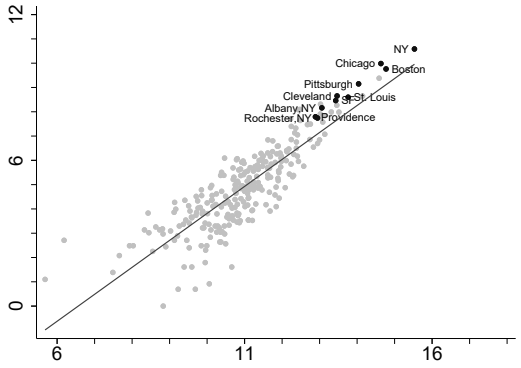
Figure 3: Joint distribution of population and patents over time.



(a) Patents vs Pop. 2010



(b) Patents vs Pop. 1950



(c) Patents vs Pop. 1900

Note: *Each panel shows a scatter plot and OLS regression line. (a)* $\ln(Patents)$ *vs.* $\ln(Population)$ *2010;* $\beta = 1.41, s.e. = 0.053$. *(b)* $\ln(Patents)$ *vs.* $\ln(Population)$ *1950;* $\beta = 1.35, s.e. = 0.046$. *(c)* $\ln(Patents)$ *vs.* $\ln(Population)$ *1900;* $\beta = 1.11, s.e. = 0.041$.

3. Urban Scale Economies for Output

We would like to calculate the total value of output for a counterfactual version of the US in which certain cities are constrained to be smaller than they are. We treat MSAs as the real world analog of our theoretical cities, and index them by $i = 1,...,n$. To represent the non-MSA sector of the economy, we use the index $i = 0$. (We often abuse language and refer to i as indexing "cities.") Y_{it} denotes output of location i in decade t, L_{it} population, K_{it} physical capital, ℓ_{it}^Y the fraction of the population engaged in the production of output, h_{it}^Y the human capital of workers engaged in producing output, and A_{it} is city-level productivity in producing output. We assume that a city transforms inputs into outputs according to

$$Y_{it} = A_{it} \left(K_{it}\right)^{\gamma} \left(h_{it}^Y \ell_{it}^Y L_{it}\right)^{1-\gamma}. \tag{1}$$

We are interested in understanding how changing the size distribution of cities would affect aggregate output. To proceed, we decompose A_{it} into three components. The first is a time-specific national component common to all cities, $\bar{A}_t$. As discussed further below, we identify the growth of $\bar{A}_t$ over time with technological progress. The second term is a city specific scale effect that depends on population, $\widetilde{A}_{it}$. This is the agglomeration or urbanization economy so often estimated in the urban economics literature. We refer to this term an "urban scale effect for output" to distinguish it from a similar effect for patents. Finally, $\widehat{A}_{it}$ is a city-decade specific idiosyncratic term. This term reflects transitory idiosyncratic shocks to a location's productivity. We assume that $\sum_i \widehat{A}_{it} = n + 1$, so that its mean value is one in each period. Altogether, we have,

$$A_{it} = \widehat{A}_{it} \bar{A}_t \widetilde{A}_{it}. \tag{2}$$

We assume that the urban scale effect for output depends on city population. Because the rural sector is large and unitary, we must find some other way to assess scale effects for the rural area. To proceed, we note that the minimum population for an MSA is 50,000, and assign rural areas the same scale effect as a city of this size, regardless of how many people live in rural areas. To implement this, we define $\widetilde{L}$ as the population that is relevant for urban scale effects.

$$\widetilde{L}_{it} = \begin{cases} L_{it} & \text{for } i > 0 \text{ (MSAs)} \\ \underline{L} & \text{for } i = 0 \text{ (rural areas).} \end{cases} \tag{3}$$

For MSAs, $\widetilde{L}$ is the same as population, but for the aggregate rural area we cap urban scale effects at the level of the smallest MSA, that is, at $\underline{L} = 50{,}000$. As a practical matter, the size of the rural population is invariant across all of the counterfactuals we conduct in the body of the paper, and so this assumption will not affect our main results.

This assumption is required in Appendix A, however, and introducing it here allows consistent notation throughout.

Using this notation, we specify the scale effect as

$$\widetilde{A}_{it} = \widetilde{L}_{it}^{\sigma_A} \tag{4}$$

We discuss the values we use for the parameter σ_A later in this section. Combining equations (1), (2), and (4) results in a production technology that nests production functions commonly used to study systems of cities, e.g., Desmet and Rossi-Hansberg (2013), Duranton and Puga (2023).

We assume that physical capital is freely mobile among cities and that this equalizes its marginal product. This implies that,

$$\frac{Y_{it}}{K_{it}} = \frac{Y_t}{K_t} \quad \forall i \tag{5}$$

Substituting (5) into the production function (1) and rearranging, we have

$$Y_{it} = A_{it}^{1/(1-\gamma)} \left(\frac{K_t}{Y_t}\right)^{\gamma/(1-\gamma)} h_{it}^Y \ell_{it}^Y L_{it} \tag{6}$$

Summing over all cities, we get aggregate output,

$$Y_t = \left(\frac{K_t}{Y_t}\right)^{\gamma/1-\gamma} \sum_i A_{it}^{1/(1-\gamma)} h_{it}^Y \ell_{it}^Y L_{it}. \tag{7}$$

We would like to compare the observed, or "base" case, to an alternative where some cities take counterfactual sizes. When necessary, we indicate the value of variable X in the two cases with superscripts, X^{base} and X^{alt}.

We assume that the aggregate ratio of capital to output, K/Y, does not differ between the two cases at a point in time. The simplest justification for holding the K/Y ratio fixed is if the country is open to the world capital market with the interest rate not varying across scenarios. Alternatively, Romer (2012) shows that if capital accumulates via a fixed investment rate (as in Solow (1957)) then the capital to output ratio is constant along any balanced growth path.[4] Empirically, Feenstra et al. (2015) show that across countries capital-output ratios do not vary systematically with income. Similarly Jones and Vollrath (2020) show the relative constancy of this ratio over time within countries that are arguably close to their balanced growth paths.

[4]If differences between our scenarios were purely in terms of the level of productivity but not its growth rate, this condition would hold. However, in later sections, we will allow for reduced agglomeration to affect the growth rate of aggregate productivity, and so the condition is no longer exact. If limiting agglomeration slowed technological progress, then the K/Y ratio would rise, partially offsetting the effect of slower productivity growth on the level of output. Thus differences in output between the baseline and alternative cases, which we find to be small under our assumption of fixed K/Y, would be even smaller if we did not make that assumption.

Assuming a fixed K/Y ratio allows us to incorporate "induced capital accumulation" (Klenow and Rodriguez-Clare, 1997, Hall and Jones, 1999): we expect that a change in city level productivity to affect the level of income, and hence the quantity of investment. This change in investment then feeds back to affect output.

We restrict attention to alternative cases where a city's population is unchanged but the size of the urban scale effect on productivity ($\widetilde{A}_{it}$) is reduced to that of a smaller city. All other characteristics of the city, h_{it}, ℓ_{it}^{Y}, and the city-decade component of productivity, $\widehat{A}_{it}$, remain constant. We can also imagine this occurring if the observed population of the city L_{it}^{base} is divided into $\frac{L_{it}^{base}}{L_{max}}$ daughter cities, each with population L_{max}, with human capital equally divided among them, and with all of the daughter cities having the same values of $\widehat{A}_{it}$ and ℓ_{it}^{Y} as the original city. We assume that non-metropolitan output does not change between base and alternative cases.

Because congestion effects are not part of the production process of equation (1), and because production in the absence of the agglomeration effect is CRS, perfect mobility of all factors of production would lead to an equilibrium in which all production took place in the city with the highest value of $\widehat{A}_{it}$. We are thus implicitly considering equilibrium population levels that are partly determined by an unspecified congestion process.

With these assumptions in place, we can use equation (7) to compare aggregate output in economies with different values of A_{it}. Multiplying each term in the sum on the right hand side of (7) by $\left(\frac{A_{it}^{base}}{A_{it}^{base}}\right)^{1/(1-\gamma)}$ and using equation (7) again to simplify, we have

$$Y_t^{alt} = \sum_i Y_{it}^{base} \left(\frac{A_{it}^{alt}}{A_{it}^{base}}\right)^{1/(1-\gamma)}.$$

Dividing by Y_t^{base} gives

$$\frac{Y_t^{alt}}{Y_t^{base}} = \sum_i \frac{Y_{it}^{base}}{Y_t^{base}} \left(\frac{A_{it}^{alt}}{A_{it}^{base}}\right)^{1/(1-\gamma)}. \tag{8}$$

We would like to evaluate the effect on the output of a particular city of constraining its productivity to the that of a city of size no greater than L_{max}. Because city size enters a city's TFP, A_{it}, only through the static scale effect of equation (4), the ratio of observed to counterfactual city productivity is,

$$\frac{A_{it}^{alt}}{A_{it}^{base}} = min\left(1, \left(\frac{L^{max}}{\widetilde{L}_{it}}\right)^{\sigma_A}\right). \tag{9}$$

Using equation (9) and (8) together, we can evaluate aggregate output for a counterfactual system of cities in which all cities with population about the threshold level L_{max} have their productivity reduced to that of a city of the threshold size. The resulting

change in aggregate output is,

$$\frac{Y_t^{alt}}{Y_t^{base}} = \sum_i \frac{Y_{it}^{base}}{Y_t^{base}} \min\left(1, \left(\frac{L^{max}}{\widetilde{L}_{it}}\right)^{\sigma_A/(1-\gamma)}\right) \tag{10}$$

In equation (10) we see the advantage of restricting attention to our particular counterfactuals. For these counterfactuals we can evaluate the change in aggregate output without measures of city-specific physical and human capital or of the other parts of city productivity term, $\widehat{A}_{it}$ and $\bar{A}_t$. We require only city population and output.[5]

A *Human Capital Extension*

The fact that output per capita is higher in big cities, as we see in figure 2(a), is a robust finding of the empirical literature on agglomeration economies. A more difficult question has been estimating the share of the raw correlation that should be attributed urban scale effects vs. the share due to the sorting of more productive people into bigger cities.

The initial approach to this problem was to estimate the relationship between city size and wages conditional on individual characteristics, e.g., Combes et al. (2008) or Glaeser and Gottlieb (2008). Including individual characteristics typically reduces the wage elasticity of city size by one third to one half, and the resulting conditional elasticity is interpreted as the causal effect of city size on the level of productivity.

Following Glaeser and Maré (2001), recent research (De la Roca and Puga, 2017, Duranton and Puga, 2023) follows workers over time and finds that the productivity of a worker increases more rapidly in bigger cities. An effort to account for differences in worker productivity across cities suggests that most of the difference can be accounted for by an increase in the rate of worker productivity growth as city size rises.

Summing up, the older literature asks whether urban workers are more productive because they are different than other workers when they arrive in the city, while the more recent literature asks whether urban workers are more productive because they become different from other workers after they arrive in the city. While the conceptual difference is clear, distinguishing the two cases empirically is obviously tricky, and research on the question is in its early stages. With that said, the evidence in De la Roca and Puga (2017) and Duranton and Puga (2023) suggests that most differences in worker productivity arise after people arrive in the city, not before. This implies that the relationship between output and city size that we see in the raw data, e.g., Figure 2, is actually close to the

[5]Estimates of city-specific productivity, $\widehat{A}_{it}$, for example, face a series of econometric problems. Does a particular city produce high output relative to its measured human capital because it has a high idiosyncratic productivity due to location or institutions, or because we do not properly measure the quality of human capital? This problem will recur in our analysis of city level research productivity. These problems are carefully described in Combes et al. (2010) and Glaeser and Gottlieb (2008).

causal effect of city size on output, but that the city size effect operates both through an effect on TFP and through an effect on human capital. We here extend our model to include both of these effects.

We assume that human capital is a function of city-decade specific inputs (years of education and their quality), which we denote S_{it}. We further allow the Mincerian return to these inputs, denoted ϕ_{it} to vary at the city-decade level. Finally, we allow an urban scale effect similar to the one for producing output, represented by the parameter σ_H :

$$h_{it}^{Y} = exp(\phi_{it} S_{it}) \widetilde{L}_{it}^{\sigma_H}, \tag{11}$$

Next, we use equation (7) to write aggregate output for a counterfactual case where city size is capped at L_{max} and multiply the right hand side by

$$\left(\frac{A_{it}^{base}}{A_{it}^{base}} \right)^{1/(1-\gamma)} \frac{h_{it}^{Y,base}}{h_{it}^{Y,base}}. \tag{12}$$

Following the same logic that leads from equation (7) to (10), we arrive at the corresponding expression for aggregate output when human capital production is subject to scale effects,

$$\frac{Y_t^{alt}}{Y_t^{base}} = \sum_i \frac{Y_{it}^{base}}{Y_t^{base}} min \left(1, \left(\frac{L_{max}}{\widetilde{L}_{it}} \right)^{\frac{\sigma_A}{1-\gamma} + \sigma_H} \right) \tag{13}$$

Comparing equation (13) to equation (10), we see that the two expressions are identical save for the interpretation and magnitude of the exponent on the term $\left(\frac{L_{max}}{\widetilde{L}_{it}} \right)$. Therefore, for the purpose of evaluating counterfactual scenarios, we evaluate the model with or without scale effects in the production of human capital by varying the magnitude and interpretation of this exponent.[6]

B *Parameterization*

Given the data already in hand, evaluating equation (13) requires only that we evaluate $\frac{\sigma_A}{1-\gamma} + \sigma_H$. We follow the standard in the growth literature and set the capital share of output, γ, to 0.33. To evaluate σ_A and σ_H, we rely on the large literature estimating the relationship between city size and productivity. Table 1 lists estimates of these parameters derived from six prominent empirical papers.

[6]Our description of the effect of city size on human capital simplifies the problem by assuming that human capital accumulated in a city can never migrate to another city. A better, but much less tractable, description of the process would allow people to accumulate human capital in one city and employ it in another. Implementing such a model would require vastly more data than the exercise we conduct. A tractable alternative to our approach would be to allow human capital to be freely mobile across cities in parallel to physical capital. In this case, rather that being "too attached" to the cities where it is accumulated, human capital is "too unattached" to the people who accumulate it.

Ciccone and Hall (1996) is an early effort to estimate urban scale effects and, like us, explicitly accounts for both physical and human capital. Although their estimation strategy is indirect, the parameters that Ciccone and Hall (1996) estimate correspond closely to ours. Their benchmark estimate implies that $\sigma_A = 3.4\%$ and that σ_H is indistinguishable from zero.[7] These values together imply $\frac{\sigma_A}{1-\gamma} + \sigma_H = 5.1\%$.

Most other efforts to estimate urban scale effects rely on regressions of the logarithm of wages or output on city size. To use these estimates for our purpose, note that from (8) we can derive the relationship between city size and output as,

$$\frac{\partial \log Y_{it}}{\partial \log L_{it}} = 1 + \frac{\sigma_A}{1-\gamma}. \tag{14}$$

If we make the further assumption that wages reflect the marginal productivity of labor, then we can also derive the relationship between city size and wages,

$$\frac{\partial \log w_{it}}{\partial \log L_{it}} = \frac{\sigma_A}{1-\gamma}. \tag{15}$$

The left hand side expressions in both of these equations are quantities that can be measured empirically and are the subject of much of the empirical literature on agglomeration effects. To estimate either $\frac{\partial \log Y_{it}}{\partial \log L_{it}}$ or $\frac{\partial \log w_{it}}{\partial \log L_{it}}$, one must distinguish between the quantity of interest, the pure effect of scale, and several potential confounders: the propensity of more productive people to sort into cities, the possibility that people accumulate human capital more quickly in cities, and the possibility that people accumulate at places that are intrinsically more productive. The literature has proposed a variety of solutions to these problems (see Rosenthal and Strange (2004) and Combes and Gobillon (2015) for surveys).[8]

Each of Combes et al. (2008), De la Roca and Puga (2017) and Duranton and Puga (2023) relies on a panel of individual workers to examine the relationship between wages and city size. In French data, Combes et al. (2008) find that the city size elasticity of wages is about 5.1%. After controlling for individual fixed effects, this drops to about 3.7%. Combes et al. (2008) implicitly assume that worker characteristics are unaffected by the size of the city where they work, and using equation (15) their estimates imply $\sigma_A = 2.5\%, \sigma_H = 0$ and that $\frac{\sigma_A}{1-\gamma} + \sigma_H = 3.7\%$

[7] Using their benchmark estimate of $\widehat{\theta} = 1.052$ (from Table 1) in their equation (20) and our assumption that the capital share is 0.33, we have that (their notation) $\gamma = 1.034$. Inspection of their equation (3) confirms that this parameter corresponds to our σ_A. Similarly, the benchmark estimate of η is not distinguishable from zero, and inspection of their equation (3) confirms that this parameter corresponds to our σ_H.

[8] We note that the literature is generally careful to distinguish between the effects of city size and city density. To simplify our analysis, as is common in much of the theoretical literature, we abstract from this distinction and treat the two concepts as interchangeable.

Table 1: Estimates of σ_A

σ_A	σ_H	$\frac{\sigma_A}{1-\gamma} + \sigma_H$	Source	Data
3.4%	~ 0	5.1%	Ciccone and Hall (1996)	US, State output, 1988
2.5%	$\equiv 0$	3.7%	Combes et al. (2008)	French, Ind. wages, 1976-98
1.5%	2.9%	5.1%	De la Roca and Puga (2017)	Spanish, Ind. wages, 2004-9
2.9%	3.1%	7.6%	Duranton and Puga (2023)	US, Ind. wages, ca. 1979-2020
2.7%	$\equiv 0$	4.1%	Glaeser and Gottlieb (2008)	US, Ind. wages, 2000
8.6%	$\equiv 0$	13%	Glaeser and Gottlieb (2009)	US, MSA output, 2000

Note: *Various estimates of the static scale effect, σ_A and the human capital scale effect, σ_H from the literature. "$\equiv 0$" indicates a quantity implicitly assumed to be zero.*

Using Spanish data, De la Roca and Puga (2017) conduct a similar exercise and find that the city size elasticity of wages is about 5.1%. After controlling for individual fixed effects, this drops to about 2.2%. Unlike, Combes et al. (2008), however, De la Roca and Puga (2017) attribute the 2.9% difference to more rapid accumulation of human capital in larger cities rather than sorting. Again using (15), these estimates suggest $\sigma_A = 1.5\%$, $\sigma_H = 2.9\%$, and $\frac{\sigma_A}{1-\gamma} + \sigma_H = 5.1\%$ Duranton and Puga (2023) replicate De la Roca and Puga (2017) for the panel of US workers described by the NLSY79 and find that the city size elasticity of wages is about 7.6%. This drops to 4.4% after controlling for individual fixed effects, with the 3.1% difference attributed to more rapid human capital accumulation in larger cities. These estimates suggest $\sigma_A = 2.9\%$, $\sigma_H = 3.1\%$, and $\frac{\sigma_A}{1-\gamma} + \sigma_H = 7.6\%$.

The individual level data employed in Combes et al. (2008), De la Roca and Puga (2017) and Duranton and Puga (2023) allows state of the art decomposition of scale effects into human capital/sorting and pure scale effects. However, these papers are based on French, Spanish and the highly selected NLSY sample of US workers. The final two papers in table 1 are based on representative samples of US data. Glaeser and Gottlieb (2008) look at the relationship between wages and city size using a large cross-section of US workers. They estimate that the city size elasticity of wages is about 4.1%. These estimates suggest $\sigma_A = 2.7\%$, $\sigma_H = 0$, and $\frac{\sigma_A}{1-\gamma} + \sigma_H = 4.1\%$.

Glaeser and Gottlieb (2009) estimate the relationship between city level output and population using data on US cities in 2000, finding that city size elasticity of output is about 13%. This estimate does not correct for the possibility of sorting or more rapid urban human capital accumulation in cities. These estimates suggest $\sigma_A = 8.6\%$, $\sigma_H = 0$, and $\frac{\sigma_A}{1-\gamma} + \sigma_H = 13.0\%$.

The estimates of $\frac{\sigma_A}{1-\gamma} + \sigma_H$ presented in table 1 range from about just under 4% to 13%. However, four of the six estimates are within about 1% of the bottom end of this range. Given this, our preferred value of $\frac{\sigma_A}{1-\gamma} + \sigma_H$ for our calculations is 4%. With that said,

Table 2: Output in 2010 for two counterfactual size caps and values of σ_A.

$\frac{\sigma_A}{1-\gamma} + \sigma_H$	$L_{max} = 1m$	$L_{max} = 100k$
0.04	0.94	0.84
0.08	0.88	0.72
0.12	0.83	0.62

Note: *Each cell reports output relative to the baseline for a particular cap on city size and value of the exponent in equation* (13)*. For the purpose of this calculation, the rural population is treated as an extra MSA whose output is constant across scenarios and the capital share of output, γ, is equal to 0.33.*

given that there is still some variation around this estimate, we also consider 8% and 12% in our calculations.

To evaluate equation (13) we use the data on city level output and population described above. We evaluate the static effect of agglomeration for data from the year 2010, considering two possible values of maximum city size, $L_{\max}$: 1,000,000 and 100,000. Note that even our mildest comparative static, capping city size at 1,000,000 involves a catastrophic reorganization of the economy. The smallest US city with a population above 1m in 2010 was Fresno CA, the 52nd largest MSA in country. In 2010 the largest 52 cities housed 58% of the population and produced 66% of output. A cap of 100,000 would require reorganizing 261 MSAs.

4. Urban Scale Economies for Patents

The analysis in Section 3 takes the city invariant component of productivity, $\bar{A}_t$, as given. Changes in this parameter over time reflect technological progress, which we now examine. We proceed in three steps. First, we consider the relationship between city size and the production of patents. Second, we consider the relationship between patents and effective research effort, a term that we define more precisely below. Combining the first two steps we can investigate the relationship between city sizes and effective research effort, decade by decade. Finally, we consider the relationship between effective research effort and changes in $\bar{A}_t$. The first of these three steps is taken in this section, while the second and third are in Section 5. Putting the three steps together, we describe the relationship between the distribution of city sizes and the speed of technological progress.

Our analysis of the relationship between city sizes and patents parallels the treatment of the relationship between city sizes and output. The number of patents produced in a city-decade, P_{it}, depends on the size of the research labor force (specifically, city population multiplied by the share of people working in R&D, ℓ_{it}^R), the human capital of those research workers, h_{it}^R, and a city-decade patent productivity multiplier, B_{it},

according to the function,[9]

$$P_{it} = B_{it} h_{it}^{R} \ell_{it}^{R} L_{it}. \tag{16}$$

Summing over cities within a year, we have aggregate patent production,

$$P_t = \sum_i B_{it} h_{it}^{R} \ell_{it}^{R} L_{it}. \tag{17}$$

City-decade patent productivity can be decomposed into three components: a time specific national component common to all cities, $\bar{B}_t$; a city specific agglomeration effect that depends on population, $\widetilde{B}_{it}$; and, a city-decade specific idiosyncratic term, $\widehat{B}_{it}$. More formally,

$$B_{it} = \widehat{B}_{it} \bar{B}_t \widetilde{B}_{it}. \tag{18}$$

We model the scale effect in producing patents in the same way we did for output, but with a different value of returns to scale parameter,

$$\widetilde{B}_{it} = \widetilde{L}_{it}^{\sigma_B}. \tag{19}$$

We do not restrict the relationship between city-specific output productivity (the $\widehat{A}_{it}$'s) and city-specific patent productivity (the $\widehat{B}_{it}$'s). Places can be good at one but not the other. Nor do we restrict the relationship between the quality of human capital used to produce output, h_{it}^{Y} and that used to produce patents, h_{it}^{R}. Two cities may have the same numbers of Ph.D.s working in production but different numbers of Ph.D.s working in research.

As in the previous section, we consider the thought experiment where the urban scale effect on research productivity, $\widetilde{B}_{it}$, take the value that would hold if the city were constrained to maximum size L_{max}. To evaluate the resulting change in counterfactual patent production, we use the same approach that as in our analysis of counterfactual output, adjusting for the fact absence of capital. To proceed, first use (17) to write aggregate research output for the counterfactual case. Second, multiply the right hand side by $\frac{B_{it}^{base}}{B_{it}^{base}}$. Third, rearrange and use equations (18) and (19) to get

$$P_t^{alt} = \sum_i P_{it}^{base} min\left(1, \left(\frac{L_{max}}{\widetilde{L}_{it}}\right)^{\sigma_B}\right). \tag{20}$$

Finally, divide both sides by P_t^{base} to get

$$\frac{P_t^{alt}}{P_t^{base}} = \sum_i \frac{P_{it}^{base}}{P_t^{base}} min\left(1, \left(\frac{L_{max}}{\widetilde{L}_{it}}\right)^{\sigma_B}\right). \tag{21}$$

[9]To simplify the model, we assume that physical capital is not used for the production of research output.

In words, equation (21) allows us to calculate the level of patenting under a counterfactual system of cities from observed city and national patenting, observed city populations, and assumed counterfactual city populations.

In fact, equation (21) describes the largest possible change that could result from a change to city sizes. To see this, consider the case in which there is a city of two million people, of whom 20,000 are engaged in research. It seems likely (see next section) that agglomeration effects in research depend on the number of other researchers in a city, rather than the number of people overall. This means that one could imagine splitting the parent city into two daughter cities, each with one million people, but with one daughter city containing all 20,000 researchers. In that case, patent production would not fall at all. By dividing up the resources devoted to research proportionally with population, as in equation (21), we maximize the effect of reductions in city size on the output of patents.

A *Parameterization*

The key parameter required for the calculation in equation (21) is σ_B, the effect of city size on patent production. A large literature establishes that, for people working in knowledge intensive activities, proximity to other people working in similar industries has important effects on productivity, and also that the benefits of proximity fall off rapidly with distance. For example, Arzaghi and Henderson (2008) show that a few hundred meters of distance from an incumbent firm has a large impact in the location choice of an entrant, while Atkin et al. (2022) shows the importance of face-to-face contact for patent citations. Similarly, Carlino and Kerr (2015) use results in Rosenthal and Strange (2003) to calculate that the benefits of proximity decrease about five times more quickly with distance for software production than for metal fabrication. There is also evidence that inventive or innovative activity is much more likely to cluster together than it would if firms chose locations at random, e.g. Inoue et al. (2019). For a useful survey of both literatures, see Carlino and Kerr (2015) and Kerr and Kominers (2015). These papers strongly suggest the existence of scale effects and suggest that they are more important for innovation and invention than for most other types of economic activity, but they are less helpful for thinking about how scale effects vary with the size of a city.

Carlino et al. (2007) applies more directly to our case. This paper estimates a cross-sectional regression of patents per person on employment density in US MSAs around 2000. They they estimate that the elasticity of patents per person to employment density is 17-20%. Finally, Moretti (2021), constructs a panel of US inventors and their patenting activity by year, sector, and BEA economic area (slightly larger than an MSA). Controlling for inventor fixed effects, this paper estimates that the elasticity of inventions per inventor

Table 3: Patents during 2000-9 for three counterfactual size caps and values of σ_B.

σ_B	$L_{max} = 1m$	$L_{max} = 100k$
0.06	0.93	0.83
0.20	0.82	0.58

Note: *Each cell reports the share of total patents during 2000-2009 relative totals reported in the CUSP data ((Berkes, 2018)), for a particular cap on city size and value of σ_B. For the purpose of this calculation, the rural population is treated as an extra MSA whose patents are constant across scenarios.*

with respect to the number of inventors in a sector-cluster is 5% to 9%, depending on specification, with a preferred estimate around 6%.

Summing up, patenting seems to increase with city size at least as rapidly as does output or wages, and probably more quickly. Moretti (2021) is the only estimate based on disaggregated panel data, and suggests values of $\sigma_B \in [0.05,0.09]$. Carlino et al. (2007) uses only cross-sectional data, and so is less able to address reverse causation and sorting than Moretti (2021), but suggests $\sigma_B \in [0.17,0.20]$.

For our baseline calculation, we rely on the Moretti (2021) estimate of $\sigma_B = 6\%$ because it is based on higher quality data and and an econometric strategy that is better able to control for unobservable attributes of people and cities. With this said, consistent with the estimates from Carlino et al. (2007), we also consider the much larger value $\sigma_B = 20\%$.

Table 3 shows national patent production for 2000-9 in the alternative case where cities are limited in size relative to the observed case. We consider a range of values of L_{max} as in table 2. When scale economies in patenting are set at our base-case value of $\sigma_B = .06$ and we cap city size at one million, patent output falls by only 7% relative to baseline. This magnitude is similar the static urban scale effect on output production shown in table 2 for a similar value of σ_A. We find this surprising. Because patenting is more concentrated in larger cities than output, we expected that restricting city size would be more harmful to patenting than to output. In fact, this intuition appears not to be economically important. Larger declines in aggregate patenting are possible, but require the catastrophic counterfactual changes associated with L_{max} equal to 100,000, or the Carlino et al. (2007) value of σ_B estimated on cross-sectional data rather than than the smaller value derived from panel data. In the most extreme case, where we consider the largest plausible value of σ_B and cap city size at 100,000, patenting falls by 42%.

We can also examine the evolution of patenting over time. For this purpose, we restrict attention to our baseline value of $\sigma_B = 0.06$ and calculate how patenting changes in each of our two counterfactual systems of cities. Figure 4 presents our results. As one would expect, restricting city size to a fixed level would have had a much smaller effect on patenting early in the 20th century, when cities were much smaller.

Figure 4: Share of total patents produced under two hypothetical city networks

Note: *Counterfactual patents as a fraction of actual patents reported in CUSP when city sizes are capped at 1m (diamonds) and 100k (squares). Calculations assume* $\sigma_B = 0.06$.

5. From Patents to the Level of Technology

The previous section analyzes how patenting in every decade would have differed from its observed value in a counterfactual scenario where city sizes were restricted. We now ask how such a restriction would affect the 2010 level of the time-specific national component of productivity, $\bar{A}_t$, which we identify as the level of technology. More specifically, we can observe in the data actual path of technology $(\bar{A}^{\text{base}}_{1900}, ...\bar{A}^{\text{base}}_{2010})$, and we assume that that the levels of technology in the base and alternative cases are the same in 1900. Our goal is to construct the post-1900 path for $\bar{A}^{alt}$. On the alternative path, restricted city sizes would have led to less patenting and slower technological progress.

Our starting point is the framework relating research and technological progress laid out in Bloom et al. (2020). The key equation is

$$\frac{\dot{A}(t)}{A(t)} = \alpha S(t)^{\lambda} A(t)^{-\beta}. \tag{22}$$

Here, $S(t)$ is research effort and $A(t)$ is the level of technology, both in continuous time.

In their application to the aggregate US economy, $S(t)$ is gross domestic investment in intellectual property products from the National Income and Product Accounts deflated, by a measure of the nominal wage for high-skilled workers. The parameter λ captures the "stepping on toes" effect, whereby a the rate of technological progress may not scale linearly with research effort. The parameter β captures the extent to which ideas become harder to find as more of them have been discovered, the "fishing out effect".

We make two changes to this framework. First, we allow for the input to technological progress to depend not only on research spending, but also on the degree of agglomeration. Specifically, we define R_t as *effective research effort*, which is S_t adjusted for agglomeration effects in the cities where research takes place. The level of R_t will differ between the base and alternative cases because of limitations on city size. The second change we make is to allow for a time-varying term η_t, in the equation that relates effective research effort and technological progress. This change allows us to solve for an implicit level of effective research effort from the observed time series for productivity.

Making these two changes, and switching to discrete time, equation (22) becomes

$$\Delta \bar{A}_t / \bar{A}_t = \alpha \eta_t R_t^{\lambda} \bar{A}_t^{-\beta}, \tag{23}$$

and manipulating slightly,

$$\bar{A}_t^{\text{base}} = \bar{A}_{t-1}^{\text{base}} + \left[\alpha \eta_{t-1} (R_{t-1}^{\text{base}})^{\lambda}\right] (\bar{A}_{t-1}^{\text{base}})^{1-\beta} \tag{24}$$

Using this equation, and data on $\bar{A}$ by decade, we solve for a series for $\alpha \eta_t (R_t^{base})^{\lambda}$.

To construct the counterfactual series for effective research effort in the case where city sizes were restricted, R_t^{alt}, we use the machinery relating patents to city sizes from Section 4. Specifically, we use the ratio of P^{alt} / P^{base} for every decade that was presented in Figure 4. In mapping from this change in patents to a change in effective research effort, we face two immediate obstacles. First, the relationship between research effort and patenting is not constant over time. For example, between 1910 and 1960, the pace of patent filings by US residents was roughly constant at around 30,000 per year, while the effective number of researchers increased by a factor of more than eight (Berkes, 2018, Bloom et al., 2020). Second, the relationship between patents, on the one hand, and technological progress, on the other, is not stable in either levels or growth rates. For example, the number of patent granted in the US increased more than five-fold between 1980 and 2020, having less than doubled in the previous four decades, but there was no corresponding jump in the rate of TFP growth (USPTO, 2024).

To overcome the first of these obstacles, we assume that the amount of effective research effort required to produce a patent is constant only within a decade, that is:

$$P_{it} = \mu_t R_{it}, \tag{25}$$

where μ_t describes the time varying proportionality between effective research effort and patents.

The second of these problems is overcome via the time-varying term η_t in equation (23) that relates effective research effort and technological progress.

Under these two assumptions,

$$\frac{\alpha\eta_t R_t^{alt}}{\alpha\eta_t R_t^{base}} = \left(\frac{P_t^{alt}}{P_t^{base}}\right)^{\lambda} \tag{26}$$

The resulting values of $\alpha\eta_t(R_t^{alt})^{\lambda}$ can then be accumulated forward using the difference equation,

$$\bar{A}_t^{\text{alt}} = \bar{A}_{t-1}^{\text{alt}} + \left[\alpha\eta_{t-1}(R_{t-1}^{\text{alt}})^{\lambda}\right](\bar{A}_{t-1}^{\text{alt}})^{1-\beta}. \tag{27}$$

This yields the desired path of $\overline{A}$ under the alternative case, and in particular its value in 2010, $\bar{A}_{2010}^{alt}$.

Implementing this process requires values for the two key parameters describing the process of technological progress in equation (23). In their baseline case, Bloom et al. (2020) assume $\lambda = 1$, so that the stepping on toes effect does not operate. Under this assumption, they estimate that $\beta = 3.1$ As an alternative they consider $\lambda = 3/4$. In this case, they estimate that $\beta = 2.4$ In the analysis that follows, we use both pairs of parameterizations. In addition, we also consider a "naive" parameterization of $\lambda = 1, \beta = 0$, where both the stepping on toes effect and the negative effect of current technology on the ease of finding new technologies are absent. Our measure of the level of technology in the base case, $\bar{A}^{base}$ is constructed based on estimates of decadal TFP growth from Bloom et al. (2020) and Gordon (2016).[10]

Two issues regarding our calculations require further discussion. First is that the estimates of TFP growth used by Bloom et al. and Gordon, reflect not only technological progress, which is what $\bar{A}$ is supposed to be, but also the static urban scale effect discussed in Section 3. In the presence of static urban scale effects, the growth of cities will cause a standard measure of TFP to overstate the speed of technological progress. We elaborate in Appendix A. Under mildly restrictive assumptions we construct a series for technological progress that corrects observed TFP for the growth rate of urban scale effects. This correction is small for the period 1900-2010, in the range of one-tenth of a percent per year. We could easily use this adjusted series as our measure of $\bar{A}_t$. However, to simplify our exposition and to maintain comparability with other literature, we instead equate observed TFP growth with technological progress.

The second issue is how to think about technological progress in the rest of the world. In practice, ideas cross borders easily, and so if a reduction in city size had led to less new technology being invented in the US, this deficiency would have largely been made up

[10]Specifically the data are taken from the file AggregateBLSIPP.m in the replication package of Bloom et al. (2020). Their data for the period after 1950 comes from the Bureau of Labor Statistics Private Business Sector multifactor productivity growth series, adding back in the contributions from R&D and IPP. Their data for prior to 1950 come from Gordon (2016)

by innovation from abroad. To address this issue we could impose the same restrictions on global city sizes as we impose in the US. Using global data on city populations and research output, we could then perform an analysis of the effect of this size restriction. Unfortunately, data on city research effort at the global level are not available. As the best practical alternative to an analysis of the entire world, we assume that technological progress in the US results entirely from US research effort.

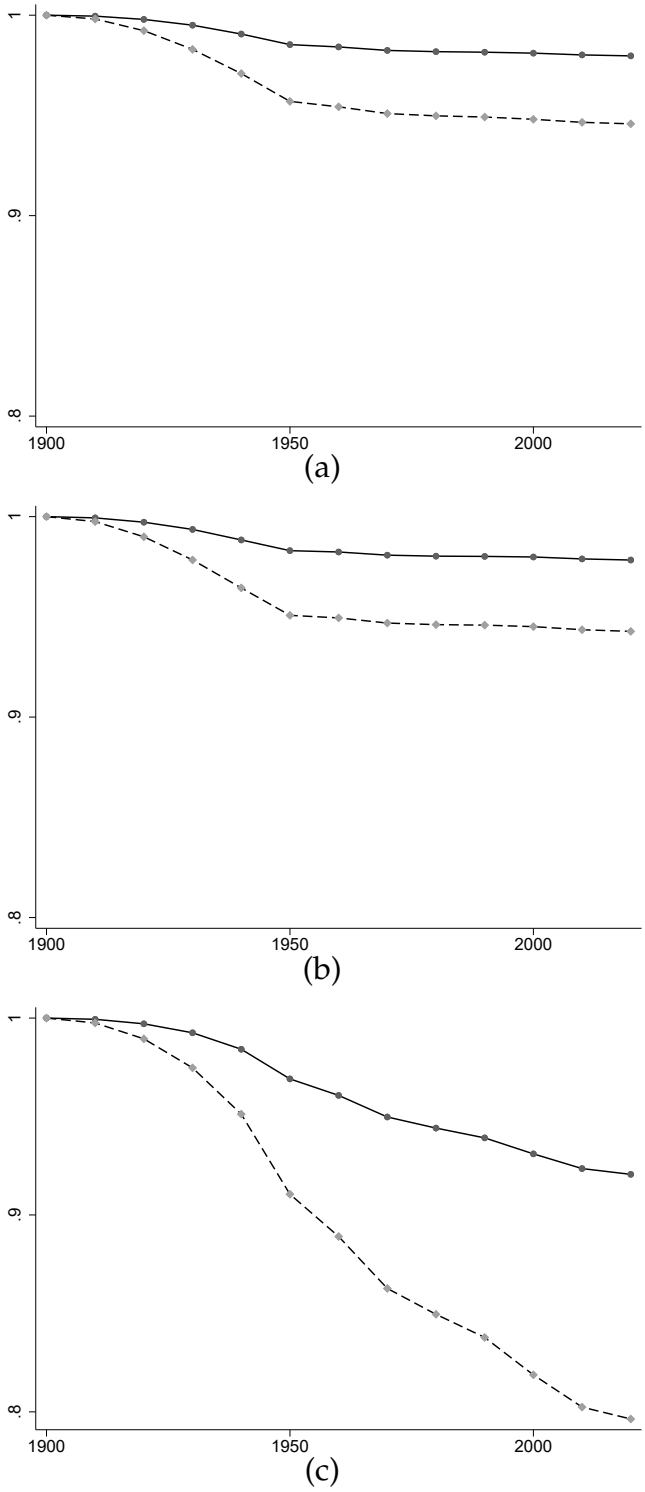
Figure 5 shows our results. Each panel reports the ratio of $\bar{A}_t^{\text{alt}} / \bar{A}_t^{\text{base}}$ in the counterfactual cases where city populations are restricted to one million or 100,000. The three panels differ from each other in parameters that govern the relationship between effective research effort and TFP. Figure 5(a) and (b) use the two sets of parameters considered by Bloom et al. (2020), $(\lambda = 1, \beta = 3.1)$ and $(\lambda = .75, \beta = 2.4)$. Figure 5(c) we considers a "naive" parameterization of $\lambda = 1, \beta = 0$, where the stepping on toes effect and the negative effect of current technology on the ease of finding new technologies are both absent.

As the figure shows, the cumulative effect of reduced research output has a small effect on the level of productivity for 2010 under either of the parameterizations used by Bloom et al. (2020). If we restrict attention to the empirically relevant cases where $\lambda = 0.75, \beta = 2.4$ or $\lambda = 1, \beta = 3.1$, then when $\sigma_B = 0.06$ and city size is limited to one million, $\bar{A}$ in 2020 is only about two percent lower than the baseline in both alternative cases. When we restrict city size to 100,000, the impact is still less than 6%. This seems somewhat puzzling, given that Figure 4 shows that research output in the counterfactual cases is between 5% and 20% lower than in the base case, depending on which scenario one is looking at, for all of the decades of the twentieth century.

The resolution to this puzzle is exactly the negative effect of the technology level $\bar{A}$ on the speed of technological progress. Less effective research effort early in the century leads to a lower level of $\bar{A}$, which in turn makes research later in the century more productive at generating technological progress than in the base case. Consistent with this intuition, the bottom panel of Figure 5 illustrates the results of a "naive" parameterization where $\lambda = 1$ and $\beta = 0$ and each successive unit of research effort yields the same increment to the level of technology. In this case, if city size is restricted to one million, $\bar{A}$ in 2020 is 8% below its baseline level, while if city size is restricted to 100,000, the reduction is about one fifth.

All of the calculations in Figure 5 are based on our preferred value of σ_B (the urban scale effect in R&D) of 6%. Table 4 shows the sensitivity of our results to the alternative value of 20% discussed above. We focus on the case where city size in the alternative scenario is limited to one million, and consider the same combinations of λ and β that

Figure 5: Counterfactual trajectories of national productivity.



Note: *Ratio of counterfactual to observed productivity, $\bar{A}_t^{alt}$ / $\bar{A}_t^{base}$, by decade for three different counterfactuals. City sizes are capped at 1m (squares) and 100k (diamonds). Panels differ in assumptions about the relationship between research output and productivity growth; (a)* $\lambda = 0.75, \beta = 2.4$*, (b)* $\lambda = 1, \beta = 3.1$*, (c)*$\lambda = 1, \beta = 0$*. We assume* $\sigma_B = 0.06$ *and* $\sigma_A = 0.08$*.*

Table 4: $\bar{A}_{alt}/\bar{A}_{base}$ for L_{max} = 1,000,000

Parameters	$\sigma_B = .06$	$\sigma_B = .20$
$\lambda = 1$ and $\beta = 3.1$	0.979	0.935
$\lambda = .75$ and $\beta = 2.4$	0.980	0.939
$\lambda = 1$ and $\beta = 0$	0.924	0.790

Note: *Each cell reports* $\bar{A}_t^{alt}/\bar{A}_t^{base}$ *for 2010 when maximum city size is limited to one million and* $\sigma_A = 0.8$.

we consider in Figure 5. To the extent that this effect is relatively small under our baseline parameterization, it would take a very large adjustment of σ_B to produce a large negative effect on productivity.

6. Combining Static and Dynamic Effects

We can now consider the combined effects of reduced aggregate productivity due to slower technological progress ($\bar{A}_t$) and lower static productivity from urban scale effects (the $\widetilde{A}_{it}$s) that would result from a limitation on city sizes. Combining equations (1), (2), (10) and (27) we can write the total change in counterfactual output due to both the types of urban scale effects,

$$\frac{Y_t^{alt}}{Y_t^{base}} = \left(\frac{\bar{A}_t^{alt}}{\bar{A}_t^{base}}\right)^{1/(1-\gamma)} \sum_i \frac{Y_{it}^{base}}{Y_t^{base}} min\left(1, \left(\frac{L_{max}}{\widetilde{L}_{it}}\right)^{\frac{\sigma_A}{1-\gamma}+\sigma_H}\right) \tag{28}$$

Performing the calculation described by equation (28), Table 5 shows output in the alternative case where city size is limited to one million relative to the base case of actual city sizes. We show values for all of the combinations of parameters that we consider above.

For the base-case set of parameters, i.e. $\frac{\sigma_A}{1-\gamma} + \sigma_H = 0.04$, $\sigma_B = .06$, and Bloom *et al.*'s preferred values, output in the alternative case is 8% lower if city sizes are restricted to on million. If we pick values for scale effect parameters at the upper end of the plausible range, $\frac{\sigma_A}{1-\gamma} + \sigma_H = 0.12$ and $\sigma_B = 0.20$, then the reduction in output is 22%. Given the importance often assigned to large cities as drivers of economic growth, these strike us as relatively small effects. For example, on the basis of PPP data from 2022, Canada's per capita GDP is 20% below the US, about the same as in the counter factual case just considered.

The results in Table 5 can also be transformed to describe the growth rate of output rather than its level. Recall that the experiment we are consider imposes a cap on city sizes starting in 1900 – the point in time when our baseline and alternative scenarios diverge. Using data from the Maddison Project, GDP per capita in the United States

Table 5: Output Relative to Baseline 2010

	$\sigma_B = 0.06$			$\sigma_B = 0.20$		
$\frac{\sigma_A}{1-\gamma} + \sigma_H$	0.04	0.08	0.12	0.04	0.08	0.12
$\lambda = 1.00$, $\beta = 3.1$	0.917	0.863	0.816	0.877	0.825	0.780
$\lambda = 0.75$, $\beta = 2.4$	0.919	0.865	0.817	0.880	0.829	0.783
$\lambda = 1$, $\beta = 0$	0.866	0.815	0.770	0.741	0.697	0.658

Note: *Counterfactual output as a share of realized output in 2010 when counterfactual city size is capped at 1m for different parameter values, Cells in this table are calculated by multiplying the appropriate entries of tables 2 and 4.*

increased by a factor of 6.1 between 1900 and 2010, corresponding to an annual growth rate of 1.66%. If output in the year 2010 had been 92% of its observed value, corresponding to our base-case parameters, then annual growth would have been 1.58%. Using the parameters representing the upper end of plausible scale effects, the annual growth rate would instead have been 1.43%.

Comparing the last line of Table 5 with the two above shows the importance of the fishing out effect in the Bloom *et al.* production function for technology. When we turn this effect off by setting $\beta = 0$, the decline in output is between 25% and 110% larger. (The relative importance of the fishing-out effect is largest when σ_A is small, so that static scale effect are not important and similarly when σ_B is large, so that scale effects on research productivity are large.)

What could be missing from our analysis that would make the effect of limiting urban scale larger? It is possible that we have simply used incorrect estimates of the scale effects that we include in our model. An alternative is that there are effects of urban scale, either static or dynamic, that we have failed to account for entirely. We can think of several such possibilities.

It may be that there are effects of large cities on human capital that go beyond the effect on wages of individuals who live in or move to those cities. For example, if having New York as a very large city makes everyone in the country have higher human capital, this would be an additional effect. Importantly, we *do* include in our analysis the channel by which New York being a big city affects production of new technology, as proxied by patents. So this would have to be a different type of knowledge production.

A second possibility is that big cities are important, not because of high population *per-se*, but because having a city be very populous allows a lot of people to take advantage of a particularly good location. This channel is ruled out in the approach we take. Specifically, we assume that the location-time specific component of productivity, $\hat{A}_{it}$ that each person experiences is the same in the baseline and alternative cases. Indeed, because of difficulties in measurement, we don't even try to estimate how $\hat{A}_{it}$ is related to city size. However, if one took this view, the conclusion would be that big cities are

important for growth only because the number of intrinsically good places to locate a city is limited, contradicting the spirit of much of the literature on agglomeration.

A related possibility is that the existence of large cities allows for higher mobility in response to transitory productivity shocks than would be present in the counterfactual case of more smaller cities. For example, if a particular location has a very good productivity shock, many more people can move there in a world of unrestricted city sizes than in a world where sizes are limited.

Finally, it is worth pointing out that our analysis is explicitly about the importance of *big* cities, rather than urbanization in general. Our alternative scenario maintains the same non-urban share of the population as is observed in the baseline.

7. Conclusion

To assess the effect of agglomeration economies on economic growth in the United States, we consider the effect of counterfactual restrictions on city size over the period from 1900 to 2010 on GDP per capita in the year 2010. We allow for both a static effect of city size on productivity and a dynamic effect of city size on research output, which then accumulates over time to determine the level of productive technology.

We conclude that the effects of restricting city size are surprisingly small – or put differently, that there is surprisingly little benefit from agglomeration. To give an example, consider the case in which city size is limited to one million people. In this case, holding the level of technology constant and using our standard set of parameters, we estimate that the loss of the static productivity effect reduces output in 2010 to 94% of its baseline level. In addition, over the 110 year period that we consider, the compounded effect of restricting city size is to reduce the level of technology by 2%, to 98% of its baseline level. Multiplying these effects, 2010 output in the case with limited agglomeration would have been 8% lower than the baseline, and the 1900-2010 GDP growth rate in the alternative scenario would have been less than a tenth of a percentage point lower than the baseline (i.e. 1.58% vs. 1.66%). While this is certainly not a trivial effect, it suggests to us that the urban scale effect was not the primary engine of economic growth.

As with any quantitative conclusion, there are many possible reasons why ours could be wrong. One possibility is that we have incorrectly parameterized either the scale effect of city size on productivity or the similar scale effect on research. But moving to the very highest end of the range of parameters estimated in the literature does not reverse our finding.

A second possibility is that there are effects of urban scale, either static or dynamic, that we have failed to account for. In the previous section we discussed some of these in detail.

A third possibility is that in examining our particular counterfactual, we have done violence to what people mean when they say that cities are engines of growth. Concretely, we assume the *only* economic effect of limiting city sizes would be via the urban scale effect. A skeptic might point out that if city sizes were limited, there would have to be more cities, and that some of these cities might not have the same fundamental productivity (the term we call $\widetilde{A}$) as the actually observed cities. This might be due to the new cities not being in locations that are as desirable as the cities that we actually observe. Our answer to this particular critique is that if it is correct, it is not so much cities themselves that are engines of economic growth, but rather good locations in which to put cities.

A final possibility is that we are being too narrow in our interpretation of the phrase "engine of growth." If urban scale effects explain one-tenth of US economic growth, maybe that qualifies them as being an engine of growth.

References

Arzaghi, M. and Henderson, J. V. (2008). Networking off Madison Avenue. *The Review of Economic Studies*, 75(4):1011–1038.

Atkin, D., Chen, M. K., and Popov, A. (2022). The returns to face-to-face interactions: Knowledge spillovers in silicon valley. Technical report, National Bureau of Economic Research.

Berkes, E. (2018). Comprehensive Universe of US patents (CUSP): Data and facts. *Unpublished, Ohio State University*.

Bloom, N., Jones, C. I., Van Reenen, J., and Webb, M. (2020). Are ideas getting harder to find? *American Economic Review*, 110(4):1104–1144.

Carlino, G. and Kerr, W. R. (2015). Agglomeration and innovation. *Handbook of regional and urban economics*, 5:349–404.

Carlino, G. A., Chatterjee, S., and Hunt, R. M. (2007). Urban density and the rate of invention. *Journal of urban economics*, 61(3):389–419.

Caselli, F. (2005). Accounting for cross-country income differences. In Aghion, P. and Durlauf, S. N., editors, *Handbook of Economic Growth*, volume 1 of *Handbook of Economic Growth*, pages 679–741. Elsevier.

Ciccone, A. and Hall, R. (1996). Productivity and the density of economic activity. *American Economic Review*, 86(1):54–70.

Combes, P.-P., Duranton, G., and Gobillon, L. (2008). Spatial wage disparities: Sorting matters! *Journal of urban economics*, 63(2):723–742.

Combes, P.-P., Duranton, G., Gobillon, L., and Roux, S. (2010). Estimating agglomeration economies with history, geology, and worker effects. In *Agglomeration economics*, pages 15–66. University of Chicago Press.

Combes, P.-P. and Gobillon, L. (2015). The empirics of agglomeration economies. In *Handbook of regional and urban economics*, volume 5, pages 247–348. Elsevier.

De la Roca, J. and Puga, D. (2017). Learning by working in big cities. *Review of Economic Studies*, 84(1):106–142.

Desmet, K. and Rossi-Hansberg, E. (2013). Urban accounting and welfare. *American Economic Review*, 103(6):2296–2327.

Duranton, G. and Puga, D. (2023). Urban growth and its aggregate implications. *Econometrica*, 91(6):2219–2259.

Fajgelbaum, P. D. and Gaubert, C. (2020). Optimal spatial policies, geography, and sorting. *The Quarterly Journal of Economics*, 135(2):959–1036.

Feenstra, R. C., Inklaar, R., and Timmer, M. P. (2015). The next generation of the Penn World Table. *American Economic Review*, 105(10):3150–82.

Forstall, R. and NBER (1995). US Decennial County Population Data, 1900-1990. Technical report, National Bureau of Economic Research. Accessed, January 2, 2024, https://www.nber.org/research/data/census-us-decennial-county-population-data-1900-1990.

Glaeser, E. L. and Gottlieb, J. D. (2008). The economics of place-making policies. Technical report, National Bureau of Economic Research.

Glaeser, E. L. and Gottlieb, J. D. (2009). The wealth of cities: Agglomeration economies and spatial equilibrium in the united states. *Journal of economic literature*, 47(4):983–1028.

Glaeser, E. L. and Maré, D. C. (2001). Cities and skills. *Journal of labor economics*, 19(2):316–342.

Gordon, R. (2016). *The rise and fall of American growth: The US standard of living since the civil war*. Princeton university press.

Hall, R. E. and Jones, C. I. (1999). Why do some countries produce so much more output per worker than others? *The Quarterly Journal of Economics*, 114(1):83–116.

Hsieh, C.-T. and Moretti, E. (2019). Housing constraints and spatial misallocation. *American economic journal: macroeconomics*, 11(2):1–39.

Hulten, C. R. (2010). Growth accounting. In Hall, B. H. and Rosenberg, N., editors, *Handbook of the Economics of Innovation, Volume 2*, volume 2 of *Handbook of the Economics of Innovation*, pages 987–1031. North-Holland.

Inoue, H., Nakajima, K., and Saito, Y. U. (2019). Localization of collaborations in knowledge creation. *The Annals of Regional Science*, 62:119–140.

Jones, C. I. and Vollrath, D. (2020). Solow and balanced growth / revisiting accounting. In *Online Study Guide for Economic Growth, fourth edition*.

Kerr, W. R. and Kominers, S. D. (2015). Agglomerative forces and cluster shapes. *Review of Economics and Statistics*, 97(4):877–899.

Klenow, P. J. and Rodriguez-Clare, A. (1997). The neoclassical revival in growth economics: Has it gone too far? *NBER macroeconomics annual*, 12:73–103.

Moretti, E. (2021). The effect of high-tech clusters on the productivity of top inventors. *American Economic Review*, 111(10):3328–3375.

Romer, D. (2012). *Advanced Macroeconomics*. McGraw-Hill Irwin, New York.

Rosenthal, S. S. and Strange, W. C. (2003). Geography, industrial organization, and agglomeration. *Review of Economics and Statistics*, 85(2):377–393.

Rosenthal, S. S. and Strange, W. C. (2004). Evidence on the nature and sources of agglomeration economies. In *Handbook of regional and urban economics*, volume 4, pages 2119–2171. Elsevier.

Solow, R. M. (1957). Technical change and the aggregate production function. *The Review of Economics and Statistics*, 39(3):312–320.

USDOC/BEA/RD (2023). Gross domestic product (gdp) by county and metropolitan area. Technical report. Accessed, January 2, 2024, `https://www.bea.gov/sites/default/files/2023-12/lagdp1223.xlsx`.

USPTO (2024). US Patent Activity Calendar Years 1790 to the Present. Technical report. Accessed, Nov. 7, 2024, https://www.uspto.gov/web/offices/ac/ido/oeip/taf/h_counts.htm.

Appendix A. Adjusting the Level of Technology for Urban Scale Economies in Output

This appendix pursues the issue of how taking urban scale effects into account changes the measurement of the speed of technological progress. Literature in this area generally proceeds by via the technique of growth accounting, whereby changes in output expected due to observed factor accumulation are subtracted from the observed series of changes in output to derive growth of total factor productivity (TFP). Growth of TFP is in turn assumed to be the same as technological progress. While there are many reasons why there might be slippage between these two concepts, we focus on the effect specific to our paper: If cities are getting larger over time, then urban scale effects will raise measured TFP even in the absence of technological change; or more generally, TFP growth will overstate technological progress. Using the machinery of Section 3, we can correct this problem.

We begin by rewriting the production function, explicitly including the determinants of productivity. For convenience, we also define human capital used in production at the city level as $H_{it} = h_{it}^{Y} \ell_{it}^{Y} L_{it}$:

$$Y_{it} = \widehat{A}_{it} \bar{A}_t \widetilde{A}_{it} K_{it}^{\gamma} H_{it}^{1-\gamma} \tag{A1}$$

Incorporating urban scale effects, the production function can be rewritten as

$$Y_{it} = \widehat{A}_{it} \bar{A}_t \widetilde{L}_{it}^{\sigma_A} K_{it}^{\gamma} H_{it}^{1-\gamma} \tag{A2}$$

We assume that the allocation of factors across cities satisfies

$$(K_i,H_i,L_i) = \alpha_i(\overline{K},\overline{H},\overline{L}), \text{ for all } i \text{ and } \alpha_i > 0 \tag{A3}$$

It follows immediately that $\Sigma_i \alpha_i = n+1$ and that $\alpha_i = L_i/\overline{L}$.

We can thus rewrite city output yet again as

$$Y_{it} = \widehat{A}_{it} \bar{A}_t \widetilde{L}_{it}^{\sigma_A} \frac{L_{it}}{\overline{L}_t} \overline{K}^{\gamma} \overline{H}^{1-\gamma} \tag{A4}$$

Next, let A^* be traditionally measured TFP. The conventional definition for A^* is

$$A_t^* = \frac{\sum_i \widehat{A}_{it} \bar{A}_t \widetilde{L}_{it}^{\sigma_A} \frac{L_{it}}{\overline{L}} \overline{K}^{\gamma} \overline{H}^{1-\gamma}}{(n+1)\overline{K}^{\gamma} \overline{H}^{1-\gamma}} = \frac{\bar{A}_t}{(n+1)\overline{L}_t} \sum_i \widehat{A}_{it} \bar{A}_t \widetilde{L}_{it}^{\sigma_A} L_{it} \tag{A5}$$

This can be rewritten (recalling that $\sum_i \hat{A}_i = n+1$ requires that $\overline{\widehat{A}_i} = 1$):,

$$A_t^* = \bar{A}_t \frac{Cov(\hat{A}_{it}, \widetilde{L}_{it}^{\sigma_A} L_{it}) + \overline{\widetilde{L}_{it}^{\sigma_A} L_{it}}}{\overline{L}_t} \tag{A6}$$

This can be rewritten as

$$A_t^* = \bar{A}_t \frac{\overline{\widetilde{L}_{it}^{\sigma_A} L_{it}}}{\bar{L}_t} \left[1 + Corr(\hat{A}_{i,t}, \widetilde{L}_{it}^{\sigma_A} L_{it}) CV(\hat{A}_{it}) CV(\widetilde{L}_{it}^{\sigma_A} L_{it})\right] \tag{A7}$$

where CV indicates coefficient of variation. The second term in brackets represents the effect on aggregate productivity of the relationship between city size and the component of city productivity that is unaffected by size. It is thus almost certainly positive. Unfortunately, because we can't measure city level productivity $\hat{A}_i$, we have no way of evaluating this term or its changes over time. In calculating technological change, we make the assumption that this term is constant. We can thus derive a simple expression for the relationship between TFP growth and aggregate technological change:

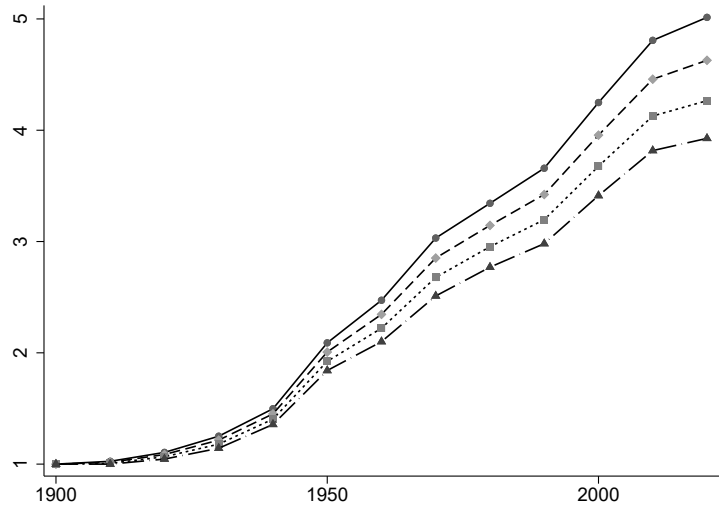
$$\Delta ln(\bar{A}_t) = \Delta ln(A_t^*) - \Delta ln\left(\frac{\overline{\widetilde{L}_{it}^{\sigma_A} L_{it}}}{\bar{L}_t}\right) \tag{A8}$$

Equation (A8) shows how to correct aggregate TFP growth to account for increased scale effects over time.

Figure 6 implements equation (A8). The highest, solid black line in the figure reports the uncorrected level of TFP, A^*. Each of the three bottom three lines reports a corrected level of technology based on a different estimate of the strength of urban scale economies on output, σ_A. Respectively, the small dashed line, dotted and large dashes describe $\bar{A}_t$ for $\sigma_A = 0.04, 0.08$, and 0.12.

Unsurprisingly, the gap between the adjusted and unadjusted level of technology grows over the course of the 20th century as cities become larger. By 2010, the adjusted level of technology is 93%, 86% and 79% of its unadjusted level when σ_a is $0.04, 0.08$, and 0.12. Alternatively, from 1900 to 2010, the growth rate of the unadjusted level of technology was 1.43% and the corresponding adjusted rates were 1.36%, 1.29%, and 1.22%. Note that these are *not* counterfactual calculations. These are adjustments to the growth rate that result from accounting for changes in the importance of urban scale effects for output.

Figure 6: TFP and Adjusted Technology for Different Urban Scale Effects for Output



Note: *Figure shows raw measure of the level of TFP, A_t^*, as the highest line. Three lines below report, from top to bottom, the adjusted level of technology, $\bar{A}_t$, when $\sigma_A = 0.04, 0.08,$ and 0.12.*