Are Big Cities Important for Economic Growth?*

Matthew A. Turner† and David N. Weil‡

May 2024

Abstract: Agglomeration is often described as an engine of economic growth. We quantitatively assess this statement, focusing in particular on urban scale economies in two dimensions: total factor productivity and the productivity of invention. The former is a static effect that makes production in bigger cities more efficient. The latter works dynamically, slowing the rate of productivity growth if there is less agglomeration. We use MSA-level patent and population data since 1900 to ask how much lower output would be in the US if agglomerations had been limited in size to populations of one million, one hundred thousand, or fifty thousand. Overall, we find that such limitations would have had a surprisingly small effects on output today.

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

*The authors are grateful to Enrico Berkes for generously sharing the CUSP patent database.

†Brown University, Department of Economics, Box B, Brown University, Providence, RI 02912. email: Matthew_Turner@Brown.edu. Also affiliated with PERC, IGC, NBER, PSTC, S4.

‡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 agglomerations. In 2010, the 27 cities in the top decile 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 quantitatively evaluate this idea. Our approach follows the analysis of the role of railroads in US economic growth of (Fogel, 1964). Prior to Fogel's work it had widely been noted that by the late 19th century, railroads were carrying the vast bulk of inter-regional trade, which was in turn a vital driver of growth. The natural conclusion was that railroads were thus a necessary contributor to that growth. Fogel's innovation was to note that even though railroads were in practice the dominant carrier or freight, in a world where railroads did not exist, it would have been possible for that same freight traffic to flow, only at higher cost – which he showed by constructing the transit 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 in which so much economic activity takes place today had not existed. This question cannot be answered simply by observing how much output is produced in cities or how much inventive activity takes place there. Had the 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 agglomeration effects in two specific dimension: total factor productivity and the productivity of invention. The former is a static effect that makes production in larger cities more efficient. The latter works dynamically, speeding productivity growth in larger cities. We begin with existing estimates of the magnitudes of these effects. Using a straightforward growth model, we consider counterfactual scenarios where the degree of agglomeration – 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.

Our approach follows the literature on growth accounting that began with Solow (1957). This approach takes as given growth in population, human capital, and physical capital as well as, in our case, the size distribution of cities. This contrasts with the full general equilibrium approach that is more common in the urban literature. The general equilibrium approach is more explicit about the drivers of the size distribution of cities, but requires strong assumptions. Our growth accounting approach requires remarkably

weak assumptions and allows an immediate mapping from data to results. For example, the analysis in Duranton and Puga (2019) is based on a general equilibrium model that features agglomeration effects on both labor productivity and human capital accumulation, an urban rent gradient, commuting costs, and politically-determined restrictions on development. Counterfactual city size distribution can then be generated by considering exogenous alternations to city planning restrictions. However, in this paper, the major driver of long-run growth, which is technological change, is completely exogenous. By contrast, our analysis considers the effect of the city size distribution and technological change.

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 agglomeration and applies it to study the effect of agglomeration statically on total factor productivity in the cross section of cities. We specifically consider counterfactual scenarios in which agglomerations in the US are limited in population to one million, one hundred thousand, or fifty thousand individuals. Section 4 then takes the same approach to study technological progress, specifically using data on MSA-level patents to assess the impact of agglomeration 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 that would have resulted from limitations on city sizes. Section 6 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 (2019), along with a single rural area that aggregates all non-metropolitan counties. We construct decadal population data by combining population data in replication files from Duranton and Puga (2019)

with 1900-1990 county population data from Forstall and NBER (1995). This results in an MSA by decade panel of MSA population stretching from 1850 to 2010, with 1860-1890 missing.[1]

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.[2] 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 1850 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. Non-metropolitan counties accounted for 19% of population, 14% of output and 6% of patents. Large cities have higher per-capita output and patent production than small cities. Non-metropolitan counties are less productive than cities.

Figures 1(b) and 1(c) repeat the analysis of Figure 1(a) for the years 1900 and 1950. Because the BEA output data is not available until 2000, we impute 1900 and 1950 output levels from 2000 output and contemporaneous population using equation (13) below. 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. Panel (a) plots the relationship between the log of output and the log of population for 2010. The tight linear relationship of the logs implies an elasticity of output per capita with respect to population of 8%. This is slightly larger than the 13% elasticity reported by Glaeser and Gottlieb (2009) for the same regression using data for 2000 and slightly different MSA definitions.

Panels (b-d) describe the relationship between the log of MSA patents and log population for 2010, 1950 and 1900. Three features of these plots seem noteworthy. First, the relationship between patenting and city size in 2010 is much noisier than it is for output.

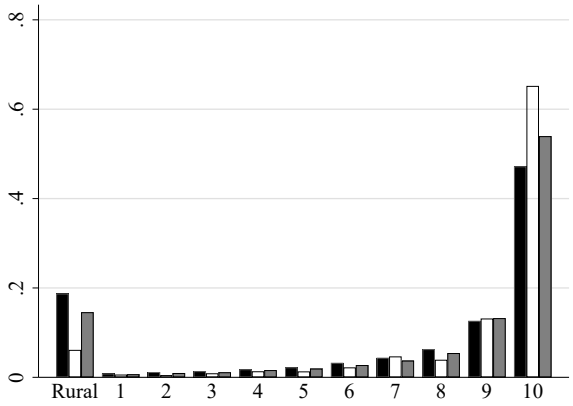
[1]Our sample of MSAs decreases slightly in the early part of our sample. This largely reflects the fact that some had not yet joined the union and so census data and county boundaries do not exist. For example, Arizona, New Mexico and Oklahoma all joined the US after 1900.

[2]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.

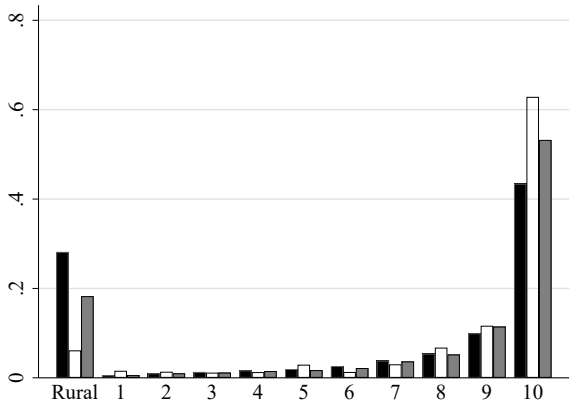
Second, the relationship between city size and patenting is much steeper than for output. The slope of the patents versus population regression line in 2010 is 1.45 versus 1.08 for output in panel (a). Third, the relationship between patenting and city size becomes much flatter as we go back in time. The slope of the regression lines in 1950 and 1900 are 1.35 and 1.11. versus 1.45 in 2010.

Finally panel (d) plots the log of patents against the log of output. In light of results so far, the fact that that slope of the best-fit line is greater than one is unsurprising. More interesting is that the relationship between output and patents is quite noisy. MSAs that produce more output tend to produce more patents, but there are also MSAs that are quite specialized in one or the other.

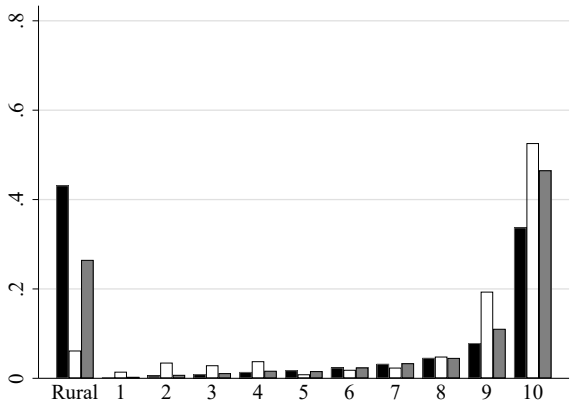
Figure 1: Distribution of population, output and patents by city size in 1900, 1950 and 2010



(a) 2010



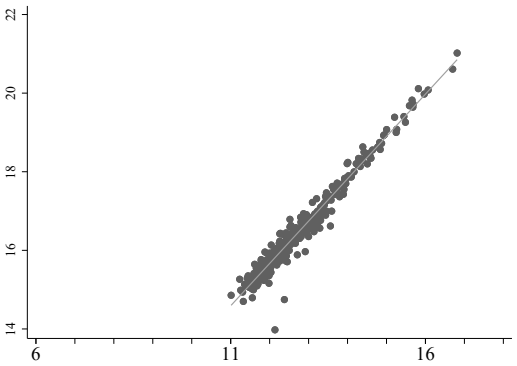
(b) 1950



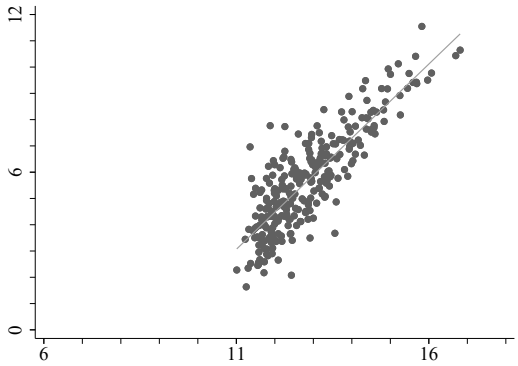
(c) 1900

Note: (a) Share of population, patents and total output by deciles of city size and rural status for 2010. (b) Same as (a) but for 1950 and population is imputed using the logic described in equation 13. (c) Same as (b) but for 1900. In each panel, black bar is population share, gray bar is output share, and white bar is patent share. Population, output and patents are all concentrated in the largest cities. Patenting is more concentrated than output, and output is more concentrated than population. Big cities become progressively more important over time.

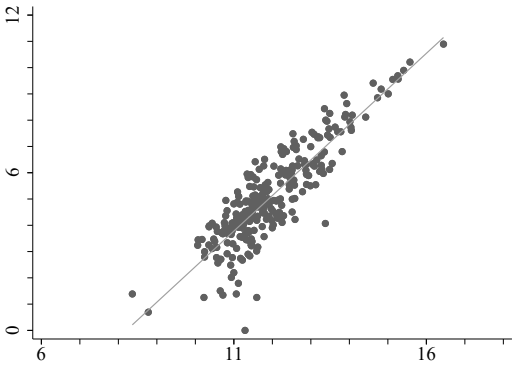
Figure 2: Joint distribution of output, patents and city size.



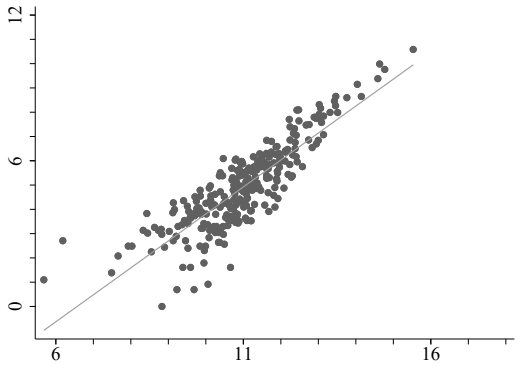
(a) Output vs, Pop. 2010



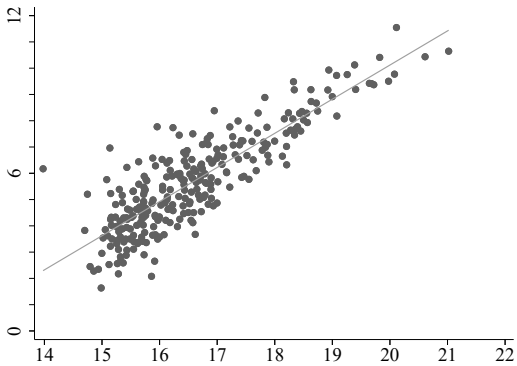
(b) Patents vs Pop. 2010



(c) Patents vs Pop. 1950



(d) Patents vs Pop. 1900



(e) Patents vs Output. 2010

Note: Each panel shows a scatter plot and OLS regression line. (a) ln(Output) vs. ln(Population) 2010; $\beta = 1.08, \text{s.e.} = 0.013$. (b) ln(Patents) vs. ln(Population) 2010; $\beta = 1.41, \text{s.e.} = 0.053$. (c) ln(Patents) vs. ln(Population) 1950; $\beta = 1.35, \text{s.e.} = 0.046$. (d) ln(Patents) vs. ln(Population) 1900; $\beta = 1.11, \text{s.e.} = 0.041$. (e) ln(Patents) vs. ln(Output) 2010; $\beta = 1.29, \text{s.e.} = 0.047$. Outputs and patents both increase with city size. The relationship between patents and population is noisier than between output and population, and becomes weaker as we go back in time. Patents and output are also strongly related, though some cities are quite specialized in output or patents.

3. Agglomeration and TFP in a Cross Section

We would like to calculate the total value of 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 $i = 1,...,N$. Y_{it} denotes output of city i in decade t, L_{it} population, K_{it} physical capital, ℓ_{it}^Y the fraction of the labor force 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 an MSA 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 important are changes in the sizes of cities for aggregate output. To proceed, we decompose A_{it} into three components: a time specific national component common to all cities, $\bar{A}_t$, a city specific agglomeration effect that depends on population, $\widetilde{A}_{it}$, and city-decade specific idiosyncratic term, $\widehat{A}_{it}$:

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

Finally, we assume agglomeration economies in the production of output depend on city population according to,

$$\widetilde{A}_{it} = L_{it}^{\sigma_A}. \tag{3}$$

This production technology nests those commonly used to study systems of cities, e.g., Desmet and Rossi-Hansberg (2013), Duranton and Puga (2019).

We assume that physical capital is freely mobile among cities, to equalize its marginal product. This implies that,

$$\frac{Y_{it}}{K_{it}} = \frac{Y_t}{K_t} \quad \text{for all } i \tag{4}$$

Substituting (4) into the production function (1) and rearranging, we have

$$Y_{it} = A_{it}^{1/(1-\gamma)} \left(\frac{K_t}{Y_t}\right)^{\gamma/(1-\gamma)} h_{it}^Y \ell_{it}^Y L_{it} \tag{5}$$

Summing over all cities, we get aggregate output,

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

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, is invariant across cases. There are two foundations for this assumption. First, if capital is accumulated with a fixed investment rate, as in Solow, the the capital to output ratio is constant along any balanced growth path.[3] Assuming a constant capital to output ratio is convenient for our analysis. However, the stronger assumption that our economy is on a balanced growth path and the saving rate is fixed also solves a conceptual problem. We expect a change in city level productivity to affect income and hence capital accumulation. This raises the possibility that a change in city size affects the level of output indirectly through the level of capital. The assumption that the capital labor ratio is fixed accounts for this possibility. There is also empirical evidence for a constant capital labor ratio

We restrict attention to alternative cases where a city's population is unchanged but the size of the urban scale effect on productivity ($\widetilde{A}$) is reduced to that of a smaller city. All other characteristics of the city, h_{it}, ℓ_{it}^{Y}, and the time-city specific dimension 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 also assume that non-metropolitan output does not change between realized and counterfactual cases.

Restricting attention to this particular class of counterfactuals is important for two reasons. First, as we will see, it is convenient for our analysis. Second, it relieves us of the problem of measuring h_{it}, ℓ_{it}^{Y}, and the time-city specific dimension of productivity, $\widehat{A}_{it}$, each of which poses difficult econometric problems.[4]

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 implicitly considering equilibrium population levels that are partly determined by an unspecified congestion process.

With these assumptions in place, we can use (6) to compare aggregate output in economies with different values of A. Multiplying each term in the sum on the right hand side of (6) by $\left(\frac{A_{it}^{base}}{A_{it}^{base}}\right)^{1/(1-\gamma)}$ we have

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

[3]See Romer (2012).

[4]Estimates city-specific productivity in producing output, 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 measure the quality of human capital? This problem will recur in our analysis of city level research productivity.

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

Dividing by Y_t^{base} gives

$$\frac{Y_t^{alt}}{Y_t^{base}} = \sum_i \frac{Y_{it}^{base}}{Y_t^{base}} \left(\frac{A_{it}^{alt}}{A_{it}^{base}} \right)^{1/(1-\gamma)} \tag{7}$$

Equation (7) is central to our analysis. It allows us to calculate the change in output relative to the observed base case, using only information on realized city level output, realized city level productivity, and counterfactual city level productivity.

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 (3), the ratio of observed to counterfactual city productivity is,

$$\frac{A_{it}^{alt}}{A_{it}^{base}} = min \left(1, \left(\frac{L_{max}}{L_{it}} \right)^{\sigma_A} \right). \tag{8}$$

Using equation (8) and (7) 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,

$$\frac{Y_t^{alt}}{Y_t^{base}} = \sum_i \frac{Y_{it}^{base}}{Y_t^{base}} min \left(1, \left(\frac{L_{max}}{L_{it}} \right)^{\sigma_A/(1-\gamma)} \right). \tag{9}$$

In equation (9) 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.

A *Human capital extension*

The fact that people are more productive 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 scale effects and 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 slope of log wage versus log city size by one third to one half, and the resulting residual slope is interpreted as the causal effect of city size on the level of productivity.

More recent research (De la Roca and Puga (2017) and 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 variance in the rate of worker productivity growth as city size varies.

This has two implications. First, the relationship between output and city size that we see in the raw data, e.g., figure 2, is actually close to the causal effect. Second, that the relationship between city size and the level of productivity operates through two channels, a productivity effect like the one we describe above, and a human capital production effect. We here extend our model to include the second 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 :

$$h_{it}^{Y} = exp(\phi_{it} S_{it}) L_{it}^{\sigma_h}, \tag{10}$$

Next, we use equation (6) to write aggregate output for a counterfactual case where city size is capped at L_{max} and multiply the right hand side by

$$\left(\frac{A_{it}^{base}}{A_{it}^{base}}\right)^{1/(1-\gamma)} \frac{h_{it}^{Y,base}}{h_{it}^{Y,base}}. \tag{11}$$

Following the same logic that leads from equation (6) to (9), we arrive at the corresponding expression for aggregate output when human capital production is subject to scale effects,

$$\frac{Y_t^{alt}}{Y_t^{base}} = \sum_i \frac{Y_{it}^{base}}{Y_t^{base}} min\left(1, \left(\frac{L_{max}}{L_{it}}\right)^{\frac{\sigma_A}{1-\gamma}+\sigma_h}\right). \tag{12}$$

Comparing equation (12) to equation (9), we see that the two expressions are identical save for the interpretation and magnitude of the exponent on the term $\left(\frac{L_{max}}{L_{it}}\right)$. 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.

B *Parameterization*

Throughout, we set the capital share of output, $\gamma = 0.33$, which is standard in the growth literature.

A key parameter in this analysis is σ_A, which measures agglomeration benefits in productivity. To estimate σ_A one must distinguish between the quantity of interest, pure

effect of scale, and; 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 productive. The literature has proposed a variety of solutions to these problems (see Rosenthal and Strange (2004) and Combes and Gobillon (2015) for surveys). Regardless of technique and setting, estimates usually fall between 3% and 10%. Table 1 lists six prominent, relevant estimates.

It is common to estimate agglomeration economies from the relationship between wage and city size or density; wage data is easily available and is easily linked to measures of human capital.[5] 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 on average a worker's wage increases by about 5.1% when city size doubles. After controlling for individual fixed effects, this drops to about 3.7%, suggesting that $\sigma_A = 3.7\%$ with the 1.2% difference due the sorting of more productive workers into larger cities. Using Spanish data, De la Roca and Puga (2017) conduct a similar exercise and finds that on average a worker's wage increases by about 5.1% when city size doubles. 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. Duranton and Puga (2023) replicate De la Roca and Puga (2017) for the panel of US workers described by the NLSY79 and find that on average a worker's wage increases by 7.6% when city size doubles. 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.

While the individual level data employed in Combes et al. (2008), De la Roca and Puga (2017) and Duranton and Puga (2023) allows the state of the art decomposition of scale effects into human capital/sorting and pure scale effects, they are based on French, Spanish and the highly selected NLSY sample of US workers. Like this paper, the other three papers in table 1 are based on representative samples of US data. Glaeser and Gottlieb (2008) looks at the relationship between wages and city size using a large cross-section of US workers. They estimate $\sigma_A = 4.1\%$. While they cannot control for individual fixed effects in their cross-sectional data, they experiment widely with econometric technique and find little variation in this estimate. Ciccone and Hall (1996) is an early effort to estimate σ_A and does so indirectly using US state (not city) level data on output and population density. They estimate that $\sigma_A = 5.2\%$. Glaeser and Gottlieb (2009) estimates the relationship between city level output and population using data on

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

Table 1: Estimates of σ_A

σ_A	σ_h	Source	Data
3.7%	1.2%	Combes et al. (2008)	French, Ind. wages, 1976-98
2.2%	2.9%	De la Roca and Puga (2017)	Spanish, Ind. wages, 2004-9
4.5%	3.1%	Duranton and Puga (2023)	US, Ind. wages, ca. 1979-2020
4.1%	.	Glaeser and Gottlieb (2008)	US, Ind. wages, 2000
5.2%	.	Ciccone and Hall (1996)	US, State output, 1988
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.

US cities in 2000, and finds that output increases by 13% when city size doubles. This estimate does not correct for the possibility of sorting or more rapid urban human capital accumulation in cities.

We consider three values of σ_A in our calculations, 4%, 8% and 12%. These values approximately bound the estimates reported in table 1. Our preferred estimate for the pure scale effect is $\sigma_A = 4\%$. To allow for the possibility of more rapid urban human capital accumulation, we (like Duranton and Puga (2023)) consider $\sigma_A = 8\%$ for our baseline estimates. To check robustness, we also consider $\sigma_A = 12\%$, which we regard as close to the largest defensible estimate of this parameter.

To evaluate equation (9) we use the data on city level output and population described above.

We first evaluate the static effect of agglomeration for data from the year 2010. In addition to our baseline value of $\sigma_A = 0.08$, we also consider $\sigma_A = 0.04$ and 0.12. We also consider three possible values of maximum city size, $L_{\max}$: 1,000,000, 100,000, and 50,000. 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. A cap of 100,000 would require reorganizing 261 MSAs,while the cap of 50,000 affects every MSA.

Table 2 presents results. Rows report the value of equation (9), the ratio of counterfactual to realized aggregate output, as the strength of static agglomeration economies increase. Columns describe different counterfactual systems of cities. Moving from column 1 to 3, we consider systems of cities in which cities are constrained to be smaller and agglomeration economies are less important. We see that counterfactual output is 94% of realized output when cities are allowed to be as large as 1m and agglomeration economies take their smallest value. This share declines to 88% when the strength of agglomeration economies is largest. It is only when we consider the large values of σ_A

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

σ_A	$L_{max} = 1m$	$L_{max} = 100k$	$L_{max} = 50k$
0.04	0.94	0.84	0.82
0.08	0.88	0.72	0.68
0.12	0.83	0.62	0.57

Note: Each cell reports the share of total output relative totals reported in the 2010 BEA data, for a particular cap on city size and value of σ_A. For the purpose of this calculation,the rural population is treated as an extra MSA whose output is constant across scenarios and capital share of output, γ, is equal to 0.33.

and restrict cities to be no larger than 50 or 100k that we begin to see 30 to 40% declines in output. With this said, if we consider has increased by about a factor of 13 since 1900, and rely on central estimates of σ_A that are no larger than 8%, then it seems hard to conclude that agglomeration economies are more than a moderately important contributor to the overall increase in output over this period.

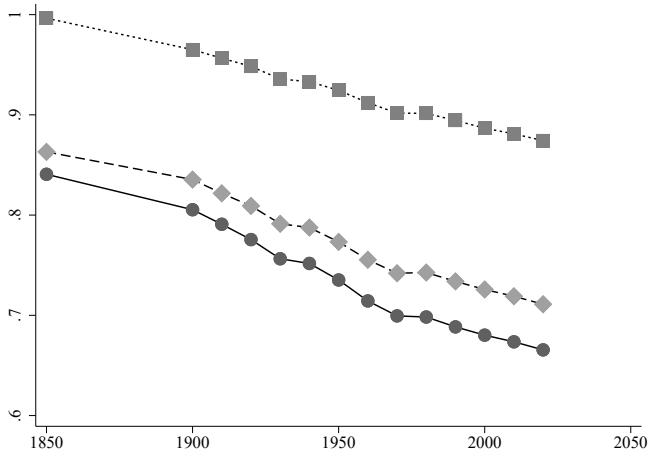
These results can be extended historically, although with some difficulty. The BEA data on MSA output begins in to 2000 and so we must impute historical values. To do this, we assume all quantities in the production function take their 2000 values except population, which takes its realized value for each decade. In this case, we can impute each city's output as

$$Y_{it}^{base,2000} \equiv= Y_{it}^{base} \left(\frac{L_{it}}{L_{i2000}} \right)^{1-\gamma+\sigma_A}. \tag{13}$$

We perform this calculation using the BEA data for 2000 and our population data from 1850 to 2020, and figure 1 reports imputed output levels for 1950 and 1900 when $\gamma = 0.33$ and $\sigma_A = 0.08$.

Given imputed levels of output from equation (13), we can calculate the hypothetical output that would result from imposing a cap on city size using equation (9), just as we did for measured 2010 output data in table 2. Figure 3 presents these results. Recalling that restricting populations to 1m requires that we reorganize the largest 52 cities in the country, it is only with catastrophic reorganizations of the economy that we see large effects on output in Table 2 and Figure 3. If we allow cities as large as 1m and a central estimate of σ_A of 8%, the effects on output are about 6%, about half again as large as the 2008 financial crisis. Even our most extreme counterfactual, where we reduce all cities to 50k, and consider σ_A at the upper end of plausible estimates, we reduce output by about 40%, not quite as large as the current gap between the US and France.

Figure 3: Share of total output produced under three hypothetical city networks, imputed baseline output



Note: Counterfactual output as a fraction of actual outcome when city sizes are capped at 1m (squares), 100k (diamonds), and 50k (circle). Calculations are based on city output imputed from the BEA data for 2000 and population data, and assume $\sigma_A = 0.08$ and $\gamma = 0.33$.

4. Productivity Growth

The analysis of the previous section takes the city invariant component of TFP, $\bar{A}_t$ as given. We now turn to this parameter, and in particular to the effect of agglomeration on the speed of technological progress. This channel obviously has the potential to produce a much larger effect on current output than the channel of static productivity from agglomeration, because improvements in technology accumulate over time.

We proceed in parallel with our approach in the previous section, although as will be seen we have to make adjustments to deal with the dynamics of technological change.

Define R_t as research output at time t. The use of this new terminology is required because research output will not map directly into the speed of technological progress. Specifically, as will be seen in Section 4, the speed of technological progress depends on both research output and the level of technology itself. However, at a given point in time, we assume that cross-sectional variation in research output among cities will produce proportional variation in patents per city. Specifically, defining P_{it} as patents, we assume that[6]

[6]Bloom et al. (2020) stress that the relationship between patents, on the one hand, and productive ideas, on the other, is unlikely to be stable over time. Among the reasons for this are changes in what can be patented. In our setting, this instability would be reflected in a changing value of μ_t that relates patents to research output. In the application below, we do not assume that μ_t is constant. Rather, our only assumption is that cross-city variation in patents at a point in time is proportional to cross-city variation in research output.

$$P_{it} = \mu_t R_{it} \tag{14}$$

Research output in a city depends on the size of the research labor force, its human capital, h_{it}^R, and a city-decade research productivity multiplier, B_{it}, according to the function,[7]

$$R_{it} = B_{it} h_{it}^R (1 - \ell_{it}^Y) L_{it}. \tag{15}$$

Summing over cities within a year, we have aggregate research output,

$$R_t = \sum_{i=1}^{N} B_{it} h_{it}^R (1 - \ell_{it}^Y) L_{it}. \tag{16}$$

City-decade research 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,

$$B_{it} = \widehat{B}_{it} \bar{B}_t \widetilde{B}_{it}. \tag{17}$$

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

$$\widetilde{B}_{it} = L_{it}^{\sigma_B}. \tag{18}$$

We do not restrict the relationship between city-specific output productivity (the $\widehat{A}_{it}$s) and city-specific research productivity (the $\widehat{B}_{it}$s). Places can be good at one but not the other. We also do not restrict the relationship between the quality of human capital used to producing output, h_{it}^Y and that used in producing research, h_{it}^R. For example, two cities might have the same numbers of Ph.D.s working in production, while they have radically different numbers of Ph.D.s working in research.

As in the previous section, we consider the thought experiment of having 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 research output, we use the same argument that we used in our analysis of counterfactual output, adjusting for the fact that capital does not play a role in the production of research output.

This argument proceeds in four steps. First, use (16) 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 (17) and (18) to get

$$R_t^{alt} = \sum_i R_{it}^{base} min\left(1, \left(\frac{L_{max}}{L_{it}}\right)^{\sigma_B}\right). \tag{19}$$

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

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

$$\frac{R_t^{alt}}{R_t^{base}} = \sum_i \frac{R_{it}^{base}}{R_t^{base}} min\left(1, \left(\frac{L_{max}}{L_{it}}\right)^{\sigma_B}\right). \tag{20}$$

Recalling our assumption (in equation (14)) that observed patents are proportional to research output within each decade, we can evaluate equation (20) by replacing research output with patents .

A *Research output in a cross section*

We begin by considering the effect of limiting city sizes at a point in time, holding constant the time-specific component of research output, $\bar{B}$ constant. As our measure of research output, we use patents at the MSA level.

It is worth pointing out that this approach to examining the effect of limiting city sizes on research productivity represents something of an extreme case. To see why, consider the case in which there is a city of two million people, of whom 20,000 are engaged in R&D. Agglomeration effects in research presumably depend on the number of other researchers in a city, rather than the number of people overall. Thus 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, research output would not fall at all. By contrast, in dividing up the resources devoted to R&D proportionally with population, we have made the assumption that maximizes the effect of limiting agglomeration in research productivity.

The key parameter required for our calculation is σ_B, the effect of city size on research productivity. 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 the 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; 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.

These papers are less helpful for thinking about how scale effects vary with the size of a cluster or a city. Atkin et al. (2022a) takes a useful step in this direction. This paper

estimates the effect of face-to-face contact, measured by cellphone proximity, on patent citations. A back of the envelope calculation based on these estimates suggests that 25% reduction in workforce would cause a 17% decline in citations. Averaging, this gives an elasticity of about 0.8. This is a huge effect, but calculated indirectly and based on a sample of tech workers in Silicon Valley. 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 find that doubling employment density increases patents per person by 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 doubling the number of inventors in the same year-sector-cluster increases an inventors productivity by 5% to 9%, depending on specification.

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 $\sigma_B \in [0.17,0.20]$. Atkin et al. (2022b) suggests a still larger estimate, but is based on a highly selected sample where the effects would be expected to be large. For our baseline calculation, we consider the preferred estimate of $\sigma_B = 6\%$ from Moretti (2021), and also consider the much larger value $\sigma_B = 20\%$.

Table 3 shows national research output for 2020 in the alternative case where cities are limited in size relative to the base case of observed research output. 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, research output falls by only 7% relative to baseline. This magnitude is similar the static urban scale effect on output production shown in table 2 if we used a similar value of σ_A. We find this surprising. If, as we had expected, patenting were more concentrated in larger cities than output, then reducing the opportunity for agglomeration economies to operate by restricting city size would be more harmful to patenting than output. Larger declines in aggregate patenting are possible, but require the catastrophic counterfactual changes associated with L_{max} equal to 100k or 50k, 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 50k, the output of patents falls by 48%.

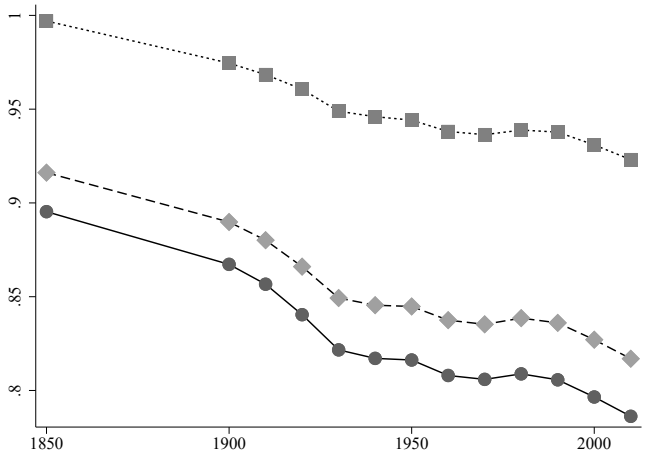
We can also examine the evolution of research output over time. For this purpose, we restrict attention to our baseline value of $\sigma_B = 0.06$ and calculate how patenting changes

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

σ_B	$L_{max} = 1m$	$L_{max} = 100k$	$L_{max} = 50k$
0.06	0.93	0.83	0.80
0.20	0.82	0.58	0.52

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.

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



Note: Counterfactual patents as a fraction of actual patents reported in CUSP when city sizes are capped at 1m (squares), 100k (diamonds), and 50k(circles). Calculations assume $\sigma_B = 0.06$.

in each of our three counterfactual systems of cities in each decade for which we have both population and patent data. Figure 4 presents our results. This figure corresponds closely to figure 3, which reports changes in aggregate output for different networks over time, except that figure 3 is based on imputed output levels for the observed system of cities, while figure 4 does require this imputation.

5. From Research Output to the Speed of Technological Progress

Our goal is to calculate how the speed of technological progress would differ if the US were less urbanized. An immediate issue that arises is how to think about the rest of the world. In practice, ideas easily cross borders, and so if there were less effective R&D taking place in the US, the decline in research productivity would to a large extent simply

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

Parameters	$\sigma_B = .06$	$\sigma_B = .20$
$\lambda = 1$ and $\beta = 3.1$	.979	.935
$\lambda = .75$ and $\beta = 2.4$	.980	.939
$\lambda = 1$ and $\beta = 0$	.924	.790

Note: Each cell reports the ratio of the time-specific component of aggregate productivity, $\bar{A}$, for the year 2010 in the case where maximum city size is limited to one million, relative to the base case in which city size is not limited.

be made up by a larger fraction of innovation taking place abroad. (maybe cite recent Bloom, Jones, etc. paper here).

One way of dealing with this issue would be to assume that the same restriction on agglomeration that we impose on the US was imposed on the whole world. 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 output at the global level is not available. As an alternative, we make the assumption technological progress in the US results only from R&D in the US. Another set of assumptions that would produce the same result would be that new technologies flow freely across borders and that the reduction in R&D input that take place in the rest of the world is of the same magnitude as that in the US. We view this as a reasonable approximation.

Bloom et al. (2020) empirically explore the relationship between productivity growth and aggregate R&D in the US over the period 1930-2015. Adapting their formulation of the research production function to our notation:[8]

$$\frac{dln(\bar{A})}{dt} = \alpha R_t^{\lambda} \bar{A}_t^{-\beta} \tag{21}$$

The parameter λ captures the "stepping on toes" effect, whereby a the speed of technological progress may not scale linearly with research output. The parameter β captures the extent to which ideas become harder to find as more of them have been discovered. They take as their baseline assumption $\lambda = 1$ (no stepping on toes effect), and under this assumption estimate that $\beta = 3.1$ As alternative they consider $\lambda = 3/4$, in which case they estimate that $\beta = 2.4$ In the analysis that follows, we use both pairs of parameterizations. (Along balanced growth paths, the ratio of the growth rate of productivity to the growth rate of research output is determined solely by the ratio of λ to β. However, we will be looking along transition paths where the both parameter values matter independently.

We want to derive the time path of the city-invariant component of productivity $\bar{A}_t$ under the assumption that city sizes were limited.

[8]In Bloom et al. (2020), the input into research is the number of researchers. We replace that with our measure of research output adjusted for the impact of urban scale effects.

Consider a set of observed values of $\bar{A}^{\text{base}}$ in the baseline case where city sizes where not restricted – that is, what actually happened. Call these $\bar{A}_1^{\text{base}}, \bar{A}_2^{\text{base}}, ...$ We want to derive an alternative pathway of this component of productivity, $\bar{A}^{\text{alt}}$ in the case where city sizes were restricted. These are $\bar{A}_1^{\text{alt}}, \bar{A}_2^{\text{alt}}, ...$ We assume that in period one, the level of $\bar{A}$ in the two scenarios were equal.

The time periods that we examine will be decades. We take the discrete version of the equation for technological progress, and further substitute our variable for city-invariant productivity, $\bar{A}$, as the measure of technology:

$$\Delta \bar{A}_t / \bar{A}_t = \alpha R_t^{\lambda} \bar{A}_t^{-\beta} \tag{22}$$

Rewriting this separately for the base and alternative cases,

$$\bar{A}_t^{\text{base}} = \bar{A}_{t-1}^{\text{base}} + \alpha (R_{t-1}^{\text{base}})^{\lambda} (\bar{A}_{t-1}^{\text{base}})^{1-\beta} \tag{23}$$

$$\bar{A}_t^{\text{alt}} = \bar{A}_{t-1}^{\text{alt}} + \alpha (R_{t-1}^{\text{alt}})^{\lambda} (\bar{A}_{t-1}^{\text{alt}})^{1-\beta} \tag{24}$$

To calculate the path of $\bar{A}$ in the alternative case, we proceed as follows. In the base case, given a series of values for $\bar{A}_t^{\text{base}}$ we can back out a series for R_t^{base}. Equation (20) then gives us the ratio of research output in the alternative case where city sizes are restricted relative to the base case where they are not. [9] This allows us to produce a series for R_t^{alt}. Under the assumption that $A_{base,0} = A_{alt,0}$, we can then generate a full time series for $\bar{A}$ in the alternative case relative to the base case, by forward iteration of equation (**??**).

Figure 5 shows the result of this calculation. In addition to the two sets of parameters considered by Bloom et al. (2020), $(\lambda = 1, \beta = 3.1)$ and $(\lambda = .75, \beta = 2.4)$, we also consider a "naive" parameterization of $\lambda = 1, \beta = 0$, which would imply that both the stepping on toes effect and the negative effect of current technology on the ease of finding new technologies are absent.

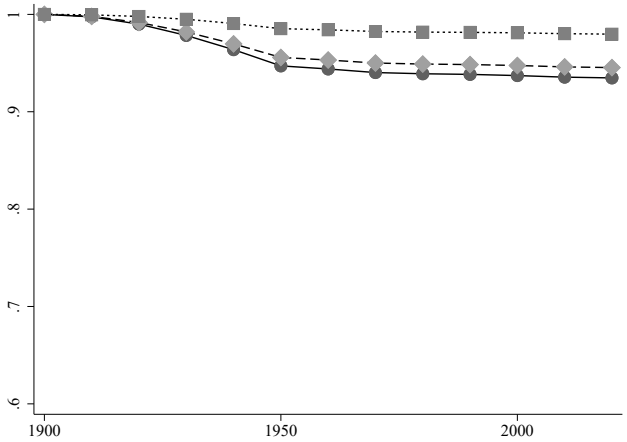
As the figure shows, the cumulative effect of reduced research output turns out to have a remarkably small effect on the level of productivity in the year 2020 under either of the parameterizations used by Bloom et al. (2020). When city size is limited to one million, productivity in the year 2020 is only two percent lower in the alternative case than in the baseline. Even when city size is limited to 50 thousand, the impact on the level of productivity is on the order of seven percent. This seems somewhat puzzling, given that Figure 4 shows that research output in the counterfactual cases is between 5%

[9] Formally, what we back out is the series for $\alpha(R_{t-1}^{\text{base}})^{\lambda}$ and what we then construct for the alternative case is the series for $\alpha(R_{t-1}^{\text{alt}})^{\lambda}$. Because we are interested only in the ratio of these two objects, the value of α is irrelevant.

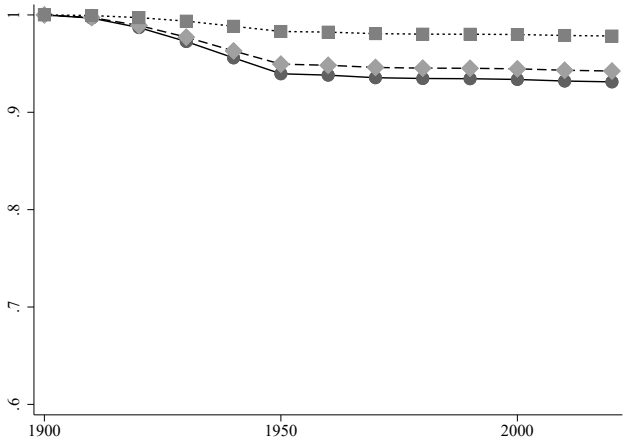
and 20% lower than in the base case, depending on which scenario one is looking at, for all of the decades of the twentieth century.

The resolution to this puzzle is exactly the negative effect of the technology level $\bar{A}$ on the speed of technological progress that is at the center of the model in Bloom et al. (2020). Less research output early in the century would have led to a lower level of $\bar{A}$, which would have in turn made research later in the century lead to faster technological progress that it did in the base case. This can be seen by examining the top panel of Figure 5 where we use the "naive" parameterization in which the effect just described is shut down. In this case, even if city size is restricted to one million, productivity in 2020 is 8% below its baseline level, while if city size is restricted to fifty thousand, the reduction in productivity in 2020 is roughly one quarter.

Figure 5: Counterfactual trajectories of national productivity.



(a)



(b)

(c)

Note: Counterfactual 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), 100k (diamonds), and 50k(circles). 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 $\sigma_B = 0.06$ throughout.

Table 4 shows the sensitivity of this result to value of σ_B, the parameter that measures the agglomeration effect in R&D. 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 were examined in Figure 5. As the table shows, the cumulative effect of limiting city size on productivity is roughly linear in the value of σ_B. To the extent that this effect is relatively small under our baseline parameterization, it would take a very large adjustment in the parameter to produce large negative effects 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 $\tilde{A}_{it}$s) that would result from a limitation on city sizes. Equation **??** puts together the results from the previous sections.

$$\frac{Y_t^{alt}}{Y_t^{base}} = \left(\frac{\bar{A}_t^{alt}}{\bar{A}_t^{base}}\right)^{1/(1-\gamma)} \sum_i \frac{Y_{it}^{base}}{Y_t^{base}} min\left(1, \left(\frac{L_{max}}{L_{it}}\right)^{\sigma_A/(1-\gamma)}\right) \tag{25}$$

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 were considered above.

For the base-case set of parameters, i.e. $\sigma_A = 0.08$ and $\sigma_B = .06$, and using Bloom *et al.*'s preferred values for technology production, output in the alternative case would be 14% lower if city sizes had been restricted than in the baseline case of observed sizes. If we pick parameters representing the upper end of plausible scale effects, that is $\sigma_A = 0.12$ and $\sigma_B = 0.20$, the reduction in output is 22%. These strike us a relatively small effects, given the importance that often assigned to large cities as drivers of economic growth. For example, using PPP data from 2022, Canada's per capita GDP was 80% of the US level.

Table 5: Output Relative to Baseline 2010

Parameters	$\sigma_B = 0.06$			$\sigma_B = 0.20$		
	$\sigma_A = 0.04$	$\sigma_A = 0.08$	$\sigma_A = 0.12$	$\sigma_A = 0.04$	$\sigma_A = 0.08$	$\sigma_A = 0.12$
$\lambda = 1.00$ and $\beta = 3.1$	0.917	0.863	0.816	0.877	0.825	0.780
$\lambda = 0.75$ and $\beta = 2.4$	0.919	0.865	0.817	0.880	0.829	0.783
$\lambda = 1$ and $\beta = 0$	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.

Comparing the last line of Table 5 with the two above it shows that an important role in moderating the impact of city size limitation is being played by the fishing out effect embodied in the Bloom *et al.* production function for technology. When this effect is turned off by setting $\beta = 0$, the decline in output comparing the alternative case to the baseline 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.)

The results in Table 5 can easily be transformed to examine the growth rate of output rather than its level. Recall that the experiment that we are considering is imposing a cap on city sizes in the US starting in the year 1900 – this is the point in time in which our baseline and alternative scenarios diverge. Using data from the Maddison Project, GDP per capita in the United States 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 86% of its observed value, the annual growth rate would instead have been 1.52%.

7. Conclusion

In order to assess the effect of agglomeration on economic growth in the United States, we have considered the effect of counter factually limiting city sizes starting in the year 1900 on GDP per capita in the year 2020. 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.

Our conclusion is the the effects of limited city size would have been surprisingly small – or put differently, that there was surprisingly little benefit from agglomeration. To give an example, consider the case in which city size was limited to one million people. Our estimate of the static productivity effect is that in this case (holding the level of technology constant), output would have been 88% of its baseline level, using

our standard set of parameters. The dynamic effect of limiting city size in this fashion over the 120 year period that we consider would be that the level of technology would have been 98% of its baseline level. Multiplying these effects, output in the case with limited agglomeration would have been 14% lower that the baseline. GDP growth in the alternative scenario would have been 0.014 percentage points lower than if city size had not been limited (i.e. 1.52% 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 on research. Moving to the very highest end of the range of parameters estimated in the literature does not substantially reverse our finding, but it is always possible that the literature has been wildly off base.

A second possibility is that there are effects of urban scale, either static or dynamic, that we have failed to account for.

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 on which to put cities.

A final possibility is that we are being too broad 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. (2022a). The returns to face-to-face interactions: Knowledge spillovers in silicon valley. Technical report, National Bureau of Economic Research.

Atkin, D., Chen, M. K., and Popov, A. (2022b). 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.

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

De la Roca, J. and Puga, D. (2017). Learning by working in big cities. *Review of Economic Studies*.

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

Duranton, G. and Puga, D. (2019). Urban growth and its aggregate implications. Technical report, National Bureau of Economic Research.

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

Fogel, R. W. (1964). *Railroads and American economic growth*. Johns Hopkins Press Baltimore.

Forstall, R. and NBER (1995). U.s. decennial county population data, 1900-1990. Technical report. 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.

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

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

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

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.

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`.