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 and to seminar audiences at UCLA 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 agglomerations. 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 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 urban scale 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 effect works dynamically and has an impact external to a particular city. New technologies are more efficiently created in large cities, but these technologies raise productivity everywhere, with technological progress accumulating over time. We begin with existing estimates of the magnitudes of these effects. Using a straightforward 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.

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 on 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 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, 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 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 that would have resulted from limitations on city sizes. In Section 6, we combine the static and dynamic effects to calculate the overall impact 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 (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 1900 to 2010.[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 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. 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, 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. 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 smaller than the 13% elasticity reported by Glaeser and Gottlieb (2009) for the same regression using data for 2000 and slightly different MSA definitions.

[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.

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. 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 (e) 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.

3. Agglomeration Economies and Output 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 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 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 changing the size distribution of cities would affect 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, does not differ between the two cases at a point in time. Feenstra et al. (2015) show that across countries capital-output ratios do not vary systematically with income. Similarly Jones and Vollrath (2023) show the relative constancy of this ratio over time within countries that are arguably close to their balanced growth paths.

Assuming a fixed K/Y ratio allows us to incorporate the effect of "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. 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.[3]

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 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

[3]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.

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 realized and counterfactual cases.

Restricting attention to this particular class of counterfactuals 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_{it}. Multiplying each term in the sum on the right hand side of (6) by $\left(\frac{A_{it}^{base}}{A_{it}^{base}}\right)^{1/(1-\gamma)}$ and using equation (6) 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

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

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}

[4]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 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).

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 urban 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 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) 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 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 worker 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 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 :

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

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.[5]

B *Parameterization*

Given the data already in hand, evaluating equation (12) 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.

Ciccone and Hall (1996) is an early effort to estimate urban scale effects and, like us, explicitly accounts for both capital and human capital. Although their estimation strategy is indirect, the parameters that Ciccone and Hall (1996) estimate correspond closely to

[5]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.

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
2.2%	2.9%	5.1%	De la Roca and Puga (2017)	Spanish, Ind. wages, 2004-9
4.5%	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.

ours. Their benchmark estimate implies that $\sigma_A = 3.4\%$ and that σ_h is indistinguishable from zero.[6] 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 (7) we can derive the relationship between city size and output as,

$$\frac{\partial \log Y_{it}}{\partial \log L_{it}} = 1 + \frac{\sigma_A}{1-\gamma}. \tag{13}$$

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

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

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, 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).[7]

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

[6]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.

[7]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.

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 assumes that worker characteristics are unaffected by the size of the city where they work, and using equation (14) their estimates imply $\sigma_A = 2.5\%, \sigma_H = 0$ and that $\frac{\sigma_A}{1-\gamma} + \sigma_h = 3.7\%$

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

To evaluate equation (12) 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 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

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.

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 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.

Recalling that restricting city 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. 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. Our most extreme counterfactual, where we reduce all cities to 50k, and consider σ_A at the upper end of plausible estimates, reduces output by about 40%.

4. Agglomeration Economies and Patents in a Cross Section

The analysis of the previous section takes the city invariant component of TFP, $\bar{A}_t$, 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 $\bar{A}_t$. Putting the three steps together, we describe the relationship between the distribution of city sizes and the speed of technological progress.

Our approach to modeling the relationship between city sizes and patents parallels our approach to the relationship between city sizes and output in the previous section. 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,[8]

$$P_{it} = B_{it} h_{it}^R \ell_{it}^P L_{it}. \tag{15}$$

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

$$P_t = \sum_{i=1}^{N} B_{it} h_{it}^R \ell_{it}^P L_{it}. \tag{16}$$

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,

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

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,

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

We do not restrict the relationship between city-specific output productivity (the $\widehat{A}_{it}$'s) and city-specific patent 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 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 of having the urban scale effect on research productivity, $\widetilde{B}_{it}$, take the value that would hold if the city were

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

constrained to maximum size L_{max}. To evaluate the resulting change in counterfactual patent 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

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

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

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

In words, equation (20) 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 (20) 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 (20), 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 (20) is σ_B, the effect of city size on patent output. 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 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 $\sigma_B \in [0.17,0.20]$.

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 output for 2020 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 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, patenting falls by 48%.

We can also examine the evolution of patenting over time. For this purpose, we restrict

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.

attention to our baseline value of $\sigma_B = 0.06$ and calculate how patenting changes in each of our three counterfactual systems of cities in each decade for which we have both population and patent data. Figure 3 presents our results.

5. Agglomeration, Patents, and Effective Research Effort

The weight of evidence suggest that the contribution of a patent to TFP growth has changed over time. Berkes (2018) describes annual patent filings from the late 19th to early 21st century. From about 1900 until the late 20th century, annual patent filings by US residents were approximately constant at around 50k per year. Then, over the course of a few decades, from about 1980-2000, filings roughly doubled, before stabilizing at the higher level. That TFP growth does not show a corresponding jump suggests that the increment of innovation in an average patent has not been constant over time. Moreover, the literature on innovation, surveyed in Bloom et al. (2020), establishes that the quantity of resources required to produce a patent has also not been constant over time.

If every patent embodied an idea causing an equal increment to technology, then we could use the calculations reported in figure 3 as basis for calculating the relationship between city size and technological progress. That patents are not reliable as a measure of innovation across time means that completing the chain requires two more links: One connecting patenting to what we call *effective reserach effort*, and a second from effective research effort to technological progress. We here describe the first, and turn to the second in the next section.

We define effective research effort in a city decade as the number of researchers employed in the city, adjusted for urban scale economies in research effort. This follows Bloom et al. (2020), who estimate the relationship between TFP growth and the number of workers employed in research, a quantity they refer to as *research effort*. We differ from Bloom et al. (2020) in that we adjust research effort to reflect urban scale economies.

Although the research effort required to create a patent changes over time, there is no reason to think that it varies systematically within a period, up to the evidence

for urban scale economies described above. This statement has a simple mathematical representation. That is, if R_{it} is effective research output in city i at time t, then cross-city variation in patents at each time is proportional to cross-city variation in effective research effort within the same period. More formally,

$$P_{it} = \mu_t R_{it}. \tag{21}$$

where μ_t describes the decade specific proportionality between effective research effort and patents. For completeness, also define R_t aggregate quantity research effort in a decade t.

If we further assume that urban scale economies operate on research effort in exactly the same way that they operate on patents, then we can substitute equation (21) into (20), to get

$$\frac{R_t^{alt}}{R_t^{base}} = \sum_i \frac{P_{it}^{base}}{P_t^{base}} min\left(1, \left(\frac{L_{max}}{L_{it}}\right)^{\sigma_B}\right). \tag{22}$$

This equation makes an important step in our analysis by completing the logic connecting city size to effective research effort. With it, we can use our patent data and a realized measure of effective research effort to calculate changes in effective research effort under a counterfactual city size distribution. More concretely, given a realized value of effective research effort in any decade, we can adjust it for counterfactual city size distributions using the scaling factors presented on the vertical axis of figure 3.

A *Agglomeration, Effective Research Effort, and TFP*

Our goal is to calculate how the rate of technological progress would change if the largest US cities were smaller. To proceed, we must first think about technological progress in the rest of the world. In practice, ideas cross borders easily, and so if a reduction in city size leads to less new technology being invented in the US, this deficiency would largely be made up by innovation from abroad.

To address this issue we might assume that the same restrictions on city sizes that we impose in the US holds elsewhere as well. 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 is 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.

Bloom et al. (2020) estimate the relationship between productivity growth and aggregate research effort in the US over the period 1930-2015. Adapting their formulation to our notation and implicitly exchanging 'research effort' for 'effective research effort', we

have,

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

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.

To calculate the change in the level of productivity under a counterfactual system of cities, consider a series of observed values of $\bar{A}^{\text{base}} = (\bar{A}_1^{\text{base}}, \bar{A}_2^{\text{base}}, ...)$ in the baseline case where city sizes are not restricted. We calculate this series from the historical TFP data presented in Gordon (2017) for 1900-1950 and Bloom et al. (2020) thereafter.[9] Denote the corresponding counterfactual path where city sizes are restricted as $\bar{A}^{\text{alt}} = (\bar{A}_1^{\text{alt}}, \bar{A}_2^{\text{alt}}, ...)$.

Manipulating equation (23) slightly and rewriting separately for base and alternative cases we have,

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

$$\bar{A}_t^{\text{alt}} = \bar{A}_{t-1}^{\text{alt}} + \alpha (R_{t-1}^{\text{alt}})^{\lambda} (\bar{A}_{t-1}^{\text{alt}})^{1-\beta}. \tag{25}$$

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 Bloom et al. (2020) consider $\lambda = 3/4$. In this case, they estimate that $\beta = 2.4$ In the analysis that follows, we use both pairs of parameterizations.

We can now calculate the path of $\bar{A}$ in the alternative case in three steps. In the first step, we solve equation (24) for effective research effort, R_t^{base}, and then use observed $\bar{A}_t^{\text{base}}$ to construct the base case time series of aggregate effective research effort. In the second step, we use equation (20) to calculate the effective research output, R_t^{alt}, in the alternative case where city sizes are restricted from its base case counterpart. [10] Finally, under the assumption that $\bar{A}_1^{\text{base}} = \bar{A}_1^{\text{alt}}$, we generate a time series for $\bar{A}^{\text{alt}}$ in the alternative case relative to the base case by forward iteration of equation (25) using R_t^{alt}.

Figure 4 shows the result in series for effective research effort. In this figure, the .... TBD.

[9] There is a slight mismatch between our approach that that of Bloom et al. (2020). In their paper, change in aggregate TFP is assumed to be identical to technological progress, which in our notation is the growth of $\bar{A}$. However, in our setting, TFP can also rise because of a change in the average level of the agglomeration term $\tilde{A}_{it}$. In our setting, the right way to measure technological progress would be to start with measured TFP growth and to subtract productivity growth that was due to increased agglomeration. At present we do not do this because we don't have the city-level output series that would be required to construct the latter series.

[10] 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.

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.

Figure 5 shows the final results of our calculation. Each plot in each panel reports the ratio $\bar{A}_t^{\text{alt}}/\bar{A}_t^{\text{base}}$ in a particular counterfactual case. Each panel of the figure reports on three counterfactuals with cities restricted to 1m, 100k and 50k. Finally, 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 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, $\bar{A}$ 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 is on the order of seven percent. This seems somewhat puzzling, given that Figure 3 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 that is at the center of the model in Bloom et al. (2020). Less effective research effort 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 more productive in terms of generating technological progress than it was the base case. This can be seen by examining the bottom 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, $\bar{A}$ in 2020 is 8% below its baseline level, while if city size is restricted to fifty thousand, the reduction is roughly one quarter.

All of the calculations in Figure 5 are based on our preferred value of σ_B, the parameter that measures the urban scale effect in R&D, of 6%. Table 4 shows the sensitivity of this result to the alternative value of 20% that was 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 were examined in Figure5. 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 26 puts together the results from the previous sections.

$$\frac{Y_t^{alt}}{Y_t^{base}} = \left(\frac{\bar{A}_t^{alt}}{\bar{A}_t^{base}}\right)^{1/(1-\gamma)} \sum_i \frac{Y_{it}^{base}}{Y_t^{base}} min\left(1, \left(\frac{L_{max}}{L_{it}}\right)^{\frac{\sigma_A}{1-\gamma}+\sigma_H}\right) \tag{26}$$

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. $\frac{\sigma_A}{1-\gamma} + \sigma_H = 0.04$ and $\sigma_B = .06$, and using Bloom *et al.*'s preferred values for technology production, output in the alternative case would be 8% 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 $\frac{\sigma_A}{1-\gamma} + \sigma_H = 12$ 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

	$\sigma_B = 0.06$			$\sigma_B = 0.20$		
Parameters	$\sigma_A = 0.04$	$\sigma_A = 0.08$	$\sigma_A = 0.12$	$\sigma_A = 0.04$	$\sigma_A = 0.08$	$\sigma_A = 0.12$
$\lambda = 1.00$ 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.

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%.

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.)

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 possibilities that fall under this heading.

One possibility is 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.

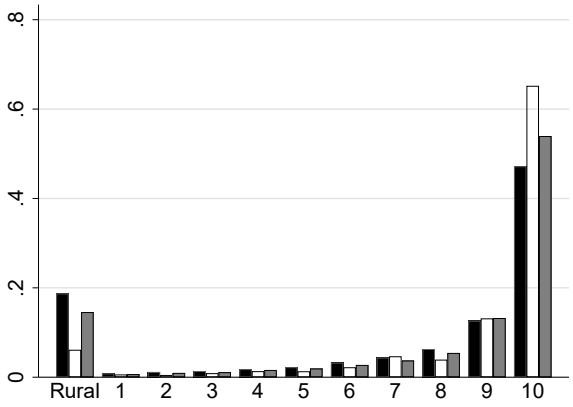
A second possibility is that big cities are important not because of high population *per-se*, but because having a city be very populous allows a lot of people to take advantage of a particularly good location. This channel is ruled out in the approach we take. Specifically, we assume that the location-time specific component of productivity, $\hat{A}_{it}$ that each person experiences is unchanged in going from the baseline to the alternative case. 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 in which to have a city is limited, which seems to us to contradict 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

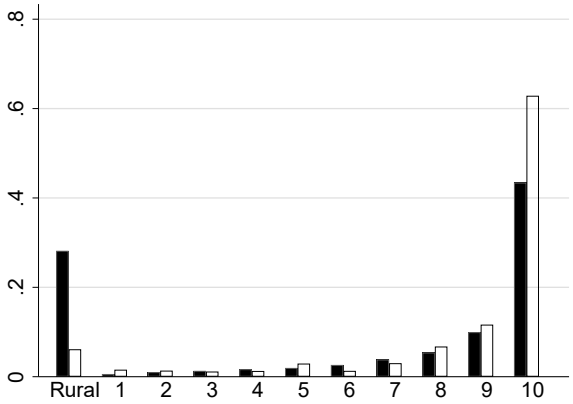
productivity shock, many more people can move there in a world of unrestricted city sizes than in a world where sizes are limited.

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.

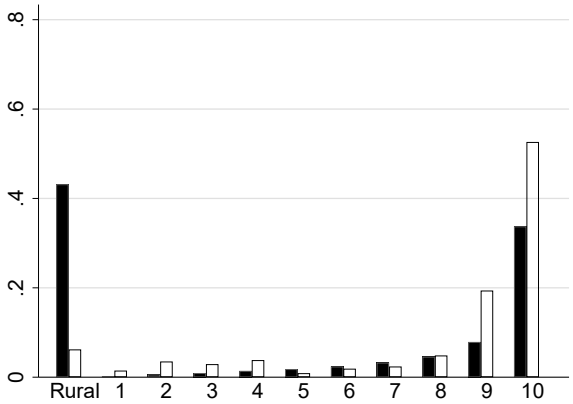
Figure 1: Distribution of population, output and patents by city size in 1900, 1950 and 2010



(a) 2010



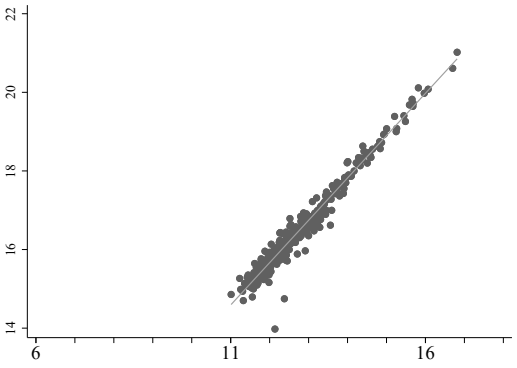
(b) 1950



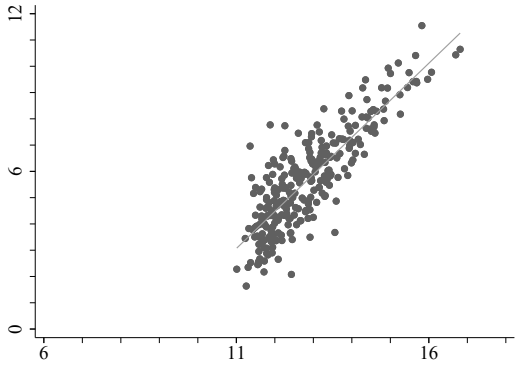
(c) 1900

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.

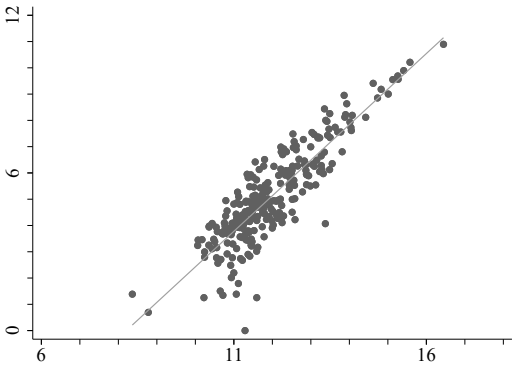
Figure 2: Joint distribution of output, patents and city size.



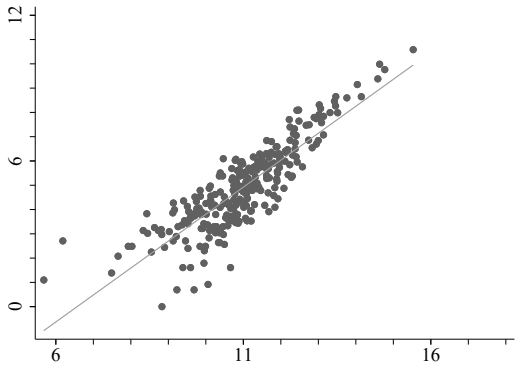
(a) Output vs, Pop. 2010



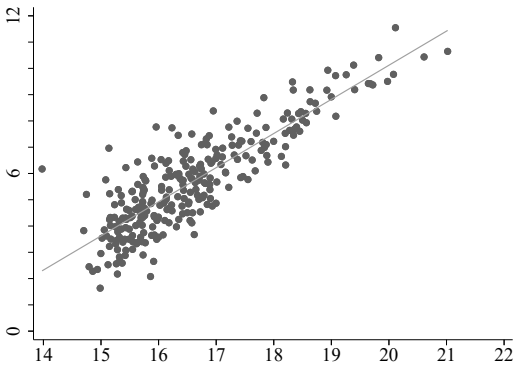
(b) Patents vs Pop. 2010



(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; $\beta = 1.29, \text{s.e.} = 0.047$.

Figure 3: Share of total patents produced under three hypothetical city networks

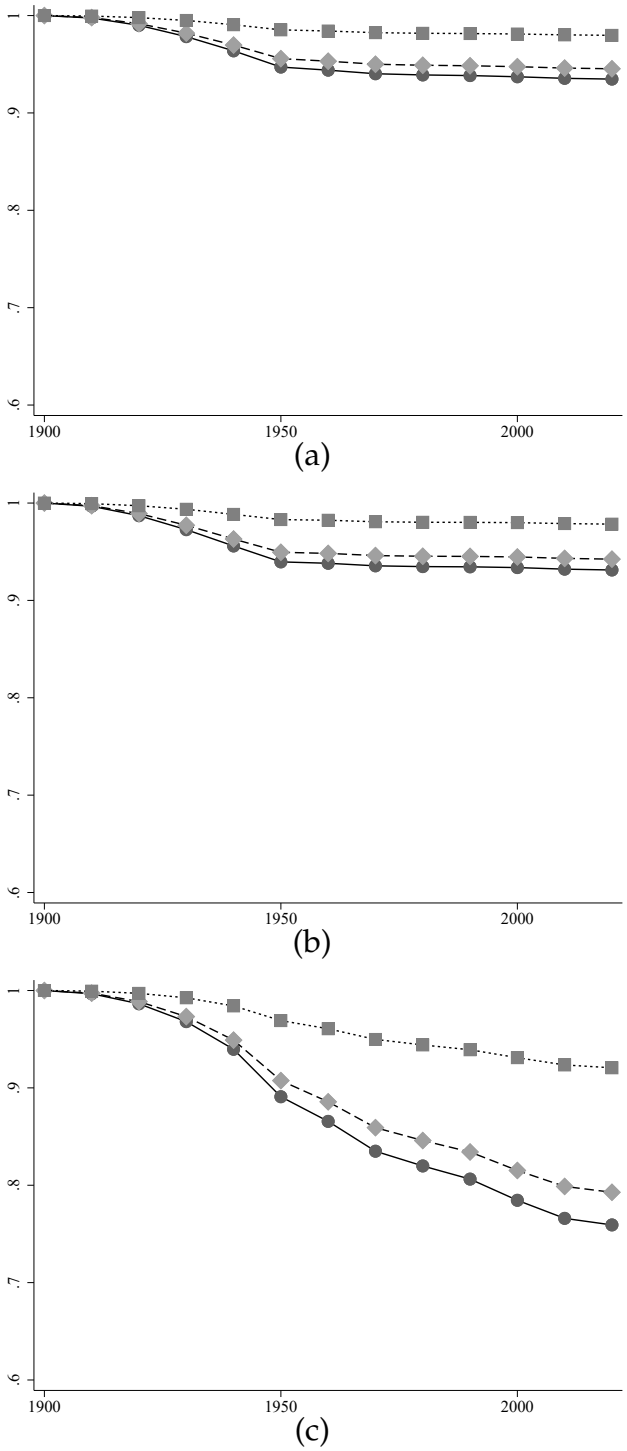
Note: Counterfactual patents as a fraction of actual patents reported in CUSP when city sizes are capped at 1m (squares), 100k (diamonds), and 50k(circles). Calculations assume $\sigma_B = 0.06$.

Figure 4: Observed, Imputed and effective research effort over time

TBD

Note:

Figure 5: Counterfactual trajectories of national productivity.



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.

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.

Our conclusion is the 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 resulting loss the static productivity effect reduces output in 2010 is 88% of its baseline level. Over the 120 year period that we consider, the dynamic effect of restricting city size is that the level of technology falls by 2% to 98% of its baseline level. Multiplying these effects, 2010 output in the case with limited agglomeration would have been 14% lower that the baseline and the 1900-2010 GDP growth rate in the alternative scenario would have been 0.014 percentage points lower than the baseline (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 reverse our finding, but it is always possible that the literature we rely on for our parameter estimates is 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. 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 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. (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.

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.

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.

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.

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.

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

Gordon, R. (2017). *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.

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. (2023). Capital and returns in growth study guide.

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.

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.