CENTRE *for* ECONOMIC PERFORMANCE

**CEP Discussion Paper No 892**

**October 2008**

# Urbanization and Structural Transformation

**Guy Michaels, Ferdinand Rauch and Stephen J. Redding**

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

**Abstract**
This paper presents new evidence on urbanization using sub-county data for the United States from 1880-2000 and municipality data for Brazil from 1970-2000. We show that the two central stylized features of population growth for cities – Gibrat's Law and a stable population distribution - are strongly rejected when both rural and urban areas are considered. Population growth exhibits a U-shaped relationship with initial population density, and only becomes uncorrelated with initial population density at the high densities found in predominantly urban areas. We provide evidence that the explanation for these patterns lies in different employment growth dynamics in the agricultural and non-agricultural sectors and the process of structural transformation away from the agricultural sector.

Keywords: urbanisation, economic development, urban population, rural population
JEL Classifications: R11, R51

This paper was produced as part of the Centre's Globalisation Programme. The Centre for Economic Performance is financed by the Economic and Social Research Council.

**Acknowledgements**
This paper was produced as part of the Globalization Programme of the ESRC-funded Centre for Economic Performance (CEP). We thank Yu-Hsiang Lei, Ralph Ossa, and Bart Thia for excellent research assistance. We also thank Robin Burgess, Alan Manning, Rachel Ngai, Chris Pissarides and seminar participants at LSE for helpful comments. Responsibility for any opinions, results and errors lies with the authors alone.

Guy Michaels is a Research Associate with the Centre for Economic Performance, London School of Economics. Ferdinand Rauch is an Occasional Research Assistant at the Centre for Economic Performance, LSE. Stephen Redding is a Programme Director at the Centre for Economic Performance, LSE.

Published by
Centre for Economic Performance
London School of Economics and Political Science
Houghton Street
London WC2A 2AE



ISBN 978-0-85328-299-0

# 1 Introduction

Urbanization – the concentration of population in cities and towns – is one of the most striking features of economic development.[1] The share of the world's population living in cities grew from less than one tenth in 1300, to around one sixth in 1900 and to almost one half today.[2] While this transition from rural to urban is largely complete in developed countries such as the United States, the urbanization process continues apace in developing countries such as Brazil, China and India. In China alone, 240 million people are expected to migrate from rural to urban areas by 2025, helping to raise the share of the world's population living in cities to 60 percent by 2030.[3] This shift from rural to urban settlement has wide-ranging implications for infrastructure, public health, the environment and economic performance.[4] As a result, urbanization is central to many policy debates and is viewed as a key part of economic development.

In this paper, we provide new evidence on urbanization using a novel dataset that encompasses both rural and urban areas. Our analysis proceeds in three stages. We first outline six stylized facts about the dynamics of population and employment. We next develop a simple model that provides a parsimonious explanation for these six stylized facts in terms of structural transformation away from the agricultural sector. Finally, we show that the model can account both qualitatively and quantitatively for the observed features of urbanization.

Although the urbanization process involves large-scale reallocations of population from rural to urban areas, much of our existing knowledge about this process comes from city or metropolitan area data that exclude rural areas. The two stylized features of population growth that have received most attention in existing research – the extent to which population growth is independent of population size ("Gibrat's Law") and a stable population distribution that exhibits an approximately linear relationship between log rank and log size with a unit coefficient ("Zipf's Law") – have both largely been studied in urban population samples.[5] This exclusion of rural areas is a concern because historically in developed countries and in developing countries today these areas account for a large

[1] The US Census Bureau defines an urban area as territory consisting of core census blocks with a population density of at least 1,000 people per square mile and surrounding census blocks with a population density of at least 500 people per square mile (Census 2000d).

[2] The historical figures are from Bairoch (1988) and the present-day figures from United Nations (2005).

[3] The estimates for China are from Mckinsey (2008) and those for the world are again from United Nations (2005).

[4] There is a large empirical literature documenting higher productivity in urban than rural areas (see for example the survey by Rosenthal and Strange 2004). Similarly, an extensive body of research examines the relationship between urbanization and income inequality (see for example Kuznets 1955 and Black and Henderson 1999).

[5] For empirical evidence on Zipf's Law for a large number of countries, see Rosen and Resnick (1980) and Soo (2005). For the classic treatments of Gibrat's Law, see Simon (1955) and Gabaix (1999).

share of the overall population. The potential importance of augmenting urban population samples to include less densely-populated areas is suggested by the fact that the entry of new cities is a recurrent issue in the urban growth literature (see for example Black and Henderson 1999, 2003), departures from Zipf's Law are observed in the lower as well as the upper tail of the city-size distribution (as shown by Rossi-Hansberg and Wright 2007), and the functional form of the city-size distribution has been found to be sensitive to the minimum population size at which the distribution is truncated (see in particular Eeckhout 2004).

To provide evidence on urbanization for both rural and urban areas, we construct a new dataset that is a partition of the surface area of US states. This dataset exploits information on sub-county units, which are commonly referred to as Minor Civil Divisions (MCDs), and extends for more than a century from 1880 to 2000.[6] Our data include information on both population and employment by industry and are characterized by the following six stylized facts. Despite substantial US population growth, as reflected in an increased mass of densely-populated areas over time, we also find an increased mass of sparsely-populated areas over time. As a result there is an unstable population distribution, which exhibits polarization: the difference in density between densely and sparsely-populated areas increases over time (Stylized Fact 1).

While our data confirm previous findings that Gibrat's Law holds for densely-populated urban areas, we show that this feature of population growth is strongly rejected when we include both rural and urban areas (Stylized Fact 2). For this more comprehensive range of locations, population growth is decreasing in initial population density at low densities, and then increasing in initial population density at intermediate densities, before becoming uncorrelated with initial population density at high densities in urban areas.[7] Although a natural explanation for the decreasing relationship between population growth and initial population density at low densities is mean reversion, the explanation for the increasing relationship at intermediate densities is less transparent.

Our third stylized fact is that the correlation between population growth and initial population density is systematically related to differences in employment structure between agriculture and non-

[6] We exclude Alaska, Hawaii, Oklahoma, North Dakota, and South Dakota, which had not attained statehood in 1880, and therefore are either not included in the 1880 census or did not have stable county boundaries at that time. Additionally, we use county data for some states where sub-county units are not comparable over time. We discuss in further detail below the construction of our data and the robustness of our results to the sample and specification.

[7] While the existing literature on cities concentrates on the relationship between population growth and population size, we focus on the relationship between population growth and population density to control for differences in land area across sub-county units. Although our results are qualitatively the same if we instead use population size, the population density specification is more appropriate if land area varies across sub-county units and is derived directly from our theoretical model.

agriculture (manufacturing and services). In particular, the share of agriculture in employment drops steeply in the range where population growth and initial population density are positively correlated. Our fourth stylized fact is that there is a higher variance in the distribution of employment per square kilometer in non-agriculture than in agriculture (so non-agricultural employment is more concentrated across space). Finally, our last two stylized facts are concerned with differences in employment dynamics in the two sectors. In agriculture, employment growth is decreasing in initial population density (Stylized Fact 5). In contrast, in non-agriculture, employment growth is largely uncorrelated with initial population density (Stylized Fact 6).

As our data span a long historical time period during which the economic environment in the United States changed considerably, we undertake a wide range of robustness checks to sample and econometric specification. We estimate our baseline specification non-parametrically to allow for a flexible relationship between population growth and initial population density. We show that our results are robust to the inclusion of state fixed effects, which in this cross-section specification control for changes in institutions and other characteristics of states that can affect population growth. Among several robustness checks, we find a similar pattern of results when we restrict the sample to a subset of the former thirteen colonies that have similar organizations of administrative functions at the county and sub-county level and the most stable administrative boundaries over time. Additionally, we find a similar pattern of results if we aggregate sub-county units within the immediate vicinity of a city to construct larger economic units, suggesting that our results are not driven by suburbanization around the boundaries of existing cities. Finally, while counties are relatively coarse spatial units for examining the transition from rural to urban, we also find the same pattern of results at the county level.

Most importantly, we also replicate our entire analysis for Brazil for the period 1970-2000. Like the United States during the nineteenth and twentieth centuries, Brazil experienced rapid industrialization during 1970-2000, and therefore we would expect the mechanisms emphasized in our model to apply. Even though these data are for a different country with distinct institutions and physical geography, and even though these data are collected at a different level of spatial aggregation and for a different time period, we find strikingly similar results to those we document for the United States. This similarity of the results in a quite different context reassures us that our findings are not driven by idiosyncratic features of the data or the institutional environment in the United States.

To make sense of our empirical findings, we develop a simple general equilibrium model of the distribution of population across locations that distinguishes between agriculture and non-agriculture. Workers are geographically mobile and the distribution of population across space is determined by the requirement that workers are indifferent between locations. Within each location, land is allocated endogenously to residential and commercial use depending on its relative return in the two types of activities. Land used commercially can be employed in either the agricultural or non-agricultural sector, and in equilibrium it is allocated to the sector in which it has the higher value marginal product. As idiosyncratic shocks to productivity in each sector and location occur, workers move across locations to arbitrage away real wage differences, and within locations land endogenously switches between agricultural and non-agricultural use.

To match the observed employment dynamics for agriculture and non-agriculture, we assume that the stochastic process for productivity in agriculture is mean reverting due to the influence of natural endowments such as climate and soil, whereas the stochastic process for productivity in non-agriculture exhibits constant proportional growth. As a result agricultural productivity is bounded from above, while non-agricultural productivity is unbounded, which generates a non-agricultural employment share that is positively correlated with population density in relatively dense areas. This positive correlation is further enhanced in the model by greater land intensity and weaker agglomeration forces in agriculture than in non-agriculture.

As in the large macroeconomics literature concerned with unbalanced growth, structural transformation away from agriculture occurs because productivity growth is more rapid in non-agriculture than in agriculture and there is inelastic demand between the two goods.[8] Given inelastic demand, more rapid agricultural productivity growth leads to a more than proportionate decline in the relative price of the agricultural good, which in turn leads to a reallocation of employment from agriculture to non-agriculture.[9] Combining more rapid employment growth in non-agriculture, a non-agricultural employment share that is positively correlated with population density in relatively dense areas and different employment dynamics in the two sectors, the model generates the U-shaped relationship between population growth and initial population density observed in our data.

Our paper is related to a large body of work in urban economics and economic geography. Recent research on the relationship between city growth and size includes Cordoba (2008), Duranton

[8] See in particular Baumol (1967), Galor (2005), Galor and Weil (2000), Goodfriend and McDermott (1995), Ngai and Pissarides (2007) and Rogerson (2008).

[9] A straightforward extension of the model also incorporates the other leading explanation for structural transformation in the macroeconomics literature, namely non-homothetic preferences and technological progress that raises real incomes. See among others Echevarria (1997), Gollin et al. (2002) and Matsuyama (2002).

(2007), Eeckhout (2004), Gabaix (1999) and Rossi-Hansberg and Wright (2007). While the Pareto distribution (of which Zipf's Law is a special case) is viewed as a good approximation to the observed distribution of city populations, Rossi-Hansberg and Wright (2007) find systematic departures from the Pareto distribution in the upper and lower tails. Evidence of departures from a Pareto distribution is also found by Eeckhout (2004), who uses data on Census Designated Places (CDPs) in the United States to provide evidence that the population distribution is log normal, as implied by Gibrat's Law of constant proportional growth.[10] While Gibrat's Law is generally seen as a good approximation to observed city population growth rates, there is also some evidence of departures from Gibrat's Law, as found for example by Black and Henderson (2003), González-Val et al. (2008) and Soo (2007).[11] Two empirical issues in the existing literature on cities are the treatment of new cities and the minimum population size to be considered as a city, both of which are addressed in our approach by considering the entire distribution of population across both rural and urban areas.[12]

Our focus on the reallocation of economic activity from agriculture to non-agriculture connects with theories of new economic geography, including Fujita et al. (1999) and Krugman (1991). Although reductions in trade costs in these models can result in a polarization of population across space, they do not provide natural explanations for why Zipf's Law and Gibrat's Law are good approximations to the observed distributions of city populations and growth (see for example the discussion in Davis and Weinstein 2002). While a large literature has examined the empirical determinants of the distribution of economic activity across states and counties in the United States, much of this literature abstracts from the reallocation of economic activity from agriculture to non-agriculture.[13] Closer in spirit to our work is Caselli and Coleman (2001), who examine structural transformation and the convergence of incomes between Southern and Northern US states. Also related is Desmet and Rossi-Hansberg (2007), who examine differences in patterns of employment growth between the manufacturing and service sectors using US county data, and relate these differences to technological diffusion and the age of sectors. Neither paper examines the relationship between structural transformation and urbanization – an analysis for which our newly-constructed

[10] See also Eeckhout (2008) and Levy (2008).

[11] Research on the empirical determinants of city growth includes among others Glaeser et al. (1992), da Mata et al. (2007) Ioannides and Overman (2004), as surveyed in Gabaix and Ioannides (2004). The role of industrial specialization is emphasized in Henderson (1974).

[12] Within the cities literature, Henderson and Wang (2007) and Henderson and Venables (2008) examine the emergence of new cities as a source of growth in the urban population, while Henderson (2005) and Williamson (1965) examine the relationship between the share of the urban population and economic development.

[13] See Beeson et al. (2001), Ellison and Glaeser (1999), Glaeser (2008), Kim (1995) and Rappaport and Sachs (2003) among others.

sub-county data are especially well suited.

In addition to the macroeconomic literature discussed above, our research is related to the development and economic history literatures. Early work documenting the importance of structural change to economic development is surveyed in Syrquin (1988), while more recent research on the interlinkages between industrial and agricultural development is reviewed in Foster and Rosenzweig (2008). Influential work on the history of urban development in the United States includes Kim (2000) and Kim and Margo (2004), although for reasons of data availability this research has again largely concentrated on cities.

The remainder of the paper is organized as follows. Section 2 discusses our main dataset for the United States, outlines our empirical strategy, presents our main empirical findings, and reports the results of a number of robustness checks. Section 3 presents the results of an additional robustness check using Brazilian data. Section 4 outlines our theoretical model and Section 5 shows that it can quantitatively account for the patterns observed in our data. Section 6 concludes.

# 2 US Data and Stylized Facts

## 2.1 Data and Samples

This section begins by introducing the US data that we use in this paper and the samples that we construct. We then document a set of stylized facts that shed light on the dynamics of urban and rural population growth from 1880-2000.

In order to analyze these dynamics, we require data on land area, population, and sectoral employment for geographic units that are consistent over time. Since we are interested in both rural and urban areas, we also require that these geographic units partition the land area that we analyze. In other words, we want a dataset that covers the entire population and all the land - from the largest cities to the smallest farms. And since we are interested in examining rural and urban population dynamics, we prefer that our geographic units be fine enough to separate urban areas from rural ones.

While these criteria may seem natural, it is not easy to find an existing dataset that satisfies them all. The literature on urban growth in the US has often analyzed counties or Metropolitan Statistical Areas (MSAs), which are groups of counties. And although counties satisfy most of our requirements, they often pool together urban centers with their surrounding countryside. So while we include counties in our analysis, we are also interested in data that provide finer spatial aggregation.

One dataset that is less aggregated than the county dataset includes incorporated places - this is the dataset used by Eeckhout (2004). But while this data is useful for studying urban growth dynamics, it does not contain information on many rural areas, where the majority of US population lived before the 20th century.

Since existing datasets are not fully satisfactory for our purposes, we construct a new data set using minor civil divisions (MCDs). MCDs have been used to report population in parts of the US, especially in the Northeast, since the first census in 1790 (see Census 2000c). But as we discuss below, we are interested not only in population but also in sectoral employment. And since the earliest available digitized employment data for MCDs comes from the 1880 Census, we chose 1880 as the starting year for our analysis. Over time, MCDs became a standard tool for partitioning counties throughout (almost) the entire US.[14] It is this feature of MCDs that makes them so suitable for our analysis: they provide the finest level of geographical disaggregation for which we can analyze urbanization and structural transformation over more than a century.

The most common types of MCDs are towns and townships, but in some areas election precincts, magisterial districts, parishes, election districts, plantations, reservations or boroughs were used as MCDs (even this long list is not exhaustive). As some of these names suggest, in many states MCD boundaries coincide with those of local government bodies. In New England in particular, MCDs are actively functioning units of local government, in many cases since the $17^{th}$ Century. But in other states MCDs are often statistical entities with few (or no) other functions (see Chapter 8 Census 2000c). Given the variation in their functions, it is not surprising that the size and shape of MCDs also vary from state to state. For example, in the Midwest MCDs are often follow a chessboard patterns with squares of 6 miles per side; this design dates back to the Land Ordinance of 1785 and the Northwest Ordinance of 1787 (see Prescott 2003). As one travels West or South, the size of MCDs tends to grow, and they tend to become less regular and less stable over time.

To address concerns about differences in the geographical and institutional organization of MCDs, we report results for subsets of states with similar geographical and institutional organizations. To overcome changes in MCD boundaries, we aggregate some MCDs to create geographic units that are stable over time. This aggregation process involved considerable work using historical maps and gazetteers, and it is described in further detail in the appendix. To give the reader a brief idea

[14] In many Western states sub-county units were initially called MCDs but were reclassified as census county divisions (CCDs) in 1950, when the map of sub-county units in many of these states was redrawn. For simplicity, we refer to both MCDs and CCDs as MCDs (see chapter 8, Census 2000c).

of the aggregation process, we matched the approximate centroid of each 1880 and 1940 MCD to the 2000 MCD in which it fell. We then aggregated any 2000 MCD that did not contain at least one 1880 MCD and one 1940 MCD to the nearest 2000 MCD that did. This aggregation process enables us to track the evolution of population at a fine level of spatial detail over 60-year intervals.[15] One reason for restricting ourselves to these years is that adding more years would have forced us to aggregate further. But perhaps more importantly, we only know the employment structure of MCDs for 1880 (using the individual-level census records from the North Atlantic Population Project) and for the very recent censuses, such as 2000 (using data from the US census American Factfinder tool, see Census 2000b). Since our analysis focuses on agricultural and non-agricultural employment and growth dynamics, adding more years for which we don't have this information would have not contributed much. Finally, we used the 2000 census to calculate the land area in each geographic unit.

The extent of aggregation required to construct time-consistent units varied considerably by state. In some states, especially in the Northeast and the Midwest, MCDs corresponded to local administrative units that were very stable over time, so little aggregation was required. We therefore divided states into samples: little aggregation was required in A states, more was needed in B states, and more still in C states. The geographic distribution of states across these three groups is shown in Map 1. In choosing our baseline sample, we sought to include as many states as possible while limiting the extent of aggregation, since the aggregation process might entail some imprecision. We therefore choose the A and B states for our baseline sample, for which $1-1$ matches between the 1880 and 2000 censuses involving no aggregation exceeded 70 percent.[16] But as we discuss below, we also construct alternative samples that either include more states (by using county-level data) or restrict our sample to A states, where very little aggregation was required. In our baseline sample there are, on average, 13 units ("MCDs") per county. The average unit spans $115km^2$, with a population of 2,400 in 1880 and 8,800 in 2000.

[15] While all MCDs in our baseline sample of "A and B" states have non-zero population in all three years of our sample, there are 7 MCDs in the C states that have zero population in 1880. These are dropped when we construct population growth rates.

[16] As in most cases our geographic units consist of a single MCD, we refer for simplicity to these units as "MCDs", even though they sometimes consist of multiple MCDs.

## 2.2 Empirical Strategy

We are interested in characterizing the population density distribution and the relationship between population growth and the initial population density distribution. In both cases, we adopt a nonparametric approach that imposes minimal structure on the data.

To characterize the population density distribution, we divide the range of values for log population density, $x$, into discrete bins of equal size $\delta$. We index MCDs by $m$ and bins by $b \in \{1, ..., B\}$. Denoting the set of MCDs with log population density in bin $b$ by $\Phi_b$ and denoting the number of MCDs with log population densities within this set by $n_b$, we estimate the population density distribution, $\hat{g}(x_m)$, as follows:

$$\hat{g}(x_m) = \frac{n_b}{n}, \qquad n = \sum_{b=1}^{B} n_b, \qquad \text{for } x_m \in \Phi_b. \tag{1}$$

This corresponds to a simple histogram, which yields a consistent estimate of the true underlying probability density function (Scott 1979). We choose bin sizes of $\delta = 0.1$ log points, which provides a fine discretization of the space of values for log population density, while preserving a relatively large number of MCDs within bins.[17] Although this approach provides a simple and flexible characterization of the population density distribution, which connects closely with the other components of our analysis below, we also find similar results using related non-parametric approaches such as kernel density estimation (Silverman 1986).

To characterize the relationship between population growth and the initial population density distribution, we follow a similar approach. We approximate the continuous function relating population growth to initial population density using a discrete-step function consisting of mean population growth within each initial population density bin:

$$y_{mt} = f(x_{mt-T}) = \sum_{b=1}^{B} \phi_b I_b, \qquad \phi_b = \frac{1}{n_b} \sum_{m \in \Phi_b} y_{mt}, \qquad \text{for } x_m \in \Phi_b. \tag{2}$$

where $m$ again indexes MCDs, $b$ again indexes bins and $t$ indexes time. In this specification, bins are defined over initial population density, $x_{mt-T}$; $y_{mt}$ is average population growth from time $t-T$ to $t$; and $I_b$ is an indicator variable equal to one if $x_{mt-T} \in \Phi_b$ and zero otherwise.

[17] In all the graphs throughout this paper we remove the top and bottom one percent of the observations from the graphical representation, but not from the regressions. The bins at these extremes of the distribution contain few observations and have correspondingly large standard errors. Hence they tend to cloud rather than to illuminate the true picture.

This specification corresponds to a regression of population growth on a full set of fixed effects for initial population density bins. We report both mean population growth and the 95 percent confidence intervals around mean population growth for each initial population density bin. The confidence intervals are based on heteroscedasticity robust standard errors adjusted for clustering by county, which allows for correlated errors across MCDs within counties. While this non-parametric specification allows for a flexible relationship between population growth and initial population density, we again find similar results using other related non-parametric approaches, such as locally weighted linear least squares regression (Cleveland 1979) and kernel regression (Härdle 1990). A key advantage of the specifications in (1) and (2) is that we can preserve the same discrete bins when analyzing the population density distribution, the relationship between population growth and the initial population density distribution, and the variables for the agricultural and non-agricultural sector discussed below.

As our model yields predictions for the functional form of the relationship between population growth and initial population density, we also estimate parametric versions of specification (2) of the form:

$$y_{mt} = \rho x_{mt-T} + u_{mt}, \tag{3}$$

where $\rho$ is a parameter, $u_{mt}$ is a stochastic error, and we again report standard errors clustered by county.

Finally, our model highlights a relationship between population dynamics and employment dynamics in the agricultural and non-agricultural sectors. Therefore, in addition to the specifications for population in (1)-(3), we also estimate related specifications for employment in the agricultural and non-agricultural sectors.

## 2.3 Stylized Facts

To better understand the process of urbanization and structural transformation in the US from 1880-2000, we present a set of 6 stylized facts. These facts highlight the instability of the spatial distribution of economic activity over this time period, which lies in stark contrast to the stability documented within the sample of cities in the literature on urban growth. These facts also suggest that this instability is closely related to the different growth dynamics of the agricultural and non-agricultural sectors, and to structural transformation away from agriculture.

We begin by reporting a number of descriptive statistics for our baseline sample of "A and

B" states in Column (1) of Table 1. Figures 1-6 then display the results of the non-parametric specifications (1) and (2) for population and for employment in the agricultural and non-agricultural sectors. Our first stylized fact is that the distribution of log population density across MCDs has become more dispersed from 1880-2000. As shown in Panel A of Column (1) in Table 1, the standard deviation of the distribution of log population density increased over this period from 0.97 to 1.56, which is both statistically significant and larger than the increase in mean log population density. Figure 1 confirms this increase in dispersion by displaying the results from specification (1). Although the US population increased substantially between 1880 and 2000, as reflected in Figure 1 in an increased mass of densely-populated areas, the figure also shows an increased mass of sparsely-populated areas. The population density distribution therefore exhibits polarization with some low-density areas depopulating as other higher-density areas experience rapid population growth. This instability of the overall distribution of population stands in sharp contrast to the stability of the distribution of city sizes (e.g. Duranton 2007). Existing research for cities finds that the population size distribution appears to follow Zipf's Law for large cities (e.g. Gabaix 1999) or a lognormal distribution for a wider range of city sizes (Eeckhout 2004).[18]

Second, Gibrat's law that population growth and population size are uncorrelated is clearly violated. While Gibrat's law has been confirmed in a number of studies for several countries (e.g. Gabaix 1999, Eeckout 2004, Eaton and Eckstein 1997, Ioannides and Overman 2004), other studies find evidence of violations of Gibrat's Law even for urban population samples (e.g. Black and Henderson 2003, González-Val et al. 2008, Soo 2007). In contrast to these studies, we examine population growth for both urban and rural areas. In Figure 2, we display the results from our population growth specification (2), where the dark solid line denotes mean population growth within each initial population density bin and the lighter dashed lines denote the 95 percent confidence intervals. As shown in the figure, log population density in 1880 is strongly predictive of population growth from 1880-2000. A similar relationship is found if we replace initial population density with initial population, as discussed further below. As Figure 2 shows, for low population densities, there is a negative correlation between population density in 1880 and subsequent population growth. But above the threshold of log population density of about 2, population density in 1880 is positively correlated with subsequent population growth. The magnitudes of these differences are large: MCDs

[18] The null hypotheses that the 1880 and 2000 distributions are drawn from a Pareto distribution with an unchanged shape parameter or from a lognormal distribution with an unchanged mean and variance are strongly rejected.

with log density of about 0 or 4 in 1880 experienced population growth at a rate of about 1 percent from 1880-2000. By contrast, MCDs with population density around 2 barely grew on average. As shown in Panel B of Column (1) in Table 1, these differences are statistically significant. We also note that at levels of population density above 4 population density seems to be largely uncorrelated with population growth; this is the range that typically includes urbanized areas. Hence this finding is broadly consistent with the literature that documents Gibrat's law for cities. And yet for most of the population density distribution, and in the range that includes most of the 1880 population, we see a strong positive correlation between initial population density and subsequent growth.

Third, the share of agriculture in employment drops steeply in the range where population density in 1880 and subsequent growth are positively correlated. Figure 3 presents the results from specification (2) using the share of agriculture in employment in 1880 as the left-hand side variable rather than population growth. As shown in the figure, the agricultural employment share in 1880 drops from about 0.8 for MCDs with log density of 2 to about 0.2 for MCDs with log density of 4. Panel C of Column (1) in Table 1 shows that this difference is statistically significant. For denser MCDs the share continues to decline, but at a much slower rate.[19]

Fourth, the distribution of employment per square kilometer across MCDs has a lower standard deviation in agriculture than in non-agriculture in both 1880 and 2000. As shown in Panel D of Column (1) in Table 1, this difference is statistically significant at conventional critical values. Figure 4 presents the results from specification (1) for employment in agriculture and non-agriculture in 1880 and 2000. As shown in the figure, the employment density distribution in agriculture also has thinner tails than its non-agricultural counterpart.[20] Therefore there are more observations with extreme low and high values of employment density for non-agriculture than for agriculture, reflecting the greater spatial concentration of non-agricultural employment. Furthermore, a comparison of Figures 1 and 4 suggests that 1880 population was distributed in a similar way to 1880 agricultural employment, while 2000 population is more spatially concentrated and distributed in a similar way to 2000 non-agricultural employment. This is perhaps not surprising, since agriculture's share in employment in the average MCD declined over this period from 63 percent to 6 percent, and its share in overall employment fell from 12 percent to 0.5 percent.

[19] The share of employment in total population was about 0.33 in 1880 and 0.48 in 2000. In both years, it was relatively stable across the population density distribution, suggesting that employment dynamics are a reasonable predictor of population dynamics.

[20] We also find that non-agricultural employment per square kilometer is more unequally distributed than agricultural employment in both 1880 and 2000 using standard measures of inequality such as the Gini Coefficient, the Theil Index, the difference between the 90th and 10th percentiles, and the difference between the 99th and 1st percentiles.

Fifth, agricultural employment growth appears to follow a mean-reverting process. To document this stylized fact, we consider the subsample of MCDs for which agriculture accounted for more than 80 percent of 1880 employment. Although the share of agricultural employment in this subsample was over 88 percent in 1880, it fell to below 10 percent in 2000, and hence this subsample does not entirely capture agricultural dynamics alone. Nevertheless, since this subsample was at least initially mostly agricultural, it is likely to capture the main features of agricultural growth.[21] Figure 5 displays the results from non-parametric specification (2) for this subsample using agricultural employment growth as the left-hand side variable. As apparent from the figure, densely-populated MCDs in this subsample exhibited much slower growth of agricultural employment from 1880-2000 than sparsely populated MCDs. Panel E of Column (1) in Table 1 reports the results from parametric specification (3) for this subsample, again using agricultural employment growth as the left-hand side variable. This confirms our finding of mean reversion: the coefficient on log population density in 1880 in the parametric specification is $-0.006$ and significant (p-value $< 0.001$). From the size of this coefficient, each additional log point of population density in 1880 is associated on average with just over half a percentage point lower rate of agricultural employment growth. We find very similar results if we instead relate agricultural employment growth to log agricultural employment density in 1880: the coefficient on initial log agricultural employment density is $-0.006$ and statistically significant.

Sixth, in contrast to agricultural employment, non-agricultural employment growth is uncorrelated with 1880 population density. To demonstrate this, we consider the subsample of MCDs for which agriculture accounted for less than 20 percent of 1880 employment. In this subsample the share of non-agricultural employment was higher than 90 percent in 1880 and higher than 98 percent in 2000. Figure 6 displays the results from non-parametric specification (2) using non-agricultural employment growth as the left-hand side variable, while Panel F of Column (1) in Table 1 reports the results from the analogous parametric specification (3). As apparent from the figure, non-agricultural employment grew at about 1.2 percent per year. This positive growth rate is very different from the (mostly) negative growth rates of agricultural employment shown in Figure 5. Moreover, in sharp contrast to the results for the agricultural sector, non-agricultural employment growth is uncorrelated with 1880 population density. As reported in Panel F of Column (1) in Table

[21] We also find mean reverting processes when we consider population growth (rather than employment growth) for both 1880-2000 and 1880-1940 for the same agricultural subsample. For the second of these periods, agriculture remains an important employer in much of the US for both the start and end year.

1, the coefficient on log population density in 1880 in the parametric specification is $-0.0002$ and statistically insignificant (p-value = 0.515). We also find very similar results if we instead relate non-agricultural employment growth to log non-agricultural employment density in 1880. The coefficient on log non-agricultural employment density is $-0.00021$, which is more than an order of magnitude smaller than the corresponding coefficient for the agricultural sector, and statistically insignificant.

## 2.4 Robustness of the Stylized Facts

Having documented the 6 stylized facts for our preferred sample of MCDs, we now examine their robustness to different samples and specifications. The results of these robustness checks are summarized in Columns (2) to (8) of Table 1, while Figure 7 replicates the non-parametric population growth specification (2) displayed in Figure 2 for each of the robustness checks.

One potential concern about our preferred sample is that imperfect matching of MCDs across censuses could have affected our estimates. For example, some of the population and employment of MCDs with intermediate densities could have been assigned to MCDs with either higher or lower densities. To address this concern, the second column of Table 1 shows that all of our stylized facts remain intact when we restrict the sample to MCDs in the "A states" (to which we henceforth refer as the restricted sample). In this restricted sample match rates are well over 90 percent, so imperfect matching is unlikely to be the cause of our finding. Figure 7 also shows non-parametrically that the U-shape we document in the second stylized fact is still strongly apparent in this sample.

Another possible concern is that we use a level of aggregation that is too fine. For example, people could live in one MCD and commute to work in another MCD, which could in turn influence the correlation between population growth and population density. As a first step to address this concern, we replicate our analysis using county-level data, since fewer people commute across county boundaries than across MCD boundaries. The third column of Table 1 uses county-level data for 45 states and Washington DC.[22] The fourth column restricts the county sample to our baseline "A and B" states. And the fifth column reports results using a hybrid sample of MCDs for states where matching was possible and counties for other states.

Our results are robust across all three specifications with two exceptions. The first stylized fact does not hold in Column (3), where the standard deviation of log population is not statistically sig-

[22] As noted in footnote 6, we exclude Alaska, Hawaii, Oklahoma, North Dakota and South Dakota, which had not attained statehood in 1880, and therefore are either not included in the 1880 census or did not have stable county boundaries at that time.

nificantly different between 1880 and 2000. The sixth stylized fact does not hold in Columns (3) and (5), where we find some evidence of mean reversion in both agriculture and non-agriculture. These exceptions are perhaps not surprising because the samples in Columns (3) and (5) include Western states that were not yet fully settled in 1880. Early settlement dynamics in these states, around the time of the "Closing of the frontier" (identified in the 1890 Census), are likely to be quite different from those elsewhere. As the Western states include areas that were largely uninhabited in 1880, they have correspondingly high standard deviations of log population in 1880, consistent with the exception to stylized fact 3. Relatedly, the future settlement of areas that were largely uninhabited in 1880 provides a natural explanation for mean reversion that is unrelated to employment structure, consistent with the exception to stylized fact 6. Despite these caveats, our results show that in areas that were well-settled by 1880, the stylized facts are robust to aggregating MCDs up to the county level.[23]

While county-level data are consistent with our results, a further concern is that the aggregation they provide is insufficient around large cities. Some large cities (Metropolitan Statistical Areas (MSAs)) span multiple counties and may be characterized by commuting across county boundaries. Additionally, the suburbanization that occurred in the second half of the twentieth century can extend beyond county boundaries and could have influenced population dynamics in the neighborhood of large cities. To address these concerns, we undertake further aggregation. One possibility would be to aggregate counties based on 20th-century definitions of MSAs, but these definitions are themselves endogenous to population growth during our sample period. Therefore we instead aggregate MCDs based on 1880 characteristics using a flexible approach that allows us to consider various levels of aggregation. Starting with our baseline sample, we identify as "cities" MCDs that had 100,000 people or more in 1880. To each of these cities we add the land area, population, and employment of any MCD whose geographic centroid lies within 25 kilometers of their centroid.[24] We label the resulting sample a suburban sample, since it pools together large urban centers with

[23] To further test whether our results are affected by the US's Westward expansion, we restricted our baseline "A and B" sample to states that were part of the original 13 colonies. All the stylized facts are robust to this restriction, except part of stylized fact 3 (the downward slope of the u-shape). We do not find that population growth for log density 0 is significantly larger than for log density 2. But this finding is not surprising, since only two MCDs fall in the category of log population density 0 in this restricted sample. When we further restrict our sample to A states within the 13 colonies (New York and New England, except Maine), the remaining stylized facts all hold, except that we find no significant mean reversion in the agricultural subsample (stylized fact 5). But this is probably again due to small sample size. There are only 78 observations (in 48 counties) in the agricultural subsample for A states that were part of the original colonies (out of 4439 observations for this sample), reflecting the relatively urban character of these states.

[24] When two or more cities and their surrounding areas overlapped, we merged them together.

their surrounding suburbs. As shown in Column (6) of Table 1 and Panel D of Figure 7, all of our stylized facts hold in this suburban sample. We also experimented with other levels of aggregation, including defining "cities" as MCDs with 50,000 or more inhabitants in 1880 and using a distance threshold of 50 kilometers, and again found a similar pattern of results.

As a further robustness check, we examined whether the upward-sloping relationship between population growth and initial population density observed in Figure 2 for densities in between 2 and 4 is robust to restricting the sample to MCDs with an above median distance to one of our "cities." Re-estimating our non-parametric specification (2) for this subsample, in which the distance to a "city" is greater than 170 kilometers, we continue to find a strong upward-sloping and highly statistically significant relationship between population growth and initial population density for log densities in between 2 and 4. Therefore commuting and suburbanization in the neighborhood of large cities do not appear to be driving the pattern of violations of Gibrat's Law observed in our data.[25]

Although we examine the relationship between population growth and initial population density to control for variation in land area across MCDs, existing research concentrates on the relationship between population growth and initial population size. A further concern is therefore that the violation of Gibrat's Law could be driven by the use of initial population density rather than initial population size. To address this concern, Column (7) of Table 1 and Panel E of Figure 7 display results using log initial population size. Given that log population is measured in different units from log population density, we do not expect the inflection point at which the population growth relationship switches from being downward-sloping to upward-sloping relationship to occur at the same numerical values, and therefore the statistical tests based on values of 0, 2 and 4 in Table 1 do not apply to this specification. Nonetheless, we observe the same qualitative pattern, and each of our stylized facts holds if we use initial log population size instead of initial log population density.

A final concern is that the observed relationship between population growth and initial population density could be influenced by omitted locational fundamentals, such as institutions and natural endowments. While institutions and endowments are captured in our model in so far as they influence location-specific productivities in the agricultural and non-agricultural sector, the empirical concern is that locational fundamentals have a direct effect on population growth and are correlated

[25] While suburbanization is primarily associated with the use of the automobile as a form of mass transit, we also note that we find a very similar pattern of empirical results for the period 1880-1940, prior to the large-scale dissemination of the automobile after the end of the Second World War.

with initial population density. To explain our results, these locational fundamentals would need to have a non-linear relationship with population growth and initial population density, to have the same non-linear relationship with the share of agricultural employment and initial population density, and to have differential effects on the correlation between employment growth and initial employment in the agricultural and non-agricultural sectors. To provide evidence that such a direct effect of locational fundamentals is not driving our results, we first regress each of our left-hand side variables (population growth, the share of agriculture in employment, and employment growth in agriculture and non-agriculture) on state fixed effects (to control for state policies and institutions) and on measures of proximity to natural endowments (rivers, lakes and coastlines, and mineral endowments).[26] We next take the residuals from these regressions and implement our tests for Gibrat's Law (stylized fact 2), the share of agriculture in employment (stylized fact 3) and the relationship between employment growth and initial employment in agriculture and non-agriculture (stylized facts 5 and 6). As shown in Column (8) of Table 1 and Panel F of Figure 7, these four stylized facts are robust to controlling for locational fundamentals.[27]

Taken together, the evidence presented in this section shows that our stylized facts are robust characteristics of the US growth experience in the 20th Century. But are they also relevant for more recent experiences of structural transformation in other countries? To shed more light on this issue, we next examine urbanization and structural transformation in Brazil.

# 3 Brazilian Data and Stylized Facts

## 3.1 Data and Samples

The most populous country in the Western Hemisphere after the US is Brazil. Like the US, Brazil is divided into states. And just as US states are divided into counties, Brazilian states are divided into municipalities. Since municipality boundaries have changed over time, the Instituto de Pesquisa Econômica Aplicada (IPEA) has created "áreas mínimas comparáveis" (AMCs), geographic units that are much more stable over time. The 5,507 municipalities that existed in 1997 were pooled

[26]As an additional robustness check to further address the concern about institutional differences, we also re-estimated our baseline specification for the subset of the A states that were part of the original 13 colonies. Within this subset of the A states, MCDs are towns and townships with similar administrative functions. Once again, we find a similar pattern of results, as discussed in footnote 23 above.

[27]As the relationship between population and locational fundamentals can change over time, and as the relationship between employment and location fundamentals can differ between the agricultural and non-agricultural sectors, we do report standard deviations for log population and employment after controlling for locational fundamentals (stylized facts 1 and 4).

into 3,659 AMCs, which allow us to consistently analyze data from 1970-2000.[28] Although we could analyze Brazilian data before 1970, this would entail considerable further aggregation of municipalities, which would make it harder to distinguish urban from rural areas. Therefore we choose 1970 as the starting point for our analysis. It is worth noting that agriculture's share in employment in the average AMC declined from 71 percent to 43 percent from 1970-2000, and its share in overall employment fell from 46 percent to 20 percent. In other words, the period we analyze involved considerable structural transformation.

The average Brazilian AMC spans 2,323$km^2$, with a population of 25,817 in 1970 and 46,421 in 2000. While AMCs are on average larger than the units that we analyze in our US sample, the difference is due in part to the fact that the interior regions of Brazil have larger and more sparsely populated AMCs. Therefore, while our baseline sample uses all of Brazil, we also demonstrate the robustness of our results to using a restricted sample that includes the Northeast, Southeast and South regions in Brazil only. In these areas, the average AMC spans 923$km^2$, and had a population of 26,013 in 1970 and 44,125 in 2000. These units are still substantially bigger than in the US sample, suggesting that it might be harder to separate urban from rural areas in Brazil.

## 3.2 Stylized Facts

Having described Brazilian AMCs, we now examine whether their population dynamics are characterized (at least qualitatively) by the same stylized facts as for US MCDs. Panel A in Figure 8 and Table 2 shows that the standard deviation of log population density across Brazilian AMCs increased from 1970-2000, confirming our first stylized fact. Additionally, Panel B in the same Figure and Table shows that low density areas and high density areas grew faster than areas of intermediate density. Therefore the U-shaped relationship between population growth and initial population density, characterized in stylized fact 2, also holds for Brazil. One quantitative difference between Brazil and the US is, however, that the increasing segment of this U-shape is not 2-4 (as in the US), but rather 4-6. This difference partly reflects differences in the relative distribution of agricultural and non-agricultural employment in Brazil and the US, as evident in Figures 4 and 9 (Panel D).

Furthermore, Panel C in Figure 8 and Table 2 shows that the increasing segment of the U-shaped population growth relationship is located in the same range of initial population densities where a sharp decline in agriculture's share of employment is observed, as in the US (stylized fact 3). This

[28] New municipalities were created after 2000, but the 1997 municipalities were used in the 2000 Census, the latest Census that we analyze in this paper.

provides further corroborating evidence that the U-shape is indeed related to employment structure. Panel D in Figure 8 and Table 2 also confirms that agricultural employment has a lower standard deviation than non-agricultural employment (stylized fact 4). Finally, the last two stylized facts - that agricultural employment is mean reverting and non-agricultural employment is uncorrelated with initial density, are also confirmed for Brazil, as shown most clearly in the final two panels of Table 2 and also in Figure 8.[29]

In summary, we find a striking similarity in the relationship between population growth and employment structure in Brazil and the United States. This similarity of the results in two quite different contexts and time periods suggests that our results are unlikely to be driven by idiosyncratic features of the data and institutional environment for an individual country but rather capture more systematic features of the relationship between urbanization and structural transformation.

# 4 The Model

In this section we develop a simple theoretical model that generates the main stylized features of population growth found in our empirical work and shows how they can be explained by the process of structural transformation from agriculture to non-agriculture. The model is a natural extension of existing research on the distribution of population across space (Eeckhout 2004) to incorporate a distinction between agriculture and non-agriculture. The population and employment structure of locations are determined by their productivities in each sector, which evolve stochastically over time. Residential and commercial land use provide forces for the dispersion of population and employment, while productivity differences and agglomeration forces in non-agriculture provide forces for the concentration of population and employment. Structural transformation away from agriculture occurs as a result of faster productivity growth in the agricultural sector and inelastic demand between the two goods.

## 4.1 Preferences and Endowments

Time is discrete and is indexed by $t$. The economy consists of locations $i \in \{1, ..., I\}$, which are grouped in our data into larger statistical units called MCDs. Each location is endowed with a quantity of land $H_i$, which can be used residentially or commercially. We denote the population of

[29] For Brazil, to ensure a sufficient sample size, we construct the non-agricultural subsample using AMCs that have an agricultural employment share in 1970 of less than less than 0.4 (instead of less than 0.2 for the US). Nonetheless, if we also use a threshold of less than 0.2 for Brazil, we continue to find no statistically significant relationship between non-agricultural employment growth and initial population density.

each location by $S_{it}$ and the total population of the economy by $S_t = \sum_{i \in I} S_{it}$. While the total population of the economy is a parameter of the model, the population of each location is determined endogenously through labor mobility across locations. Workers are infinitely-lived and are endowed with one unit of labor, which is supplied inelastically with zero disutility, so that employment equals population for each location.

Workers' derive utility from consumption of goods, $C_{it}$, and residential land use, $h_{Uit}$, and for simplicity we assume that the utility function takes the Cobb-Douglas form:[30]

$$U\left(C_{it}, h_{Uit}\right) = C_{it}^{\alpha} h_{Uit}^{1-\alpha}, \qquad 0 < \alpha < 1, \tag{4}$$

The goods consumption index, $C_{it}$, includes consumption of agriculture, $c_{Ait}$, and non-agriculture, $c_{Nit}$, and is assumed to take the constant elasticity of substitution (CES) form:

$$C_{it} = \left[\psi_{At} c_{Ait}^{\rho} + \psi_{Nt} c_{Nit}^{\rho}\right]^{1/\rho}, \qquad 0 < \kappa = \frac{1}{1-\rho} < 1, \quad \psi_{At}, \psi_{Nt} > 0, \tag{5}$$

where $\psi_{At}$ and $\psi_{Nt}$ are preference parameters that capture the relative strength of consumer preferences for the agricultural and non-agricultural goods. Following the macroeconomics literature on unbalanced growth (e.g. Ngai and Pissarides 2007), we assume that agricultural and non-agricultural consumption are complements, so that the elasticity of substitution between the two goods, $\kappa$, is strictly less than one.[31]

## 4.2 Production Technology

The non-agricultural and agricultural goods are produced under conditions of perfect competition and are assumed to be costlessly tradeable across locations. The two sectors differ in terms of their production technology and the stochastic process for the evolution of productivity. Output of the non-agricultural good, $Y_{Nit}$, depends on labor input, $L_{Nit}$, land input, $H_{Nit}$, the location's productivity parameter, $\theta_{Nit}$, and a positive local externality that reflects agglomeration forces in

[30] For empirical evidence using US data in support of the constant housing expenditure share implied by the Cobb-Douglas functional form, see Davis and Ortalo-Magne (2008).

[31] The assumption of an elasticity of substitution between agriculture and non-agriculture of less than one is consistent with empirical findings of larger changes in nominal consumption shares than in real consumption shares (see for example Kravis et al. 1983).

the non-agricultural sector, $S_{it}^{\eta}$:[32]

$$Y_{Nit} = S_{it}^{\eta}\theta_{Nit}L_{Nit}^{\mu}H_{Nit}^{1-\mu}, \qquad 0 < \mu < 1,\ 0 < \eta < 1. \tag{6}$$

Non-agricultural productivity in each location evolves stochastically as follows:

$$\theta_{Nit} = \Gamma_{Nt}\left(1 + \sigma_{Nit}\right)\theta_{Nit-1}, \tag{7}$$

where $\Gamma_{Nt}$ is a component of non-agricultural productivity that is common across locations and so captures secular changes in technology over time. In contrast, $\sigma_{Nit}$ is a component of non-agricultural productivity that is idiosyncratic to individual locations, which is assumed to be independently and identically distributed with mean zero, and bounded support satisfying $1 + \sigma_{Nit} > 0$.

Output of the agricultural good, $Y_{Ait}$, depends on labor input, $L_{Ait}$, land input, $H_{Ait}$, and the location's productivity parameter, $\theta_{Ait}$:[33]

$$Y_{Ait} = \theta_{Ait}L_{Ait}^{\gamma}H_{Ait}^{1-\gamma}, \qquad 0 < \gamma < \mu < 1, \tag{8}$$

where we make the natural assumption that agriculture is land intensive: $\gamma < \mu$. Agricultural productivity in each location evolves stochastically as follows:

$$\theta_{Ait} = \Gamma_{At}\left(1 + \sigma_{Ait}\right)\theta_{Ait-1}^{\nu}, \qquad 0 < \nu < 1, \tag{9}$$

where $\Gamma_{At}$ is a component of agricultural productivity that is common across locations and captures secular changes in technology. The parameter $\sigma_{Ait}$ is a component of agricultural productivity that is idiosyncratic to individual locations, which is assumed to be independently and identically distributed with mean zero, and bounded support satisfying $1 + \sigma_{Ait} > 0$.

The idiosyncratic variation in $\{\sigma_{Nit},\, \sigma_{Ait}\}$ across locations reflects differences in their suitability for production in a sector given their natural endowments and the current production technology. To match the observed differences in employment dynamics between the agricultural and non-agricultural sectors, we assume that the stochastic process for productivity in agriculture is mean-reverting, $0 < \nu < 1$, whereas the stochastic process for productivity in non-agriculture exhibits constant proportional growth. This combination of assumptions is consistent with the view

[32] One interpretation of these agglomeration forces is knowledge spillovers that are increasing in the total population of a location, though other interpretations are also possible. In contrast to Eeckhout (2004), there is no negative commuting externality and so all of location's labor can be used productively. The reason for this choice of model structure is that diminishing marginal returns to labor already provide a force for the dispersion of population even without a negative commuting externality. Introducing such a negative externality would merely strengthen the model's dispersion forces without substantively altering the predictions of the model.

[33] As seems reasonable, we assume that there are no agglomeration forces in the agricultural sector, although all our results require is that agglomeration forces in agriculture are weaker than those in non-agriculture.

that the relative productivity of locations is more constrained by the persistent physical geography of soil and climate in agriculture than in non-agriculture.

We assume for simplicity that land allocated to commercial use in each location can be employed in either agricultural or non-agricultural production, but cannot be simultaneously employed in both. Therefore, as each location specializes completely in either the agricultural or the non-agricultural good, labor input in a sector is either equal to zero or the location's population: $L_{Nit} = S_{it}$ and $L_{Ait} = 0$ or $L_{Nit} = 0$ and $L_{Ait} = S_{it}$.

## 4.3 Consumer Equilibrium

Workers and firms are perfectly mobile across locations and can therefore relocate instantaneously and at zero cost. After observing the vector of agricultural and non-agricultural productivity shocks in period $t$, $\sigma_{At}$ and $\sigma_{Nt}$, each worker chooses their location to maximize their discounted stream of utility. As relocation is costless, this problem reduces to the static problem of maximizing their instantaneous flow of utility. Each worker chooses location, $i_t \in \{1, ..., I\}$, consumption of the agricultural good, $c_{Ait}$, consumption of the non-agricultural good, $c_{Nit}$, and residential land use, $h_{Uit}$, to maximize their utility taking the population distribution as given. The first-order conditions to this maximization problem imply that the worker allocates the constant shares of expenditure $\alpha$ and $(1-\alpha)$ to goods consumption and residential land use respectively:

$$C_{it} = \frac{\alpha \pi_{it}}{P_t}, \qquad h_{Uit} = \frac{(1-\alpha)\,\pi_{it}}{r_{it}},$$

where $\pi_{it}$ denotes the representative worker's income; $r_{it}$ is the rental rate on land; $P_t$ denotes the dual goods price index, $P_t = \left[\psi_{At}^{\kappa} p_{At}^{1-\kappa} + \psi_{Nt}^{\kappa} p_{Nt}^{1-\kappa}\right]^{\frac{1}{1-\kappa}}$, which with costless trade in goods is the same across locations $i$ at a given point in time $t$.

From the goods consumption index (5), the equilibrium share of goods consumption expenditure allocated to agriculture, $\lambda_{At}\left(p_{Nt}/p_{At}\right)$, and non-agriculture, $1 - \lambda_{At}\left(p_{Nt}/p_{At}\right)$, depends on relative prices:

$$\lambda_{At}\left(\frac{1}{p_{At}}\right) = \frac{p_{At} c_{Ait}}{P_t C_{it}} = \left[1 + \left(\frac{\psi_{Nt}}{\psi_{At}}\right)^{\kappa} \left(\frac{1}{p_{At}}\right)^{1-\kappa}\right]^{-1}, \tag{10}$$

where we have chosen the non-agricultural good as the numeraire so that $p_{Nt} = 1$.

We assume that expenditure on land is redistributed lump sum to the workers residing in a location. Therefore aggregate income equals the total value of production (including all payments to labor and land used in production) plus aggregate expenditure on residential land use:

$$\pi_{it} S_{it} = p_{Jt} Y_{Jit} + (1-\alpha)\,\pi_{it} S_{it} = \frac{p_{Jt} Y_{Jit}}{\alpha}. \tag{11}$$

where we use $J \in \{A, N\}$ to indicate land use in location $i$.

Finally, the relative price of the agricultural good is determined by the requirement that the economy-wide share of the agricultural good in revenue equals its economy-wide share in goods consumption expenditure:

$$\frac{p_{At} \sum_{i \in I_{At}} Y_{Ait}}{p_{At} \sum_{i \in I_{At}} Y_{Ait} + \sum_{i \in I_{Nt}} Y_{Nit}} = \lambda_{At} \left( \frac{1}{p_{At}} \right) \tag{12}$$

where $I_{Jt} \subset I$ denotes the subset of locations specializing in sector $J$ at time $t$.

## 4.4 Equilibrium Land Use and Factor Prices

With perfectly competitive factor markets, both labor and land are paid their value marginal product, and payments to labor and land exactly exhaust the value of output. Whether commercial land is used in agriculture or non-agriculture is determined by whichever sector offers the higher value marginal product:

$$J_{it} = \begin{cases} A & \text{if } (1-\mu)\,\theta_{Nit} S_{it}^{\eta+\mu} H_{Nit}^{-\mu} < (1-\gamma)\, p_{At} \theta_{Ait} S_{it}^{\gamma} H_{Ait}^{-\gamma} \\ N & \text{otherwise} \end{cases}, \tag{13}$$

where we have used complete specialization, and where we solve for commercial land use in the non-agricultural sector, $H_{Nit}$, and commercial land use in the agricultural sector, $H_{Ait}$, below.

The equilibrium rental rate for land in each location, $r_{it}$, is determined by the requirement that total land demand equals land supply in that location, and hence in general varies across locations. The utility function (4) implies that workers devote a constant share of their income to expenditure on residential land, while the production function (8) implies that firms devote a constant share of their revenue to expenditure on commercial land. Combining these results with total income (11), the requirement that total land demand equals land supply implies that the equilibrium rental rate for land in each location is as follows:

$$r_{it} = \begin{cases} \frac{[(1-\alpha)+(1-\mu)\alpha]\pi_{it} S_{it}}{H_i} & \text{if } J_{it} = N \\ \frac{[(1-\alpha)+(1-\gamma)\alpha]\pi_{it} S_{it}}{H_i} & \text{if } J_{it} = A \end{cases}. \tag{14}$$

The allocation of land to residential and commercial use can be determined by combining land market clearing (14), the share of worker's expenditure devoted to residential land use, and the share of firm revenue devoted to commercial land use. As agriculture and non-agriculture have different land intensities, the equilibrium allocation of land to residential and commercial use depends on which good is produced:

$$H_{Kit} = \chi_{KJ} H_i. \tag{15}$$

where $K \in \{U,\ A,\ N\}$ indexes the allocation of land to residential, agricultural and non-agricultural use respectively and $J \in \{A,\ N\}$ indexes which good is produced. Thus $\chi_{KJ}$ denotes the equilibrium fraction of land allocated to activity $K$ when good $J$ is produced. The explicit expressions for $\chi_{KJ}$ as a function of model parameters alone are provided in the appendix.

## 4.5 Equilibrium Location Choices

Perfect mobility of labor implies that upon the realization of the productivity shocks in agriculture and non-agriculture, workers are indifferent across alternative locations. The equilibrium distribution of population across locations is therefore determined by the requirement that all workers obtain the same indirect utility:

$$\frac{\pi_{it}}{P_t^{\alpha} r_{it}^{1-\alpha}} = \frac{\pi_{kt}}{P_t^{\alpha} r_{kt}^{1-\alpha}} = V_t,$$

for all locations $i$ and $k$ populated in equilibrium.[34] While the income of the representative consumer, $\pi_{it}$, and the rental rate on land, $r_{it}$, in each location depend on the good produced, the population mobility condition holds for all populated locations irrespective of the good produced.

Substituting for the equilibrium rental rate for land (14) and for total income (11), the population mobility condition can be re-written in the following form:

$$\frac{\pi_{it}}{[1-\mu\alpha]^{\frac{1-\alpha}{\alpha}}} \left(\frac{S_{it}}{H_i}\right)^{-\frac{(1-\alpha)}{\alpha}} = P_t V_t^{1/\alpha} = \frac{\pi_{kt}}{[1-\gamma\alpha]^{\frac{1-\alpha}{\alpha}}} \left(\frac{S_{kt}}{H_k}\right)^{-\frac{(1-\alpha)}{\alpha}}, \tag{16}$$

for a non-agricultural location $i$ and an agricultural location $k$.

In order to rule out a degenerate population distribution in which all workers concentrate in a single location, we require that the model's dispersion forces from diminishing marginal returns to labor in production and residential land use are sufficiently strong relative to the agglomeration forces in non-agriculture. A sufficient condition to rule out such a degenerate population distribution can be derived from (16) and is $\eta < (1-\mu) + (1-\alpha)/\alpha$. As we observe a non-degenerate distribution of population in the data, we focus on parameter values for which this inequality is satisfied.

[34] As noted above, all MCDs in our baseline sample of "A and B" states have non-zero population in all three years of our sample. As we assume for simplicity that workers reside in the location where they are employed, one way of generating zero population in some locations in the model is to assume that these locations are unsuitable for production and therefore have zero productivity in both sectors.

## 4.6 Population Density and Growth

The equilibrium population density of each non-agricultural location, $S_{Nit}$, can be determined from the population mobility condition (16), equilibrium income (11), the non-agricultural production technology (6), the equilibrium fractions of land allocated to residential and commercial use (15), and complete specialization:

$$\frac{S_{Nit}}{H_i} = \Lambda_{Nt}^{\xi_N} \theta_{Nit}^{\xi_N} H_i^{\eta \xi_N}, \qquad \xi_N \equiv \frac{1}{(1-\mu) + \frac{1-\alpha}{\alpha} - \eta} > 0, \tag{17}$$

where $\Lambda_{Nt}$ is constant across all non-agricultural locations and is defined in the appendix. Similarly, the equilibrium population density of each agricultural location, $S_{Ait}$, is:

$$\frac{S_{Ait}}{H_i} = \Lambda_{At}^{\xi_A} \theta_{Ait}^{\xi_A}, \qquad \xi_A \equiv \frac{1}{(1-\gamma) + \frac{1-\alpha}{\alpha}} > 0, \tag{18}$$

where $\Lambda_{At}$ is constant across all agricultural locations and is defined in the appendix.

From (17) and (18), the equilibrium population of a location is increasing in its productivity and land area. Furthermore, as agriculture is land-intensive relative to non-agriculture, $\gamma < \mu$, and non-agriculture exhibits agglomeration forces, $\eta > 0$, a non-agricultural location has a higher population for a given land area and productivity than an agricultural location. Combining (17) and (18) with the stochastic processes for productivity in the two sectors, (7) and (9), the model replicates our empirical findings that population growth in non-agricultural locations is uncorrelated with initial population density, while population growth in agricultural locations is negatively correlated with initial population density:

$$\ln\left(\frac{S_{Nit}}{S_{Nit-1}}\right) = \ln\left(\frac{S_{Nit}/H_i}{S_{Nit-1}/H_i}\right) = \vartheta_{Nt} + \xi_N \ln\left(1+\sigma_{Nit}\right), \tag{19}$$

$$\ln\left(\frac{S_{Ait}}{S_{Ait-1}}\right) = \ln\left(\frac{S_{Ait}/H_i}{S_{Ait-1}/H_i}\right) = \vartheta_{At} + \xi_A \ln\left(1+\sigma_{Ait}\right) - (1-\nu)\ln\left(\frac{S_{Ait-1}}{H_i}\right), \tag{20}$$

where $0 < \nu < 1$; $\vartheta_{Nt}$ and $\vartheta_{At}$ are constant across locations that produce a given good in both $t$ and $t-1$, and explicit expressions for them are provided in the appendix.

While the relationships in (19) and (20) are for locations that remain specialized in a given good over time, a key feature of our model is that it allows for endogenous switches in land use. By far the most frequent transitions observed in our data are from agricultural to non-agricultural land use. Combining equilibrium population densities for the two sectors, (17) and (18), with the stochastic process for non-agricultural productivity (7), we obtain the following expression for population growth in locations undergoing such a transition in land use:

$$\ln\left(\frac{S_{Nit}}{S_{Ait-1}}\right) = \vartheta_{NAt} + \xi_N \ln\left(1+\sigma_{Nit}\right) + \ln\left(\frac{\theta_{Nit-1}^{\xi_N}}{\theta_{Ait-1}^{\xi_A}}\right) + \eta\xi_N \ln\left(H_i\right), \tag{21}$$

where $\vartheta_{NAt}$ is constant across locations that switch from agricultural to non-agricultural land use and its explicit expression is provided in the appendix.

Transitions in land use in the model are driven by exogenous changes in technology, through the parameters $\sigma_{Nit}$, $\sigma_{Ait}$, $\Gamma_{Nt}$, and $\Gamma_{At}$, as well as the resulting endogenous changes in the relative price of the agricultural good, $p_{At}$. Transitions from agricultural to non-agricultural land use occur when changes in technology and relative prices cause the value marginal product of land in the non-agricultural sector to rise above its value marginal product in the agricultural sector in (13). Since initial relative productivity in the two sectors, $\theta_{Nit-1}/\theta_{Ait-1}$, determines initial specialization in agriculture, and since initial productivity in agriculture, $\theta_{Ait-1}$, determines initial population density, population growth for locations undergoing a transition in land use in (21) is in general correlated with initial population density.

## 4.7 Structural Transformation

The model provides a natural explanation for the preponderance of land use transitions from agriculture to non-agriculture in terms of more rapid technological progress in agriculture than in non-agriculture.[35] To illustrate this, suppose that there is a secular increase in productivity in agriculture that is common across all locations (a rise in $\Gamma_{At}$), while the common component of productivity in non-agriculture remains constant ($\Gamma_{Nt}$ unchanged).

To evaluate the impact of this change in relative productivity, we begin by considering its impact at the initial allocation of factors of production across sectors and locations. From the production technologies, (6) and (8), the secular increase in productivity in agriculture leads at the initial allocation of factors of production to a proportionate increase in the relative supply of agriculture. From the goods market clearing condition (12) and the equilibrium expenditure share (10), this increase in the relative supply of agriculture in turn leads to a decline in its relative price. Furthermore, as agriculture and non-agriculture are complements, the decline in relative price is more than proportionate to the increase in relative supply and hence more than proportionate to the increase

[35] This hypothesis of uneven technological progress across sectors is one of the central explanations for structural transformation in the macroeconomics and economic history literatures: see for example Baumol (1967), Galor (2005), Galor and Weil (2000), Matthews et al. (1982), Ngai and Pissarides (2007), and Rogerson (2008).

in productivity in agriculture. From the population mobility condition (16), this combination of increased productivity in agriculture with a more than proportionate decline in its relative price leads to a reallocation of employment from agriculture to non-agriculture. The theory appendix characterizes this process of structural transformation in closed form for the two location case.

Although not directly captured by our analysis so far, the model can also be extended to accommodate the other main explanation for structural transformation in the macroeconomics and economic history literatures: rising real incomes and non-homothetic preferences.[36] Non-homothetic preferences can be incorporated into the model by allowing the weighting parameters for non-agriculture and agriculture, $\psi_N$ and $\psi_A$, in the goods consumption index to be functions of real income, as for example in Sato (1977). In particular, suppose that $\psi_N = \psi(V_t)$ and $\psi_A = 1 - \psi(V_t)$, where $\psi : \Re^+ \rightarrow (0, 1)$ is a non-decreasing function of real income, $V_t$. Under this assumption, the share of consumer expenditure allocated to non-agriculture at a given relative price (10) is non-decreasing in real income.

Now, suppose that there is a secular increase in productivity in both sectors (a rise in $\Gamma_{At}$ and $\Gamma_{Nt}$) that raises the common value of real income across all locations. To evaluate the impact of such an increase in real income, we again begin by considering its impact at the initial allocation of factors of production across sectors and locations. From the goods market clearing condition (12) and the equilibrium expenditure share (10), a rise in $\psi_N$ and a fall in $\psi_A$ for given output of the two goods reduces the equilibrium relative price of the agricultural good, which in turn reduces income in agricultural locations relative to non-agricultural locations. From the population mobility condition (16), this change in relative prices and incomes leads to a reallocation of employment from agriculture to non-agriculture. The theory appendix again characterizes this process of structural transformation in closed form for the two location case.

## 4.8 MCD Population Growth

While our analysis of population dynamics so far has focused on locations, the MCDs observed in our data are aggregations of locations, and so can contain mixtures of agricultural and non-agricultural locations. Population growth for the MCD as a whole can be written as a weighted average of population growth in locations that produce the non-agricultural good in both periods, population growth in locations that produce the agricultural good in both periods, population growth

[36] See for example Echevarria (1997), Gollin et al. (2002), and Matsuyama (2002).

in locations that switch from agriculture to non-agriculture, and population growth in locations that switch from non-agriculture to agriculture:

$$\begin{aligned}\frac{S_{mt}-S_{mt-1}}{S_{mt-1}} \quad &= \quad \omega_{NNmt-1}\frac{S_{NNmt}-S_{NNmt-1}}{S_{NNmt-1}}+\omega_{AAmt-1}\frac{S_{AAmt}-S_{AAmt-1}}{S_{AAmt-1}} \\ &\quad +\omega_{ANmt-1}\frac{S_{ANmt}-S_{ANmt-1}}{S_{ANmt-1}}+\omega_{NAmt-1}\frac{S_{NAmt}-S_{NAmt-1}}{S_{NAmt-1}},\end{aligned} \tag{22}$$

where $\omega_{NNmt-1}$ denotes the share of locations that produce the non-agricultural good in both time periods in the MCD's population at time $t-1$; $S_{NNmt}$ is the total population at time $t$ of all locations within the MCD that produce the non-agricultural good in both time periods; the other variables are defined analogously; and $\omega_{NNmt-1}+\omega_{AAmt-1}+\omega_{ANmt-1}+\omega_{NAmt-1}=1$.

From equation (22), the correlation between population growth and initial population density across MCDs depends partly on the relationship between population growth in each of the four groups of locations and initial population density, and also depends on the relationship between the initial population share of each group and initial population density. To characterize the way in which initial population shares are related to initial population density, note that idiosyncratic productivity shocks in the non-agricultural and agricultural sectors, (7) and (9) respectively, are drawn from fixed distributions with bounded support. Together with the assumption that agricultural productivity exhibits mean reversion, this implies that agricultural productivity has bounded support. In contrast, non-agricultural productivity exhibits constant proportional growth and therefore is unbounded from above (although it is bounded below by zero as $1+\sigma_{Nit}>0$). Therefore the differences in the stochastic processes for productivity between the two sectors imply that the most densely-populated MCDs have on average higher shares of locations that produce the non-agricultural good. This positive correlation between non-agricultural specialization and initial population density is reinforced in the model by the fact that agriculture is more land-intensive than non-agriculture, $\gamma<\mu$, and non-agriculture exhibits agglomeration forces, $\eta>0$. Together these assumptions imply a higher population for a given land area and productivity when a location produces the non-agricultural good than when it produces the agricultural good (equations (17) and (18)).

We are therefore in a position to characterize the relationship between population growth and initial population density across MCDs. On the one hand, for MCDs containing only locations that produce the non-agricultural good in both time periods, the analysis of the previous section implies that population growth is uncorrelated with initial population density. Therefore the model replicates our finding that population growth in the most densely-populated MCDs is largely consistent with Gibrat's Law. On the other hand, for MCDs containing only locations that produce the agricultural good in both time periods, the analysis of the previous section implies that population growth is negatively correlated with initial population density. Therefore, at low population densities where the agricultural sector dominates, the model replicates our finding of mean reversion and a downward-sloping relationship between population growth and population density.

In between these two extremes, the positive relationship between the share of locations producing the non-agricultural good and initial population density, combined with higher employment growth in non-agriculture than agriculture as a result of structural transformation, implies that population growth must at some point become increasing in initial population density until the MCD contains exclusively locations that produce the non-agricultural good in both time periods. Therefore the model also replicates the upward-sloping relationship between population growth and initial population density observed at intermediate densities.

### 4.9 Structural Transformation and the Six Stylized Facts

The model provides a parsimonious explanation for the six stylized facts and explains how they are related. This explanation comprises differences in the stochastic processes for productivity in agriculture and non-agriculture and structural transformation away from agriculture. On the one hand, constant proportional growth in non-agricultural productivity generates the lack of correlation between non-agricultural employment growth and initial population density (stylized fact 6). On the other hand, mean reversion in agricultural productivity gives rise to the decreasing relationship between agricultural employment growth and initial population density (stylized fact 5).

These differences in the stochastic processes for productivity growth in the two sectors in turn imply that non-agricultural productivity is unbounded from above, whereas agricultural productivity is bounded from above. As a result there is a higher standard deviation of employment in non-agriculture than in agriculture (stylized fact 4) and the share of employment in agriculture is negatively correlated with population density (stylized fact 3).

Finally, the combination of mean reversion in agriculture, an agricultural employment share that is decreasing in population density and higher employment growth in non-agriculture than in agriculture generates the U-shaped relationship between population growth and initial population density (stylized fact 2). The upward-sloping segment of this U-shaped relationship in turn explains the increased dispersion of population (stylized fact 1), since along the upward-slope more densely-populated locations experience faster population growth than less-densely populated locations.

# 5 Quantitative Predictions

In this section, we use relationships from the model to provide evidence that structural transformation can account not only qualitatively but also quantitatively for the relationship between population growth and population density observed in our data. The quantitative analysis builds on four key components of the model. First, MCD population growth can be written as a weighted average of employment growth in agriculture and non-agriculture.[37] Second, the share of agricultural employment in the population is negatively correlated with population density. Third, the relationship between employment growth and population density differs between agricultural and non-agricultural locations. Fourth, the relationship between employment growth and initial population density depends on whether a location continues to produce the same good in both time periods or whether it endogenously switches between goods.

To illustrate the explanatory power of each of these components of the model, we generate a sequence of counterfactual predictions for MCD population growth, each of which uses progressively more components of the model. We next compare the predicted relationship between population growth and initial population density from each of these counterfactuals to the actual relationship observed in the data. We undertake this comparison in two ways. First, we estimate our non-parametric specification (2) and display the results for predicted and actual population growth graphically in Figure 9. Second, to provide further evidence on the predictive power of the model, we regress actual on predicted population growth and include a number of control variables. We first undertake the analysis using our US data before examining whether the model can also quantitatively account for our results using the Brazilian data. For brevity, we concentrate on results for the US data with our baseline sample of "A and B" states. However, we find a qualitatively similar pattern with the other samples, as expected from the robustness checks above, and as discussed further below.

As a first step, Counterfactual 1 uses the property that MCD population growth is a weighted average of employment growth in agriculture and non-agriculture and makes the following assump-

[37] As the model abstracts from the labor force participation decision, total employment equals total population. Therefore we undertake the quantitative analysis using employment data and compare the model's predictions to observed population growth. Empirically, we do not variation in labor force participation to a major source of differences in population dynamics across MCDs, as noted above and discussed further below.

tions: (a) a common rate of employment growth within each sector across all MCDs, (b) the same share of agricultural employment in the population across all MCDs, and (c) no switching between agriculture and non-agriculture. To measure employment growth in locations that produce the agricultural good in both periods, we use average agricultural employment growth in the agricultural sample from Table 1. Similarly, to measure employment growth in locations that produce the non-agricultural good in both periods, we use average non-agricultural employment growth in the non-agricultural sample from Table 1.[38] As Counterfactual 1 assumes both the same employment growth rates within each sector and the same agricultural employment share for all MCDs, it predicts the same rate of population growth for all MCDs, as shown in Figure 9.[39]

Counterfactual 2 is the same as Counterfactual 1, except that it allows the agricultural employment share to vary across MCDs by using the 1880 value of this variable for each MCD. Therefore, in this second counterfactual, the cross-section variation in population growth is predicted solely from the cross-section variation in the initial agricultural employment shares combined with common values of average employment growth within each sector for all MCDs. As evident from Figure 9, the employment share of an MCD in agriculture and non-agriculture in 1880 goes a good way towards explaining its population growth from 1880-2000, providing strong evidence for the importance of structural transformation in shaping observed population dynamics.

Counterfactual 3 is the same as Counterfactual 2, except that it allows for mean reversion in agriculture. To calculate average employment growth within each sector in Counterfactuals 1 and 2, we regress employment in each sector on a constant using the agricultural and non-agricultural samples from Table 1. In Counterfactual 3, we allow for mean reversion in agriculture by augmenting the employment growth regression for this sector with initial population density. The results of the regressions for agriculture and non-agriculture are reported in Columns (1) and (2) of Table 3. As shown in Figure 9, enriching the model in this way makes the downward-sloping relationship between population growth and initial population density observed at low densities more pronounced.[40]

[38] Recall that the agricultural and non-agricultural samples comprise MCDs in which agriculture accounts for more than 80 percent and less than 20 percent of MCD employment respectively.

[39] From 1880-2000, the annualized logarithmic growth rate of total population for our baseline sample of "A and B" states is 1.1 percent, as compared with a value of around 1.4 percent for the US as a whole. These figures are somewhat larger than the average annualized logarithmic growth rate across MCDs of around 0.4 percent in Figure 9 for two main reasons. First, population growth rates are correlated with population density across MCDs in Figure 9. Second total population is the weighted average of population in each MCD, and Jensen's Inequality implies that the log of a weighted average is not equal to the weighted average of the log.

[40] As a robustness check, we also augmented the non-agricultural employment growth regression with initial population density, which although not shown in Figure 9 had no visible effect, since from Table 1 employment growth is largely uncorrelated with initial population density in non-agriculture. Finally, we experimented with allowing for richer forms of scale dependence within each sector by introducing polynomials in initial population density, which

Counterfactual 4 differs from Counterfactuals 1-3, because it allows for the possibility that locations can endogenously switch from agricultural to non-agriculture, as suggested by the model. In Counterfactuals 1-3, we measured the common value of employment growth within the two sectors using employment growth in the most and least agricultural MCDs, which contain locations least likely to switch between sectors. In contrast, Counterfactual 4 takes into account the possibility of switching by allowing for a more flexible relationship between population growth and initial patterns of specialization in agriculture and non-agriculture.

In particular, in Counterfactual 4, we regress total employment growth in each MCD on the 1880 agricultural employment share, the 1880 log population density, and the interaction term between these two variables.[41] The inclusion of the initial agricultural employment share captures the role of structural transformation in shaping population growth, while the inclusion of initial log population density allows for the possibility of mean reversion in non-agriculture, and the inclusion of the interaction term between the two variables captures the extent to which mean reversion in agriculture differs from that in non-agriculture.

The regression results are reported in Columns (3) and (4) of Table 3. As expected, the initial share of agricultural employment is strongly negatively correlated with total employment growth, the initial log population density is statistically insignificant as is consistent with an absence of mean reversion in non-agriculture, and the interaction term is negatively signed and statistically significant reflecting mean reversion in agriculture. Given the lack of significance of the main effect of log population density, we exclude this variable in the final column of Table 3, which is used to generate the predicted values for total population growth for Counterfactual 4.

The results of estimating our non-parametric specification (2) using these predicted values for total population growth are shown in Figure 9 alongside the estimates using actual population growth. As apparent from the figure, actual population growth rates are substantially more variable than predicted population growth rates and the actual data exhibit a sharper change in slope than the predicted values from the counterfactual. Nonetheless, the model closely replicates the observed pattern of violations of Gibrat's Law: the downward sloping relationship between population growth and initial population density at low densities, the upward sloping relationship at intermediate densities,

also had little effect on the relationship between predicted population growth and initial population density.

[41] Since variation in labor participation rates across MCDs is not a major source of differences in population dynamics, the results of Counterfactual 4 are very similar if we replace total employment growth on the left-hand side of the regression with total population growth.

and the largely flat relationship at high densities. The mean reversion in population growth rates at low initial population densities evident in Counterfactual 3 is further enhanced in Counterfactual 4, consistent with the idea that some of the mean reversion is the result of switches from agriculture to non-agriculture. Additionally, mean predicted population growth for Counterfactual 4 is closer to mean actual population growth, because some of the higher employment growth in non-agriculture is associated with these switches in land use, which are allowed for in Counterfactual 4.

To provide further evidence on the model's ability to explain the observed patterns of population growth, and to compare its performance against alternatives, Table 4 reports the results of regressions of actual against predicted population growth from the model using our preferred Counterfactual 4. Whereas the non-parametric estimates that were displayed in Figure 9 are means for population size bins, the regressions exploit variation across individual MCDs. To provide a benchmark, we begin in Column (1) by regressing actual population growth rates on a constant. In Column (2), we augment that regression with the predicted population growth rates from the counterfactuals. Clearly there are many idiosyncratic factors affecting the population growth of individual MCDs that are not captured by our model, which results in a much larger variance of actual than of predicted population growth rates, as reflected in the regression R2. Nonetheless, the coefficient on predicted population growth is positive, highly statistically significant and statistically indistinguishable from one.[42] Therefore, despite the much greater variance in the actual population growth rates, there is a close correspondence between actual and predicted population growth.

In Columns (3) to (5) of Table 4, we report a number of robustness checks for our baseline sample of "A and B" states, in which we show that the explanatory power of the model is robust to the inclusion of a number of control variables. After including measures of proximity to natural endowments, state fixed effects and county fixed effects, we continue to find a positive coefficient on predicted population growth that is large in magnitude and statistically significant. Columns (6) to (8) take the most demanding of these specifications, including proximity to natural endowments and county fixed effects, and show that the same pattern of results holds for the more restrictive sample of A states, the county sample and the suburban sample.

As the regressions in Columns (1) through (8) of Table 4 are estimated across MCDs, they exploit in part variation across the initial population density bins used in our non-parametric specification

[42] The standard errors in Table 4 are adjusted for predicted population growth being generated in a prior regression (Pagan 1984) and clustered on county.

(2) as displayed in Figure 9. As a final step in our analysis of the model's quantitative predictions, we examine whether it can explain variation in population growth not only across but also within population density bins. In Column (9) of Table 4, we therefore augment the baseline specification from Column (2) with a full set of fixed effects for initial population density bins. Even focusing solely of variation within initial population density bins, we continue to find a positive coefficient on predicted population growth that is large in magnitude and statistically significant. Column (10) of Table 4 shows that we continue to find the same pattern of results if we further augment this specification with our measures of proximity to natural endowments and county fixed effects.

As an additional test of the model's explanatory power, the remainder of this section shows that we also find a very similar pattern of results for Brazil. The four counterfactuals are constructed in the same way for Brazil as for the United States.[43] The employment growth regressions used in these counterfactuals for Brazil are reported in Table 5 (analogous to Table 3 for the US). Having constructed the four counterfactuals, Figure 10 displays the results of estimating our non-parametric specification (2) for Brazil using actual and predicted population growth. As for the US, controlling for the initial agricultural employment share has considerable predictive power for patterns of population growth (Counterfactual 2). Controlling for mean reversion in agriculture generates the downward-sloping relationship between population growth and initial population density at low densities (Counterfactual 3). Finally, a more flexible relationship between population growth and initial patterns of specialization to allow for switches from agriculture to non-agriculture again enhances the explanatory power of the model (Counterfactual 4).

Following the same structure as for the US, Table 6 reports the results of regressions of actual against predicted population growth from the model using our preferred Counterfactual 4. While actual population growth again has a much higher variance than predicted population growth, the coefficient on predicted population growth is positive, highly statistically significant and statistically indistinguishable from one at the 5 percent level once state fixed effects are included in Column (3).[44] Therefore we again find a close correspondence between actual and predicted population growth. In Columns (4) and (5), we show that the model has explanatory power within as well as across population density bins by including a full set of fixed effects for population density bins. Finally,

[43] As noted above, to ensure a sufficient sample size, we construct the non-agricultural subsample for Brazil using AMCs that have an agricultural employment share in 1970 of less than less than 0.4 (instead of less than 0.2 for the US).

[44] Again the standard errors are adjusted for predicted population growth being generated in a prior regression (Pagan 1984).

while Columns (1)-(5) include all AMCs, we find a similar pattern of results in Columns (6)-(10), where we restrict attention to AMCs in the Northeast, Southeast and South of Brazil, which are smaller in geographic scope and are therefore likely to permit a finer discrimination between rural and urban areas.

Overall, there is considerable evidence that the model can match the quantitative as well as the qualitative patterns of population growth in both the United States and Brazil. Given the substantial differences between the countries and time periods considered, the consistency of the results provides strong corroborating evidence in support of the model.

# 6 Conclusion

While as recently as the nineteenth century less than one tenth of the world's population lived in cities, urban residents now account for a growing majority of the world's population. Arguably few other economic changes have involved as dramatic a transformation in the organization of society. In this paper, we present new evidence of six stylized facts about urbanization and develop a simple theoretical model that accounts for these facts both qualitatively and quantitatively.

Making use of a new source of sub-county data for the United States from 1880-2000, we find an unstable population distribution that exhibits polarization. This polarization reflects a population growth rate that is at first decreasing in initial population density at low densities, before increasing in initial population density at intermediate densities, and finally becoming largely uncorrelated with initial population density at high densities characteristic of urban areas.

Our model explains these systematic departures from Gibrat's Law of constant proportional growth in terms of differences in productivity dynamics between agriculture and non-agriculture. While agricultural productivity is mean reverting, due to the influence of natural endowments such as climate and soil, non-agricultural productivity exhibits constant proportional growth. Over time, faster productivity growth in agriculture than in non-agriculture and inelastic demand between the two goods leads to structural transformation and a decline in the share of agriculture in employment.

At low population densities, where agricultural employment dominates, mean reversion in agriculture generates the observed decreasing relationship between population growth and initial population density. In contrast, at high population densities, where non-agricultural employment dominates, population growth is largely uncorrelated with initial population density. In between, faster employment growth in non-agriculture than in agriculture combined with a positive correlation between the non-agricultural employment share and population density leads to the observed increasing relationship between population growth and initial population density.

This pattern of empirical results is robust across a wide range of specifications and samples. Moreover, we find a strikingly similar pattern of results for Brazil from 1970-2000 as for the United States from 1880-2000. The ability of our model to quantitatively account for our empirical findings in these two quite different contexts provides strong evidence in its support. Our findings suggest that structural transformation is not only central to macroeconomic issues, such as growth and employment, but also has important microeconomic implications for the organization of economic activity within countries. As the reallocation of population from rural to urban areas has wide-ranging implications for public policy, urbanization is likely to remain a central policy issue as large developing countries such as Brazil, China and India experience rapid structural change.

# A Data Appendix

## A.1 US Data: Sources and Variable Definitions

Most of the sources of data that we used for US MCDs come from the US Census. We also used numerous other sources, including historical maps and gazetteers, as described below.

Data on MCD employment, population, land area and location in 2000 comes from the American Factfinder of the US Census Bureau (Census 2000b). The 3 sectors we use (agriculture, manufacturing, and services) are defined using the following industry classification. "Agriculture" includes agriculture, forestry, fishing and hunting; "manufacturing" includes mining and construction as well as manufacturing; and "services" include trade, transportation, warehousing, information, finance, insurance, real estate, professional, scientific, management, administrative, education, health, arts, entertainment, accommodation and food services.

The population information for 1940 comes from the 1940 Census Files (Census 1940). Unfortunately, we have no employment data for 1940 at the MCD level, restricting their suitability for our analysis. The same files contain a full set of maps that allowed us to identify the location of 1940 MCDs.

The 1880 population and employment data come from the North Atlantic Population Project (NAPP 2006). We use the 1950 occupation and industry classifications as provided by NAPP. We classify people for whom industry information is available into 3 categories: agriculture, manufacture and services. Agricultural workers are those with industry classifications $105 - 126$, which are mainly agriculture, forestry and fishing. Manufacturing workers are those with industry classifications $206-$ $499$, and services include all other NAPP entries, except in the cases where industry was illegible, missing, not reported, or not available.

Some people identified themselves as part of the labor force, but did not report their industry, wrote it in an illegible way or were unclassifiable. These amounted to about $15$ percent of the workers classified above. In order to categorize these workers we use their self reported occupations. If we classified most of the workers in a given occupation for which we did have industry information into, say, services, we also assigned all the workers in that occupation who did not report an industry to services. While this process may have introduced some error, for the vast majority of occupations one of the three sectors of agriculture, manufacturing and services accounts for a large majority of employment.

To determine the geographic location of MCDs in 1880 we used a variety of sources. For states for which 1880 maps of MCDs were available, we georeferenced those maps. For states for which 1880 maps of MCDs were not available, we started with the aforementioned 1940 maps and worked backwards through the microfilms for the 1930, 1920, 1910, 1900 and 1890 censuses, where changes to the names and organization of MCDs are documented in footnotes. Finally, we supplemented this information with additional maps and gazetteers as reported in Appendix Table A2.

Each 1940 and 1880 MCD was allocated to a 2000 MCD using the procedure discussed in Appendix A.2 below. Finally, we created geographic control variables using maps from ESRI (1999). These geographical control variables are dummy variables equal to one if an MCD borders the ocean, if the distance between the centroid of an MCD and the closest river is less than 50 kilometers, if the distance between the centroid and the closest lake is less than 50 kilometers, and if the MCD contains coal.

## A.2 US Data: Algorithm for Linking 1880 and 1940 MCDs to 2000 MCDs

In matching MCDs from 1880, 1940, and 2000 we strove to cover all the population and land area within each state in each of the 3 censuses, while consistently matching MCDs over time. This raised 6 challenges. First, some MCDs were renamed. Second, some MCDs merged over time. Third, in some areas county boundaries were redrawn, such that MCDs were reassigned to other counties. Fourth, in some areas the census did not provide sufficient geographical information. Fifth, some MCDs were split. Sixth, in some areas MCD boundaries were redefined.

In order to deal with these challenges we aggregate some of the MCDs, and this process of aggregation required us to identify the geographic location of contemporary and historical MCDs. We started with a digital Geographic Information Systems (GIS) map from the Bureau of the Census of MCDs in 2000 (see Census [2000a]). For the earlier censuses we assigned coordinates to the MCDs ourselves, using the 1940 MCD maps provided by the Bureau of the Census (Census 1940) and a variety of historical maps and gazetteers for 1880 (see Appendix Table A2). Using these historical sources, we assigned geographic coordinates to MCDs in 1940 and 1880.[45]

To do so, we georeferenced the historical maps to the digital 2000 map using ArcGIS software. We used a conformal conic projection for the digital map, since it best fitted the historical maps from the census. We then assigned the centroids manually in a point-shapefile. In total we assigned around 22,000 coordinates for 1880 and around 50,000 for 1940. In some states we were not able to assign coordinates, and these states are not divided into sub-county units in the final dataset. The geographical distribution of these states can

[45] When assigning the coordinates for 1940 we generally used the approximate geographic centroid of the MCD, except for the case that the MCD was dominated by a single town. In this case we used the coordinates of the town. The definition of an MCD as being dominated by a single town was that the census mentioned exactly one town within the MCD.

be seen in Map 1 (these states are labeled as having "counties only" or "no data") or in Appendix Table A2. For the other states we were able to determine the location of all MCDs in 1940 given the high quality of the maps provided by the census (Census 1940). For 1880 we were are able to determine the location of the vast majority of MCDs, with the main reason for unmatched 1880 MCDs being that the digital 1880 MCD data (NAPP 2006) contained entries with missing names. Out of the 22,000 MCDs listed by NAPP for 1880, only 150 MCDs remained unmatched (see more discussion of these below).

We used the coordinates assigned to 1880 and 1940 MCDs to create geographic units that are stable over time and to which we could assign the data with reasonable confidence. To do so, we linked the 1880 and 1940 MCDs to the 2000 MCD in which their coordinates fell. In some cases multiple 1880 or 1940 MCDs fell into a single 2000 MCD, in which case we aggregated them into the single 2000 MCD.

We then merged together 2000 MCDs if they shared the same state, county and name. We did that because these were often cases where one MCD denoted a town and another denoted the surrounding area, and changes to the boundaries between the town and its surrounding area over time complicate the allocation of population to the two areas separately. This first step of the aggregation process involved 1163 aggregations.

In the second step of the aggregation process, we aggregated some MCDs to the county level (using 1880 and 1940 county definitions). We did so in counties for which we could not find all the 1880 MCDs on the map due to missing names. This second step involved the aggregation of 85 counties.

In the third step of the aggregation process, concerned 2000 MCDs that had not been matched to both 1880 and 1940 MCDs (we will refer to them as "uncovered MCDs"). These uncovered MCDs can exist either if an older MCD was split (such that there are multiple MCDs in 2000 where there used to be one), or if boundaries were redrawn. In both cases we used proximity as a guide to solving the problem of uncovered MCDs. We determined the location of all 2000 centroids, and matched each uncovered MCDs to the closest 2000 MCD within the same county (by 1880 definition) that was not uncovered. Overall we had about 6,000 uncovered MCDs that we matched with the other 19,000 non-uncovered ones.

Finally, we manually aggregated some additional units to deal with changes in municipal boundaries. We merged the MCDs Bronx, Brooklyn, Manhattan, Queens, and Staten Island, since they all constitute parts of New York City. We also merged Saint Louis, Missouri, with its neighboring county, from which it split off at one point. Finally, we merged the MCDs Peoria and West Peoria in Illinois.

## A.3 US Data: Samples

This section discusses in further detail the samples that we use to analyze US population dynamics. Our baseline sample comprises "A and B" states (10,864 observations), and we also use a sample of "A" states (4,439 observations), a county sample (2,496 observations), a hybrid sample (19,229 observations) and a suburban sample (10,674 observations). In discussing these samples in this section of the appendix, we refer to the geographical units created by the aggregation of MCDs using the procedure above as "units of analysis."

The "A and B" sample consists of states in which the ratio of the number of "units of analysis" to the number of MCDs in both 1880 and 2000 is larger than 0.7. This restricts the extent to which we aggregate MCDs (a process that may involve imprecisions due to changes in boundaries), while maintaining a sizeable number of states. This sample consists of 15 states (plus Washington DC), most of which are found in the North-East and Mid-West of the US, as shown in Map 1 and listed in Table A1.

The "A" sample is more restrictive: it only uses states for which both ratios above are larger than 0.9. These are the states in which there is a close correspondence between the 1880, 1940 and 2000 MCDs. This sample includes 8 states and Washington DC, and apart from Indiana and Iowa all of these are part of the original 13 colonies, as shown in Map 1 and listed in Table A1.

The county sample tackles the problem of representativeness, by expanding the number of states that we use. The tradeoff is that in this sample we analyze data at a higher level of spatial aggregation. We exclude Alaska, Hawaii and Oklahoma, which were not included in the 1880 census. We also exclude North and South Dakota, which had not attained statehood in 1880, and therefore did not have stable county boundaries at that time. For all other states, 1880 and 1940 counties are linked to 2000 counties using the centroids of the 1880 and 1940 counties.

The hybrid sample combines MCD and county data, and uses for each state the smallest unit for which we have data – MCDs in 30 states and counties in the remaining states.

In the suburban sample we pooled together cities with their suburban areas to form larger units. To define city boundaries, we mark units with population larger than 100,000 in 1880 as city centres. Then we draw a radius of 25 kilometers around each, and combine these "city" MCDs with all other MCDs whose centroids fall within the circle. Overlapping circles are further aggregated. In our baseline "A and B" sample, the number of observations is reduced by 190 units (from 10,864 to 10,674) as a result of this procedure. As a further robustness test, we also repeated the analysis using other thresholds, such as defining a "city" based on an 1880 population of larger than 50,000 and using a distance threshold of 50 kilometers.

## A.4 Brazilian Data: Sources and Variable Definitions

Our data for Brazil uses as units of analysis 3659 "áreas mínimas comparáveis" (AMCs), groups of municipalities that are broadly consistent from 1970-2000. The Brazilian data we use mostly come from Instituto de Pesquisa Econômica Aplicada (IPEA), and Brazilian Census micro data compiled by Instituto Brasileiro de Geografia e Estatística (IBGE). Data on AMC employment and population in both 1970 and 2000 comes from the Brazilian census data (Brazil Census 1970, 2000). Data on AMC land area in 2000 comes from IPEA (2008).

While obtaining population data is straightforward, calculating sectoral employment involved some choices in classification of workers into agriculture, manufacturing, and services. In particular, in the US the logging sector is not considered part of agriculture, but in Brazil it proved more difficult to consistently separate logging from the rest of the agricultural sector for both 1970 and 2000. We therefore pooled the Brazilian logging industry with its agricultural sector. Using 1970 industry definitions, we classified people employed in industries 111-222 as agricultural workers, those in industries 300-352 as manufacturing workers, and those in industries 411-928 as service workers. In 2000, agricultural workers were those with industry classifications 01101-05002, manufacturing workers are those with industry classifications 10000-37000 or 45001-45999, and services workers are those with industry classifications 40010-41000 or 50010-93092.

To analyze Brazilian population dynamics, we also use a subsample of only AMCs in the states of the Northeast, Southeast, and South official regions of Brazil, since AMCs in these regions are relatively small, allowing a clearer distinction between rural and urban areas. It is also less likely that those areas were not fully settled in 1970. The three regions in this subsample cover about 90 percent of Brazil's AMCs, 36 percent of its land area and 91 percent of its population in 1970. In some of the robustness checks we also use a set of state fixed effects. To generate these fixed effects we use the 2000 classification of 27 Brazilian states, although we note that some state boundaries did change during the period since 1970. In particular, Mato Grosso do Sul was separated from Mato Grosso in the 1970s; Guanabara and Rio de Janeiro merged in 1975 under the name of Rio de Janeiro; and Tocantins was formed in 1988 out of the northern part of Goiás.

In addition to state fixed effects, some of our specifications also use a range of geographic controls. These include indicators for substantial mineral deposits of oil, nickel, manganese, iron, gold, copper, cobalt, and aluminum (bauxite). We also construct an indicator for whether an AMC borders on the ocean, or whether its centroid lies within 50 kilometers of a river. Finally, we construct a variable indicating if an AMC's centroid is covered with tropical or subtropical moist broadleaf forest, or for if it is situated in the Amazonas area. The river shapefile is from ArcView Database Access (ESRI 1999). The broadleaf forest, minerals, and oil and gas shapefiles are from the GlobalGIS DVD (GIS 2003).

# B Theory Appendix

## B.1 Theoretical Derivations

The equilibrium fractions of land that are used residentially and commercially in each location as a function of whether the agricultural or non-agricultural good is produced (equation (15)) are:

$$\begin{array}{lll} \chi_{UN} = \left[\frac{(1-\alpha)}{(1-\mu)\alpha+(1-\alpha)}\right] & \chi_{AN} = 0 & \chi_{NN} = \left[\frac{(1-\mu)\alpha}{(1-\mu)\alpha+(1-\alpha)}\right] \\ \chi_{UA} = \left[\frac{(1-\alpha)}{(1-\gamma)\alpha+(1-\alpha)}\right] , & \chi_{AA} = \left[\frac{(1-\gamma)\alpha}{(1-\gamma)\alpha+(1-\alpha)}\right] , & \chi_{NA} = 0 \end{array} .$$

In equations (17) and (18):

$$\Lambda_{Nt} \equiv \chi_{NN}^{1-\mu} / \left( \alpha \left[ V_t P_t^{\alpha} \right]^{1/\alpha} \left[ 1 - \mu\alpha \right]^{\frac{1-\alpha}{\alpha}} \right)$$

$$\Lambda_{At} \equiv \left( \left( p_{At} \left( \chi_{AA} \right)^{1-\gamma} \right) / \left( \alpha \left[ V_t P_t^{\alpha} \right]^{1/\alpha} \left[ 1 - \gamma\alpha \right]^{\frac{1-\alpha}{\alpha}} \right) \right)$$

In equations (19), (20), and (21):

$$\vartheta_{Nt} \equiv \ln \left( \left( V_t / V_{t-1} \right)^{-\xi_N/\alpha} \left( P_t / P_{t-1} \right)^{-\xi_N} \Gamma_{Nt}^{\xi_N} \right)$$

$$\vartheta_{At} \equiv \ln \left( \left( \frac{\chi_{AA}^{1-\gamma}}{\alpha \left[ (1-\gamma\alpha) \right]^{\frac{1-\alpha}{\alpha}}} \right)^{\xi_A (1-\nu)} \left( \frac{p_{At}}{p_{At-1}^{\nu}} \right)^{\xi_A} \left( \frac{V_t}{V_{t-1}^{\nu}} \right)^{-\xi_A/\alpha} \left( \frac{P_t}{P_{t-1}^{\nu}} \right)^{-\xi_A} \Gamma_{At}^{\xi_A} \right)$$

$$\vartheta_{NAt} \equiv \ln \left( p_{At-1}^{-\xi_A} \left( \frac{\chi_{NN}^{(1-\mu)\xi_N}}{\chi_{AA}^{(1-\gamma)\xi_A}} \right) \left( \frac{\left( \alpha \left[ (1-\gamma\alpha) \right]^{\frac{1-\alpha}{\alpha}} \right)^{\xi_A}}{\left( \alpha \left[ 1-\mu\alpha \right]^{\frac{1-\alpha}{\alpha}} \right)^{\xi_N}} \right) \left( \frac{V_t^{-\xi_N/\alpha}}{V_{t-1}^{-\xi_A/\alpha}} \right) \left( \frac{P_t^{-\xi_N}}{P_{t-1}^{-\xi_A}} \right) \Gamma_{Nt}^{\xi_N} \right)$$

## B.2 Results for the Two Location Case

In the case where the economy consists of two locations, $i \in \{1, 2\}$, the process of structural transformation from agriculture to non-agriculture can be characterized in closed form. The specification for the consumption goods index satisfies the Inada conditions, and therefore both the agricultural and non-agricultural good will be consumed in equilibrium, which requires both goods to be produced in equilibrium. Since commercial land in each location can be allocated to either non-agricultural production or agricultural production but not both, it follows that one location will specialize in non-agriculture and the other location will specialize in agriculture.

The pattern of specialization is determined by the two locations' relative productivities in non-agriculture and agriculture and their relative supplies of land. Without loss of generality, we suppose that location 1 specializes in non-agriculture and location 2 specializes in agriculture, which can be ensured by the appropriate choice of relative productivities for the two locations in the two sectors. With complete specialization in each location, the allocation of labor to non-agriculture also corresponds to the allocation of labor to location 1, and similarly the allocation of labor to agriculture corresponds to the allocation of labor to location 2.

In the two location case, general equilibrium in the economy can be characterized by the following three relationships. First, we require that the population mobility condition (16) holds, which can be written as:

$$\begin{aligned} S_{N1t}^{\eta-(1-\mu)-\frac{(1-\alpha)}{\alpha}} H_1^{(1-\mu)+\frac{(1-\alpha)}{\alpha}} &= \Phi \frac{\chi_{AA}^{1-\gamma}}{\chi_{NN}^{1-\mu}} \frac{p_{At}\theta_{A2t}}{\theta_{N1t}} \left(S_t - S_{N1t}\right)^{-(1-\gamma)-\frac{(1-\alpha)}{\alpha}} H_2^{(1-\gamma)+\frac{(1-\alpha)}{\alpha}}, \quad (23) \\ \Phi &\equiv \left(\frac{[1-\mu\alpha]}{[(1-\gamma\alpha)]}\right)^{\frac{1-\alpha}{\alpha}}, \end{aligned}$$

where we have used equilibrium income (11), the equilibrium allocation of land within each location (15), and labor market clearing. Recall that $\eta < (1-\mu) + (1-\alpha)/\alpha$.

Second, we require that goods markets clear for the economy as a whole, which requires the share of the agricultural good in revenue to equal the share of the agricultural good in expenditure:

$$\frac{1}{1+\left(\frac{1}{p_{At}}\right)\left(\frac{Y_{N1t}}{Y_{A2t}}\right)} = \frac{1}{1+\left(\frac{\psi_{Nt}}{\psi_{At}}\right)^{\kappa}\left(\frac{1}{p_{At}}\right)^{1-\kappa}}, \qquad 0 < \kappa < 1, \tag{24}$$

where we have used equilibrium expenditure shares (10).

Third, relative output of non-agriculture and agriculture is determined by the production technologies:

$$\frac{Y_{N1t}}{Y_{A2t}} = \frac{\theta_{N1t} S_{N1t}^{\eta+\mu} \chi_{NN}^{1-\mu} H_1^{1-\mu}}{\theta_{A2t}\left(S_t - S_{N1t}\right)^{\gamma} \chi_{AA}^{1-\gamma} H_2^{1-\gamma}}, \tag{25}$$

where we have again used equilibrium land use within each location (15) and labor market clearing.

Now consider a secular increase in productivity in non-agriculture relative to agriculture that is common across all locations: that is a rise in $\Gamma_{At}/\Gamma_{Nt}$ and hence in $\theta_{A2t}/\theta_{N1t}$. We begin by considering the impact of this change in relative productivity at the initial equilibrium allocation of labor between sectors and locations and at the initial equilibrium prices. From the production technology (25), the increase in relative productivity in agriculture implies a decrease in the relative supply of the non-agricultural good. Next consider the impact of this change in relative supply on relative prices at the initial allocation of labor between sectors and locations. From the goods market clearing condition (24), the decrease in the relative supply of the non-agricultural good requires a rise in the relative price of the non-agricultural good in order to re-equate relative demand and supply. Furthermore, since non-agriculture and agriculture are complements, $0 < \kappa < 1$, the rise in the relative price of the non-agricultural good must be more than proportionate to the decrease in the relative supply of the non-agricultural good. Finally, consider the combined impact of the change in technology and relative goods prices on the equilibrium allocation of labor between the two sectors and locations. From the population mobility condition (23), a decrease in relative productivity in non-agriculture combined with a more than proportionate increase in the relative price of non-agriculture leads to a reallocation of labor from agriculture to non-agriculture and from location 2 to location 1.

Additionally, as discussed in the main text, the model can be extended to incorporate non-homothetic preferences by allowing the weighting parameters for agriculture and non-agriculture in the CES goods consumption index, $\psi_A$ and $\psi_N$, to depend on real income. When combined with technological progress that raises the common value of real income across all locations, such non-homothetic preferences provide a complementary explanation for structural transformation. Suppose for example that technological change in both sectors increases real income and reduces the strength of consumer preferences for agriculture relative to non-agriculture, which corresponds to a rise in $\psi_N/\psi_A$ in the goods market clearing condition (24). By a similar line of reasoning to that used above, this rise in $\psi_N/\psi_A$ leads a reallocation of labor from agriculture to non-agriculture and from location 2 to location 1.

While the two location example allows us to characterize structural transformation in closed form, it cannot of course capture a change in the number of locations producing each goods, since locations are completely specialized, and both goods are consumed and produced in equilibrium. Therefore one location produces the non-agricultural good, while the other location produces the agricultural good. To incorporate changes in the number of locations producing each good, requires us to move to the general case of the model with arbitrary numbers of locations, as considered in the main text above.

# References


Bairoch, P (1988) "Cities and Economic Development: From the Dawn of History to the Present", University of Chicago Press.

Baumol, William J. (1967) “Macroeconomics of Unbalanced Growth: The Anatomy of Urban Crises,” *American Economic Review*, 57(3), 415-26.

Beeson, Patricia E., David N. DeJong and Werner Troesken (2001) “Population growth in U.S. counties, 1840–1990,” *Regional Science and Urban Economics*, 31(6), 669-699.

Black, Duncan and Vernon Henderson (1999) “A Theory of Urban Growth,” *Journal of Political Economy*, 107(2), 252-284.

Black, Duncan and Vernon Henderson (2003) “Urban Evolution in the USA,” *Journal of Economic Geography*, 3, 343-372.

Brazil Census (1970, 2000), Integrated Public Use Microdata Series — International (IPUMS International)

Caselli, Francesco and Wilbor John Coleman II (2001) “The U.S. Structural Transformation and Regional Convergence: A Reinterpretation,” *Journal of Political Economy*, 109(3), 584-616.

Census (1940), US Census Bureau, Sixteenth US Census of Population and Housing,
*www.census.gov/prod/www/abs/decennial/1940.htm*

Census (2000a), US Census Bureau, US Census of Population and Housing, Shapefile downloaded from
*www.census.gov/main/www/cen2000.html*

Census (2000b), US Census Bureau, American Factfinder, available at
*factfinder.census.gov*

Census (2000c), US Census Bureau, Bureau’s Geographic Areas Reference Manual, downloaded from
*http://www.census.gov/geo/www/garm.html*.

Census (2000d), US Census Bureau, "Census 2000 Basics", Chapter 4: Geographic Areas,
*http://www.census.gov/mso/www/c2000basics/chapter4.htm*.

Cleveland, William S. (1979) “Robust Locally Weighted Regression and Smoothing Scatterplots,” *Journal of the American Statistical Association*, 74(368), 829-836.

Cordoba, Juan-Carlos (2008) “On the Distribution of City Sizes,” *Journal of Urban Economics*, 63(1), 177-197.

Davis, Morris A. and François Ortalo-Magne (2008) “Household Expenditures, Wages, Rents,” University of Wisconsin-Madison, mimeograph.

Davis, Donald R. and David Weinstein (2002) "Bones, Bombs, and Break Points: The Geography of Economic Activity ", *American Economic Review*, 92(5), 1269-1289.

Desmet, Klaus and Esteban Rossi-Hansberg (2007) “ Spatial Growth and Industry Age,” *NBER Working Paper*, 13302.

Duranton, Gilles (2007) “Urban Evolutions: The Fast, the Slow, and the Still,” *American Economic Review*, 97(1), 197-221.

Eaton, Jonathan and Zvi Eckstein (1997) "Cities and growth: Theory and evidence from France and Japan", *Regional Science and Urban Economics*, 27, 443-374

Echevarria, Cristina (1997) “Changes in Sectoral Composition Associated with Economic Growth,” *International Economic Review*, 38(2), 431-452.

Eeckhout, Jan (2004) “Gibrat’s Law for (All) Cities,” *American Economic Review*, 94(5), 1429-1451.

Eeckhout, Jan (2008) “Gibrat’s Law for (All) Cities: Reply,” *American Economic Review*, forthcoming.

Ellison, Glenn and Edward L. Glaeser (1999) “The Geographic Concentration of Industry: Does Natural Advantage Explain Agglomeration?” *American Economic Review*, 89(2), 311-316.

ESRI (1999), Arc View Database Access (Version 2.1a), Environmental Systems Research Institute, Inc.

Foster, Andrew D. and Mark R. Rosenzweig (2008) “Economic Development and the Decline of Agricultural Employment,” Chapter 47 in (eds) T. Schultz and John Strauss, *Handbook of Development Economics*, Volume 4(5).

Fujita, Masahisa, Paul Krugman, and Anthony J. Venables (1999) "The Spatial Economy: Cities, Regions and International Trade," Cambridge, MA: MIT Press.

Gabaix, Xavier (1999) “Zipf’s Law for Cities: An Explanation,” *Quarterly Journal of Economics*, 114(3), 739-767.

Gabaix, Xavier and Yannis Ioannides (2004) “The Evolution of City Size Distributions,” *Handbook of Regional and Urban Economics*, Volume 4, in (eds) Vernon Henderson and Jacques Thisse, North-Holland.

Galor, Oded and David Weil (2000) “Population, Technology, and Growth: From the Malthusian Regime to the Demographic Transition and Beyond,” *American Economic Review*, 90(4), 806-828.

Galor, Oded (2005) “From Stagnation to Growth: Unified Growth Theory,” in (eds) Philippe Aghion and Steven Durlauf, *Handbook of Economic Growth*, Volume 1A.

GIS (2003), Global GIS DVD-Global Coverage DVD 1st ed. U.S. Geological Data Series DDS-62 A-H Data Collections and Custom Programming the United States Geological Survey Published by the American Geological Institute, copyright 2003.

Goodfriend, Marvin and John McDermott (1995) “Early Development,” *American Economic Review*, 85(1), 116-133.

Glaeser, Edward L. (2008) “The Rise of the Sunbelt,” *Southern Economic Journal*, 74(3), 610-643.

Glaeser, Edward L., Hedi D. Kallal, Jose A. Scheinkman and Andrei Shleifer (1992) “Growth in Cities,” *Journal of Political Economy*, 100(6), 1126-1152.

Gollin, Douglas, Stephen Parente and Richard Rogerson (2002) “The Role of Agriculture in Development,” *American Economic Review*, 92(2), 160-164.

González-Val, Rafael, Luis Lanaspa and Fernando Sanz (2008) “New Evidence on Gibrat’s Law for Cities,” *MPRA Working Paper*, 10411.

Härdle, Wolfgang (1990) "Applied Nonparametric Regression", Econometric Society Monographs, Cambridge University Press: Cambridge.

Henderson, J. Vernon (1974) “The Sizes and Types of Cities,” *American Economic Review*, 64(4), 640-56.

Henderson, J. Vernon (2005) “Urbanization and Growth,” in (eds) Philippe Aghion and Steven Durlauf, *Handbook of Economic Growth*, North Holland.

Henderson, J. Vernon and Anthony J. Venables (2008) “The Dynamics of City Formation,” *NBER Working Paper*, 13769.

Henderson, J. Vernon and Hyoung Gun Wang (2007) “Urbanization and City Growth: the Role of Institutions,” *Regional Science and Urban Economics*, 37(3), 283-313.

Ioannides, Yannis and Henry G. Overman (2004) “Spatial Evolution of the U.S. Urban System,” *Journal of Economic Geography*, 4(2), 131–156.

IPEA (2008) - Instituto de Pesquisa Econômica Aplicada, Data Section, available at *http://www.ipeadata.gov.br*

Kim, Sukkoo (1995) “Expansion of Markets and the Geographic Distribution of Economic Activities: The trends in U.S.,” *Quarterly Journal of Economics*, 110(4), 881-908.

Kim, Sukkoo (2000) “Urban Development in the United States, 1690-1990,” *Southern Economic Journal*, 66(4), 855-880.

Kim, Sukkoo and Robert A. Margo (2004) "Historical Perspectives on U.S. Economic Geography" in (eds) Vernon Henderson and Jacques Thisse, *Handbook of Regional and Urban Economics*, Volume 4, North-Holland.

Kravis, Irving B., Alan W. Heston and Robert Summers (1983) "The Share of Services in Economic Growth," in F. Gerard Adams and Bert G. Hickman, eds., *Global Econometrics: Essays in Honor of Lawrence R. Klein*, Cambridge: MIT Press, 188-218.

Krugman, Paul (1991) "Increasing Returns and Economic Geography," *Journal of Political Economy*, 99(3), 483-499.

Kuznets, Simon (1955) "Economic Growth and Income Inequality," *American Economic Review*, 45, 1-28.

Levy, Moshe (2008) "Gibrat's Law for (All) Cities: A Comment," *American Economic Review*, forthcoming.

da Mata, Daniel, Uwe Deichmann, J. Vernon Henderson, Somik V. Lall and Hyoung G. Wang (2007) "Determinants of City Growth in Brazil," *Journal of Urban Economics*, 62(2), 252-272.

Matthews, Robert C., Charles H. Feinstein and John C. Odling-Smee (1982) "British Economic Growth 1856-1973", Stanford: Stanford University Press

Matsuyama, Kiminori (2002) "The Rise of Mass Consumption Societies," *Journal of Political Economy*, 110, 1035-1070.

Mckinsey (2008) "Preparing for China's Urban Billion", Mckinsey Global Institute: New York.

Ngai, Rachel and Chris Pissarides (2007) "Structural Change in a Multisector Model of Growth," *American Economic Review*, 97(1), 429-443.

NAPP 2006, North Atlantic Population Project and Minnesota Population Center. *NAPP: Complete Count Microdata*. NAPP Version 1.0 [computer files]. Minneapolis, MN: Minnesota Population Center [distributor], 2006. [*www.nappdata.org*]

Pagan, Adrian (1984) "Econometric Issues in the Analysis of Regressions with Generated Regressors," *International Economic Review*, 25(1), 221-247.

Prescott, Samuel T. (2003) "Federal Land Management - Current Issues and Background", Nova Publishing

Rappaport, Jordan and Jeffrey Sachs (2003) "The United States as a Coastal Nation," *Journal of Economic Growth*, 8(1), 5-46.

Rogerson, Richard (2008) "Structural Transformation and the Deterioration of European Labor Market Outcomes", *Journal of Political Economy*, 116(2), 235-259.

Rosen, Kenneth T. and Mitchel Resnick (1980) "The Size Distribution of Cities: An Examination of the Pareto Law and Primacy", *Journal of Urban Economics*, 8(2), 165-186.

Rosenthal, Stuart S. and William C. Strange (2004) "Evidence on the Nature and Sources of Agglomeration Economies," *Handbook of Urban and Regional Economics*, (eds) J. Vernon Henderson and Jacques-Francois Thisse.

Rossi-Hansberg, Esteban and Mark L. Wright (2007) "Urban Structure and Growth," *Review of Economic Studies*, 74(2), 597-624.

Sato, Ryuzo (1977) "Homothetic and Non-Homothetic CES Production Functions," *American Economic Review*, 67(4), 559-569.

Scott, David W. (1979) "On Optimal and Data-based Histograms," *Biometrika*, 66(3), 605-610.

Silverman, B. W. (1986) "Density Estimation for Statistics and Data Analysis", Chapman and Hall: London.

Simon, Herbert A. (1955) "On a Class of Skew Distribution Functions," *Biometrika*, 42(3/4), 425–440.

Soo, Kwok Tong (2005) "Zipf's Law for Cities: A Cross-country Investigation", *Regional Science and Urban Economics*, 35(3), 239-263.

Soo, Kwok Tong (2007) "Zipf's Law and Urban Growth in Malaysia," *Urban Studies*, 44(1), 1-14.

Syrquin, Moshe (1988) "Patterns of Structural Change," Chapter 7 in (eds) H. Chenery and T. N. Srinivasan, *Handbook of Development Economics*, 1, 205-273.

United Nations (2005) "Word Urbanization Prospects - The 2005 Revision," published October 2006, accessible via *www.un.org/esa/population/publications/WUP2005/2005wup.htm.*

Williamson, Jeffrey G. (1965) "Regional Inequality and the Process of National Development: A Description of the Patterns," *Economic Development and Cultural Change*, 13(1), 3-45.

# Table 1: US – Robustness of stylized facts

| | | (1) Baseline: A and B states | (2) Only A states | (3) Counties, 45 states and DC[1] | (4) Counties, A and B sample | (5) Hybrid Sample, 45 states and DC[2] | (6) Suburban A and B states[3] | (7) Log pop, not log density | (8) Baseline with geo controls[4] |
|---|---|---|---|---|---|---|---|---|---|
| Panel A | Standard deviation of log population density in 1880 ($\sigma_1$) | 0.967 | 1.025 | 1.757 | 0.963 | 1.272 | 0.932 | 0.833 | . |
| | Standard deviation of log population density in 2000 ($\sigma_2$) | 1.556 | 1.631 | 1.450 | 1.303 | 1.687 | 1.484 | 1.475 | . |
| | $H_0$: $\sigma_1 = \sigma_2$, vs. $H_1$: $\sigma_1 < \sigma_2$, p-value | <0.001 | <0.001 | 1.000 | <0.001 | <0.001 | <0.001 | <0.001 | . |
| | <u>Stylized Fact 1</u>: Distribution of log population density across geographic units became more dispersed from 1880-2000 (population became more concentrated) | Yes | Yes | No[5] | Yes | Yes | Yes | Yes | .[4] |
| Panel B | Mean population growth at log population density 0 ($\beta_g(0)$) | 0.013 | 0.012 | 0.016 | 0.019 | 0.010 | 0.013 | . | 0.013 |
| | Mean population growth at log population density 2 ($\beta_g(2)$) | 0.001 | -0.001 | 0.007 | 0.007 | 0.002 | 0.001 | . | 0.005 |
| | Mean population growth at log population density 4 ($\beta_g(4)$) | 0.009 | 0.010 | 0.014 | 0.014 | 0.011 | 0.008 | . | 0.011 |
| | $H_0$: $\beta_g(0) = \beta_g(2)$, $H_1$: $\beta_g(0) > \beta_g(2)$, p-value | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | . | <0.001 |
| | $H_0$: $\beta_g(2) = \beta_g(4)$, $H_1$: $\beta_g(2) < \beta_g(4)$, p-value | <0.001 | <0.001 | 0.001 | 0.011 | <0.001 | <0.001 | . | <0.001 |
| | <u>Stylized Fact 2</u>: U-shaped relationship between population growth from 1880-2000 and log population density in 1880 | Yes | Yes | Yes | Yes | Yes | Yes | .[6] | Yes |
| Panel C | Percent of agricultural in total employment at log population density 2 ($\beta_{sa}(2)$) | 0.767 | 0.762 | 0.691 | 0.618 | 0.738 | 0.767 | . | 0.743 |
| | Percent of agricultural in total employment at log population density 4 ($\beta_{sa}(4)$) | 0.227 | 0.189 | 0.195 | 0.185 | 0.228 | 0.220 | . | 0.236 |
| | $H_0$: $\beta_{sa}(2) = \beta_{sa}(4)$, $H_1$: $\beta_{sa}(2) > \beta_{sa}(4)$, p-value | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | . | <0.001 |
| | <u>Stylized Fact 3</u>: Share of agriculture in employment falls in the range where population density distribution in 1880 is positively correlated with population growth 1880-2000 | Yes | Yes | Yes | Yes | Yes | Yes | .[6] | Yes |
| Panel D | Standard deviation of agricultural employment in 1880 ($\sigma_{1a}$) | 0.820 | 0.722 | 1.677 | 0.810 | 1.084 | 0.815 | 0.820 | . |
| | Standard deviation of non-agricultural employment in 1880 ($\sigma_{1na}$) | 1.520 | 1.631 | 1.784 | 1.272 | 1.779 | 1.478 | 1.520 | . |
| | $H_0$: $\sigma_{1a} = \sigma_{1na}$, vs. $H_1$: $\sigma_{1a} < \sigma_{1na}$, p-value | <0.001 | <0.001 | 0.001 | <0.001 | <0.001 | <0.001 | <0.001 | . |
| | Standard deviation of agricultural employment in 2000 ($\sigma_{2a}$) | 0.858 | 0.853 | 0.806 | 0.617 | 0.936 | 0.854 | 0.858 | . |
| | Standard deviation of non-agricultural employment in 2000 ($\sigma_{2na}$) | 1.623 | 1.689 | 1.530 | 1.359 | 1.767 | 1.552 | 1.623 | . |
| | $H_0$: $\sigma_{2a} = \sigma_{2na}$, vs. $H_1$: $\sigma_{2a} < \sigma_{2na}$, p-value | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | . |
| | <u>Stylized Fact 4</u>: Standard deviation of non-agricultural employment is larger than standard deviation of agricultural employment in both years | Yes | Yes | Yes | Yes | Yes | Yes | Yes | .[4] |
| Panel E | Regress agricultural employment growth on log population density and intercept in subsample of units with agricultural employment share > 0.8 in 1880, report slope coefficient ($\beta_a$) | -0.0060 | -0.0077 | -0.0067 | -0.0054 | -0.0066 | -0.0060 | -0.0056 | -0.0055 |
| | $H_0$: $\beta_a = 0$, $H_1$: $\beta_a \neq 0$, p-value | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| | <u>Stylized Fact 5</u>: Agricultural employment does not follow Gibrat's law (employment growth depends on population density) | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Panel F | Regress non agricultural employment growth on log population density and intercept in subsample of units with non-agricultural employment share < 0.2 in 1880, report slope coefficient ($\beta_{na}$) | -0.0002 | -0.0006 | -0.0016 | -0.0006 | -0.0010 | -0.0004 | -0.0002 | -0.0005 |
| | $H_0$: $\beta_{na} = 0$, $H_1$: $\beta_{na} \neq 0$, p-value | 0.515 | 0.287 | <0.001 | 0.096 | <0.001 | 0.286 | 0.515 | 0.0954 |
| | <u>Stylized Fact 6</u>: Non-agricultural employment follows Gibrat's law (employment growth does not depend on population density) | Yes | Yes | No[7] | Yes | No[7] | Yes | Yes | Yes |

Note: This table reports robustness tests of the 6 stylized facts using US data. All the regressions and tests reported in the table use robust standard errors clustered by county.

[1] The county sample includes all US states except Alaska, Hawaii, North Dakota, Oklahoma, and South Dakota, which had not attained statehood in 1880 and did not have stable county boundaries at that time.

[2] The hybrid sample uses the smallest geographical units available for each state. We use MCDs for the states in samples A, B, and C, and counties elsewhere. This sample excludes Alaska, Hawaii, North Dakota, Oklahoma, and South Dakota, as explained in the footnote above.

[3] In the Suburban Sample we merge any MCD with more than 100,000 inhabitants in 1880 to all the MCDs whose centroids lie within 25 kilometers of its centroid.

[4] The geographic control variables are state fixed effects, an indicator for the presence of coal, and indicators for the unit bordering on the ocean and for its centroid being within 50 kilometers from a lake or a river. As these specifications include controls, we do not test stylized facts 1 and 4, which involve measuring standard deviations.

[5] Since this sample includes many states that were not fully settled in 1880, many near-empty areas increase the standard deviation of the population density distribution in that year. When we restrict the analysis to counties in states A and B only, the stylized fact does hold (see column 4). This is reassuring, since our model is concerned with long-run equilibria, which is likely to a better characterization of the longer-settled A and B states.

[6] In this sample we do not expect the turning point of the U and the fall of the agriculture share at coefficient 2, and hence do not report these coefficients. The figures qualitatively show that there is a U-shape whose minimum coincides with the drop in agricultural employment.

[7] Since this sample includes many states that were not fully settled in 1880, many near-empty areas increase the standard deviation of the population density distribution in that year. The future settlement of areas that were near empty in 1880 is also likely to cause mean reversion that is unrelated to employment structure.

# Table 2: Brazil – Robustness of stylized facts

| | | (1) All of Brazil (AMCs) | (2) As (1) but with state fixed effects | (3) As (1) but with geo controls[1] | (4) As (2) but with geo controls | (5) Brazil sub-sample[2] | (6) As (5) but with state fixed effects | (7) As (5) but with geo controls | (8) As (6) but with geo controls |
|---|---|---|---|---|---|---|---|---|---|
| Panel A | Standard deviation of log population density in 1970 ($\sigma_1$) | 1.222 | .[3] | 1.222 | .[3] | 1.009 | .[3] | 1.009 | .[3] |
| | Standard deviation of log population density in 2000 ($\sigma_2$) | 1.323 | . | 1.323 | . | 1.197 | . | 1.197 | . |
| | $H_0$: $\sigma_1 = \sigma_2$, vs. $H_1$: $\sigma_1 < \sigma_2$, p-value | <0.001 | . | <0.001 | . | <0.001 | . | <0.001 | . |
| | Stylized Fact 1: Distribution of log population density across geographic units became more dispersed from 1970-2000 (population became more concentrated) | Yes | . | Yes | . | Yes | . | Yes | . |
| Panel B | Mean population growth at log population density 0 ($\beta_g(0)$) | 0.0239 | 0.0239 | 0.0239 | 0.0239 | 0.0146 | 0.0146 | 0.0146 | 0.0146 |
| | Mean population growth at log population density 4 ($\beta_g(4)$) | 0.0079 | 0.0134 | 0.0116 | 0.0146 | 0.0079 | 0.0053 | 0.0090 | 0.0100 |
| | Mean population growth at log population density 6 ($\beta_g(6)$) | 0.0214 | 0.0271 | 0.0265 | 0.0305 | 0.0214 | 0.0190 | 0.0240 | 0.0258 |
| | $H_0$: $\beta_g(0) = \beta_g(4)$, $H_1$: $\beta_g(0) > \beta_g(4)$, p-value | <0.001 | 0.015 | 0.002 | 0.016 | <0.001 | <0.001 | 0.001 | 0.006 |
| | $H_0$: $\beta_g(4) = \beta_g(6)$, $H_1$: $\beta_g(4) < \beta_g(6)$, p-value | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| | Stylized Fact 2: U-shaped relationship between population growth from 1970-2000 and log population density in 1970 | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Panel C | Percent of agricultural in total employment in 1970 at log population density 4 ($\beta_{sa}(4)$) | 0.6710 | 0.6710 | 0.6710 | 0.6710 | 0.6710 | 0.6710 | 0.6710 | 0.6710 |
| | Percent of agricultural in total employment in 1970 at log population density 6 ($\beta_{sa}(6)$) | 0.1677 | 0.1933 | 0.1459 | 0.1689 | 0.1677 | 0.1933 | 0.1447 | 0.1686 |
| | $H_0$: $\beta_{sa}(4) = \beta_{sa}(6)$, $H_1$: $\beta_{sa}(4) > \beta_{sa}(6)$, p-value | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| | Stylized Fact 3: Share of agriculture in employment falls in the range where population density distribution in 1970 is positively correlated with population growth 1970-2000 | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Panel D | Standard deviation of agricultural employment in 1970 ($\sigma_{1a}$) | 0.8933 | .[3] | 0.8933 | .[3] | 0.8869 | .[3] | 0.8869 | .[3] |
| | Standard deviation of non-agricultural employment in 1970 ($\sigma_{1na}$) | 1.4157 | . | 1.4157 | . | 1.4287 | . | 1.4287 | . |
| | $H_0$: $\sigma_{1a} = \sigma_{1na}$, vs. $H_1$: $\sigma_{1a} < \sigma_{1na}$, p-value | <0.001 | . | <0.001 | . | <0.001 | . | <0.001 | . |
| | Standard deviation of agricultural employment in 2000 ($\sigma_{2a}$) | 1.0176 | . | 1.0176 | . | 0.9954 | . | 0.9954 | . |
| | Standard deviation of non-agricultural employment in 2000 ($\sigma_{2na}$) | 1.3754 | . | 1.3754 | . | 1.3642 | . | 1.3642 | . |
| | $H_0$: $\sigma_{2a} = \sigma_{2na}$, vs. $H_1$: $\sigma_{2a} < \sigma_{2na}$, p-value | <0.001 | . | <0.001 | . | <0.001 | . | <0.001 | . |
| | Stylized Fact 4: Standard deviation of non-agricultural employment is larger than standard deviation of agricultural employment in both years | Yes | . | Yes | . | Yes | . | Yes | . |
| Panel E | Regress agricultural employment growth on log population density and intercept in subsample of units with agricultural employment share > 0.8 in 1970, report slope coefficient ($\beta_a$) | -0.0038 | -0.0036 | -0.0022 | -0.0037 | -0.0042 | -0.0031 | -0.0028 | -0.0031 |
| | $H_0$: $\beta_a = 0$, $H_1$: $\beta_a \neq 0$, p-value | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| | Stylized Fact 5: Agricultural employment does not follow Gibrat's law (employment growth depends on population density) | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Panel F | Regress non agricultural employment growth on log population density and intercept in subsample of units with agricultural employment share < 0.4 in 1970, report slope coefficient ($\beta_{na}$) | 0.00126 | 0.00176 | 0.00027 | 0.00090 | 0.0013 | 0.00156 | 0.00030 | 0.00059 |
| | $H_0$: $\beta_{na} = 0$, $H_1$: $\beta_{na} \neq 0$, p-value | 0.124 | 0.0503 | 0.758 | 0.342 | 0.108 | 0.074 | 0.729 | 0.521 |
| | Stylized Fact 6: Non-agricultural employment follows Gibrat's law (employment growth does not depend on population density) | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |

Note: This table reports robustness tests of the 6 stylized facts using data on Brazilian municipalities (Áreas Mínimas Comparáveis (AMCs)). All the regressions and tests reported in the table use robust standard errors.

[1] The geographic controls are twelve dummy variables indicating the presence of oil, nickel, manganese, iron, gold, copper, cobalt, and aluminum, whether the AMC borders the ocean, lies within 50 kilometers of a river, has its centroid covered with tropical or subtropical moist broadleaf forest, or is contained in the Amazonas area.

[2] This subsample uses only AMCs in the states of the Northeast, Southeast, and South official regions of Brazil, since AMCs in these regions are relatively small, allowing a clearer distinction between rural and urban areas. The three regions in this subsample cover about 90 percent of Brazil's AMCs, 36 percent of its land area and 91 percent of its population in 1970.

[3] As these specifications include controls, we do not test stylized facts 1 and 4, which involve measuring standard deviations.

## Table 3: US – Constructing the counterfactuals

| Employment growth rate, 1880-2000 | (1)<br>For counterfactual 3<br>Non-agric. | (2)<br><br>Agric. | (3)<br>For counterfactual 4<br>Total | (4)<br><br>Total |
|---|---|---|---|---|
| Constant | 0.011<br>(0.001) | -0.005<br>(0.001) | 0.014<br>(0.001) | 0.014<br>(0.001) |
| Log population density in 1880 | | -0.006<br>(0.000) | -0.0002<br>(0.0003) | |
| Share of agriculture 1880 | | | -0.008<br>(0.002) | -0.007<br>(0.001) |
| (Share of agriculture in 1880) x (log population density in 1880) | | | -0.0010<br>(0.0005) | -0.0013<br>(0.0004) |
| Number of Observations | 755 | 3,074 | 10,856 | 10,856 |
| $R^2$ | 0 | 0.31 | 0.063 | 0.063 |
| Sample: | A and B, non-agric | A and B, agric | A and B | A and B |

Note: This table reports the regressions used to construct counterfactuals 3 and 4 for the US data. We construct counterfactual 3 using the predicted values of sectoral employment growth from the regressions reported in columns (1) and (2), as described in the text of the paper. We construct counterfactual 4 using the predicted values of employment growth from the regression reported in column (4), as described in the text of the paper. The non-agricultural subsample used in column (1) includes MCDs from our baseline A and B Sample for which agriculture's share of 1880 employment was less than 0.2. The agricultural subsample used in column (2) includes MCDs from our baseline A and B Sample for which agriculture's share of 1880 employment exceeded 0.8. Robust standard errors in parentheses are clustered by county.

## Table 4: US – Quantifying the predictive power of counterfactual 4

| | (1)<br>Inter-cept only | (2)<br>As (1) but with predicted growth | (3)<br>As (2) but with geo controls[1] | (4)<br>As (3) but with state fixed effects | (5)<br>As (4) but with county fixed effects | (6)<br>As (5) but with A Sample only | (7)<br>As (5) but with county sample | (8)<br>As (5) but with suburban sample | (9)<br>As (2) but with log pop density bins[2] | (10)<br>As (5) but with log pop density bins |
|---|---|---|---|---|---|---|---|---|---|---|
| Actual population growth regression | | | | | | | | | | |
| Predicted population growth | | 1.041<br>(0.059) | 0.799<br>(0.055) | 0.752<br>(0.058) | 0.811<br>(0.049) | 0.908<br>(0.062) | 0.952<br>(0.055) | 0.812<br>(0.057) | 0.648<br>(0.079) | 0.469<br>(0.046) |
| Intercept | 0.475<br>(0.034) | -0.026<br>(0.045) | | | | | | | | |
| $R^2$ | 0 | 0.098 | 0.183 | 0.284 | 0.246 | 0.35 | 0.546 | 0.28 | 0.151 | 0.284 |
| Number of observations | 10,864 | 10,864 | 10,864 | 10,864 | 10,864 | 4,439 | 2,496 | 10,674 | 10,864 | 10,864 |
| Regression used to generate predicted population growth | | | | | | | | | | |
| Share of agriculture in 1880 | | -1.05<br>(0.017) | -1.038<br>(0.017) | -0.982<br>(0.016) | -0.963<br>(0.014) | -0.953<br>(0.018) | -0.44<br>(0.051) | -0.83<br>(0.016) | -1.217<br>(0.046) | -1.003<br>(0.023) |
| Share of agriculture 1880 x population density 1880 | | -0.162<br>(0.006) | -0.156<br>(0.006) | -0.147<br>(0.006) | -0.14<br>(0.006) | -0.154<br>(0.008) | -0.797<br>(0.018) | -0.197<br>(0.006) | -0.077<br>(0.017) | -0.104<br>(0.009) |
| F – statistic[3] | | 5236 | 4873 | 6652 | 8151 | 5738 | 1241 | 5971 | 3266 | 8429 |

Note: This table shows the predictive power of counterfactual 4 for various specifications using US data. The upper panel of the table reports the regressions of actual population growth on predicted population growth. The lower panel of the table reports the regression whose fitted values are used for predicted population growth. The left-hand side variable in the lower panel of the table is total employment growth. Robust standard errors clustered by county are in parentheses. The standard errors in the upper panel of the table have been adjusted for the fact that predicted population growth is generated using a prior regression (Pagan 1984).

[1] The geographic control variables are state fixed effects, an indicator for the presence of coal, and indicators for observations bordering on the ocean and for observations whose centroid lies within 50 kilometers of a lake or a river.

[2] The log population density bin fixed effects included in these regressions are a full set of dummy variables for MCDs having population densities within intervals of 0.1 log points. For example, all MCDs with log population density from 0.1 to 0.2 are grouped together in bin 0.1.

[3] The F-value reported is for an F-test that the coefficients on the share of agriculture and the interaction term are jointly equal to zero in the prior regression used to generate predicted population growth.

## Table 5: Brazil – Constructing the counterfactuals

| Employment growth rate 1970-2000 | (1) For counterfactual 3 Non-agric. | (2) For counterfactual 3 Agric. | (3) For counterfactual 4 Total | (4) For counterfactual 4 Total |
|---|---|---|---|---|
| Constant | 0.039 | 0.00216 | 0.045 | 0.043 |
| | (0.001) | (0.00111) | (0.004) | (0.001) |
| Log population density in 1970 | | -0.0038 | -0.0005 | |
| | | (0.0004) | (0.0008) | |
| Share of agriculture 1970 | | | -0.0317 | -0.0291 |
| | | | (0.0044) | (0.0016) |
| (Share of agriculture in 1970) x (log population density in 1970) | | | -0.0037 | -0.0043 |
| | | | (0.0009) | (0.0004) |
| Number of Observations | 384 | 1,651 | 3,659 | 3,659 |
| $R^2$ | 0 | 0.059 | 0.262 | 0.262 |
| Sample: | AMCs non-agric. | AMCs agric. | AMCs | AMCs |

Note: This table reports the regressions we used to construct counterfactuals 3 and 4 for the Brazilian municipalities (Áreas Mínimas Comparáveis (AMCs)) data. We construct counterfactual 3 using the predicted values of sectoral employment growth from the regressions reported in columns (1) and (2), as described in the text of the paper. We construct counterfactual 4 using the predicted values of employment growth from the regression reported in column (4), as described in the text of the paper. The non-agricultural subsample used in column (1) includes AMCs for which agriculture's share of 1970 employment was less than 0.4 due to the small sample size using a threshold of 0.2 (but results are similar using a 0.2 threshold). The agricultural subsample used in column (2) includes AMCs for which agriculture's share of 1970 employment exceeded 0.8. Robust standard errors are in parentheses.

## Table 6: Brazil – Quantifying the predictive power of counterfactual 4

| | (1) Intercept only | (2) As (1) but with predicted growth | (3) As (2) but with geo controls[1] | (4) As (3) but with state fixed effects | (5) As (4) but with subsample[4] only | (6) As (2) but with log pop density bins[2] | (7) As (4) but with log pop density bins |
|---|---|---|---|---|---|---|---|
| Actual population growth | | | | | | | |
| Predicted population growth | | 1.024 | 0.968 | 1.112 | 1.122 | 0.909 | 0.915 |
| | | (0.035) | (0.035) | (0.040) | (0.042) | (0.036) | (0.036) |
| Intercept | 0.269 | 0.010 | | | | | |
| | (0.009) | (0.010) | | | | | |
| $R^2$ | 0 | 0.196 | 0.315 | 0.378 | 0.350 | 0.287 | 0.385 |
| Number of observations | 3,659 | 3,659 | 3,659 | 3,659 | 3,659 | 3,659 | 3,659 |
| Regression used to generate predicted population growth | | | | | | | |
| Share of agriculture in 1970 | | -0.810 | -0.821 | -0.755 | -0.885 | -0.693 | -0.708 |
| | | (0.013) | (0.015) | (0.015) | (0.017) | (0.044) | (0.045) |
| Share of agriculture 1970 x population density 1970 | | -0.122 | -0.125 | -0.129 | -0.073 | -0.158 | -0.159 |
| | | (0.003) | (0.003) | (0.004) | (0.005) | (0.012) | (0.012) |
| F – statistic[3] | | 5460 | 5651 | 4728 | 4088 | 4042 | 4212 |

Note: This table shows the predictive power of counterfactual 4 for various specifications using the Brazilian municipalities (Áreas Mínimas Comparáveis (AMCs)) data. The upper panel of the table reports the regression of actual population growth on predicted population growth. The lower panel of the table reports the regression whose fitted values are used for predicted population growth. The left-hand side variable in the lower panel of the table is total employment growth. Robust standard errors are in parentheses. The standard errors in the upper panel of the table have been adjusted for the fact that predicted population growth is generated using a prior regression (Pagan 1984).

[1] The geographic controls are twelve dummy variables indicating the presence of oil, nickel, manganese, iron, gold, copper, cobalt, and aluminum, whether the AMC borders the ocean, lies within 50 kilometers of a river, has its centroid covered with tropical or subtropical moist broadleaf forest, or is contained in the Amazonas area.

[2] The log population density bin fixed effects included in these regressions are a full set of dummy variables for MCDs having population densities within intervals of 0.1 log points. For example, all AMCs with log population density from 0.1 to 0.2 are grouped together in bin 0.1.

[3] The F-value reported is for an F-test that the coefficients on the share of agriculture and the interaction term are jointly equal to zero in the prior regression used to generate predicted population growth.

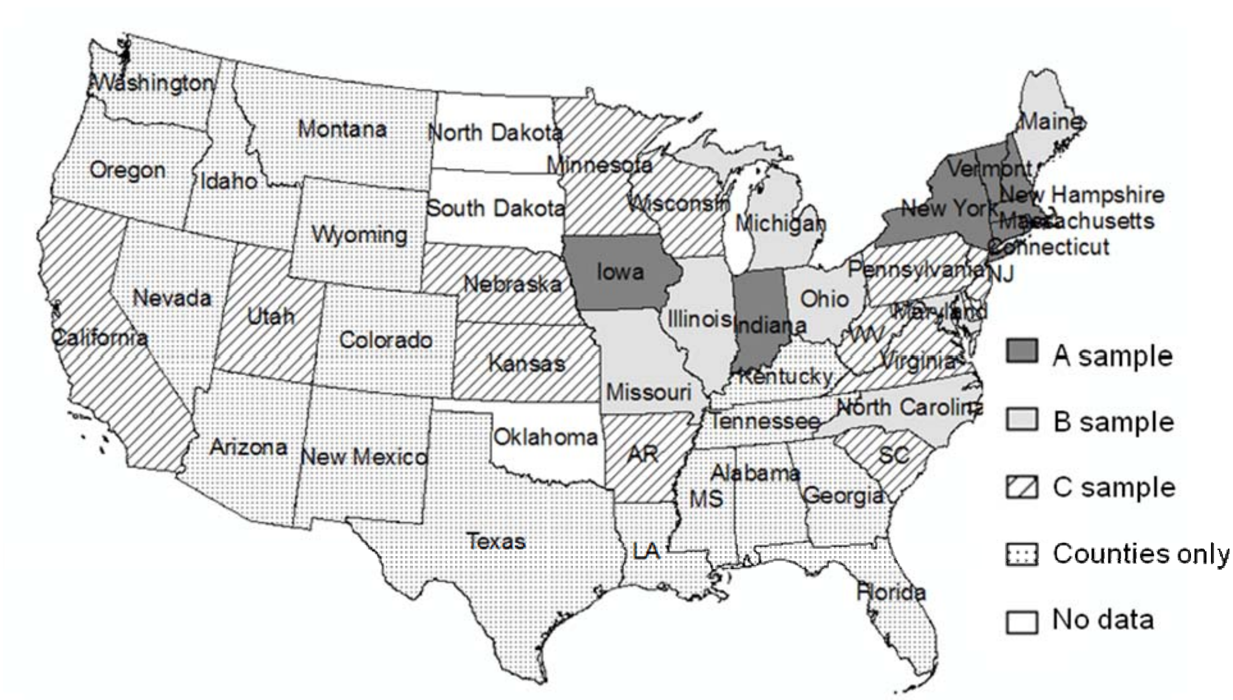
[4] This subsample uses only AMCs in the states of Northeast, Southeast, and South official regions of Brazil, since AMCs in these regions are relatively small, allowing a clearer distinction between rural and urban areas. The three regions in this subsample cover about 90 percent of Brazil's AMCs, 36 percent of its land area and 91 percent of its population in 1970.

## Appendix Table A1: US – MCD match quality by state

| | 1880 MCDs | 1940 MCDs | 2000 MCDs | Final (pooled) MCDs | Final/ 1880 | Final/ 2000 | Class- ification | Aggreg- ation Step 1 | Aggreg- ation Step 2 | Aggreg- ation Step 3 |
|---|---|---|---|---|---|---|---|---|---|---|
| Arkansas | 887 | 1482 | 1330 | 806 | 0.91 | 0.61 | C | 0 | 0 | 524 |
| California | 419 | 504 | 387 | 133 | 0.32 | 0.34 | C | 0 | 69 | 185 |
| Connecticut | 167 | 169 | 169 | 167 | 1.00 | 0.99 | A | 0 | 0 | 2 |
| Delaware | 33 | 31 | 27 | 21 | 0.64 | 0.78 | C | 0 | 0 | 6 |
| Washington DC | 1 | 1 | 1 | 1 | 1.00 | 1.00 | A | 0 | 0 | 0 |
| Georgia | 1232 | 1648 | 577 | 505 | 0.41 | 0.88 | C | 0 | 0 | 72 |
| Illinois | 1583 | 1638 | 1708 | 1446 | 0.91 | 0.85 | B | 0 | 104 | 158 |
| Indiana | 1011 | 1015 | 1009 | 997 | 0.99 | 0.99 | A | 0 | 0 | 12 |
| Iowa | 1545 | 1676 | 1654 | 1509 | 0.98 | 0.91 | A | 7 | 0 | 145 |
| Kansas | 1066 | 1686 | 1492 | 982 | 0.92 | 0.66 | C | 43 | 135 | 375 |
| Maine | 574 | 712 | 530 | 487 | 0.85 | 0.92 | B | 1 | 0 | 43 |
| Maryland | 236 | 302 | 293 | 212 | 0.90 | 0.72 | B | 0 | 15 | 66 |
| Massachusetts | 344 | 349 | 351 | 337 | 0.98 | 0.96 | A | 0 | 0 | 14 |
| Michigan | 1110 | 1428 | 1425 | 1044 | 0.94 | 0.73 | B | 104 | 0 | 381 |
| Minnesota | 1220 | 2911 | 2506 | 897 | 0.74 | 0.36 | C | 269 | 485 | 1124 |
| Missouri | 1134 | 1303 | 1379 | 1099 | 0.97 | 0.80 | B | 0 | 28 | 252 |
| Nebraska | 654 | 1506 | 1198 | 526 | 0.80 | 0.44 | C | 36 | 328 | 344 |
| New Hampshire | 245 | 249 | 258 | 236 | 0.96 | 0.91 | A | 1 | 0 | 22 |
| New Jersey | 265 | 563 | 549 | 254 | 0.96 | 0.46 | C | 17 | 0 | 295 |
| New York | 969 | 1006 | 986 | 913 | 0.94 | 0.93 | A | 27 | 21 | 48 |
| North Carolina | 871 | 1027 | 1053 | 834 | 0.96 | 0.79 | B | 2 | 0 | 219 |
| Ohio | 1373 | 1445 | 1548 | 1320 | 0.96 | 0.85 | B | 31 | 44 | 184 |
| Pennsylvania | 1995 | 2567 | 2469 | 1671 | 0.84 | 0.68 | C | 110 | 167 | 631 |
| Rhode Island | 36 | 39 | 39 | 36 | 1.00 | 0.92 | A | 0 | 0 | 3 |
| South Carolina | 411 | 574 | 296 | 246 | 0.60 | 0.83 | C | 0 | 0 | 50 |
| Utah | 219 | 422 | 90 | 55 | 0.25 | 0.61 | C | 0 | 20 | 15 |
| Vermont | 248 | 252 | 251 | 243 | 0.98 | 0.97 | A | 4 | 0 | 8 |
| Virginia | 434 | 473 | 544 | 362 | 0.83 | 0.67 | C | 0 | 17 | 165 |
| West Virginia | 326 | 352 | 240 | 208 | 0.64 | 0.87 | C | 0 | 0 | 32 |
| Wisconsin | 952 | 1808 | 1646 | 902 | 0.95 | 0.55 | C | 255 | 0 | 744 |

Note: This table shows the number of MCDs in each state in each of the three census years (1880, 1940, and 2000). It also reports the number of observations in the final dataset, and ratios of this number to 1880 MCDs and to 2000 MCDs. The table reports the classification of the match quality in each state (the classification is A if both ratios are ≥ 0.9, B if both ratios are ≥ 0.7 but one or more of them is less than 0.9, and C otherwise). The baseline sample we use consists of the A and B states. This and other samples used in robustness checks are described in the paper. The table also reports the number of MCDs aggregated in each step of the data creation process. In the first step we merged together 2000 MCDs with identical state, county and names; in the second step we pooled together MCDs in 1880 counties when we could not identify the location of some of the MCDs; and in the third step we pooled together 2000 MCDs that were not matched to data from 1880 or 1940 to the nearest 2000 MCD that lay within the same 1880 county. For a detailed discussion of the matching and aggregation process, see the Data Appendix. Five states are excluded from our analysis (Alaska, Hawaii, Oklahoma, North Dakota and South Dakota), because they had not attained statehood in 1880 and are either not included in the 1880 census or did not have stable county boundaries at that time. The remaining 16 states are included in some of our robustness checks using county data.

## Map 1: US MCD data by state and county




Note: This map shows the geographical distribution of states for the various samples. Our baseline sample consists of A and B states. The classification A, B and C corresponds to the quality of the match rate between 1880 and 2000 MCDs. In states classified as A (Connecticut, DC, Indiana, Iowa, Massachusetts, New Hampshire, New York, Rhode Island, Vermont), the 1-1 match rate between 1880 and 2000 MCDs is larger than 0.9. In states classified as B (Illinois, Maine, Maryland, Michigan, Missouri, North Carolina, Ohio), the match rate is larger than 0.7. In states classified as C (Arkansas, California, Delaware, Georgia, Kansas, Minnesota, Nebraska, New Jersey, Pennsylvania, South Carolina, Utah, Virginia, West Virginia, Wisconsin), 1880 MCD data are available but the match rate is lower than 0.7. For states in the counties sample (Alabama, Arizona, Colorado, Florida, Idaho, Kentucky, Louisiana, Mississippi, Montana, Nevada, New Mexico, Oregon, Tennessee Texas, Washington, Wyoming), 1880 MCD data are not available. We exclude Alaska, Hawaii, Oklahoma, North Dakota, and South Dakota, which had not attained statehood in 1880, and therefore are either not included in the 1880 census or did not have stable county boundaries at that time.

## Appendix Table A2: US Geographical sources by state

| State | Map sources |
|---|---|
| Alabama: | No sufficient 1880 map found. Only county data available. |
| Alaska: | Not included in the 1880 census and so excluded from the dataset. |
| Arizona: | In 1880 precincts were not separately returned by the enumerators. Only county data available. |
| Arkansas: | United States Library of Congress map collection, Rand McNally and Co: "Arkansas Administrive Railroad and Township Map", 1898. |
| California: | National Archives and Records Administration (NARA), Washington DC, Microfilm publication A3378: "Enumeration District Maps of the Twelfth trough Sixteenth Censuses of the United States, 1900 - 1940", County Maps on Microfilm Roll numbers 4 to 6. Additionally (for the counties of Colusa, Napa, Solano and Ventura): Blum, George W., California Book Map, Compiled and Published by Geo. W. Blum, 330 Pine St., S.F. Edward Denny and Co., Agents. Copyrighted 1895 By Geo. W. Blum, San Francisco, Cal., available at David Rumsey Map Collection (www.davidrumsey.com) |
| Colorado: | In 1880 precincts were not separately returned by the enumerators. Only county data available. |
| Connecticut: | Mitchell, Samuel A. "Township map of the States of Massachusetts, Connecticut and Rhode Island, Drawn and engraved by W.H. Gamble, Philadelphia. Copyright 1887 by Wm. M. Bradley and Bro. (1890)", Publisher: John Y. Huber and Co, available at David Rumsey Map Collection (www.davidrumsey.com). |
| Delaware: | Library of Congress, "The Township Map of Delaware", Mc Connell School supply company, copyright McConnell (Philadelphia), 1990. |
| Florida: | No sufficient 1880 map found. Only county data available. |
| Georgia: | National Archives and Records Administration (NARA), Washington DC, Microfilm publication A3378: "Enumeration District Maps of the Twelfth trough Sixteenth Censuses of the United States, 1900 - 1940", County Maps on Microfilm Roll numbers 11 and 12. |
| Hawaii: | Not included in the 1880 census and so excluded from the dataset. |
| Idaho: | In 1880 precincts were not separately returned by the enumerators. Only county data. |

| | |
|---|---|
| Illinois: | Library of Congress, Rufus Blanchard (cartographer), "Blanchard's township map Illinois", 1867. Additionally Mitchell, Samuel Augustus: "County and Township map of the State of Illinois", (1880) available at David Rumsey Map Collection (www.davidrumsey.com). |
| Indiana: | Representative Districts Indiana. Published by Baskin, Forster and Co. Lakeside Building Chicago, Ills. 1876. Engraved and Printed by Chas. Shober and Co. Props. of Chicago Lithographing Co.), Andreas, A. T., 1839-1900. Additionally: Gazetteer from United States Geological Survey (geonames.usgs.gov). |
| Iowa: | Sectional map of Iowa showing civil and congressional townships, all towns, post offices, railroads, streams. Compiled by D.W. Ensign, published by A.T. Andreas, Chicago, Ills., 1875. (Lakeside Building, Chicago, Ills. Engraved and printed by Chas. Shober and Co., Props. of Chicago Lithographing Co.), available at David Rumsey Map Collection (www.davidrumsey.com). |
| Kansas: | The official state atlas of Kansas compiled from government surveys, county records and personal investigations. Philadelphia. L.H. Everts and Co. 1887. Copyright, 1887, L.H. Everts and Co. (with view:), Additionally: Gazetteer from United States Geological Survey (geonames.usgs.gov). |
| Kentucky: | No sufficient map found. Only county data available. |
| Louisiana: | No sufficient map found. Only county data available. |
| Maine: | Mitchell, Samuel A. "Township map of the State of Maine", Drawn and engraved by W.H. Gamble, Philadelphia. Copyright 1887 by Wm. M. Bradley and Bro. (1890)", Publisher: John Y. Huber and Co, available at David Rumsey Map Collection (www.davidrumsey.com). |
| Maryland: | Gazetteer from United States Geological Survey (geonames.usgs.gov). |
| Massachusetts: | Mitchell, Samuel A. "Township map of the States of Massachusetts, Connecticut and Rhode Island", Drawn and engraved by W.H. Gamble, Philadelphia. Copyright 1887 by Wm. M. Bradley and Bro. (1890)", Publisher: John Y. Huber and Co, available at David Rumsey Map Collection (www.davidrumsey.com). |
| Michigan: | Mitchell, Samuel A. "Township map of the States of Michigan and Wisconsin", Drawn and engraved by W.H. Gamble, Philadelphia. Copyright 1887 by Wm. M. Bradley and Bro. (1890)", Publisher: John Y. Huber and Co, available at David Rumsey Map Collection (www.davidrumsey.com). |
| Minnesota: | United States Library of Congress, "Map of the state of Minnesota", The Anderson Publishing Company, Arthur Gibson. |
| Mississippi: | No sufficient map found. Only county data available. |
| Missouri: | Gazetteer from United States Geological Survey (geonames.usgs.gov). |
| Montana: | In 1880 precincts were not separately returned by the enumerators. Only county data available. |
| Nebraska: | The official state Atlas of Nebraska. Compiled from government surveys, county records and personal investigations. Philadelphia, Everts and Kirk, 1885. Copyright, 1885, Everts and Kirk. Additionally gazetteer from United States Geological Survey (geonames.usgs.gov). |
| Nevada: | In 1880 precincts were not separately returned by the enumerators. Only county data available. |
| New Hampshire: | County and township map of Vermont and New Hampshire. Copyright 1887 by William M. Bradley and Brother, John Y. Huber Company, Publishers, Philadelphia and St. Louis. (1890), available at David Rumsey Map Collection (www.davidrumsey.com). |
| New Jersey: | Johnson's new Illustrated (Steel Plate) Family Atlas, With Descriptions, Geographical, Statistical, And Historical. Compiled, Drawn, and Engraved Under The Supervision Of J.H. Colton And A.J. Johnson. New York: Johnson And Browning, Formerly (Successors To J.H. Colton And Company) No. 133 Nassau Street. 1860, available at David Rumsey Map Collection (www.davidrumsey.com). |
| New Mexico: | In 1880 precincts were not separately returned by the enumerators. Only county data are available. |
| New York: | Johnson's new Illustrated (Steel Plate) Family Atlas, With Descriptions, Geographical, Statistical, And Historical. Compiled, Drawn, and Engraved Under The Supervision Of J.H. Colton And A.J. Johnson. New York: Johnson And Browning, Formerly (Successors To J.H. Colton And Company,) No. 133 Nassau Street. 1860, available at David Rumsey Map Collection (www.davidrumsey.com). |
| North Carolina: | National Archives and Records Administration (NARA), Washington DC, Microfilm publication A3378: "Enumeration District Maps of the Twelfth trough Sixteenth Censuses of the United States, 1900 - 1940", County Maps on Microfilm Roll numbers 44 and 45. |
| North Dakota: | Had not attained statehood in 1880 and did not have stable county boundaries at that time. Excluded from the dataset. |
| Ohio: | United States Library of Congress, "Colton's Ohio", published by J.H. Colton (New York), 1898. |
| Oklahoma: | Not included in the 1880 census and so excluded from the dataset. |
| Oregon: | No sufficient map found. Only county data available. |
| Pennsylvania: | United States Library of Congress, "Pennsylvania administrative Township map", Rand McNally and Co. (Publishers), 1898. |
| Rhode Island: | Mitchell, Samuel A. "Township map of the States of Massachusetts, Connecticut and Rhode Island", Drawn and engraved by W.H. Gamble, Philadelphia. Copyright 1887 by Wm. M. Bradley and Bro. (1890)", Publisher: John Y. Huber and Co, available at David Rumsey Map Collection (www.davidrumsey.com). |
| South Carolina: | National Archives and Records Administration (NARA), Washington DC, Microfilm publication A3378: "Enumeration District Maps of the Twelfth trough Sixteenth Censuses of the United States, 1900 - 1940", County Maps on Microfilm Roll number 58. |
| South Dakota: | Had not attained statehood in 1880 and did not have stable county boundaries at that time. Excluded from the dataset. |
| Tennessee: | No sufficient map found. Only county data available. |
| Texas: | No sufficient map found. Only county data available. |
| Utah: | No additional map source was found, but the 1880 MCDs could be identified using the 1940 MCD maps. |
| Vermont: | County and township map of Vermont and New Hampshire. Copyright 1887 by Wm. M. Bradley and Bro., John Y. Huber Company, Publishers, Philadelphia and St. Louis. (1890), available at David Rumsey Map Collection (www.davidrumsey.com). |
| Virginia: | Gazetteer from United States Geological Survey (geonames.usgs.gov). |
| Washington: | In 1880 precincts were not separately returned by the enumerators. Only county data available. |
| West Virginia: | White's political map of West Virginia. Drawn and engraved by W.H. Gamble, Philadelphia. Entered according to Act of Congress in the year 1873 by M. Wood White in the Office of the Librarian of Congress at Washington, 1873, available at David Rumsey Map Collection (www.davidrumsey.com). |
| Wisconsin: | Mitchell, Samuel A. "Township map of the States of Michigan and Wisconsin", Drawn and engraved by W.H. Gamble, Philadelphia. Copyright 1887 by Wm. M. Bradley and Bro. (1890)", Publisher: John Y. Huber and Co. Additionally county maps on Lincoln county and Marathon county (Map of Wisconsin showing congressional and judicial districts. Copyright 1877 by Snyder, Van Vechten and Co. (Compiled and published by Snyder, Van Vechten and Co., Milwaukee. 1878)). Both available at David Rumsey Map Collection (www.davidrumsey.com). |

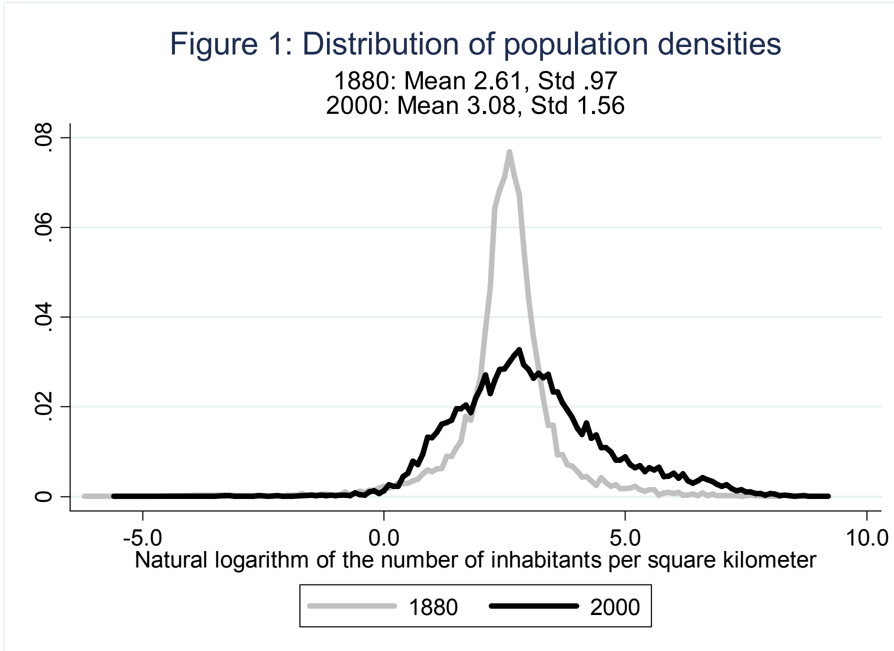
Wyoming: In 1880 precincts were not separately returned by the enumerators. Only county data available.

The states excluded from our analysis are: Alaska, Hawaii and Oklahoma (which are not included in the 1880 census) and North and South Dakota (which had not attained statehood in 1880 and did not have stable county boundaries at that time).

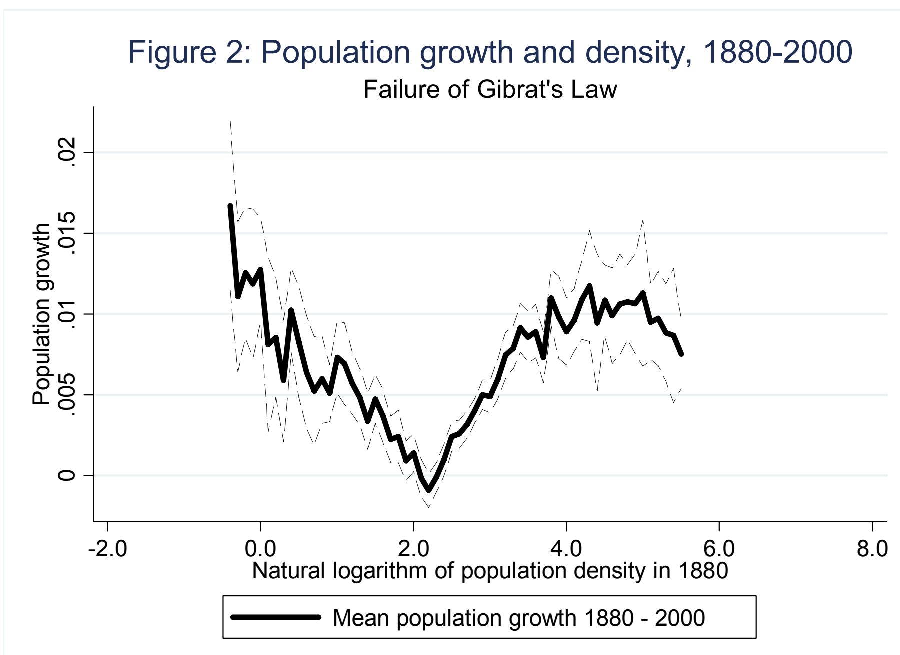
The included states for which we could not create data at the sub-county level are: Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Washington and Wyoming (for these 10 states the 1880 census contains a note saying that: "As the precincts in the different counties were not separately returned by the enumerators, the counties cannot be published in detail", which precludes obtaining information on sub-county divisions); the states of Alabama, Florida, Kentucky, Louisiana, Mississippi, Oregon, Tennessee and Texas (in these states we were unable to find sufficient maps to determine the location of the 1880 MCDs). The main problem with these states was that many of the MCD entries contained numbers instead of names, such as "Beat 1" or "Precinct 5". These entries are much harder to find on maps than names, since there are many competing numbering schemes applied to maps of this period, and number schemes are changed more frequently than names.

All MCDs with coordinates are present in one of the sources listed above apart from the following MCDs. These were found in the footnotes of the censuses 1890 – 1930 (the number indicates the county):

FERGUSON, 430, Arkansas; BLACKWOOD, 610, Arkansas; TREMONT, 1030, Arkansas; BRIDGE BEND, 1170, Arkansas; SPRING CREEK, 1490, Arkansas; HAPPY CAMP, 150, California; MOUNTAIN, 150, California; HOT SPRING, 490, California; SOUTH FORK, 490, California; HOT SPRINGS, 550, California; OSO FLACO, 790, California; SALINAS, 790, California; SAN JOSE, 790, California; SALMON, 930, California; SOUTH, 930, California; CASCADE, 1030, California; LASSEN, 1030, California; BUCKEYE, 1130, California; FAIRVIEW, 1130, California; MERRITT, 1130, California; ALLEN, 650, Illinois; ALLIN, 1130, Illinois; " ", 1770, Indiana; ELWOOD, 110, Kansas; KIOWA, 110, Kansas; LAKE CITY, 110, Kansas; MEDICINE LODGE, 110, Kansas; SUN CITY, 110, Kansas; MILLROOK, 670, Kansas; MILLROOK, 710, Kansas; Anthony, 750, Kansas; NOBLE, 1310, Kansas; VALLEY, 1310, Kansas; TWIN MOUND, 1370, Kansas; LUDWICK, 1510, Kansas; WEST WATERVILLE, 110, Maine; MUSCLE RIDGE, 130, Maine; MUSCONGUS ISLAND, 150, Maine; PINE, 170, Maine; BURBANK T, 210, Maine; MOUNT KINEO, 210, Maine; PERKINS, 230, Maine; HOLDEN, 250, Maine; PLEASANT VALLEY, 410, Maryland; DISTRICT 6, CLOBOURNES, 450, Maryland; CENTER, 1010, Minnesota; TOWNSHIP 105, RANGE 42, 1010, Minnesota; TOWNSHIP 103, RANGE 42, 1050, Minnesota; TOWNSHIP 103, RANGE 43, 1050, Minnesota; EAST BATTLE LAKE, 1110, Minnesota; TOWNSHIP 132, 1110, Minnesota; TOWNSHIP 133, RANGE 49, 1110, Minnesota; TOWNSHIP 135, RANGE 42, 1110, Minnesota; TOWNSHIP 136, RANGE 36, 1110, Minnesota; TOWNSHIP 136, RANGE 37, 1110, Minnesota; TOWNSHIP 137, RANGE 36, 1110, Minnesota; TOWNSHIP 137, RANGE 37, 1110, Minnesota; TOWNSHIP 137, RANGE 38, 1110, Minnesota; RESERVE, 1230, Minnesota; TOWN 111 RANGE 38, 1270, Minnesota; DULUTH (I), 1370, Minnesota; ONEOTA, 1370, Minnesota; SAHLMARK, 1490, Minnesota; TOWNSHIP 124 RANGE 44, 1490, Minnesota; TOWNSHIP 125 RANGE 44, 1490, Minnesota; TOWNSHIP 126 RANGE 44, 1490, Minnesota; MORITZIUS (I), 1710, Minnesota; OTIS, 1730, Minnesota; TOWNSHIP 114 RANGE 46, 1730, Minnesota; GERMAN, 170, Missouri; BENTON, 430, Missouri; GALLOWAY, 430, Missouri; MARION, 430, Missouri; WESTPORT, 950, Missouri; GERMAN, 1230, Missouri; EAST, 1430, Missouri; LYNN, 1490, Missouri; OAK GROVE, 1490, Missouri; MARION, 1530, Missouri; FOURCHEE, 1810, Missouri; CARONDELET, 1890, Missouri; CENTRAL, 1890, Missouri; JEFFERSON, 1950, Missouri; COURT-HOUSE ROCK, 330, Nebraska; SCOTT, 350, Nebraska; CEDAR VALLEY, 810, Nebraska; PLATTE, 810, Nebraska; SPRING CREEK, 830, Nebraska; SPRINGBROOK, 830, Nebraska; CAPITAL, 1090, Nebraska; MIDLAND, 1090, Nebraska; BOHNART, 1290, Nebraska; SPRING VALLEY, 1290, Nebraska; JOHNSON CREEK, 1510, Nebraska; GRANT, 1690, Nebraska; LISBON, 450, New York; GRAMPION 10 Utah; TERRACE 30 Utah; HILLSDALE, 210, Utah; LITTLE PINTO, 210, Utah; TINTIC, 230, Utah; BELLEVUE, 250, Utah; DUNCANS RETREAT, 250, Utah; GRAFTON, 250, Utah; JOHNSON, 250, Utah; PAH REAH, 250, Utah; SHUNESBURG, 250, Utah; KANYON, 290, Utah; MEADOWVILLE, 330, Utah; FREEDOM, 390, Utah; PETTY, 390, Utah; VERMILLION, 410, Utah; WILLOW BEND, 410, Utah; HAYTSVILLE, 430, Utah; BATESVILLE, 450, Utah; JACOB CITY, 450, Utah; MILL, 450, Utah; HEBRON, 530, Utah; PINTO, 530, Utah; PRICE CITY, 530, Utah; SILVER REEF, 530, Utah; LYNNE, 570, Utah; WALKER, 1950, Virginia; LEE, 130, West Virginia; SULLIVAN, 530, Wisconsin; CARPENTER, 670, Wisconsin; BRANNAN, 990, Wisconsin.



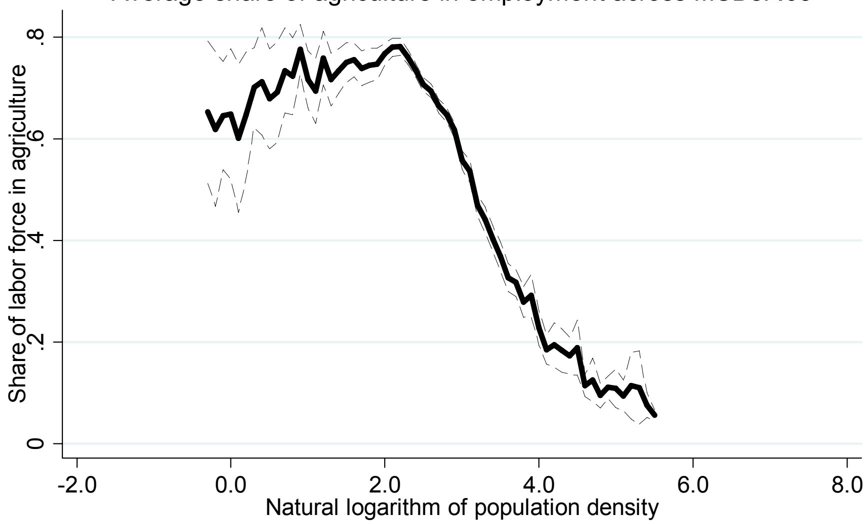

Note: This figure shows the distribution of log population per square kilometer in 1880 and 2000 estimated using non-parametric specification (1) for the sample of "A and B" states. Population density bins are defined by rounding down log population density for each MCD to the nearest single digit after the decimal point. For example, all MCDs with log population density from 0.1 to 0.2 are grouped together in bin 0.1. See the data appendix for further details.




Note: The solid line shows mean population growth rate from 1880-2000 within each population density bin based on estimating non-parametric specification (2) for the sample of "A and B" states. Population density bins are defined by rounding down log population density for each MCD to the nearest single digit after the decimal point. The dashed lines show 95 percent confidence intervals based on robust standard errors clustered by county. Since population density bins at the extreme ends of the distribution typically contain at most one observation, the figure (but not the estimation) omits the 1 percent most and least dense MCDs in 1880. See the data appendix for further details.

Figure 3: Share of agriculture and population density in 1880

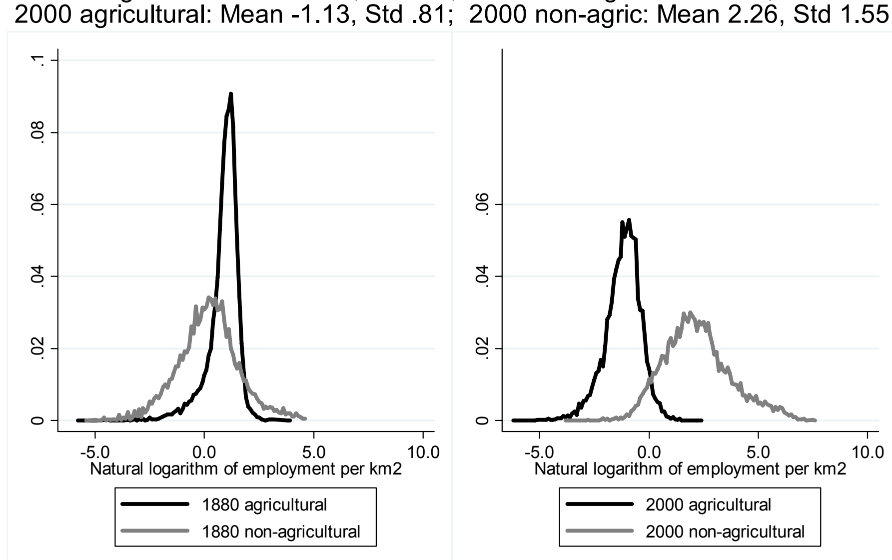
Share of agriculture in total employment: .35
Average share of agriculture in employment across MCDs: .63




Note: The solid line shows the mean share of agriculture in 1880 employment within each population density bin based on estimating non-parametric specification (2) for the sample of "A and B" states. Population density bins are defined by rounding down log population density for each MCD to the nearest single digit after the decimal point. The dashed lines show 95 percent confidence intervals based on robust standard errors clustered by county. Since population density bins at the extreme ends of the distribution typically contain at most one observation, the figure (but not the estimation) omits the 1 percent most and least dense MCDs in 1880. See the data appendix for further details.

Figure 4: Density of employment

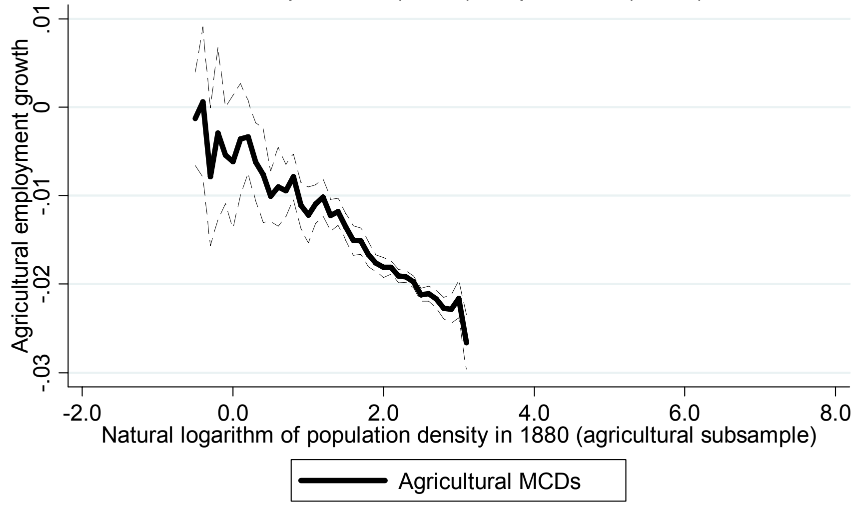
1880 agricultural: Mean .92, Std .71;  1880 non-agric: Mean .23, Std 1.4
2000 agricultural: Mean -1.13, Std .81;  2000 non-agric: Mean 2.26, Std 1.55




Note: This figure shows the distribution of log agricultural employment and log non-agricultural employment (employment in manufacturing and services) per square kilometer in 1880 and 2000 estimated using non-parametric specification (1) for the sample of "A and B" states. Employment density bins are defined by rounding down log employment density for each MCD to the nearest single digit after the decimal point. Since population density bins at the extreme ends of the distribution typically contain at most one observation, the figure (but not the estimation) omits the 1 percent most and least dense MCDs in 1880. See the data appendix for further details.

## Figure 5: Agricultural growth and population density

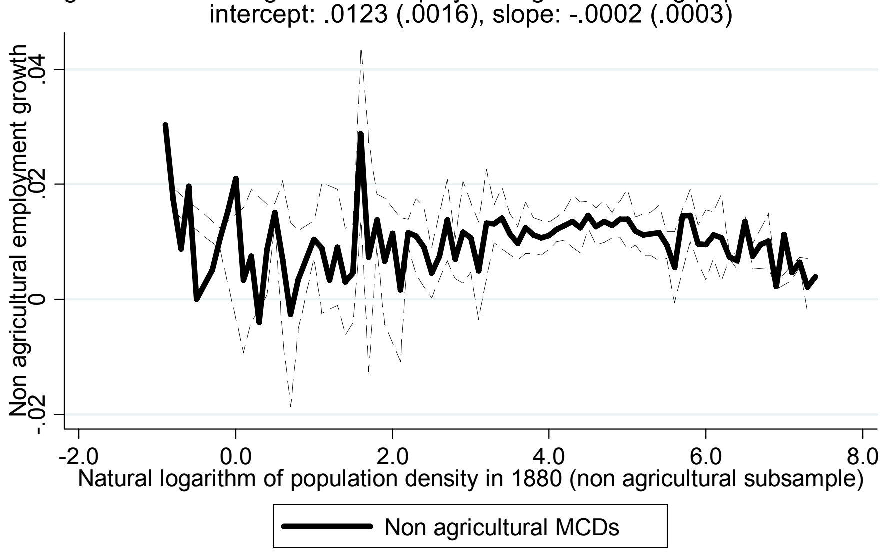
Regression of agricultural employment growth on log population density:
intercept -.0053 (.0009), slope: -.006 (.0004)




Note: The solid line shows the mean growth rate of agricultural employment from 1880-2000 within each population density bin based on estimating non-parametric specification (2) for the agricultural subsample (an agricultural share in 1880 employment of greater than 0.8) within "A and B" states. Population density bins are defined by rounding down log population density for each MCD to the nearest single digit after the decimal point. The dashed lines show 95 percent confidence intervals based on robust standard errors clustered by county. Since population density bins at the extreme ends of the distribution typically contain at most one observation, the figure (but not the estimation) omits the 1 percent most and least dense MCDs in 1880. See the data appendix for further details.

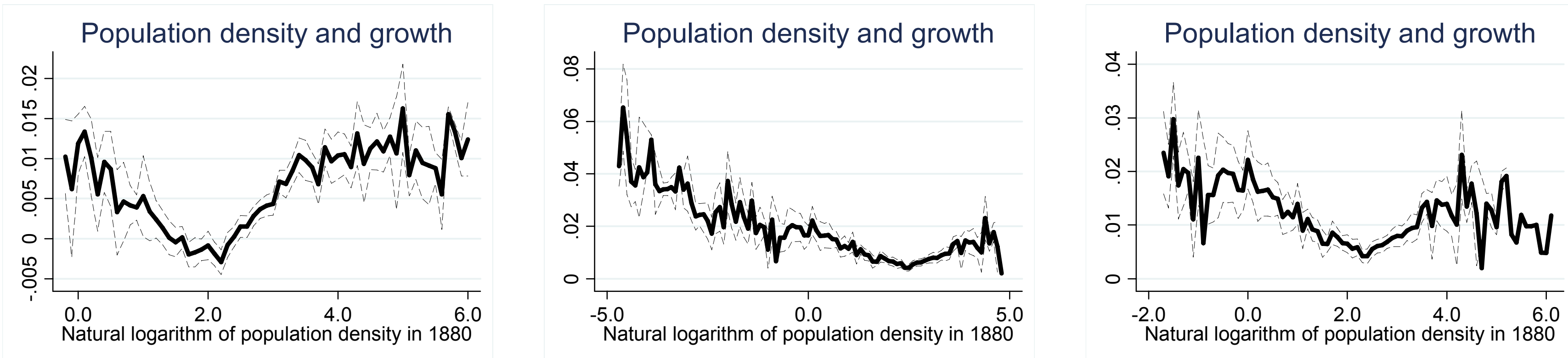
## Figure 6: Non agricultural growth and population density

Regression of non agricultural employment growth on log population density:
intercept: .0123 (.0016), slope: -.0002 (.0003)




Note: The solid line shows the mean growth rate of non-agricultural employment (employment in manufacturing and services) from 1880-2000 within each population density bin based on estimating non-parametric specification (2) for the non-agricultural subsample (an agricultural share in 1880 employment of less than 0.2) within "A and B" states. Population density bins are defined by rounding down log population density for each MCD to the nearest single digit after the decimal point. The dashed lines show 95 percent confidence intervals based on robust standard errors clustered by county. Since population density bins at the extreme ends of the distribution typically contain at most one observation, the figure (but not the estimation) omits the 1 percent most and least dense MCDs in 1880. See the data appendix for further details.

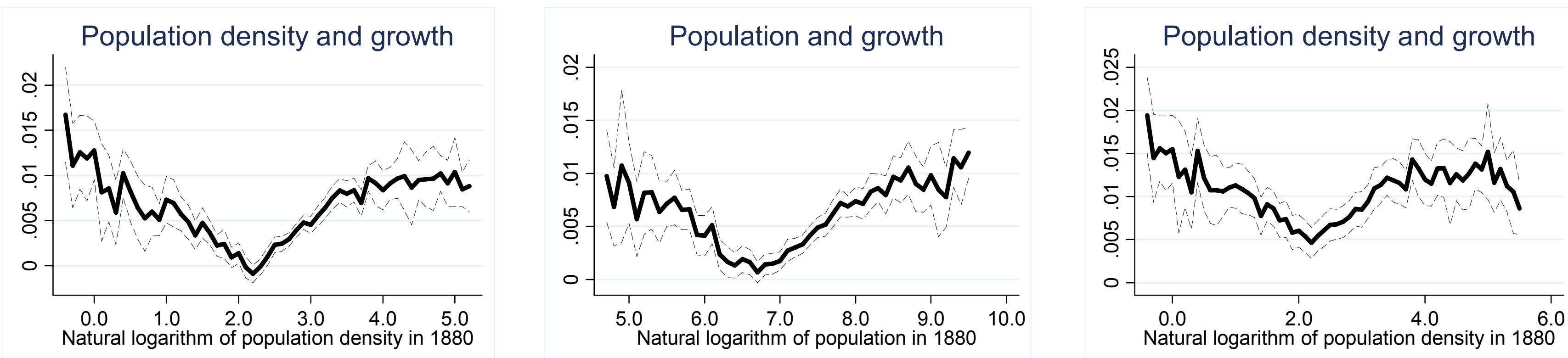
Figure 7: Robustness of failure of Gibrat’s Law




Panel A: A states sample

Panel B: Counties sample

Panel C: Hybrid sample



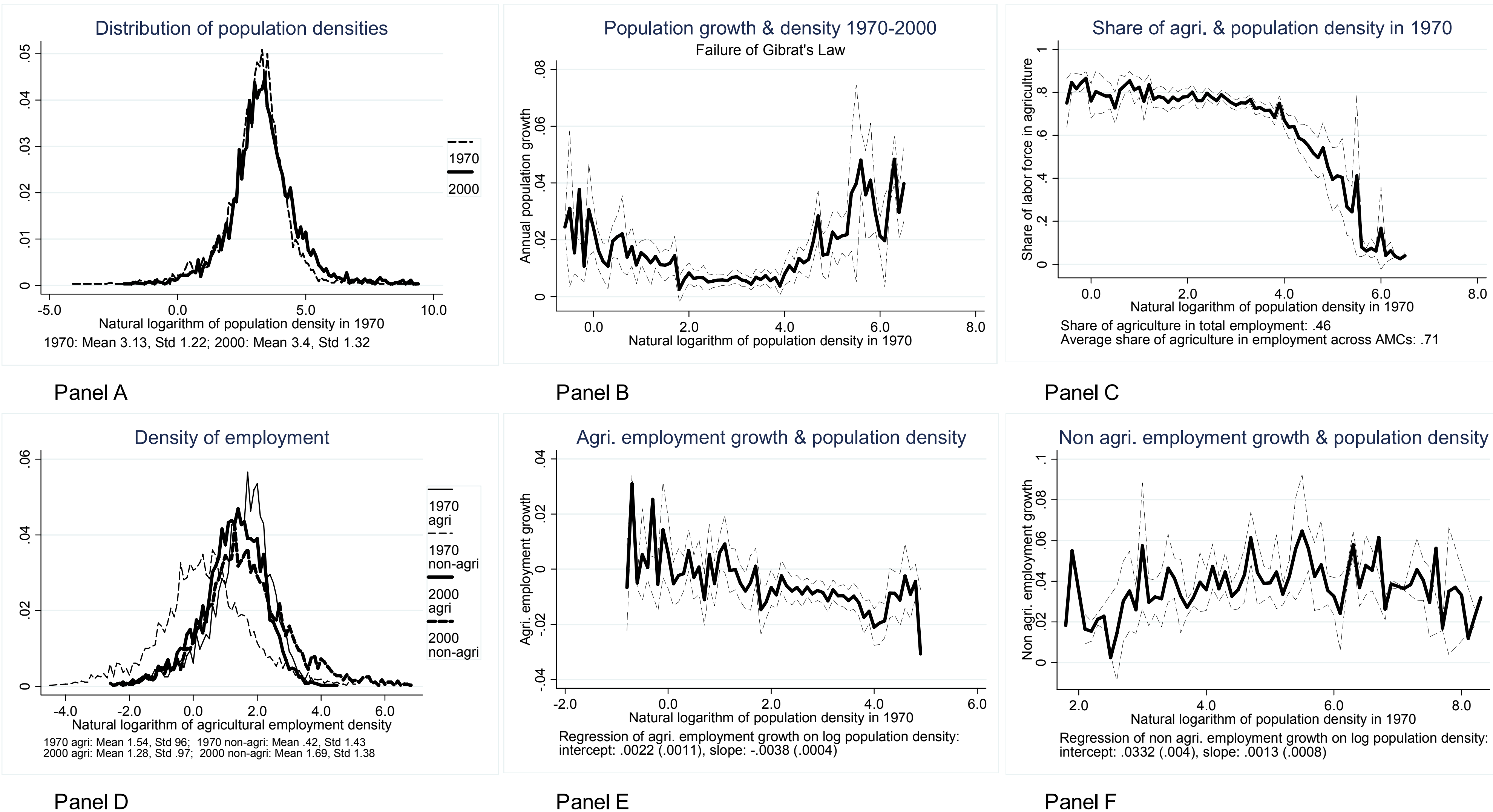

Panel D: Suburban sample

Panel E: A and B, population not densities

Panel F: A and B sample, state controls

Note: This figure shows the robustness of the failure of Gibrat’s Law (Figure 2) by reproducing it for other samples. The various samples used here are described in the appendix. Since population density bins at the extreme ends of the distribution typically contain at most one observation, the figure (but not the estimation) omits the 1 percent most and least dense MCDs in 1880.

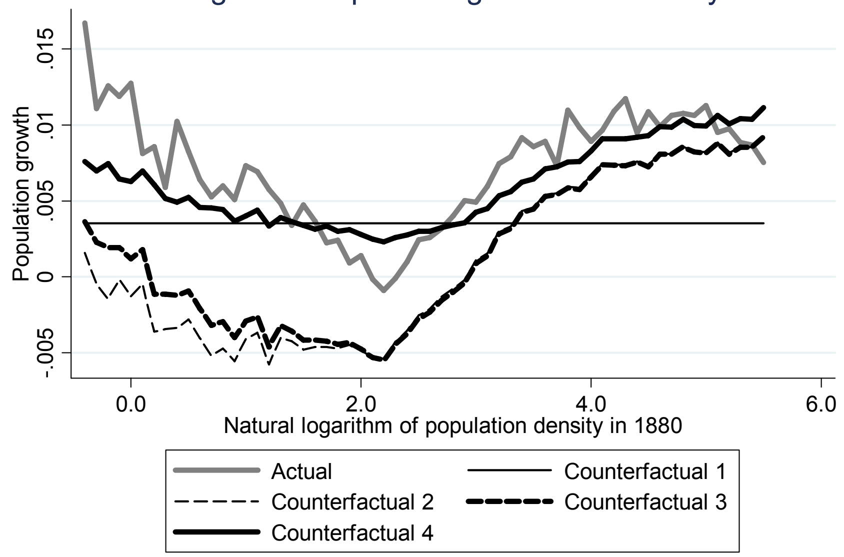
Figure 8: Brazilian Results




Note: This figure reproduces Figures 1 to 6 but uses data on Brazilian municipalities (Áreas Mínimas Comparáveis (AMCs)) instead of US data. The agricultural subsample in Panel E is defined as in the US. The non-agricultural subsample in Panel F includes AMCs for which agriculture's share of 1970 employment was less than 0.4 rather than 0.2 as in the US, because of the small sample size using a 0.2 threshold (though results are similar using the lower threshold).

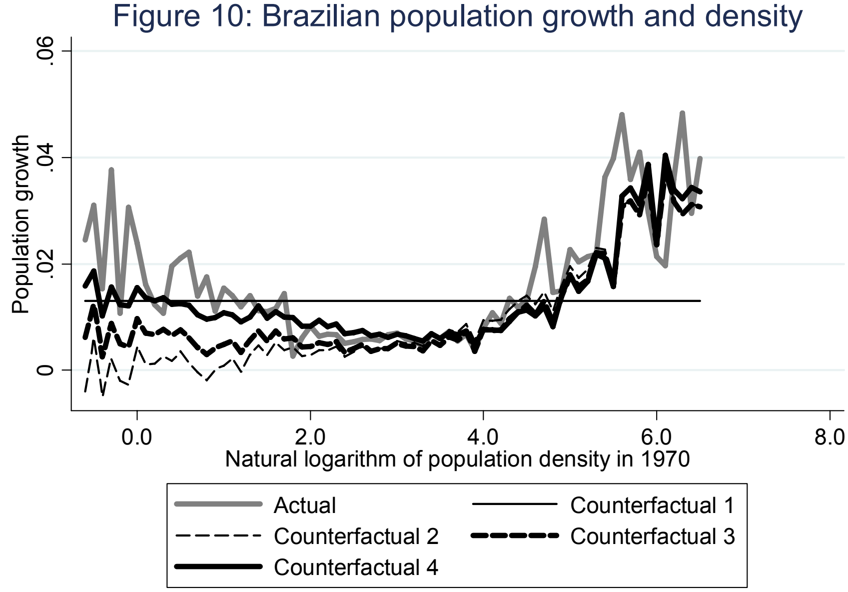
.

Figure 9: Population growth and density




Note: This figure shows mean actual and predicted population growth from 1880-2000 within each population density bin based on estimating non-parametric specification (2) for the sample of "A and B" states. Population density bins are defined by rounding down log population density for each MCD to the nearest single digit after the decimal point. Since population density bins at the extreme ends of the distribution typically contain at most one observation, the figure (but not the estimation) omits the 1 percent most and least dense MCDs in 1880. Counterfactuals 1-4 use progressively more components of the model to generate predicted population growth as discussed in the paper and data appendix.

Figure 10: Brazilian population growth and density




Note: This figure reproduces Figure 9 but uses data on Brazilian municipalities (Áreas Mínimas Comparáveis (AMCs)) instead of US data. The agricultural subsample used in Counterfactuals 1-3 is defined as in the US. The non-agricultural subsample used in Counterfactuals 1-3 includes AMCs for which agriculture’s share of 1970 employment was less than 0.4 rather than 0.2 as in the US, because of the small sample size using a 0.2 threshold (though results are similar using the lower threshold).

**CENTRE FOR ECONOMIC PERFORMANCE**
**Recent Discussion Papers**

| | | |
|---|---|---|
| 891 | Nicholas Bloom<br>Christos Genakos<br>Ralf Martin<br>Raffaella Sadun | Modern Management: Good for the Environment or Just Hot Air? |
| 890 | Paul Dolan<br>Robert Metcalfe | Comparing willingness-to-pay and subjective well- being in the context of non-market goods |
| 889 | Alberto Galasso<br>Mark Schankerman | Patent Thickets and the Market for Innovation: Evidence from Settlement of Patent Disputes |
| 888 | Raffaella Sadun | Does Planning Regulation Protect Independent Retailers? |
| 887 | Bernardo Guimaraes<br>Kevin Sheedy | Sales and Monetary Policy |
| 886 | Andrew E. Clark<br>David Masclet<br>Marie-Claire Villeval | Effort and Comparison Income Experimental and Survey Evidence |
| 885 | Alex Bryson<br>Richard B. Freeman | How Does Shared Capitalism Affect Economic Performance in the UK? |
| 884 | Paul Willman<br>Rafael Gomez<br>Alex Bryson | Trading Places: Employers, Unions and the Manufacture of Voice |
| 883 | Jang Ping Thia | The Impact of Trade on Aggregate Productivity and Welfare with Heterogeneous Firms and Business Cycle Uncertainty |
| 882 | Richard B. Freeman | When Workers Share in Profits: Effort and Responses to Shirking |
| 881 | Alex Bryson<br>Michael White | Organizational Commitment: Do Workplace Practices Matter? |
| 880 | Mariano Bosch<br>Marco Manacorda | Minimum Wages and Earnings Inequality in Urban Mexico. Revisiting the Evidence |
| 879 | Alejandro Cuñat<br>Christian Fons-Rosen | Relative Factor Endowments and International Portfolio Choice |
| 878 | Marco Manacorda | The Cost of Grade Retention |
| 877 | Ralph Ossa | A ‘New Trade’ Theory of GATT/WTO Negotiations |

| | | |
|---|---|---|
| 876 | Monique Ebell<br>Albrecht Ritschl | Real Origins of the Great Depression: Monopoly Power, Unions and the American Business Cycle in the 1920s |
| 875 | Jang Ping Thia | Evolution of Locations, Specialisation and Factor Returns with Two Distinct Waves of Globalisation |
| 874 | Monique Ebell<br>Christian Haefke | Product Market Deregulation and the U.S. Employment Miracle |
| 873 | Monique Ebell | Resurrecting the Participation Margin |
| 872 | Giovanni Olivei<br>Silvana Tenreyro | Wage Setting Patterns and Monetary Policy: International Evidence |
| 871 | Bernardo Guimaraes | Vulnerability of Currency Pegs: Evidence from Brazil |
| 870 | Nikolaus Wolf | Was Germany Ever United? Evidence from Intra- and International Trade 1885 - 1993 |
| 869 | L. Rachel Ngai<br>Roberto M. Samaniego | Mapping Prices into Productivity in Multisector Growth Models |
| 868 | Antoni Estevadeordal<br>Caroline Freund<br>Emanuel Ornelas | Does Regionalism Affect Trade Liberalization towards Non-Members? |
| 867 | Alex Bryson<br>Harald Dale-Olsen | A Tale of Two Countries: Unions, Closures and Growth in Britain and Norway |
| 866 | Arunish Chawla | Multinational Firms, Monopolistic Competition<br>and Foreign Investment Uncertainty |
| 865 | Niko Matouschek<br>Paolo Ramezzana<br>Frédéric Robert-Nicoud | Labor Market Reforms, Job Instability, and the Flexibility of the Employment Relationship |
| 864 | David G. Blanchflower<br>Alex Bryson | Union Decline in Britain |
| 863 | Francesco Giavazzi<br>Michael McMahon | Policy Uncertainty and Precautionary Savings |
| 862 | Stephen Hansen<br>Michael F. McMahon | Delayed Doves: MPC Voting Behaviour of Externals |

**The Centre for Economic Performance Publications Unit**
**Tel 020 7955 7284 Fax 020 7955 7595 Email info@cep.lse.ac.uk**
**Web site http://cep.lse.ac.uk**