CROSS-SECTIONAL EVOLUTION OF THE US CITY SIZE DISTRIBUTION [1]

September 15, 2000

Henry G. Overman
Department of Geography and Environment
London School of Economics
Houghton Street, London WC2A 2AE, UK

and

Yannis M. Ioannides
Department of Economics
Tufts University
Medofr, MA 02155, USA

Abstract

We report nonparametrically estimated stochastic transition kernels for the evolution of the distribution of US metropolitan area populations, for the period 1900 to 1990. These suggest a fair amount of uniformity in the patterns of mobility during the study period. The distribution of city sizes is predominantly characterised by persistence. Additional kernel estimates do not reveal any stark differences in intra-region mobility patterns.

We characterise the nature of intra-size distribution dynamics by means of measures that do not require discretisation of the city size distribution. We employ these measures to study the degree of mobility within the US city size distribution and, separately, within regional and urban subsystems. We find that different regions show different degrees of intra-distribution mobility. Second-tier cities show more mobility than top-tier cities.

This version: September 14, 2000. dynxevo3.tex

Proposed running head: Cross-Sectional Urban Evolution

JEL Classification codes: R12, C14.

Key words: City Size Evolution, City Size Distribution, Cross-sectional evolution.

[1] We thank two referees and the editor for insightful comments and suggestions, Danny Quah for giving us access to his tSrF statistical package, Linda H. Dobkins for her assistance with the data, and Timothy Harris for valuable research assistance. Ioannides thanks the National Science Foundation and the MacArthur Foundation for financial support.

Corresponding author: Henry G. Overman, Department of Geography and Environment, London School of Economics, Houghton St, London, WC2A 2AE. Email: h.g.overman@lse.ac.uk. Tel: +44-207-955-6581. Fax +44-207-955-7412.

1. Introduction

Empirical studies of the distribution of city sizes have a long and distinguished history. At least 80 years ago, it was observed that the distribution of cities within an urban system is often remarkably well approximated by a Pareto distribution. This observation has generated a vast body of empirical work aimed at testing this and related propositions. Much of this work has concentrated on testing the rank–size rule first proposed by Zipf [29].[2] This large empirical literature has, in turn, led to the development of a number of theoretical models which attempt to generate this apparent regularity. This collection of models are essentially statistical – they seek to generate, rather than explain, the regularity. To do this, they abstract from underlying economic or social processes that drive the evolution of city sizes. The importance of the rank–size rule in framing the discussion about the distribution of city sizes has had two important implications for the literature on the development of the urban system. First, it has led to the acceptance of simplistic models that downplay important economic and social forces but that are capable of replicating the regularity. Second, it has relegated work on other aspects of the distribution to a distant second place. This paper is primarily concerned with these other aspects of the distribution.

With respect to the first implication, recent work by a number of theorists, who developed the so–called new economic geography, highlight the problems that the rank size rule has presented for theoretical work. In common with an older theoretical literature, these authors have emphasised the interplay of agglomeration and dispersion forces as key in determining city sizes. However, they have also emphasised the fact that "when it comes to the size distribution of cities, [...] the problem we face is that the data offer a stunningly neat picture, one that is hard to reproduce in any plausible (or even implausible) theoretical model" [See Fujita, Krugman and Venables[13], Chapter 12]. For the earlier literature, see Simon [27], Krugman [18], and Gabaix [14], who propose models capable of generating regularities in the distribution of city sizes.

The second implication has received very little direct attention. The empirical work on the rank size rule is essentially involved with one particular characteristic of the distribution of city sizes – the shape of that distribution. In contrast, this paper examines intra–distribution dynamics. It asks questions about how cities develop relative to the rest of the urban system, both in terms of (ordinal) rankings and relative sizes. We propose a number of techniques for characterising this

[2]The rank–size rule (or Zipf's law) states that the city size distribution follows a Pareto distribution with exponent one. See Overman [20], Chapter 4.

intra–distribution mobility.

We do not see characterising this intra–distribution mobility as a substitute of direct tests of either the economic or the stochastic models of the development of the urban system. Economic models are only infrequently asked to predict the shapes of distribution of endogenous variables of interest, so there is no reason to be unduly demanding with regard to the dynamics of the distribution of city sizes. To the extent that economic models help us understand the economic forces that might promote agglomeration or drive dispersion, failure to match empirical regularities on city sizes should not lead to an outright rejection of those models. However, given that the aim of stochastic models is to help us understand the nature of the process that might produce the rank–size rule, it would seem important that these models also deliver on other aspects of the city size distribution. Stochastic models which generate the shape of the distribution, but only at the expense of unrealistic intra–distribution dynamics, may well be uninformative about the processes at work.

This paper proceeds as follows. Section 2 reviews some of the related empirical and theoretical literature. Section 3 briefly describes the data. Section 4 develops a number of empirical tools which can be used to analyse intra–distribution dynamics. We use these tools to examine the evolution of the US city size distribution from 1900 to 1990. Section 5 concludes.

2. Related literature

There is a vast empirical literature on the distribution of city sizes. A very selective account follows, which seeks to highlight the main issues and those most closely related to the empirical work in this paper. A number of extensive surveys exist: Carroll [7] covers earlier work in some detail; Cheshire [8] provides a survey of more recent work.

At least as early as Auerbach [3] it had been proposed that the city size distribution could be closely approximated by a Pareto distribution. Thus, if we rank cities from the largest (rank 1) to the smallest (rank N) then r(p), the rank for a city of size (population) p, obeys: $r(p) = Ap^{-\alpha}$. Or, taking logs

$$\ell n r(p) = \ell n A - \alpha \ell n p. \tag{1}$$

Rank may equivalently be measured by the countercumulative of the size distribution.

Zipf [29] went further. He proposed that, not only did the distribution of city sizes follow a Pareto distribution, but that it took a special form of the distribution where $\alpha = 1$.This expression of the regularity is known as the 'rank–size" rule (or Zipf's rule) and has formed the starting point for much of the empirical literature. It implies that the second biggest city is half the size of the largest, the third biggest is a third the size of the largest, etc.

Rosen and Resnick [25] brought together the questions from a large body of literature developed from the 1950s to the 1970s.[3] They highlighted the importance of the definition of the lower threshold size for cities[4] and considered how the urban system might be best defined. We will return to this second issue briefly in section 4.2 below.

A further two decades of work has followed with two key conclusions. The first, less controversial, is that the city size distribution is reasonably well approximated by a Pareto distribution, at least for the largest cities. The second, far more controversial, is that the exponent of the Pareto Law, coefficient α in Eq. (1), is close to one. Some authors, notably Krugman [18], have argued that the combined evidence suggests that the rank size rule holds for a number of different samples over a number of different time periods. Others, such as Alperovich [2], reject this stronger second conclusion, but accept the first. The debate still rages. Dobkins and Ioannides [10] obtain good fits, by performing OLS of the countercumulative of the city size distribution against the logarithm of population (along the lines of Equ. (1)) and by maximum likelihood directly in terms of $r(p) = Ap^{-\alpha}$. Their estimates for US cities, for 1900 to 1990, show α decreasing over time. In common with other work, they find that the exponent α is around one for a sub–sample of the largest cities, but below one for the whole sample. However, when they compare the fit of the Pareto Law with a nonparametric one obtained with the generalized validation criterion, they find that the fit of the Pareto Law is poor for a substantial portion of the distribution, thus raising doubts about the validity of the strict rank-size rule. Black and Henderson [5] use similar (though not identical) data, and reject Eq. (1) as they find a significant quadratic term for $\ell n p$, as well.

These last two papers also consider a number of issues related to the intra–distribution mobility characteristics of the city size distribution. Both build on Eaton and Eckstein [11], who use transition probability matrices to characterise the evolution of the French and Japanese urban systems and find that both those systems are characterised by parallel growth. Cities tend to grow at the same rate, maintaining their place in the relative distribution and consequently showing little intra–distribution mobility. In contrast, Dobkins and Ioannides [10] find that the US system is characterised by the

[3] Key contributions included Allen [1], Madden [19], and Berry [4].

[4] This is a recurring theme in the urban systems literature. See Black and Henderson [5] and Dobkins and Ioannides [10] for a recent discussion.

entry of new cities and a higher degree of mobility. Black and Henderson [5] confirm this result. They show that new entry means that cities tend to be more mobile up the distribution, but less mobile down the distribution. The expected transition time from lowest state to highest is around 500 years. Movement in the opposite direction takes, on average, 5500 years. This paper builds on these three papers to provide a more detailed characterisation of the intra–distribution mobility of cities within the US city size distribution.

3. Data

There are a variety of ways to define cities.[5] In this paper, we use contemporaneous Census Bureau definitions of metropolitan areas, with adaptations for availability. From 1900 to 1950, we have metropolitan areas defined by the 1950 census.[6] That is, for years previous to 1950, we use reconstructions from Bogue [6] of what populations would have been in each metropolitan area in each year if the cities had been defined as they were in 1950. For each decennial year from 1950 to 1980, we use the metropolitan area definitions that were in effect for those years. Between 1980 and 1990, the Census Bureau redefined metropolitan areas in such a way that the largest US cities would seem to have taken a huge jump in size, and several major cities would have been lost. While this might be appropriate for some uses of the data, it would introduce "artificial" intra–distribution mobility for the 1980–1990 period. Therefore, Dobkins and Ioannides reconstructed the metro areas for 1990, based on the 1980 definitions, much as Bogue did earlier. We believe that this gives the most consistent definitions of US cities (metropolitan areas) that we are likely to find.

The method raises a question as to which cities, as defined or reconstructed, should be included. In the years from 1950 to 1980, we use the Census Bureau's listing of metropolitan areas. Although the wording of the definitions of metropolitan areas has changed slightly over the years, the number 50,000 is a minimum requirement for the core area within the metropolitan area.Therefore, we used 50,000 as the cutoff for including metropolitan areas as defined by Bogue prior to 1950. Consequently we have a changing number of cities over time, from 112 in 1900 to 334 in1990. While it is often difficult to deal with an increasing number of cities econometrically, we think that this is a key aspect of the US system of cities, and thus worthy of accommodating.

[5]This section draws extensively from Dobkins and Ioannides [10].

[6]Technically, a metropolitan area must contain either a city of at least 50,000,or an urbanized area of at least 50,000 and total metropolitan population of 100,000 (75,000 in New England).

We also use information on regional location defined according to the Census Bureau division of the country into nine regions. We recombine these regions into five regions, when necessary. Table 1 provides summary statistics of the data for each census year. Table 2 provides additional statistics for the whole sample in 1990.

There are two important distinctions between our data and the data used by Black and Henderson [5] for the same time period. First, they define the geographical area of a city as the collection of counties that form that city in 1990. They then use the urban population of each of these counties to give city size in each census year from 1900–1990. This gets around problems relating to changing definitions of metro areas between 1950 and 1990 that apply to our data. However, it introduces an additional source of mismeasurement relating to the use of contemporaneous definitions of urban population that may change throughout the period. It also means that collections of small towns in areas that will become cities are treated identically to genuine metro areas of a similar size. Second, they use a relative cut–off point to define when a city enters into the sample, whereas we use an absolute cut–off point. Their use of a relative cut–off point in combination with metro areas defined on the basis of urban populations means that their sample will tend to overstate the number of functional metro–areas in any given sample period. In contrast, our approach based on an absolute cut–off point will tend to understate the number of functional metro–areas. Black and Henderson show that estimates of intra–distribution mobility are sensitive to the choice of an absolute versus relative cut–off. However, a–priori there is no reason to prefer one definition over another.

4. Intra–distribution dynamics

As Quah [21] has forcefully argued, typical cross-section or panel data econometric techniques do not allow inference about patterns in the intertemporal evolution of the entire cross-sectional *distribution*. Such techniques do not allow us to consider the impact over time of one part of the distribution upon another, i.e., of the development of large cities as a group upon smaller cities. Making such inferences requires that one model and estimate directly the full dynamics of the entire distribution of cities. In contrast, typical panel data analyses involve efficient and consistent estimation of models where the error consists of components reflecting individual effects (random or fixed), time effects and purely random factors. The evolution of urbanization and suburbanization may affect individual cities so drastically as to render conventional methods of accounting for attrition totally inappropriate.

Examination of evolving cross-sectional distributions are most appropriate when the sample of interest is the entire distribution, and individual observations are used to describe the entire distribution of population of metropolitan areas in the US. We may elaborate further the process of evolution of the system of cities by considering alternative scenaria that articulate the spatial context. Consider first a situation where cities of uniform sizes are uniformly spread over space. Appearance of new cities that are randomly scattered over space is likely not to alter the pattern of uniformity. To the extent that geographical proximity leads invariably to agglomeration, this setting implies creation of larger cities of uniform sizes. Consider, alternatively, cities of uniform sizes scattered over space but in a way that exhibits clustering. Appearance of additional cities of uniform sizes makes it more likely that ever larger cities will be created through the agglomeration of existing ones.

Recall that our data consist of only ten cross-sections, one for each of the ten census years since 1900, with 112 metropolitan areas and 334 in 1990. We use these data to examine intra-distribution dynamics by first considering nonparametrically the long run transition patterns in the US city size distribution. Next we introduce measures of intra-distribution mobility in the form of suitably defined statistics of dispersion and serial correlation in changes in rankings. Finally, we examine patterns in the intra-distribution dynamics within different groupings of cities, that is, in terms of geographical regions and hierarchical tiers.

4.1 Intra–distribution mobility

We will consider two inter-related types of intra–distribution mobility: changes in the rankings of cities and changes in their relative sizes. Previous studies of intra–distribution mobility have studied both types of mobility without clearly distinguishing between implications of those two different concepts. Eaton and Eckstein [11], Black and Henderson [5] and Dobkins and Ioannides [10] consider the size of cities relative to the mean city size. They then discretize the state space of relative city sizes by defining discrete intervals[7] and calculate the transition matrices corresponding to this discretization. Only Dobkins and Ioannides [10] consider mobility in terms of the rankings of cities by discretizing the state space in each period on the basis of quantiles (bottom 10%, second

[7] For example, Dobkins and Ioannides [10] divide the state space in terms of bounds defined by .30, .50, .75, 1.00, 2.00 and 20.00 times the contemporaneous mean.

10%, ..., top 10%). They argue that this gives a more detailed insight into the intra–distribution dynamics, without making it clear that the mobility that they are studying is subtly different.

To see why the distinction is important, we need to think about what the two types of exercises tell us. Considering the first type of mobility, that is, in terms of city sizes relative to the mean, allows us to answer a number of interesting questions about the long run city size distribution. Thus, one can examine whether the distribution has a tendency to become uniform (flatten out), to collapse to a single point (all cities converge to the same size), or to, say, become bimodal. To do this, after discretising the state space and calculating the transition probability matrix, one would calculate the ergodic distribution of the associated markov process(assuming that it exists). All three of the above mentioned papers do exactly that. Black and Henderson [5] emphasize that the long run ergodic distribution is remarkably close to the current distribution.

Notice that all of these are questions about what mobility implies for the overall shape of the distribution. These exercises also tell us about changes in the rankings of cities. Take any two neighbouring discrete states. If some cities move up from the lower state to the higher state, while others move from the higher state to the lower state, then the rankings of those cities must have changed. There are, however, two problems with this method of characterising the changing rankings of cities. First, the discretisation of the state space means that we do not observe what happens within each discrete state. Only when cities move between states do we get information on mobility. Second, observing the movement of an individual city between states does not necessarily imply a change in rankings. This is where the second approach of Dobkins and Ioannides [10] based on quantiles differs from the approach based on an (arbitrary) fixed discretisation. For this second approach, movement of one city up a discrete state, *must* be accompanied by a corresponding move down a state by another city and vice–versa. Thus all movements between states correspond to changes in rankings. However, this second approach still suffers from the fact that we do not observe mobility and changes of rankings within cells.

The problem of movements within cells arises because we discretize a continuous state space in order to calculate transition probability matrices. Previous attempts to characterise the intra–distribution dynamics of the city size distribution also face two other related problems. The first is that there are a group of very large cities whose mobility characteristics may be different from the rest of the system. Including these cities may over–emphasise the degree of persistence in the distribution. A simple solution would be to exclude those cities from the sample and recalculate the transition matrices. However, this brings us to the second problem, that the number of cities is such that we can only discretize the state space into relatively few discrete states. For example, Black and Henderson report results for a five–state markov process, but the top state is occupied by the very immobile largest cities, leaving four states to capture the dynamics of the remaining cities. Such a limited number of states may lead us to underestimate the degree of mobility. Dobkins and Ioannides [10] allow for ten discrete quantile states. However, that number of states leaves very few cities in each state, and mobility may be overstated due to the movements of a very few cities. Finally, a large number of states for a small number of cities means that small changes in the discretization may lead to large changes in our estimates of the degree of persistence or mobility.

4.1.1 Estimations Danny Quah has proposed, in a series of papers starting with Quah [20] and including notably Quah [21; 22; 23], a set of tools[8] for analysing evolving distributions which avoid the need to discretize the state space. He suggests calculating a non–parametric estimate of the underlying continuous transition kernel. Let f_t denote the density function (distribution) of $P_{i,t}$, the population of city i at time t. Let us assume that the intertemporal evolution of f_t may be described in general by

$$f_{t+1} = \mathcal{M}(f_t, \varepsilon_{t+1}), \tag{2}$$

where $\mathcal{M}$ is an operator that maps (f_t, ε_{t+1}) to a probability measure, and ε_{t+1} is an appropriately defined stochastic function representing random shocks. E.g., the random growth model in Simon [27] may be considered as a special case of processes consistent with specification (2). We may estimate such a law of motion for the evolution of city sizes by estimating nonparametrically the probability distribution function of city i population in time $t+1$, conditional on its population at time t, $f(P_{i,t+1}|P_{i,t})$. Overman [20], Appendix C, presents technical issues necessary to establish that stochastic kernel estimation techniques may be used to estimate transition, and more generally mapping, probability functions.[9]

[8] Danny T. Quah's program tSrF is available at http://econ.lse.ac.uk/~dquah/tsrf.html.

[9] A more precise definition of the intertemporal evolution of f_t would be the following. Let $\mathbf{b}(R,\mathcal{R})$ the Banach space under the sup norm of bounded measurable functions on $(R,\mathcal{R})$. Given a stochastic kernel $\mathcal{M}$, define the operator T mapping $\mathbf{b}(R,\mathcal{R})$ to itself by:

$$\forall f \in \mathbf{b}(R,\mathcal{R}), \forall y \in \mathcal{R} : (Tf)(y) = \int f(x)\mathcal{M}(y,dx).$$

$\mathcal{M}(y,\cdot)$ is a probability measure, and therefore the image Tf is the forwards conditional expectation. In our case, when the kernel is applied to city sizes in time t, $P_{i,t}$, it is the conditional distribution of $P_{i,t+1}$, given $P_{i,t}$. For more details, see tSrF*Manual* in http://econ.lse.ac.uk/~dquah/tsrf.html.

We estimate stochastic kernels for the cross-sectional evolution of the city size distribution which mitigate some of the problems described above. Figure 2 presents nonparametrically estimated kernels for transition probabilities along the lines of Equ. (2) above, obtained by using the techniques developed by Quah, *op. cit.*. The underlying data at time t are the (logarithm of) population of each city relative to the mean city size at time t. These results illuminate the extent of mobility, as we discuss shortly below.

The specifics of the estimation of stochastic kernels are as follows. First, we derive a non–parametric estimate of the joint distribution of the population of city i in two successive periods, $f(P_{i,t}, P_{i,t+1})$, where $P_{i,s}$ is city i population at time s.[10] We then numerically integrate under this joint distribution with respect to $P_{i,t+1}$ to get the marginal distribution of population at time t, $f(P_{i,t})$.[11] Next we estimate $f(P_{i,t+1}|P_{i,t})$, the distribution of population size in a given period conditional on population size in the previous period, by dividing through $f(P_{i,t}, P_{i,t+1})$ by $f(P_{i,t})$:

$$\hat{f}(P_{i,t+1}|P_{i,t}) = \frac{\hat{f}(P_{i,t}, P_{i,t+1})}{\hat{f}(P_{i,t})}. \tag{3}$$

Under regularity conditions, this gives us a consistent estimator for the conditional distribution. See Rosenblatt [26], Silverman [27] and Yakowitz [29] for details. The stochastic kernels plot this conditional distribution for all values of $P_{i,t}$.[12]

We report in Figure 2 a selection of our results, from the periods 1910–1920 and 1980–1990. To interpret the diagram, take a cross–section from any point on the 1910 axis, parallel to the 1920 axis. The cross–section is the distribution of relative city sizes in 1920 conditional on the city size in 1910. For example, cutting across from zero in 1910 gives us the distribution of city sizes in 1920, conditional on the city having mean size in 1910.[13] Extreme persistence would imply that all cities of (say) twice mean size in 1910, would be approximately twice mean size in 1920. In this case, the conditional distribution for twice mean size cities would be very tightly centred around log 2. In contrast, no persistence in city sizes would imply that cities of (say) twice mean size could occupy any point in the distribution in 1920. In this case, the conditional distribution for twice mean

[10] All densities are calculated nonparametrically using a Gaussian Kernel with bandwidth set as per section 3.4.2 of Silverman [26]. The range is restricted to the positive interval using the reflection method proposed in Silverman [26]. Calculations were performed with Danny Quah's tsrf econometric shell (available from http://econ.lse.ac.uk/ dquah/).

[11] We could also estimate the marginal distribution $f(P_{i,t})$ using a univariate kernel estimate. The asymptotic statistical properties of both estimators are identical, and in practice tend to produce very similar estimates.

[12] We note that this approach admits the nonparametric conditional mean estimate $P_{i,t+1} = g(P_{i,t}) + v_{i,t}$ as a special case, although it does not give actual numerical estimates.

[13] Recall that our underlying data are the (logarithm of) population of each city relative to the mean city size at time t.

size cities would just be the (unconditional) distribution of city sizes. Thus, extreme persistence is characterised by a stochastic kernel tightly centred around the diagonal. Extreme mobility would be characterised by a stochastic kernel centred around zero.

Figure 2 shows that city sizes are highly persistent. Nearly all the mass is concentrated around the diagonal implying that, from decade to decade, cities hardly move relative to the mean. This is even clearer if one looks at the contour plots presented in Figure 2. These contour plots work exactly like the contours on a more standard map, connecting points at similar heights on the stochastic kernel. The contour plots clearly demonstrate the high degree of persistence, reflected in the concentration of the mass around the diagonal, at both the beginning and end of the period.

Figure 2 shows that there is almost the same degree of mobility at both the beginning and end of the sample. We cannot directly test for the stationarity of the underlying markov process in our nonparametric setting, although χ^2 tests performed by Black and Henderson [5] do not reject stationarity. If so, we can pool across time periods to get a better estimate of the underlying transition process. Figure 3 shows such pooled transition kernels, for the entire US and the pooled data for regions, where each city population is taken relative to the regional mean.

These stochastic kernels give a pictorial representation which allows us to compare mobility across samples and time periods. They suffer from two problems, however. First, when estimated for smaller samples, the degree of precision is reduced, giving the appearance of more mobility. Second, they do not give us statistics with which to compare mobility across different samples. We could discretize the state space, estimate transition matrices and calculate the standard mobility indices – but then we are back to the problems with the previous literature that we have highlighted above.

Instead, we proceed by reporting on Figure 4 the "cross profile" plots, a graphical device proposed by Dolado *et al.* [9] and also used by Quah [24]. The top left hand corner of Figure 4 shows such a cross–profile plot for all the cities that exist in 1920. The vertical axis is the logarithm of the city size relative to the mean. The horizontal axis marks city rank according to size in 1920. Reading upward from the bottom of the figure, the plots are for 1920, 1940, 1960 and 1980 respectively. For 1920, cities have been ranked in order of increasing size thus the cross–profile plot is monotone rising. We then maintain the same horizontal ordering of cities for each of the plots in subsequent years. That is, the ordering of cities is fixed according to their 1920 rankings. As cities change rankings year to year, the plots cease to be monotone rising and instead become "jagged". Thus, the extent of choppiness (or jaggedness) depends on the degree of intra–distribution mobility. The shape of the plots gives us information on both types of mobility that we discussed above. If the cross–profile plots were always monotone rising, but the slope increased over time, then city population ranks are invariant, but the spread of the distribution is increasing. If the cross profile plot becomes jagged, then cities are changing rankings over time.

We add precision by calculating a number of statistics which capture features of the changes in the distributions that we have described above. We report these statistics in Tables 3, 5 and 7. The first measure, *Slope,* gives the OLS estimated slope of the *resorted*[14] cross–profile plot at each point in time, that is, the slope of the regression of the logarithm of city size against its rank. By resorting the data in each year and by comparing the respective estimates, we may capture the extent to which the spread of the distribution is changing over time. This gives an idea of the degree of changes in inequality in the city size distribution. When *Slope* has the value v, then being 10 cities larger means having a population e^{10v} times higher. For the whole sample, this measure decreases slightly over time, but stabilizes towards the end of the period.

Additional insight on the degree of intra–distribution mobility is provided by two new measures, *SerCorr* and *Variation*. They are intended to capture the changing choppiness of the cross–profile plot and are defined as follows. Let (r) denote the ranking in 1920 when the cities are ordered in terms of increasing size. Thus, $r = 1$ for the smallest city; $r = 2$ for the second smallest city etc. Then, for each period, *SerCorr* is the first–order serial correlation coefficient of sequential changes across this ordering. If $p_{(r)}$ is the relative population of the city with rank r, then sequential changes are defined as the difference in relative sizes of the cities with two successive rankings:

$$\Delta^* p_{(r)} = p_{(r)} - p_{(r-1)}. \tag{4}$$

Then, *SerCorr* is defined as the“serial” (along the ranking in 1920) correlation coefficient:

$$SerCorr = \frac{\sum_r (\Delta^* p_{(r)} - E[\Delta^* p])(\Delta^* p_{(r-1)} - E[\Delta^* p])}{\sum_r (\Delta^* p(r) - E[\Delta^* p])^2}, \tag{5}$$

where $E[\Delta^* p]$ is the average of $\Delta^* p_{(r)}$ across all rankings. As with all correlation coefficients, the definition of *SerCorr* ensure that it lie between -1 and $+1$. If the cross–profile plot is a straight line, then *SerCorr* is zero one regardless of the slope of that cross–profile plot.[15] Usually, *SerCorr* differs

[14]That is, OLS is done after sorting cities according to their current ranks. This is necessary, because the smallest (largest) city in 1920, is not necessarily the smallest (largest) city in later decades.

[15]To see this, note that the slope of the cross-profile at any point is proportional to $\Delta^* p_{(r)}$ by definition. Thus a constant slope implies $\Delta^* p_{(r)} = \Delta^* p_{(r-1)} = E[\Delta^* p]$.

from one, because the relative sizes do not differ uniformly across the ranking. If the cross-profile plot is monotone rising and convex, then *SerCorr* is positive.[16] If the cross-profile plot is increasing and concave, then *SerCorr* will be negative. It becomes smaller (or more negative) when the choppiness of the cross–profile plot increases.

The other measure of intra–distribution mobility is *Variation*, defined as the mean–square variation in relative "growth rates" across two pairs of successive rankings. In contrast, the variance sums up squared deviations from the mean, and *SerCorr* sums up the products of deviations from the mean for all pairs of successive rankings. That is, if N is the number of cities, then:

$$Variation = \frac{1}{N-2} \sum_{r} (\Delta^* p_{(r)} - \Delta^* p_{(r-1)})^2. \tag{6}$$

Variation is non–negative and becomes larger with increasing variance of the cross–profile plot. As with *SerCorr*, *Variation* is zero when the cross-profile is a straight line, that is, when the relative size of cities with successive rankings are constant, regardless of the slope of the profile. However, it will be positive for all other cross–profile plots, regardless of whether or not they are monotonically increasing.

Table 3 provides these summary statistics for the cross–profile of all cities existing in 1920. The intra–distribution dynamics for the whole sample settle down rapidly: *SerCorr* has value -0.612, -0.619 and -0.609 in 1960, 1980 and 1990 respectively. One can see from the cross–profile plot that this does not mean that the profile is actually frozen in time. Rather, the ongoing "churning " of the distribution, that is the changing in the rankings of cities, has characteristics that are stable. This is consistent with our earlier observation on the stationarity of the Markov process for city transitions. However, now we are directly examining the mobility properties of the entire distribution. Notice that *Variation* shows ongoing increase over time, which reflects increasing amplitudes of changes in relative sizes, as evidenced by the cross-profile plots in the upper left hand corner of Figure 4. However, the estimated slope of the cross-plot diminishes over time, implying that the relative sizes across successive rankings decrease over time.

Our results for the cross–profile of cities that exist in 1920 suggests that the churning characteristics of the distribution are relatively stable over time and parameterised by a value of *SerCorr* around -.6. Both these statistics and the estimated stochastic kernels indicate the degree of mobility that characterises the evolution of the US city size distribution. Models that seek to explain the

[16] Again, because slope of the cross-profile at any point is proportional to $\Delta^* p_{(r)}$, the convexity implies that $\Delta^* p_{(r)}$ is increasing across the sample.

evolution of that distribution could use these figures as upper bound benchmarks.[17] These tools can also be used to compare the mobility patterns of different groupings of cities. It is to this issue that we now turn.

4.2 Regional urban sub–systems

A key issue, seldom addressed in the rank–size literature, is the appropriate definition of the urban system. Our approach allows us to characterise the degree of mobility within different urban sub–systems. Here, we demonstrate the technique by considering the evolution of nine subsystems defined by the US Census regions.[18] Regional analysis of the US system of cities is particularly interesting in view of US economic history. Not all of the continental U.S. was settled at the same time, and urban development since the beginning of the twentieth century has sharp regional patterns. Note that Kim [15] argues that the census regions are likely to serve well as economic regions.[19] Kim [16] relates regional economic patterns to changes in cities' industrial employment shares and other characteristics. Table 4 provides summary stastics for the regional urban systems.

The picture on the right hand side of Figure 3 shows the stochastic kernel for the evolution of city size relative to the average city size of cities in the same region. This stochastic kernel is estimated assuming that the transition process is stationary over time and identical across regions. This allows us to pool observations across both dimensions. It appears that the pattern of mobility of cities within their regional subsystems is not much different from the pattern of mobility relative to the US-wide average city size. However, remember that this result is conditional on the assumption that we can pool observations across both regional subsystems and across time. The results of the cross–profile plots suggest that this is not a valid assumption. In fact, there may actually be substantial differences between regional subsystems.

Figure 4 shows the cross–profile plots for eight of the nine regions.[20] The cross–profile plots are

[17]We would argue upper bound, as the actual urban system is hit by shocks that presumably increase mobility relative to the underlying economic mechanisms captured by current theoretical models.

[18]The nine regions are New England(ned); Middle Atlantic (mad); South Atlantic (sad), East South Central (escd); East North Central (encd), West North Central(wncd); West South Central (wscd); Mountain (mtd); Pacific (pad). These regions may not correspond exactly to functional urban sub–systems at all times during the period of study. However, they provide a convenient division that allows us to demonstrate the general approach. Their geographic definitions are indicated on Figure 1.

[19]Kim [15], p. 7–9, discusses the original intention of the definition of U.S. regions as delineating areas of homogeneous topography, climate, rainfall and soil, but subject to requirement that they not break up states. By design, the definitions were particularly suitable for agriculture and resource-based economies. The role of those industries as inputs to manufacturing would make them likely to serve well as economic regions.

[20]We exclude the Mountain region which only has 4 cities in 1920.

for the years 1920, 1940, 1960 and 1980 as before. Because of the varying numbers of cities in each region the plots are hard to compare visually. However, some stark differences do immediately jump out. For example, compare the cross–profile plots for the South Atlantic and Mid Atlantic regions. Both regions have similar numbers of cities[21], but the transition dynamics appear very different. The measures in Table 5 allow a more direct comparison. For example, we see that the visual impression that the Mid Atlantic region shows more churning than the South Atlantic, is actually driven by higher variation, rather than increased churning. Thus, *SerCorr* has similar values for the two regions, but *Variation* is much higher for the South Atlantic. We know that the South Atlantic has gained, in the second half of the twentieth century, larger cities relative to the Mid Atlantic. To take another example, we see that the West South Central and West North Central regions show a higher level of churning than all the other regions. In contrast to other US regions, which typically experienced roughly monotonic changes in their shares of larger cities, those two regions saw their shares increasing and then subsequently decreasing during the study period. The cross–profile plots suggest some evidence that there are differences in intra–distribution mobility within the different regional subsystems. Some areas of the US have urban systems that are characterised by far higher intra–distribution mobility.

4.3 City tiers

Classical hierarchical theories of cities divide cities in to tiers, depending on the functions of each city. More recent theoretical work has incorporated insights from this older literature in to the new economic geography literature. These theoretical analyses suggest that the highest-tier cities, which are most diversified, may display different patterns of evolution from lower tier cities. See Fujita, Krugman and Venables [13] for details. In this section, we examine whether the intra–distribution dynamics do appear to differ substantially among tiers.

In order to construct the tiers, we took as our basic classification a listing of US cities by "function" (nodal centers) from Knox [17]. We amended the top tier slightly to include Atlanta, Chicago, Denver, Houston, Los Angeles, New York City, Miami, San Francisco, Seattle and Washington D.C.. The next classification is the regional nodal centres, which includes fourteen large cities: Baltimore, Boston, Cincinnati, Cleveland, Columbus, Dallas, Indianapolis, Kansas City MO, Minneapolis, New Orleans, Philadelphia, Phoenix, Portland OR, and St. Louis. The

[21] 24 and 21 respectively.

third classification is the sub–regional nodal centres. This comprises nineteen cities: Birmingham, Charlotte, Des Moines, Detroit, Hartford, Jackson MS, Little Rock, Memphis, Milwaukee, Mobile, Nashville, Oklahoma City, Omaha, Pittsburgh, Richmond, Salt Lake City, Shreveport, Syracuse and Tampa. The remaining cities are placed in the lowest tier. Table 6 gives summary statistics for each tier.[22]

From the classification, it is clear that the number of cities in the different tiers differs substantially between tiers. Thus the top tier comprises ten cities, the second tier fourteen cities, the third tier nineteen cities and the lowest tier the remaining 291 cities. With such small numbers of cities within the top three tiers, it makes no sense to calculate stochastic kernels for each tier. Instead, in Figure 5 we show the cross–profile plots for each of the four tiers. Table 7 gives the corresponding measures.

The table shows that the top tier actually shows a surprising degree of mobility. The top tier exhibits, consistently over the century, the largest estimated slopes, which imply that relative sizes increase with rankings. Looking at the cross–profile plot suggests that this mobility is mainly due to changes in the relative sizes and rankings of cities at the lower end of the tier. By 1940, the rankings of the top four cities appear set, although they still display mobility with respect to relative sizes.[23] For the bottom five cities, there is a surprising degree of mobility both in terms of rankings and relative sizes. Results for the second tier are again surprising. It is actually this second tier of cities that show remarkable stability, both in terms of relative size and rankings. The measures and the shape of the cross-profile plot show that this is easily the most stable subsystem that we have studied. Mobility patterns for the third tier lie somewhere between the first and second. Finally, the fourth tier shows the highest degree of mobility. In standard analysis using transition probability matrices, nearly all the action for the top three tiers would be disguised by the fact that they all fall in the top discrete state. Our results here suggests that there are interesting differences in mobility for subsystems of cities that usually fall within this highest state.

[22] The time-invariant nature of this classification is particularly problematic in certain instances. For example, despite its high population, San Diego would be in tier four in the early years, after having entered in 1910, but should be in a higher tier later on in the century. However, geographers might consider San Diego as rather special, because it has functioned in a very specialized way as a naval base before its more diversified growth in recent years. It is for this reason that it is in tier four in this analysis.

[23] Relative sizes are now defined with respect to the average city size for cities in the same tier.

5. Conclusions

This paper has studied a number of aspects of intra–distribution mobility for the US city size distribution. Characterising the nature of such intra–distribution mobility should help guide the two different theoretical strands that seek to explain the evolution of urban systems. For the literature that attempts to generate urban systems that obey the rank size rule, these results provide benchmarks for the upper level of intra–distribution mobility that would ensure these models are consistent with real world intra–distribution dynamics. The results on regional subsystems and urban hierarchies also prompt questions for the literature that tries to model the economic mechanisms that may govern the evolution of urban systems. Are there economic forces that can explain the apparent differences in the nature of intra–distribution mobility between different regional sub–systems? More interestingly, what explains different patterns in churning and changes in relative rankings within groups of cities at different levels of the US urban hierarchy?

6. References

1. G. R. P. Allen, The 'courbe des populations:'a Further Analysis, *Bulletin of the Oxford University Institute of Statistics,* 16, 179–189 (1954).

2. G. Alperovich, An Explanatory Model of the City-Size Distribution: Evidence from Cross-country Data, *Urban Studies*, 30, 1591–1601 (1993).

3. F. Auerbach, Das Gesetz der Bevölkerungskonzentration, *Petermanns Geographische Mitteilungen*, 59, 74–76 (1913).

4. B. J. L. Berry, City Size Distributions and Economic Development, *Economic Development and Cultural Change,*9, 593–587 (1961).

5. D. Black and J. V. Henderson, Urban Evolution in the USA, processed, London School of Economics (1999).

6. D. J. Bogue, "Population Growth in Standard Metropolitan Areas 1900-1950," Scripps Foundation in Research in Population Problems, Oxford, Ohio (1953).

7. G. Carroll, National City Size Distributions: What Do we Know after 67 Years of Research, *Progress in Human Geography,* 6, 1–43 (1982).

8. P. Cheshire, Trends in Sizes and Structure of Urban Areas, in "Handbook of Regional and Urban Economics" (P. Cheshire, and E. S. Mills, eds.), Elsevier Science, B. V., Amsterdam (1999).

9. J. J. Dolado, J. M. Gonzales-Paramo, and J. M. Roldan, Convergencia Economica entre las Provincias Espanolas: Evidencia Empirica (1955–1989), *Monedo y Credito,* 198, 81–119 (1994).

10. L. H. Dobkins and Yannis M. Ioannides, Dynamic Evolution of the U.S. City Size Distribution, in "The Economics of Cities" (J.-F. Thisse and J.-M. Huriot, eds.), Cambridge University Press, Cambridge (2000).

11. J. Eaton and Z. Eckstein, Cities and Growth: Theory and Evidence from France and Japan," *Regional Science and Urban Economics,* 27, 4–5, 443–474, (1997).

12. M. Fujita, P. R. Krugman and A. J. Venables, "The Spatial Economy: Cities, Regions and International Trade," MIT Press, Cambridge, MA (1999).

13. X. Gabaix, Zipf's Law for Cities: An Explanation, *Quarterly Journal of Economics,* CXIV, 739–767 (1999).

14. S. Kim, Economic Integration and Convergence: U.S. Regions, 1840–1987,National Bureau of Economic Research, working paper No. 6335, December (1997).

15. S. Kim, Urban Development in the United States, 1690–1990,National Bureau of Economic Research working paper No. 7120, May (1999).

16. P. L. Knox, "Urbanization: An Introduction to Urban Geography," Prentice Hall, Englewood Cliffs, NJ (1994).

17. P. R. Krugman, "The Self-Organizing Economy," Blackwell Publishers, Oxford (1996).

18. C. H. Madden, On Some Indications of Stability in the Growth of Cities in the United States, *Economic Development and Cultural Change,* IV, 3, 236–252 (1956).

19. H. G. Overman, "Empirical Studies on the Location of Economic Activity and its Consequences," Ph. D. dissertation, submitted to the University of London (1999).

20. D. T. Quah, Empirical Cross-section Dynamics and Economic Growth, *European Economic Review*, 37(2/3), 426–434 (1993).

21. D. T. Quah, Regional Convergence Clusters across Europe, *European Economic Review,* 40, 951–958 (1996).

22. D. T. Quah, Empirics for Growth and Distribution: Stratification, Polarization and Convergence Clubs, *Journal of Economic Growth,* 2, 1, 27–59 (1997a).

23. D. T. Quah, Regional Cohesion from Local Isolated Actions: Historical Outcomes, discussion paper 378, Centre for Economic Policy Research, London, (1997b).

24. K. Rosen and M. Resnick, The Size Distribution of Cities: an Examination of the Pareto Law and Primacy, *Journal of Urban Economics,* 8, 165–186 (1980).

25. M. Rosenblatt, Curve Estimates, *Annals of Mathematical Statistics,* 42, 1815–1842 (1971).

26. B. W. Silverman, “Density Estimation for Statistics and Data Analysis,” Chapman and Hall, London (1986).

27. H. Simon, On a Class of Skew Distribution Functions, *Biometrika*, 44, 425–440 (1955).

28. S. Yakowitz, Nonparametric Density Estimation, Prediction, and Regression for Markov Sequences, *Journal of American Statistical Association,* 80, 215–221 (1985).

29. G. K. Zipf, “Human Behavior and the Principle of Least Effort,” Addison-Wesley, Cambridge, Massachusetts (1949).

Table 1. Descriptive statistics: decennial data 1900–1990

1	2	3	4	5	6	7	8
Year	US Pop. (000)	US Urban Pop. (000)	Mean Size	Median Size	GNP billion $	Distance miles	Nearest miles
1900	75,995	29,215	259952	121830	71.2	802.5	70.9
1910	91,972	39,944	286861	121900	107.5	863.8	68.3
1920	105,711	50,444	338954	144130	135.9	864.0	66.2
1930	122,775	64,586	411641	167140	184.8	876.9	64.8
1940	131,669	70,149	432911	181490	229.2	884.9	64.4
1950	150,697	85,572	526422	234720	354.9	890.8	65.3
1960	179,323	112,593	534936	238340	497.0	940.4	56.9
1970	203,302	139,419	574628	259919	747.6	981.3	52.5
1980	226,542	169,429	526997	232000	963.0	998.7	45.9
1990	248,710	192,512	577359	243000	1277.8	1005.3	45.5

All figures are taken from *Historical Statistics of the United States from Colonial Times to 1970*, Volumes 1 and 2, and *Statistical Abstract of the United States, 1993.* Column 6: GNP adjusted by the implicit price deflator, constructed from sources above; 1958=100. Column 7 gives the average distance to all other cities. Column 8 gives the average distance to the nearest city. Distances are calculated as great circle distances based on latitudes and longitudes from the Times Atlas 1997 edition.

Table 2. Descriptive statistics for all cities – 1990

Variable	Mean	Std. Dev.	Skewness	Kurtosis	Min	Max
Population (000)	479.5	1001.5	6.6	58.8	50.7	9,372.0
Log(Population)	12.4028	0.9895	1.0	4.1	10.8343	16.374
Growth Rate (%)	10.62	41.98	-1.1	5.8	-.999	1.8752
Education (%)	57.1085	20.9284	-0.4	1.8	11.80	92.73
Real Wage ($)	3197.92	1132.37	0.2	2.3	1020.00	7311.00
New England	8.8					
Mid Atlantic	12.8					
South Atlantic	16.7					
East North Central	20.3					
East South Central	6.6					
West North Central	9.1					
West South Central	12.2					
Mountain	4.6					
Pacific	8.8					

Data on education and real wage are taken from *Historical Statistics of the United States from Colonial Times to 1970, Vol.1 and 2, and* Statistical Abstract of the United States, 1993.Educational percentage refers to the mean percent of 15 to 20 age cohort in school. Mean real annual earnings, by city proper or metro area, are in dollars, deflated by the Consumer Price Index,1967 = 100.

	Slope	*SerCorr*	*Variation*
1920	0.020	0.391	0.063
1940	0.019	−0.479	0.513
1960	0.017	−0.612	1.324
1980	0.015	−0.619	1.708
1990	0.015	−0.609	1.924

Table 3. Whole sample cross profile statistics

	East North Central	East South Central	Mid Atlantic	Mountain	New England	Pacific	South Atlantic Central	West North Central	West South
No. cities									
1900	25	9	20	2	12	7	17	13	7
1910	35	9	21	3	12	12	22	14	11
1920	36	10	21	4	12	12	24	15	15
1930	36	11	21	5	12	12	26	16	18
1940	36	11	21	6	12	12	27	16	19
1950	36	11	21	6	12	12	27	16	21
1960	41	13	24	13	22	15	33	18	31
1970	47	15	25	14	25	23	38	20	36
1980	58	20	35	18	28	35	60	26	42
1990	58	23	36	21	28	36	62	27	43
Mean pop.									
1900	233,300	121,400	561,100	120,000	330,200	166,700	149,700	188,800	99,820
1910	232,700	151,500	704,400	143,500	403,100	213,100	153,000	228,600	114,000
1920	309,100	164,000	835,300	151,400	469,800	303,400	181,900	257,200	135,000
1930	410,800	201,000	1,019,000	163,200	526,800	482,200	212,100	287,700	178,400
1940	431,400	227,200	1,071,000	166,800	541,600	570,000	247,300	305,200	196,100
1950	511,600	293,700	1,180,000	248,300	598,300	868,900	329,100	357,500	259,000
1960	577,200	320,300	1,162,000	257,600	336,000	1,099,000	400,700	408,300	292,300
1970	617,500	349,700	1,224,000	336,700	341,600	985,200	470,200	436,000	335,000
1980	546,500	378,000	920,500	398,400	337,900	812,500	436,300	386,000	397,000
1990	546,200	367,900	951,900	437,100	366,600	985,500	525,900	407,900	455,000

Table 4. Regions

	Slope	*SerCorr*	*Variation*
East North Central			
1920	0.084	0.279	0.169
1940	0.087	0.117	0.276
1960	0.085	−0.355	0.589
1980	0.083	−0.529	0.788
1990	0.085	−0.544	0.867
East South Central			
1920	0.213	−0.304	0.185
1940	0.203	−0.276	0.209
1960	0.190	−0.879	0.468
1980	0.148	−0.296	0.396
1990	0.142	−0.397	0.449
Mid Atlantic			
1920	0.159	0.448	0.250
1940	0.154	0.240	0.388
1960	0.146	−0.013	0.612
1980	0.134	−0.355	0.945
1990	0.128	−0.272	0.983
North East			
1920	0.211	−0.025	0.497
1940	0.212	0.019	0.549
1960	0.241	−0.346	1.206
1980	0.219	−0.352	1.210
1990	0.226	−0.400	1.332
Pacific			
1920	0.242	−0.427	0.581
1940	0.222	−0.461	1.117
1960	0.162	−0.243	1.318
1980	0.120	−0.216	1.395
1990	0.095	−0.237	1.489
South Atlantic			
1920	0.090	0.376	0.148
1940	0.082	0.048	0.329
1960	0.079	−0.254	0.882
1980	0.073	−0.275	1.265
1990	0.074	−0.304	1.549
West North Central			
1920	0.198	−0.071	0.308
1940	0.193	−0.296	0.484
1960	0.188	−0.672	1.120
1980	0.184	−0.622	1.255
1990	0.188	−0.671	1.579
West South Central			
1920	0.120	−0.154	0.222
1940	0.134	−0.857	0.781
1960	0.148	−0.821	1.032
1980	0.164	−0.797	1.539
1990	0.172	−0.713	1.601

Table 5. Sub–regions cross profile statistics

	top tier	second tier	third tier	fourth tier
No. cities				
1900	9	13	19	74
1910	9	13	19	98
1920	9	14	19	107
1930	10	14	19	114
1940	10	14	19	117
1950	10	14	19	119
1960	10	14	19	167
1970	10	14	19	200
1980	10	14	19	279
1990	10	14	19	291
Mean pop.				
1900	974,300	588,200	210,200	127,500
1910	1,384,000	736,400	247,600	138,400
1920	1,755,000	826,800	334,100	156,400
1930	2,168,000	988,300	449,900	180,000
1940	2,406,000	1,039,000	487,700	190,400
1950	2,975,000	1,233,000	590,400	229,800
1960	3,368,000	1,495,000	747,900	262,200
1970	3,885,000	1,795,000	872,300	294,400
1980	3,980,000	1,928,000	994,700	300,100
1990	4,526,000	2,121,000	1,084,000	332,000

Table 6. Tiers

	Slope	*SerCorr*	*Variation*
top tier			
1920	0.431	0.050	0.560
1940	0.394	−0.353	1.059
1960	0.302	−0.441	1.077
1980	0.191	−0.475	1.040
1990	0.169	−0.546	1.094
second tier			
1920	0.212	0.032	0.286
1940	0.181	−0.247	0.306
1960	0.123	−0.204	0.346
1980	0.073	−0.162	0.404
1990	0.055	−0.191	0.497
third tier			
1920	0.150	0.360	0.243
1940	0.132	−0.025	0.465
1960	0.109	−0.232	0.782
1980	0.082	−0.426	1.073
1990	0.073	−0.462	1.163
fourth tier			
1920	0.019	−0.334	0.036
1940	0.017	−0.625	0.491
1960	0.012	−0.649	1.279
1980	0.010	−0.636	1.597
1990	0.010	−0.632	1.828

Table 7. Functional tiers cross profile statistics

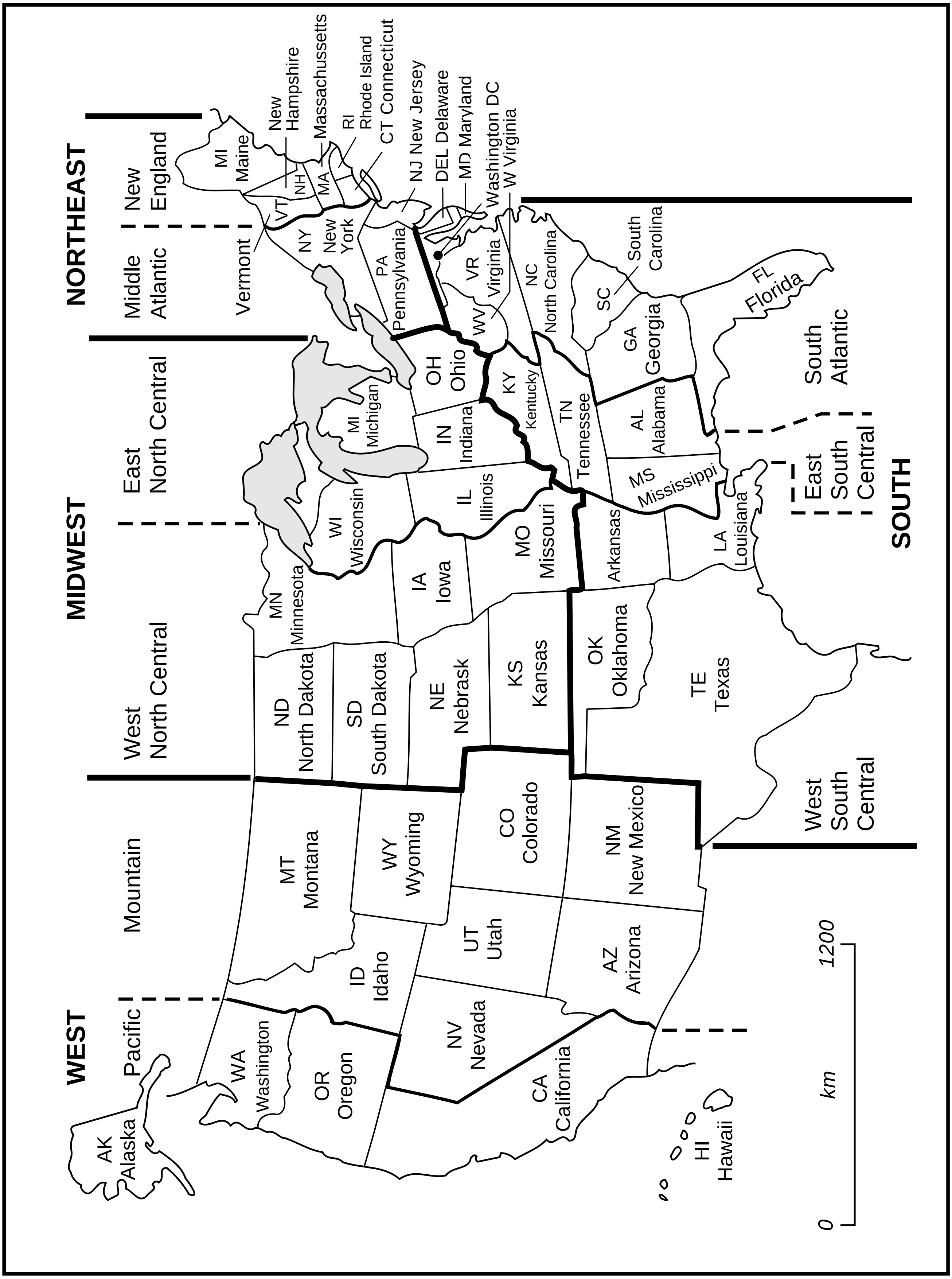

Figure 1. Census regions

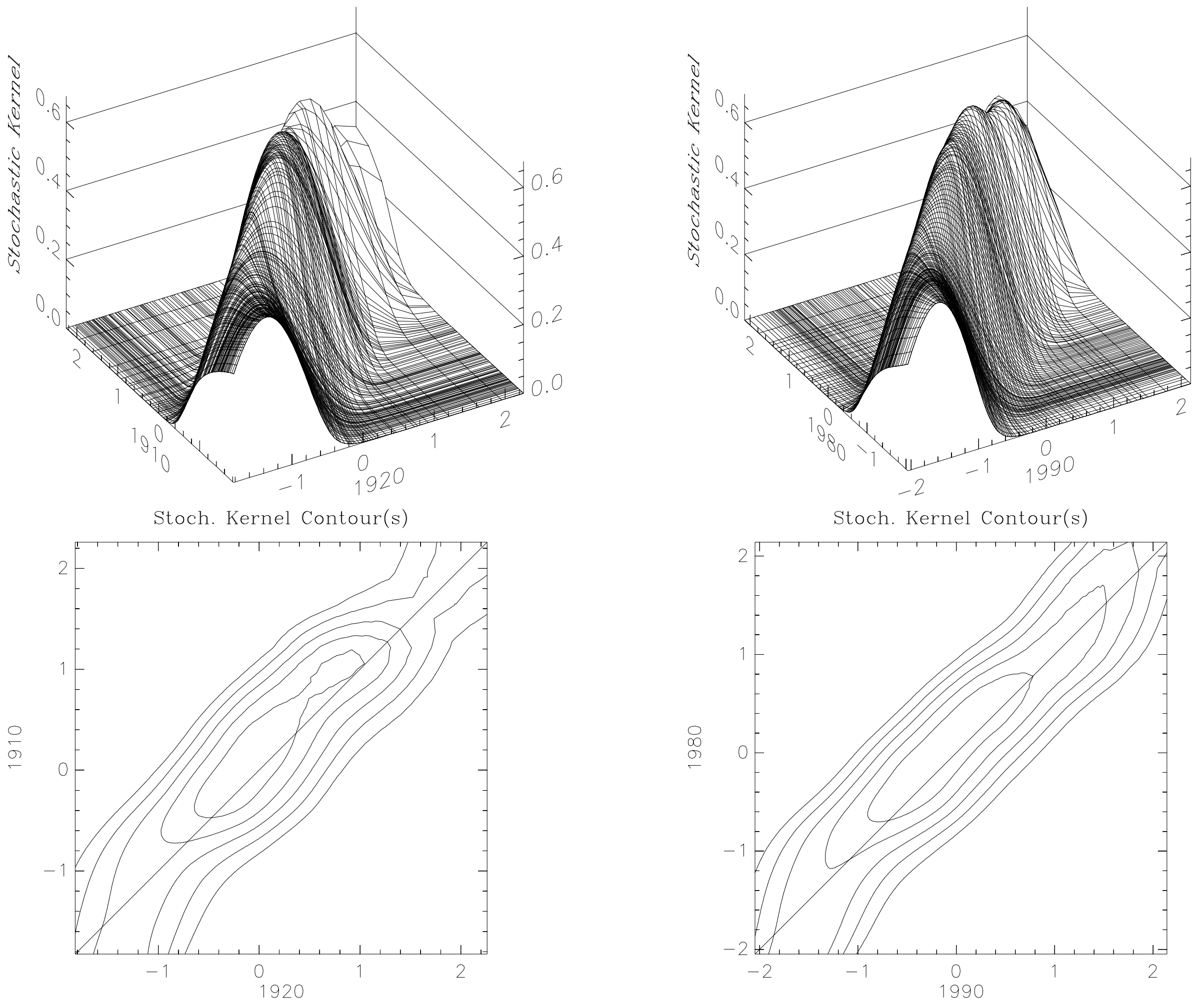

Figure 2. Selected decades transition kernels

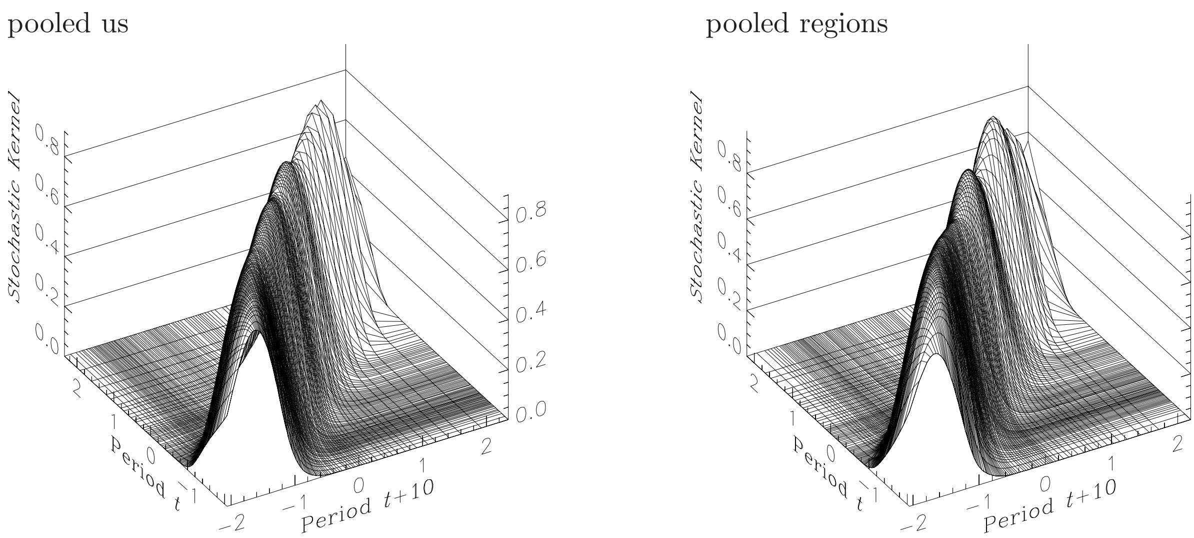

Figure 3. Pooled US and regions transition kernels

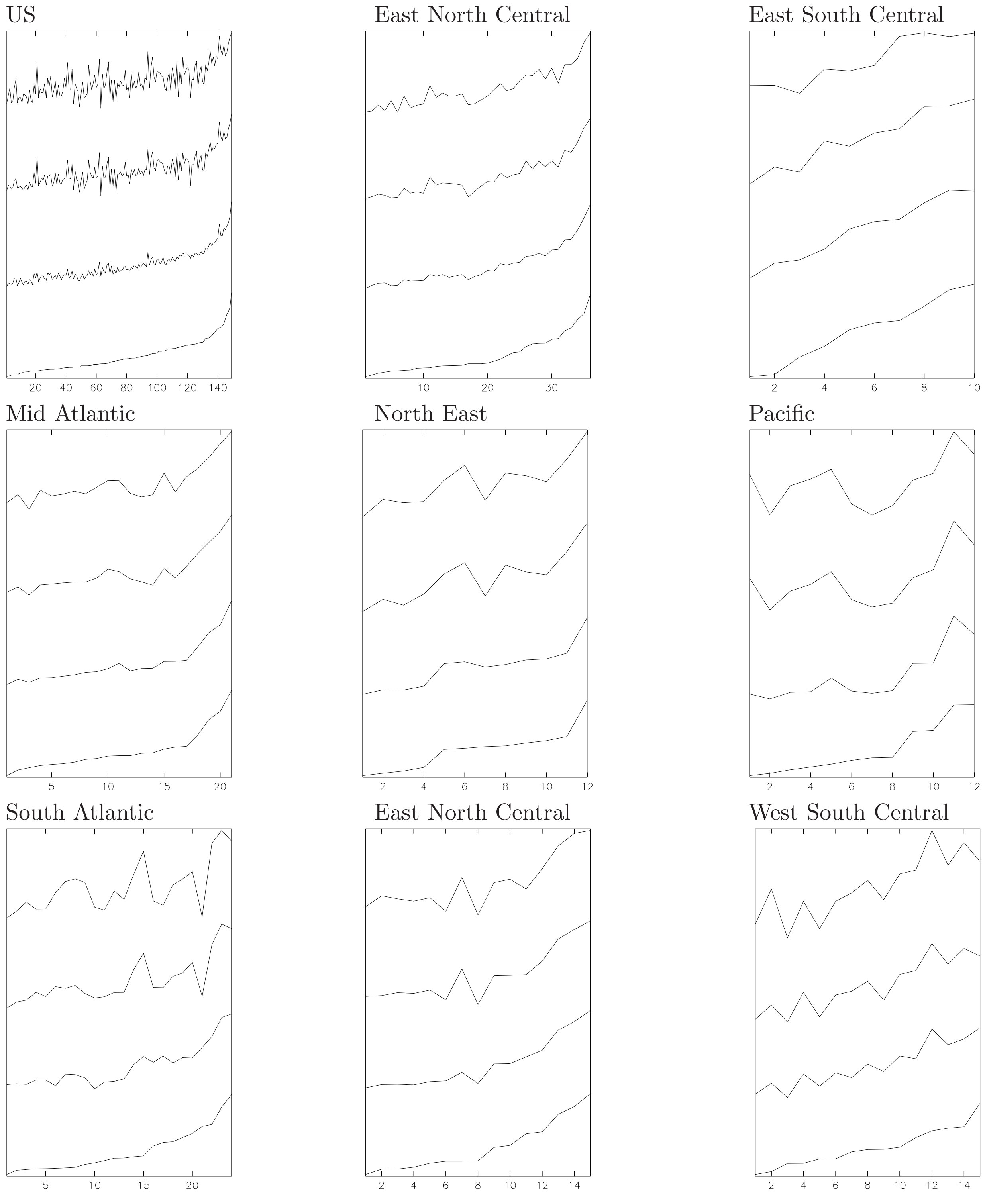

Figure 4. US and US Census–regions cross profile plots

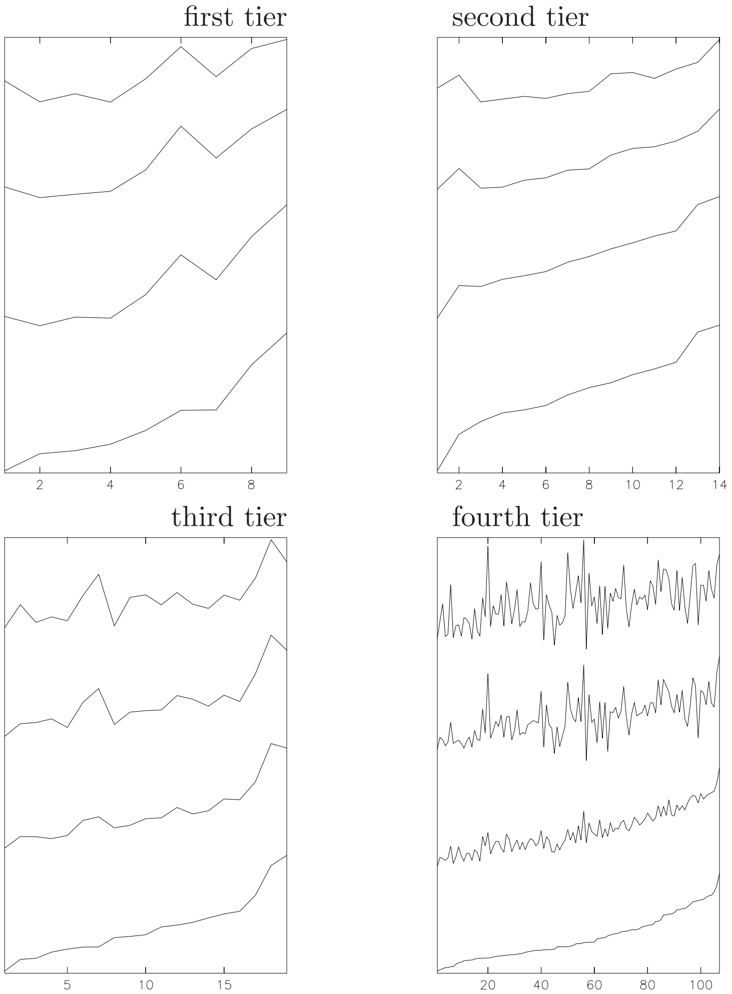

Figure 5. Functional tiers cross profile plots