Journal of Economic Literature
Vol. XXXVI (September 1998), pp. 1426–1464

Urban Spatial Structure

ALEX ANAS
RICHARD ARNOTT
and
KENNETH A. SMALL[1]

1. *Introduction*

AN INTERVIEW WITH Chicago's current mayor, Richard M. Daley:

> "New York is too big this way," the mayor says, raising a thick hand over his head. Stretching both arms out at his sides, he adds, "Los Angeles is too big this way. All the other cities are too small. We're just right." (Jeff Bailey and Calmetta Coleman 1996, p. 6)

Mayor Daley's remarks reflect a widespread fascination with the roles that urban size and structure play in people's lives. Academic as well as other observers have long sought explanations for urban development patterns and criteria by which to judge their desirability. Furthermore, as we shall see, understanding the organization of cities yields insights into economy-wide growth processes and sheds light on economic concepts of long-standing interest: returns to scale, monopolistic competition, vertical integration, technological innovation, innovation diffusion, and international specialization. Cities also are prime illustrations of some newer academic interests such as complex structural evolution and self-organization.

In this essay we offer a view of what economics can say about and learn from urban spatial structure. In doing so, we reach into neighboring disciplines, but we do not aspire to a complete survey even of urban economics, much less of the related fields of urban geography, urban planning, or regional science. Our focus is on describing and explaining urban spatial structure and its evolution.

This is a particularly interesting time to study urban structure because cities' growth patterns are undergoing qualitative change.[2] For two centuries at least, cities have been spreading out. But in recent decades, this process of decentralization has taken a more polycentric form, with a number of concentrated employment centers making their mark on both employment and population

[1] Anas: University of Buffalo; Arnott: Boston College; Small: University of California at Irvine.
Acknowledgments: The authors would like to thank John Pencavel, three referees, Robert Bacon, Amihai Glazer, Peter Gordon, Robert Johnston, Cassey Lee Hong Kim, and David Pines for helpful comments on earlier drafts, and Alexander Kalenik for assistance in the preparation of Figure 1. We also thank the University of California Transportation Center for financial assistance.

[2] Throughout this essay we use the word "city," or the name of a particular city, to mean an entire urban region; other terms with similar meanings are " metropolitan area" and "urban area."

distributions. Most of these centers are subsidiary to an older central business district (CBD), hence are called "subcenters." Some subcenters are older towns that gradually became incorporated into an expanded but coherent urban area. Others are newly spawned at nodes of a transportation network, often so far from the urban core as to earn the appellation "edge cities" (Joel Garreau 1991). There is some evidence, discussed later, that the employment centers within a given urban region form an interdependent system, with a size distribution and a pattern of specialization analogous to the system of cities in a larger regional or national economy.

At the same time, rampant dispersion of economic activity has continued outside centers altogether, prompting Peter Gordon and Harry Richardson (1996) to proclaim that Los Angeles, at least, is "beyond polycentricity." But even sprawl is far from homogeneous, and geographers have perceived patterns that conform to the mathematics of highly irregular structures such as fractals. Whether such irregularity is really new, or even increasing, is not so clear, as we shall see in the next section; but urban economics helps us understand the order that may be hidden in such patterns.

An important source of current change in urban structure is the changing economic relationships within and between firms. Telecommunications, information-intensive activities, deregulation, and global competition have all contributed to changes in the functions that firms do in-house, and in how those functions are spatially organized. Some internal interactions can now be handled via telecommunications with remote offices which already perform routine activities such as accounting. Some vertical interactions are now more advantageously made as external transactions among separate firms, possibly requiring even more frequent face-to-face communications because of the need for contracting. Allen Scott (1988, 1991) describes how such "vertical disintegration" has shaped the geographical structure of a number of industries in southern California, including electronics, animated films, and women's clothing. Meanwhile, firms are developing new interactive modes which are neither market nor hierarchy, but rather constitute what Walter Powell (1990) calls a "network" organizational form, characterized by "relationship contracting" and having unknown implications for locational propensities.

The research agenda that emerges from these observations is heavy on economies of agglomeration, a term which refers to the decline in average cost as more production occurs within a specified geographical area. One class of agglomeration economies is intrafirm economies of scale and scope that take place at a single location. Another class is positive technological and pecuniary externalities that arise between economic agents in close spatial proximity[3] due, for example, to knowledge spillovers, access to a common specialized labor pool, or economies of scale in producing intermediate goods. Agglomeration economies may be dynamic as well as static, and are suspected of giving cities a key role in generating aggregate economic growth (Jane Jacobs 1984; Edward Glaeser et al. 1992).

Any agglomerative or "centripetal" force, even one caused just by a unique resource such as a harbor, places a premium on land at certain locations. This encourages spatially concentrated capital formation (buildings) and accentu-

[3] Some authors reserve the term "agglomeration economies" only for this second class.

ates the need to produce at discrete points in space because of increasing returns to scale in production (David Starrett 1974). Because of these pervasive externalities and nonconvexities, economic analysis when applied to urban geography yields results that differ in important and interesting ways from results of other branches of economics. Agglomeration economies also create first-mover advantages and regional specializations that are important in international trade (Paul Krugman 1991a), and some first-mover disadvantages that prevent optimal dynamic growth paths from being realized. We discuss these in Section 5.

Agglomeration economies are, of course, not new. As eloquently expositied by Raymond Vernon (1960) and Benjamin Chinitz (1961), they are at the heart of our current understanding of central business districts. But recent changes in the technology of agglomeration, due to advances in information processing and telecommunications, may profoundly alter the pattern of spatial development (Jess Gaspar and Glaeser 1998). Understanding these new forces will help us understand newly emerging forms of urban structure as well as basic determinants of industrial structure and interregional and international trade.

While our focus is on explaining urban spatial structure as a result of market processes, we touch on two related issues as well. The first concerns the role of government. Government policies—notably land-use controls and the provision of transportation infrastructure—play a major role in shaping cities. What can we say about optimal policy? The second issue concerns the importance of space in economics. Accounting for location yields new insights into economic phenomena that are normally analyzed in aspatial models. But what is the level of spatial resolution at which such phenomena are best analyzed?

2. *History and Description of Urban Spatial Structure*

We begin with a sketch of how urban form has evolved in modern times, followed by some observations on how to measure its characteristics.

2.1. *Recent Evolution of Urban Form*

The spatial structure of modern cities was shaped, in large measure, by advances in transport and communication. The history of urban development in North America since colonial times allows us to document aspects of this process (Charles Glaab and Theodore Brown 1967).

Prior to about 1840, most cities were tied to waterways such as harbors, rivers, and canals. The average cost of processing freight fell sharply with the quantity processed at a particular port, creating substantial scale economies at harbors or river junctions with access to the sea. Similarly, as railroads competed with waterways later in the 19th century, scale economies in rail terminals created accessibility advantages near them as well. Meanwhile intra-urban freight transport took place mainly by horse and wagon, which was time consuming and unreliable in bad weather. These conditions favored the growth of a single manufacturing district located near the harbor or railhead, with residences surrounding it (Leon Moses and Harold Williamson 1967).

In the last quarter of the century, the telegraph greatly speeded the flow of information from city to city (Alexander Field 1992). But economies of scale prevented it from being used much within a city—instead, messengers remained the primary means of intra-city

business communication. The high cost of intra-urban communication meant that even light manufacturing and service industries tended to concentrate within the central manufacturing core, as shown for New York by Chinitz (1960). But this small core area was far from homogeneous; rather, it was divided into districts, each specialized in an activity such as commercial banking, pawnbrokerage, or light or heavy manufacturing. In late nineteenth-century Chicago, four-fifths of the city's jobs were located within four miles of State and Madison streets, according to Raymond Fales and Moses (1972), who go on to show how a pattern of specialized districts arose due to agglomerative forces within industries and the linkages among them.

Before 1850, personal transport within the city was mainly by foot and horse-drawn carriage, causing the great majority of rich and poor alike to live close to the city center. For the most part, the rich outbid the poor for the most central and hence most convenient sites, so that income declined markedly with distance from the CBD, as is documented in studies of Milwaukee, Pittsburgh, and Toronto (Stephen LeRoy and Jon Sonstelie 1983).

Between 1850 and 1900, the advent of horse-drawn and then electric streetcars enabled large numbers of upper- and middle-income commuters to move further out. This migration gave rise to "streetcar suburbs," residential enclaves organized around a station on a radial streetcar line (Sam Warner 1962). Toward the turn of the century, subways further contributed to this pattern in the largest cities. Thus developed a spatial structure now known as the "nineteenth century city," consisting of a compact production core surrounded by an apron of residences concentrated around mass transport spokes.

The next big changes were the dissemination of the internal combustion engine and the telephone in the early twentieth century. Gradually the horse and wagon were replaced by the small urban truck, and the messenger by the telephone. For example, in the single decade from 1910 to 1920, truck registrations in Chicago increased from 800 to 23,000, while horse-drawn vehicle registrations dropped almost by half. Moses and Williamson (1967) estimate that variable costs and travel time for the truck were less than half those for the horse and wagon. The truck and the telephone allowed businesses to spread outward from the center, thereby taking advantage of lower land values while maintaining their links to the central port or railhead. Thus central business districts expanded. In Chicago, firms that moved in 1920 located on average 1.5 miles from the core, as opposed to 0.92 miles in 1908 (Moses and Williamson 1967).

The automobile, at first restricted to richer families, rapidly increased in importance with assembly-line production of the Model T Ford starting in 1908. Cars broadened the coverage of motorized personal transport, causing the areas between the streetcar suburbs to be settled and the residential apron to expand. The automobile competed successfully with mass transit, despite transit fares remaining flat in nominal terms from the beginning of the century until approximately World War II; it did this mainly by providing speed, privacy, and convenience, although it was also facilitated by an active program of building and upgrading public roads (Paul Barrett 1983).

As assembly-line production became widespread, the lower capital-land ratios characterized by flat buildings increased the attractiveness of locations where land was cheap. Nevertheless,

even at mid-century many producers outside the core were bound to the central harbors and rail terminals for inter-city shipments. Eventually, however, this link was weakened by the creation of suburban rail terminals and the declining cost of inter-city trucking, the latter facilitated by the interstate highway system. These developments, coming primarily after World War II, enabled manufacturing to leapfrog out to the outermost suburbs. Central cities began their painful transition from manufacturing to service and office centers.

Due to the durability of the urban capital stock and urban infrastructure, cities in the modern American landscape bear proof of the lasting impacts of these developments. Large cities of the eastern seaboard and the midwest, such as Boston or Detroit, still contain streets and buildings dating from the heyday of their harbor and rail operations and from the subsequent era of radial mass transportation systems. Even Chicago, the great metropolis of the midwest, was first established as one of the last and western-most of the waterway cities—its later importance as a rail and air hub derived from its already well established position by the beginning of the rail era (William Cronon 1991). Further west, however, the spatial pattern of many urban settlements was first shaped by the railroad. Major cities, such as Oklahoma City, Denver, Omaha, and Salt Lake City, grew up around rail nodes and developed compact CBDs centered on rail terminals. In contrast, the even later automobile-era cities such as Dallas, Houston, and Phoenix have spatial structures determined mainly by the highway system. Los Angeles is an intermediate case: partly a western rail terminus and partly a set of residential communities populated by rail-based migration from the American midwest, its many towns became connected to each other by high-speed highways and eventually merged into one vast metropolis.

The most recent phase is the growth of "edge cities" in the suburban and even the outermost reaches of large metropolitan areas, both old and new (Garreau 1991). An edge city is characterized by large concentrations of office and retail space, often in conjunction with other types of development, including residential, at the nodes of major express highways. Most are in locations where virtually no development, possibly excepting a small town, existed prior to 1960. In many cases, the initial design and construction was the product of a single development company, even a single individual. Edge cities are made possible by ubiquitous automobile access, even when they are located at a transit station, as occasionally happens.[4]

Cities in western Europe have evolved somewhat differently. Being much older, many still have centers which started out as medieval towns. There is a greater mixture of residences and businesses in the core, possibly because of the rich cultural amenities there. Apartment buildings are more common and public transportation more important. Nevertheless, as in North America, there has been massive suburbanization and the emergence of edge cities.

2.2. *Describing Urban Structure*

Using basic land-use data, scholars have sought to describe the regularities and irregularities of urban structure.

[4] The huge Walnut Creek office and retail complex 22 miles east of San Francisco, which developed in the 1970s and 1980s, has at its center a station of the Bay Area Rapid Transit system which opened in the early 1970s. Yet, the automobile accounts for 95 percent of commuting trips to the complex, and presumably an even higher proportion of other trips (Robert Cervero and Kang-Li Wu 1996, Table 5).

We are particularly interested in the degree of spatial concentration of urban population and employment. We distinguish between two types of spatial concentration. At the city-wide level, activity may be relatively *centralized* or *decentralized* depending on how concentrated it is near a central business district. The degree of centralization has been studied mainly by estimating monocentric density functions, and is discussed in Section 3. At a more local level, activities may be *clustered* in a polycentric pattern or *dispersed* in a more regular pattern. It is this clustering that has captured the recent attention of both theoretical and empirical economists.

Defining such clusters precisely, however, is not so easy. If one uses three-dimensional graphics to plot urban density across two-dimensional space, one is struck by how jagged the picture becomes at finer resolutions. An example is presented in Figure 1, which plots 1990 employment density in Los Angeles County (a portion of the Los Angeles urban region) using a single data set plotted at three different degrees of spatial averaging.[5] A similar lesson from the fractal approach discussed below is that within a fixed area, development that appears relatively homogenous at a coarse scale may actually contain a great deal of fine structure. Where fine structure is present, it becomes somewhat arbitrary to say how large a concentration of employment is required to define a location as a subcenter. Even an isolated medical office has a high employment density when viewed at the scale of the building footprint, but we would not call it a subcenter. What about a cluster of twenty medical offices? What if this cluster is adjacent to a hospital and a shopping center? The distinction between an organized system of subcenters and apparently unorganized urban sprawl depends very much on the spatial scale of observation.

We consider three approaches to describing the fine structure of urban development. The first two are ways of mathematically describing distributions of points in space. The third is the basis for extensions of monocentric density functions to a polycentric pattern.

The first approach, called *point pattern analysis*, defines various statistics involving distances between observed units of development (R.W. Thomas 1981). These statistics are then compared with theoretical distributions. One such comparison distribution is that resulting from perturbations of a regular lattice, such as is postulated by one variant of central place theory (Walter Christaller 1966) in which development occurs in a hierarchy of centers, each with a hexagonal market area. Another comparison distribution is that resulting from purely random location, which can be described as a Poisson process. An example of the use of point pattern analysis is the search for population clusters in the Chicago area by Arthur Getis (1983).

A more recent approach to describing urban spatial patterns is based on the idea that they resemble *fractals*, geometric figures which display ever-finer structure when viewed at finer resolutions. Mathematically, a fractal is the limiting result of a process of repeatedly replicating, at smaller and smaller

[5] The data (available on request) are plotted on a 121 X 131 kilometer square locational grid, with a spatial smoothing function used to compute the smoothed average density at each grid point from the raw data for nearby zones. If zone i's centroid is distance D_i from the grid point, its density is weighted proportionally to $[1-(D_i/R)]^2$, where R is the smoothing radius within which zone densities are allowed to affect a given grid point. In the three plots shown in the figure, R takes values equal to $2\sqrt{2}$, $4\sqrt{2}$, and $6\sqrt{2}$ kilometers respectively.

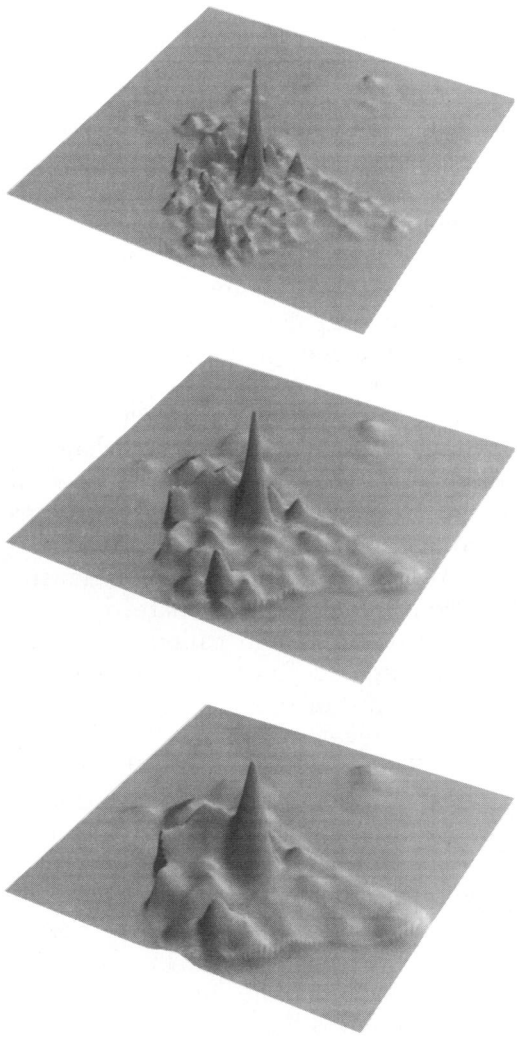

Figure 1. Employment Density, Los Angeles County, 1990, at Different Resolutions.
Source: Authors' plots of data from Southern California Association of Governments.

scales, the same geometric element. Thus the fractal has a similar shape no matter what scale is employed for viewing it. If the original element is one-dimensional, the fractal's length becomes infinite as one measures it at a finer and finer resolution; the classic example is a coastline. One plus the elasticity of measured length with respect to resolution is known as the *fractal dimension*. For example, a coastline might have length L when measured on a map that can just resolve 100-meter features, and $Lx10^{D-1}$ when 10-meter features can be seen; its fractal dimension would then be D, at least within that resolution range. A perfectly straight coastline has fractal dimension one, since its length does not increase with the level of resolution.

Geographers have used fractals to examine the irregularity of the line marking the outer edge of urban development in a particular urban region. Michael Batty and Paul Longley (1994, pp. 174–79) use data on land development in Cardiff, Wales, to define such a boundary to an accuracy as fine as 11 meters. Their best estimates of the fractal dimension of this boundary are between 1.15 and 1.29. (By way of comparison, Britain's coastline has fractal dimension 1.25, Australia's 1.13.) Surprisingly, they find that the fractal dimension of Cardiff's outer edge of development declined slightly over the time period examined (1886 to 1922), a period of significant transport improvements, mainly in the form of streetcars. They conclude that "the traditional image of urban growth becoming more irregular as tentacles of development occur around transport lines is not borne out" (p. 185).

More significantly, one can use fractals to represent two-dimensional development patterns, thereby capturing irregularity in the interior as well as at the boundary of the developed area. For example, a fractal can be generated mathematically by starting with a large filled-in square, then selectively deleting smaller and smaller squares so as to create self-similar patterns at smaller and smaller scales. Such a process simulates the existence of undeveloped land

inside the urban boundary. The fractal dimension D for this situation can be measured by observing how rapidly the fraction of zones containing urban development falls as zonal size is decreased, i.e., as resolution becomes finer. (More precisely, D is twice the elasticity of the number of zones containing any development with respect to the total number of zones into which the fixed urban area is divided.) We call this dimension the *areal fractal dimension*; it can vary from 0, indicating that nearly all the interior space is empty when examined at a fine enough resolution, to 2, indicating that each coarsely-defined zone that contains development is in fact fully developed. Long narrow development would have $D = 1$ (since as we increase the total number N of zones into which a well-defined region is divided, the number of zones containing any development would grow only as $\sqrt{N}$).

Batty and Longley (1994, Table 7.1) report estimated areal fractal dimensions for many cities around the world, with the result most often in the range 1.55 to 1.85. For Paris in 1981 the estimate is 1.66. For Los Angeles in the same year, it is 1.93, tied with Beijing for the highest among the 28 cities reported. This latter estimate implies that the fraction of area developed is almost constant at different scales, indicating a relative absence of fine-structure irregularities in development patterns. Apparently Los Angeles has grown in a more homogeneous manner than Cardiff or Paris.

Time series observations of London from 1820 to 1962, and of Berlin from 1875 to 1945, suggest that the areal fractal dimension has been increasing steadily throughout these time periods. This lends further support to the conclusion that urban growth during the industrial era has made development patterns somewhat more regular, at least in western Europe. Batty and Longley suggest that a possible reason is the more extensive imposition of land-use controls and other forms of urban planning.

Unfortunately, the estimated areal fractal dimension of a city is quite sensitive to just how the land-use data are summarized (Batty and Longley, p. 236). Another problem is that in such a measurement, a city's fine structure is assumed to look like a miniature of the coarse structure, whereas in fact the processes operating at the micro and macro scales are probably very different: fine structure may reflect local zoning rules or developers' detailed design strategies, while coarse structure may reflect regional planning, regional transportation facilities, or land speculation based on anticipated regional growth. Nevertheless, the fractal approach highlights the inadequacy of a deterministic view of development, adopted especially in earlier economic models, in accounting for the irregularities in urban structure. As we discuss in Section 5, more recent advances such as random utility theory enable us to deal with irregularities in a way that is better suited to economic modeling.

Most urban economists have used more intuitive, if simplified, depictions of urban structure, identifying one or more *employment centers* and estimating how these centers affect employment and population densities around them. Much of the early literature on subcenters used criteria based on local knowledge in planning organizations or real estate firms. More recent work has used objective definitions based on employment data for a large number of zones within a metropolitan area (John McDonald 1987). Genevieve Giuliano and Kenneth Small (1991) define a "center"—either a main center (the one containing the CBD) or a subcenter—as

a cluster of contiguous zones, all with gross employment density exceeding some minimum $\overline{D}$, and together containing total employment exceeding some minimum $\overline{E}$. Thus a center contains a peak of employment density, yet substantial intermixing of population is not precluded. This definition facilitates comparisons across cities and among the various centers within a city, including the main center. But as we shall see in Section 4, where we describe some empirical uses of such definitions, the exact pattern of centers so defined may be quite sensitive to the choice of cutoff values $\overline{D}$ and $\overline{E}$. Once again, we find that urban structure is inconveniently irregular and scale-dependent—features that are important clues to the scale-dependent processes governing agglomeration in the modern world.

3. *The Monocentric City Model*

The monocentric city model was the most influential depiction of urban structure for at least two decades, following its formulation by William Alonso (1964) as an adaptation of Johann von Thünen's (1826) theory of agricultural land use. The model was quickly broadened to include production, transport, and housing, and has been generalized in many ways since.[6] It has proved extremely fertile because it provides a rigorous framework for analyzing the spatial aspects of the general-equilibrium adjustments that take place in cities, and for empirically measuring and comparing the degree of centralization across cities and time periods. In this section we present the basic model and illustrate how it can be used to explain historic trends in the suburbanization of households.

[6] The key initial steps were taken by Edwin Mills (1967, 1972) and Richard Muth (1969). For an excellent synthesis see Masahisa Fujita (1989).

3.1 *The Basic Model*

In the model's simplest form, the city is envisaged as a circular residential area surrounding a central business district (CBD) of radius x_c , in which all jobs are located. The theory distinguishes between an *open* city with perfectly elastic population size (due to costless migration) and a *closed* city with fixed population. We deal here with the closed case. Each of N identical households receives utility $u(z,L)$ from a numeraire good z and a residential lot of size L.[7] A household located x miles from the CBD incurs annual transport cost $T(x)$, normally interpreted as commuting cost to the CBD. Each household has exogenous income y which must cover expenditures on the numeraire good, land at unit rent $r(x)$, and transport. Normally $T(x)$ is interpreted as including the value of travel time, so y must include the value of some time endowment.

We define the residential *bid rent* $b(x,\overline{u})$ at location x as the maximum rent per unit land area that a household can pay and still receive utility $\overline{u}$:

$$b(x,\overline{u}) = \max_{z,L} \frac{y - T(x) - z}{L} \quad \text{s.t. } u(z,L) \geq \overline{u}. \tag{1}$$

By the envelope theorem, the slope of the bid-rent function is

$$\frac{db(x,\overline{u})}{dx} = -\frac{T'(x)}{L[y-T(x),\overline{u}]}, \tag{2}$$

where $L[\cdot]$ is the solution to the maximization in (1). Equation (2) is one of the most basic results of the monocentric model, and is entirely intuitive. A household located a small additional distance dx from the CBD incurs additional transport cost $T'(x)dx$. To keep this household

[7] The model is readily extended to explicitly treat housing as a produced commodity, with lot size as one of its inputs. Jan Brueckner (1987) provides a nice analysis of the resulting comparative statics.

indifferent between the two locations, lot rent must be lower at the more distant location by the same amount: that is, $Ldb = -T'(x)dx$.

For each household, there is a family of residential bid-rent functions, indexed by $\bar{u}$. Households are treated as identical and costlessly mobile. Hence, they all obtain the same utility in equilibrium, and the equilibrium rent function $r(x)$ coincides with one of these bid-rent functions. To determine which one, two conditions are needed. First, there is an arbitrage condition at the city boundary (whose value x^* is yet to be determined): residential rent there must equal the rent on land in nonurban use, r_A. (This opportunity cost of land, often called "agricultural rent," is assumed not to vary with location.) Second, all households must be accommodated, which means the integral of household density ($1/L$) over the residential area must equal the number of households:

$$\int_{x_c}^{x^*} \frac{\varphi(x)}{L[y - T(x), \bar{u}]}\, dx = N, \tag{3}$$

where $\varphi(x)dx$ is the residential land area between x and $x + dx$.[8] These two conditions provide two equations in the unknowns x^* and $\bar{u}$; we denote the solution for $\bar{u}$ by u^e.

The land rent at any location is the maximum of the bid rents there:

$$r(x) = \max[b(x,u^e), r_A] = \begin{cases} b(x,u^e) & \text{for } x \le x^* \\ r_A & \text{for } x > x^*. \end{cases} \tag{4}$$

This expresses the principle that, in the land market, each piece of land goes to the highest-bidding use. This principle is the basis for generalizing the model to more than one type of household or to other sectors bidding on land outside the CBD; in such generalizations, the market rent function is the upper envelope of applicable bid-rent functions.

The comparative statics of the model were first fully worked out by William Wheaton (1974). To illustrate their derivation, consider the case of an increase in household population, N. This causes no change in the family of bid-rent functions (1) or in the lot-size function $L[\cdot]$ corresponding to any given net income and utility. But from (3) the higher population would create excess demand for land if the solution were unchanged. Equilibrium is reestablished with higher densities, lower utility, a steeper rent function, and an expanded outer boundary.

Land use in the simple monocentric model is efficient—that is, the equilibrium density pattern is Pareto optimal (Fujita 1989). This is basically because there are no externalities; land-use decisions are based entirely on tradeoffs between desire for space and recognition of commuting costs, both of which are purely private. The need for commuting is exogenous in the model, so no agglomerative effects are present. Of course, these nice properties disappear in more realistic models with congestion, air pollution, neighborhood quality effects, and economies of agglomeration—the last being of prime interest in this essay.

Several comments are in order about the limitations of the monocentric model. The model implicitly assumes that businesses have steeper bid-rent functions than residents, so that all jobs are centrally located. But most of its results can follow from the weaker assumption that employment is dispersed in a circularly symmetric manner, so long as it is less dispersed than residences—that is, within any circle there

[8] If all urban land is used for residential purposes and the city is circular, then $\phi(x) = 2\pi x$.

are more jobs than resident workers. In this case the wage varies over location so as to offset differences in commuting costs (Robert Solow 1973; Michelle White 1988), and commuters still choose to travel radially inward to work.

The model is also easily extended to incorporate different groups of residents. For example, it can predict the pattern of residential location by income. In order to do this, marginal transport cost $T'(x)$ has to be reinterpreted to include the shadow value of time, which turns out to be its dominant component in modern developed nations. (Deriving this shadow value endogenously would require adding leisure and a time budget to the model.) Because this shadow value rises with income, so does marginal transport cost. If $T'(x)$ is less elastic with respect to income than is lot size $L[\cdot]$, equation (2) predicts that rich households will have flatter bid-rent functions than poor households and hence will locate more peripherally. Whether this condition holds for a typical U.S. city is under some dispute (Wheaton 1977).

A more fundamental limitation is that the model is static. Two interpretations are possible, both unrealistic. One is that the model describes a stationary state with durable housing, which a real city would approach asymptotically. The other is that the model describes short-term equilibrium at a point in time, with perishable housing being continuously replaced. The trouble with both interpretations is that the typical lifetimes of buildings greatly exceed the time over which the model's parameters can be expected to remain unchanged.

3.2 *Explanations of Post-war Suburbanization*

What has the monocentric model enabled us to say about the dramatic changes in urban structure over the last century and a half? It obviously throws no light on the trend toward polycentricity. If it applies to anything, it should help explain the broad population decentralization that has occurred in most cities of the world (Mills and Jee Peng Tan 1980). To see how the model performs, we need to quantify the empirically observed trends and provide some plausible parameters for the model.

Pioneered by Colin Clark (1951), researchers have estimated urban population density functions for an enormous range of places and times.[9] In most of this work, a negative exponential function is assumed: $D(x) = D_0 e^{-\gamma x}$ where $D(x)$ is population density at distance x from the CBD and D_0 and γ are positive constants. The negative exponential function is convenient because it is easy to estimate after taking logarithms. The constant $\gamma \equiv -D'/D$ is the proportional rate at which population density falls with distance and is known as the *density gradient*. It is a useful index of population centralization.

Two of the strongest empirical regularities relating to urban spatial structure can be concisely stated using the gradient as defined earlier. First, density declines with distance from the center—the density gradient is positive. Second, virtually all cities in the developed world and most others elsewhere have decentralized over the last century or more—the density gradient has declined over time. Table 1 provides just a tiny sampling of empirical support for these assertions; corroborating evidence is provided for Japan by Mills and Katsutoshi Ohta (1976), for Latin America by Gregory Ingram and Alan Carroll

[9] McDonald (1989) and Mills and Tan (1980) provide good surveys of methodology and results, respectively.

TABLE 1
SOME ESTIMATES OF POPULATION DENSITY GRADIENTS

City	Year	Density Gradient (per mile)
London	1801	1.26
	1841	0.93
	1901	0.37
	1931	0.27
	1939	0.23
	1961	0.14
Paris	1817	2.35
	1856	0.95
	1896	0.80
	1931	0.76
	1946	0.34
Frankfurt	1890	1.87
	1933	0.92
Birmingham, UK	1921	0.80
	1938	0.47
Rangoon	1931	1.16
	1951	0.55
New York	1900	0.32
	1940	0.21
	1950	0.18
Chicago	1880	0.77
	1900	0.40
	1940	0.21
	1956	0.18
Los Angeles	1940	0.27
Boston	1900	0.85
	1940	0.31
Sydney	1911	0.48
	1954	0.26
Christchurch	1911	1.61
	1951	1.34

Source: Clark (1967, pp. 349–51), converted from km to miles.

(1981), and for a number of developing nations by Mills and Tan (1980). Any persuasive theory of urban spatial structure should accord with these facts.

Urban economists' standard explanation for decentralization is a combination of rising incomes and declining transport costs, both of which cause the density gradient to decline according to the monocentric model. The second part of this explanation is not entirely satisfactory, however, because the largest portion of transport cost is user time, whose value tends to rise with wages, creating a strong force counteracting improvements in travel speeds. It is therefore worth taking a closer look at the magnitudes of the parameters governing the density gradient.

In order to most conveniently match theory with empirical measurement, we first consider a specific set of assumptions that lead to the negative exponential population density function.[10] Suppose the utility function is Cobb-Douglas, $u(z,L) = z^{\alpha}L^{1-\alpha}$. Suppose also that the ratio of marginal transport cost to income net of transport cost, $T'/(y-T)$, is constant across locations—reflecting the fact that congestion is least in peripheral locations from which total commuting cost is greatest. Then the population density function is negative exponential with gradient

$$\gamma = \frac{\alpha T'/y}{(1-\alpha)[1-(T/y)]}. \tag{5}$$

Land rent is also negative exponential, with gradient γ/α, while net income y-T and marginal transport cost T' are each negative exponential with gradient $(1-\alpha)\gamma/\alpha$.

Using empirically plausible point estimates for the right-hand side of (5), from parameters appropriate for U.S. cities around 1970, we can calculate a

[10] See Yorgos Papageorgiou and David Pines (1989) for a more complete discussion. The original derivation of the negative exponential relied on unitary price elasticity of demand for housing and Cobb-Douglas production of housing (Muth 1969, ch. 4). We instead provide conditions on the utility function and on transport costs, which to the best of our knowledge is novel. Alex Anas and Ikki Kim (1992) generate negative exponential densities by incorporating an income distribution.

gradient of $\gamma = 0.234$ per mile.[11] By way of comparison, Edmonston (1975, Table 5.5) and Mills and Ohta (1976) report average values of 0.38 and 0.12, respectively, for various samples of U.S. cities in 1970. So our "guesstimate" of (5) is near the average of their estimates.

How does (5) do in explaining decentralization in U.S. cities? Comparisons across decades are tenuous, but we can very roughly ask whether changes in incomes and transport costs could account for the changes in γ observed between 1950 and 1970. According to our model, from 1950 to 1970 the gradient should have fallen from 0.318 to 0.234 or by 26 percent.[12] By comparison, Edmonston reports a 41 percent decline in the density gradient for a sample of cities over that period. Again, the simple model appears to be in the right ball park.[13]

However, there are some unsatisfactory aspects to the attempt to explain density gradients in this way. Peter Mieszkowski and Mills (1993) give a cogent account. First, attempts to explain differences in gradients across cities and across times have not been very successful at isolating transport costs as an explanatory factor; this may be because such costs are inaccurately measured and are strongly correlated with income. Second, many of the density gradient estimates are based on just two observations, population in the central city and in the suburbs, along with the area covered by the central city; but this method appears to be highly inaccurate in certain cases, particularly in smaller cities. Third, because of lack of land-use data at a fine scale, most of the empirical work uses gross density (population divided by total land area) although the theory would be better represented by net density (population divided by residential land area); unfortunately, using gross density may drastically overstate the size of density gradients because the outer reaches of an urban area contain much higher proportions of undeveloped land (Mieszkowski and Barton Smith 1991). Finally, a strong negative correlation is observed between the density gradient and total population, with larger cities more decentralized; whereas the monocentric model predicts either no correlation or, in our version, a mild positive correla-

[11] Housing costs were probably around 20% of after-tax income net of commuting cost, and land costs about 20% of housing costs (Small 1981, p.320), giving $1-\alpha = 0.04$. We assume that each commuter had nine hours daily for commuting plus work, and that income was taxed at a constant rate τ. We assume that the average one-way commute was 10 miles and took place at a speed of 25 miles per hour, requiring 48 minutes of round trip per day. Suppose that the only cost of travel is time, valued at the after-tax wage rate. So total daily commuting cost averaged over x is $(48/60)w$, while marginal daily commuting cost (per mile of one way trip) is one-tenth as large. Figuring in taxes: $y = (1-\tau)9w$, $T = (1-\tau)(48/60)w$, and $T' = (0.10)T$. Hence $T/y = 0.0889$ and $T'/y = 0.00889$. (This implies that commuting time is, on average, about 9% of the consumer's time endowment, which is quite plausible.) Hence, $\gamma = (0.96/0.04)(0.00889)/(1-0.0889) = 0.234$. To be better aligned with the empirical evidence (see Small 1992, pp. 44, 84), we would have to recognize that travel time is valued at $w/2$, or somewhat less than the after-tax wage rate; and also that there is a variable money cost of automobile commuting equal to about half the time cost. These corrections happen to approximately cancel, so do not change the gradient estimate by much.

[12] We assume that $1-\alpha$ remained at 0.04 throughout the period. LeRoy and Sonstelie (1983, Table 4) estimate that real income rose approximately 88% over those two decades while real marginal transport costs (including the value of time) rose only 43%. (They give nominal figures which we deflated by the CPI. We have estimated the mean by interpolating between their figures for the 25th and 75th percentiles.) Then, the 1950 value of γ predicted by equation (5) is found by replacing the 1970 value of (T'/y) by $[(T'/1.43)/(y/1.88)] = 1.315\,(T'/y)$, and similarly for T/y. The result is $\gamma = (0.96/0.04)\,(1.315)(0.00889)/[1 - (1.315 \text{ times } 0.0889)] = 0.318$.

[13] More refined predictions could be made using available extensions of the simple monocentric

tion.[14] Mills and Tan (1980) suggest that the observed negative correlation, "though not a consequence of the model, is strongly suggested by common sense" because larger cities support outlying employment subcenters (p. 315). This of course is an appeal to forces outside the monocentric model.

Probably the most serious deficiency of the monocentric model as an explanation of urban decentralization is its failure to account for the durability of housing. David Harrison and John Kain (1974) observed that cities tend to grow outward by adding rings of housing at a density which reflects contemporaneous economic conditions, with the density of earlier rings remaining unchanged due to housing durability. The same phenomenon is demonstrated by Mieszkowski and Smith (1991), who show that the density of developed residential land (i.e. net density) in Houston is approximately constant all the way to the outer edge of the metropolitan area. A variety of dynamic versions of the monocentric model with durable housing has been constructed. In such models, the density gradient depends not only on the past time path of income and transport costs, but also on developers' expectations and the prospects for redevelopment. Explanations for observed density gradients are correspondingly complex.

Employment density functions can be estimated in the same way as population density functions, although data on the location of jobs are less readily available and less reliable than those for population. The general conclusion from the empirical literature is that the density gradient is larger for jobs than for households, but has been falling faster (Mieszkowski and Mills 1993). This evidence weakly supports the hypothesis that jobs have been following people; but there are many other reasons for jobs to have decentralized, as described in Section 2.

Other possible explanations of population decentralization, especially in the U.S., include variants of a "flight from blight" hypothesis. First is deteriorating central housing quality, due to style or technological obsolescence combined with rational decisions by owners to run down housing quality. Second is the existence of racial preferences combined with the tendency of poorer African-Americans to live in central cities. Third is negative neighborhood externalities associated with many poor neighborhoods. Fourth is the working out of Tiebout mechanisms for providing local public goods (Charles Tiebout 1956), resulting in better-off residents with high demands for local public goods abandoning the central city and excluding the poor from the suburbs through minimum lot-size zoning. All these explanations imply that the poor live near downtown and the rich are pushed or pulled out to the suburbs. The implied effect on the value of density gradients is, however, ambiguous.

4. *The Polycentric City: Empirical Descriptions*

We now turn to one of the most interesting features of modern urban landscapes—the tendency of economic activity to cluster in several interacting centers of activity. This section describes empirical findings. The next reviews theoretical models of polycentric-

model. For example, accounting for income differences would increase the predicted density gradient if parameters are such that higher income people live more peripherally, since they also choose more land per dwelling for a given rent.

[14] Looking at the outer boundary, rising population does not change marginal transport cost but it does increase total transport cost, hence lowering the denominator in (5) and causing γ to rise.

ity. Throughout, we use "center" to mean either the main center or a subcenter.

It is not hard to discover subcenters lurking in spatial employment or population data for most large cities. Giuliano and Small (1991) provide a review of studies, and new ones are steadily appearing. Here we consider some tentative generalizations about the nature and role of subcenters in U.S. cities, for which polycentricity has been examined in greater detail than anywhere else.

(i) *Subcenters are prominent in both new and old cities*. Evidence is emerging that in each of the largest metropolitan areas in the United States, twenty or so subcenters can be identified using the criteria described in Section 2 with minimum gross density ($\overline{D}$) of 10 employees per acre and minimum total employment ($\overline{E}$) of 10,000. Giuliano and Small (1991) find 29 such centers in Los Angeles in 1980, and add three smaller outlying centers with prominent density peaks. Daniel McMillen and McDonald (1998b) find 15 subcenters outside the city limits of Chicago meeting an identical criterion, using a combination of 1980 and 1990 data. Cervero and Wu (1997) find 22 such centers in the San Francisco Bay Area for 1990.

Each of these studies covers a Consolidated Metropolitan Statistical Area (CMSA), a census concept that is the most inclusive of the various types of metropolitan areas defined in official U.S. statistics. For example, San Francisco's CMSA includes nine counties, from the Napa Valley wine country in the north to San Jose and Silicon Valley in the south.[15]

(ii) *The number of subcenters and their boundaries are quite sensitive to definition*. Both the Los Angeles and the Chicago studies mentioned above find that with changes in density cutoffs, certain employment clusters could be viewed either as several large subcenters or as one mega-center. In the Chicago data, for example, the criteria just listed produce a huge subcenter near O'Hare Airport, with 420,000 employees,[16] whereas doubling the density cutoff breaks this subcenter into five smaller ones. The Los Angeles case, discussed in the next subsection, shows even more sensitivity to subcenter definition.

Such sensitivity is not surprising considering the observations made in Section 2. The urban landscape is highly irregular when viewed at a fine scale, and how one averages these local irregularities determines the look of the resulting pattern. It may be that the patterns that occur at different distance scales are influenced by different types of agglomeration economies, each based on interaction mechanisms with particular requirements for spatial proximity. This observation applies also to clustering at a regional scale such as the U.S. eastern

[15] Smaller urban regions, and a few large ones like that surrounding Washington, D.C., are not classified as CMSAs but rather as Metropolitan Statistical Areas (MSAs). Both CMSAs and MSAs are collections of whole counties (except in New England) that are highly integrated; the MSA is closest to what before 1983 was defined as a Standard Metropolitan Statistical Area (SMSA). The CMSA typically combines several adjacent areas formerly classified as SMSAs, most of which are now called Primary Metropolitan Statistical Areas (PMSAs). For example, the New York–Northern New Jersey–Long Island CMSA consists of 11 PMSAs including New York (New York City plus three adjacent counties), Nassau-Suffolk (two counties constituting Long Island), and Newark (five counties in New Jersey). The Los Angeles–Anaheim–Riverside CMSA consists of four PMSAs: Los Angeles County, Riverside and San Bernardino Counties, Orange County, and Ventura County. See U.S. Bureau of the Census (1996, pp. 937–945). Because we are not interested in municipal boundaries, in this essay we generally designate a CMSA just by the name of its largest city.

[16] O'Hare airport is annexed to the City of Chicago, despite its being surrounded entirely by suburbs. For this reason employment at the airport itself is missing in these data, which cover only the suburbs.

seaboard and the core industrialized complex of northwestern Europe.

(iii) *Subcenters are sometimes arrayed in corridors.* In the 1980 Los Angeles data, the four largest centers and one smaller one form an arc extending through the downtown area, Hollywood, and Century City all the way to the Pacific Ocean. The five centers are tenuously separated by zones just failing the density cutoff; a slight lowering of the cutoff causes the centers to become joined into one 19-mile-long center containing over 17 percent of the entire region's employment.

There is even an example where a corridor, rather than a set of point centers, seems to best explain surrounding density patterns. This is the Houston Ship Channel, a 20-mile-long canal lined by manufacturing plants and connecting central Houston (starting just two miles from the CBD) to Galveston Bay (Steven Craig, Janet Kohlhase and Steven Pitts 1996).

Both these examples of corridor development follow older established transportation facilities. Indeed, the corridor shape is quite familiar from urban history: as we have already seen, "streetcar suburbs" were prominent a century ago and less. Some of these communities and their associated transportation facilities later became the focus for development and redevelopment that were more automobile-oriented and more job-intensive. Similarly, at a regional scale, large metropolitan areas have sometimes grown together into a corridor-like "megalopolis" following an older inter-regional travel corridor, such as that between Boston and Washington.

(iv) *Employment centers help explain surrounding employment and population.* Several studies have established that point or corridor subcenters, as described above, help explain surrounding patterns of employment density, population density, and land values.

Three functional forms have been suggested as appropriate to generalize monocentric formulations to a polycentric structure (Eric Heikkila et al. 1989). All generalize the negative exponential function $D(x) = Ae^{-\gamma x}$ of Section 3.2, but each uses a different assumption about how the occupant of a given land parcel values access to multiple centers. They are:

$$D_m = \underset{N}{Max_n}\left\{A_n \exp\left(-\gamma_n x_{mn}\right)\right\} \tag{6}$$

$$D_m = A\prod_{n=1}^{N} \exp\left(-\gamma_n x_{mn}\right) \tag{7}$$

$$D_m = \sum_{n=1}^{N} A_n \exp\left(-\gamma_n x_{mn}\right) \tag{8}$$

where D_m is density at location m, x_{mn} is distance of location m to center n, and A, A_n, and γ_n are coefficients to be estimated.

The first, (6), assumes that centers are viewed as perfect substitutes; each center therefore generates its own declining bid-rent function for surrounding land, and land-use density at any point is determined by the highest of these bid-rent functions. In other words, what matters at any location is only the center with the largest influence at that point, and space is divided into strictly separate zones of influence as in the model of White (1976). We are not aware of any empirical support for this form, however, and it is rarely used in applied work.

The assumption in (7) is that centers are complements. The occupant of a given location requires access to every center in the area. This specification is easy to estimate after taking logarithms. It seems rather robust in practice, although it has a rather extreme property, that great distance from even one subcenter can entirely prevent development at location m. A modification of

(7) that replaces $-\gamma_n x_{mn}$ by γ_n / x_{mn} overcomes this difficulty, and seems to fit even better.[17]

An intermediate case is the additive form (8), used by Gordon, Richardson, and H.L. Wong (1986) and by Small and Shunfeng Song (1994). It is based on the idea that the accessibility of a location is determined by the sum of exponentially declining influences from various centers. Here every center has an influence as in (7), but unlike in (7) a center's influence becomes negligible at large distances. Unfortunately, estimation of (8) requires nonlinear estimation and often produces convergence problems.

Considerable success has been attained using these models to explain density and land-value patterns in Los Angeles, Chicago, San Francisco, and a few other places. The pioneering studies were Daniel Griffith (1981) and Gordon, Richardson and Wong (1986). Small and Song (1994) are able to explain roughly 50 to 75 percent of the variance in employment or population density across the entire Los Angeles region using equation (8) with five centers for 1970 and eight centers for 1980. In all cases the special case of monocentricity is soundly rejected. The population density patterns fit well even though population data were not used to determine the locations of the centers used in the specification. Small and Song also show that monocentric density estimates fit poorly, especially in the later year, reinforcing the belief that polycentricity is an increasingly prominent feature of the landscape.

[17] McDonald and Paul Prather (1994), McMillen and McDonald (1998a, b). A different modification replaces the distances x_{mn} to specific centers n in (7) with distance to the nearest center, the second nearest center, and so forth. Rena Sivitanidou (1996) uses this form successfully to explain Los Angeles office and commercial land values.

(v) *Subcenters have not eliminated the importance of the main center.* Whenever a downtown center and one or more subcenters have been defined using the same criteria, downtown has more total employment, higher employment density, and usually a larger statistical effect on surrounding densities and land prices than does any subcenter. Because so many people believe that big-city downtowns are passé, it is worth reviewing this evidence in some detail.

Let us begin with Chicago. In explaining 1980 employment density patterns outside the city limits of Chicago, three large subcenters are found by McDonald and Prather (1994) to have exerted an important influence, but none has a t-statistic even one-fourth as large as does the CBD. In a remarkable study of land values over a century and a half, McMillen (1996) finds a clear and marked land-value peak at the CBD for each of 10 different years from 1836 to 1990, despite the steady rise in importance of centers several miles to the northwest.

In their study of San Francisco, Cervero and Wu (1997) list the sizes of the 22 centers emerging from the Giuliano-Small criterion described earlier. The largest and densest by far is the one containing downtown San Francisco. This center accounts for 15 percent of the region's employment. Silicon Valley is the second largest center, and the third (despite Gertrude Stein) is centered in downtown Oakland.

Now consider Los Angeles, famous for its sprawl. Garreau (1991) names more actual plus emerging "edge cities" there than in any other metropolitan area in the United States.[18] Yet of the

[18] Garreau's definition of an edge city includes five criteria: 5,000,000 square feet of office space; 600,000 square feet of retail space; a daily inflow of commuters; a "local perception as a single end destination for mixed use"; and a location that was

centers identified by Giuliano and Small (1991), the one containing downtown Los Angeles dominates by nearly any measure. It contained 469,000 employees, more than double the next largest center and nearly ten times the size of the largest "edge city" in the region, known as South Coast Metro. The downtown center, much larger than the traditionally defined CBD, contained one-tenth of the region's employment and nearly one-third of the employment in all centers combined. Small and Song (1994) try alternative center locations in monocentric models of employment and population density, finding that the downtown center gives the best fit (although Los Angeles Airport comes close in the case of population).

(vi) *Most jobs are outside centers.* Remarkably, centers account for less than half of all employment in the areas studied: 47 percent in San Francisco, one-third in Los Angeles, and less than one-fourth in suburban Chicago.[19] The polycentric pattern, interesting and important though it may be, coexists with a great deal of local employment dispersion. Furthermore, the population distribution can be explained much better by a model that accounts for distance to all employment rather than just to employment in centers, even if that model is constrained to have fewer parameters in total (Song 1994).

Nevertheless, we think Gordon and Richardson (1996) are premature in suggesting that dispersion has made the polycentric city a phenomenon of the past. Their results show that newer growth is more dispersed than earlier growth, but this has always been true. The crucial but unanswered questions are whether older centers remain vital and, when not, whether they are replaced by newer ones.

Another thing we do not know is whether subcenters fill essential niches in the local economy out of proportion to the sheer numbers of people working or shopping there. Certainly there is suggestive evidence that they do. Edge cities, for example, are well known as important sites of office location, indicating that they serve as nodes of information exchange. More generally, Giuliano and Small (1991) and McMillen and McDonald (1998b) find that different centers have quite different industry-mix characteristics, with some centers very specialized and others resembling the CBD in their diversity. Indeed, in Los Angeles, even the size distribution of centers closely follows the "rank-size rule" characterizing the distribution of city sizes within a nation.[20] Further empirical research on the economic roles that subcenters play would appear to us to have a high payoff.

(vii) *Commuting is not well explained by standard urban models, either monocentric or polycentric.* Bruce Hamilton (1982) was the first to note that the stan-

residential or rural thirty years previously (Garreau 1991, p. 425). He allows for some element of judgment in deciding on boundaries and on when two nearby edge cities should be counted as one. An "emerging" edge city is an area showing signs that it will soon become an edge city.

[19] This last statement is for 1990 employment using the more restricted definitions for the subcenters near O'Hare and Evanston, as preferred by McMillen and McDonald (1998b). Total 1990 employment in suburban subcenters was 558,600, from their Table 1. Total 1990 suburban employment was 2,381,900, from Daniel McMillen, private correspondence. Unfortunately certain data sources are incompatible between the City of Chicago and the rest of the CMSA; as a result many studies have used one or the other, making us unable to make statements for the entire CMSA.

[20] This rule, also known as Zipf's law, postulates that the cumulative fraction of cities of size N or greater is proportional to $1/N$. See Kenneth Rosen and Mitchel Resnick (1980) for a thorough empirical investigation. See Krugman (1996) for a thoughtful discussion of possible reasons for this amazingly robust empirical relationship.

dard assumption of people commuting up a land-rent gradient cannot come close to explaining actual commuting patterns in the United States or Japan. Starting from the distributions of jobs and employee residences as functions of distance to the CBD, Hamilton calculates the average commuting distance when everyone commutes inward along a ray, as is implied by the monocentric model with dispersed employment. This procedure predicts average commutes of about one mile, understating actual average commutes by a factor of seven! Nor is the problem just monocentricity; letting density patterns be polycentric does not eliminate the discrepancy (Giuliano and Small 1993). In fact, even allowing for all the spatial irregularities of job and housing locations, average commutes are still three times as long, both in time and distance, as they would be if jobs and employees were matched so as to minimize average commuting distance as is implied by deterministic residential location models with identical individuals (Small and Song 1992).

It appears that at least in auto-dominated cities, there is more "cross-commuting," in which commuters pass each other in opposite directions, than there is commuting "up the rent gradient." Cross-commuting does not occur under standard assumptions, because if it did, people could reduce commuting costs without incurring higher rents, simply by interchanging houses. Naturally we don't expect the real world to fit the monocentric model perfectly, but being off by a factor of seven or even three is hard to swallow, considering the central role that commuting plays in the standard models.

There are several possible explanations for why people do not eliminate these extra commuting costs by moving. People have idiosyncratic preferences for particular locations, due to the different mixes of local amenities and to practical or sentimental attachments; two-worker households have to compromise between locations convenient to a job; frequent job changes and substantial moving costs cause people to choose locations convenient to an expected array of possible future jobs rather than just their current job; and racial and income segregation affect housing choices. All these explanations require job specialization, for otherwise people could get around the constraints by choosing a suitable job location. No one of these explanations is likely to explain the entire discrepancy, but perhaps all can together.

At a more fundamental level, these observations suggest that heterogeneity of preferences and of job opportunities is extremely important in explaining urban residential location decisions. For example, adding idiosyncratic taste heterogeneity to a standard monocentric model results in greater decentralization (Anas 1990).

The upshot of the empirical work on subcenters is that some patterns stand out despite a great deal of irregularity and dispersion. Downtowns are still important; major employment centers still exist and exert influence over surrounding population and employment distributions; but density and commuting patterns contain much randomness. We now turn to theoretical explanations of these facts. Because the theories could apply at regional as well as urban scales, the same analytical framework should also aid in the understanding of the regional clustering, both within and across national boundaries, that so vitally affects national cohesion and international trade.

5. *Theories of Agglomeration and Polycentricity*

Why do employment concentrations within cities exhibit the complex pat-

terns discussed in the previous sections? To fix ideas, imagine first a "backyard economy" with no patterns—just a uniform distribution of economic activity over space. This would be the equilibrium under certain restrictive assumptions: land is homogeneous, production of each good exhibits constant returns to scale, goods and people are costly to transport, and there is no interaction over space. To understand agglomeration, we can ask, following Papageorgiou and T. Smith (1983): What are some alternative assumptions that would make this uniform distribution of activity unstable? The classical answers are spatial inhomogeneities and internal scale economies in production. More recent answers involve scale economies external to firms, including those arising from spatial contacts and imperfect competition. When any of these alternative assumptions holds in an environment where transport and communication costs are not too high, spatial agglomeration can occur.

In this section we explore each of these alternative assumptions in turn. We then consider dynamics, and finally examine some approaches to agglomeration from outside economics.

5.1 *Spatial Inhomogeneities*

Locations differ in factors such as soil, climate, mineral deposits, and access to waterways. Given such sources of Ricardian comparative advantage, trade arises and production specializes by location, unless transport costs are prohibitively high—in which case the backyard economy persists but with backyards that differ from one another.

Thus even with constant returns to scale in production, spatial inhomogeneities can give rise to towns (Marcus Berliant and Hideo Konishi 1996). An example is a mineral deposit which attracts workers to a mine. Miners have to be clothed and fed; depending on the structure of transport costs, some stages of the production or processing of clothing and food are performed locally. If the cost of shipping unprocessed ore is high, ore processing also occurs locally. A similar example is a town forming at a river rapids, since transshipment activity creates a demand for other goods causing local production—early Montreal is one such case.

Spatial inhomogeneities can create subcenters as well as central business districts. For example, a CBD may form on a harbor and a secondary employment center may form at the site of a river landing. The early model of White (1976) stressed such causes of subcenter formation.

5.2. *Internal Scale Economies*

The second classical explanation for agglomeration is economies of scale in some production process. An important example is scale economies in the loading and unloading of goods. Even in the absence of a natural advantage such as a protected harbor, port activities would tend to concentrate for this reason, a tendency which helped produce the port or railhead orientation of the nineteenth century city (Moses and Williamson 1967; Mills 1972). The advent of containerization has, if anything, intensified the economies of scale in port operations; trucking, by contrast, appears to require only small-scale loading and unloading equipment, so its terminal operations are widely dispersed along major highways.

Another source of scale economies is the production of local public goods (Joseph Stiglitz 1977), as suggested by many of the classic explanations for the historical origin of cities—the city as temple, citadel, capitol, marketplace, granary, or theater. Their counterparts in modern cities include civic buildings,

water works, and monuments. Because such infrastructure is durable and lumpy, numerous man-made inhomogeneities emerge as an urban area grows and some become the sites around which new agglomerations form.

There are also scale economies in private production. A larger plant may have lower average production costs, but also higher average transport costs since inputs have to be gathered from, and outputs distributed to, a larger area. The efficient scale and hence the efficient market area are larger the greater is the degree of increasing returns and the lower are unit transport costs (Starrett 1974). The diseconomy from transport tends to balance the scale economies present in production, resulting in an equilibrium without the requirement that the production process itself have a U-shaped average cost curve—rather, the average production plus distribution cost is U-shaped.

5.3 *External Scale Economies*

We have seen that a public or private good produced under increasing returns can lead to agglomeration. Now suppose there are two private goods, each produced by a different firm, and that one of them, which is costly to ship, is used in the production of the other. This interindustry linkage may cause aggregate costs to be lower if the two firms co-locate. This is just one example of economies of scale that are external to individual firms, resulting in this case from transport costs. Other examples include contact externalities among consumers and market linkages between firms and consumers.

External scale economies between firms are called *economies of localization* if between firms in the same industry, and *economies of urbanization* if across industries. Economies of localization cause cities to be specialized; economies of urbanization cause them to be diversified. Empirical work has found strong evidence of localization economies and somewhat weaker evidence of urbanization economies.[21] Typically this work measures a production or cost function for firms in a given industry with a shift factor depending on local aggregate activity, either in the same industry (localization economies) or in all industries (urbanization economies).

External economies may also be dynamic, affecting not only the level of unit costs but also the rate at which they fall over time. An obvious example is technical progress spurred by knowledge transfer, along the lines suggested by Paul Romer (1986). The prevalence of dynamic external economies is emphasized by Jacobs (1969) in describing the growth of cities, both early and modern, and by AnnaLee Saxenian (1994) in explaining the recent contest between Boston and Silicon Valley for dominance in computer electronics. There is some evidence that urbanization economies contribute to economic growth through endogenous technical change (Ó hUallacháin 1989; Glaeser et al., 1992).

One type of external economy that can be either localization or urbanization is *economies of massed reserves* (E.A.G. Robinson 1931; Hoover 1948), also called statistical economies of scale. In a world with firm-specific shocks, a firm with a specialized job

[21] Randall Eberts and McMillen (forthcoming) provide a good review. Glenn Ellison and Glaeser (1997) derive a general index of the geographical concentration of an industry that distinguishes between that due to the random distribution of finite-sized plants and that due to agglomerative forces other than internal scale economies. They find that for the U.S., roughly half the observed employment concentration is due to such randomness and internal scale economies; as to the other half, most industries show a mild degree of agglomeration while a few show a marked degree.

vacancy is more likely to find a match with an unemployed worker when the labor market is larger; likewise, specialized capital that is unemployed due to a firm's closing is more likely to be successfully redeployed the larger the number of other firms using similar types of capital (Robert Helsley and William Strange 1991). Another type is *information exchange* within or between industries, for example, learning about the efficacy of new techniques by observing the successes and failures of competitors. Yet another type derives from *education*: because labor specialization encourages investment in human capital, larger cities have more educated work forces which may in turn result in more experimentation, more innovation, greater adaptability, and improved management skills.

How do inter-firm externalities affect spatial structure? We can learn a lot just by specifying how their strength varies with spatial proximity, even without describing their source. Using such a "pure externality" approach, Fujita and Hideaki Ogawa (1982) consider a closed market economy on a line segment with a fixed number of workers, each of whom consumes a single produced good and a residential lot of fixed size. They assume an equal number of firms, each employing one worker and occupying an industrial lot of fixed size. Workers commute to their jobs at a constant cost per unit distance. Firms benefit from proximity to other firms, as described by a *location potential function* in which the external productivity benefit conferred by one firm on another falls off with the distance between them according to a negative exponential function with a fixed decay rate. All agents are price takers. If commuting costs are very high, equilibrium entails a completely mixed land use pattern with all workers living adjacent to their job sites—the backyard economy again; if commuting costs are very low and the decay rate is small, agglomeration benefits dominate and firms cluster around one location giving rise to a monocentric city; and if commuting costs are moderate and the decay rate is high (so that a firm benefits a lot from nearby firms but not much from more distant firms) then equilibrium is polycentric. This model produces multiple equilibria—for example with one, three, or five centers—under the same set of parameter values, suggesting that a city's structure at a point in time may be path-dependent even when the durability of structures is ignored. Also, the comparative statics of this model are catastrophic—i.e., the solution changes discontinuously as parameter values are varied.

What might lie behind a location potential function? One possibility is simply *spatial contact*. Consider, for example, a very basic *fixed interaction model* in the spirit of Robert Solow and William Vickrey (1971) or E. Borukhov and Oded Hochman (1977). The city's geography is described by a finite space such as a line segment or a disc, with a geometric center but no predetermined economic center. The city is populated by homogeneous agents (either firms or households but not both), each of whom occupies a lot of unit size and interacts by traveling the same fixed number of times to visit every other agent. These abstract interactions can be interpreted as social contact, information acquisition, search, or exchange.[22]

[22] In actual cities, many such interactions are face to face. Because formal contracting is costly, much contracting takes place informally; this requires honest dealing, and honesty is communicated by body language and eye contact. The fact that humans have developed unconscious signals of their intentions, as well as the ability to decipher those signals, can be explained by theories of evolutionarily stable strategies as postulated by John Maynard-Smith (1976). See also Robert

Equilibrium is characterized by equal profits or, in the case of individuals, equal utilities. In equilibrium, the geometric center is the most accessible point; so rents peak there, declining monotonically towards the edge of the space. If the model is extended so that lot size is responsive to rent, population or employment density shows the same monotonic pattern. Unlike in the monocentric model of Section 3, however, this equilibrium is not efficient because interdependence among agents creates an externality. If an agent moves to a more accessible location, she imparts an external benefit on all other agents by reducing the average cost of their contacting her, which is in addition to the reduction in cost she obtains in contacting them.[23] Since she does not value the benefit conferred on others, she will choose a less central location than is optimal. Hence, the city is too dispersed.

Presumably, the agents interact because they receive a benefit from doing so—for example, each pair of agents may exchange valuable but unpriced information. Then there is a second externality at the margin of the city's population, because adding a new agent confers a benefit on other agents that the new agent fails to capture. The city is therefore too small as well as too dispersed.

Contacts in the above models are non-market interactions between consumers or between firms. In Anas and Kim's (1996) general equilibrium model, contacts are instead market interactions and they occur between consumers and firms—specifically, purchases on shopping trips. Goods are differentiated by location. Each retail firm produces at a particular location under competitive conditions using land and labor, and sells its product on site. Having a taste for variety, a consumer shops everywhere products are sold, with the number of shopping trips to a particular location depending on its accessibility to that consumer's residence. Hence, this is a *flexible interaction model*, in which the attenuation of shopping trips with distance plays a role akin to that of the location potential function. Firms and consumers use varying amounts of land, and transportation is characterized by congestion. The model determines equilibrium rents, wages, and retail prices, all as functions of location with respect to the geometric center.

In the absence of external scale economies, firms and households in the Anas-Kim model are intermixed and dispersed around this geometric center, with commercial and residential densities declining with distance from it. But now suppose there is an external scale economy that operates within a particular shopping district. When the scale economy is large and congestion not too severe, there is a unique, stable equilibrium with firms in a single central district surrounded by consumers. As the cost of congestion increases, the monocenter becomes unstable and two or more smaller shopping districts emerge. Again we observe multiple equilibria, path dependence, and catastrophic transitions.

5.4 *Imperfect Competition*

When firms compete imperfectly they impose a variety of pecuniary externali-

---

Frank (1988). Another reason for face-to-face interaction is that much creative activity is facilitated by conversation in a social setting (Jacobs 1969; Saxenian 1994).

[23] That is, the benefit from lowering the cost of a given contact is mutual, so both agents cannot capture it fully through transaction prices. This easily misunderstood point is made by Tjalling Koopmans and Martin Beckmann (1957). It is true that any transactions that are socially desirable could be elicited by sufficient side payments—but this amounts to internalizing the externality. Short of that, any pricing rule that allocates the cost of the interaction in a specified way leaves one or both parties short of the full incentive to interact.

ties on one another. In aspatial contexts this can create critical-mass effects as in some "big push" models of industrialization (Kevin Murphy, Andrei Schleifer and Robert Vishny 1989). In spatial contexts, imperfect competition can cause agglomeration in an analogous way. Indeed, from Harold Hotelling (1929) on, one of the central issues addressed by spatial competition theory is the circumstances under which firms have incentives to co-locate. Jean Gabszewicz and Jacques-François Thisse (1986) provide a review.

If economies of scale internal to the firm are large, the number of firms in the industry will be small. Given the resulting market power, determining equilibrium location patterns entails game-theoretic considerations. In such *spatial oligopoly models*, firms may compete in price, product quality, product mix, and location, conferring market advantages and disadvantages on each other. Such firms are typically conceived to be retailers or, more recently, developers (J. Vernon Henderson and Eric Slade 1993). Typically, product variety is assumed to be valued because of convex preferences, idiosyncratic preferences, or specialized intermediate goods. Such models easily produce externalities: suppose, for example, that expansion of the market occurs, causing one more firm to enter and the accessibility or variety of products to be thereby enlarged; this creates additional consumer surplus that is not fully captured by the entrant.

Agglomeration may arise in situations of spatial oligopoly, depending on the balance of advantages and disadvantages of clustering. In the model of Norbert Schulz and Konrad Stahl (1996), shoppers trade off the higher transport costs from traveling to a larger activity center (which on average is farther away from consumers) against the benefits from the increased product variety to be found there (which in their model lowers search costs). Retailers, in turn, trade off the larger potential volume of customers at a center offering the advantages of product variety against the lower degree of monopoly power achieved there. This type of model leads one to expect more homogeneous products to be sold in smaller centers, and more differentiated products, as well as big ticket items, to be sold in larger centers. The result is a hierarchy of centers analogous to the hierarchy of cities in the central place theories of Christaller (1933) and August Lösch (1940). The pattern is further complicated by complementarities that arise if consumers purchase multiple goods on a single trip, giving retailers of different goods an added incentive to locate in the same place (Robert Bacon 1984).

When economies of scale are less important but product variety is still valued, firms are more numerous and so may engage in *monopolistic competition*, in which strategic considerations are absent. One particular model of such a situation, by Avinash Dixit and Stiglitz (1977), has been used by others to derive results on agglomeration which can be interpreted as applying either at an intraurban or regional scale (e.g. Hesham Abdel-Rahman and Fujita 1990; Fujita and Tomoya Mori 1997). In two models by Krugman (1991b, 1993), co-location of all of the monopolistically competitive firms at a single point is a stable outcome when transport costs are low. Fujita (1988) has shown that introducing a land market into such models causes the agglomeration of firms to spread out as firms economize on rent, and generates a variety of possible equilibria in which residential and commercial land uses can be either mixed or segregated, monocentric or polycentric,

depending on the structure of transport costs and consumer preferences.

5.5 *Stability, Growth, and Dynamics*

Recall that some of the static models we have discussed display multiple equilibria and catastrophic comparative statics. Adding a dynamic adjustment mechanism should then produce a model in which complex and interesting spatial patterns evolve over time.[24] The two-location model of Anas (1992) provides a simple illustration. Each location is a potential center, containing a fixed amount of land. Total population is N. Individuals at any location maximize a utility function depending on per-capita output at that location and on per-capita land consumption there. Per-capita land consumption at location i decreases with the number of people n_i there, but localization economies cause per-capita output at i to rise with n_i. Writing the resulting utility as $V(n_i)$, assume that $V(n_i)$ is inverted U-shaped with a maximum at n^*, and that $V(N) > V(0)$.

Our assumptions guarantee that the monocentric outcome, with all population in one center, is an equilibrium. So is the symmetric duocentric outcome with two centers, each of size $N/2$. A duocentric equilibrium is characterized by the condition $V(n_1) = V(N\text{-}n_1)$, so that no one has an incentive to move.

Consider, however, a dynamic adjustment mechanism in which migration occurs from a low- to a high-utility location. If $N < 2n^*$, the symmetric duocentric equilibrium is unstable because a small perturbation (i.e. a randomly sized group migration) that makes one center larger gives it a localization advantage, causing it to grow larger still until it absorbs all the population. Thus when the city is small, it is monocentric. But which of the two locations becomes the monocenter is determined by chance.

Larger cities are more interesting. If $N > 2n^*$ two things happen. First, the symmetric duocentric equilibrium is now stable and in fact Pareto superior to the monocentric one. Second, while the monocentric equilibrium remains locally stable, it is upset by a random migration of n' or more people from the monocenter to the other location, where n' is a number which depends on N. That is, it takes a certain threshold size n' to make a viable subcenter in the presence of an initial monocenter.[25] This suggests that some sort of coordination is required to move from the less efficient to the more efficient structure.

As it happens, n' is a decreasing function of N. We can now see what happens to a small city that grows. Suppose that in each time period, a randomly sized group migration from one location to the other occurs with probability proportional to the utility differential between the two locations. (The microfoundations for such fluctuations could, for example, include random migrations by small groups or herd behavior caused by signalling phenomena.) When total population is small, there will be just one center. As population grows, the one center remains but becomes less and less stable. Eventually a group migration produces a viable subcenter, which then grows rapidly until there are two equal-sized centers; but chances are this will not occur until well after the initial monocenter becomes inefficiently large. This suggests that a grow-

[24] The multiplicity of equilibria, their stability, and the patterns of path-dependence are analyzed explicitly in Fujita and Ogawa (1982) and in Anas and Kim (1996). These properties are implicitly present in the models of Papageorgiou and Smith (1983), Fujita (1988), Krugman (1991b, 1993) and Fujita and Mori (1997).

[25] Hence two cities of size n' and $N-n'$ are another duocentric equilibrium, this one asymmetric; but it is locally unstable.

ing CBD can become too large because of coordination failures among potential outmigrants.

The process of "edge city" formation envisioned by Henderson and Arindam Mitra (1996) is one way in which subcenters can be sized and timed more efficiently. In their model firms decide whether to relocate from the monocenter to a new edge city. The essential innovation is the introduction of a developer who helps the migration process along by internalizing some of the external benefits that migrants to the edge city confer on each other. The developer is engaged in a game with the city government, which exercises influence over conditions in the original center. Henderson and Mitra examine the strategic considerations facing the developer, finding a rich set of possible decisions concerning the location and size for an edge city. The developer internalizes some of the externalities, but introduces new ones due to strategic effects. The role of developers is only just beginning to receive attention in the economic literature, but clearly it is quite important in practice.[26]

5.6 *Noneconomic Dynamic Models*

The existence of multiple centers, the irregularity of spatial forms, and the unpredictability of how they evolve are important features of the modern urban landscape. Similar properties are also known to arise in a variety of nonlinear dynamic processes in chemistry, physics, and biology. As a result, some of the more interesting infusions of ideas into urban economics and urban geography are coming from those fields. In particular, urban structure is proving to be a fertile application of generalized concepts such as chaos, complexity, fractals, dissipative structures, and self-organization. All involve some form of positive feedback (Brian Arthur 1990), which in the urban growth context takes the form of development at one location somehow enhancing the development potential of nearby locations. This, of course, is just another description of agglomeration economies; the difference is that this strain of literature has emphasized the dynamic analytics of such feedback mechanisms rather than their economic underpinnings. In this sense it resembles many macroeconomic models.

These models typically explore systems that are out of equilibrium, an approach now also established in evolutionary economics (Richard Nelson 1995) and one that is amply justified by the durability of urban structures. Unfortunately, the models often lack prices and so may neglect forces tending toward the restoration of equilibrium. But are spatial interactions mediated through prices more important than unpriced spatial influences and externalities? Since unpriced externalities probably play a dominant role in shaping urban spatial structure, the challenge posed by the noneconomic models cannot be easily dismissed. What follows is a sampler of these noneconomic models from a quite eclectic literature centered mostly in geography and regional science. We attempt to extract some basic insights which are useful to economic models.

Markovian models explain the transitions of micro units from one state to another: development or redevelopment of a parcel of land, household migrations, and the birth or death of firms. Agglomeration effects imply that individual transition probabilities depend on the number of actors in each state, as in interactive Markov chain models (John Conlisk 1992). A model

[26] It is also important for equilibrium in Tiebout models of local public goods, as demonstrated by Henderson (1985).

whose macro features depend on the particular realization of stochastic transitions is a model in which history matters, just as recent work has shown that it matters in other fields of economics (Paul David 1985; Arthur 1989) and just as it matters in the economic models with multiple equilibria discussed earlier.

Looked at more abstractly, positive feedback reinforces certain perturbations in the urban system and can therefore amplify some random fluctuations. Such fluctuations are driving forces in dynamic theories of *self-organization*. In some circumstances fluctuations result in sudden shifts from one relatively stable state to another, a phenomenon resembling punctuated equilibria in biological evolution (Niles Eldredge and Stephen Jay Gould 1972). Krugman (1996) uses Fourier analysis to decompose a random perturbation (such as the irregular spatial pattern of employment changes caused by building a large plant) into an infinite series of regularly spaced fluctuations at different spatial frequencies. A physical analogy is the decomposition of the sound of plucking a violin into a set of audible harmonic frequencies known as a tone and overtones. Just as the violin body amplifies some frequencies and dampens others, the urban system causes some of the regular spatial fluctuations to be magnified (as with an influx of new firms in a regular pattern) and others to be suppressed (as with the closing of unsuccessful firms due to unfavorable location patterns vis-à-vis their competitors). The result of selective amplification is recognizable macro spatial features such as a tendency toward a particular spacing among urban subcenters. By understanding the properties of the "amplifier," which is just a set of dynamic equations, we obtain insight into the varying spatial scales at which agglomeration or congestion effects occur. Some such effects are based on personal interaction, producing the classic CBD. Others are based on daily or weekly trip-making, yielding spatial structures at scales up to an hour or so of travel. Others are based on interregional or international trade, yielding size hierarchies of cities at a national, continental, or even global scale.

Diffusion and Percolation are dynamic physical processes in which the evolution of a macro state, such as the flow of water through porous rock, is governed by microscopic obstructions whose precise locations are random. (An urban development analogy would be a new firm seeking to assemble a large land parcel in an area with many small parcels that are randomly occupied.) Relationships between such macro quantities as water pressure and average flow can be derived from the statistical properties of the obstructions, even though the exact pattern of pathways is random. Electrical conductivity and magnetization of minerals operate in somewhat similar ways (Armin Bunde and Shlomo Havlin 1996). A. Stewart Fotheringham, Batty and Longley (1989) propose that in an analogous way, discrete lumps of development arrive randomly at the edge of a metropolitan area and seek suitable vacant sites. Agglomeration is posited by requiring that a new lump may settle only on the edge of an existing cluster of development. The resulting patterns of developed land are fractals, and Batty and Longley (1994) use this model to derive the fractal patterns which, as noted in Section 2, they believe characterize urban development.

Hernan Makse, Havlin and Eugene Stanley (1995) propose a model with somewhat stronger agglomeration tendencies known as *correlated percola-*

tion, in which the development probability for a given site increases with the proximity of other occupied sites and decreases with distance from an exogenous monocenter. Simulations yield growth patterns that resemble, at least impressionistically, the historical development of Berlin from 1875 to 1944, which especially in the later years showed a high degree of irregularity. Perhaps the main advantage of such models is the tools they offer for analyzing irregularity—for example, the fitting of power laws to the size distributions of local spatial fluctuations.

Per Bak and Kan Chen (1991) have shown that many dramatic physical phenomena, including avalanches and earthquakes, occur when the dynamics of a system push it to an ordered state that is just on the edge of breakdown. Given such a state of *self-organized criticality*, small fluctuations cause chain reactions whose sizes typically obey a power-law distribution. Krugman (1996) hints that the interactions among economic agents may produce similar states in cities, as well as in other economic situations, and that this may explain the prevalence of sudden transitions such as the extremely rapid growth of new edge cities. Extensions of economic models that produce sudden growth, such as those of Krugman (1996) and Anas (1992), could perhaps produce temporary states of self-organized criticality with testable statistical properties.

Regional scientists have long been interested in models in which the attractiveness of a location, for example a shopping center, is enhanced by large size. As already discussed, such models are capable of generating bifurcations, in which small shifts of parameter values produce qualitatively different equilibrium configurations, some stable and some not. Peter Allen and collaborators have put some of the same ideas into dynamic models intended to describe urban or regional growth processes that may be far from equilibrium. This work is part of a more general movement, inspired by Ilya Prigogine, to describe systems that maintain organized structure against the ravishes of entropy. Such systems are called *dissipative structures* (G. Nicholis and Prigogine 1977; John Foster 1993).

Allen's models are based upon interdependent growth equations for population and employment which incorporate both agglomeration economies and congestion diseconomies. For example, in the model of Allen and M. Sanglier (1981), employment S in a given region and sector obeys a dynamic equation in which dS/dt is proportional to $S \bullet (E\text{-}S)$, where E is a measure of "potential employment demand." This potential demand is in turn determined by other equations in the system that account for the location's relative attractiveness, crowding, and a rather arbitrary "natural carrying capacity." Thus existing employment attracts new employment, but eventually the location becomes saturated. The authors create simulations in which random fluctuations cause the spontaneous creation of centers, which subsequently grow along a path resembling a logistic curve. Most simulations lead to a stable but not necessarily unique steady state. Constraints such as zoning regulations, if added early in the simulation, can affect which of the possible steady states occurs. This model and related ones have been calibrated for a number of cities and regions in Belgium, France, Senegal, and the United States (Allen 1997).

Most of the noneconomic models described here lack a price system and any explicit description of rational economic decision-making. Furthermore, their

dynamic behavior is backward- rather than forward-looking. Thus, for all their tantalizing success in portraying the complexity in the dynamics of urban structure, they fail to incorporate economic explanations. Fortunately, they tend to be based on the behavior of individual units and so are not fundamentally incompatible with economic reasoning. This suggests that advances might be achieved by some merging of modeling techniques. Either economic behavior might be inserted rigorously into existing noneconomic models, or attractive analytical features from those models might be blended into existing models in urban economics.

An example of the first approach is by Hsin-Ping Chen (1996), who shows that a rigorous microeconomic model can generate macro-level equations like those of Allen and Sanglier. Chen's model contains land and labor prices, development and abandonment decisions, and other recognizable microeconomic constructs, all within a framework of agglomeration economies and congestion. She produces abstract simulations much like those of Allen and Sanglier, and in other work (Chen 1993) makes a plausible case for replicating the 1970–80 growth of the Los Angeles region with a calibrated version of the model.

6. *The Welfare Economics of Urban Structure*

In defense of the low-density development that increasingly characterizes modern cities, Gordon and Richardson (1997) have argued that the urban spatial structure generated by market forces reflects the will of the people—or more precisely, that it is a successful and largely desirable adaptation to the forces of urban growth and congestion. Planners, in contrast, typically have little faith in either the efficiency or the equity of market-determined outcomes, and advocate detailed land use planning. To evaluate these conflicting points of view we need to explore the welfare economics of urban land use. In this section we attempt to show how some of the prominent policy questions can be illuminated, if not answered, by building on the theoretical models and empirical observations of the previous sections.

6.1 *Can Agglomeration Economies Be Internalized?*

We have seen that although agglomeration economies are the raison d'étre of most cities, their exact nature is in flux and only partially understood. Our current understanding of them is based on a variety of factors including Smithian specialization, idiosyncratic matching, interaction, and innovation. Because these notions are broad ones, no one has really succeeded in coming to grips with how they affect the industrial organization of the modern city. Why, if there are economies of scale, is production not undertaken by a single large firm? Why do some forms of interaction occur within firms, while others operate through the market and yet others take place informally? And why do some interactions appear to require face-to-face contact while others can be effected via telecommunication? The answers given to these questions often refer to transactions costs, incomplete contracts, trust, and flexibility.

Does the market—broadly speaking—deal efficiently with agglomeration economies? The standard answer is negative. If scale economies are internal to firms, then efficient pricing cannot be supported by competition. If they are external, firms will under-employ those business practices that contribute social value to their neighbors.

The standard argument, however, neglects that efficiency could be achieved by competition among private city-developers who would set up efficient cities, thereby internalizing the agglomeration economies. Each city would operate at minimum average cost—a point of locally constant returns to scale—with increasing returns in the production of goods being balanced by decreasing returns in the production of accessible land, due to the higher costs of transport and communications in larger cities. Under marginal-cost pricing, the losses from production of goods would be just offset by profits on the production of accessible land, which are manifested as land rents—a variant of the Henry George Theorem (Richard Arnott and Stiglitz 1979). When developers make decisions concerning the internal structure of edge cities, they are to a limited extent playing this role. We do not, however, observe developers trading cities in a competitive market; so it is doubtful that agglomeration economies can be fully internalized in this way. Government intervention can help in principle, but until the sources of market failure are better understood it risks making things worse instead of better—as has also been argued in the international trade context (Krugman 1987).

6.2 *How Efficient Is Subcenter Formation?*

We have seen how agglomeration economies tend to create clusters of economic activity within a city and how these clusters influence surrounding residential densities. Given the rich nature of interactions within urban areas, such clusters play a variety of roles. What can we say about the optimality of the resulting pattern?

Our theoretical review suggests that urban subcenters, like cities themselves, are formed from the tension between agglomerative and dispersive forces. Both sets of forces entail strong externalities—external economies producing the agglomerative tendencies, and congestion or nuisance externalities limiting the size and density of the agglomeration that is achieved. The first set of externalities is largely positive, suggesting an inadequate private incentive to join an agglomeration and hence excessive dispersion. The second set consists of negative externalities, so may cause too many activities to locate close together. Since different externalities operate at different scales, it is quite possible for the spatial pattern of economic activity to be too centralized at one scale (e.g. cities that are too big) and too dispersed at another (e.g. subcenters that are too small). To further complicate matters, the externalities are linked. For example, downtown congestion, along with the excessive residential decentralization caused by underpriced transport, may give rise to excessive employment decentralization (because jobs follow households), which may in turn spawn excessively large secondary agglomerations.

The two-location model of Anas (1992), reviewed in the previous section, illustrates these problems in a dynamic setting. As the population of the first center grows, there comes a time when it is optimal for a mass of population to move to the second location. Since, however, the social gains from relocation exceed the private gains, under atomistic migration the second center will not be established until probably much later. According to this reasoning, some collective action is needed not only to establish the second center at the right time but also to protect it until it becomes stable and self-sustaining. In principle, a private developer has a profit incentive to form the

second center at the right time;[27] but in a more realistic model with multiple locations, the strategic rivalry among potential developers, each trying to create a subcenter, results in other inefficiencies (Henderson and Slade 1993). There may therefore be a role for government in assisting subcenter formation—for example by providing infrastructure, regulating or subsidizing developers, or subsidizing firm location. On a regional or national scale, analogous issues have been raised in the debates over France's "pôles de croissance," Britain's New Towns policy, and policies of less developed nations to divert growth away from their "primate" cities which contain large percentages of the national urban populations.

The comparison between optimal and market-determined spatial structure is further complicated by history dependence. The most obvious source is the durability of structures and infrastructure. But as we have seen, even in the absence of durability one can have multiple stable equilibria, with some more efficient than others and with history determining which obtains. On balance, therefore, it appears formidably difficult to ascertain how the actual size distribution and composition of subcenters differs from the optimum under realistic situations. While there is certainly scope for ameliorative government action, a precise prescription of good planning in this arena remains elusive.

6.3 *Does Traffic Congestion Cause Excessive Decentralization?*

In the basic monocentric-city model, urban spatial structure is efficient. It is reassuring that the Invisible Hand can work with respect to the location of economic activities. Unfortunately, this efficiency property is not very robust theoretically, and is of questionable practical relevance because of the pervasiveness of externalities in actual cities. One of the most serious is traffic congestion.

The congestion externality arises because the user of a motor vehicle does not pay for its marginal contribution to congestion. Consequently, the private cost of travel during peak periods falls short of the social cost. Travel is misallocated across transport modes, routes, and times of the day, and overall travel may be excessive too. As is well known, this externality can be internalized by means of a congestion toll equal to the marginal congestion externality evaluated at the optimum. However, optimal congestion tolls are charged nowhere and congested travel is underpriced almost everywhere. Uncongested travel, by contrast, may be considerably overpriced, especially in nations with high fuel taxes.

What does this imply about urban form? Even in today's complex urban structures, the most severe congestion continues to occur on radial travel to and from the central business district (CBD), and it is here that underpricing is most severe. If urban structure is fundamentally shaped by commuting costs to the CBD, as postulated by the monocentric model, then such underpricing causes the city to be more spread out than is optimal. This excessive residential decentralization is compounded by a less obvious effect: underpricing travel distorts land values in a way that encourages planners to allocate too much downtown land to roads (Arnott 1979). To see why, suppose the only cost associated with a road is the oppor-

[27] On a smaller scale, James Rauch (1993) shows how the developer of an industrial park, in which there are production complementaries between firms, can achieve efficiency by subsidizing the first firms moving into the park in order to attract additional tenants. Shopping centers employ a similar strategy by giving rental discounts to anchor stores.

tunity cost of the land it uses. Now let the planner employ the following "naive" cost-benefit rule: at each location, expand the road until the incremental travel-cost saving from further expansion equals the residential market value of the incremental land required. However, the market value of residential land reflects only the *private* transport-cost savings from a more central location, not the *social* savings which—because of underpriced congestion—are greater. The market therefore undervalues downtown residential land, so that application of the naive rule results in too much land there being devoted to roads. Another way of viewing it is that the naive rule ignores the contribution to congestion of "induced traffic," i.e., traffic caused by land-use changes induced by the highway investment. Wheaton (1978) has argued that such a mechanism resulted in massive overbuilding of urban highways in the U.S. during the 1950's and 1960's.

This reasoning, of course, must be modified when one takes into account non-central employment. As congestion builds near the city center, some centrally located employers respond by moving out of the CBD and closer to their workers and customers, with agglomerative forces causing some of this employment to become clustered in subcenters. As the metropolitan area evolves from a monocentric to a dispersed or polycentric structure, average travel times and congestion levels are reduced. This phenomenon is empirically documented by Gordon, Ajay Kumar, and Richardson (1989) and Gordon and Richardson (1994), and occurs in simulations based on the theoretical model of Anas and Kim (1996).

Clearly, however, the process of decentralization does not occur efficiently because the congestion externality remains. Highly accessible land is still underpriced and hence is developed at inefficiently low density. So the resulting land use pattern is likely to be inefficiently dispersed (not clustered enough). It is more difficult to say if the pattern is also inefficiently decentralized (too spread out from the center) because the timing of polycentric development depends on how the land development industry is organized. If the industry is dominated by a few large developers, then timing is affected by strategic interdependence; whereas if there are instead many small developers, timing is influenced by coordination failure and the dynamics of herd behavior.

Possible second-best policies to correct excessive decentralization, if such is the case, include more sophisticated cost–benefit analysis of transport projects, minimum-density controls, and greenbelts. In fact, policies in the United States have worked in exactly the opposite direction, as emphasized by Anthony Downs (1992) and others. Subsidies for home ownership, subsidized highway construction and maintenance, and minimum-lot-size residential zoning are just some of the measures which have increased decentralization, even while keeping the poor excessively concentrated in the central cities. In response to the ongoing transformation in urban form, the planning community has tended to advocate policies aimed at reversing decentralization, reducing automobile use, and revitalizing the downtown core—for example building mass transit facilities or downtown convention centers. But because the pricing errors of the past have been cast in brick and asphalt, such policies are very expensive and have limited effectiveness.

6.4 *When Are Land-Use Controls Justified?*

Given the many externalities revealed by our theoretical review, it is tempting

to conclude that only very comprehensive and detailed planning can overcome the resulting inefficiencies. Because the externalities are so poorly understood, however, attempted cures may well do more harm than the disease. The brief discussion below illustrates the complexity of determining one aspect of optimal policy: land-use planning.

First, consider *incompatible land uses*. Cities are awash in very localized externalities, from the smells from a fish shop to the blockage of ocean views by neighbors' houses. Mills and Hamilton (1994, pp. 252–54) argue that they are not significant, but that may be because the worst have been eliminated by zoning. Because pricing solutions in this context would be extremely cumbersome, zoning is a potentially valuable tool for dealing with incompatible land uses. However, it can easily be overdone; for example, the complete separation of retail and residential land uses results in visual monotony and unnecessary auto travel.

Second, consider *preservation of open space*. Greenbelts and urban parks are potentially valuable public goods, and government intervention is probably the only viable way to ensure their provision. It is important to recognize, however, that *someone* is implicitly bearing the cost of designating areas off-limits to development. The increased scarcity of residential land induced by greenbelts drives up land rents and hence housing rents. So the bucolic landscapes surrounding London and Paris arguably come at the cost of miserable and badly overcrowded neighborhoods for the poor. Where such controls divert growth from the entire metropolitan area, they may improve the local environment but against this must be weighed the environmental cost of growth elsewhere. In other situations, greenbelts are likely to spawn exurban development further out, which raises another set of issues for growth management.

Third, consider *urban sprawl*, a pejorative term often used for leapfrogging in development. This appears inefficient at first glance. But what some planners see as haphazard development may well be the seeds of future agglomerations, and the land left vacant can be developed later at higher density than is justified today.

Another argument for greenbelts or growth boundaries is *maintenance of viable central cities*. Critics of current development patterns argue, with some justification, that misguided policies have produced excessively decentralized cities at great cost in duplicative infrastructure and with disastrous results for the poor who live in concentrations of blight (David Rusk 1993; Downs 1994; Myron Orfield 1997). Some of these authors argue for growth boundaries to force new development into the central cities in hopes of revitalizing them. But such gross restrictions may well have perverse distributional consequences: the prior owners of land within the boundary enjoy windfall gains at the expense of nascent businesses and new home buyers, while inner-city renters—who are disproportionately poor—must pay more for housing.

Finally, consider *exclusionary zoning*. Many suburban municipalities enforce minimum-lot-size restrictions, largely in order to exclude lower-income residents who would pay less in property taxes while receiving the full benefits of the local public goods. Such restrictions may also be designed to exclude undesired socioeconomic, racial, or ethnic groups. Exclusionary zoning probably adds considerably to decentralization as well as fostering social stratification,

segregation in education, and racial division. By forcing the poor to live in central cities, it also limits their access to suburban blue-collar jobs, a phenomenon known as spatial mismatch (Kain 1968). These are all reasons why higher levels of government might want to encourage suburban municipalities to be more receptive to high-density housing targeted to lower-income residents.

6.5 *Summary: The Role of Government Policy*

As in so much of economic policy analysis, it is hard to make overall recommendations about the scope of government intervention. Theory provides clear instances of market failure, against which must be balanced the likelihood and severity of government failure. An interesting object lesson is Paris, whose urban form has been strongly influenced by government intervention to limit central building heights and to channel exurban development towards planned satellite towns. The result is a city regarded by many as extremely attractive and vital. Others prefer the convenience, lower cost, and ease of interaction of Los Angeles, which Paris would probably come to resemble absent government policy.

What seems clear to us is that cities are complex entities in which market forces are both powerful and beneficial in many ways, obvious and subtle. These market forces sometimes need to be controlled or channeled, yet they tend to find their own way of thwarting such restrictions. Whether a particular government policy is enlightened intervention or misguided meddling will inevitably be debated case by case.

7. *Conclusion*

And so we see that cities are strongly shaped by agglomeration economies, especially external scale economies. Cities teem with positive and negative externalities, all acting with different strengths, among different agents, at different distances. Some people need to interact frequently face-to-face; others carry out routine actions remotely via telecommunications but must meet periodically to create and renew trust; still others learn crucial information by overhearing conversations at restaurants, bars, parties, or meetings. Consumers want to purchase some goods often, other infrequently; some want to see and touch goods, others to hear about them from a friend; for some any variety will do, for others a specific variety is required. The pedestrian and car traffic generated by one firm as a side effect can make or break another firm's business, as window shoppers stop at an intriguing display or as disreputable patrons scare away a neighbor's potential workers, residents, or customers. Together these many interactions, helped by history and a good deal of chance, produce the spatial structure that we see. Is it any wonder that spatial patterns are complex, that they occasionally display sudden change, or that tractable models can capture only a portion of their rich variegation?

Agglomeration economies have resisted attempts to fully understand their microfoundations. This is illustrated by urban economists' lack of confidence in forecasting the effects of the communications revolution on urban spatial structure. On the theoretical side, we do not know the scale at which the various forces work or what kinds of equilibria the simultaneous interaction of many forces will produce; nor do we have reliable models of dynamic growth paths with random shocks. We also do not know which external economies will be internalized through private initia-

tive. On the empirical side, despite the increasing sophistication of studies relating a firm's productivity to the size and industrial composition of the city in which it is located, we do not really know the specific forces that produce these relationships, nor just how they depend on industry mix, industrial policy, local public goods, or zoning.

Complicating matters even more are the longevity of urban structures, including public infrastructure, and the stability of certain equilibria even when other equilibria exist that would make everyone happier. Urban structure locks in past forces that may have little bearing today. Precious little traffic now uses the locks on the Erie Canal that are the namesake of Lockport, New York; yet that is where its downtown remains. Other downtowns may be overcrowded because no developer has managed to assemble land or obtain zoning variances needed to establish a much-needed satellite center.

We have seen that forces that are ascendant throughout the world are producing decentralization and dispersion at a citywide scale, and agglomeration at a local scale. Will Paris and Tokyo, then, go the way of Los Angeles? To a large extent they already have; in both, as in cities throughout the developed world, automobile-age development has created a vast periphery of residential suburbs with outlying commercial, office, and industrial centers. Apparently these patterns are not just the product of crazy Americans in love with their cars. Yet Paris and Tokyo have each preserved a distinctive city center, in part through strict zoning and by valuing historic preservation. In addition Paris, like other cities including London and Seoul, has seen the shape of decentralization and dispersion altered by central-government policies that zone large tracts of outlying land as greenbelts and create satellite cities. Our review suggests that such policies can in principle elicit more efficient growth paths; but that serious undesirable side effects are likely. As for the city centers, whether the desire to maintain their special character can stave off the forces of economic change depends both on politics and on the ultimate preferences of the citizenry.

References

Abdel-Rahman, Hesham and Masahisa Fujita. 1990. "Product Variety, Marshallian Externalities, and City Sizes," *J. Reg. Sci.*, 30:2, pp. 165–83.

Allen, Peter M. 1997. *Cities and Regions as Self-Organizing Systems: Models of Complexity.* Amsterdam: Gordon and Breach Science Pub.

——— and M. Sanglier. 1981. "A Dynamic Model of a Central Place System-II," *Geographical Analysis*, 13:2, pp. 149–64.

Alonso, William. 1964. *Location and Land Use.* Cambridge, MA: Harvard U. Press.

Anas, Alex. 1990. "Taste Heterogeneity and Urban Spatial Structure: The Logit Model and Monocentric Theory Reconciled," *J. Urban Econ.*, 28:3, pp. 318–35.

———. 1992. "On the Birth and Growth of Cities: Laissez-Faire and Planning Compared," *Reg. Sci. Urban Econ.*, 22:2, pp.243–58.

Anas, Alex and Ikki Kim. 1992. "Income Distribution and the Residential Density Gradient," *J. Urban Econ.*, 31:2, pp. 164–80.

———. 1996. "General Equilibrium Models of Polycentric Urban Land Use with Endogenous Congestion and Job Agglomeration," *J. Urban Econ.*, 40:2, pp. 232–56.

Arnott, Richard J. 1979. "Unpriced Transport Congestion," *J. Econ. Theory*, 21:2, pp. 294–316.

——— and Joseph E. Stiglitz. 1979. "Aggregate Land Rents, Expenditure on Public Goods, and Optimal City Size," *Quart. J. Econ.*, 93:4, pp. 471–500.

Arthur, W. Brian. 1989. "Competing Technologies, Increasing Returns, and Lock-In by Historical Events," *Econ. J.*, 99:394, pp. 116–31.

———. 1990. "Positive Feedbacks in the Economy," *Sci. Amer.*, 262:2, pp. 92–99.

Bacon, Robert W. 1984. *Consumer Spatial Behavior.* Oxford, UK: Clarendon Press.

Bailey, Jeff and Calmetta Y. Coleman. 1996. "Despite Tough Years, Chicago Has Become a Nice Place to Live," *Wall Street Journal*, Aug. 21, 135:37, pp. 1, 6.

Bak, Per and Kan Chen. 1991. "Self-Organized Criticality," *Sci. Amer.*, 264:1, pp. 46–53.

Barrett, Paul. 1983. *The Automobile and Urban Transit: The Formation of Public Policy in Chi-*

cago, 1900–1930, Philadelphia: Temple U. Press.

Batty, Michael and Paul Longley. 1994. *Fractal Cities: A Geometry of Form and Function*. London: Academic Press.

Berliant, Marcus and Hideo Konishi. 1996. "The Endogenous Foundations of a City: Population Agglomeration and Marketplaces in a Location-Specific Production Economy," working paper, U. Rochester.

Borukhov, E. and Oded Hochman. 1977. "Optimum and Market Equilibrium in a Model of a City without a Predetermined Center," *Environment & Planning* A, 9:8, pp. 849–56.

Brueckner, Jan. 1987. "The Structure of Urban Equilibria: A Unified Treatment of the Muth-Mills Model," in *Handbook of Regional and Urban Economics, Vol. II: Urban Economics*. Edwin S. Mills, ed. North-Holland, Amsterdam, , pp. 821–45.

Bunde, Armin and Shlomo Havlin. 1996. "Percolation I," in *Fractals and Disordered Systems*. Armin Bunde and Shlomo Havlin, eds. Berlin: Springer-Verlag, pp. 59–113.

Cervero, Robert and Kang-Li Wu. 1996. "Subcentering and Commuting: Evidence from the San Francisco Bay Area, 1980–1990," Working Paper, Inst. of Urban & Regional Development, U. C. Berkeley.

———. 1997. "Polycentrism, Commuting, and Residential Location in the San Francisco Bay Area," *Environment and Planning A*, 29:5, pp. 865–86.

Chen, Hsin-Ping. 1993. *Theoretical Derivation and Simulation of a Nonlinear Dynamic Urban Growth Model*. Ph.D. Dissertation, Dept. of Econ., U. California Irvine.

———. 1996. "The Simulation of a Proposed Nonlinear Dynamic Urban Growth Model," *Annals Reg. Sci.*, 30:3, pp. 305–19.

Chinitz, Benjamin. 1960. *Freight and the Metropolis*, Cambridge, MA: Harvard U. Press.

———. 1961. "Contrasts in Agglomeration: New York and Pittsburgh," *Amer. Econ. Rev., Papers & Proceedings*, 51:2, pp. 279–89.

Christaller, Walter. 1933. *Central Places in Southern Germany*, C.W.Baskin. Trans. London: Prentice-Hall [1966].

Clark, Colin. 1951. "Urban Population Densities." *J. Royal Statistical Society: Series A*, 114:4, pp. 490–96.

———. 1967. *Population Growth and Land Use*, London: Macmillan.

Conlisk, John. 1992. "Stability and Monotonicity for Interactive Markov Chains," *J. Math. Sociology*, 17:2–3, pp. 127–43.

Craig, Steven G., Janet E. Kohlhase and Steven C. Pitts. 1996. "The Impact of Land Use Restrictions in a Multicentric City," working paper, U. Houston.

Cronon, William. 1991. *Nature's Metropolis: Chicago and the Great West*, NY: Norton.

David, Paul A. 1985. "Clio and the Economics of QWERTY," *Amer. Econ. Rev.*, 75:2, pp. 332–37.

Dixit, Avinash K. and Joseph E. Stiglitz. 1976. "Monopolistic Competition and Optimum Product Diversity," *Amer. Econ. Rev.*, 67:3 , pp. 297–308.

Downs, Anthony. 1992. *Stuck in Traffic: Coping with Peak-Hour Traffic Congestion*. Washington, D.C.: Brookings Institution.

———. 1994. *New Visions for Metropolitan America*. Washington, D.C. and Cambridge, MA: Brookings Institution and Lincoln Institute of Land Policy.

Eberts, Randall W. and Daniel P. McMillen. Forthcoming. "Agglomeration Economies and Urban Public Infrastructure," in: *Handbook of Regional and Urban Economics, Volume 3: Applied Urban Economics*, Paul Cheshire and Edwin S. Mills, eds. Amsterdam: North-Holland.

Edmonston, Barry. 1975. *Population Distribution in American Cities*. Lexington, MA: D.C. Heath.

Eldredge, Niles and Stephen Jay Gould. 1972. "Punctuated Equilibria: An Alternative to Phyletic Gradualism," in *Models in Paleobiology*. T.J.M. Schopf, ed. San Francisco: Freeman, Cooper & Co., pp. 82–115.

Ellison, Glenn and Glaeser, Edward L. 1997. "Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach," *J. Polit. Econ.*, 105:5, pp. 889–927.

Fales, Raymond and Leon N. Moses. 1972. "Land Use Theory and the Spatial Structure of the Nineteenth Century City," *Papers & Proceedings Reg. Sci. Association*, 28, pp.49–80.

Field, Alexander J. 1992. "The Magnetic Telegraph, Price and Quantity Data and the New Management of Capital," *J. Econ. History*, 52:2, pp.401–13.

Foster, John. 1993. "Economics and the Self-Organisation Approach: Alfred Marshall Revisited," *Econ. J.*, 103:419, pp. 975–91.

Fotheringham, A. Stewart, Michael Batty, and Paul A. Longley. 1989. "Diffusion-Limited Aggregation and the Fractal Nature of Urban Growth," *Papers & Proceedings Reg. Sci. Association*, 67, pp. 55–69.

Frank, Robert H. 1988. *Passions Within Reason: The Strategic Role of the Emotions*. NY: Norton.

Fujita, Masahisa. 1988. "A Monopolistic Competition Model of Spatial Agglomeration : Differentiated Products Approach," *Reg. Sci. & Urban Econ.*, 18:1, pp.87–124.

———. 1989. *Urban Economic Theory: Land Use and City Size*, Cambridge, UK: Cambridge U. Press.

——— and Tomoya Mori. 1997. "Structural Stability and Evolution of Urban Systems," *Reg. Sci. & Urban Econ.*, 27:4–5, pp.399–442.

Fujita, Masahisa and Hideaki Ogawa. 1982. "Multiple Equilibria and Structural Transition of Non-monocentric Urban Configurations," *Reg. Sci. & Urban Econ.*, 12:2, pp.161–96.

Gabszewicz, Jean Jaskold and Jacques-François Thisse. 1986. "Spatial Competition and the Location of Firms," in *Location Theory*, by Jean

Jaskold Gabszewicz, Jacques-François Thisse, Masahisa Fujita and Urs Schweizer. Vol. 5 of *Fundamentals of Pure and Applied Economics* series. Chur, Switzerland: Harwood Academic Publishers, pp. 1–71.

Gaspar, Jess and Edward L. Glaeser. 1998. "Information Technology and the Future of Cities," *J. Urban Econ.*, 43:1, pp. 136–56.

Garreau, Joel. 1991. *Edge City: Life on the New Frontier*. NY: Doubleday.

Getis, Arthur. 1983. "Second-Order Analysis of Point Patterns: The Case of Chicago as a Multi-Center Urban Region," *Professional Geographer*, 35:1, pp. 73–80.

Giuliano, Genevieve and Kenneth A. Small. 1991. "Subcenters in the Los Angeles Region," *Reg. Sci. & Urban Econ.*, 21:2, pp. 163–82.

———. 1993. "Is the Journey to Work Explained by Urban Structure?" *Urban Studies*, 30:9, pp. 1485–500.

Glaab, Charles N. and Theodore Brown. 1967. *A History of Urban America,* London: Macmillan Press.

Glaeser, Edward L. et al. 1992. "Growth in Cities," *J. Polit. Econ.*, 100:6, pp. 1126–52.

Gordon, Peter, Ajay Kumar, and Harry W. Richardson. 1989. "The Influence of Metropolitan Spatial Structure on Communting Time," *J. Urban Econ.*, 26:2, pp. 138– 151.

Gordon, Peter and Harry W. Richardson. 1994. "Congestion Trends in Metropolitan Areas," in *Curbing Gridlock: Peak-Period Fees to Relieve Traffic Congestion, Volume 2: Commissioned Papers*, Transportation Research Board *Special Report* 242. Committee for Study on Urban Transportation Congestion Pricing, National Research Council. Washington, D.C.: National Academy Press, pp. 1–31.

———. 1996. "Beyond Polycentricity: The Dispersed Metropolis, Los Angeles, 1970–1990," *J. Amer. Planning Association*, 62:3, pp. 289–95.

———. 1997. "Are Compact Cities a Desirable Planning Goal?" *J. Amer. Planning Association*, 63:1, pp. 95–106.

——— and H.L. Wong. 1986. "The Distribution of Population and Employment in a Polycentric City: The Case of Los Angeles," *Environ. & Planning A*, 18:2, pp. 161–73.

Griffith, Daniel A. 1981. "Evaluating the Transformation from a Monocentric to a Polycentric City," *Professional Geographer*, 33:2, pp. 189–96.

Hamilton, Bruce W. 1982. "Wasteful Commuting," *J. Polit. Econ.*, 90:5, pp. 1035–53.

Harris, B. and A.G. Wilson. 1978. "Equilibrium Values and Dynamics of Attractiveness Terms in Production-Constrained Spatial-Interaction Models," *Environ. & Planning A*, 10:4, pp. 371–88.

Harrison, David, and Kain, John F. 1974. "Cumulative Urban Growth and Urban Density Functions," *J. Urban Econ.*, 1:1, pp. 61–98.

Heikkila, E. et al. 1989. "What Happened to the CBD-Distance Gradient?: Land Values in a Policentric City," *Environ. & Planning A*, 21:2, pp. 221–32.

Helsley, Robert W. and William C. Strange. 1991. "Agglomeration Economies and Urban Capital Markets," *J. Urban Econ.*, 29:1, pp. 96–112.

Henderson, J. Vernon. 1985. "The Tiebout Model: Bring Back the Entrepreneurs," *J. Polit. Econ.*, 93:2, pp. 248–64.

——— and Arindam Mitra. 1996. "The New Urban Landscape: Developers and Edge Cities," *Reg. Sci. & Urban Econ.*, 26:6, pp. 613–43.

Henderson, J. Vernon and Eric Slade. 1993. "Development Games in Non-monocentric Cities," *J. Urban Econ.*, 34:2, pp.207–29.

Hoover, Edgar M. 1948. *The Location of Economic Activity*. NY: McGraw-Hill.

Hotelling, Harold. 1929. "Stability in Competition," *Econ. J.*, 39:1, pp. 41–57.

Ingram, Gregory K. and Alan Carroll. 1981. "The Spatial Structure of Latin American Cities," *J. Urban Econ.*, 9:2, pp. 257–73.

Jacobs, Jane. 1969. *The Economy of Cities*. NY: Random House.

———. 1984. *Cities and the Wealth of Nations: Principles of Economic Life*. NY: Random House.

Koopmans, Tjalling C., and Martin Beckmann. 1957. "Assignment Problems and the Location of Economic Activities," *Econometrica*, 25:1, pp. 53–76.

Krugman, Paul. 1987. "Is Free Trade Passé?" *J. Econ. Perspectives*, 1:2, pp. 131–44.

———. 1991a. *Geography and Trade*. Cambridge, MA: M.I.T. Press.

———. 1991b. "Increasing Returns and Economic Geography," *J. Polit. Econ.*, 99:3, pp. 483–99.

———. 1993. "First Nature, Second Nature and Metropolitan Location," *J. Reg. Sci.*, 33:2, pp.129–44

———. 1996. *The Self-Organizing Economy*. Cambridge, MA: Blackwell.

LeRoy, Stephen F. and Jon Sonstelie. 1983. "Paradise Lost and Regained: Transportation Innovation, Income, and Residential Location," *J. Urban Econ.*, 13:1, pp. 67– 89.

Lösch, August. 1940. *The Economics of Location*, W.H. Woglom and W.F. Stolper. Trans. New Haven: Yale U. Press [1954].

Makse, Hernan A., Shlomo Havlin, and H. Eugene Stanley. 1995. "Modelling Urban Growth Patterns," *Nature*, Oct., 377, pp. 608–12.

Maynard Smith, John. 1976. "Evolution and the Theory of Games," *Amer. Scientist*, 64:1, pp. 41–45.

McDonald, John F. 1987. "The Identification of Urban Employment Subcenters," *J. Urban Econ.*, 21:2, pp. 242–58.

———. 1989. "Econometric Studies of Urban Population Density: A Survey," *J. Urban Econ.*, 26:3, pp. 361–85.

——— and Paul J. Prather. 1994. "Suburban Employment Centres: The Case of Chicago," *Urban Studies*, 31:2, pp. 201–18.

McMillen, Daniel P. 1996. "One Hundred Fifty

Years of Land Values in Chicago: A Nonparametric Approach," *J. Urban Econ.*, 40:1, pp. 100–24.

——— and John F. McDonald. 1998a. "Population Density in Suburban Chicago: A Bid-Rent Approach," *Urban Studies*, 55:7, pp. 1119– 30.

———. 1998b. "Suburban Subcenters and Employment Density in Metropolitan Chicago," *J. Urban Econ.*, 43:2, pp. 157–80.

Mieszkowski, Peter and Edwin S. Mills. 1993. "The Causes of Metropolitan Suburbanization," *J. Econ. Perspectives*, 7:3, pp. 135–47.

——— and Barton Smith. 1991. "Analyzing Urban Decentralization: The Case of Houston," *Reg. Sci. & Urban Econ.*, 21:2, pp. 183–99.

Mills, Edwin S. 1967. "An Aggregative Model of Resource Allocation in a Metropolitan Area," *Amer. Econ. Rev.*, 57, pp. 197–210.

———. 1972. *Studies in the Structure of the Urban Economy*, Baltimore: Johns Hopkins Press.

——— and Bruce W. Hamilton. 1994. *Urban Economics*, New York: Harper Collins.

Mills, Edwin S. and Katsutoshi Ohta. 1976. "Urbanization and Urban Problems," in *Asia's New Giant: How the Japanese Economy Works*. Hugh Patrick and Henry Rosovsky, eds. Washington: Brookings Institution, pp. 673–751.

Mills, Edwin S. and Jee Peng Tan. 1980. "A Comparison of Urban Population Density Functions in Developed and Developing Countries," *Urban Studies*, 17:3, pp. 313–21.

Mirrlees, James A. 1972. "The Optimum Town," *Swedish J. Econ.*, 74:1, pp. 114–35.

Moses, Leon and Harold F. Williamson, Jr. 1967. "The Location of Economic Activity in Cities," *Amer. Econ. Rev.*, 57:2, pp. 211–22.

Murphy, Kevin M., Andrei Schleifer, and Robert W. Vishny. 1989. "Industrialization and the Big Push," *J. Polit. Econ.*, 97:5, pp. 1003–26.

Muth, Richard F. 1969. *Cities and Housing*. Chicago: The U. of Chicago Press.

Nelson, Richard R. 1995. "Recent Evolutionary Theorizing About Economic Change," *J. Econ. Lit.*, 33:1, pp. 48–90.

Nicholis, G., and Ilya Prigogine. 1977. *Self-Organisation in Non-equilibrium Systems: From Dissipative Structures to Order through Fluctuations*. NY: John Wiley.

Ó hUallacháin, Breandán. 1989. "Agglomeration of Services in American Metropolitan Areas," *Growth & Change*, 20:3, pp. 34–49.

Orfield, Myron. 1997. *Metropolitics: A Regional Agenda for Community and Stability*. Washington, D.C. and Cambridge: Brookings Institution and Lincoln Institute of Land Policy.

Papageorgiou, Yorgos Y. and David Pines. 1989. "The Exponential Density Function: First Principles, Comparative Statics, and Empirical Evidence," *J. Urban Econ.*, 26:2, pp. 264–68.

——— and T.R. Smith. 1983. "Agglomeration as Local Instability in Spatially Uniform Steady-States," *Econometrica*, 51:4, pp. 1109–19.

Powell, Walter W. 1990. "Neither Market nor Hierarchy: Network Forms of Organization," in *Res. Organizational Behavior*, 12, pp. 295–336.

Rauch, James E. 1993. "Does History Matter Only When It Matters Little? The Case of City-Industry Location," *Quart. J. Econ.*, 108:434, pp. 843–67.

Robinson, E.A.G. 1931. *The Structure of Competitive Industry*. Cambridge, UK: Cambridge U. Press.

Romer, Paul M. 1986. "Increasing Returns and Long-Run Growth," *J. Polit. Econ.*, 94:5, pp. 1002–37.

Rosen, Kenneth T. and Mitchel Resnick. 1980. "The Size Distribution of Cities: An Examination of the Pareto Law and Primacy," *J. Urban Econ.*, 8:2, pp. 165–86.

Rusk, David. 1993. *Cities Without Suburbs*. Baltimore: Johns Hopkins U. Press.

Saxenian, AnnaLee. 1994. *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*. Cambridge, MA: Harvard U. Press.

Schulz, Norbert and Konrad Stahl. 1996. "Do Consumers Search for the Highest Price? Oligopoly Equilibrium and Monopoly Optimum in Differentiated-Products Markets," *Rand J. Econ.*, 27:3, pp. 542–62.

Scott, Allen J. 1988. *Metropolis: From the Division of Labor to Urban Form*. Berkeley: U. California Press.

———. 1991. "Electronics Assembly Subcontracting in Southern California: Production Processes, Employment, and Location," *Growth and Change*, 22:1, pp. 22– 35.

Sivitanidou, Rena. 1996. "Do Office-Commercial Firms Value Access to Service Employment Centers? A Hedonic Value Analysis within Polycentric Los Angeles," *J. Urban Econ.*, 40:2, pp. 125–149.

Small, Kenneth A. 1981. "A Comment on Gasoline Prices and Urban Structure," *J. Urban Econ.*, 10:3, pp. 311–22.

———. 1992. *Urban Transportation Economics*, Vol. 51 of *Fundamentals of Pure and Applied Economics* series. Chur, Switzerland: Harwood Academic Publishers.

——— and Shunfeng Song. 1992. " 'Wasteful' Commuting: A Resolution," *J. Polit. Econ.*, 100:4, pp. 888–98.

———. 1994. "Population and Employment Densities: Structure and Change," *J. Urban Econ.*, 36:3, pp. 292–313.

Solow, Robert M. 1973. "On Equilibrium Models of Urban Location," in *Essays in Modern Economics*. Michael Parkin with A.R. Nobay, eds. London: Longman, pp. 2–16.

——— and William S. Vickrey. 1971. "Land Use in a Long Narrow City," *J. Econ. Theory*, 3:4, pp. 430–47.

Song, Shunfeng. 1994. "Modelling Worker Residence Distribution in the Los Angeles Region," *Urban Studies*, 31:9, pp. 1533–44.

Starrett, David A. 1974. "Principles of Optimal Location in a Large Homogeneous Area," *J. Econ. Theory*, 9:4, pp. 418–48.

Stiglitz, Joseph E. 1977. "The Theory of Local Public Goods," in *The Economics of Public Services*. Martin S. Feldstein and Robert P. Inman, eds. London: Macmillan. pp. 274–333.

Thomas, R.W. 1981. "Point Pattern Analysis," in *Quantitative Geography: A British View*. N. Wrigley and R.J. Bennett, eds. London: Routledge & Kegan Paul, pp. 164–76.

Tiebout, Charles M. 1956. "A Pure Theory of Local Expenditures," *J. Polit. Econ.*, 64:5, pp. 416–424.

U.S. Bureau of the Census. 1996. *Statistical Abstract of the United States: 1996*. Washington, D.C.: U.S. Government Printing Office.

Vernon, Raymond. 1960. *Metropolis 1985*, Cambridge, MA: Harvard U. Press.

von Thünen, Johann Heinrich. 1826. *Der Isolierte Staat in Beziehung auf Landwirtschaft und Nationalokonomie*. Hamburg: F. Perthes.

Warner, Sam Bass Jr. 1962. *Streetcar Suburbs: The Process of Growth in Boston :1870–1900*, Cambridge, MA: Harvard U. Press.

Wheaton, William C. . 1974. "A Comparative Statics Analysis of Urban Spatial Structure," *J. Econ. Theory*, 9:2, pp. 223–37.

———. 1977. "Income and Urban Residence: An Analysis of Consumer Demand for Location," *Amer. Econ. Rev.*, 67:4, pp. 620–31.

———. 1978. "Price-Induced Distortions in Urban Highway Investment," *Bell J. Econ.*, 9:2, pp. 622–32.

White, Michelle J. 1976. "Firm Suburbanization and Urban Subcenters," *J. Urban Econ.*, 3:4 , pp. 323–43.

———. 1988. "Location Choice and Commuting Behavior in Cities with Decentralized Employment," *J. Urban Econ.*, 24:2, pp.129–52.