Urban Spatial Structure

Alex Anas

Department of Economics
State University of New York at Buffalo
Amherst, NY 14260

Richard Arnott

Department of Economics
Boston College
Chestnut Hill, MA 02167

Kenneth A. Small

Department of Economics
University of California, Irvine
Irvine, CA 92697-5100

Working Paper
March 1997

UCTC No. 357

The University of California Transportation Center
University of California at Berkeley

The University of California Transportation Center

The University of California Transportation Center (UCTC) is one of ten regional units mandated by Congress and established in Fall 1988 to support research, education, and training in surface transportation. The UC Center serves federal Region IX and is supported by matching grants from the U.S. Department of Transportation, the California Department of Transportation (Caltrans), and the University.

Based on the Berkeley Campus, UCTC draws upon existing capabilities and resources of the Institutes of Transportation Studies at Berkeley, Davis, Irvine, and Los Angeles; the Institute of Urban and Regional Development at Berkeley; and several academic departments at the Berkeley, Davis, Irvine, and Los Angeles campuses. Faculty and students on other University of California campuses may participate in Center activities. Researchers at other universities within the region also have opportunities to collaborate with UC faculty on selected studies.

UCTC's educational and research programs are focused on strategic planning for improving metropolitan accessibility, with emphasis on the special conditions in Region IX. Particular attention is directed to strategies for using transportation as an instrument of economic development, while also accommodating to the region's persistent expansion and while maintaining and enhancing the quality of life there.

The Center distributes reports on its research in working papers, monographs, and in reprints of published articles. It also publishes *Access*, a magazine presenting summaries of selected studies. For a list of publications in print, write to the address below.

University of California Transportation Center

108 Naval Architecture Building
Berkeley, California 94720
Tel: 510/643-7378
FAX: 510/643-5456

I. 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.' (Bailey and Coleman, 1996, p. 6)

Mayor Daley is catering to 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 about 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 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 or urban planning. Our focus on internal structure should provide Mayor Daley a more complete basis for comparing Chicago's density to that of New York, or its degree of centralization to that of Los Angeles. (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.")

This is a particularly interesting time to study urban structure because cities' growth patterns are undergoing qualitative change. For many decades, even centuries, cities have been spreading out. But recently 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 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" (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 of centers altogether, prompting Gordon and Richardson (1996) to proclaim that Los Angeles, at least, is "beyond polycentricity." But even sprawl is far from homogeneous, and geographers have perceived patterns of irregularity so pronounced as to fit in certain ways 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 W. 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 agglomeration economies — those positive externalities that arise between firms because their interactions are facilitated by spatial proximity in ways not fully captured in transactions prices. Agglomeration economies place a premium on land at accessible locations; this in turn accentuates the nonconvexity in production sets that is inherent in the indivisibility of location (Starrett, 1974). Because of pervasive externalities and nonconvexities, economic analysis when applied to urban geography yields results that differ in important and interesting respects from results of other branches of economics. Agglomeration economies also create first-mover advantages and regional specializations that are important in international trade (Krugman, 1991a), and some first-mover

disadvantages that prevent optimal dynamic growth paths from being realized. Furthermore, they are suspected of giving cities a key role in generating aggregate economic growth (Jacobs, 1984).

Agglomeration economies are of course not new. As eloquently exposited by Vernon (1960) and Chinitz (1961), they are at the heart of our current understanding of central business districts. But recent events are creating new types of agglomeration economies, mediated by the properties of a world where information is even more important, transportation is faster, and long-distance communication is cheaper. 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. We believe agglomeration economies are amenable to microeconomic analysis, and we show how such analysis provides a rich set of explanations for polycentric urban structure.

While our focus is on describing and explaining urban spatial structure, we address two related issues as well. The first concerns the appropriate role of government in cities. Spatial structure is determined by the balancing of centripetal agglomerative forces and centrifugal forces related to crowding. To a large extent, these forces operate outside markets — for example, agglomeration economies are mostly external to firms, and congestion is mostly unpriced. What policies, then, can help internalize the pervasive externalities operating in cities without sacrificing the benefits of the Invisible Hand? The second issue concerns the importance of space in economics. Does the study of urban spatial structure yield new insights into economic phenomena that are normally analyzed in aspatial models? What is the level of spatial resolution at which economic activity is best analyzed?

II. 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 about the complexity encountered in measuring its characteristics.

A. 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.[1]

Prior to about 1840, the beginning of the railroad era, cities were tied to waterways such as harbors, rivers and canals. Freight moved most efficiently by barge, and the average cost of processing freight fell sharply with the quantity processed at a particular port. Cities therefore had a small number of water ports, usually just one. Railroads competed with waterways in the latter part of the 19th century, and scale economies in rail terminals were similar to those in harbors.

Moses and Williamson (1967) observed that intra-urban freight costs in the 19th century were high relative to intra-urban personal transport costs as well as high relative to interurban freight costs. These costs caused manufacturers to locate near the harbor or railhead, and residences to spread. Meanwhile, cities were located at great distances from each other.

In the last quarter of the century, the telegraph greatly speeded the flow of information from city to city, but economies of scale prevented the telegraph from being used much within a city (Field, 1992). Instead, messengers remained the primary means by which businesses communicated with each other within a city. Similarly, scale economies in railroad shipping restricted the use of railroads within cities. Intra-urban freight transport took place mainly by horse and wagon, which was time consuming and unreliable in bad weather.

These costly technologies of communication and intra-urban freight caused businesses 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 heavy manufacturing. Fales and Moses

[1]For a history of North American urbanization, see Glaab and Brown (1967).

(1972) showed empirically how in Chicago, this pattern of districts could be explained by a combination of intra-industry agglomeration economies and inter-industry linkages.

Lower factor prices for land and labor could be obtained in satellite areas, but for most firms these savings were outweighed by higher communication and freight costs. Firms also remained close to the harbor or railhead because of the durability of existing structures. The great Chicago fire of 1873 removed such constraints, making most firms footloose; studying the relocation patterns of these firms, Fales and Moses found that they located more peripherally than before the fire, while maintaining their linkages to the rail and water terminals and other central firms.

Until about 1850, personal transport within the city occurred by walking, horse-drawn carriages, horse-drawn streetcars, and in a few cases diesel trains. All except walking were very expensive and confined to a small elite, 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, causing a distinct pattern of income declining with distance from the CBD as documented in studies of Milwaukee, Pittsburgh, and Toronto (LeRoy and Sonstelie, 1983).

Between 1850 and 1880, the advent of electric streetcars and trolleys 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 (Warner, 1962). Toward the turn of the century subways further contributed to this pattern in the largest cities. Thus developed a pattern now known as the "nineteenth century city," consisting of a compact production core surrounded by an apron of residences concentrated around mass transport spokes. Fales and Moses (1972) report that 80% of the jobs in late nineteenth-century Chicago were located within a four mile radius of State and Madison streets.

The next big changes were the introduction of motorized freight transport and the telephone, both in the early part of the 20th century. The horse and wagon was replaced by the small urban truck. Moses and Williamson (1967) report that truck registrations in Chicago increased from 800 in 1910 to 23,000 in 1920 while, in the same period, horse-drawn vehicle registrations dropped from 58,000 to 31,000. They also estimate that both variable costs and travel time for the truck were less than half those for the horse and wagon. The telephone, unlike the telegraph, permitted easy point to point use within a city. The truck and the telephone allowed businesses to spread outward from the center and from each other, while still maintaining their links to the central port or railhead,

thereby taking advantage of lower land values and expanding the central business districts. In Chicago, firms that moved in 1920 located on average 59 percent further from the core than in 1908 — 1.46 as opposed to 0.92 miles.

At about the same time, automobiles improved the efficiency of personal transport, causing the areas between the streetcar suburbs to be settled and the residential apron to expand. However, automobile ownership was at first restricted mainly to richer families. As they acquired cars and suburbanized, relative house rents in the central cities must have fallen, benefiting the poorer residents. The automobile competed successfully with mass transit despite the transit fare 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 (Barrett, 1983).

The monocentric character of cities persisted well into the 20th century, because producers who located outside the core, thanks to the truck and telephone, were still bound to the central harbors and rail terminals. Although the automobile expanded the residential apron of the monocentric city, it reinforced the monocentric orientation of export industries, as improved labor access to the center and higher relative land values in the suburbs kept most export industries from suburbanizing.

Monocentricity persisted until the widespread use of the interurban truck, along with the interstate highway system and the establishment of suburban rail terminals. These developments came primarily after World War II, in the midst of massive suburbanization by the auto-owning population. They caused employment and production to leapfrog out to the farther suburbs in order to take advantage of cheaper suburban land and of proximity to suburban highway interchanges, rail terminals and suburban labor pools. Employment suburbanization drew manufacturing from the mostly multistory buildings of the central cities to the flat buildings and assembly plants built on cheap land near interstate highways. Central cities were transformed from manufacturing to service and office centers, even as office buildings and service activities also suburbanized.

Due to the durability of the urban capital stock and urban infrastructure, many 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, show strong evidence of origins tied to harbor and rail terminals, and development patterns tied to early radial mass transportation

systems. Chicago, the great metropolis of the midwest, was established as one of the last and westernmost of the waterway cities. It was already important by the beginning of the railroad era, so the railroads were brought through Chicago, which then made an extremely important rail hub. If, by historical accident, railroads had emerged before Chicago became well established, the great midwestern metropolis might have been located inland, perhaps at Springfield or Indianapolis (Cronon, 1991).

Further west, the spatial pattern of urban settlement 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 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 most outer reaches of large metropolitan areas, both old and new (Garreau, 1991). An edge city is characterized by very 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.[2] The automobile orientation is also reflected in the internal structure of edge cities. Large, campus-style office buildings are located singly or in small clusters, with arterial highways handling movement between clusters and often even between individual buildings. Edge cities take advantage of further cost reductions in telecommunications and transport, facilitating interaction with other parts of the urban area while retaining the advantages of cheap land and proximity to rural amenities.

[2]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% of commuting trips to the complex, and presumably an even higher proportion of other trips (Cervero and Wu, 1996, Table 5).

Cities in western Europe have evolved somewhat differently. Being much older, many still have centers which started out as medieval towns. There is a higher degree of mixture of residences and businesses in the core, possibly because of the rich cultural amenities there. Apartment buildings and subway systems are more common, partly for historical reasons and partly because government policy has favored compact development. Nevertheless, as in North American cities, there has been massive suburbanization and the emergence of edge cities.

B. Describing Urban Structure

To describe urban structure one must make use of basic data on land uses. Using such data, scholars have sought to describe the regularities and irregularities of urban structure. 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. At a more local level, activities may be *clustered* in a polycentric pattern or *dispersed* in a more regular pattern.

Abstract Statistical Approaches

Geographers have developed abstract methods, which facilitate realistic description but fall short of useful theorizing. We discuss two such methods here, then briefly describe economists' descriptive attempts to define and identify *subcenters*, those employment clusters outside the CBD.

One approach, called *point pattern analysis*, defines various statistics involving distances between observed units of development. 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 central place theory (Christaller, 1966) in which development is in centers organized hierarchically to maximize the market area of each. Another comparison distribution is that resulting from purely random location, which can be formulated as a Poisson process. This random pattern implies known probability distributions for such measures as the average distance from each point to its nearest neighbor (Thomas, 1981, p. 169).

For example, an observed average nearest-neighbor distance smaller than that for a random pattern indicates clustering, a possible definition of the existence of one or more centers or subcenters. Contrariwise, an average distance larger than random implies a tendency toward some regular spatial pattern, such as a uniform density or perhaps the hexagonal lattice pattern of central place theory. The trouble is, clustering and regularity may be present simultaneously, and either may occur in many varieties. To say more, we need additional statistics such as distance to second, third, and fourth nearest neighbors. The analysis quickly becomes complex and hard on intuition.

An example of the use of point pattern analysis is the search for population clusters in the Chicago area by Getis (1983). Getis first uses census-tract data from 1970 to approximately represent the residences of each 10,000 people by a single point in space. (He does not indicate exactly how this is achieved, so we do not know how much arbitrary judgment went into this phase of the analysis.) Getis then asks whether the resulting pattern of population could have arisen from overlapping areas of influence of a set of centers. To answer this question, he computes the average number of points $K(x)$ within distance x of any given point, for various values of x. Applying corrections for boundary effects, he demonstrates that at distances x up to 0.7 miles, $K(x)$ is smaller than would be expected under the Poisson process; whereas at greater distances it is larger than expected, with the largest deviation occurring at about 8 miles. The implication is that Chicago's population tends to be constrained to regular patterns or uniform densities at close distances but to be clustered when viewed at a scale of 8 miles. Such clustering is consistent with one or more employment centers exerting an attraction felt substantially at distances on the order of 8 miles.

Fractals

A more recent approach to describing urban spatial patterns is based on the idea that they resemble fractals. Mathematically, a fractal is the limiting result of a process of repeatedly replicating, at smaller and smaller 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. The elasticity of measured length with respect to resolution is known as the *fractal dimension*. So for example a coastline might have length L when measured on a map that

can resolve 100-meter features, and $Lx10^D$ 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. Batty and Longley (1994, pp. 174-179) 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, the deviation from 1.0 indicating the degree of irregularity. (By way of comparison, Britain's coastline has fractal dimension 1.25, Australia's 1.13.)[3] Surprisingly, they find that the fractal dimension of Cardiff's boundary declined slightly over the time period examined (1886 to 1922), a period of significant transport improvements (mainly 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 development with respect to the total number of zones into which the fixed urban area is divided.) This dimension 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 the number of developed zones grows as $\sqrt{N}$ as the total number N of zones is increased.

Batty and Longley (1994, Table 7.1) report estimated fractal dimensions for many cities around the world, with the result most often in the range 1.55 to 1.85. Paris in 1981 had a fractal

[3]Batty and Longley (1994), p. 167.

dimension estimated at 1.66. For Los Angeles in the same year, the estimated fractal dimension is 1.93, tied with Beijing for the highest among the 28 cities reported. This estimate implies that the fraction of area developed is almost constant at different scales, indicating a relative absence of fine-structure irregularities in development patters. 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 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 imposition of greater land-use controls or other forms of urban planning.

Unfortunately the estimated fractal dimension of a city is quite sensitive to just how the land-use data are summarized (Batty and Longley, p. 236). Of course similar problems afflict point pattern analysis (discussed above) and the estimation of urban density functions (which we describe in the next section). Another problem with the fractal approach is that 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 very different: fine structure may reflect local zoning rules or developers' detailed design strategies, while coarse structure may reflect regional planning, regional economic base, transportation facilities, large-scale geographical features, or land speculation based on anticipated regional growth.

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. More recent advances, especially random utility theory, do a better job of incorporating irregularities and noise into economic models. Thus there is hope that the powerful explanatory insights of economics can be exploited without sacrificing so much of the descriptive realism found in urban geography. Such approaches are examined in section IV.

Defining Subcenters

The methods discussed to this point lack any obvious connection to behavioral models explaining how city structure develops. In order to better accommodate such theorizing, urban economists have tended to use somewhat more concrete, if simplified, depictions of urban structure.

Most often these involve identifying one or more employment centers and estimating how these centers affect employment and population densities around them. Monocentric models have one employment center, polycentric more than one.

But how are such centers to be defined? 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.[4] Similarly, a lesson from the fractal approach 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.

In practice, much of the early literature on subcenters used criteria based on the 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 (McDonald, 1987).[5] Giuliano and 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 $\bar{D}$, and together containing total employment exceeding some

[4]The data are plotted on a square locational grid, with a spatial smoothing function used to compute the average density at each grid point from the raw data for nearby zones. If zone i 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. In the three plots shown in the figure, R takes values equal to $2\sqrt{2}$, $4\sqrt{2}$, and $6\sqrt{2}$ kilometers.

[5]Zonal definitions vary but are typically census tracts or similarly sized areas used for transportation planning. Usually some attempt is made to eliminate undevelopable land from the zonal definitions, but this is not always possible.

minimum $\bar{E}$. 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 IV, the exact pattern of centers so defined may be quite sensitive to the choice of cutoff values $\bar{D}$ and $\bar{E}$. 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.

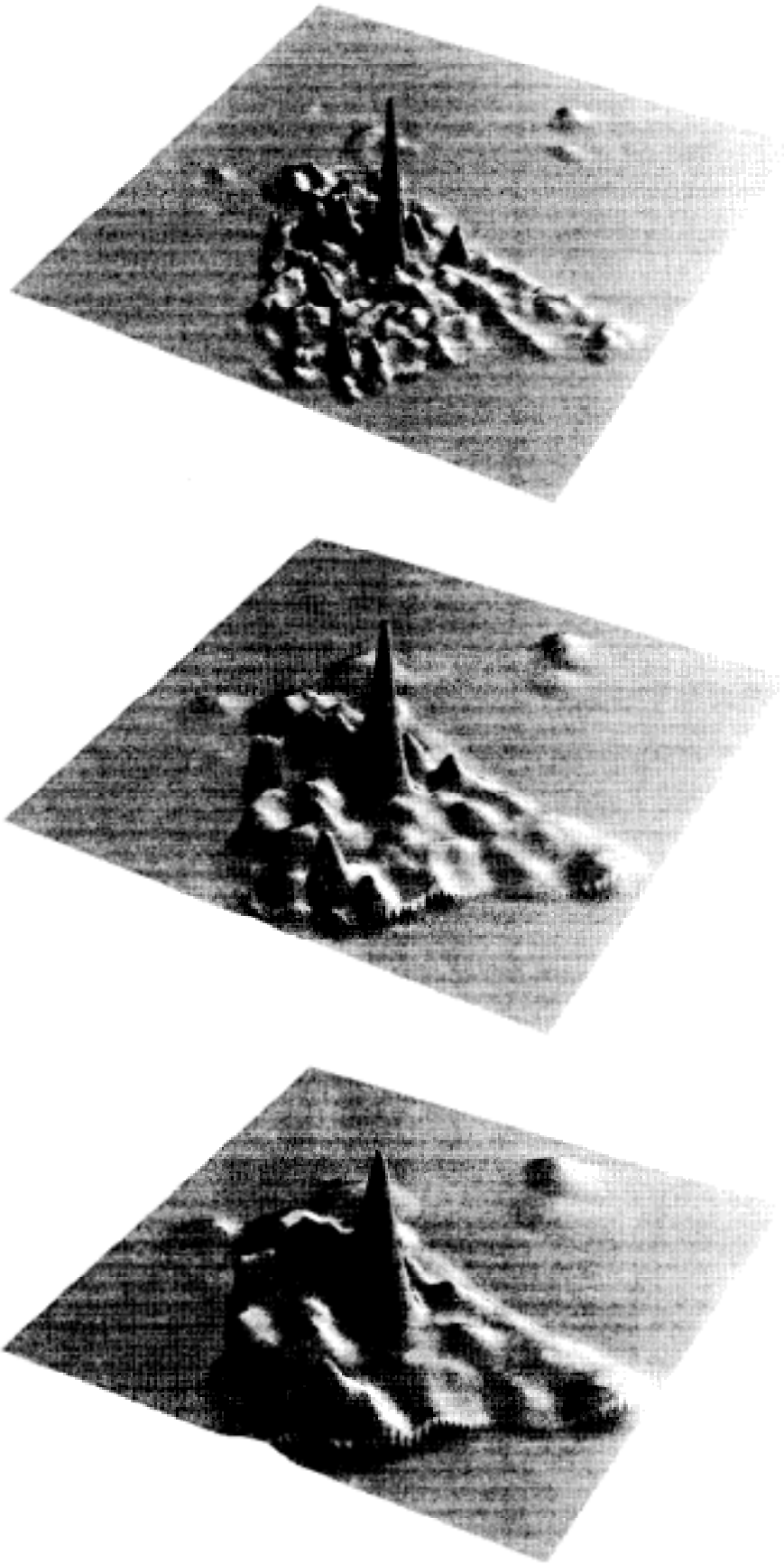

Figure 2. Employment Density, Los Angeles County, 1990, at Differing Resolutions
Source: Authors' plots of data from Southern California Association of Governments

III. The Monocentric City Model

The monocentric city model was formulated by Alonso (1964) as an adaptation of von Thünen's (1826) theory of agricultural land rent and land use to the urban case. It was almost immediately broadened to include production, transport, and housing and has been generalized in many ways since.[1] In this section we present the basic model and examine how it can be used to explain historic trends in the suburbanization of households.

A. The basic model

The city is envisaged as a circular residential area surrounding a central business district (CBD) 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. N identical households live at different distances from the CBD, each receiving utility $u(z,L)$ from a numeraire good z and a residential lot of size L. A household located x miles from the CBD incurs annual transport costs $T(x)$, normally interpreted as commuting cost to the CBD. Households each have exogenous income y which must cover expenditures on the numeraire good, land at unit price $r(x)$, and transport.

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

$$b(x,\bar{u}) = \max_{z,L} \frac{y - T(x) - z}{L} \; s.t. u(z,L) \geq \bar{u}. \qquad (1)$$

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

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

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

where $L(\bullet)$ is the solution to the maximization step 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 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 $\bar{u}$. Since all households are identical, the equilibrium rent function $r(x)$ coincides with one of these bid-rent functions. To determine which one, we can examine two conditions. 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 non-urban 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:

$$\int_0^{x^*} \frac{\phi(x)}{L[y - T(x),u]}dx = N, \tag{3}$$

where $\phi(x)dx$ is the land area[2] between x and $x+dx$. 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:

[2]For the simplest situation in which all urban land is used for residential purposes, the city is fully circular and $\phi(x)=2\pi x$.

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

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 Wheaton (1974). To illustrate their derivation, consider an increase in population, N. This causes no change in the family of bid-rent functions (1) or in the lot-size function $L(\bullet)$ corresponding to any given net income and utility. But from (3) the higher population does create excess demand for land. Equilibrium is reestablished with higher densities, lower utility, steeper bid-rent functions, and an expanded outer boundary.

Since the household can combine his residential lot with some of his other goods to produce housing, the above model treats housing implicitly. The extension to make this treatment explicit is straightforward. Brueckner (1987) provides a nice analysis of the resulting comparative statics. Land rent, housing rent, household density, and housing density all decline monotonically with distance from the CBD. A rise in income or a fall in marginal transport cost causes the household and housing density functions to flatten, whereas a rise in agricultural rent or in population causes them to steepen.

Land use in the simple monocentric model is efficient — that is, the equilibrium density pattern is Pareto optimal (Mirrlees, 1972; 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, of course, 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 are more jobs than resident workers. In this case the wage varies over location so as to offset differences in commuting costs (Brueckner, 1978; White, 1988). Because commuters still choose to travel radially inward to work, (2) applies and so do most results depending on the steepness of the rent and density functions.

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, transportation cost $T'(x)$ has to be reinterpreted to include the shadow value of the individual's time, which turns out to be its dominant component in modern developed nations. (Strictly speaking this would require adding leisure and a time budget to the model.) Because this shadow value rises with income, so does marginal transport cost $T'(x)$. If $T'(x)$ nevertheless rises more slowly with income than does lot size $L(\bullet)$, equation (2) predicts that rich households will have flatter bid rent functions than poor households and hence will locate more peripherally. Whether or not this condition holds for a typical U.S. city and therefore explains the observed pattern of higher-income groups locating more peripherally, on average, 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. We return to the durability question in the next subsection.

B. 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? If it applies to anything, it should help explain the broad population decentralization trends that have occurred in most cities of the world (Mills and Tan, 1980). To see how the model performs, we need to quantify the empirically observed trends and

provide some plausible parameters for the model.

Table 1.
Some Estimates of Population Density Gradients

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

Source: Clark (1968, pp. 349-351), converted from km to miles.

Pioneered by Clark (1951), researchers have estimated urban population density functions for an enormous range of places and times.[3] 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 linear in the logarithms of D and x, and is therefore easy to estimate. The constant $\gamma = -D'/D$ is the proportional rate at which population density falls with distance, known as the *density gradient*. It is a useful index of population dispersion.

[3]McDonald (1989) and Mills and Tan (1980) provide good surveys of methodology and results, respectively. Because of lack of data on land use at a fine scale, most of this work uses gross density, i.e. population divided by total land area, although the theory would be better represented by net density, i.e. population divided by residential land area. There seems to be no evidence that this affects the results systematically.

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

The standard explanation for decentralization among urban economists is rising incomes and declining transportation costs, both of which cause the density gradient to decline according to the monocentric model. However, the second part of this explanation is not entirely satisfactory because a large portion of transportation 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 specific assumptions that lead to the negative exponential population density function.[4] 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

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

Using empirically plausible values for the quantities on the right-hand side of (5), we can calculate the gradient and compare it with direct empirical estimates. Consider first the parameters appropriate for U.S. cities around 1970. Expenditure on urban housing was probably about 20% of

[4]See Papageorgiou and 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, chpt. 4). We instead provide conditions on the more primitive parameters of the model.

after-tax income net of commuting costs, and the ratio of land rent to housing rent also about 20% (Small, 1981, p. 320). This implies $1-\alpha=0.04$.

To "guesstimate" y, T and T', we will assume that each commuter has nine hours daily that can be allocated between commuting and work. Assume also that non-wage income is on average 10% of wage income and that all money income (wage plus non-wage) is taxed at an average tax rate of 25%. We will assume that the average one-way commute is 10 miles and takes place at an average speed of 25 miles per hour, thus requiring 48 minutes of round trip per day. The consensus of studies suggests that the value of travel time is about half the gross (before-tax) wage rate (Small, 1992, p.44). Money cost for a typical automobile commute (excluding insurance, parking and capital costs) is about half the time cost (Small, 1992, p. 84). So, total daily commuting cost is $T = (48/60)x(1/2)x(3/2)w = 0.6w$, where w is the hourly gross wage rate;[5] while marginal daily commuting cost T' (per mile of one way trip) is one-tenth as large. It also follows from the above assumptions that after-tax daily economic income is $y = (1-0.25)x(1.10)x[9-(48/60)]w = 6.765w$. Hence, $T/y = (0.6w/6.765w) = 0.0887$. This says that commuting cost is, on average, about 9% of after-tax economic income. Hence, $\gamma = (0.96/0.04)x(0.00887)/[1-0.0887] = 0.234$ per mile. 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 of parameters across decades are tenuous, but we can very roughly ask whether changes in incomes and transportation costs could account for the changes in γ between 1950 and 1970. Let us then presume that the expenditure share of land $1-\alpha$ remained at 0.04 throughout the period. LeRoy and Sonstelie (1983, Table 4) estimate that real income rose approximately 88% over those two decades, whereas real marginal transport costs (including the value of time) rose only 43 percent.[6]

---

[5]These assumptions imply that if wage income is a constant proportion of total net income, the income-elasticity of transportation cost is 0.6, well within the range of estimates of the income-elasticity of housing demand (hence of demand for lot size if housing is produced with Cobb-Douglas technology). This is why Wheaton (1977) argues that bid-rent curves for rich and poor are very similar in slope, casting doubt on the model's ability to explain location patterns by income.

[6]They give nominal figures, which we deflate by the Consumer Price index. We have estimated the mean by interpolating between their figures for the 25th and 75th percentiles.

Presumably this is because road improvements kept time costs from rising as fast as incomes, and money costs fell or at least did not rise in real terms. Then, the 1950 value of γ predicted by equation (5) is found by replacing the 1970 value of (t'/y) by $[(T'/1.43)/(y/1.88)]$, and similarly for T/y. The result is $\gamma = (0.96/0.04)x(1.315)x(0.00887)/[1-(1.315x0.0887)] = 0.317$. Hence, from 1950 to 1970, the gradient fell from 0.317 to 0.233, or by 26 percent. By comparison, Edmonston reported a 41 percent decline in density gradient for a sample of U.S. cities over that period. Again, the simple model appears to be in the right ball park.

However, there are some unsatisfactory aspects to the attempt to explain density gradients in this way. Mieszkowski and Mills (1993) give a cogent account. For one, attempts to explain differences in gradients across cities and across times have not been very successful at getting transportation costs to work; this may be because such costs are inaccurately measured and are strongly correlated with income. For another, 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. Finally, a strong inverse correlation is observed between the density gradient and total population, with larger cities more dispersed; whereas our model predicts a mild positive correlation.[7] 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.

Needless to say, more refined predictions could be made using available extensions of the simple monocentric model. For example, accounting for income differences would steepen the predicted density function if parameters are such that higher income people live more peripherally, since they also choose more land per dwelling for a given land rent (Anas and Kim, 1992). As another example, LeRoy and Sonstelie (1983) note that automobiles first were used by higher-income people, thereby flattening their bid-rent curves compared to those of poor people and encouraging high-income suburbanization; whereas by 1970 automobiles had diffused throughout the

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

income distribution so that the bid-rent curves of rich and poor became more similar in slope. In fact, they suggest that after 1970 the bid-rent functions of some of the rich became steeper that those of the poor, causing the gentrification movement of the 1970s in which upper-income groups moved into selected inner-city neighborhoods.

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. Harrison and Kain (1974) observed that cities tend to grow outwards 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. Dynamic versions of the monocentric model with durable housing have been constructed, leading to results that have conflicting implications for the value of density gradients compared to those predicted from the basic model. In spatial models with durable housing, the density gradient depends not only on the past time path of income and transport costs, but also on developers' expectations over time and the prospects for redevelopment. Explanations for observed density gradients are correspondingly complex.

Though data on the location of jobs are less readily available and less reliable than those on the location of population, employment density functions can be estimated in the same way as population density functions. 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 II.

Other possible explanations of decentralization, variants of a "flight from blight" hypothesis, were excluded by the assumptions in the basic monocentric model. *First* is deteriorating central housing quality, due to style or technological obsolescence combined with rational decisions by owners to maintain older housing at less than constant quality over time. *Second* is racial preferences combined with the tendency of poorer African-Americans to live in central cities. *Third* are negative neighborhood externalities associated with many poor neighborhoods. *Fourth* is the working out of Tiebout mechanisms for providing local public goods (Tiebout, 1956), resulting in poor cities being abandoned by better-off residents with a high demand for such goods and an incentive to use minimum lot-size zoning to exclude the poor. All these explanations imply that the poor live near downtown and the rich are pushed or pulled out to the suburbs. This raises suburban

land rents and lowers suburban lot sizes, thereby increasing suburban densities and causing the population density to fall off less rapidly with distance from the CBD than it would in the standard model.

There is one remaining set of explanations for the decline in the density gradient, which has received less attention than it probably merits: development restrictions. Development restrictions in central areas typically take the form of maximum density restrictions which may preclude redevelopment. Those in the suburbs restrict the amount of land zoned for residential development (e.g., green belts), which drives up the value of residential land thereby inducing construction at higher density. In many less developed nations, land use policy has had the effect of creating high-density squatter settlements on the outskirts of cities. These policies cause population density to fall off less rapidly with distance from the CBD than it otherwise would.

C. Assessment

Many researchers dismiss the monocentric model entirely as lacking realism, arguing that it describes the city of a bygone era. This view is somewhat extreme: the model provides key insight into the two most pervasive facts about urban structure: (1) that densities decline, albeit non-monotonically, with distance from the center, and (2) that most cities have been steadily decentralizing for a century or more.

But there is no disputing that the traditional CBD is becoming less important, and that employment in the modern city has a spatial pattern that is both dispersed and polycentric. Perhaps for these reasons, the model does not explain well certain other important facts, especially the tendency of larger cities to be more decentralized. Also the statistical fit of monocentric models to disaggregated data is rather poor and becoming poorer (Small and Song, 1994).

Our assessment is that the monocentric model has been an excellent conceptual tool for thinking about an urban economy, particularly about the role of commuting costs. It facilitates accounting for general-equilibrium effects and it appears to identify some powerful determinants of urban structure. But it provides no more than a useful starting point in explaining the observed spatial structure of modern cities.

IV. The Polycentric City

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. We begin with a description of empirical findings, then consider possible theoretical underpinnings for such a pattern. Throughout, we use "center" to mean either the main center or a subcenter.

A. Empirical Descriptions of Polycentric Forms

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 the United States, for which polycentricity has been examined in greater detail than anywhere else. Because many of the same forces are at work in other nations, especially those with highly developed economies, we expect that similar trends characterize them as well. However, some of these trends may be masked by the existence of older built-up areas and by stricter land-use controls.

(i) *Subcenters are prominent in both new and old cities*. Evidence is emerging that for large metropolitan areas in the United States, twenty or so subcenters can be identified at minimum gross density ($\bar{D}$) of 10 employees per acre and minimum total employment ($\bar{E}$) 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. McMillen and McDonald (1996a) find 15 subcenters outside the city limits of Chicago meeting an identical criterion, but modify it to cause two very large centers to divide into seven; Cervero and Wu (1997) find 22 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.[1]

[1]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

(ii) *The number of subcenters and their boundaries are quite sensitive to definition*. Both the Los Angeles and the Chicago studies mentioned above found that with changes in density cutoffs, certain employment clusters could be viewed either as several large subcenters or as one gigantic mega-center. In the 1990 Chicago data, for example, the criteria just listed produce a single subcenter surrounding O'Hare Airport, which incorporates around 16 percent of all suburban Chicago employment; 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 II. 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.

(iii) *Subcenters are sometimes arrayed in corridors*. In the 1980 Los Angeles data, the four largest centers and one smaller one are close together and arrayed in an arc extending from a few miles inland from the CBD all the way to the Pacific Ocean. This arc (which is most definitely not a beltway) more or less follows Wilshire Boulevard and includes the downtown area, Hollywood, Century City, Westwood, and Santa Monica. 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 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

(..continued)
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, Statistical Abstract of the United States, 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.

by manufacturing plants and connecting central Houston (starting just two miles from the CBD) to Galveston Bay (Craig et al., 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 was 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 (Heikkila et al., 1989). Each is based on a different assumption about how the occupant of a given land parcel interacts with multiple centers.

The first 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). Density D_m at location m then depends on distance r_{mn} to each center n according to a function such as:

$$D_m = \underset{n}{Max}\left\{A_n \exp(-b_n r_{mn})\right\}, \tag{6}$$

where A_n and b_n are coefficients to be estimated. We are not aware of any empirical support for this form, however, and it is rarely used in applied work.

An alternative assumption is that centers are complements. The occupant of a given location then requires access to every center in the area. Density might then be specified as the product of influences of the N centers, as follows:

$$D_m = A\prod_{n=1}^{N}\exp(-bnr_{mn}) \; . \tag{7}$$

This specification seems rather robust in practice, although it has a rather extreme property, namely that great distance from even one subcenter can prevent development entirely. A modification that substitutes b_n/r_{mn} for $-b_n r_{mn}$ in (7) overcomes this difficulty and seems to fit well.[2]

An intermediate case is the additive form, used by Gordon et al. (1986) and Small and Song (1994):

$$D_m = \sum_{n=1}^{N} A_n \exp(-b_n r_{mn}) \; . \tag{8}$$

Here every center has an influence, but unlike in (7) a center's influence becomes negligible at large distances.

Each of equations (6)-(8) contains the monocentric model as a special case. An advantage of (6) and (8) is that each center has its own magnitude and rate of decay of influence. On the other hand, (7) has the advantage of being linear in parameters after taking logarithms of both sides, whereas estimation of (8) by nonlinear least squares often results in convergence problems.

Another form with intermediate substitutability is defined by replacing the distances r_{mn} to specific centers n in (7) with distance to the nearest center, the second nearest center, and so forth. This could approximate the result of having several complementary types of centers, with centers of a given type being close substitutes. Sivitanidou (1996) uses this form successfully to explain Los Angeles office and commercial land values, although the form (7) fits about equally well.

Considerable success has been attained using these models to explain density and land-value patterns in Los Angeles and Chicago. The pioneering study was Gordon et al. (1986). A recent example is Small and Song (1994), who are able to explain roughly 50 to 75 percent of the variance in employment or population density across the entire Los Angeles region using (8) with 5 centers for 1970 and 8 centers for 1980. In all cases the special case of monocentricity is soundly rejected.

---

[2]McDonald and Prather (1994), McMillen and McDonald (1996a,b).

It is particularly important to note that 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 more poorly in 1980 than in 1970, reinforcing the belief that polycentricity is an increasingly prominent feature of the landscape.

(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 in suburban 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 (McDonald and Prather, 1994). 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 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[3]) 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.[4] Yet of the 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

[3]She is alleged to have said of Oakland that "there is no 'there' there."

[4]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 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.

and nearly ten times the size of the largest "edge city" in the region, known as South Coast Metro.[5] 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) test monocentric models of both employment and population density in Los Angeles assuming a variety of alternative center locations. The downtown center gives the best fit, although Los Angeles Airport comes close in the case of population. They also fit polycentric density models with five and eight centers, finding the downtown center to have by far the greatest influence (as measured by statistical significance) in the case of employment. For population, by contrast, Los Angeles Airport has slightly greater influence.[6]

(vi) *Most jobs are outside centers*. When all is said and done, centers account for less than half the employment in the areas studied: 47 percent in San Francisco, one-third in Los Angeles, and barely over one-fifth in the Chicago suburbs.[7] 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 tends disproportionately to spill outside previously defined centers and subcenters, but this has always been true: it does not tell us whether this newer growth continues to produce agglomerative forces that will result in the birth of yet more subcenters.

We do not know whether subcenters fill essential niches in the local economy that would lend them importance beyond the sheer numbers of people working or shopping there. Certainly there is

---

[5]This center, in Orange County, includes manufacturing and office complexes in parts of Irvine, Costa Mesa, Newport Beach, and Santa Ana. It borders John Wayne Airport and a large regional mall called South Coast Plaza.

[6]This is not due to its oceanside location, which was controlled for independently in one specification.

[7]Unfortunately certain data sources are incompatible between the City of Chicago and its suburbs (i.e. the rest of the CMSA). As a result some studies have used only one or the other, making us unable to make statements for the entire CMSA.

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 (1996a) find that different centers have quite different industry-mix characteristics, with some centers quite 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.[8] Further empirical research on the economic roles that subcenters play would appear to us to have a great payoff.

(vii) *Commuting is not well explained by standard urban models, either monocentric or polycentric*. Hamilton (1982) was the first to note that the standard assumption of people commuting up a land-price gradient cannot come close to explaining actual commuting patterns in the United States or Japan. Given the actual degree of dispersion of jobs and residences, a monocentric model produces commutes of just a mile or so, understating actual 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, people still incur far longer commutes, both in time and distance, than they would if they were minimizing the sum of housing rents and commuting costs, holding lot size constant (Hamilton, 1982; Small and Song, 1992).[9] Yet that is what they must do under the standard model of urban economics reviewed in Section III, with a deterministic utility function depending solely on a numeraire good and housing.

It appears that at least in auto-dominated cities, there is more "cross-commuting," in which

[8]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 Rosen and Resnick (1980) for a thorough empirical investigation. See Krugman (1996) for a thoughtful discussion of possible reasons for this amazingly robust empirical relationship.

[9]A counter-example to the prevalence of this so-called "excess commuting" appears to be Tokyo, where much commuting is by public transit and so average commuting times are much longer than in the U.S. Tokyo has more than twice the total employment of Los Angeles, and average employment density is somewhat larger. The density gradients for employment are about the same, but for residences Tokyo's is much higher, consistent with the notion that people there place more value on proximity to work places. Tokyo's population is also more homogenous, possibly removing a barrier to short work trips in racially and economically diverse U.S. cities. See Merriman et al. (1995).

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 three — Small and Song's estimate of actual relative to predicted commuting for Los Angeles — 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 residences, due to local amenities or to practical or sentimental attachments formed from years of living there. Two-worker households have to compromise between locations convenient to each job. Frequent job changes and substantial residential moving costs cause people to choose locations convenient to an expected array of possible future jobs rather than just their current job (Crane, 1996). Racial and income segregation constrain housing choices. All these explanations require the existence of 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 opportunities is extremely important in explaining urban residential location decisions. Fortunately, researchers have made considerable headway in adding heterogeneity to urban models, and the results suggest that heterogeneity affects the resulting structure and not just individual decisions. For example, at a very abstract level, adding heterogeneity to a standard monocentric model results in greater dispersion (Anas, 1990). Heterogeneity in zonal-based empirical models is naturally represented through a discrete-choice formulation, such as logit, of the various decisions that economic actors make about location, land development, and redevelopment (Anas, 1986).

B. Theories of Agglomeration and Polycentricity

Why do employment concentrations within cities exhibit the complex shapes identified above? Explanations center on *agglomeration economies*. These are pervasive scale economies, many of them external to firms and households, which manifest themselves through spatial proximity. There are many types of agglomeration economies operating at various levels of spatial

resolution, including the interurban scale. The compounded effects of these economies generate complex spatial patterns such as those of Figure 1.

Agglomeration economies are also believed to cause cities to exist in the first place. At the scale of major economic regions, cities are linked by traded goods and by factor mobility, and geographers such as Christaller (1933) and Losch (1940) have sought to explain the spatial and size distribution of cities on the basis of such trade interactions.

At the urban scale, factor mobility is much greater and interactions are more spatially intensive. Firms interact with suppliers (backward linkages), customers (forward linkages), and each other (sideways linkages). For example, Schwartz (1992) has shown how companies located throughout the large metropolitan areas of New York, Los Angeles, and Chicago purchase business services predominantly from firms located in the respective central cities.

Linkages cause external economies between firms within or across given industries. The resulting economies are called *economies of localization* in the former case and *economies of urbanization* in the latter. The former are established empirically by using industry size to explain the productivity of firms in that industry, as for example in Henderson (1986a, 1988) and Moomaw (1988). One expects localization economies to produce specialized cities, of which abundant evidence also exists (Henderson, 1988). Economies of urbanization, which produce diversified cities, are more difficult to isolate but several studies have found evidence of them.[10] There is some evidence that urbanization economies contribute to economic growth through the encouragement and diffusion of innovations (Jacobs, 1984; O h'Uallacháin, 1989; Glaeser et al., 1992).

Specialization à la Adam Smith is another important agglomeration economy which operates at the scale of an entire urban area. The specialization of firms, combined with change and uncertainty such as that caused by a business cycle, create what has been called "economies of massed reserves" (Robinson, 1958), by which larger concentrations of specialized jobs, labor or equipment make it less likely that a household or a firm will be unable to fulfill an unexpected need. Hence, for example, urban areas function as unified markets which facilitate idiosyncratic matching of firms and workers, or of firms and customers. Agglomeration at the urban scale also derives from the fact that human interaction at close proximity fosters new ideas and creative insights and probably encourages formal education and training. Greater education may in turn result in more experimentation, more innovation, more rapid diffusion of innovation, greater adaptability, and improved management

[10]Sveikauskas (1975), Moomaw (1988), Ciccone and Hall (1996).

skills.

Agglomeration economies create a centripetal tendency in cities, causing agents to cluster in either large or small groups to facilitate interaction and save costs. There are many centripetal mechanisms other than those already mentioned. For example, people cluster to enjoy the human environment of cities as a public good, to lower the cost of supplying local public goods, or to economize on search and trading costs. Retail trade concentrates to facilitate shopping when consumers have imperfect information about the products of different firms.

Centripetal forces of agglomeration are balanced by centrifugal tendencies which limit the extent of spatial clustering. The most fundamental centrifugal tendency comes from the limitation of geography: land at any location is in limited supply. Other centrifugal tendencies are created by congestion, by disamenities associated with urban activities such as pollution and by idiosyncratic preferences for different locations.

Below, we discuss how these forces can be modeled explicitly and how they result in the formation of urban centers and subcenters.

1. Spatial Contact Models: Monocenters with Dispersed Agents

Models of spatial contact generate a peaking of rent and land use density, just like the monocentric model of the previous section, but without imposing a prespecified employment site. Rather, central peaking emerges solely from the interdependence of economic activity, via forward, backward, or sideways linkages.

Consider first a very basic framework, as in Solow and Vickrey (1971). Geography is described as a finite space, such as a line segment or disc, with a geometric center but no predetermined economic center. Now consider homogeneous agents who must interact through sideways linkages by traveling to one another's locations each day. Define the *accessibility* of a location x as the inverse of the mean travel cost for someone located there, in which the cost of contacting each other location from x is weighted by the relative frequency of such contacts. Each agent maximizes utility which depends on goods and on residential lot size. In equilibrium the utilities of all agents are equalized. Because the geometric center is the most accessible point, rents and densities peak there, declining monotonically and symmetrically toward the edges of the space. If lot size is responsive to price, this means that density also declines monotonically from the center. (If there were no

geometric center, as for example in models confined to the perimeter of a circle, the symmetric equilibrium would instead be a uniform distribution of densities and land rent.)

This simple model generates a monocentric residential pattern, yet it departs from the standard monocentric city in a very important way: the equilibrium is not optimal. This is because the interdependence among agents creates an externality, as noted by Borukhov and Hochman (1977). If agent A chooses a more accessible location, an external benefit is imparted by reducing the average contact cost of the other agents. This is in addition to agent A's own cost reduction, which is internalized. Since agent A does not value the benefit conferred on others, A will choose a less central location than is socially optimal and the equilibrium city will be too dispersed.

A second externality may operate at the margin of city population once we account for the reason agents contact each other. Homogeneous agents can be given an explicit benefit from interaction by endowing them with a taste for variety in interaction. Then, adding a new agent causes each existing agent to want to interact with that new agent. If that creates a benefit to the existing agents that is not somehow captured by the new agent through the price system, there is insufficient incentive for new agents to join the city and the equilibrium population is too small.

The motivations for interaction become more compelling when we consider two or more types of forward- or backward-linked agents. Doing so also allows us to investigate how different groups interact in land markets to determine location patterns. For example, firms might outbid households everywhere within a central area, thereby endogenously generating the monocentric city model; or firms and households might both locate in a dispersed pattern but with production more centralized than housing, thereby generating a simple extension of the monocentric model mentioned in the previous section.

Many models of this type use a one-dimensional geography for simplicity, following the lead of the Solow-Vickrey model mentioned earlier. Although unsuitable for realistic simulation, such models allow for many of the same patterns of mixed or separated land uses as two-dimensional models and hence provide most of the same insights. Two examples — Fujita (1988) and Anas and Kim (1996) — nicely illustrate the way different patterns are generated depending on parameter values representing such variables as transportation costs and taste for variety. Both use a straight line segment, which in Fujita is continuous and in Anas and Kim is discrete. In both, households visit a firm (retailer) in order to purchase its unique brand of good. In Fujita's partial equilibrium model, firms are monopolistically competitive and free entry leads to a spatial Chamberlinian equilibrium.

Firms and consumers occupy one unit of land each. Consumers have a taste for variety of brands, as in Dixit and Stiglitz (1977), hence travel to each firm, purchasing a fixed amount on each trip. Since consumers are never satiated by variety, what limits the number of brands is the fixed cost of a firm's entry into the market. In Anas and Kim's fully closed general equilibrium model, consumers are also workers of the firms so their location decisions are influenced by both commuting and shopping travel, leading to an equilibrium in which rent, wage, and retail price are all functions of location.

These models can produce a variety of patterns, depending on parameter values. A fully separated equilibrium like that obtained in the monocentric model is one possibility, but it occurs only with parameter values that appear unrealistic. With realistic parameter values, a partially separated equilibrium obtains, with mixed production and housing towards the center and just housing towards the periphery. Another possibility is a fully integrated equilibrium in which production and housing are mixed throughout the city. In both models, land rents are highest at the geometric center. Typically wages, when explicitly modeled, also peak at the geometric center; but it is possible for this to be reversed: in Anas and Kim's model land rent can fall so rapidly with distance from the center that peripheral firms substitute sharply away from labor, causing labor's marginal product to be higher at those peripheral locations. It is also possible in their model, if production is highly land intensive, for firms to be more dispersed than residences, a pattern we might think of as explaining the location of suburban shopping centers.

2. Endogenous subcenters: agglomeration and polycentricity

Early polycentric models such as by White (1976) treated the location of production centers as exogenous, providing conditions under which a firm would choose to locate in a secondary center in order to take advantage of lower land rents and cheaper labor — cheaper because the firm can attract workers who otherwise have to incur large costs to commute to the CBD. As more firms locate at the subcenter, the wage they must offer rises and the subcenter's labor area grows. All residents within the subcenter's labor area commute to the subcenter, and land rent within the labor area declines as a function of distance from the subcenter. All those outside the subcenter's labor area commute to the CBD.

In this section, we consider models that take the further step of explaining both the location and the size of subcenters. In order to generate endogenous clustering of economic activity, we need

to consider centripetal forces which are stronger than those which operate in the spatial dispersion models. The literature has demonstrated a variety of ways in which such strong centripetal forces can arise.

Export orientation

Consider first a national economy on a featureless space, based on primitive agriculture or home-based manufactures. Production requires inputs of land and labor and is constant returns to scale. Assume that the economy is self-sufficient: it neither imports nor exports. In such a world, all production occurs in the backyard of each consumer. There is no transport. As long as consumers do not interact socially, land use densities are everywhere uniform.

Now, open this economy to trade: consumers import goods produced in other regions and pay for these by exporting their backyard crafts. This gives rise to transport, and to terminals which take advantage of scale economies in loading and unloading. Mills (1972) formalized the argument — advanced earlier by Moses and Williamson (1967) -- that, in such an economy, urban structure emerges as the concentration of export goods production around the terminals, provided that the intra-urban cost of moving export goods is substantially higher than the cost of moving people. Commodities for which this relationship holds are produced in factories clumped around terminals. Workers employed in the terminal and in the factories are spread out and commute to them. Non-traded goods continue to be produced in backyards. These relationships of relative transport costs are thought to have been the causes behind the core-dominated nineteenth century style city. Note that in such cities, the size of the manufacturing core and, hence, of the city would be determined primarily by the efficient scale in terminal operations.

Scale economies in production

Instead of trade, suppose our backyard economy becomes subject to increasing returns to scale in production. Provided that the degrees of such returns to scale are sufficiently high relative to the cost of transporting people and goods, it is now more efficient to concentrate production in a discrete number of regions in space, which emerge as centers or subcenters. This is because the lower production costs from having larger and fewer plants more than offsets the higher costs of goods distribution and commuting. The greater the degree of returns to scale and the smaller the cost of transport, the fewer centers will be optimal.

The argument was formalized independently by Serck-Hanssen (1969) and Starrett (1974). Starrett showed that there is an optimum (cost-minimizing) scale at which firms should operate, and at this optimal scale the value of production times the local degree of increasing returns to scale equals total differential land rent (land rent in excess of agricultural land rent). This, in turn, determines the number and spacing of identical centers over a homogeneous space.

Forward and backward linkages

A different insight for spatial agglomeration comes from trade theory which has long emphasized forward linkages between a firm and its customers, and backward linkages between a firm and its suppliers (including workers). Krugman (1993) develops such a model to explain the location of an urban concentration in a rural hinterland. But with only slight modifications, his model could also explain why production of some goods within a city will be concentrated in space, while other urban goods will be produced in a dispersed manner.

Krugman's is an unusual model of location without land and, hence, without rents. The immobility of land is proxied by assuming that peasant farmers are uniformly distributed and immobile. Food is produced by peasants under constant returns to scale and is transported freely to urban areas. Manufactures, on the other hand, are differentiated and are produced by mobile urban labor with a fixed amount of labor needed to start production. Manufacturers are Chamberlinian monopolistic competitors, in the manner of Dixit and Stiglitz (1977). Both farmers and urban laborers have a taste for variety and consume the product of each manufacturer. Under these conditions, the forward linkages are to a manufacturer's urban and rural customers, while the backward linkages are to urban laborers (who are also the urban customers).

To economize on the costs of delivering to customers, firms and their laborers clump together to form cities or — in the intraurban case — manufacturing centers. How many such centers emerge is determined by the level of unit transport costs. Under higher transport costs, there are more centers, and under lower transport costs, there is just one center.

The basic insight of Krugman's model is that when the laborers of an industry are also its customers, lower transport costs from the co-location of firms confers an external scale economy among firms. This is an example of a principle now well understood from international trade theory: with monopolistic competition, a pecuniary externality creates real scale economies.

Scale Economies in Retailing

We saw earlier that in the model of Anas and Kim (1996), congestion and taste for variety creates a centrifugal force generating dispersion of retail activity. But suppose we now postulate that the number of shopping trips made to location k by a consumer residing at i attenuates with the full price of a trip and is also directly influenced by total retailing output at k, expressing the convenience of shopping at large shopping centers. Then retail stores have an incentive to cluster into subcenters, a tendency balanced by the centrifugal forces. Anas and Kim show that both monocentric and polycentric equilibria can exist under the same parameter values, indicating that history determines which pattern occurs in long-run equilibrium.

The monocentric equilibrium has the highest welfare ranking and the widest margin of stability when the pure preference for larger shopping centers is sufficiently high relative to a traffic congestion parameter. As the level of congestion increases, the stability and welfare position of the monocentric pattern deteriorate, causing land use patterns with subcenters to become more stable and to acquire higher welfare rankings. Eventually, with a sufficiently high congestion level, retail centers are completely mixed with residences in order to maximize firms' access to customers and labor.

Pure externalities

Some authors have treated agglomeration as a pure nonpecuniary external effect in production or consumption, without specifying exactly what causes the externality. Such models have a general appeal, because they are broadly consistent with many different specifications of the external effects. The external effects are assumed to confer scale economies by lowering production costs or by influencing consumer demands.

A good example is the model of Fujita and Ogawa (1982). Firms benefit from other firms near them by means of a "locational potential" function; this function is meant to capture informational spillovers (sideways communication externalities) among firms, but in fact it can represent any external benefit of one firm on its neighbors. Firms are distributed over a continuous linear space, and the positive externality conferred by a firm at y on a firm at x is postulated to attenuate with the distance between the two firms. Thus the productivity of every firm depends on its distance from all other firms.

When the model is simulated, Fujita and Ogawa find that for a given strength of the spillover

externality, a sufficiently high unit transport cost is needed to maintain an interspersed pattern. Lowering unit transport costs from such a starting point causes various multinucleated patterns to emerge. These include patterns in which there are two, three, or more exclusive business districts and patterns in which there are exclusively residential and business areas coexisting with mixed areas (quasi-CBDs). There are multiple equilibria under the same parameter values and as population grows, transitions from one equilibrium to another follow catastrophic paths.

3. Stability, Growth, and Dynamics

Although we have described the models in this section as generating static equilibria, the same mechanisms can be used to study the stability of equilibria and dynamic spatial patterns. Due to the multiplicity of equilibria and the catastrophic nature of the comparative statics, it is not surprising that such models may result in periods of instability and rapid change, and in history-dependent steady states.

One of the simplest examples is the two-location model of Anas (1992). Each location is a potential center, containing a fixed amount of land. Individuals maximize a utility function which depends on per-capita output and per-capita land consumption. Localization economies cause per-capita output to rise with the number of people n_i at location i, but per-capita land consumption varies inversely with n_i. Writing the resulting utility as $V(n_i)$, assume functional forms are such that there is some value n^* for n_i which maximizes $V(\bullet)$.

Equilibrium is characterized by $V(n_1) = V(N-n_2)$. There are either three or five equilibria depending on whether total population N is less or greater than $2n^*$. A symmetric equilibrium with two equal-size centers always exists, because then every agent is satisfied with his choice; but this equilibrium is locally unstable if $N<2n^*$ because then a small fluctuation in size gives the localization advantage to the larger city, causing it to grow still larger. Two stable monocentric equilibria, with everyone concentrated at one location or the other, also exist provided that atomistic defection is sufficiently discouraging.

What makes this model especially interesting is the presence of two asymmetric equilibria that occur if $N>2n^*$. In these equilibria, one center is too large and the other two small, relative to the utility-maximizing size n^*. These asymmetric equilibria may be thought of as polycentric patterns with a large center and a small subcenter. They are unstable because any fluctuation enhances the

attractiveness of the smaller center and reduces that of the larger one.

However, the polycentric pattern, even though unstable in this case, plays a key role in the stability analysis of the monocentric or symmetric duocentric equilibria. With $N>n^*$, the symmetric duo-centric equilibrium is always better than the monocentric one; yet the monocentric equilibrium is stable against any fluctuation up to size n^s, where n^s is the size of the smaller of the two centers in the asymmetric equilibrium. This is because if any clump of population smaller than n^s leaves the monocenter to establish a new subcenter, the latter will still be less attractive than the monocenter so people will tend to migrate back. Only if it is of size at least n^s is the new subcenter viable, and then it will tend to grow until it attracts half the population. The size n_s of a viable subcenter becomes smaller the larger the total population, because a single large monocenter is so overcrowded that even a small subcenter becomes a viable competitor.

The dynamics of the system follow from these same observations. Suppose in each time period there are random migrations from one location to the other occur, but with probability proportional to the utility differential offered by the other location. When total population is small, there will be just one center. As population grows, the one center remains but becomes stable against smaller and smaller fluctuations. Eventually a fluctuation produces a viable subcenter, which then grows rapidly until there are two equal-size centers.

These fluctuations are not unlike the process of "edge city" formation envisioned by Henderson and Mitra (1996), for whom the "individuals" are firms and the possible locations are not fixed but are constrained by the existence of a fixed distribution of residences around an existing monocenter. Most important, Henderson and Mitra provide an agent, called a developer, to help the migration process along. They examine carefully the strategic considerations facing the developer, finding a rich set of possible decisions about where and how large an edge city to build.

4. Non-Economic Dynamic Models

The existence of multiple centers, the irregularity of spatial forms, and the unpredictability of how they evolve are challenges forced by observations of modern urban structure. 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 can come from these 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 (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 consequences of such feedback mechanisms rather than their economic underpinnings.

These models typically explore systems that are out of equilibrium, an approach now well established in evolutionary economics (Nelson, 1995) and 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. What follows is a sampler from a quite eclectic literature centered mostly in geography.

Markovian Transitions

One approach is to model probabilistic transitions of micro units from one state to another. Examples include the development or redevelopment of a parcel of land, a household migration decision, and the birth or death of a firm. Agglomeration effects imply that individual transition probabilities depend on the number of actors in each state, making this an example of an interactive Markov chain (Conlisk, 1992). From the individual transition probabilities, one can derive a "master equation" which describes the evolution of the probability distribution function giving the likelihood of each possible combination of micro states (Fischer et al., 1990).

In some cases the system evolves toward one or more stationary states in which macro variables are time-invariant. Conlisk (1992) provides some general conditions. If the transition probabilities are exponential in utility differences, for example, those states are described by a multinomial logit model. Such a formulation is therefore a natural generalization of the discrete-choice approach to modeling dispersion discussed earlier. But the current formulation is richer because it describes dynamics. Thus we can now describe how starting conditions, the particular realization of stochastic variables, and other details of the dynamics determine which stationary state is achieved and what happens along the way.

A model whose macro features depend on the particular realization of stochastic variables is a model in which history matters, just as recent work has shown that it matters in other fields of economics (David, 1985; Arthur, 1989). The regional shopping center could have succeeded in any

of several locations, and perhaps only the perspicuity of one individual made the difference that ultimately fixed the location of the next edge city. This in turn may determine whether a steady state with few or with many centers is reached.

Self-Organization

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 these dynamic theories. In some circumstances they result in sudden shifts from one relatively stable state to another, a phenomenon resembling punctuated equilibria in biological evolution (Eldredge and Gould, 1972). But only certain stationary states are consistent with the underlying dynamic system, and it is the task of self-organization theory to describe these possibilities as functions of general properties of the system.

Krugman (1996) attempts to do this by using Fourier analysis to decompose a random fluctuation (such as the employment pattern resulting from building a large plant on a particular site) into an infinite series of regularly spaced fluctuations (such as the pattern from the simultaneous startup of many small firms along a regularly spaced lattice), the infinite series consisting of 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 spatial fluctuations to be magnified (as with a further influx of new firms in the same pattern) and others to be suppressed (as with the closing of unsuccessful firms due to unfavorable location patterns vis-a-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 can understand the underlying reasons for these regularities.

This kind of analysis provides insight into the effect of 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 inter-regional or international trade, yielding size hierarchies of cities at a national, continental, and recently even a global scale.

Diffusion and Percolation

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 the inmigration of firms to an area consisting of many small land parcels, randomly occupied, when each firm requires several contiguous vacant parcels. 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 (Bunde and Havlin, 1996).

Fotheringham et al. (1989) propose that in a somewhat analogous way, discrete clumps of development arrive randomly at the edge of a metropolitan area and seek suitable vacant sites. Agglomeration is posited by requiring that a new clump may settle only on the edge of an existing cluster of development. The resulting patterns of developed land are similar to the pathways by which water percolates through rock or electrons flow through partially conductive materials. Such pathways are well known to be fractals, and Batty et al. (1994) use this model to simulate the fractal patterns which, as noted in Section II, they believe characterize urban development.

Makse et al. (1995) propose a model with somewhat stronger agglomeration tendencies. Known as correlated percolation, the model postulates a development probability for a give site which increases with the proximity of other occupied sites, but otherwise declines with distance from an exogenous monocenter. Simulations yield growth patterns that resemble, at least impressionistically, the historical development of Berlin from 1875 to 1945, which especially in the later years showed the high degree of irregularity typical of large modern cities. It is difficult to see clearcut centers in the visual displays. However, a statistical analysis of the sizes of interconnected regions confirms that they follow a power law that plausibly approximates that of real municipalities in the Berlin region.

Self-Organized Criticality

Per Bak and several colleagues have shown that many 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 (Bak and Chen, 1991). This can happen when the system is subject to bifurcations due to parameter boundaries that determine qualitatively different states, and when those

parameters are themselves endogenous. Small fluctuations then cause chain reactions whose sizes typically obey a power-law distribution. Krugman (1996) hints that the interactions among economic agents may produce such a condition 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. However, no explicit mechanism has been developed, nor has this type of explanation been integrated with existing models that produce sudden growth, as in Anas (1988) and Krugman (1991a).

Logistic Growth

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 was also the case in Anas and Kim, 1996). Such models are capable of generating bifurcations, in which small shifts of parameter values produce qualitatively different equilibrium configurations, some stable and some not (Harris and Wilson, 1978).

Peter Allen and collaborators from the Free University of Brussels have put some of the same ideas into purely dynamic models intended to describe growth processes that may be far from equilibrium. These 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 Sanglier (1981a) employment S in a given region and sector obeys a dynamic equation in which dS/dt is proportional to $S \bullet (E-S)$, where E is a measure of potential employment demand. This potential demand is in turn determined by other equations in the system that include the location's relative attractiveness, crowding, and a rather arbitrary "natural carrying capacity." Thus existing employment attracts new employment, but eventually 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 including Bastogne, Belgium (Allen and Sanglier, 1981b) and Rouen, France (Pumain et al., 1987). A version was even built for the U.S. Department of Transportation.

Toward Convergence with Economic Models

Most of the noneconomic models described here lack a price system and any explicit description of rational economic decision-making. Furthermore, behavior is typically myopic. Thus, for all their tantalizing ability to portray complexity in the dynamics of urban structure, they fail to incorporate many insights from urban economic models.

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 into existing non-economic models, or attractive features of those models might be added to existing models within urban economics.

An example of the first approach is 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 postulates, 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.

V. The Welfare Economics of Urban Structure

In defense of the sprawling, low-density development which increasingly characterizes modern cities, Gordon and Richardson (1986, 1996) argue that the urban spatial structure generated by market forces reflects the will of the people. Planners, in contrast, typically have little faith in either the efficiency or equity of market-determined urban spatial structure, and advocate detailed land use planning. To evaluate these conflicting points of view we need to explore the welfare economics of urban land use. We begin within the context of the basic monocentric model, then consider the implications of agglomeration economies and polycentricity.

A. Excessive Suburbanization in the Monocentric City

Urban spatial structure in the most basic monocentric-city model is efficient, as noted earlier. It is reassuring that the Invisible Hand can work with respect to the location of economic activities. Unfortunately, this efficiency property is of questionable practical relevance because of the pervasiveness of externalities in actual cities. Here we focus on one that is particularly important and most extensively studied within the monocentric framework: 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 mode, route, and time, and overall travel may be excessive also. 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, congestion tolls are charged almost nowhere and as a result 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 structure? The most severe congestion continues to occur, even in today's complex urban structures, on radial travel to and from the central business district (CBD). Hence it is here that underpricing is most severe. If urban structure is fundamentally shaped by marginal commuting costs to the CBD, as postulated by the monocentric model, then such underpricing causes rent and density functions to be flatter and the city to extend to a larger radius.

This holds even relative to the second-best optimum, i.e., the optimum conditioned on an absence of congestion pricing, because at the margin people contemplating a close-in residence are not willing to pay as much extra rent for it as the social cost savings that would be realized if they reduced their commute.

This excessive residential decentralization is compounded by a less obvious effect, working through the land market. Underpricing travel distorts land values in a way that encourages planners to allocate too much land to roads. To see why, suppose the only cost associated with a road is the opportunity cost of the land it uses. Now let the planner employ the following "naive" cost-benefit rule: at each location, expand the road by using more land until the incremental travel cost saving from further expansion equals the incremental market value of land in residential use. This rule is a fair characterization of current practice: while cost-benefit analysis is often undertaken for road projects, it typically accepts market land prices as valid when computing costs. With unpriced congestion, the market value of residential land at central locations is less than its shadow value as just explained. The naive rule therefore uses too low a land price to trade off against travel-cost savings, and its application results in too much central land being devoted to roads. Wheaton (1978) has argued that the failure of cost-benefit practice to take into account the underpricing of urban auto travel resulted in massive overbuilding of urban highways in the U.S., especially in the 1950's and 1960's.

What then is the appropriate role of government with respect to urban spatial structure, from the perspective of the monocentric model? If automobile travel cannot be priced efficiently then government intervention may be warranted to correct the resulting excessive decentralization. Possible policies include second-best 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 Downs (1992) and others. Subsidies for home ownership, subsidized highway construction and maintenance, fragmentation of local government, and minimum-lot-size zoning are just some of the powerful forces by which government intervention tends to cause more rather than less dispersion in U.S. metropolitan areas. We do not mean to imply, however, that such government policies are the main reason for ongoing decentralization — the phenomenon is far more universal than any particular set of policies.

While government intervention can be beneficial, excessive or inappropriate intervention can be harmful. For example, planners are fond of using land-use controls to combat urban sprawl. But with durable housing it may be efficient for development to "leap-frog" over vacant land in order to leave that land free for later development at higher density than is economically justified today. Many planners also advocate policies, such as building mass transit facilities or downtown convention centers, to reverse the excessive decentralization that has resulted from underpricing urban auto travel. But because the pricing errors of the past have been cast in brick and asphalt, such policies may compound the damage by creating still more inefficiencies.

B. Economies of Agglomeration and Welfare

We have seen that although agglomeration economies are the *raison d'être* 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 soft, 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 some others operate through the market, while 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.

Given such likely causes of agglomeration economies, does "the market" — broadly speaking — deal efficiently with them? The standard answer is negative. If the agglomeration economies are internalized, then efficient pricing cannot be supported by perfect competition. If they remain external, firms will underemploy those business practices that contribute social value to their neighbors.

The standard argument neglects, however, the possibility that efficiency could be achieved by private city-developers who would set up optimally-sized cities, thereby internalizing the agglomeration economies, and who compete with other such developers in a regional or national market. Each optimally-sized city would operate at 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 lots (because of transport costs). Under marginal-cost pricing the losses from goods production are just offset by the profits from the production of lots, which are manifested as land rents. (This is a variant of the Henry George Theorem.) To a limited extent the developers of edge cities are playing this role. We do not, however, observe developers trading cities in a competitive market, and we suspect that the assumptions of the implicit model on which the above argument is based are significantly unrealistic in some respect. No one, however, has provided a persuasive alternative model.

Of course, cities have always been full of very localized externalities, from the smells of household waste to the blockage of ocean views by neighbors' apartments. In principle, land use controls may be justified to deal with such cases. Just how important these spillovers are empirically is subject to some debate, with Mills and Hamilton (1994, pp. 252-254), for example, arguing that they are quantitatively small. The city of Houston, one of very few in the U.S. to lack explicit zoning laws, affords a chance for some interesting empirical studies. We do not take a position on this question except to note that such "neighborhood externalities," resulting from the close interactions among urban denizens, are not minor aberrations but are inherent in the nature of cities.

C. Welfare Economics and Polycentric Structures

We have seen how agglomeration economies tend to create clusters of economic activity, which in turn influence surrounding residential densities. Within an urban area, such clusters may play roles similar to the regional hierarchy of cities derived in the central place theory of Christaller (1933) and Losch (1940). But given the rich nature of interactions within urban areas, they play many other roles as well. What can we say about the optimality of the resulting urban structure?

Our theoretical review has suggested that urban subcenters, like cities themselves, are based on a tension of centripetal and centrifugal forces. Both forces entail strong externalities: external economies producing the agglomerative tendencies, and congestion or nuisance externalities that limit the size and density of agglomeration that is achieved. The first set of externalities is largely positive, suggesting an inadequate private incentive to join an agglomeration. The second set consists of negative externalities, so may cause too many activities to locate close together. But as

we have already seen one of the negative externalities, traffic congestion, also tends to cause residential decentralization (because the action that actually creates the externality, commuting, is associated with living further away rather than close in). Furthermore, residential decentralization and downtown congestion encourage employment decentralization, further eroding the private incentive to maintain healthy central agglomerations — but perhaps creating incentives for welfare-improving secondary agglomerations.

This last possibility is illustrated by the dynamic models described in the previous section. Suppose we start with a monocentric equilibrium (i.e., everyone in one location) and the population gradually increases. As long as the perturbations in the system stem from random events we cannot predict with certainty when an additional center will become established, but over time more and more centers are likely to appear. We can see how the optimal and market growth paths differ by returning to the two-location model of Anas (1992). Anas shows that on an optimal growth path, the second center ought to be established much earlier than it is likely to be established under atomistic defection. Hence, collective action is called for to mitigate the market's failure to optimally time the establishment of a second center. Also, under the optimal path, the second center must be established when it is still too small to be stable; hence planning is needed not only in timing, but also in temporarily protecting the newly established center until it becomes stable and self-sustaining.

Such collective action may take the form of society subsidizing the formation of coalitions which would pioneer the emergence of a second center of a size big enough to insure its future stability. An alternative would be for a large scale developer with foresight to undertake initial infrastructure investments at the location of the second center, reducing the entry costs of firms or consumers relocating there; however, rivalry among developers trying to form competing subcenters causes complex strategic interdependence which results in another layer of market failure suggesting possible gains from regulation (Henderson and Slade, 1993). Yet a third strategy is to subsidize the defection of the first firm to a new subcenter site. Once that is done, interfirm linkages ensure that a sequence of other firms, requiring successively lower subsidies, could be induced to join the new agglomeration.

This subsidy issue was raised in two different contexts in the literature. Henderson (1986b) observed that cities in the United States and Brazil initially formed on the coastlines, and only later

did urbanization spread to the interior. In this situation coastal residents have an incentive to decongest their cities by subsidizing the formation of towns in the interior. Rauch (1993) considered that the developer of an industrial park, in which firms enjoy sideways linkages with one another, should subsidize the first firms moving into the park in order to subsequently attract additional tenants. This is a strategy commonly employed in shopping center developments — which are a form of small-scale planned agglomeration — by giving rental discounts to "anchor stores."

On balance, it is difficult to say whether the process of subcenter formation has created too many or too few subcenters. Since multiple stable equilibria exist under the same parameter values, historical accident can cause a metropolitan economy to get stuck on either an inefficient or an efficient equilibrium. The process of land use planning may improve welfare by promoting those incentives, regulations, and infrastructure investments that minimize the frictions and welfare losses arising from uncoordinated market actions and from historical accidents. However, a precise prescription of "good planning" in this arena remains elusive.

D. Assessment

Broadly speaking, then, we are confronted with a situation with three classes of externalities — transport congestion externalities, neighborhood externalities, and agglomeration externalities. We understand the first two classes much better than the third, although the third is probably the most important. Under these circumstances, theory provides only limited guidance concerning optimal policy. Our judgment is that piecemeal second-best policies addressing just transport congestion externalities are likely to be welfare improving. Such policies include congestion pricing, parking pricing, some measures to encourage carpools, and restricting road capacities in central areas. Land-use controls can sometimes be beneficial, but are more problematic because they tend to repress market forces. Policies designed to exploit economies of agglomeration, such as targeting public infrastructure or promoting local amenities in potential business centers, are sound in principle and may be highly beneficial in the right circumstances. Unfortunately they are also easily subverted to serve parochial business or political interests rather than overall efficiency.

As in many areas of economic policy, no blanket rule will suffice. Each situation must be understood on its own terms, but within a sound knowledge base concerning how all the parts of the

urban economy fit together. A solid understanding of urban structure is a prerequisite to the more ambitious goal of establishing such a knowledge base.

References

Allen, P.M. and Sanglier, M. "A Dynamic Model of a Central Place System — II," Geographical Analysis, April 1981a, 13(2), pp. 149-64.

Allen, P.M. and Sanglier, M. "Urban Evolution, Self-Organization, and Decisionmaking," Environment and Planning A, 1981b, 13, pp. 167-83.

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

Anas, Alex. "From Physical to Economic Urban Models: The Lowry Framework Revisited," in Advances in Urban Systems Modelling. Eds: B. Hutchinson and M. Batty. Amsterdam: North-Holland, 1986, pp. 163-72.

Anas, Alex. "Agglomeration and Taste Heterogeneity: Equilibria, Stability, Welfare and Dynamics," Regional Science and Urban Economics, Feb. 1988, 18(1), pp. 7-35.

Anas, Alex. "Taste Heterogeneity and Urban Spatial Structure: The Logit Model and Monocentric Theory Reconsidered," Journal of Urban Economics, Nov. 1990, 28(3), pp. 318-35.

Anas, Alex. "On the Birth and Growth of Cities: Laissez-Faire and Planning Compared", Regional Science and Urban Economics, June 1992, 22(2), pp. 243-58.

Anas, Alex and Kim, Ikki. "Income Distribution and the Residential Density Gradient," Journal of Urban Economics, March 1992, 31(2), pp. 164-80.

Anas, Alex and Kim, Ikki. "General Equilibrium Models of Polycentric Urban Land Use with Endogenous Congestion and Job Agglomeration", Journal of Urban Economics, September 1996, 40(2), pp. 232-56.

Arthur, W. Brian. "Competing Technologies, Increasing Returns, and Lock-In by Historical Events," Economic Journal, Aug. 1989, 99(394), pp. 116-31.

Arthur, W. Brian. "Positive Feedbacks in the Economy," Scientific American, Feb. 1990, 263(2), pp. 92-99.

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

Bak, Per and Chen, Kan. "Self-Organized Criticality," Scientific American, Jan. 1991, 264(1), pp. 46-53.

Barrett, Paul. The Automobile and Urban Transit: The Formation of Public Policy in Chicago, 1900-1930, Philadelphia: Temple university Press, 1983.

Batty, Michael and Longley, Paul. Fractal Cities. London: Academic Press, 1994.

Borukhov, E. and Hochman, Oded. "Optimum and Market Equilibrium in a City without a Predetermined Center", Environment and Planning A, 1977, 9, pp. 849-56.

Brueckner, Jan. "Urban General Equilibrium Models with Non-Central Production", Journal of

Regional Science, 1978, 18, pp. 203-15.

Brueckner, Jan. "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.), Amsterdam: North-Holland, 1987, pp. 821-45.

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

Cervero, Robert and Wu, Kang-Li. "Polycentrism, Commuting, and Residential Location in the San Francisco Bay Area," Environment and Planning A, forthcoming 1997.

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

Chen, Hsin-Ping. "The Simulation of a Proposed Nonlinear Dynamic Urban Growth Model," Annals of Regional Science, 1996, 30(3), pp. 305-19.

Chinitz, Benjamin. Freight and the Metropolis, Cambridge, Mass: Harvard University Press, 1960.

Chinitz, Benjamin. "Contrasts in Agglomeration: New York and Pittsburgh," American Economic Review, Papers & Proceedings, May 1961, 51(2), pp. 279-89.

Christaller, Walter. Central Places in Southern Germany, 1933. C.W. Baskin (trans.) London: Prentice-Hall, 1966.

Ciccone, Antonio and Robert E. Hall. "Productivity and the Density of Economic Activity", American Economic Review, March 1996, 86(1), pp. 54-70.

Clark, Colin. "Urban Population Densities." Journal of the Royal Statistical Society (Series A), 1951, 114, pp. 490-96.

Clark, Colin. Population Growth and Land Use, London: MacMillan, 1968.

Conlisk, John. "Stability and Monotonicity for Interactive Markov Chains," Journal of Mathematical Sociology, 1992, 17(2-3), pp. 127-143.

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

Crane, Randall. "The Influence of Uncertain Job Location on Urban Form and the Journey to Work" Journal of Urban Economics, May 1996, 39(3), pp. 342-58.

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

David, Paul A. "Clio and the Economics of QWERTY," American Economic Review, May 1985, 75(2), pp. 332-37.

Dixit, Avinash and Stiglitz, Joseph. "Monopolistic Competition and Optimum Product Diversity", American Economic Review, 1977, 67( ), pp. 297-308.

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

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

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

Fales, Raymond and Moses, Leon N. "Land Use Theory and the Spatial Structure of the Nineteenth Century City," Papers and Proceedings of the Regional Science Association, 1972, 28, pp. 49-82.

Field, Alexander J. "The Magnetic Telegraph, Price and Quantity Data and the New Management of Capital," Journal of Economic History, June 1992, 52(2), pp. 401-13.

Fischer, Manfred M., Haag, Günter, Sonis, Michael and Weidlich, Wolfgang. "Account of Different Views in Dynamic Choice Processes," in Spatial Choices and processes. Eds: Manfred M. Fischer, Peter Nijkamp and Y.Y. Papageorgiou. Amsterdam: North-Holland, 1990, pp. 17-47.

Fotheringham, A. Stewart, Batty, Michael and Longley, Paul A. "Diffusion-Limited Aggregation and the Fractal Nature of Urban Growth," Papers of the Regional Science Association, 1989, 67, pp. 55-69.

Fujita, Masahisa. "A Monopolistic Competition Model of Spatial Agglomeration: Differentiated Products Approach", Regional Science and Urban Economics, 1988, 18( ), pp. 87-124.

Fujita, Masahisa. Urban Economic Theory: Land Use and City Size, Cambridge, UK: Cambridge University Press, 1989.

Fujita, Masahisa and Ogawa, Hideaki. "Multiple Equilibria and Structural Transition of Non-monocentric Urban Configurations", Regional Science and Urban Economics, May, 1982, 12(2), pp. 161-96.

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

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

Giuliano, Genevieve and Small, Kenneth A. "Subcenters in the Los Angeles Region," Regional Science and Urban Economics, Vol. 21 (1991), pp. 163-82.

Giuliano, Genevieve and Small, Kenneth A. "Is the Journey to Work Explained by Urban Structure?" Urban Studies, Nov. 1993, 30(9), pp. 1485-500.

Glaab, Charles N. and Brown, Theodore. A History of Urban America, London, The MacMillan Press, 1967.

Glaeser, Edward L., Kallal, Hedi D., Scheinkman, José A. and Shleifer, Andrei. "Growth in Cities," Journal of Political Economy, Dec. 1992, 100(6), pp. 1126-52.

Gordon, Peter, and Richardson, Harry W. "Beyond Polycentricity: The Dispersed Metropolis, Los Angeles, 1970-1990," Journal of the American Planning Association, Summer 1996, 62(3), pp. 289-95.

Gordon, Peter, Richardson, Harry W. and Wong, H.L. "The Distribution of Population and Employment in a Polycentric City: The Case of Los Angeles," Environment and Planning A, 1986, 18, pp. 161-73.

Hamilton, Bruce W. "Wasteful Commuting," Journal of Political Economy, Oct. 1982, 90(5), pp. 1035-53.

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

Harrison, David, and Kain, John F. "Cumulative Urban Growth and Urban Density Functions," Journal of Urban Economics, 1974, 1, pp. 61-98.

Heikkila, E., Gordon, P., Kim, J.I., Peiser, R.B., Richardson, H.W. and Dale-Johnson, D. "What Happened to the CBD-Distance Gradient?: Land Values in a Policentric City," Environment and Planning A, 1989, 21, pp. 221-32.

Henderson, J. Vernon. "Efficiency of Resource Usage and City Size," Journal of Urban Economics, 1986a, 19(), pp. 47-70.

Henderson, J. Vernon. "The Timing of Regional Development", Journal of Development Economics, 1986b, 23(2), pp. 275-92.

Henderson, J. Vernon. Urban Development: Theory, Fact, and Illusion, New York: Oxford University Press, 1988.

Henderson, J. Vernon and Arindam Mitra. "The New Urban Landscape: Developers and Edge Cities," Regional Science and Urban Economics, Dec. 1996, 26(6), pp. 613-43.

Henderson, J. Vernon and Slade, Eric. "Development Games in Non-monocentric Cities", Journal of Urban Economics, September 1993, 34(2), pp. 207-29.

Ingram, Gregory K. and Alan Carroll. "The Spatial Structure of Latin American Cities," Journal of Urban Economics, March 1981, 9(2), pp. 257-73.

Jacobs, Jane. Cities and the Wealth of Nations: Principles of Economic Life. New York: Vintage, 1984.

Krugman, Paul. Geography and Trade. Cambridge, Mass.: M.I.T. Press, 1991a.

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

Krugman, Paul. "First Nature, Second Nature and Metropolitan Location", Journal of Regional Science, May 1993, 33(2), pp. 129-44

Krugman, Paul. The Self-Organizing Economy. Cambridge, Mass.: Blackwell, 1996.

LeRoy, Stephen F. and Sonstelie, Jon. "Paradise Lost and Regained: Transportation Innovation, Income, and Residential Location," Journal of Urban Economics, Jan. 1983, 13(1), pp. 67-89.

Losch, August The Economics of Location, 1940. W.H. Woglom and W.F. Stolper (trans.) New Haven: Yale University Press, 1954.

Makse, Hernan A., Havlin, Shlomo and Stanley, H. Eugene. "Modelling Urban Growth Patterns," Nature, 19 October 1995, 377, pp. 608-12.

McDonald, John F. "The Identification of Urban Employment Subcenters," Journal of Urban Economics, March 1987, 21(2), pp. 242-58.

McDonald, John F. "Econometric Studies of Urban Population Density: A Survey," Journal of Urban Economics, November 1989, 26(3), pp. 361-85.

McDonald, John F. and Prather, Paul J. "Suburban Employment Centres: The Case of Chicago," Urban Studies, March 1994, 31(2), pp. 201-18.

McMillen, Daniel P. "One Hundred Fifty Years of Land Values in Chicago: A Nonparametric Approach," Journal of Urban Economics, July 1996, 40(1), pp. 100-24.

McMillen, Daniel P. and McDonald, John F. "Suburban Subcenters and Employment Density in Metropolitan Chicago," working paper, Tulane University, July 1996a.

McMillen, Daniel P. and McDonald, John F. "Transportation Facilities, Suburban Employment Centers, and Population Density in Metropolitan Chicago," working paper, Tulane University, July 1996b.

Merriman, David, Ohkaward, Toru and Suzuki, Tsutomu. "Excess Commuting in the Tokyo Metropolitan Area: Measurement and Policy Simulations," Urban Studies, Feb. 1995, 32(1), pp. 69-85.

Mieszkowski, Peter and Mills, Edwin S. "The Causes of Metropolitan Suburbanization," Journal of Economic Perspectives, Summer 1993, 7(3), pp. 135-47.

Mills, Edwin S. "An Aggregative Model of Resource Allocation in a Metropolitan Area," American Economic Review, 1967. 57, pp. 197-210.

Mills, Edwin S. Studies in the Structure of the Urban Economy, Baltimore: The Johns Hopkins Press, 1972.

Mills, Edwin S. and Bruce W. Hamilton. Urban Economics, New York: Harper-Collins, 1994.

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

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

Mirrlees, James A. "The Optimum Town," Swedish Journal of Economics, March 1972, 74(1), pp. 114-35.

Moomaw, Ronald L. "Agglomeration Economies: Localization or Urbanization?", Urban Studies, 1988, 25(), pp. 150-61.

Moses, Leon N. and Williamson, Harold F. Jr. "The Location of Economic Activity in Cities," American Economic Review, 1967, 57, pp. 211-22.

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

Nelson, Richard R. "Recent Evolutionary Theorizing About Economic Change," Journal of Economic Literature, March 1995, 33(1), pp. 48-90.

O hUallacháin, Breandán. "Agglomeration of Services in American Metropolitan Areas", Growth and Change, Summer 1989, 20(3), pp. 34-49.

Papageorgiou, Yorgo Y. and Pines, David. "The Exponential Density Function: First Principles, Comparative Statics, and Empirical Evidence," Journal of Urban Economics, Sept. 1989, 26(2), pp. 264-68.

Powell, Walter W. "Neither Market nor Hierarchy: Network Forms of Organization", in Research in Organizational Behavior, 1990 (12), pp. 295-336.

Pumain, D., Saint-Julien, Th. and Sanders, L. "Application of a Dynamic Urban Model," Geographical Analysis, April 1987, 19(2), pp. 152-66.

Rauch, James E. "Does History Matter only when it Matters too Little ? The Case of City-Industry Location", The Quarterly Journal of Economics, August 1993, pp.

Robinson, E.A.G. The Structure of Competitive Industry, University of Chicago Press, Chicago, 1958.

Rosen, Kenneth T. and Resnick, Mitchel. "The Size Distribution of Cities: An Examination of the Pareto Law and Primacy," Journal of Urban Economics, Sept. 1980, 8(2), pp. 165-86.

Schwartz, Alex. "Corporate Service Linkages in Large Metropolitan Areas: A Study of New York, Los Angeles, and Chicago," Urban Affairs Quarterly, Dec. 1992, 28(2), pp. 276-96.

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

Scott, Allen J. "Electronics Assembly Subcontracting in Southern California: Production Processes, Employment, and Location," Growth and Change, Winter 1991, 22(1), pp. 22-35.

Serck-Hansen, Jan. "The Optimal Number of Factories in a Spatial Market," in: Towards Balanced International Growth, Ed. Bos, H.C. Amsterdam: North-Holland, 1969.

Sivitanidou, Rena. "Do Office-Commercial Firms Value Access to Service Employment Centers? A Hedonic Value Analysis within Polycentric Los Angeles," Journal of Urban Economics, Sept.

1996, 40(2), pp. 125-49.

Small, Kenneth A. "A Comment on Gasoline Prices and urban Structure," Journal of Urban Economics, Nov. 1981, 10(3), pp. 311-22.

Small, Kenneth A. Urban Transportation Economics, Vol. 51 of Fundamentals of Pure and Applied Economics series, Harwood Academic Publishers, 1992.

Small, Kenneth A. and Song, Shunfeng. "Population and Employment Densities: Structure and Change," Journal of Urban Economics, Nov. 1994, 36(3), pp. 292-313.

Small, Kenneth A. and Song, Shunfeng. "'Wasteful' Commuting: A Resolution," Journal of Political Economy, Aug. 1992, 100(4), pp. 888-98.

Small, Kenneth A. and Song, Shunfeng. "Population and Employment Densities: Structure and Change," Journal of Urban Economics, Nov. 1994, 36(3), pp. 292-313.

Solow, Robert M. and Vickrey, William S. "Land Use in a Long Narrow City", Journal of Economic Theory, 1971, 3, pp. 430-47.

Song, Shunfeng. "Modelling Worker Residence Distribution in the Los Angeles Region," Urban Studies, Nov. 1994, 31(9), pp. 1533-44.

Starrett, David A. "Principles of Optimal Location in a Large Homogeneous Area," Journal of Economic Theory, 1974, 9(), pp. 418-48.

Sveikauskas, Leo. "The Productivity of Cities", Quarterly Journal of Economics, 1975, 89(), pp. 393-413.

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

Tiebout, Charles M. "A Pure Theory of Local Expenditures," Journal of Political Economy, October 1956, 64(5), pp. 416-424.

Vernon, Raymond. Metropolis 1985, Harvard University Press, 1960.

Von Thunen, J. Der Isolierte Staat in Beziehung ant Landswirtschaft and Nationalekomie. Hamburg, 1826

Warner, Sam Bass Jr. Streetcar Suburbs: The Process of Growth in Boston (1870-1900), Cambridge, Mass: Harvard University Press, 1962.

Wheaton, William C. "A Comparative Statics Analysis of Urban Spatial Structure," Journal of Economic Theory, 1974, 9, pp. 223-37.

Wheaton, William C. "Income and Urban Residence: An Analysis of Consumer Demand for Location," American Economic Review, Sept. 1977, 67(4), pp. 620-31.

Wheaton, William C. "Price-Induced Distortions in American Highway Investment," Bell Journal of Economics, Summer 1978, 9(2), pp. 622-32.

White, Michelle J. "Firm Suburbanization and Urban Subcenters," Journal of Urban Economics, Oct. 1976, 3(4), pp. 323-43.

White, Michelle J. "Location Choice and Commuting Behavior in Cities with Decentralized Employment", Journal of Urban Economics, 1988, 24( ), pp.129-52.