Journal of Urban Economics **50,** 448-473 (2001)
doi:10.1006/juec.2001.2228, available online at http://www.idealibrary.com on IDEAL®

Nonparametric Employment Subcenter Identification

Daniel P. McMillen*

Department of Economics (MC 144), University of Illinois,
601 S. Morgan St., Chicago, Illinois 60607
E-mail: mcmillen@uic.edu

Received December 9, 1999; revised December 11, 2000;
published online July 18, 2001

A two-stage procedure is proposed for identifying urban employment subcenters. The first stage identifies candidate subcenters as significant positive residuals in a smoothed employment density function. Subcenters are those sites that provide significant explanatory power in the second-stage, semiparametric employment density function estimation. The procedure can be applied to either aggregated or disaggregated data, does not require detailed knowledge of the study area, and is easily reproducible by other researchers. Results are presented for five previously studied cities—Chicago, Dallas, Houston, Los Angeles, and San Francisco—and a new one, New Orleans. © 2001 Academic Press

Key Words: subcenters; employment density; population density; nonparametric; semiparametric; fourier expansion.

1. INTRODUCTION

A polycentric city has one or more employment subcenters beyond the traditional central business district (CBD). In contrast, a monocentric city can have decentralized employment, but the non-CBD employment is not grouped into subcenters. Subcenters enjoy some of the same agglomeration economies as the CBD but offer lower commuting costs for suburban workers and lower land costs for firms. Recent research suggests that large employment subcenters have pronounced effects on non-CBD employment density, housing prices, land rents, and population density.[1]

The first step in estimating the effects of subcenters on urban spatial structure is identifying the subcenters. Important contributions include [6, 8, 9, 14, 16, 24–26, 30–32, 36, 48]. Although some consensus has developed, the task has not been easy. Based on the papers just listed, a reasonable working definition of a subcenter is a site (1) with significantly larger employment density than nearby locations that has (2) a significant effect on the overall employment

* I thank Steve Craig, Janet Kohlhase, John McDonald, and Ken Small for helpful comments.

[1] Important empirical contributions on the effects of subcenters on urban spatial structure include [5, 8, 9, 12, 13, 15, 25–27, 30, 32, 33, 35–37, 43, 47, 48]. An excellent overview is provided by Anas *et al.* [1].

0094-1190/01 $35.00

density function. Difficulties arise in making the two points operational. How large is large? What is the appropriate definition of nearby? Should we condition on distance from the CBD or consider each site only within its local context? If we do condition on distance from the CBD, should the city be treated as symmetric, or should the CBD gradient vary in different areas of the city?

Existing methods generally require ample knowledge of the study area. As an example, Giuliano and Small [24] suggest a reasonable definition of a subcenter—a set of contiguous sites that each has a minimum employment density of 10 employees per acre and that together have at least 10,000 employees. Their procedure produces a list of 32 subcenters in the Los Angeles area. McMillen and McDonald [36] use the same procedure and cutoff points to identify 15 subcenters in Chicago, but because one subcenter has more than 400,000 employees and covers an area extending from O'Hare Airport in the northwest suburbs to Lake Michigan and into Lake County, McMillen and McDonald raise the cutoff points to 20 employees per acre and 20,000 total employees to produce reasonable results. The appropriate cutoff points must be obtained through a process of trial and error, guided by local knowledge. Craig and Ng [14] identify subcenters as local rises in the employment density function, estimated as a function of distance from the CBD. This procedure identifies rings around the CBD that may contain subcenters. They then use a combination of local knowledge and employment statistics to identify subcenters.

A less widely recognized problem is that the number of subcenters depends on the size of the unit of observation. Giuliano and Small [24] have 1,146 tracts covering an area of 3,536 square miles, whereas McMillen and McDonald [36] have 14,290 tracts in an area of 3,572 square miles. The small tract size leads McMillen and McDonald to work with proximity rather than contiguity, with proximity defined as within 1.5 miles. The change is necessary because a small tract size leads to many pockets with no employment in an otherwise high-density area. The definition of proximity depends on the tract size and again requires local knowledge. Large tract sizes smooth over local employment peaks and may produce fewer subcenter sites than more disaggregated data.

I propose a two-stage nonparametric procedure for identifying subcenters that is easy to implement for a variety of cities and tract sizes. The first stage uses a nonparametric estimator, locally weighted regression, to smooth employment density. The estimate of a site's employment density is obtained by weighted least squares, with more weight given to nearby tracts. Potential subcenters are sites with significant positive residuals. Thus a potential subcenter is a site with unusually large density after broad spatial trends are accounted for.

In the second stage of the procedure, a semiparametric regression procedure is used to determine whether the potential subcenters have significant effects on employment density. The nonparametric part of the regression captures the effect of distance from the CBD using a flexible Fourier form. Thus the results are conditioned on distance from the CBD, but the CBD gradient can vary spatially. The parametric part of the regression accounts for the effects of subcenter proximity on employment density. Actual subcenter sites are omitted from the regression to avoid biasing the results by including endogenously chosen employment density peaks as explanatory variables for density. The final list of subcenters includes the sites providing significant explanatory power in the employment density function. The procedure captures the idea that a subcenter is an area with an employment density that is significantly higher than would be expected based only on its distance from the CBD.

I use the procedure to identify employment subcenters in six cities: Chicago, Dallas, Houston, Los Angeles, New Orleans, and San Francisco. Chicago, Houston, and Los Angeles are chosen because they are the focus of the studies by McMillen and McDonald [36], Craig and Ng [14], and Giuliano and Small [24], respectively. Dallas is the subject of Shukla and Waddell's [47] study and is generally thought to be a polycentric city. Cervero and Wu [8, 9] use Giuliano and Small's procedure to identify subcenters in the San Francisco Bay area. New Orleans has not been studied before, but it meets my criterion of ample local knowledge. The results are similar to those of the other studies but are somewhat more likely to identify subcenters in suburban areas with low overall levels of employment density. After identifying the subcenters, I present evidence that they have significant effects on population and employment densities.

2. SUBCENTER IDENTIFICATION PROCEDURES

Many studies use more or less arbitrary definitions of subcenter locations when analyzing the effects of employment subcenters on urban spatial structure.[2] As the literature has grown, more rigorous definitions have been proposed. One strand of analysis began with Giuliano and Small [24] and has been followed since by Bogart and Ferry [6], Cervero and Wu [8, 9], Small and Song [48], and as a first stage of the analysis by McMillen and McDonald [36]. The procedure only establishes a set of candidate sites; subsequent analysis is necessary to determine whether the identified sites have statistically significant effects on urban spatial structure. As discussed in Section 1, the procedure is sensitive to the unit of analysis and is difficult to implement without the detailed local knowledge necessary to establish reasonable definitions of the cutoff points and the notion of proximity.

Another set of studies uses regression analysis to determine subcenter candidates. The regression approach has an advantage in that it is less sensitive to the unit of analysis. McDonald [31] estimates a standard monocentric employment density function for the Chicago area and identifies subcenters as clusters of positive residuals. Subsequent studies by McDonald and McMillen [32] and

[2] See, for example, [5, 15, 26, 27, 30, 47].

McDonald and Prather [33] use the same procedure. The procedure works best in a nearly monocentric city. The local rise in employment density produced by a subcenter flattens the estimated employment density function, which reduces the probability of identifying subcenters. Cities with multiple subcenters or distinctive topographical features may not have employment density functions that are symmetric about the CBD, which may produce groups of positive residuals in areas other than those associated with employment subcenters. Clusters are identified by inspection, which limits the extent to which the results are reproducible by other researchers.

Two-stage regression procedures have been proposed by Gordon *et al.* [25] and McMillen and McDonald [36]. Gordon *et al.* [25, p. 164] identify candidate sites for Los Angeles "via visual inspection of density maps." They use distance from 57 candidate sites as explanatory variables for population and employment density functions, but conclude that only six subcenters have statistically significant effects on densities. McMillen and McDonald [36] use Giuliano and Small's [24] procedure to identify 20 candidate sites for Chicago, of which 17 provide significant explanatory power in employment density function estimates. "Visual inspection" is somewhat arbitrary, while McMillen and McDonald's use of the Giuliano and Small procedure faces the same problems discussed earlier.

Craig and Ng [14] use a nonparametric procedure to obtain smoothed employment density estimates for Houston. Their use of quantile regression is a potential advantage over my approach. However, their approach is better suited to identifying *rings* of high employment density than to identifying subcenters. They estimate employment density as a function of distance from the CBD and look for local rises in the density–CBD relationship. They then inspect the rings where density is rising to find sites with unusually high density and employment. The last step again requires local knowledge to accept or reject a high density site as a subcenter. Craig and Ng's procedure is similar to McDonald's [31] in that the initial conditioning on distance from the CBD is best suited to a nearly monocentric city.

3. A TWO-STAGE NONPARAMETRIC PROCEDURE

My goal is to develop an objective, reproducible procedure that can be applied to a variety of cities, including those for which the researcher does not have detailed local knowledge. The ideal procedure:

- is appropriate for different units of analysis (e.g., square miles, census tracts, or zip codes)
- identifies statistically significant local rises in employment density
- is conditioned on distance from the CBD (because the rise in employment density required to produce a subcenter is higher closer to the CBD)

- allows for local variations in the effect of distance from the CBD on employment density (because many cities have terrain variations that cause the gradient to vary over the metropolitan area)
- provides a measure of the geographic area covered by the subcenters.

My two-stage nonparametric procedures meets the first four of these objectives, and it can be modified to accomplish the last objective.

In the first stage, a locally weighted regression (LWR) procedure is used to smooth the natural logarithm of employment density (y) over space.[3] The Appendix provides details on the estimation procedure. The idea is very simple; nearby observations are given more weight when estimating the smoothed value of y at a given site. In spatial modeling, "nearby" has a straightforward geographic definition (which has led Brunsdon *et al.* [7] to relabel LWR as "geographically weighted regression"). I use a large window size; the nearest 50% of the observations receive some weight in estimating the smoothed value of y at a site. The large window size leads to a smooth surface, but the estimates are less restrictive than a standard exponential function. For example, estimated densities can decline more rapidly on the north side of a city than on the south side. The initial smooth serves as a benchmark. Subcenters produce densities that greatly exceed the initial smooth.

The list of potential subcenters comprises those sites with residuals that are significantly greater than 0 at the 5% significance level: $(y_i - \hat{y}_i)/\hat{\sigma}_i > 1.96$, where $\hat{y}_i$ is the LWR estimate of y at site i and $\hat{\sigma}_i$ is the estimated standard error for the prediction. To avoid including many nearby sites as subcenters when significant residuals cluster together, I narrow the list of potential subcenters to sites whose predicted log-employment densities are highest among all observations with significant positive residuals in a 3-mile radius. Although the chosen radius is arbitrary, it is reasonable and is consistent with the idea that a subcenter leads to a local peak in the employment density function.

The proposed procedure for identifying candidate subcenter sites has several advantages over existing methods. Using the initial smooth of the data as a benchmark partially eliminates the sensitivity of the results to the tract size. In contrast to the Craig and Ng [14] procedure, LWR estimates can detect *local* rises in the employment density function, because only nearby observations are used in estimation. Unlike McDonald's [31] procedure, the LWR estimates account for variation in density gradients across a metropolitan area. Rather than relying on local knowledge to choose density cutoff points as in Giuliano

[3] LWR was originally proposed by Stone [49] and Cleveland [10] and has been developed further by Cleveland and Devlin [11], Fan [17, 18], Fan and Gijbels [19], and Ruppert and Wand [45]. Pagan and Ullah [40] provide an excellent overview. It is a simple extension of the Nadaraya [39] and Watson [50] kernel regression estimator. LWR has been used productively in spatial modeling in [7, 34, 35, 38, 42, 51].

and Small [24], the procedure uses statistical criteria to determine whether local employment peaks are large enough to be counted as subcenters. Visual inspection of residual clusters is avoided by using the 3-mile radius to define subcenter peaks.

The first stage only identifies candidate subcenter sites because, although it detects local rises in the employment density function, it does not determine whether the site has a statistically significant effect on the overall shape of the employment density function. The second stage uses a semiparametric procedure (Robinson [44]) to assess the significance of the candidate sites. As in Gordon *et al.* [25], we face the problem of determining the significance of multiple sites when we do not know the appropriate functional form or the true spatial extent of the subcenters.

Let D_{ij} represent the distance between observation i and candidate subcenter site j, and let $DCBD_i$ represent the distance from observation i to the central business district. With S candidate subcenter sites, the semiparametric regression is

$$y_i = g(DCBD_i) + \sum_{j=1}^{S}(\delta_{1j} D_{ij}^{-1} + \delta_{2j} D_{ij}) + u_i. \tag{1}$$

DCBD enters the equation nonparametrically, because it is a “nuisance” variable. Equation (1) accounts for its effect in a general way while retaining the ability to conduct convenient hypothesis tests on the coefficients of interest, δ_{1j} and δ_{2j}. The subcenter distance variables, D_j, enter the equation in both level and inverse forms. Levels are preferable when a subcenter’s effect is spread over a large area, whereas the inverse form is better suited to modeling a subcenter that has a more local effect. I use a stepwise estimation procedure (described later in this section) to determine which variables to include for the subcenters. Candidate subcenter sites—the observations with highest estimated density among all nearby sites with positive residuals—are omitted in Eq. (1) to eliminate the bias introduced by including endogenously chosen sites as explanatory variables. Also, any site within a mile of the CBD is not counted as a subcenter.[4]

Various alternatives can be used to estimate $g(DCBD)$. Anderson [2, 3] has used cubic splines to approximate $g(DCBD)$ in estimating population density functions. LWR or standard kernel procedures ([4, 11, 28, 29, 40]) also could be used. In practice, the choice of procedure makes little difference as long as it is highly flexible. I use a flexible Fourier form ([20–22, 41]) to approximate $g(DCBD)$. The Fourier expansion supplements a standard quadratic specification with trigonometric terms. To implement the procedure, the variable *DCBD*

[4] Otherwise, a nearby tract could be counted as a subcenter if it has a higher employment density than the tract in which the traditional CBD is located.

is first transformed to lie between 0 and 2π. Denoting the transformed variable by z, the Fourier expansion is

$$g(DCBD_i) \approx \lambda_0 + \lambda_1 z_i + \lambda_2 z_i^2 + \sum_{q=1}^{Q} (\gamma_q \cos(qz_i) + \delta_q \sin(qz_i)). \tag{2}$$

Standard information criteria can be used to choose Q. I use the Schwarz information criterion (SIC) [46], by which the optimal Q is the value that minimizes $SIC(m) = \ln \hat{\sigma}^2 + \frac{m \log(n)}{n}$, where m is the number of estimated coefficients ($m = 3 + 2Q$) and n is the number of observations. Larger values of m reduce the estimated variance but increase the second term. I omit the subcenter distance variables when choosing Q.

The flexible Fourier form is much faster than LWR or kernel regression methods, and is more flexible than cubic splines. Gallant [22] describes the approach as “semi-nonparametric” in that it uses a parametric approximation to an underlying nonparametric function. The approach is quick and easily implemented when the function to be approximated includes only a single variable, and it makes hypothesis testing easy.

A remaining problem is the possibility of a large number of candidate subcenters. Multiple distance variables produce severe multicollinearity. I use a reverse stepwise regression procedure to choose the number of subcenter distance variables. All variables enter the estimating equation in the first stage. I include $-D_j$ and D_j^{-1} as explanatory variables, so that the estimated coefficients are positive when proximity to a subcenter increases densities. The subcenter variable with the lowest t value is then deleted.[5] The smaller equation then is estimated, and the process is repeated until all subcenter distance variables are significant at the 20% level. The intercept and Fourier terms are forced to remain in the regression at each stage. The final list of subcenters includes the sites with positive coefficients on either $-D_j$ or D_j^{-1} (or both) at the end of the stepwise regression procedure.

The two-stage procedure is readily reproducible by other researchers, controls in a general way for nonlinearity, and assesses the statistical significance of the results. It is more similar in spirit to McDonald’s [31] approach than to Giuliano and Small’s [24] approach in that it relies on statistical criteria to assess the importance of subcenters. One could argue that a subcenter is simply an area with large employment. The problem then is determining the definition of “large” when more than one city is analyzed. Statistical criteria provide an objective measure of size; a subcenter is deemed large if it leads to a significant local rise in estimated employment density. The procedure does not provide a measure of subcenter size, however. Giuliano and Small’s [24] method could be used to measure subcenter size if the researcher is willing to specify critical

[5] The deleted variable can be statistically significant if its coefficient is negative.

values for density. Another possibility is to estimate a fully nonparametric version of Eq. (1) and test for a critical distance at which subcenter proximity no longer raises employment density.

4. DATA

The data come from the Urban Element of the Census Transportation Planning Package, which is produced by the Department of Transportation's Bureau of Transportation Statistics (BTS). The BTS obtained special tabulations of 1990 U.S. Census data to match standard Census data with their unit of analysis, the transportation analysis zone (TAZ). The TAZ size varies across metropolitan areas but is usually small—smaller than census tracts or zip codes. The most important variables for this study are total employment, location, and the TAZ size. Total employment comes directly from the BTS data. A Geographic Information System (GIS) program was used to measure the area of each TAZ (in square miles) and to provide coordinates for the TAZ centroids. These coordinates are used to measure distance to the CBD.

The six metropolitan areas analyzed are Chicago, Dallas, Houston, Los Angeles, New Orleans, and San Francisco. The analysis is limited to sites within 50 miles of the CBD, because the Los Angeles area extends to the Nevada state line and the other metropolitan areas include large areas of primarily agricultural land. Some summary data are provided in Table 1. Population ranges from 1.2 million in New Orleans to 13.3 million in Los Angeles, with roughly proportional ranges in employment. New Orleans is something of an

TABLE 1
City Characteristics

Variable	Chicago	Dallas	Houston	Los Angeles	New Orleans	San Francisco
Population	7,934,311	3,832,826	3,659,784	13,339,364	1,217,865	5,740,287
Employment	3,697,823	1,844,157	1,751,237	6,516,772	454,153	2,979,459
# TAZ	11,752	5,950	2,512	3,288	498	3,127
# TAZ with Population	8,796	4,663	2,303	3,201	418	2,873
# TAZ with Employment	6,708	4,584	2,128	3,288	452	2,989
Area in Square Miles	4807	5,952	6,514	5,309	2,807	4,265
Average TAZ size	0.42	1.05	2.60	1.74	5.64	1.41

outlier; in the other cities, roughly half the population is employed, compared with 37.3% in New Orleans. City areas range from 2,807 square miles in New Orleans to 6,514 square miles in Houston. The average tract size ranges from 0.42 square miles in Chicago to 5.64 square miles in New Orleans.

5. SUBCENTER LOCATIONS

The top two panels of Table 2 present standard monocentric population and employment density estimates for the six cities, with the natural log of density as the dependent variable and distance from the CBD as the explanatory variable. The population and employment density gradients are highest in New Orleans and lowest in Los Angeles. By this measure, New Orleans is the most centralized of the six cities. The bottom panel of Table 2 gives an indication of the improvement in explanatory power provided by the semiparametric estimator. An R^2 is available for the semiparametric regression, because the Fourier expan-

TABLE 2
Regression Results

	Chicago	Dallas	Houston	Los Angeles	New Orleans	San Francisco
Log population density						
OLS intercept	9.846	8.321	8.260	9.587	8.734	9.522
	(230.209)	(166.368)	(125.492)	(159.204)	(89.624)	(184.918)
OLS gradient	−0.101	−0.064	−0.078	−0.047	−0.112	−0.051
	(72.610)	(29.517)	(29.215)	(18.158)	(13.128)	(25.728)
OLS R^2	0.375	0.158	0.271	0.099	0.293	0.189
Observations	8780	4655	2303	3001	418	2844
Log employment density						
OLS intercept	8.847	7.565	7.854	8.414	8.329	8.329
	(187.220)	(129.320)	(95.188)	(116.275)	(75.911)	(133.231)
OLS gradient	−0.109	−0.088	0.121	−0.052	−0.247	−0.054
	(61.741)	(31.776)	(32.930)	(17.034)	(24.342)	(22.473)
OLS R^2	0.363	0.187	0.338	0.087	0.568	0.148
Observations	6698	4379	2128	3051	452	2908
LWR and semiparametric estimation—Log employment density						
Number of significant LWR residuals	63	34	30	58	4	24
Number of subcenters	33	28	25	19	2	22
Fourier expansion length (Q)	5	2	1	0	0	3
R^2, semiparametric regression	0.488	0.610	0.659	0.155	0.663	0.516

Note: Absolute t values are in parentheses.

sion method allows Eq. (1) to be estimated by ordinary least squares (OLS). The semiparametric R^2 is generally much higher than the simple model's R^2, which is another indication that the cities are no longer monocentric. The first-stage LWR estimates identify 63 significant positive residuals in Chicago, 34 in Dallas, 30 in Houston, 58 in Los Angeles, 4 in New Orleans, and 24 in San Francisco. The second stage of the procedure reduces the number of subcenters to 33 in Chicago, 28 in Dallas, 25 in Houston, 19 in Los Angeles, 2 in New Orleans, and 22 in San Francisco.

The significant subcenters are displayed in Figures 1–6. The locations of the Chicago subcenters are similar to those found by McMillen and McDonald [36]. The datasets are somewhat different, because the BTS includes Chicago and areas of Indiana. Thus Figure 1 includes two sites on the north side of Chicago and two locations in northwestern Indiana that were not in McMillen and

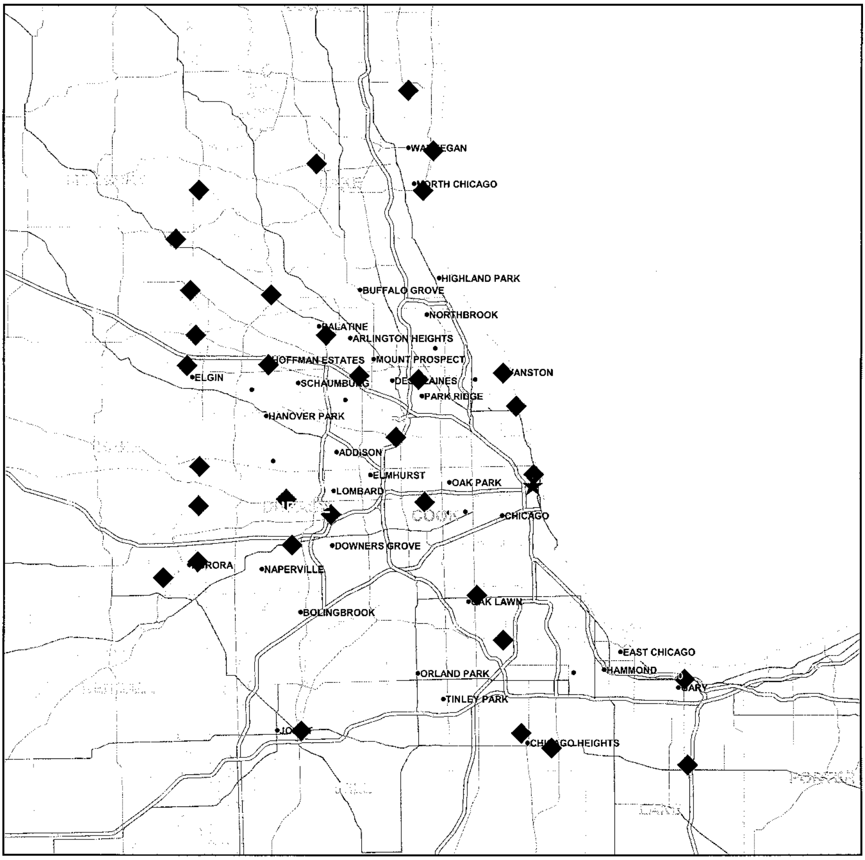

FIG. 1. Chicago.

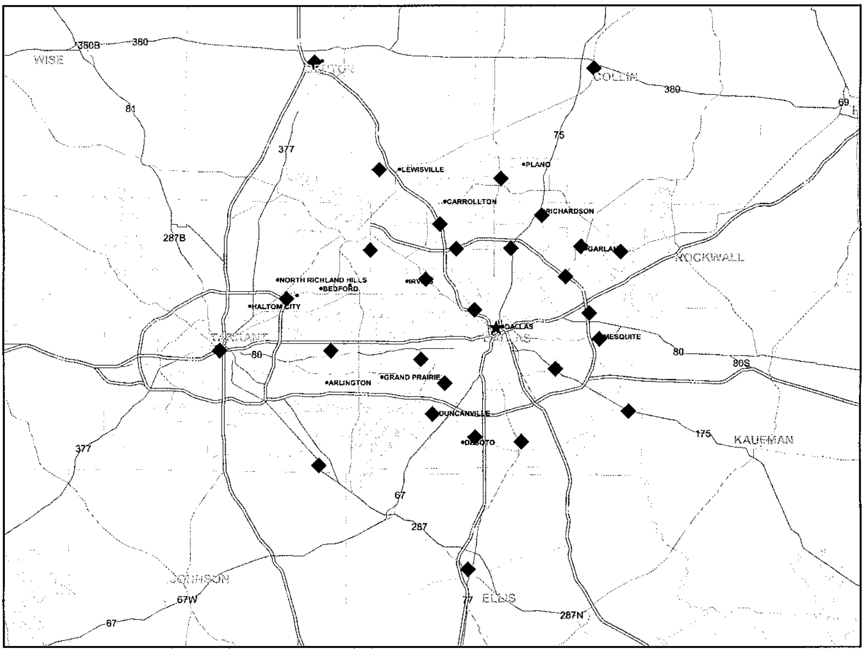

FIG. 2. Dallas.

McDonald [36]. The estimation procedure identifies sites along the Fox River to the west of Chicago that were too small to meet the subcenter criteria used in McMillen and McDonald's study but that have significant local effects on employment density. The two-stage procedure finds several subcenters near O'Hare Airport, the Schaumburg "Edge City" (Garreau [23]), and the old satellite suburbs identified by McMillen and McDonald.

Figure 2 displays the results for the Dallas-Fort Worth metropolitan area. Subcenters have not been identified formally for Dallas-Fort Worth, but some preliminary and suggestive maps are presented in Shukla and Waddell [47]. The two-stage procedure finds subcenters on the north side of Dallas, as in [47], and also finds the Fort Worth CBD.[6] In general, Figure 2 is similar to Figures 2 and 3 in [47].

Figure 3 is directly comparable to the maps produced by Craig and Ng [14] for Houston. We find nearly identical locations for subcenters in sites labeled by Craig and Ng as Galleria, Pasadena, Greenspoint, NASA, and Baytown. They find two additional subcenters, in Westheimer at Beltway and LaPorte. Figure 3 includes a subcenter close to Westheimer at Beltway, but somewhat

[6] For the Dallas-Fort Worth area, I condition only on the Dallas CBD, so it is encouraging that the estimation procedure finds Fort Worth.

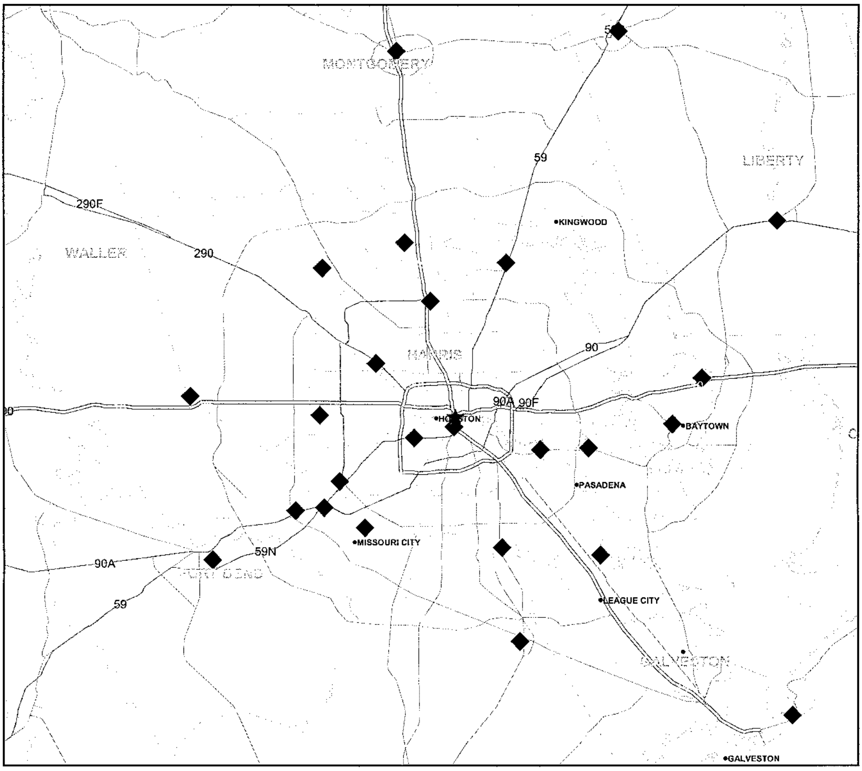

FIG. 3. Houston.

to the northwest. Only Craig and Ng's LaPorte subcenter does not have a counterpart in Figure 3. Some of the additional subcenters shown in Figure 3 are analyzed more informally in Garreau [23]. However, the nonparametric procedure identifies more subcenters having significant local effects on employment density than are identified by either Craig and Ng's formal or Garreau's informal procedure.

Figure 4 shows that New Orleans has two subcenters. The first is in suburban Metairie at the base of the causeway across Lake Pontchartrain. The second is along Airline Highway, near the Mississippi River. Both are in suburban Jefferson Parish, which has approximately the same total employment as Orleans Parish. The subcenter locations are well known in New Orleans as having significant employment concentration. Figure 5 displays the subcenters for the San Francisco Bay area. The map looks similar to Figure 3 in Cervero and Wu [9], and the sites ringing the Bay are nearly identical to those in Garreau [23].

Figure 6 is comparable to Figure 1 in Giuliano and Small [24]. Both studies find subcenters north of downtown Los Angeles, near the CBD, and in Orange

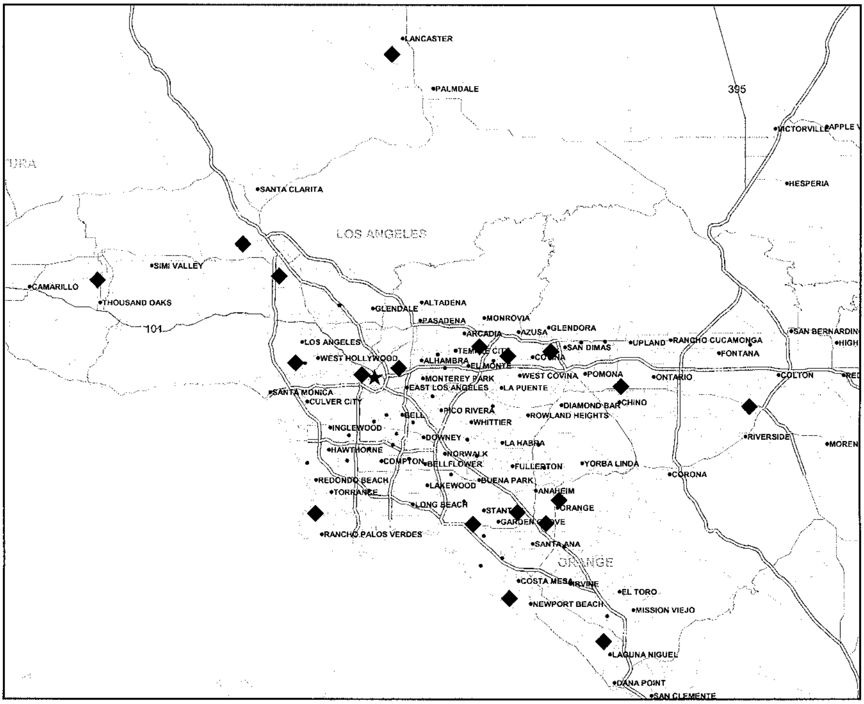

FIG. 4. New Orleans.

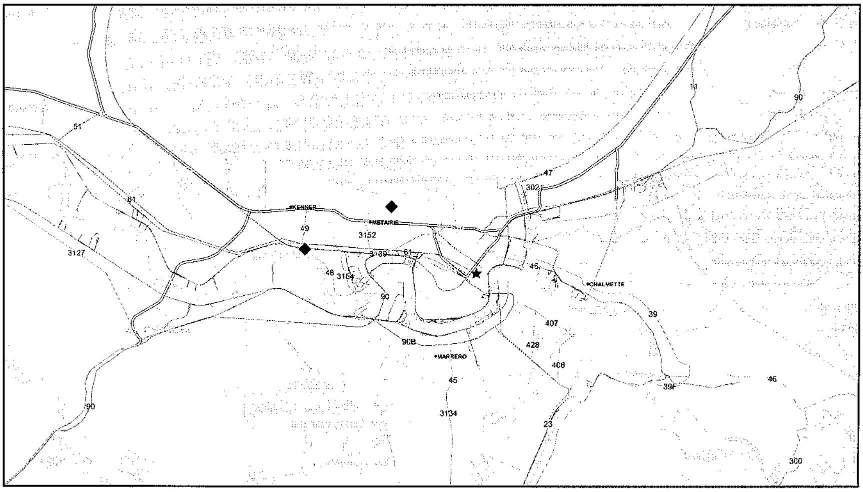

FIG. 5. San Francisco.

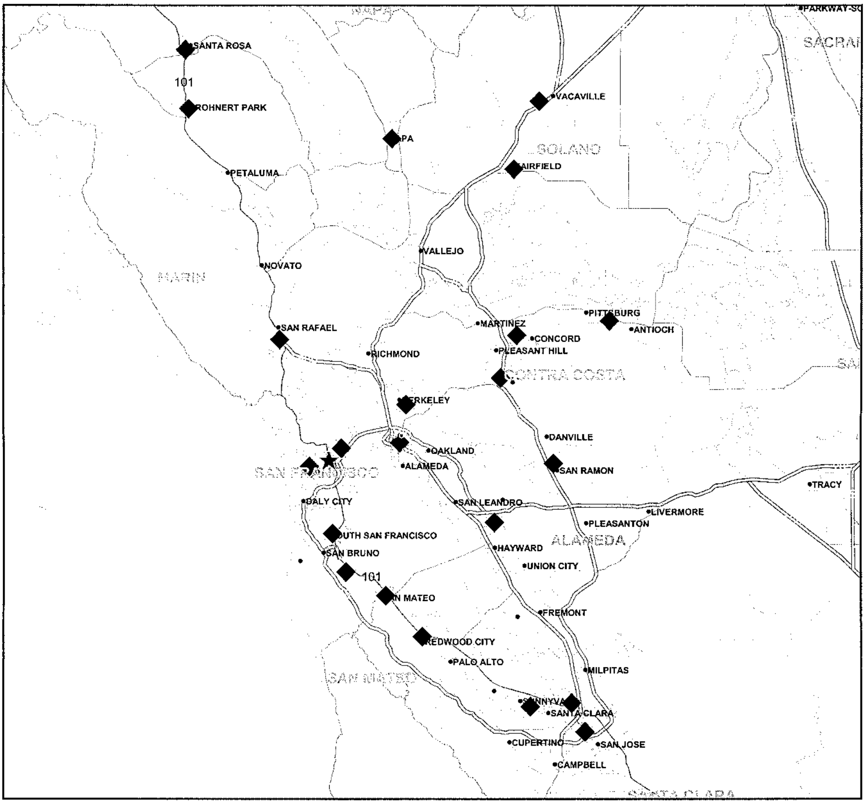

FIG. 6. Los Angeles.

County. The primary difference is that my estimation procedure does not identify sites near the Los Angeles Airport, and Giuliano and Small do not identify sites along the freeways heading east from the CBD. Most of the difference between our results is explained by my use of data from 1990 rather than 1980, and by my dropping from the subcenter list sites that do not have significant effects on the estimated employment density functions. Figure 7 shows the subcenters identified using Giuliano and Small's procedure with the 1990 BTS data. I use the cutoff points from the Small and Song [48] study; a potential subcenter site is the tract with highest employment among a group of contiguous tracts that each have at least 20 employees per acre and that together have at least 20,000 employees.[7] The procedure identifies 34 potential subcenter sites, of which 15 have significant effects in semiparametric employment density

[7] To simplify the programming, I count tracts as contiguous if they are within 1.5 miles of one another.

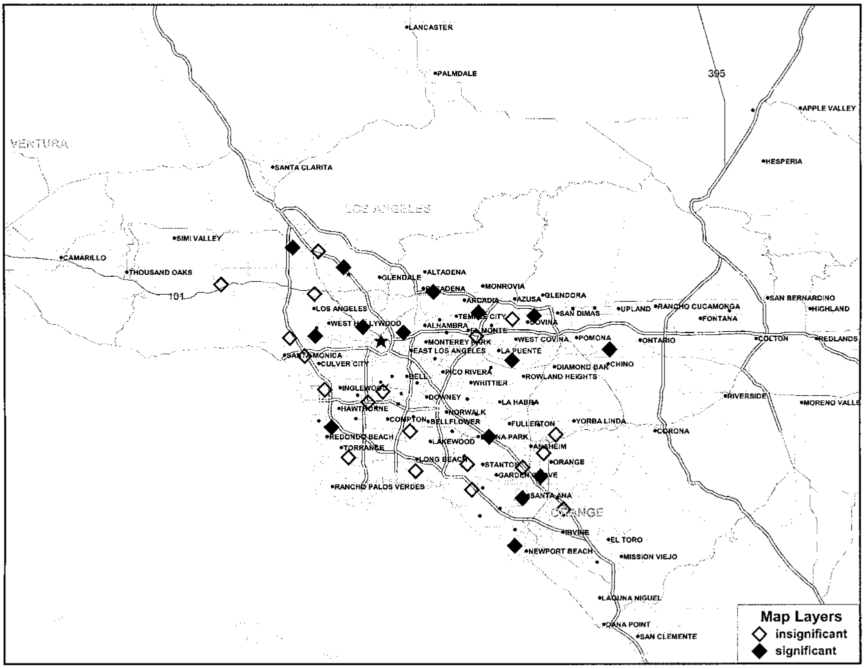

FIG. 7. Los Angeles—Giuliano–Small procedure.

functions estimated in the same way as in my two-stage procedure. As with my procedure, the site nearest the Los Angeles Airport is not statistically significant in the estimated employment density function. Unlike in the Giuliano–Small and Small–Song studies, the updated data reveal several subcenter sites along the freeways heading east from Los Angeles.

Figure 8 directly compares the results of the two procedures. Significant subcenter sites from my two-stage procedure are identified by solid diamonds, while significant sites found for 1990 with the Giuliano–Small approach are marked by a sun symbol. Solid suns are sites that both procedures identify as subcenters. Nine subcenters are identified at the same or almost the same sites by both procedures. Both procedures find sites in the eastern suburbs and in Orange County, but at somewhat different locations. The biggest difference is that the two-stage procedure identifies several sites in distant locations—Lancaster, Thousand Oaks, Riverside, and Laguna Niguel—that the Giuliano–Small procedure does not. These are sites with significant local effects on employment density but that do not have the requisite 20,000 total employees in the set of contiguous tracts.

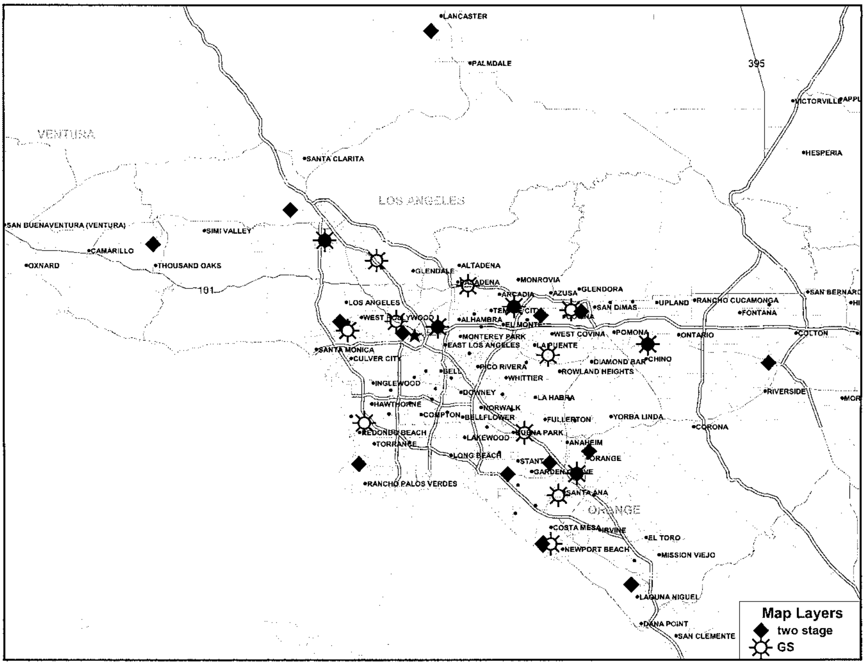

FIG. 8. Comparison of two-stage and Giuliano–Small (GS) procedures for Los Angeles.

6. AGGREGATION

One of the objectives for the two-stage procedure is to reduce the sensitivity of the results to the dataset's average tract size. Using the Giuliano–Small procedure, a cutoff point of 20 employees per acre and total subcenter employment of 20,000 may be reasonable in Los Angeles, which has an average tract size of 1.74 square miles and total employment of 6.5 million. What are reasonable cutoff points in New Orleans, Dallas, or a city that has not yet been analyzed? The cutoff points need to change when the average tract size changes. Low cutoff points will lead to extremely large subcenters in a dataset with large tract sizes, because the number of contiguous tracts with the necessary minimum employment density will be large. The same cutoff points may produce few subcenters in a dataset with small tracts, because many sites have little or no employment. The appropriate cutoff points need to be chosen somewhat arbitrarily by relying on local knowledge of what appears reasonable.

The two-stage procedure is less arbitrary and helps reduce the sensitivity of the results to the tract size. However, the results will always depend on the unit of analysis. At one extreme, we may have only two data points—the central city and its suburbs—which can at most produce one subcenter. Figure 9 shows

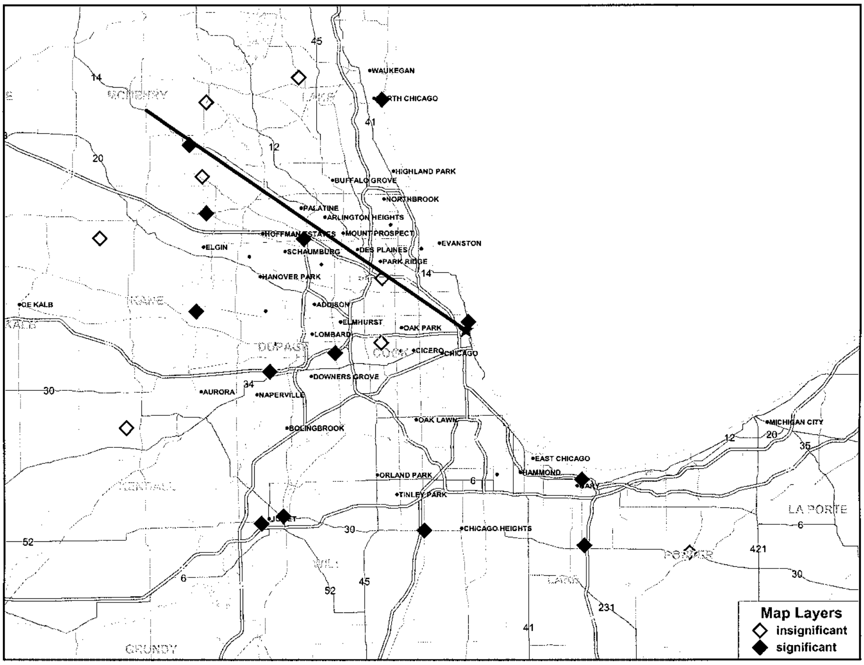

FIG. 9. Chicago—aggregated data.

the effects of aggregation on the two-stage procedure applied to Chicago (along with a ray from the CBD to the northwest, which is discussed next). Tracts are aggregated by combining the four closest tracts into one tract across the full dataset, which results in an average tract size of 1.49 square miles rather than 0.42 square miles for tracts with employment—a figure comparable to that for Los Angeles and San Francisco. The number of candidate subcenter sites falls from 63 to 23, and the number of significant sites falls from 33 to 13. Fewer small sites are identified in the western suburbs, while large sites remain significant in Gary, Schaumburg, North Chicago, and along the East–West Tollway. Nearly all of the significant subcenters in Figure 9 are in the same locations as found using the disaggregated data. The exceptions are subcenters near Hoffman Estates, northeast of Downers Grove, and along the expressway west of Chicago Heights. All of these subcenters are near sites identified using the disaggregated data.

The effects of aggregation are also seen in Figures 10 and 11, which show predicted employment densities from the second-stage semiparametric regressions for sites along the ray running northwest from the Chicago CBD. The ray runs near O'Hare Airport, Des Plaines, and Palatine and passes through the Crystal Lake subcenter in McHenry County. The estimates are similar using

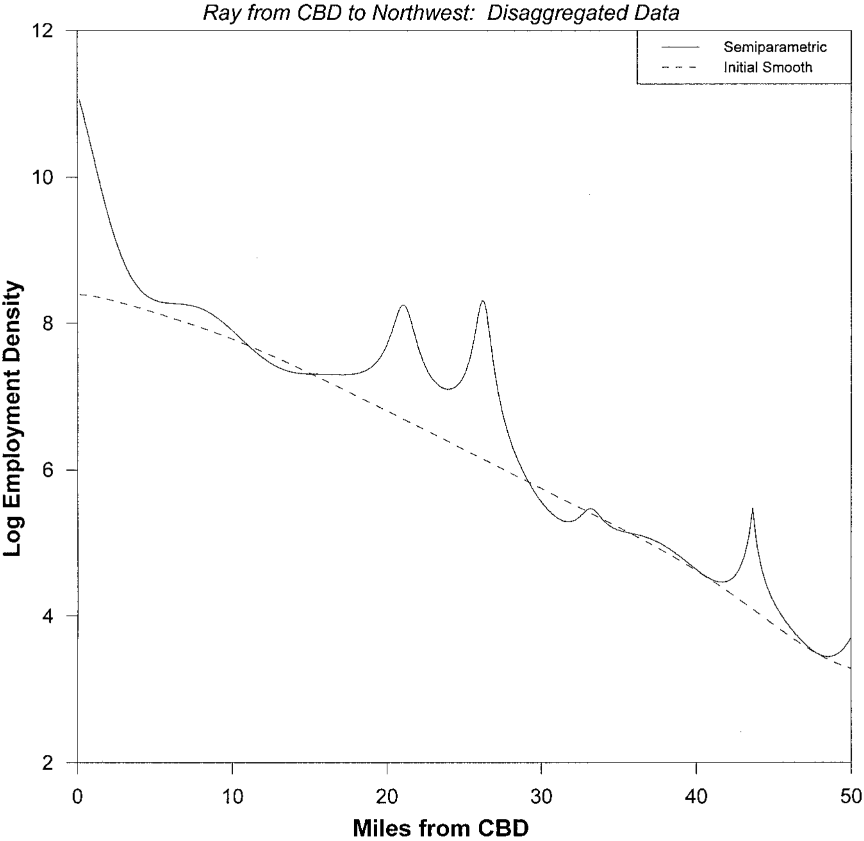

FIG. 10. Ray from CBD to Northwest: Disaggregated data.

aggregated and disaggregated data, but naturally the aggregated estimates are smoother. Both sets of estimates detect a sharp rise about 44 miles from the CBD in Crystal Lake, but the disaggregated data find two local peaks between 20 and 30 miles from the CBD rather than one local peak. These rises reflect the influence of two subcenters in the disaggregated dataset, whereas the two subcenters (near Palatine and Mt. Prospect) are merged into one subcenter when the data are combined into larger tracts.

The initial LWR-smoothed estimates are also presented in Figures 10 and 11. Note that the LWR estimates are concave, pronouncedly so in the case of the aggregated data. A simple exponential function would impose linearity on the estimates. The usefulness of the initial smooth is shown in these figures. The sharp peaks near subcenters produce large positive residuals that, as subsequent analysis demonstrates, lead to local rises in estimated log-densities around the subcenter sites.

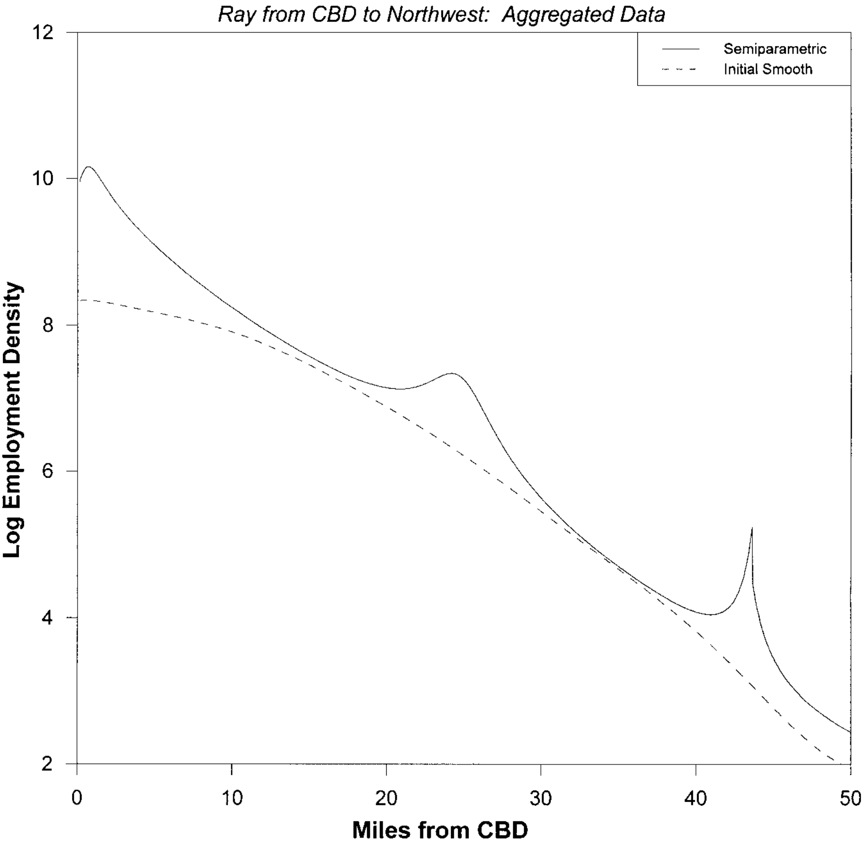

FIG. 11. Ray from CBD to Northwest: Aggregated data.

7. THE EFFECTS OF SUBCENTERS ON POPULATION AND EMPLOYMENT DENSITIES

The objective of this section is to determine the extent to which a variable measuring proximity to subcenters improves the fit of standard exponential population and employment density function estimates. The simple exponential function is $y_i = \beta_0 + \beta_1 DCBD_i + u_i$, where y_i is the natural logarithm of either population or employment density; these results were presented in Table 2. To model the effects of multiple subcenters, I follow Shukla and Waddell [47] and use a gravity variable to represent the effects of proximity to subcenters. The estimating equation becomes $y_i = \beta_0 + \beta_1 DCBD_i + \beta_2 Gravity_i + u_i$, and a positive value for β_2 implies that proximity to subcenters increases log-density.

TABLE 3
Gravity Variable Regressions

	Chicago	Dallas	Houston	Los Angeles	New Orleans	San Francisco
Log population density						
Constant	8.236	−18.964	−6.909	4.333	7.882	−4.724
	(99.363)	(23.201)	(5.807)	(7.882)	(29.549)	(7.187)
DCBD	−0.086	0.118	0.033	−0.027	−0.099	0.044
	(56.887)	(20.419)	(3.657)	(8.085)	(10.585)	(9.304)
Gravity	7.095	17.925	10.129	2.904	7.400	7.635
	(22.470)	(33.439)	(12.765)	(9.621)	(3.426)	(21.736)
R^2	0.410	0.323	0.321	0.131	0.313	0.306
Observations	8755	4636	2282	2984	416	2832
α	1.0	0.25	0.25	0.25	0.5	0.25
Log employment density						
Constant	7.055	−32.818	−20.122	2.951	5.006	−12.859
	(79.948)	(33.972)	(14.947)	(4.488)	(8.292)	(17.807)
DCBD	−0.099	0.168	0.089	−0.032	−0.216	0.087
	(56.107)	(25.651)	(8.317)	(8.044)	(19.181)	(16.504)
Gravity	8.299	26.492	18.567	3.013	18.868	11.311
	(23.564)	(41.830)	(20.803)	(8.350)	(5.566)	(29.408)
R^2	0.419	0.421	0.462	0.114	0.601	0.349
Observations	6665	4351	2103	3032	450	2891
α	1.0	0.25	0.25	0.25	0.25	0.25

Note: Absolute t values are in parentheses.

The gravity variable for observation i is

$$Gravity_i = \sum_{j=1}^{S} \frac{\hat{f}(x_j)}{D_{ij}^{\alpha}}, \tag{3}$$

where S is the number of subcenters in the metropolitan area, D_{ij} is the distance between observation i and subcenter j, α is a parameter representing the decay rate, and $\hat{f}(x_j)$ is the estimated density of the observations at the subcenter site.[8] Areas with small average tract sizes in the vicinity of site j have high values for $\hat{f}(x_j)$. Thus the term "density" represents the number of nearby observations, rather than population or employment density. Weighting the terms in the gravity variable by their estimated densities places less weight on proximity to more remote subcenters, which have large tract sizes and are more likely to be near tracts with no employment. I let α vary from 0.25 to 3.0 in increments of 0.25 and choose the one that leads to the lowest residual sum of squares for each

[8] $\hat{f}(x_j) = h^{-2}n^{-1}\sum_{i=1}^{n} \varphi((x_{1i} - x_{1j})/h)\varphi((x_{2i} - x_{2j})/h)$, where x_1 is distance east and x_2 is distance north of the city center. See the Appendix for details on the estimation procedure.

regression. Equation (3) is a convenient way to summarize the effects of many variables.

The results, presented in Table 3, show a fairly consistent pattern across the six cities. In every city, employment and population densities rise near subcenters, even after controlling for distance from the CBD. Nevertheless, distance from the CBD is an important determinant of densities in each city. The large t values on the gravity variables and the corresponding increases in R^2's across Tables 2 and 3 is evidence of the importance of treating cities as polycentric when estimating population and employment density functions.

In every case, the addition of the gravity variable results in an increase in the estimated coefficient for *DCBD*, which suggests that densities do not decrease as rapidly with distance from the CBD as is implied in Table 2. The reason for this result is that most subcenters are relatively close to the CBD, so that the correlation between *DCBD* and *Gravity* is negative. An increase in *DCBD* tends to lower densities, but it also leads to reduced access to subcenters, which further lowers densities. Thus, failing to include *Gravity* in the estimating equation leads to a downward bias in the coefficient on *DCBD*. In three cities—Dallas, Houston, and San Francisco—including *Gravity* in the regressions causes the coefficients on *DCBD* to turn *positive*. In these cities, the traditional CBD is no longer the critical determinant of the broad spatial trend in densities. A more realistic specification would be to treat the CBD as simply another of the multiple centers in these metropolitan areas.

8. CONCLUSION

This paper proposes a two-stage nonparametric procedure for identifying urban employment subcenters. The first stage uses a nonparametric procedure, locally weighted regression (LWR), to identify candidate subcenter sites. The second stage uses a semiparametric regression procedure to determine whether the subcenters have significant local effects on the overall employment density surface. My procedure adapts locally to departures from the monocentric city's featureless plain, is less sensitive to the unit of measurement than most existing procedures, is readily reproducible by other researchers, and can be applied to metropolitan areas that the researcher does not known well. It produces reasonable results for Chicago, Dallas, Houston, Los Angeles, New Orleans, and San Francisco. The results using the gravity variable representing proximity to subcenters demonstrate the importance of modeling cities as polycentric. Employment and population densities tend to fall with distance from the CBD but rise near subcenters.

The objective is to develop a procedure that can identify subcenters in many cities, including cities that the researcher does not know well. Local knowledge can improve the estimates. If the two-stage procedure fails to identify a site that is expected to be important, then distance from the site can be added to the second-stage Fourier regressions. If the procedure identifies exceedingly small sites as subcenters, then these can be deleted from the second stage. At the same time, the two-stage procedure can help identify unexpected sites that turn out to be important determinants of densities.

The two-stage procedure treats subcenters as a statistical concept. Subcenters are sites for which proximity is an important determinant of employment density. In this context, "important" is defined by standard hypothesis-testing procedures. This concept is already implicit in the literature. Gordon *et al.* [25] find 57 candidates subcenters in Los Angeles, which they narrow to 6. Giuliano and Small [24] find 32 subcenters. I find 19. As we have seen, part of the difference is a result of different time periods. But it is also clear that the results will vary when tract sizes differ, when distant tracts are omitted from the analysis, and when cutoff points and definitions of proximity differ. An important advantage of the two-stage procedure is that it can be implemented easily using standard regression procedures for a variety of cities, tract sizes, and time periods.

APPENDIX

The starting point for nonparametric estimation is the expression $y_i = m(x_i) + u_i$, where x_i is a set of explanatory variables. Using a first-order expansion of $E(y_i \mid x = x_i) = m(x_i)$ around x, we have $m(x_i) = m(x) + \frac{\partial m}{\partial x}(x^*)(x_i - x)$, where x^* lies between x_i and x. After letting $\beta(x^*) \equiv \frac{\partial m}{\partial x}(x^*)$ and $m(x) \equiv m$, the objective function for the LWR estimator is

$$\sum_{i=1}^{n}(y_i - m - \beta(x^*)(x_i - x))^2 K_i, \tag{A1}$$

which is minimized with respect to m and $\beta(x^*)$. K_i determines the weight given to observation i in forming the prediction at x. K_i is large when an observation is close to the target point. Equation (A1) produces the standard weighted least squares (WLS) regression estimator; the estimates $\hat{\beta}(x^*)$ are the coefficients from a regression of $K_i^{1/2} y_i$ on $K_i^{1/2}$ and $K_i^{1/2}(x_i - x)$, while $\hat{m}(x)$ is the predicted value at point x. Although the target point for estimation, x, is any arbitrary value, it can also represent actual data points. Separate WLS regressions are then calculated for each observation to calculate $\hat{y}_j = \hat{m}(x_j)$ for $j = 1, \ldots, n$.

The Nadaraya–Watson kernel estimator is a special case of Eq. (A1), with $\beta(x^*) = 0$. The explanatory variables help produce more accurate estimates. In identifying subcenters, I use distances east (x_1) and north (x_2) of the CBD as explanatory variables,

$$y_i = m + \beta_1(x^*)(x_{1i} - x_1) + \beta_2(x^*)(x_{2i} - x_2),$$

where x_{1i} and x_{2i} are the distance for observation i, x_1 and x_2 are the target points, and x^* is the expansion point.

In standard nonparametric models, K_i is a function of the distance between x_i and x. In spatial models, distance has a natural geographic definition and K_i is a function of the distance between observation i and the target point. Let $d_i(x)$ represent this distance and order the observations such that $d_1(x) < d_2(x) < \cdots < d_n(x)$. Although any standard kernel can be used for the weight function, K_i, the tricube has been popular in spatial models:

$$K_i = \left(1 - \left(\frac{d_i(x)}{d_q(x)}\right)^3\right)^3 I(d_i(x) < d_q(x)), \tag{A2}$$

where $I(\cdot)$ is an indicator function that equals 1 when the condition is true. All observations beyond the window of the q nearest sites are given zero weight in the estimation. Within the window, closer observations receive more weight than distant ones.

The kernel choice is less important in nonparametric estmation than the choice of window size. Fortunately, the window size choice is not critical when the objective is to identify subcenters. The key is to use a relatively large window. To see the advantages of nonparametric modeling, consider the alternative of simple OLS estimation. Following McDonald [31], candidate subcenters are those sites with significant positive residuals from a regression of y_i on $DCBD_i$. If the employment density gradient is larger on one side of the city, simple OLS estimation may lead to positive residuals on the side with the larger gradient. The excessive smoothing can lead to significantly positive residuals even when there is no local rise in density. By adjusting for local variations in the density gradient, the LWR estimator will indicate different gradients on the two sides of the city.

An extremely small window size can produce a nearly perfect fit, leading to an absence of significant residuals even though a subcenter has caused a local rise in density. Moderate smoothing provides a base: After taking into account broad spatial trends, has a local rise in density led to positive residuals in an area of the city? I use a window size of 50% of the available observations, which produces a moderate amount of smoothing.

Following Pagan and Ullah [40], the variance of the predicted value of log-employment density at location x is the first diagonal element of the following expression:

$$\sigma(x)^2 \left(\sum_{i=1}^{n} z_i w_i z_i'\right)^{-1} \left(\sum_{i=1}^{n} z_i w_i^2 z_i'\right)\left(\sum_{i=1}^{n} z_i w_i z_i'\right)^{-1}, \tag{A3}$$

where $z_i = (1, x_{1i} - x_1, x_{2i} - x_2)'$ and $w_i \equiv K_i^{1/2}$. I estimate $\sigma(x)^2$ as the predicated value from a kernel regression of $(y_i - \hat{m}(x))^2$ on x_1 and x_2. Specifically, $\hat{\sigma}^2(x) = \sum_i \omega_i e_i^2 / \sum_i \omega_i$, where $\omega_i = h^{-2}\varphi((x_{1i} - x_1)/h)\varphi((x_{2i} - x_2)/h)$, $\varphi(\cdot)$ is the standard normal density function, and $e_i^2 = (y_i - \hat{m}(x))^2$. After normalizing x_1 and x_2 to have unit variances, I choose the bandwidth, h, using the following rule: $h = 0.8^{1/6} n^{-1/6}$ (Pagan and Ullah [40]). This formulation accounts for heteroscedasticity and allows the variance to differ across space. Let $s^2(x)$ represent the estimated variance of $\hat{m}(x)$. The variance of the residual at x is $s^2(x) + \hat{\sigma}^2(x)$.

REFERENCES

1. A. Anas, R. Arnott, and K. A. Small, Urban spatial structure, *Journal of Economic Literature*, **36**, 1426–1464 (1998).
2. J. E. Anderson, Cubic spline urban density functions, *Journal of Urban Economics*, **12**, 155–167 (1982).
3. J. E. Anderson, The changing structure of a city: Temporal changes in cubic spline density patterns, *Journal of Regional Science*, **25**, 413–425 (1985).
4. W. A. Barnett, J. Powell, and G. Tauchen (Eds.), "Nonparametric and Semiparametric Methods in Econometrics and Statistics," Cambridge University Press, New York (1991).
5. B. Bender and H. Hwang, Hedonic housing price indices and secondary employment centers, *Journal of Urban Economics*, **17**, 90–107 (1985).
6. W. T. Bogart and W. C. Ferry, Employment centres in Greater Cleveland: Evidence of evolution in a formerly monocentric city, *Urban Studies*, **36**, 2099–2110 (1999).
7. C. Brunsdon, A. S. Fotheringham, and M. E. Charlton, Geographically weighted regression, *Geographical Analysis*, **28**, 281–298 (1996).
8. R. Cervero and K. Wu, Polycentrism, commuting, and residential location in the San Francisco Bay area, *Environment and Planning A*, **29**, 865–886 (1997).
9. R. Cervero and K. Wu, Subcentering and commuting: Evidence from the San Francisco Bay area, 1980–90, *Urban Studies*, **35**, 1059–1076 (1998).
10. W. S. Cleveland, Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association*, **74**, 829–836 (1979).
11. W. S. Cleveland and S. J. Devlin, Locally weighted regression: An approach to regression analysis by local fitting, *Journal of the American Statistical Association*, **83**, 596–610 (1988).
12. S. G. Craig and J. E. Kohlhase, "Employment Subcenters and the Distribution of Economic Activity," unpublished manuscript, University of Houston (1999).
13. S. G. Craig, J. E. Kohlhase, and S. C. Pitts, "The Impact of Land Use Restrictions in a Multicentric City," unpublished manuscript, University of Houston (1996).
14. S. G. Craig and P. Ng, Using quantile smoothing splines to identify employment subcenters in a multicentric urban area, *Journal of Urban Economics* **49**, 100–120 (2001).
15. D. E. Dowall and P. A. Treffeisen, Spatial transformation in cities of the developing world: Multinucleation and land-capital substitution in Bogotá, Colombia, *Regional Science and Urban Economics*, **21**, 201–224 (1991).
16. R. T. Dunphy, Defining regional employment centers in an urban area, *Transportation Research Record*, **861**, 13–15 (1982).
17. J. Fan, Design adaptive nonparametric regression, *Journal of the American Statistical Association*, **87**, 998–1004 (1992).
18. J. Fan, Local linear smoothers and their minimax efficiencies, *Annals of Statistics*, **21**, 196–216 (1993).
19. J. Fan and I. Gijbels, Variable bandwidth and local linear regression smoothers, *Annals of Statistics*, **20**, 2008–2036 (1992).
20. A. R. Gallant, On the bias in flexible functional forms and an essentially unbiased form: The Fourier flexible form, *Journal of Econometrics*, **15**, 211–245 (1981).
21. A. R. Gallant, Unbiased determination of production technologies, *Journal of Econometrics*, **20**, 285–323 (1982).

22. A. R. Gallant, Identification and consistency in seminonparametric regression, *in* “Advances in Econometrics: Fifth World Congress,” Vol. 1 (T. F. Bewley, Ed.), Cambridge University Press, New York (1987).
23. J. Garreau, “Edge City,” Doubleday, New York (1991).
24. G. Giuliano and K. A. Small, Subcenters in the Los Angeles region, *Regional Science and Urban Economics*, **21**, 163–182 (1991).
25. P. Gordon, H. W. Richardson, and H. L. Wong, The distribution of population and employment in a polycentric city: The case of Los Angeles, *Environment and Planning A*, **18**, 161–173 (1986).
26. D. Greene, Recent trends in urban spatial structure, *Growth and Change*, **10**, 29–40 (1980).
27. D. A. Griffith, Modelling urban population density in a multi-centered city, *Journal of Urban Economics*, **9**, 298–310 (1981).
28. W. Härdle, “Applied Nonparametric Regression,” Cambridge University Press, New York (1990).
29. W. Härdle and O. Linton, Applied nonparametric methods, *in* “Handbook of Econometrics,” Vol. IV (R. F. Engle and D. L. McFadden, Eds.), Elsevier, New York (1994).
30. E. Heikkila, P. Gordon, J. I. Kim, R. B. Peiser, and H. W. Richardson, What happened to the CBD-distance gradient?: Land values in a policentric city, *Environment and Planning A*, **21**, 221–232 (1989).
31. J. F. McDonald, The identification of urban employment subcenters, *Journal of Urban Economics*, **21**, 242–258 (1987).
32. J. F. McDonald and D. P. McMillen, Employment subcenters and land values in a polycentric urban area: The case of Chicago, *Environment and Planning A*, **22**, 1561–1574 (1990).
33. J. F. McDonald and P. J. Prather, Suburban employment centers: The case of Chicago, *Urban Studies*, **31**, 201–218 (1994).
34. D. P. McMillen, One hundred fifty years of land values in Chicago: A nonparametric approach, *Journal of Urban Economics*, **40**, 100–124 (1996).
35. D. P. McMillen and J. F. McDonald, A nonparametric analysis of employment density in a polycentric city, *Journal of Regional Science*, **37**, 591–612 (1997).
36. D. P. McMillen and J. F. McDonald, Suburban subcenters and employment density in metropolitan Chicago, *Journal of Urban Economics*, **43**, 157–180 (1998).
37. D. P. McMillen and J. F. McDonald, Population density in suburban Chicago: A bid-rent approach, *Urban Studies*, **35**, 1119–1130 (1998).
38. R. Meese and N. Wallace, Nonparametric estimation of dynamic hedonic price models and the construction of residential housing price indices, *Journal of the American Real Estate and Urban Economics Association*, **19**, 308–332 (1991).
39. É. A. Nadaraya, On estimating regression, *Theory of Probability and its Applications*, **9**, 141–142 (1964).
40. A. Pagan and A. Ullah, “Nonparametric Econometrics,” Cambridge University Press, New York (1999).
41. A. R. Pagan and Y. S. Hong, Nonparametric estimation and the risk premium, *in* “Nonparametric and Semiparametric Methods in Econometrics and Statistics” (W. A. Barnett, J. Powell, and G. E. Tauchen, Eds.), Cambridge University Press, New York (1991).
42. A. D. Pavlov, Space-varying regression coefficients: A semi-parametric approach applied to real estate markets, *Real Estate Economics*, **28**, 249–283 (2000).
43. H. W. Richardson, P. Gordon, M. Jun, E. Heikkila, R. Peiser, and D. Dale-Johnson, Residential property values, the CBD, and multiple nodes: Further analysis, *Environment and Planning A*, **22**, 829–833 (1990).
44. P. M. Robinson, Root-N-consistent semiparametric regression, *Econometrica*, **56**, 931–954 (1988).

45. D. Ruppert and M. P. Wand, Multivariate locally weighted least squares, *Annals of Statistics*, **22**, 1346–1370 (1994).

46. G. Schwarz, Estimating the dimension of a model, *Annals of Statistics*, **6**, 461–464 (1978).

47. V. Shukla and P. Waddell, Firm location and land use in discrete urban space: A study of the spatial structure of Dallas-Fort Worth, *Regional Science and Urban Economics*, **21**, 225–253 (1991).

48. K. A. Small and S. Song, Population and employment densities: Structure and change, *Journal of Urban Economics*, **36**, 292–313 (1994).

49. C. J. Stone, Consistent nonparametric regression, *Annals of Statistics*, **5**, 595–645 (1977).

50. G. S. Watson, Smooth regression analysis, *Sankhya, Series A*, **26**, 359–372 (1964).

51. F. Yuming and T. Somerville, Site density restrictions: Measurement and empirical analysis, *Journal of Urban Economics*, **49**, 404–423 (2001).