Using Quantile Smoothing Splines to Identify Employment Subcenters in a Multicentric Urban Area *

Steven G. Craig
Department of Economics
University of Houston
Houston, TX 77204-5882

Pin T. Ng
Department of Economics
University of Illinois
Champaign, IL 61820

Revised April 13, 2000

SUGGESTED RUNNING HEAD: Employment Density Quantile Functions

Abstract

The goal of this paper is to provide a statistically based definition of employment subcenters for multicentric urban areas. In particular, we examine the shape of the employment density function using quantile smoothing splines as a nonparametric empirical specification. This approach allows inspection of the employment gradient at the upper tail rather than the center of the employment density distribution. As a result, our definition of employment subcenters extends previous work as it allows us to condition on distance from the central business district, relies on the extent to which a subcenter influences surrounding areas, yet still emphasizes areas with high employment densities.

*We benefited from the helpful comments of two referees, the editor, participants at the 1996 South West Econometric Meetings, at the Texas Econometrics Camp, and from Janet Kohlhase and Dan McMillan. Craig also acknowledges the financial support of the U.S. Department of Education and the Center for Public Policy at the University of Houston. All the computation in this paper was performed on the Sun Ultra 2 in the Econometric Lab at the University of Illinois supported by NFS Grant SBR-9617206. All opinions expressed are the authors' alone, and any remaining errors are ours.

1 Introduction

Defining whether an employment subcenter exists is key to discussions of whether, and the extent to which, American cities have become multicentric. The literature is beginning to coalesce on a definition of an employment subcenter that focuses on the gross employment density function. Despite this progress, however, there is still a considerable degree of subjectivity pertaining to location specific information required to actually identify employment subcenters. This paper introduces *employment density quantile functions*, which will be defined more precisely in Section 2 as conditional quantile functions of log gross employment density given a specific distance from the central business district (CBD), to develop a statistically based definition of an employment subcenter. The employment density quantile functions are estimated nonparametrically by *employment density quantile splines* using the quantile smoothing splines of Koenker, Ng and Portnoy [14].

The advantage of our methodology is that it meets the criteria that have been discussed in the literature for a subcenter definition, and is not subject to some of the problems that have plagued previous attempts at a definition (McDonald [16]; Alperovich and Deutsch [1]; Small and Song, [24]; McMillen and McDonald [19]). In particular, the quantile spline focuses on the highest density tracts, and thus allows the densest tracts to define the employment subcenters. We thus can significantly reduce one of the problems of using a traditional regression based method to identify employment centers, which is that regressions identify a concentric circle rather than a point.[1] A specific advantage of using a regression method is that we are able to focus on employment density conditional on the distance from the CBD, rather than simply an absolute level of employment density. Using quantile splines shows that neither the monocentric city assumption nor the exponential functional form of

[1] As explained below, we still need to pick a point on a concentric circle, but it is likely to be a data point that determines the upper quantile regression line.

the employment density function is viable for our data, which is for Houston, Texas. Instead, the quantile spline methodology allows us to identify the multiple employment centers in the Houston metropolitan area in a statistically verifiable manner.

McDonald [16] has a clear discussion of criteria desirable to the definition of an employment subcenter. He advocates use of the employment density function rather than the population density function to identify employment subcenters[2]. While this criterion has been generally accepted, it has often been applied to simply raw employment density data regardless of other relevant covariates such as distance to the central business district; see for example Giuliano and Small [9] for Chicago, and Small and Song [24] for the Los Angeles area. A further problem with utilizing solely raw employment densities is that determining the number of subcenters is based on arbitrary size and density cut-offs that cannot be applied to other regions, and that are not verified by any economic relationship. McDonald [16], and McDonald and McMillen [18] extend the subcenter definition by comparing raw employment densities to those in surrounding areas. Their goal of determining whether an employment concentration influences surrounding areas appears to capture the core economic relationship that defines an employment subcenter. The shortfall of their approach, however, is that identification of local employment peaks has not taken into account distance from the primary employment center – the CBD. Recently, Alperovich and Deutsch [1], McMillen and McDonald [19], and McMillen [17] attempt to use an estimated employment density function to identify employment subcenters. This process extends the earlier work by allowing employment densities to be conditioned on the distance from the CBD. Their method, however, relies on the central tendency of the data. We posit, alternatively, that if it is desired to examine employment peaks, it is better to examine them directly by investigating the upper tail of the employment density distribution rather than attempting

[2]This is because the purpose of identifying employment subcenters is to examine the distribution of population in a multicentric area.

to infer peaks from the center of the data. This can be done exactly with the help of the employment density quantile splines.[3]

Our methodology for identifying employment subcenters is to use changes in the gradients of upper employment density quantile splines to define the employment subcenters. In particular, an employment subcenter should be expected to serve as an attraction to surrounding areas. Thus we look for an employment gradient greater than what would be expected from a smoothly declining density quantile spline, and use variation in the gradient to define the location of an employment subcenter.[4] Employment density quantile splines allow us to focus directly on the high density observations without regard to the central tendency of the data. We believe our methodology allows a more precise definition of employment subcenters that can be translated to alternative areas, or that does not rely on specific knowledge of an area. Further, our method avoids some of the problems of having to arbitrarily define the size and exact location of an employment subcenter, as the peak in the employment density gradient will be relatively precisely defined.

We should note we do not exactly avoid all arbitrariness. The gradient break defines a concentric circle, and we still need to pick an exact census tract at the distance defined by the gradient break. Using the upper quantile spline, however, greatly increases the chances the chosen point is one that influences the top quantile spline, and so is consistent with finding an employment concentration.[5] A related problem is the exact definition of a gradient break. While one would naturally think an employment subcenter would cause a positive gradient in

[3]Clearly if the employment density function is homoskedastic, the upper quantiles and the median (or mean) quantiles will contain the same information and will be equally useful at identifying the employment subcenters. Any heteroskedasticity in the data, however, will yield a substantial advantage to the quantile spline methodology.

[4]In our case we use a positive gradient, a break from the expected negative gradient, to define a subcenter location. Alternatively, however, any decrease in the negative gradient could be used as a definition.

[5]Conversely, a standard regression method may essentially ignore the densest observations, as they are outliers compared to the other data. The top quantile splines, such as the 95th percentile, on the other hand rely on the densest observations to define the density function (see below).

the density function due to its influence on surrounding areas, the likelihood of this depends on the size and number of observations (such as whether the observations are Census tracts, traffic analysis zones, or zip codes). The larger the area of each observation, the less likely is a positive gradient to be found.[6]

One attribute of our quantile functions approach is that we will not be using traditional conditional *mean* regression via a least squares procedure. Instead, we will utilize the conditional *quantile* (percentile) regression (with *median* as a special case), both to provide consistent estimates under a variety of conditions, and to allow us to explore the upper tail of the employment density distribution. It is well known that conditional mean regression provides an optimal (minimum variance unbiased) estimate of the true regression function when the stochastic error follows a Gaussian (normal) process. A conditional median regression, on the other hand, is more efficient (smaller variance) when the error is generated by a thick-tailed distribution disseminated through the form of outlying observations. While mean and median regressions have their own merits and drawbacks under different error distributions as estimates of the conditional central tendency, the conditional quantile regression is a more complete approach to analyzing the employment density. Rather than simply providing a description of the central tendencies in the data, quantile regression allows us to explore the relationship at the tails of the distribution. This innovation is central to understanding employment subcenters since by definition employment centers are at the upper tail of the employment density distribution. In such cases, estimates that focus on the conditional mean or median would overlook important features that are apparent in a more general conditional quantile analysis. [7]

[6] Useful data, however, is an important determinant of the effect on surrounding areas. If the unit of observation is so large to obscure such influence, it may be the data is insufficient rather than the methodology flaw. Nonetheless, as noted above, a subcenter ring could be identified by a decrease in the negative gradient rather than by a positive gradient if the unit of observation is relatively large.

[7] While regression quantiles are often used in bio-medical studies (e.g. percentile infants growth curves), their potential has yet to be fully appreciated in economics. Some of the more recent applications in economics

A further unique characteristic of the employment density quantile spline is that it avoids restrictions in the assumed production relationships inherent in the traditional negative exponential or other parametric functional specifications. There is a trade-off between a parametric and a nonparametric functional form specification, which involves the trade-off between bias and variance in the estimation. A misspecified functional form, derived or implied from an incorrect production relationship, will lead to bias in the coefficient estimates. Nonparametric specification on the other hand introduces higher variance because of the lack of imposed structure. With our fairly large data set the increase in variance can be substantially ameliorated, and we believe is a small price to pay for the potentially significant gain in bias reduction.

2 Analytical Framework

Specification of the employment density function plays a central role to all of the recent developments in identifying the number and location of employment centers in a metropolitan area (Alperovich and Deutsch [1]; Small and Song [24]; McMillen and McDonald [19]; McMillen [17]). The conventional specification is that each realization d_i of the gross employment density, given a specific distance x_i from the CBD, is generated by the following process:

$$h(d_i) = g(x_i) + \epsilon_i. \tag{1}$$

There are many alternate ways to specify the functional forms in (1). A relatively general parametric specification is the Box-Cox transformation function $h(d_i) = (d_i^{\gamma} - 1)/\gamma$ with $0 \leq \gamma \leq 1$. $g(\cdot)$ is the functional describing how distance is related to the Box-Cox transformed gross employment density, and ϵ_i is a realized random noise. It is easy to see that $h(d_i) = \ln(d_i)$ when $\gamma = 0$, and $h(d_i) = d_i$ for $\gamma = 1$.

are Hendricks and Koenker [13], Buchinsky [4], Chamberlain [5], Manning, Blumberg and Moulton [15], Goldberg [10], and Fitzenberger [8].

The traditional specification for the employment density function is the negative exponential where $g(x_i)$ is linear in x_i and $h(d_i)$ is the logarithmic transformation. For any other intermediate value of γ, the Box-Cox transformation provides a more flexible parametric specification where the model is nonlinear in γ. Even a nonlinear specification, however, such as the Box-Cox transformation, imposes constraints on the functionals. Given the complexity of models that generate multicentric areas, however, it is unclear whether the imposed structure is consistent with the underlying theoretical models.

Nonparametric regression models on the other hand impose minimal constraints on $g(\cdot)$, primarily in the form of continuity restrictions on higher order derivatives. Typically $g(\cdot)$ is assumed to be a smooth function belonging to a Sobolev space with bounded derivatives up to a certain order. The dimension of the parameters in a nonparametric regression can increase with the number of observations, thus decreasing the structure imposed on the specification. In this study, we first perform a logarithmic transformation on gross employment density to remove most of the asymmetry in the response variable. Our response y_i (logarithm of gross employment density) is then modelled as a nonparametric function $g(\cdot)$ of x_i as

$$y_i = g(x_i) + \varepsilon_i$$

In particular, we will look at the non-median quantiles to discern the shape of the density function as it pertains to the highest density areas, which presumably will most closely capture the employment subcenters affecting the distribution of employment throughout the city.

2.1 Employment Density Quantile Functions

Our method therefore is to estimate the upper quantile employment density function of log employment density conditioned on a given distance from the CBD. This methods allows us the advantages discussed above; our definition of employment centers will depend on the

high employment areas, will be conditioned on distance to the CBD, and will depend on the degree to which one area influences surrounding areas. Absent homoskedasticity, this information is not available through perusal of the conditional mean or median function of log gross employment density.

Formally, the τ-th (100τ-th) employment density quantile (percentile) function, $g_\tau(x)$, is a function of x such that

$$\tau = \int_{-\infty}^{g_\tau(x)} f_{Y|X}(y|x)\, dy$$

where $f_{Y|X}(y|x)$ is the conditional probability density function of Y (logarithmic of gross employment density) given $X = x$ (distance to CBD) and τ is any real number in the interval $[0, 1]$. The conditional τ-th quantile function dichotomizes the response variable space such that 100τ percent of the response variable falls below $g_\tau(x)$ and $100(1-\tau)$ percent above it. Examining the extreme conditional quantile functions for τ close to one enables us to analyze the higher employment density areas of a city at all given distances from CBD. Note that the conditional 50-th percentile (median) function divides the response space into two equal halves and provides an alternative conditional central tendency measure for the response variable to the conditional mean. If ϵ_i is independent and identically distributed, $g_\tau(x) = g(x) + F_\epsilon^{-1}(\tau)$, where $F_\epsilon(\epsilon)$ is the cumulative distribution function of the error. If ϵ is symmetric at zero, $g(x)$ becomes the conditional median function.

2.2 Employment Density Quantile Splines

Estimation of the employment density quantile functions is essential for examining the upper tail of the employment density distribution, as the feature most desired in the specification is the ability to focus on the tails. In this section, we explain our use of nonparametric quantile smoothing splines, shown by Koenker, Ng and Portnoy [14] to provide a consistent nonparametric estimate of $g_\tau(x)$. In particular, the smoothing spline allows a much more

flexible functional shape than a parametric function, while at the same time allowing us to examine the estimated extreme quantiles of the employment density.

The most common form of nonparametric estimation of density functions (albeit for population density) is the cubic spline (Anderson [2]; Anderson [3]; Zheng [25]). As we demonstrate below, however, the standard method of estimation using the cubic spline involves least squares and provides an estimate of the conditional mean of the employment density. Recent alternatives are the locally weighted regression technique introduced in McMillen and McDonald [19] and the two-step nearest neighbor approach introduced in McMillen [17]. These estimation techniques also involve estimates of the central tendency of the employment density function, although they do not use the entire data set for determining any point of the function. The quantile smoothing spline methodology described here provide estimates of not only the conditional median but all the employment density quantile functions and allows us to focus on the positive employment outliers. Further, the quantile smoothing splines can be computed relatively easily using linear programming methods.

Quantile smoothing spline requires specification of a smoothing parameter λ which controls the relation between "roughness" and "fidelity". Roughness (or smoothness) essentially reflects the degree of "wiggleness" of the estimated fit while fidelity measures the goodness-of-fit of the estimated function. If all of the estimation weight is on fidelity, the resulting function would simply go through each data point, and would not provide a summary picture of the data. On the other hand, traditional parametric estimation places a high weight on smoothness, at the potential expense of an appropriately flexible description of the data. The quantile smoothing spline allows the trade-off between fidelity and roughness to be made explicitly via the tuning parameter λ.

Given n pairs of observations $\{(x_i, y_i)\}_{i=1}^n$ with $a < x_1 < \cdots < x_n < b$, the τ-th L_p

quantile smoothing spline (employment density quantile spline), $\hat{g}_{\tau,L_p}(x)$, is the solution to

$$\min_{g \in \mathcal{G}_p} \text{“fidelity”} + \lambda \text{“}L_p \text{ roughness”}$$

where $\mathcal{G}_p$ is some space of smooth functions g, and τ-th refers to the 100τ-th percentile of interest. Fidelity, indicating goodness-of-fit, is described by:

$$\text{“fidelity”} = \sum_{i \in \{i | y_i - g(x_i) \geq 0\}} \tau \left(y_i - g\left(x_i\right)\right) + \sum_{i \in \{i | y_i - g(x_i) < 0\}} \left(\tau - 1\right) \left(y_i - g\left(x_i\right)\right)$$

It assigns a weight of τ to positive residuals and $(\tau - 1)$ to the negative ones. The special case of median ($\tau = .5$) smoothing spline will have fidelity measured by the following absolute (L_1) norm:

$$\text{“fidelity”} = \sum_{i=1}^{n} \left|y_i - g\left(x_i\right)\right|$$

The roughness of the fit can be defined along an entire spectrum of the L_p norm for $1 < p < \infty$. We examine two extreme alternatives for $p = 1$ and ∞. For $p = 1$, we have

$$\text{“}L_1 \text{ roughness”} = \sum_{i=1}^{n-1} \left|g'\left(x_{i+1}\right) - g'\left(x_i\right)\right|$$

which uses the total variation of the first derivatives of the fit as a measure of roughness. For $p = \infty$, we have

$$\text{“}L_\infty \text{ roughness”} = \max_x \left|g''\left(x\right)\right|$$

which uses the sharpest curvature of the fit over the entire domain of the covariate (distance to CBD) as a roughness measure. The smoothing parameter $0 < \lambda < \infty$ balances the fidelity to the data and the roughness (or smoothness) of the fit. Koenker, Ng and Portnoy [14] show that $\hat{g}_{\tau,L_1}$ is a linear smoothing spline (piecewise linear function, termed L_1) while $\hat{g}_{\tau,L_\infty}$ is a quadratic spline (piecewise quadratic function, termed L_∞). They recommend a Schwarz-type information criterion for choosing the smoothing parameter. Efficient computation of both via linear program is described in Ng [21] while consistency results can be found in

Portnoy [22]. We estimate both an L_1 and an L_∞ version of the employment density quantile splines using the S-plus implementation provided in He and Ng [11].

The quantile smoothing spline has two advantages over the cubic smoothing spline. As discussed above, the quantile spline can be used to estimate any given percentile in the distribution, not solely the central tendency of the data. Secondly, the quantile spline can be computed efficiently with linear programming techniques. The cubic smoothing spline usually solves the following optimization problem

$$\min_g \sum_{i=1}^{n} \left(y_i - g\left(x_i\right)\right)^2 + \lambda \int \left(g''\left(x\right)\right)^2 dx \tag{2}$$

where λ is again the smoothing parameter that controls the trade-off between fidelity to the data measured by the first term and the roughness measured by the second term. As is clear in the fidelity term (the first term), the resulting cubic spline uses least squared errors criterion for estimation. If the distribution is Gaussian, it provides an efficient estimate of the conditional mean function, but not information about the tails. If, alternatively, the absolute error is used (so there is no squared term in the fidelity term of the equation), the optimization problem will have to be solved by quadratic programming methods, a computationally much more complex methodology, albeit the resulting estimate will be efficient under a Cauchy (or double exponential) error distribution.

As a comparison, in addition to reporting the employment density quantile splines, we report cubic spline estimation of the employment density function below. To allow more flexibility, we select the smoothing parameter both by cross-validation, and by Craven and Wahba's [7] generalized cross-validation for the cubic splines.[8]

[8]Both methods for choosing the knots and smoothing parameter are more general, and allow more flexibility, than is the equally spaced knots method (Anderson [2]; Anderson [3]; Zheng [25]).

3 Data

We estimate employment density quantile functions for Harris County, the county containing Houston, Texas. The unit of observation is a census tract, of which there are 578.[9] Gross employment is measured as the location of work using the Journey to Work survey from the Census, which depends on resident-reported job locations.[10] Employment is therefore measured by the work locations of a random sample of individuals responding to the Census. Gross employment density is measured per acre, and distance as miles from the CBD. While land use data are available, we follow McDonald [16] and estimate the gross employment density functions. Further, Mieszkowski and Smith [20] find that gross density functions better describe the Houston area.[11]

Houston is an excellent area for examining the existence of alternative employment centers. Except in the eastern segment of the city with the Houston Ship Channel, there are few natural geographic features that affect the urban form. Further, land use restrictions in Houston are less than elsewhere, as most zoning power is granted to individual residential neighborhoods, or to enclave small cities, leaving large segments of land use in the urban area unrestricted. Finally, Houston has experienced relatively rapid growth over the last thirty years, so its shape is less affected by history than are many other areas.

4 Results: Locations of Employment Subcenters

Figure 1 presents the L_1 (piecewise linear) median employment density spline along with several of the upper tail (τ =.75, .90, .95) quantile splines for the Houston area. The raw data points are also displayed. The marks at the bottom of the figure are the locations of

[9] We have dropped Census tracts with zero employment. The quantile spline methodology we use does not require deletion of negative outliers, but taking the logarithm of zero is numerically inconvenient.

[10] We have also estimated the density function using data from Dun and Bradstreet employment data (reported by employers), and the results are virtually identical.

[11] Although their work is for population rather than employment.

the knots for the quantile splines. The 50th percentile spline is in the middle of the figure progressing up to the 95th percentile spline at the top. The piecewise linear estimation procedure characterized by the L_1 spline is evident in the linear segments of the fitted lines. We utilize the estimated 95th percentile employment density function to identify the employment subcenters.[12] As discussed earlier, this process identifies subcenters by exploiting the tracts with the densest employment.[13]

A highly noticeable feature in the figure is the heterogeneity of the error distribution. Since each of the splines traces out a portion of the conditional distribution function (measured vertically so that the 95th percentile spline has about 95% of the observations below it at any given distance from the CBD), the splines for the different percentiles should be essentially parallel if the errors are independently and identically distributed at all distances from the CBD. Instead, variance of the employment density is found to vary considerably, even when excluding very large distances from the CBD where the number of observations becomes quite small. Further, variance in the employment density does not appear to vary systematically with distance from the CBD, so there are no obvious corrections for the heteroskedasticity problem.

Also apparent in the figure is the highly nonlinear structure of the fitted quantiles, indicating that a perfectly smooth curve would miss much of the variation in employment density. Thus the nonparametric employment density quantile splines appear to offer some advantages over parametric estimation. Most important is that the central tendency of the data, even including the median estimates (the 50th quantile), is not especially useful at describing the shape of the density function in the upper tail where we expect to observe

[12] We also tried the 97.5th percentile spline and the results are qualitatively identical.

[13] While not exact (since the 95th percentile tract is estimated for every distance from the CBD), in some sense this process identifies 29 tracts (5% of the 578 observations) as candidates for employment subcenters. As discussed below, we then pick the densest tracts at the points indicated by the gradient of the density function to select the actual subcenters.

employment subcenters.

4.1 Concentric Subcenter Rings

Each panel in the upper row of Figure 2 reproduces the corresponding quantile estimate presented in Figure 1. The first from the left shows the employment density for tracts in the 95th percentile of the estimated distribution. That is, holding distance from the CBD constant, the line shows estimated employment density for the 5% densest tracts. The second panel shows the 90th percentile, and on down so the last panel illustrates the estimated median employment density function. The shaded regions illustrate the 90% confidence bands of the estimated quantiles using the studentized approach described in He and Ng [11].

The bottom row shows the gradients corresponding to each of the estimated quantiles. If the fitted function in the top panel was the traditional negative exponential, the gradient in the bottom panels would be flat, and negative. In the multicentric context, we instead expect to see the gradient become less negative at an employment subcenter, representing the employment concentration and its effect on surrounding areas. An advantage of the quantile splines is that we do not need to rely on means; instead we are able to concentrate on the high employment density areas. A further advantage, however, is not all high employment density areas are identified as a subcenter, which would be the case using raw employment density data. The estimated quantile splines shown in Figure 2, in contrast, show the usual employment density pattern, where the employment density variably rises and falls over distance. In other words, it appears some of the high density areas affect the surrounding tracts much more than others do. Thus we believe the quantile spline method is able to more fully exploit the factors that cause an area to function as an urban employment subcenter, as illustrated in the left most portion (the 95th percentile) of the panels in Figure 2.

The 95th quantile spline identifies employment subcenters because it contains all the in-

formation from the definition discussed in the introduction. Because it is the 95th quantile, the spline uses the top 5% densest tracts. It also conditions on distance from the CBD. Finally, and most crucially, the spline illustrates the extent to which one area affects the employment densities in other areas. In particular, the gradients in the bottom panel illustrate the influence of an employment subcenter over other dense employment areas by showing where employment density rises (or stops falling) even as distance from the CBD increases. A local peak of the employment density quantile spline occurs when the density quantile spline turns down so that the gradient again becomes negative. We identify subcenters as the points where the gradient of the quantile spline, presented in the bottom panels, become negative after having been positive. A less restrictive definition would not necessarily have to depend on a positive gradient, but simply one less negative. For the Houston data, identification of an employment subcenter is not sensitive to the definition of the absolute level of the gradient before an area is selected as a subcenter.[14]

Since we are looking at distances from the CBD, the locations of local peaks in the extreme quantile (95th percentile) characterize subcenters as rings. Because Houston closely resembles a flat featureless plane (with the exception of the eastern side of the city, where the Houston Ship Channel is an economically important geographic feature), subcenter rings may be appropriate like those depicted in Sasaki [23]. Figure 3 is a picture of such subcenter rings. However, we proceed a step further and use the densest tracts along a concentric circle (subcenter ring) to identify the actual employment subcenters. As discussed earlier, this is consistent with using the 95th percentile spline.[15]

[14] As discussed above, it may be that a more reasonable definition of an employment subcenter is simply that the gradient be "surprisingly large," or rising, even if negative. The actual level of the gradient can be affected in our procedure by small changes in the smoothing parameter, but identification of the subcenters is fairly robust to this parameter. Further, for different cities, identification of subcenters depends on the size of each observation, where larger areas would generally be expected to show less influence elsewhere.

[15] An interesting unresolved issue, however, is why the dense tracts along a ring are not necessarily clustered together. Thus there is still a subjective element in our identification of a specific subcenter, and the number of subcenters is not apparent by our methodology.

4.2 Identifying Potential Subcenters

Three subcenter rings containing employment subcenters are evident by examining the gradients in the first panel of the bottom row of Figure 2, corresponding to the 95th percentile spline. The first subcenter ring occurs at six miles from the CBD. This is the distance from the CBD to Interstate 610, the closest freeway that loops around the CBD. The two densest tracts, just outside the Loop, at this distance from the CBD are the Galleria and the tract just south of it. Figure 3 indicates the Galleria area (represented by $\Box$) on a map of Houston as the area just adjacent to the inner highway Loop 610 on the western side of the city. This is a retail and office center, referred to by some real estate people as "Uptown." The CBD is indicated by $\diamond$ at the center of the figure.

The second subcenter ring containing employment subcenters occurs at about thirteen miles from the CBD. The densest tracts along this ring occur in three separate areas about equidistant from the CBD. Using the thirteen mile ring, one center on the eastern side of the city is Pasadena (indicated as $\times$), located near the Houston Ship Channel. The Ship Channel area is the location of a large portion of the city's petrochemical processing center, and is actually a line stretching from within about two miles from the CBD to over twenty-three miles.[16] On the northern side of the city, the Greenspoint area (denoted as $\triangle$) is a subcenter. This area is at the confluence of highway I-45, which heads north from the CBD, and the second circumnavigational freeway (Beltway 8). It is also just adjacent to the Intercontinental Airport in the north. Finally, on the western side of the city is an office and research area on the same circumnavigational highway, Beltway 8, and a major thoroughfare –Westheimer (indicated as + in Figure 3).

The third subcenter ring as indicated by the gradient panel is the region 20.5 miles from the CBD. The twenty-one mile concentric ring corresponds to the location of major suburban

[16] See Craig and Kohlhase [6] who examine whether there are independent subcenters around the Ship Channel, or instead whether the subcenter can be characterized as a line along the Ship Channel.

development. The single densest tract at this distance includes NASA ($\divideontimes$), on the south-eastern side of the city in the Clear Lake area. Two other employment subcenters on this ring are apparent, flanking the Ship Channel. The northern area is Baytown ($\boxtimes$), and the southern area is La Porte ($\bigtriangledown$). These are both chemical and industrial areas strongly linked to the Ship Channel[17].

To show the importance of our subcenters as employment magnets in their local areas, Figure 1 and Figure 3 also show the location of the tracts near our selected subcenters. That is, the dark triangles show tracts within one mile of the Galleria, the selected subcenter on the inner concentric circle. The dark circles indicate tracts that fall within two miles of the three selected subcenters on the second concentric ring, while the dark squares indicate tracts within three miles of the two selected subcenters on the outer concentric ring.[18] As can be seen from both Figure 1 and Figure 3, most of the near points are above the median quantile, indicating as expected that the subcenters appear generally to be in the middle of larger density areas.

The second subcenter ring appears slightly different than the other two. The estimated gradient is somewhat shallower, indicating that perhaps the subcenters on this ring are less concentrated than the others. That is, the employment density at the subcenters is not as different from surrounding areas. On the other hand, this middle subcenter ring exhibits a greater influence on surrounding areas than the other two rings. That is, the area affected appears to cover a much greater distance from the CBD, and the area affected by this ring covers a larger distance from each subcenter, than is apparent in the other subcenter rings.[19]

[17] It is interesting to speculate on the importance of the distance from the CBD of the concentric rings indicated by the quantile spline. One of the most rapidly growing counties in the U.S. is Fort Bend county, which is Houston's southwestern suburb located also at about 21 miles from the CBD. Major employment growth there would be expected to show another subcenter at this distance in the year 2000.

[18] We selected the distances arbitrarily, but generally tracts farther from the CBD have larger areas (since their population density is lower, and the Census attempts to keep relatively equal populations per tract).

[19] We leave for future research the very interesting question of what it is that determines the height of the gradient of each subcenter and the extent of the area affected by a subcenter. Our point here is that we need

Our statistical procedure has therefore identified seven employment subcenters at three different distances from the CBD. We believe several features of our procedure are central to the ability to identify these subcenters. One feature is that our examination is restricted to the higher percentile elements of the spline functions. Examination of the panels to the right of the first in Figure 2 shows a less pronounced definition of the locations of the subcenter rings, in that the regions where the increase in the gradient occurs are more diffuse, and eventually disappear as we move to the right panels down the percentile curves. The inability of the median spline in the farthest right panel to identify the employment subcenters dramatically illustrates the problem with using a measure of the central tendency to identify employment subcenters. This problem is especially vivid for the middle of the three subcenter rings, as this distance exhibits the smallest change in the density function gradient in the lower percentile curves.

Estimation of an employment density function adds a further element to understanding the location of employment subcenters. The locations selected by our procedure are not necessarily the densest unconditioned employment areas. For example, while the Galleria is the second densest employment area in the city (after the CBD), the next densest area identified as an employment subcenter is the eleven mile western center, which is ranked only 17th. One of the twenty-one mile employment subcenters is Baytown, which is ranked 114th. Thus conditioning on distance to the CBD before subcenters are selected yields a much different picture, but we believe a more accurate picture, of the urban area than does examination of simply the raw employment density data.

4.3 Alternative Estimation Methods

To illustrate the difference between our employment density quantile spline procedure and others, we compare the quantile spline results with those from two other methods for es-

a consistent definition of an employment subcenter before such research is really warranted.

timating the employment density function. First, we employ the cubic spline procedure, a nonparametric method which nonetheless relies on the central tendency of the data. Second, we estimate the most flexible of the parametric methods, the Box-Cox specification. Finally, we compare all of these results to the alternative method for estimating the quantile spline, the L_∞ method. We find that both the cubic spline and Box-Cox estimation methods do not select likely candidates for employment subcenters compared to the quantile spline results under either the L_1 or L_∞ methods of calculating smoothness.

For the cubic spline we estimated (2) using both all knots (located at all the unique values of x_i), and a subset of knots. We use both cross-validation and generalized cross-validation to choose the smoothing parameter. We report here cubic spline results using cross-validation because estimates from generalized cross-validation with all knots resulted in an extremely "rough" function, leading to obscured identification of employment subcenters. Figure 4 presents estimates from the cubic spline specification. As before, the top panels present the estimated employment density functions, while the bottom panels illustrate the associated gradients. We present three versions of the cubic spline, cross-validation with all knots, cross-validation with sub-knots and generalized cross-validation with sub-knots, respectively in Figure 4. As the figure illustrates, each of these methods yield virtually identical fitted splines. Using a positive change in the gradient to select the subcenter ring location reveals three estimated subcenter ring locations. The subcenter ring located nearest to the CBD is found to occur at about four miles. This choice essentially bisects two important employment areas, Greenway Plaza (at 4.5 miles) and the Medical Center (at 3.6 miles), without pinpointing a particularly known location. The third subcenter ring location is virtually identical to that found with the quantile spline, at about twenty and a half miles from the CBD. The radius of the second subcenter ring, however, is vastly different. This ring is found to occur at a distance of about nine miles from the CBD. The nine mile distance

is an odd choice, however, as it does not define a well known area, nor does it correspond to the transportation infrastructure. Thus apparently because there is little other information to obscure the picture, high employment areas in the outer portions of the city are able to influence the central tendency estimated by the cubic spline so the subcenter ring will be revealed. The inner two subcenter rings, however, are not found to pinpoint areas that correspond to actual employment concentrations.[20]

Following Alperovich and Deutsch [1], we also obtain parametric estimates of subcenter locations using the Box-Cox transformation. An important disadvantage of this procedure, however, is that while determining employment subcenters endogenously, it is also necessary to specify the form of the interaction among employment subcenters. In a multicentric urban employment center model with x_{ij} being the distance of the i-th observation to the j-th employment subcenter, the univariate functional $g(x_i)$ in (1) is usually replaced by a specific functional form.[21] The choices are a multiplicative form indicating subcenters are complements, a maximum specification which implies that all subcenters are completely substitutable, or an additive polycentric specification which is a mixture of both.

Using either the Schwarz or Akaike information criteria for model selection that estimates subcenter interaction, we find that a 7-subcenter model yields the best fit.[22] Only one of the seven estimated subcenters coincides with those obtained using the quantile splines, namely Westheimer at Beltway (+). Another estimated subcenter is close to the NASA ($\ast$) subcenter. The one closest to the CBD is near the Medical Center (about 3.6 miles from the actual CBD) while the other three remaining estimated subcenters are not near any con-

[20] We posit that the local averaging procedure in McMillan and MacDonald (1998) may have the same problem, where identification of the subcenters may be distorted because of the influence of neighboring observations during the process of local averaging.

[21] In fact, a disadvantage of the Box-Cox specification is that to determine the subcenter location endogenously, the form of the interaction between the subcenters must be specified. A preferred alternative would be to exogenously determine the employment centers, and then separately study how they interrelate.

[22] We choose to estimate the Box-Cox function with an additive specification which allows the widest type of subcenter interaction.

centration of employment. This may be a strong indication that even the flexible Box-Cox parametric specification cannot capture the highly nonparametric conditional mean function. Alternatively, this procedure requires a specification for how the subcenters interact, which may be premature if we do not have an objective method of identifying employment subcenters. We thus believe the subcenters as defined using the quantile splines more closely reflect employment concentrations in the Houston area.

Estimation results using the alternative L_∞ quantile spline method yield similar results to those reported in Figures 1 and 2 above for the L_1 splines. Figure 5 overlays the actual employment density data for the Harris County with the estimated conditional quantile functions using the L_∞ quantile spline method, in which case smoothing is attained by a piecewise quadratic procedure rather than the piecewise linear procedure of the L_1 splines. The heteroskedasticity apparent in the L_1 splines in Figure 1 is equally apparent here, even at low distances from the CBD. Figure 6 is analogous to Figure 2, in that it presents both the estimated splines and their confidence bands in the top row for a variety of quantiles, while the panels in the bottom row present the corresponding gradients of the estimated functions. All three subcenter rings are apparent at virtually the same distances from the CBD as with the piecewise linear methodology. Also, only the highest quantile splines identify the employment centers, and quantiles toward the middle of the probability distribution fail to identify the location, and perhaps even the existence, of the subcenters. Thus it does not appear that the exact methodology is central to pinpointing employment subcenters, but the methodological goals of examination of the tails of the density function appears key.

5 Summary and Conclusion

This paper offers a method for identifying employment subcenters in a multicentric urban environment. Identification of subcenters is difficult, in that the ideal definition would be

without reference to the other elements of urban form that would be expected to be influenced by employment subcenters. We believe our method comes close to achieving this ideal, as it captures all of the essential elements needed to define an employment subcenter in a polycentric environment. Our method examines high employment areas of the city, but conditions on distance to the CBD to determine areas that are "surprisingly dense" given the remainder of the employment distribution. We examine areas with the highest employment concentrations by utilizing quantile spline estimation methods that allow us to examine solely census tracts in the top 5% of the probability distribution for each distance from the CBD. And our method relies on the influence of an employment subcenter on surrounding areas. This is accomplished because we use the gradients of the fitted employment density functions to identify areas where the gradient becomes less negative, or even positive, compared to what would be expected without an employment subcenter. A subcenter ring is indicated only when the positive (or surprisingly large) gradient again turns down. We then proceed to identify specific locations on the concentric rings to select the precise locations of employment subcenters. We apply our method to Houston, and are able to identify seven employment subcenters at three different distances from the CBD. All seven centers appear to be key areas for employment, unlike subcenters identified by other methods including a cubic spline or Box-Cox transformation.

Our statistical estimation methodology has some particular advantages for use in estimating employment density functions. It makes explicit the trade-off between goodness-of-fit and roughness of the estimated function. We explore two methods for smoothing, piecewise linear and piecewise quadratic. Both methods are shown to identify the same seven employment subcenters by our criteria. More important, perhaps, is that our statistical methodology allows us to examine employment subcenters without necessarily involving estimation of the best fit of the central tendency of the employment density.

The quantile smoothing spline estimation method relies on a median, rather than mean, type of estimation procedure. The advantage this bestows is that the shape of the estimated central tendency of the employment density function is not affected by the location of outliers in the form of local subcenters. Conversely, least squares estimation methodologies allow employment subcenters, if they exist, to profoundly shape the mean estimated density function. Thus we are able to look at tracts that are most dense, and focus on their relationship to establish the existence of employment subcenters.

While we believe our method for identifying employment subcenters has important advantages compared to alternatives, we still do not have a fully complete method to describe a subcenter. In particular, our method identifies tracts with high absolute employment levels as well as high densities. And, because we condition on distance, neither absolute nor per area employment is a requirement for identification as a subcenter. On the other hand, it is tempting to discuss the width of the density gradient as a measure of a subcenter's economic influence. Our method is univariate, however, in that it only relies on distance from the CBD. A bivariate type of analysis of some sort, in which the direction as well as distance is included, is required to completely measure the influence of a specific subcenter on its surroundings if we are not to assume that influence is equal in all directions. The bivariate quantile smoothing splines recently introduced by He, Ng and Portnoy [12] may provide a promising venue.

One fascinating attribute of the Houston metropolitan area is that the employment subcenters are found on three concentric circles, corresponding for the two closer circles to the two major circumferential highways. It may be there is an element of the urban environment that indicates the "optimal" distance from the CBD for a subcenter, such as in Sasaki [23]. This will be an excellent topic for future research as soon as a set of agreed upon criteria can be expressed for locating employment subcenters in a multicentric urban area.

References

[1] G. Alperovich and J. Deutsch, Joint estimation of population density functions and the location of the central business district, *Journal of Urban Economics*, 36, 239–48 (1994).

[2] J. E. Anderson, Cubic-spline urban-density functions, *Journal of Urban Economics*, 12, 155-67 (1982).

[3] J. E. Anderson, The changing structure of a city: temporal changes in cubic-spline urban density patterns, *Journal of Regional Science*, 25, 413-25 (1985).

[4] M. Buchinsky, Changes in the U.S. wage structure 1963–1987: application of quantile regression, *Econometrica*, 62, 405–458 (1994).

[5] G. Chamberlain, Quantile regression, censoring and the structure of wages, in "Advances in Econometrics" (C. Sims, Ed.), Elsevier, New York (1994).

[6] S. Craig and J. Kohlhase, Employment subcenters and the distribution of economic activity, unpublished, Department of Economics, University of Houston.

[7] P. Craven and G. Wahba, Smoothing noisy data with spline functions, *Numerische Mathematik*, 31, 377–403 (1979).

[8] B. Fitzenberger, A guide to censored quantile regression, in "Handbook of Statistics" (G.S. Maddala and C.R. Rao, Ed.), Volume 15, 405–437, North-Holland, New York (1997).

[9] G. Giuliano and K. A. Small, Subcenters in the Los Angeles region, *Regional Science and Urban Economics*, 21, 163–182 (1991).

[10] P. K. Goldberg, Dealer price discrimination in new car purchases evidence from the consumer expenditure survey, *Journal of Political Economy*, 104, 622-654 (1996).

[11] X. He and P. Ng, COBS: qualitatively constrained smoothing via linear program, *Computational Statistics*, 14, 315–337 (1999).

[12] X. He, P. Ng and S. Portnoy, Bivariate quantile smoothing splines, *Journal of the Royal Statistical Society, B*, 60, 537–550 (1998).

[13] W. Hendricks and R. Koenker, Hierarchical spline models for conditional quantiles and the demand for electricity, *Journal of the American Statistical Association*, 87, 58–68 (1992).

[14] R. Koenker, P. Ng and S. Portnoy, Quantile smoothing splines, *Biometrika*, 81, 673–680 (1994).

[15] W. Manning, L. Blumberg and L.H. Moulton, The demand for alcohol: the differential response to price, *Journal of Health Economics*, 14, 123-148 (1995).

[16] J. F. McDonald, The identification of urban employment subcenters, *Journal of Urban Economics*, 21, 242–258 (1987).

[17] D. McMillen, Nonparametric employment subcenter identification, unpublished, Center for Urban Real Estate, University of Illinois at Chicago.

[18] J. F. McDonald and D. McMillen, Employment subcenters and land values in a polycentric urban area: the case of Chicago, *Environment and Planning A*, 22, 1561–1574 (1990).

[19] 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).

[20] P. Mieszkowski and B. Smith, Analyzing urban decentralization, *Regional Science and Urban Economics*, 21, 183–199 (1991).

[21] P. Ng, An algorithm for quantile smoothing splines, *Computational Statistics and Data Analysis*, 22, 99–118 (1995).

[22] S. Portnoy, Local asymptotics for quantile smoothing splines, *Annals of Statistics*, 25, 414–434 (1997).

[23] K. Sasaki, The establishment of a subcenter and urban spatial structure, *Environment and Planning A*, 22, 369–383 (1994).

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

[25] X. Zheng, Metropolitan spatial structure and its determinants: a case-study of Tokyo, *Urban Studies*, 28, 87-104 (1991).

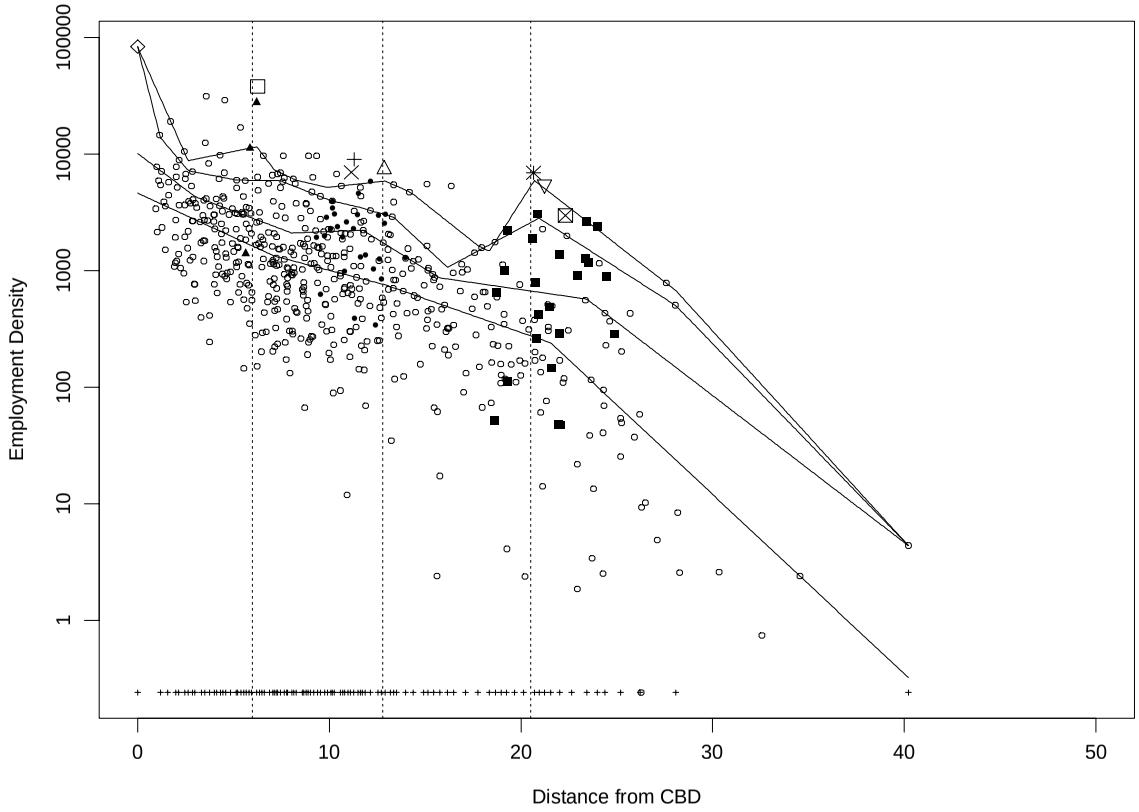

Figure 1. Various L_1 quantile splines for $\tau = .95, .90, .75, .50$ for the Harris County.

$\diamond$: CBD, $\square$: Galleria, $\times$: Pasadena, $+$: Westheimer at Beltway, Δ: Greenspoint, $*$: NASA, ∇: LaPorte, $\boxtimes$: Baytown.

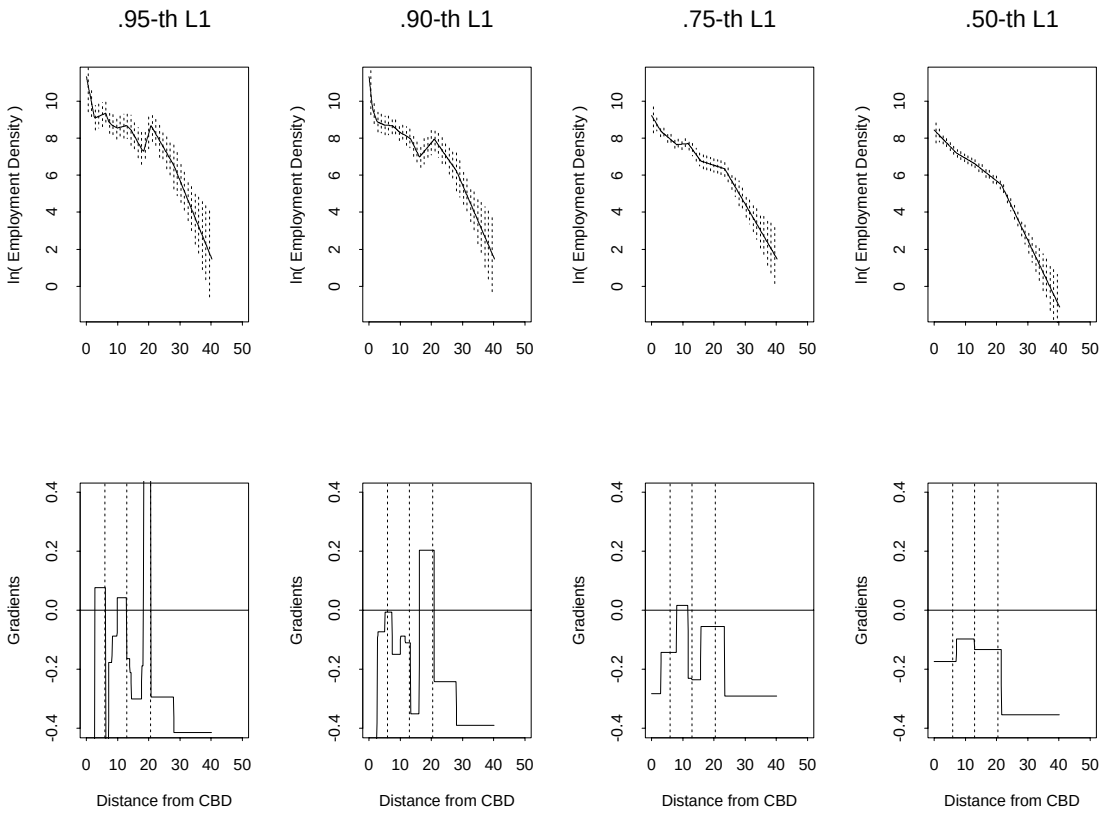

Figure 2. Top panels are the L_1 employment density quantile splines for Harris County and bottom panels are the corresponding gradients. The vertical dash lines in the bottom panels correspond to the radii (6 miles, 13 miles and 20.5 miles) of the subcenter rings.

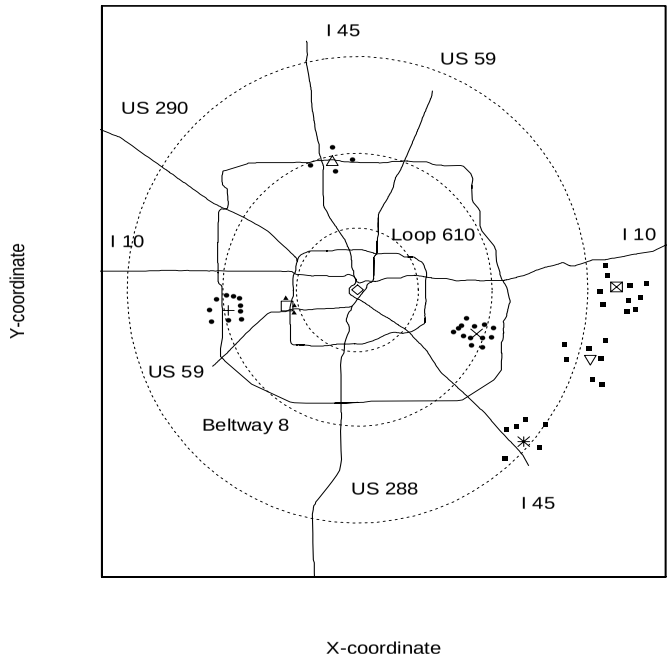

Figure 3. Locations of employment subcenters identified by the L_1 .95-th quantile spline. $\diamond$: CBD, $\square$: Galleria, $\times$: Pasadena, $+$: Westheimer at Beltway, Δ: Greenspoint, $*$: NASA, ∇: LaPorte, $\boxtimes$: Baytown.

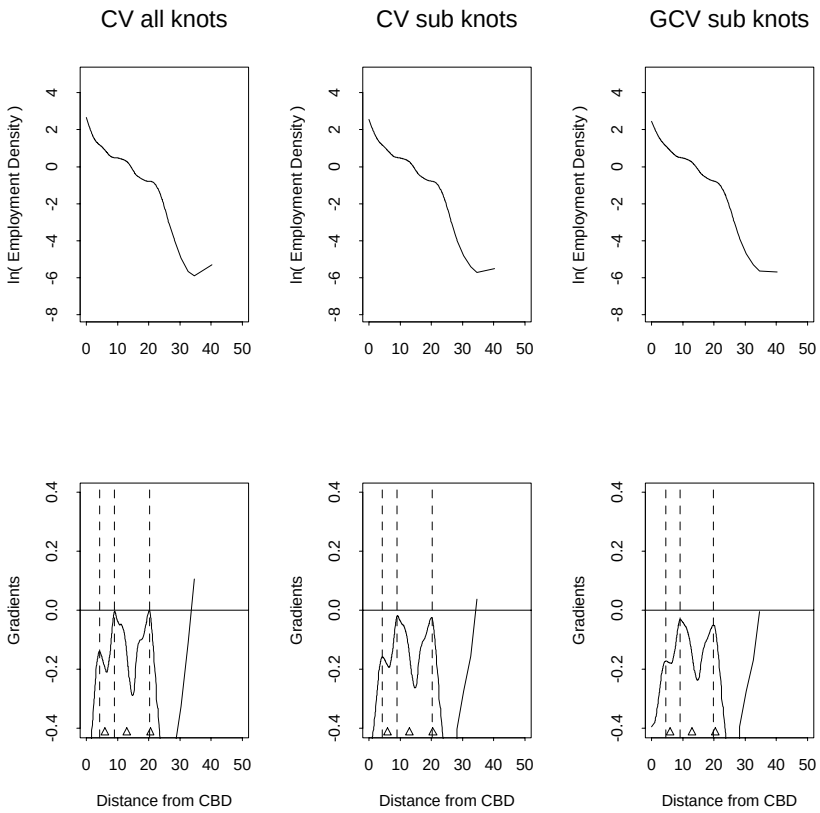

Figure 4. Top panels are the cubic smoothing splines for Harris County and bottom panels are the corresponding gradients. The vertical dash lines correspond to the distances of 4 miles, 9 miles and 20.5 miles.

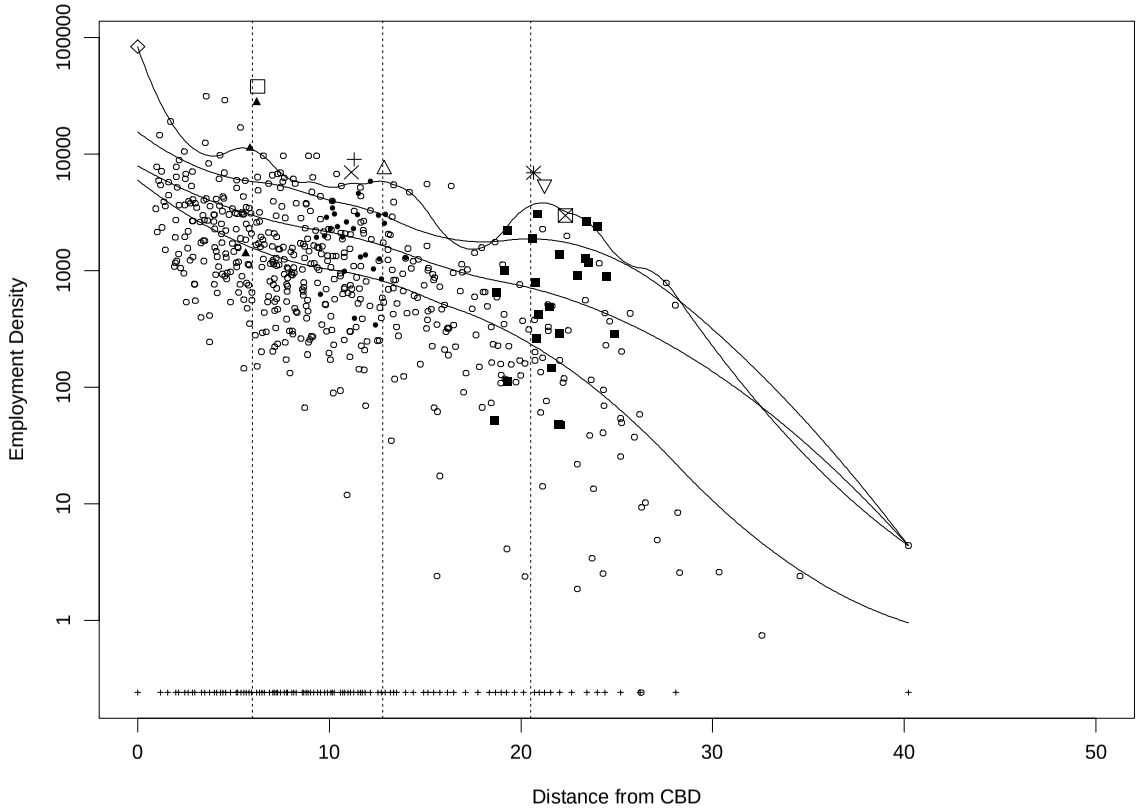

Figure 5. Various L_∞ quantile splines for $\tau = .95, .90, .75, .50$ for the Harris County.
$\diamond$: CBD, $\square$: Galleria, $\times$: Pasadena, $+$: Westheimer at Beltway, Δ: Greenspoint, $*$: NASA, ∇: LaPorte, $\boxtimes$: Baytown.

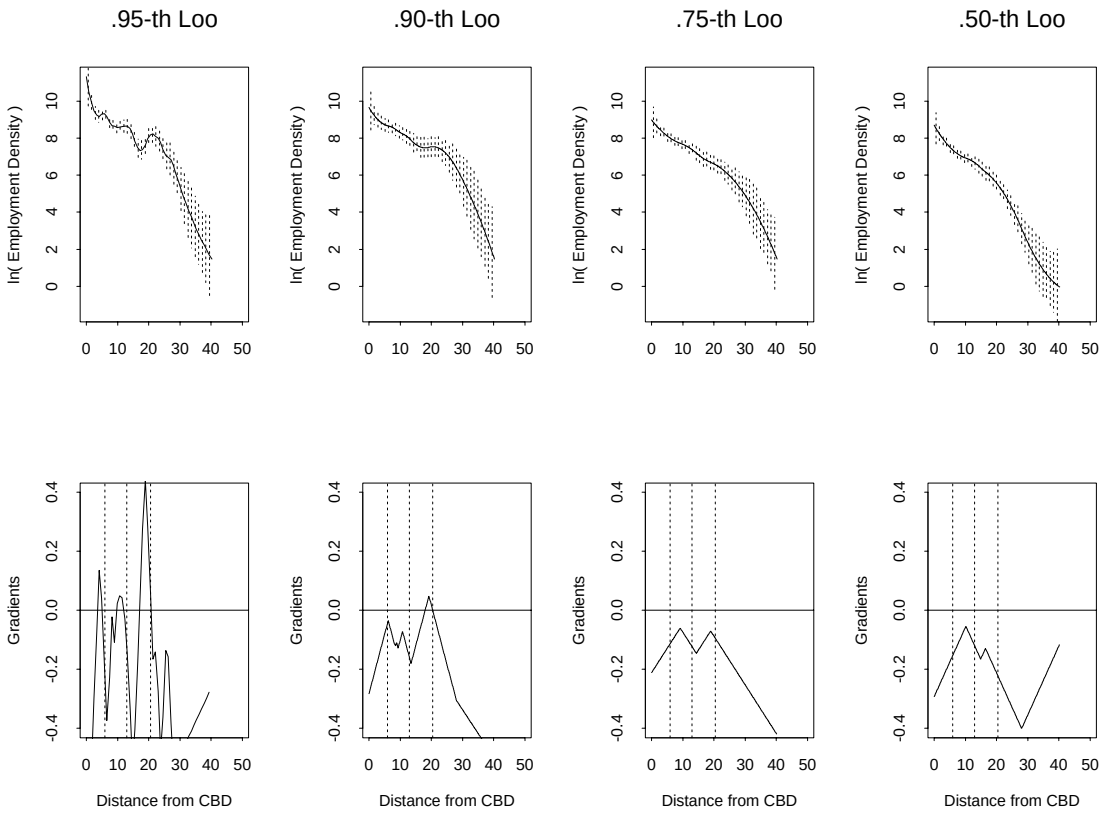

Figure 6. Top panels are the L_∞ quantile splines for Harris County and bottom panels are the corresponding gradients. The vertical dashed lines are again at distances 6 miles, 13 miles and 20.5 miles.