THE FEDERAL RESERVE BANK *of* KANSAS CITY
RESEARCH WORKING PAPERS

# How Centralized is U.S. Metropolitan Employment?

Jason P. Brown, Maeve Maloney, Jordan Rappaport, and Aaron Smalter Hall

# How Centralized is U.S. Metropolitan Employment?*

Jason P. Brown[1], Maeve Maloney[2], Jordan Rappaport[3], and Aaron Smalter Hall[4]

[1,3,4]Federal Reserve Bank of Kansas City
[2]Syracuse University

November 16, 2017

## Abstract

Centralized employment remains a benchmark stylization of metropolitan land use. To address its empirical relevance, we delineate "central employment zones" (CEZs)—central business districts together with nearby concentrated employment—for 183 metropolitan areas in 2000. To do so, we first subjectively classify which census tracts in a training sample of metros belong to their metro's CEZ and then use a learning algorithm to construct a function that predicts our judgment. Applying this prediction function to the full cross section of metros estimates the probability we would judge each census tract as belonging to its metro's CEZ. Using a high probability threshold for tract inclusion conservatively delineates a predicted CEZ for each metro. On average, the conservatively predicted CEZs account for only 12 percent of metropolitan employment in 2000. But the distribution of shares is positively skewed, with the conservatively predicted CEZs accounting for at least 20 percent of employment in 29 metros. Employment centralization is considerably higher for agglomerative occupations—those that arguably benefit most from face-to-face contact. The conservatively predicted CEZs account for at least 33 percent of agglomerative employment in 24 metros and at least 50 percent of legal employment in 79 metros.

**Keywords:** central business districts, employment density, metropolitan land use

**JEL Classification Numbers:** R12, R32, C45

---

# 1 Introduction

Centralized employment remains a benchmark stylization of metropolitan land use. In particular, a monocentric city with all employment taking place in the center of a circular area continues to be the workhorse model of urban economics (Alonso, 1964; Muth, 1969; Mills, 1967). In contrast, the majority of employment in almost all U.S. metro areas takes place outside a narrowly defined central location, both in multiple non-central clusters and spread diffusely (McMillen and Smith, 2003). Such considerable non-central employment raises the question of whether the centralized stylization remains empirically relevant.

Employment's departure from centralization is hardly new. A wide range of urban service occupations have always complemented residential location. Then, beginning in the 1950s, less-complementary jobs began following people out to the suburbs. Even so, Baum-Snow (2014) finds that the share of urban jobs that shifted to suburbs from the 1950s through 1990s was only one third the share of residents that shifted there. Moreover, the shift to the suburbs of jobs likely to benefit from agglomerative spillovers—such as in finance, insurance, and real estate—was minimal. Consistent with this, Brinkman (2016) shows that employment density in 2000 in a number of illustrative metros continued to decline sharply moving away from a central location.

The degree of employment centralization importantly affects metropolitan welfare. Jacobs (1969) and Gleaser (2011) argue that the cramming of individuals, occupations, and industries into close quarters allows for ideas to flow quickly from person to person, fostering learning and innovation. Consistent with such intensity, Glaeser and Maré (2001) find that wages in large U.S. metropolitan areas are about one third higher than wages in non-metropolitan locations. About half of this urban wage premium is likely to arise from agglomerative spillovers, which appear to be confined to a radius of just five miles (Combes, Duranton and Gobillion, 2008; Rosenthal and Strange, 2003, 2008). Exemplifying this interaction within close quarters, Arzaghi and Henderson (2008) document the extremely rapid spatial decay of spillovers and networking among advertising agencies in southern Manhat-

tan. Separately, Brinkman, Coen-Priani and Seig (2016) find that firms located in Central Business Districts (CBDs) tend to be larger and more productive than firms located elsewhere in metros. Similarly, Limehouse and McCormick (2011) find that law firms located in CBDs tend to be higher quality than law firms elsewhere in metros. And Rappaport (2017) finds that population growth from 2000 to 2015, both in the city and suburban portions of metros, was stronger in metros that had more centralized employment.

A challenge to addressing the empirical relevance of centralized employment is that there are no agreed-upon geographic delineations of where in each metro employment qualifies as central. Many empirical papers continue to use a subjective delineation of CBDs made for the 1982 Census of Retail Trade (e.g., Baum-Snow and Hartley (2016)). Some papers subjectively delineate CBDs for one or a handful of metro areas (Limehouse and McCormick, 2011; Brinkman, Coen-Priani and Seig, 2016). Other papers delineate the CBD using various algorithms, described below. And a slew of empirical papers do not state the CBD delineation they use.

The absence of an agreed-upon delineation of central employment also impedes a broader research agenda. A shared delineation would complement empirical research on a range of urban economics topics including agglomeration, land use, house prices, migration, spatial sorting, traffic congestion, and time use. More broadly, Google Scholar indexed more than 5,000 papers that were newly-written or revised in 2016 that included the term "central business district."

In this paper, we use a machine learning algorithm to delineate "central employment zones" (CEZs)—an enlargement of CBDs to include nearby concentrated employment— for 183 U.S. metropolitan areas in 2000. To do so, we first subjectively classify which census tracts in a training sample of metros belong to their metro's CEZ. The learning algorithm uses these classifications, along with hundreds of variables describing census tract characteristics, to construct a function estimating the probability we would judge census tracts as belonging to their metro's CEZ. We apply this function to the full cross section

3

of metros and use a high probability threshold for tract inclusion to delineate a predicted CEZ for each. On average, the resulting conservatively-predicted CEZs accounted for only 12 percent of metropolitan employment in 2000. But the distribution of CEZ employment shares is positively skewed, with the conservatively-predicted CEZs accounting for at least 20 percent of employment in 29 metros. Employment centralization was considerably higher for agglomerative occupations—those that arguably benefit most from face-to-face contact. The conservatively-predicted CEZs accounted for at least 33 percent of agglomerative employment in 24 metros and at least 50 percent of legal employment in 79 metros.

## 2    Defining Centrality

A prerequisite to delineating locations of centralized metropolitan employment is defining a theoretical conception of centrality. Metropolitan areas throughout much of the twentieth century, typically thought of as "cities", were conceived as having a Central Business District (CBD); the "principal commercial and retail district, forming the nucleus of the city" (Burgess, 1925). The CBD was "the region of heaviest concentration of buildings and economic activity within a city... almost exclusively of commercial, financial, retail, and service establishments...the region of greatest employment per unit land and few residences... the hub of the intracity transportation" (Muth, 1969, pp. 3-4). To guide delineation by local committees, the U.S. Census Bureau defined CBDs as "areas of high land valuation; areas characterized by a high concentration of retail businesses, offices, theaters, hotels, and service businesses, [and] areas of high traffic flow"(U.S. Bureau of the Census, 1987). More generally, CBDs were thought of as "downtown."

Updating this conception, we define central business districts as follows:

A **central business district (CBD)** is the largest cluster of relatively dense employment within a metropolitan area that is relatively accessible to a large share of the metro's workforce. Typically, it will have better transport links to locations throughout a metropolitan area compared to those of other employment clusters. Typically, a disproportionate share of its employment will be in occupations that benefit from proximity to other workers.

This definition allows for considerable flexibility in delineating actual CBDs, both in terms of the specific parcels of land included in them and in terms of distinguishing among multiple possible vicinities in which the CBD is located. For example, it allows the CBDs of smaller metros to have lower minimum employment density than the CBDs of larger ones, consistent with the Burgess and Muth definitions. Conversely, it allows for the CBDs of larger metros to be accessible to a smaller share of the metropolitan workforce than the CBDs of smaller ones. The definition deliberately avoids ambiguous judgments. In particular, it is agnostic on whether a CBD can be composed of portions that are nearby but not contiguous.

Rather than resolving narrow ambiguities, we define central employment zones:

A **central employment zone (CEZ)** is the combination of a central business district and nearby concentrated employment.

This broader conception makes sense in the context of distinguishing centralized dense employment from clusters of dense employment located further away from the CBD and from employment that is spread diffusely throughout a metropolitan area. "Midtown" locations in many modern metropolitan areas, a few miles from narrowly-conceived CBDs, are almost equally central and allow for short transit times to interact in person with downtown workers.[1] Indeed, one interpretation of CEZs is that they are equivalent to broader interpretations of CBDs such as in Holian and Kahn (2012).

Variations in employment density unambiguously identify the possible vicinities of each metro area's CEZ. For all metros we have looked at, one such vicinity overwhelmingly dom-

---

[1]Centrally-located employment also commonly takes place in wholesaling, shipping, and small manufacturing establishments at the periphery of CBDs (Muth, 1969). We exclude these from Central Employment Zones for pragmatic reasons, the learning algorithm's difficulty in identifying centrally-located tracts characterized by such employment. From a theoretical perspective, excluding such tracts may make sense to the extent that employment within them is less subject to agglomerative benefits.

inates others without having to set explicit criteria for what constitutes relatively high employment density, relatively high accessibility, a large share of a metro's workforce, and the tradeoff among these required to select a single CEZ. In contrast, more subjectivity is typically required to judge whether specific census tracts in the vicinity of the CEZ actually belong to it.

# 3    Methodology

The Census Bureau began delineating central business districts for its 1954 economic censuses, responding to demand for data on retail activity, and regularly refreshed the delineations through the 1982 economic censuses (U.S. Bureau of the Census, 1987). Local committees, representing a variety of interest groups from the central cities of metropolitan statistical ares, designated the census tracts they judged as belonging to their city's CBD. For the 1982 economic censuses, the committees delineated 456 CBDs for 455 cities in 315 metropolitan statistical areas (New York City was allowed to delineate two CBDs: one each in Manhattan and Brooklyn).

Several concerns suggest not using the 1982 Census CBDs to measure employment centralization. One is that the implicit criteria used to delineate them surely varied across the hundreds of local committees. A second concern is that metropolitan areas with multiple central cities were delineated as having multiple CBDs. A third concern is that the emphasis on retail businesses in the 1982 census definition had become anachronistic by 2000.

The main alternative approach to locating centralized employment is to subjectively specify a measurable criteria that can be uniformly applied across metropolitan areas. In some cases the criteria are straight-forward rules of thumb. For example, some papers locate CBD centroids at the city hall of the principal central city of each metro (Asabere and Huffman, 1991; Atack and Margo, 1998; Schuetz et al., 2017). However in many large cities, city hall is located in a cluster of government buildings apart from private-sector employment.

Holian and Kahn (2012) think of centralized employment as all that is located within 5 miles of the centroid returned by Google Earth for the largest principal central city of each metro. The authors note that the Google centroids approximately match their subjective assessment of CBD locations for the many metros they checked. They also note that being even a mile off from the "true" CBD centroid is likely to only modestly affect measured downtown characteristics. However, it is not clear whether 5 miles is an appropriate cutoff distance. Implicit cutoff distances of CBDs from their centroid surely vary across metros and with respect to direction. Rappaport (2014) partly addresses this by limiting CBDs to tracts within 5 miles of the Google Earth centroid that have employment density of at least 8,000 workers per square mile. But the implicit minimum density thresholds of CBDs also surely vary across metros.

In other cases, the measurable criteria and application involve statistical analysis. Redfearn (2007) delineates employment centers in the Los Angeles metro, conceived of as concentrations of employment that are significantly more dense than employment in surrounding areas. He first fits a non-parametric employment density surface using only nearby census tracts so as to retain local fluctuations in actual density. The local maxima of this surface identify possible employment centers. An iterative procedure determines the boundaries of each center by minimizing the sum of squared residuals of actual from fitted employment density of included tracts plus the sum of squared residuals of actual from average employment density of excluded tracts. The procedure does not identify one of the centers as the CBD but could easily do so with an additional criterion, such as having the highest fitted density or the most employment. Of more concern for our purposes is that delineating a CBD solely based on a geographic break in density might divide a large central cluster of employment, a portion of which has high density and a portion of which has very high density. This concern is magnified for delineating our more broadly-conceived central employment zones, which explicitly allow for breaks in employment density.[2]

---

[2]Redfearn (2007) belongs to a larger literature identifying employment subcenters of metropolitan areas. Many of the papers in this literature use clustering techniques that combine parcels of land that each exceed

# Our Approach

We follow a five-step process to delineate central employment zones for 183 metropolitan areas. First, we divide metropolitan areas by population into groups, allowing for the possibility that different characteristics identify the CEZs for each group. Second, we subjectively label census tracts in a subset of metros from each group as either belonging or not belonging to its metro's CEZ. Third, we construct a large set of variables describing characteristics affecting whether a tract belongs to its metro's CEZ. Fourth, we use a learning algorithm to construct a function for each group of metros that mimics our subjective judgment. Fifth, we apply each of the learned functions to all metros in the corresponding group, thereby generating a predicted CEZ for each metro.

## Grouping metros

The Office of Management and Budget (OMB), using data from the 2000 decennial Census, delineated 922 Core-Based Statistical Areas (CBSAs) in 2003, designating 362 of them as "metropolitan" and the remainder as "micropolitan." To split the metropolitan CBSAs into groups, we first apply an algorithm that reduces a high dimensional set of metro characteristics—including land area, population and employment, numerous measures of population and employment density, occupation shares, housing stock composition, and commuting patterns—into a low dimensional representation that captures key ways in which metros differ.[3] This reduction is analogous to calculating the first few factors in principle components analysis.

We then use a clustering algorithm to divide the metros into four groups based on mini-

---

an employment density threshold and that together meet a total employment threshold (Bogart and Ferry, 1999; Small and Song, 1993; Anderson and Bogart, 2001; Marlay and Gardner, 2010). McMillen (2001) develops a nonparametric statistical approach to delineating subcenters that is similar in spirt to Redfearn (2007). But the employment density surfaces it fits depend on census tracts' measured distance from their metro's CBD, and so the approach cannot itself identify CBDs.

[3] Variables are constructed from the 2000 decennial census summary files and from the Census Transportation Planning Product (CTPP) 2000, which re-tabulates responses to the decennial census based on place of employment. For the CTPP variables, CBSA characteristics were constructed by summing over census tracts.

mizing a measure of dissimilarity with respect to the low dimensional set of characteristics. The resulting split aligns almost perfectly with metro population, so that the main value added from the clustering routine is to suggest the appropriate population levels at which to split the groups. We drop the smallest group of metros from the analysis because of the difficulty subjectively delineating CEZs for them. This leaves a 183 metros divided into three groups: 134 small metros with population from 220 thousand to 1 million, a group of 37 medium metros with population from 1 to 4 million, and a group of 12 large metros with population above 4 million.[4]

**Subjectively Delineating Central Employment Zones**

The second step of delineating CEZs for all metros was to select a training sample of metros for each of the three groups (Table 1). We did so partly focusing on metros with which we had some familiarity and partly focusing on metros that spanned the geometric area in the clustering algorithm's two-dimensional spatial representation of each group. The second criteria is meant to insure that each of the three training samples of metros is representative of all metros in the corresponding group. Otherwise, the implicit criteria we use in delineating CEZs for the training metros may not be applicable to some of the non-training metros in a group. In the next revision of this paper, we will formally compare summary statistics of metro characteristics—the ones that make up the high-dimensional set used by the clustering algorithm—as an additional check that the representativeness criterion is satisfied for each group.

The third step of delineating CEZs for all metros was to subjectively delineate the CEZ for each training metro. More specifically, we labeled each tract, $i$, in each of the training metros, $m$, as either belonging to or not belonging to the CEZ, $y_{m,i} \in \{1, 0\}$.

---

[4]We divided the metros into four groups rather than some other number based on preliminary results from running the clustering algorithm. There is a slight overlap in population between the group of metros we dropped and the smallest group we retained: Yakima, WA was clustered in the dropped group despite having several hundred more residents than Barnstable, MA and Macron, GA, which were clustered in the retained small group.

9

| | CBSA | size group | 2000 population | total tracts | labeled CEZ tracts | labeled CEZ land (sqmi) |
|---|---|---|---|---|---|---|
| 1 | New York-Newark-Edison, NY-NJ-PA | Large | 18,309,000 | 4,470 | 61 | 4.1 |
| 2 | Chicago-Naperville-Joliet, IL-IN-WI | | 9,097,000 | 2,049 | 35 | 7.5 |
| 3 | Dallas-Fort Worth-Arlington, TX | | 5,162,000 | 1,046 | 14 | 18.6 |
| 4 | Boston-Cambridge-Quincy, MA-NH | | 4,391,000 | 920 | 9 | 2.4 |
| 5 | St. Louis, MO-IL | Medium | 2,721,000 | 553 | 18 | 10.2 |
| 6 | Pittsburgh, PA | | 2,431,000 | 721 | 28 | 8.2 |
| 7 | Denver-Aurora, CO | | 2,158,000 | 527 | 31 | 16.3 |
| 8 | Cleveland-Elyria-Mentor, OH | | 2,148,000 | 692 | 31 | 7.7 |
| 9 | Portland-Vancouver-Beaverton, OR-WA | | 1,928,000 | 426 | 25 | 12.5 |
| 10 | Kansas City, MO-KS | | 1,836,000 | 514 | 23 | 10.5 |
| 11 | Sacramento--Arden-Arcade--Roseville, CA | | 1,797,000 | 403 | 27 | 16.7 |
| 12 | Columbus, OH | | 1,613,000 | 385 | 8 | 7.6 |
| 13 | Indianapolis, IN | | 1,525,000 | 315 | 7 | 5.2 |
| 14 | Charlotte-Gastonia-Concord, NC-SC | | 1,330,000 | 267 | 12 | 7.2 |
| 15 | Nashville-Davidson--Murfreesboro, TN | | 1,312,000 | 267 | 12 | 7.3 |
| 16 | Oklahoma City, OK | | 1,095,000 | 334 | 15 | 7.6 |
| 17 | Omaha-Council Bluffs, NE-IA | Small | 767,000 | 237 | 8 | 3.4 |
| 18 | Toledo, OH | | 659,000 | 174 | 6 | 2.1 |
| 19 | Syracuse, NY | | 650,000 | 189 | 13 | 5.7 |
| 20 | Columbia, SC | | 647,000 | 144 | 14 | 11.9 |
| 21 | Charleston-North Charleston, SC | | 549,000 | 117 | 6 | 1.6 |
| 22 | Des Moines, IA | | 481,000 | 107 | 3 | 3.6 |
| 23 | Lansing-East Lansing, MI | | 448,000 | 117 | 12 | 5.4 |
| 24 | Spokane, WA | | 418,000 | 106 | 6 | 6.1 |
| 25 | Santa Barbara-Santa Maria-Goleta, CA | | 399,000 | 86 | 3 | 1.8 |
| 26 | Peoria, IL | | 367,000 | 94 | 3 | 2.2 |
| 27 | Evansville, IN-KY | | 343,000 | 85 | 4 | 1.9 |
| 28 | Ann Arbor, MI | | 323,000 | 97 | 7 | 3.0 |
| 29 | Tallahassee, FL | | 320,000 | 63 | 4 | 3.6 |
| 30 | Savannah, GA | | 293,000 | 76 | 7 | 1.1 |
| 31 | Fort Smith, AR-OK | | 273,000 | 52 | 2 | 3.4 |
| 32 | Norwich-New London, CT | | 259,000 | 62 | 1 | 0.4 |
| 33 | Bremerton-Silverdale, WA | | 232,000 | 51 | 2 | 1.1 |
| 34 | Topeka, KS | | 225,000 | 54 | 3 | 3.2 |

Table 1: **Labeled Metropolitan Areas** Table lists the metropolitan areas in the training sample for each size group.
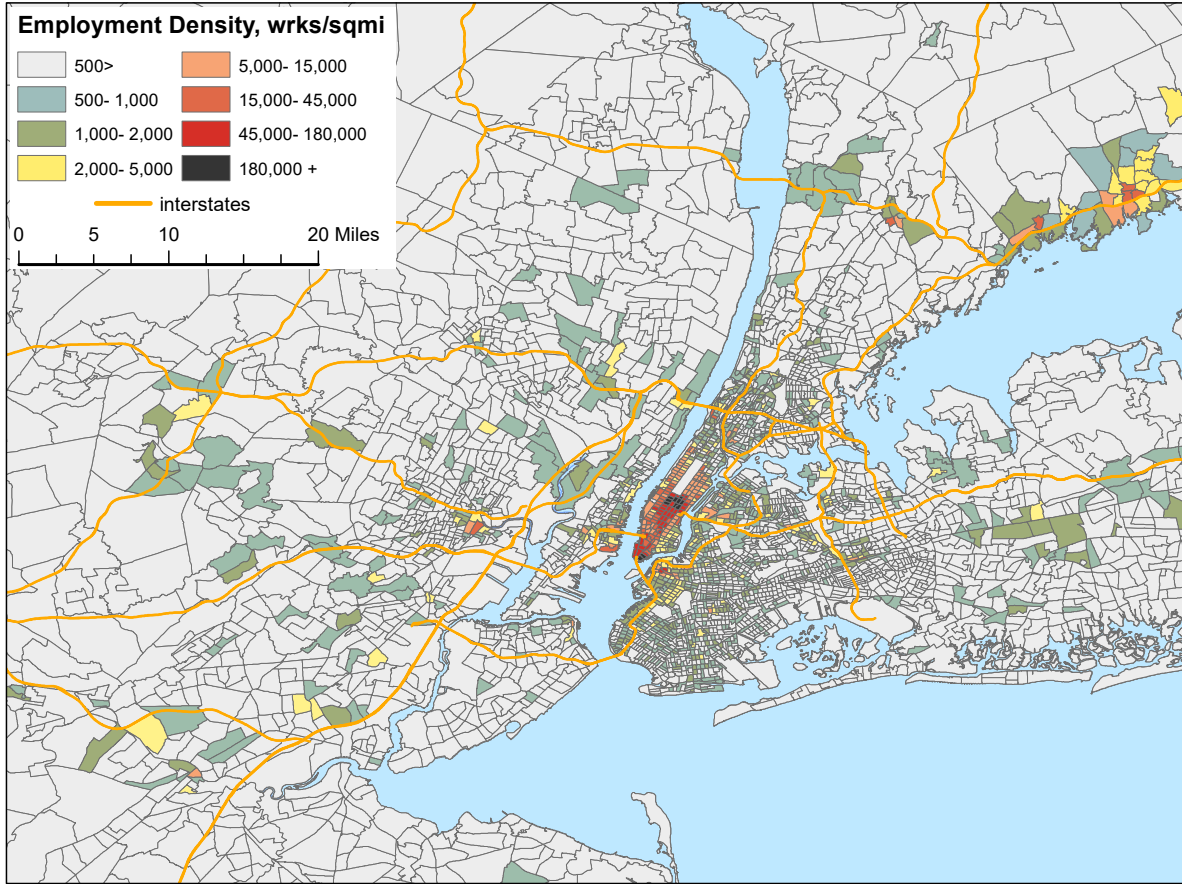
Figure 1: New York City Metropolitan Area

For each training metro, we first chose the general vicinity in which we judged the CEZ to be located. Heat maps of employment density clearly identified a single such vicinity in each metro area. For example, the geographic pattern of employment density in the New York City CBSA makes clear that the CEZ lies within the southern half of Manhattan (Figure 1). Similarly, the geographic pattern of employment density in the Kansas City CBSA makes clear that the CEZ lies in the middle of the metro, extending south from the Missouri River (Figure 2). We labeled each tract not in the CEZ vicinity as not belonging to the CEZ ($y_{m,i} = 0$).

To subjectively delineate the actual CEZ within each CEZ vicinity, we again used a heat
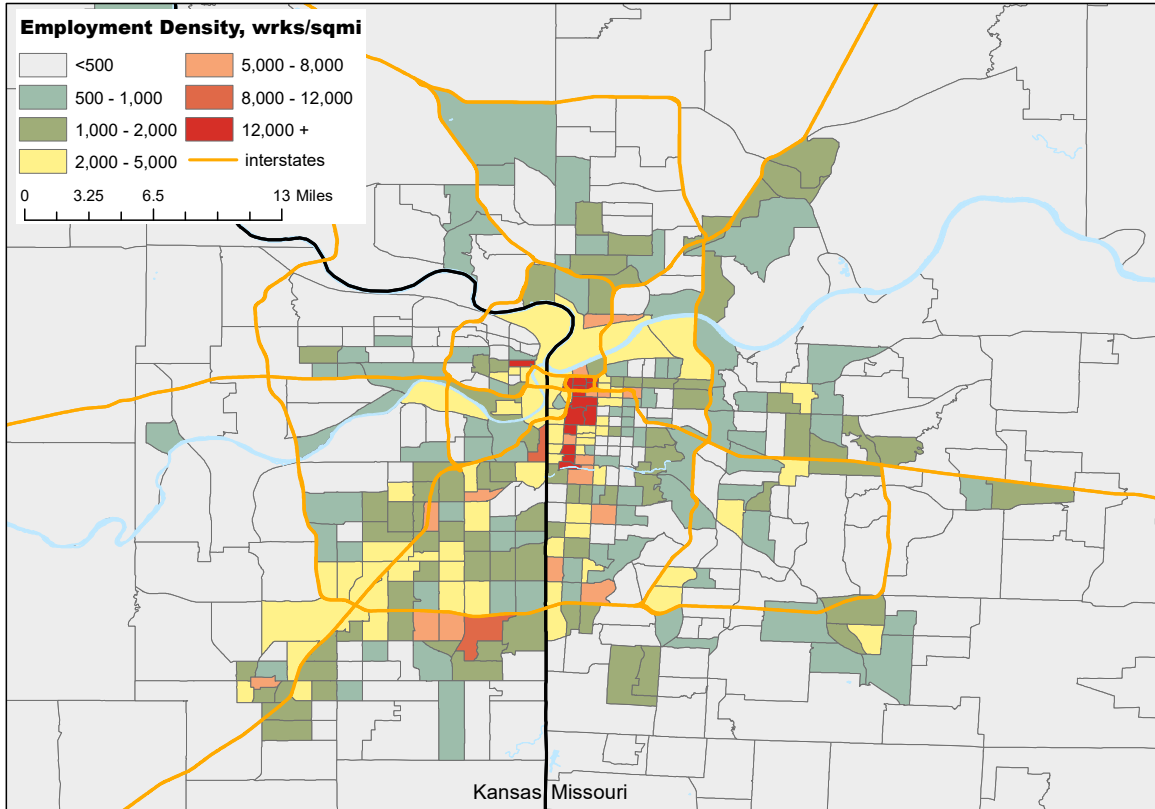
Figure 2: Kansas City Metropolitan Area

map of employment density complemented by heat maps of several other tract characteristics, such as share of employment in agglomerative occupations and the ratios of employment to population and multifamily units to single-family units. For each CEZ vicinity, we identified a core set of census tracts that we judged as clearly belonging to actual CEZ ($y_{m,i} = 1$) and another set of tracts within the vicinity that we judge as clearly not belonging to the CEZ. The former are characterized by having high employment density relative to other tracts in the vicinity and by being located close to each other. The latter are characterized by having low employment density relative to other tracts in the vicinity and by being located away from tracts with high relative density (Figures 3 and 4).

Last, we labeled each of the remaining tracts in the CEZ vicinity, which are character-
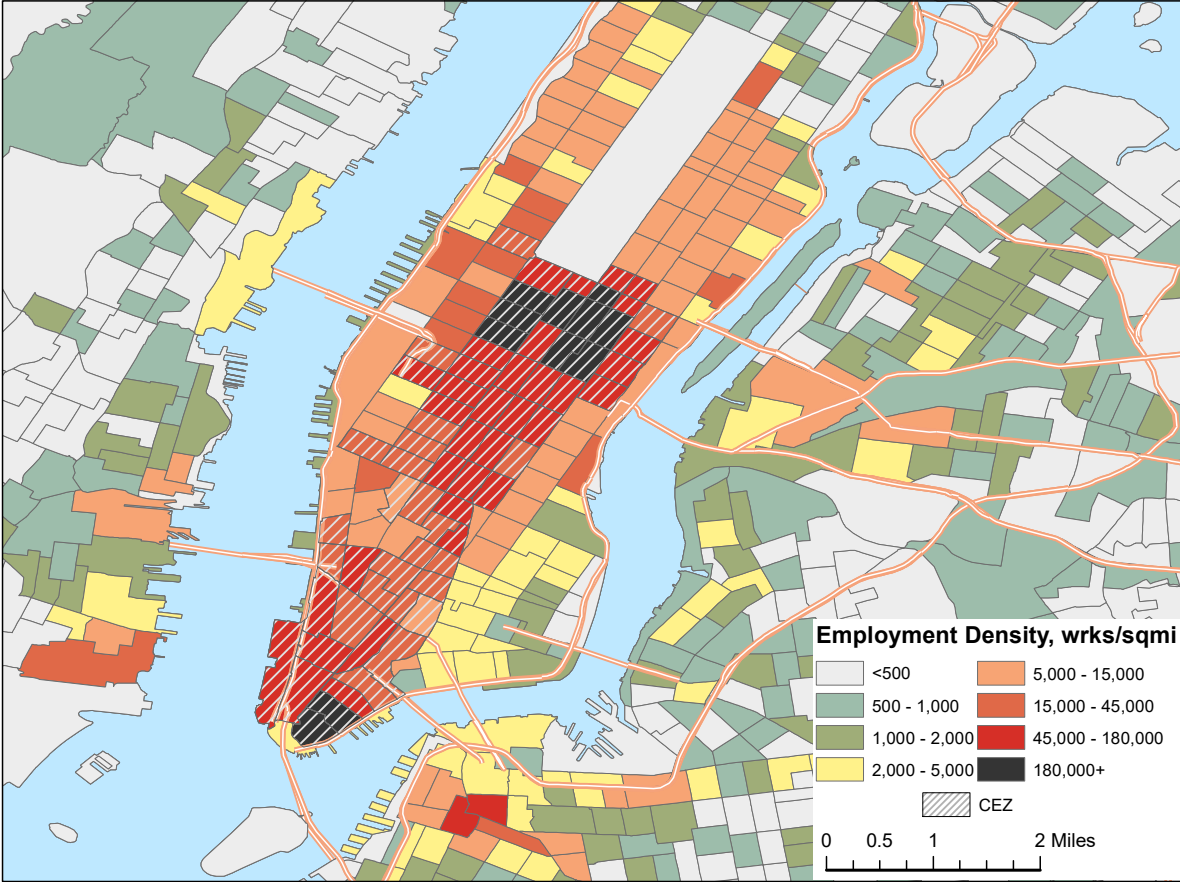
Figure 3: New York City Centralized Employment Zone

ized by several types of ambiguity. Some adjoin a tract in the core cluster but have only moderate relative employment density. Some have high relative employment density but are separated from the core cluster by one or two tracts with low relative employment density. Some connect tracts we judge to be part of the CEZ but have only moderate relative employment density. And some have a high ratio of employment to population but low relative employment density.

Several considerations guided us in resolving these ambiguities. We allowed CEZs to jump across locations from which they are excluded, such as rivers, parks, and historical districts. We leaned towards including tracts with a high ratio of employment to popula-

13
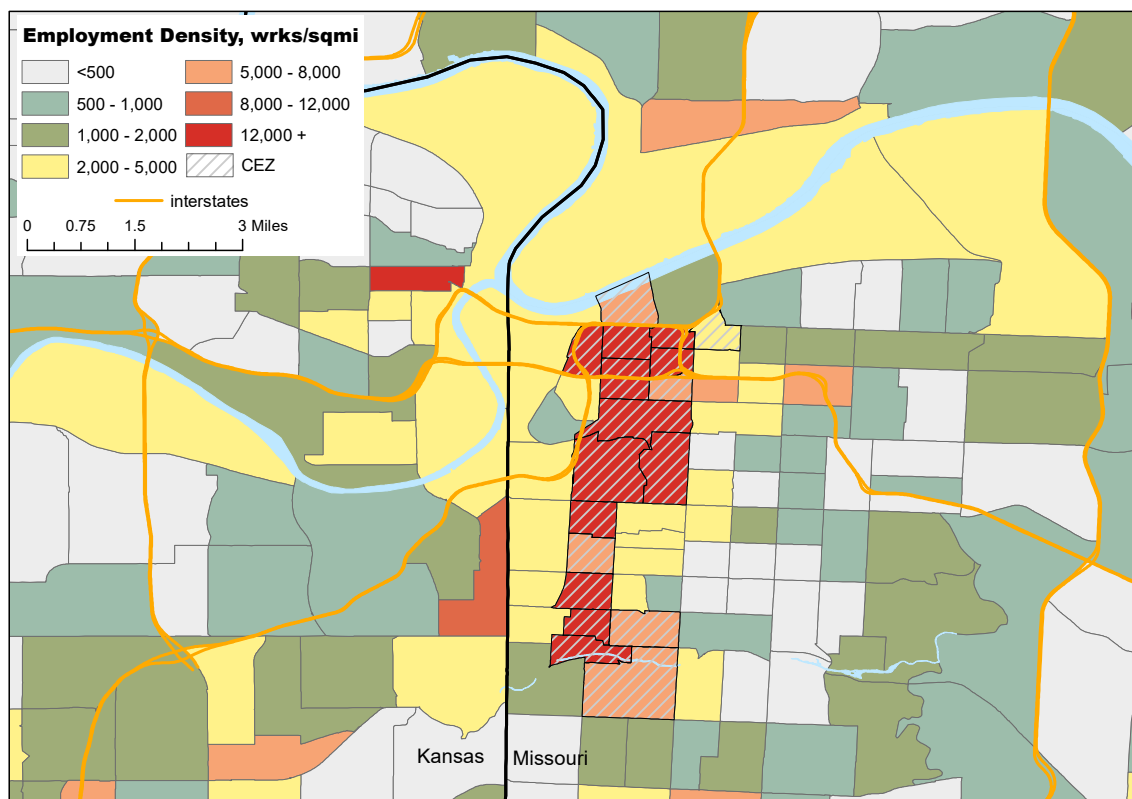
Figure 4: Kansas City Centralized Employment Zone

tion. We leaned towards including tracts where the occupational mix is skewed towards jobs in agglomerative occupations, those that arguably benefit form face-to-face contact such as business and financial operations specialists and legal occupations. Conversely, we leaned away from including tracts where the occupational mix is skewed towards jobs that complement residential location, such as education and personal care. Using satellite images from 2000, available in Google Earth, we leaned towards inclusion the higher the share of land within a tract that is devoted exclusively to employment except if such employment is by retailers surrounded by large parking lots. And we leaned towards inclusion the more residential use takes the form of apartment buildings, the rationale being that many CEZs are likely to have apartment buildings mixed in or immediately adacent. For tracts with

14

high employment density that do not adjoin the remainder of the CEZ, we leaned towards inclusion the shorter the separating distance, both in terms of geography and estimated travel time. We also leaned towards inclusion the more that residential use in separating area takes the form of apartment buildings.

Figure 3 zooms in on our classification of the CEZ vicinity in the New York City metropolitan area. It runs between the two most highly concentrated areas of employment in Manhattan, midtown and the Wall Street financial district. The excluded areas running alongside the CEZ on the east and west sides of the island are primarily residential. West Greenwich Village also juts into the CEZ on its west side.

Figure 4 zooms in on our classification of the CEZ vicinity in the Kansas City metropolitan area. We excluded a high relative density tract a mile to the west of the CEZ, partly because single-family homes occupy much of the intervening tracts and partly because the high density employment arises exclusively from a large medical center, which presumably has minimal business connections to firms and workers in the CEZ.

**Tract Variables**

The learning algorithm requires a large number of variables describing census tracts to help it predict which ones we would judge as belonging to their metro's CEZ. We constructed three sets of descriptive variables. One set characterizes individual tracts, independent of the tracts that surround them. One characterizes tracts in conjunction with neighboring tracts. And one set characterizes the distances from each tract to points in the metro that are likely to be in the vicinity of the CEZ. Together, the three sets are made up of approximately 600 variables. Table 2 gives some examples of each type, illustrating why the total number of variables is so large. In the next revision, we will allow the learning algorithm to train on the metropolitan variables used to group the metros. Although these can not help distinguish CEZ from non-CEZ tracts within the same metro, they may be help distinguish between tracts that belong to CEZ in some types of metros but not others.

15

| Variable Example | Variation of Construct | Total Variables |
|---|---|---|
| **Tract-specific** | | |
| Tract employment density | alt. measures of metro emp. | |
|   (total, agglom,): metro density (2) | density by percentiles (5) | 10 |
| Transit mode by occupation | | |
|   (total, agglom.) relative : metro (2) | alt. measures of transport. mode (3) | 6 |
| | | |
| **Neighborhood** | | |
| | alt. measures of metro emp. | |
| Tract employment density : metro density | density by percentiles (5) | 75 |
| Tract agglom. employment : | | |
|   all employment within certain radius (15) | none | 15 |
| Agglom. density | alt. measures of metro emp. | |
|   within certain radius (15) : employment density | density by percentiles (5) | 75 |
| | | |
| **Anchor-tracts** | | |
| Distance to densest tract | | |
|   within radius (5) | none | 5 |
|   within radius : metro radius (5) | none | 5 |
| Distance to tract with | | |
|   highest share of workers using public transit | none | 1 |

Note: Percentiles evaluated were 25th, 75th, 90th, 95th, and 99th. Neighborhood radii used 0.5-10 miles at 0.5 mile increments. Anchor-tracts radii used 0.5-3.0 miles at 0.5 mile increments.

Table 2: **Examples of Tract Variables Constructed by Type**

For the tract and neighborhood variables, we respectively normalize "absolute" measures of employment and population density by several measures of *metropolitan* employment and population density. For example, eight variations of tract employment density are included in the first set of variables: tract employment density relative to mean metro employment density, tract employment density relative to mean metro employment density calculated after removing tracts with outlying high density, and tract employment density relative to each of six benchmark metropolitan density percentiles (25th, 50th, 75th, 90th, 95th, and 99th).

Importantly, each of these normalizing factors is constructed weighting the "raw" employment density of each tract, tract employment divided by tract land area, by the tract's employment. In consequence, mean employment density is the mean density experienced by workers rather than the mean density experienced by tracts. Analogously, the percentiles are with respect to population and employment. For example, half of all workers in a metro

live in a tract with population density no higher than the median density and half live in a tract with population density at least as high as median density (Glaeser and Kahn, 2004; Rappaport, 2008). We expect that using only normalized density variables will effect learned criteria that better generalize across diverse metropolitan areas.

The second set of variables, which describe the neighborhoods of each tract, is meant to help the learning algorithm distinguish geographic clusters of tracts with CEZ characteristics from isolated tracts with CEZ characteristics. For example, the suburban portion of a metro may have a spike in employment density due to a business park in a single census single census tract. Combining that census tract with neighboring tracts greatly diminishes the spike. Combining CEZ census tracts with nearby ones is likely to cause less such diminishment. To allow the learning algorithm maximum flexibility, we construct neighborhoods of each tract extending to 15 benchmark radiuses ranging from 0.5 mile to 10 miles (measured by the distance from a tract's centroid to the centroids of surrounding tracts). Like the tract characteristics, the neighborhood characteristics are normalized by a number of alternative metropolitan characteristics.

The third set of variables, distances to locations with an above-average likelihood of being located in the CEZ, are meant to help identify the vicinity of the CEZ, thereby giving the learning algorithm the possibility of choosing cut distance distinguishing census tracts likely to be in the CEZ from those with CEZ characteristics likely to be in other suburban clusters. We identify a number of potential anchor tracts. For example the tract with the highest employment density serves as one potential anchor as do the tracts with the highest raw employment density when combined with neighboring tracts within 0.5, 1, 1.5, 2, and 3 miles. We also include versions of these distances normalized by a proxy of the metro radius to make cross metro comparisons more straightforward.

## Learning

As we deliberately allowed for ambiguity in defining central business districts and central employment zones, there are no "true" CEZ delineations. Instead, the relevant data generating process (DGP) is the implicit criteria we used in subjectively judging whether a tract belonged to its metro's CEZ. We employ a learning algorithm to build a function that mimics our implicit DGP, mapping observed tract characteristics, $x_{m,i} \in \mathbb{R}^K$, to a prediction of our judgment, $\hat{y}_{m,i} \in \{1, 0\}$. The mimicking prediction function will itself generally be an algorithm, possibly involving thousands of sequential operations on the observed variables. Importantly, the implicit DGP and mimicking prediction function are assumed to apply to all metros in a size group, not just to the ones we actually labeled.

The specific learning algorithm we use, LogitBoost, is grounded in a maximum likelihood framework (Friedman, Hastie and Tibshirani, 2000). It assumes the prediction function takes a logistic form: $\mathbf{p}\big(\text{we would classify } y_{m,i}{=}1 \,\big|\, x_{m,i}\big) = e^{\mathbf{F}(x_{m,i})}/(e^{\mathbf{F}(x_{m,i})} + e^{-\mathbf{F}(x_{m,i})})$, where $\mathbf{F}(x_{m,i})$ is a score function specific to each of the three size groups. Thus the probability that we would classify a tract as belonging to its metro's CEZ rises from near 0 when the score function is strongly negative to 0.5 when the score function equals 0 to near 1 when the score function is strongly positive. The prediction function for each of the three size groups can be written by stacking tract observations from all metros in the group:

$$\mathbf{p}(Y{=}1 \,|\, X) = \frac{e^{\mathbf{F}(X)}}{e^{\mathbf{F}(X)} + e^{-\mathbf{F}(X)}} = \frac{1}{1 + e^{-2\mathbf{F}(X)}} \tag{1}$$

Dropping the explicit metro index, let $\mathbf{A}$ represent the tracts in a training sample of metros, such as the labeled metros from one of the three size groups. Correspondingly, let $N_A$ equal the total number of tracts in the training metros. The likelihood function for observing the training tracts' characteristics and subjective classifications is given by,

$$\mathbf{L}(X_A, Y_A) = \prod_{\{i \in A\}=1}^{N_A} \left( \left( \frac{1}{1 + e^{-2\mathbf{F}(x_i)}} \right)^{y_i} \left( \frac{e^{-2\mathbf{F}(x_i)}}{1 + e^{-2\mathbf{F}(x_i)}} \right)^{1 - y_i} \right) \tag{2}$$

18

In a traditional logistic context, the score function takes a parametric form, $\mathbf{F}(X, \beta)$, and the parameter vector, $\beta$, can be estimated by solving the first order conditions associated with maximizing the likelihood function, $\partial \mathbf{L}(\beta | X_A, Y_A)/\partial \beta = 0$

In the present context, we assume only that some of the hundreds of characteristic variables that make up $X$ indeed affect how we classify tracts. The score function built by LogitBoost, $\mathbf{F}(X)$, is thus highly non-parametric, involving hundreds of sequential steps. Correspondingly, the LogitBoost score function is difficult to interpret, reflecting a tradeoff in return for more accurate predictions.

Let $\mathbf{F}_A(X) \equiv \mathbf{F}(X | X_A, Y_A)$ denote a score function that, subject to some constraints, is constructed to maximize the likelihood function for tracts in a set of training metros, $A$. In other words, $\mathbf{F}_A(X)$ is constructed such that applying it to the tracts in the training metros, $\mathbf{F}_A(X_A)$, maximizes $\mathbf{L}(X_A, Y_A)$ subject to the constraints.

To maximize likelihood, $\mathbf{F}_A(X_A)$ should assign a large positive score to tracts labeled as belonging to the CEZ and a large negative score for tracts labeled as not belonging to the CEZ. Unconstrained, LogitBoost can almost perfectly assign scores in this way, thereby achieving close to the maximum feasible likelihood score, $\mathbf{L}(X_A, Y_A) = 1$. Unsurprisingly, such over-fitting leads to poor performance predicting the CEZ classification of tracts in metros not included in the training sample. Hence a critical component of implementing LogitBoost is choosing constraints that prevent over-fitting.[5]

The LogitBoost algorithm builds $\mathbf{F}_A(X)$ by constructing a sequence of precursor score functions based on the training sample. First, it initializes a precursor score function to zero for all training observations, $\mathbf{F}_{A,0}(X_A) = 0$. The logistic specification implies that all tracts in the training sample have initialized probability of being in their metro's CEZ, $\mathbf{p}_{A,0}(X_A)$, equal to 0.5.

LogitBoost then iterates a pre-specified number of times, $T$, over a three-stage process.

---

[5]LogitBoost belongs to a class of learning algorithms that "boost" predictive power by iteratively applying a sub-algorithm with relatively weak predictive power. An advantage of boosting algorithms is that, with appropriate constraints, they avoid over-fitting.

For $t = \{1, 2, ..., T\}$:

1. Calculate:
$$Z_{A,t} = \frac{Y_A - \mathbf{p}_{A,t-1}(X_A)}{\mathbf{p}_{A,t-1}(X_A)(1 - \mathbf{p}_{A,t-1}(X_A))} \qquad \text{(desired adjustment to score)}$$

$$\Omega_{A,t} = \mathbf{p}_{A,t-1}(X_A)(1 - \mathbf{p}_{A,t-1}(X_A)) \qquad \text{(sample obs weight)}$$

2. Construct:
$$\mathbf{f}_{A,t}(X) \equiv \mathbf{f}_t(X | X_A, Z_{A,t}, \Omega_{A,t}) \qquad \text{(fits desired adjustment)}$$

3. Update:
$$\mathbf{F}_{A,t}(X_A) = \mathbf{F}_{A,t-1}(X_A) + \frac{1}{2}\mathbf{f}_{A,t}(X_A)$$

$$\mathbf{p}_{A,t}(X_A) = \frac{1}{1 + e^{-2\mathbf{F}_A(X)}}$$

The first calculated term, $Z_{A,t}$, represents a desired adjustment to the score function from the previous iteration, $t-1$. It takes on a simplified form that depends on whether we labeled a tract as belonging to a CEZ,

$$z_{i \in A,t} = \begin{cases} \frac{1}{\mathbf{p}_{A,t-1}(x_{i \in A})} & : \quad y_{i \in A} = 1 \\[2ex] -\frac{1}{1 - \mathbf{p}_{A,t-1}(x_{i \in A})} & : \quad y_{i \in A} = 0 \end{cases}$$

For tract observations in the training sample that belong to the CEZ, likelihood is maximized by a large positive logistic score with implied probability close to 1. Correspondingly, $z_{i \in A,t}$ is positive and becomes larger the further $\mathbf{p}_{A,t-1}(x_{i \in A})$ is below 1. In the limit, as $\mathbf{p}_{A,t-1}(x_{i \in A})$ goes to 0 for a CEZ tract, the desired adjustment goes to $\infty$. For tract observations in the training sample that do not belong to the CEZ, likelihood is maximized by a large negative logistic score with implied probability close to 0. Correspondingly, $z_{i \in A,t}$ is negative and becomes larger in absolute value the further $\mathbf{p}_{A,t-1}(x_{i \in A})$ is above 0. In the limit, as $\mathbf{p}_{A,t-1}(x_{i \in A})$ goes to 1 for a tract not in the CEZ, the residual goes to -$\infty$.

The fitted adjustment for each iteration, $\mathbf{f}_{A,t}(X) \equiv \mathbf{f}_t(X | X_A, Z_{A,t}, \Omega_{A,t})$, is constructed to

minimize subject to constraints the sum of the squared training residuals, $z_{i\in A,t} - \mathbf{f}_{A,t}(x_{i\in A})$, weighted by $\omega_{i\in A,t}$. A sample tract's weighting is maximized at $\mathbf{p}_{A,t-1}(x_{i\in A}) = 0.5$ and so prioritizes pushing $\mathbf{p}_{A,t}(x_{i\in A})$ away from the center of its distribution over (0,1).[6]

We construct the fitted adjustment functions, $\mathbf{f}_{A,t}(X)$, by "growing" regression trees.[7] These are an analog to decision trees, which sequentially split mixed observations of two or more discrete classes into subgroups that are less mixed. Regression trees instead operate on a real-valued dependent variable, sequentially splitting groups of observations into subgroups with similar values of the dependent variable. Figure 5 illustrates the first-iteration regression tree for the group of large metros.

A "root" node, leftmost in the figure, is the starting point for growing the tree. The four labeled large metros together have 119 CEZ tracts and 8,366 non-CEZ tracts. For the first iteration, these have equal relative weight, $\omega_{i\in A,1} = 0.25$ and respective values of $z_{i\in A,1}$ equal to +2 and -2. Correspondingly, the root node is associated with a weighted sum of squared residuals of $z_{i\in(A\cap 1),1}$ from the node's weighted mean, $\overline{z}_{i\in(A\cap 1),1} = -1.94$.

The regression tree algorithm splits the sample observations at the root node by finding the tract variable and associated cutoff that achieve the lowest sum of squared residuals across the resulting two children nodes. For the displayed regression tree, the optimal split is with respect to raw employment density within a 0.5 mile neighborhood normalized by the 99th percentile density of a tract's metro. The 111 CEZ tracts and 98 non-CEZ tracts for which this ratio is above 0.47 are split upward to node 2, resulting in mean value, $\overline{z}_{i\in(A\cap 2),1} = 0.12$, which corresponds to a probability, $p = 0.56$.[8] The remaining 8 CEZ tracts and 8,268 non-CEZ tracts are split downward to node 3, resulting in mean value, $\overline{z}_{i\in(A\cap 3),1} = -1.996$, which corresponds to a probability, $p = 0.018$.

---

[6] We additionally weight tract observations by the inverse square root of the number of tracts in their metro in order to keep the mimicking functions from being unduly shaped by metros with large number of tract.

[7] Alternatively, $\mathbf{f}_{A,t}(X)$ can be constructed by running a weighted ordinary-least-squares regression of $Z_{A,t}$ on $X_A$. The imposed linear functional form would serve as one constraint to help prevent over-fitting. Exclusions of any of the hundreds of tract characteristics from the right hand side of the regression would serve as other possible constraints.

[8] Note that this corresponding probability is based on the mean value of the score function at node 2 rather than the mean probability of sample observations at node 2, $\overline{p}_{i\in(A\cap 2),1}$.
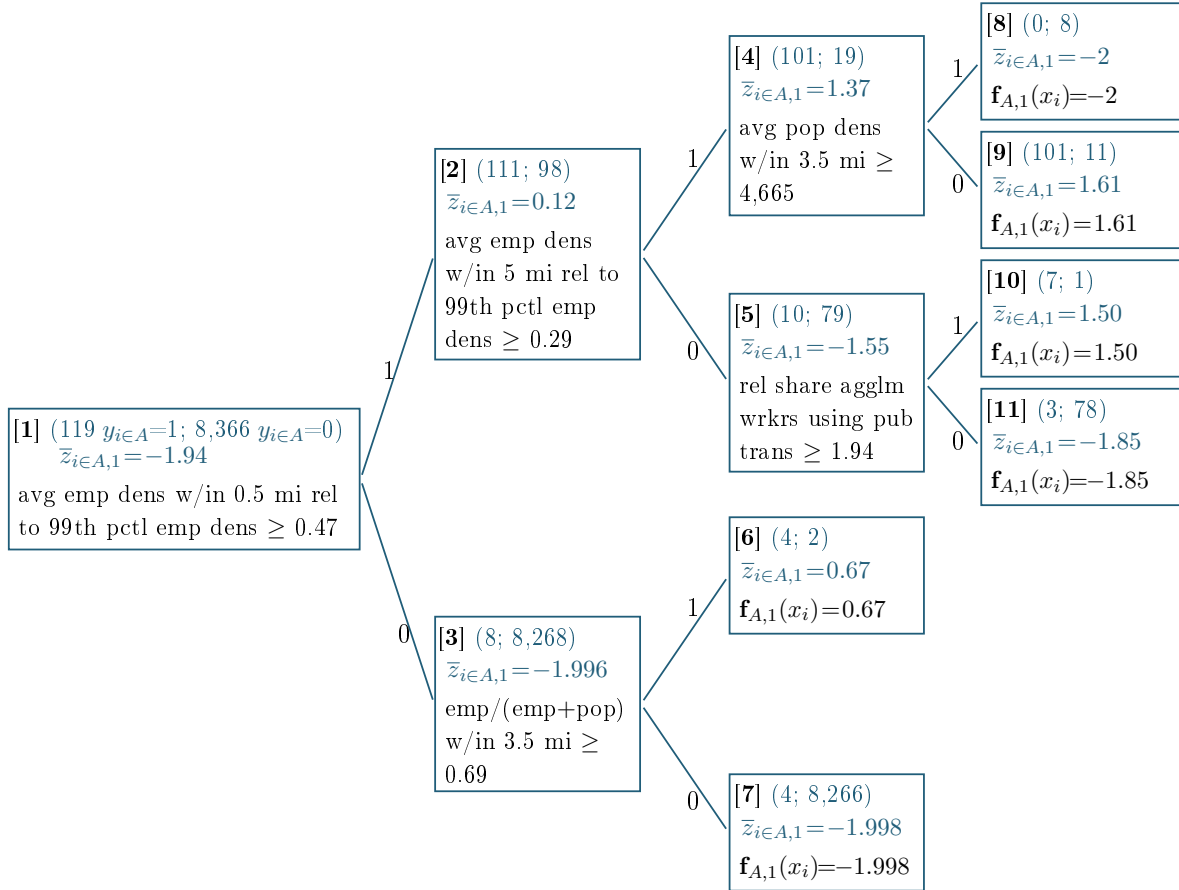
Figure 5: **Illustrative Regression Tree**. Figure shows the first-iteration regression tree constructed by LogitBoost using the 8,485 training tracts in the four labeled large metros (New York City, Chicago, Dallas, and Boston). The dependent variable, $z_{i,1}$, equals +2 for all sample tracts that belong to the CEZ and -2 for all remaining tracts. The function $f_{A,1}(x_i)$ splits tracts into the six terminal "leaf" nodes (nodes 6 through 11), assigning the mean value of sample tracts that terminated in each of these, $\overline{z}_{i \in A,1}$. Decision nodes are annotated with the characteristic variable and cutoff value used to split tracts that reach them.

The next growing step is to calculate the optimal splits at each of nodes 2 and 3 and then implement whichever reduces the sum of squared residuals by more. Splitting at node 2, based on whether raw employment density within a 5 mile neighborhood normalized by the 99th percentile density exceeds 0.29, proves best. Next, the algorithm calculates optimal splits at each of nodes 3, 4, and 5. This sequential splitting continues until the tree has undergone five splits, a first limit that constrains the algorithm from over-fitting the sample data. The fully-grown regression tree returns a fitted value of the desired adjustment, $\mathbf{f}_{A,t}(x_i) = \overline{z}_{i \in (A \cap \ell),1}$, where $\ell$ is the (right-most) "leaf" node to which a tract is routed based on its characteristics.

The constructed LogitBoost function, giving the logistic score that we would classify a tract as belonging to its metro's CEZ, equals the final iteration of the precursor function,

$$\begin{aligned} \mathbf{F}_A(X) &= \mathbf{F}_{A,T}(X) \\ &= \frac{1}{2}\sum_{t=1}^{T} \mathbf{f}_{A,t}(X) \end{aligned}$$

We set the number of iterations, $T$, to 100, a second limit that constrains the algorithm from over-fitting. Both limits, the number of splits in each regression tree and the number of iterations, are chosen to achieve the best out-of-sample fits, described below.

The final step of the LogitBoost algorithm is to make a binary prediction of how we would classify a tract based on whether the probability that corresponds to its logistic score, $\mathbf{p}_{A,t}(x_i)$, exceeds some threshold $p^*$. As with logistic regressions, $p^*$ is typically set to 0.50. For present purposes, however, we chose a higher $p^*$, thereby setting a higher bar for predicting an observation belongs to its metro's CEZ (by our subjective judgment). Implicitly, we are assigning a higher cost to falsely predicting a tract belongs to the CEZ—a type 1 error—than to falsely predicting a tract does not belong to the CEZ—a type 2 error. Doing so increases the likelihood that the high CEZ employment shares we estimate for selected metros understate rather than overstate centralization. To choose $p^*$, we looked

at the fit of the predicted inclusion probabilities of tracts in the training samples to our subjective classification.

**Fit**

We measure the fit of predictions to our subjective delineation using out-of-sample results for the labeled metros. For example, we create four training samples for the labeled large metros—$A_1$, $A_2$, $A_3$ and $A_4$—which each exclude a single metro, respectively New York, Chicago, Dallas, and Boston. We then construct a LogitBoost function for each of these training samples and apply it to the tracts in the corresponding excluded metro. Thus $\mathbf{F}_{A_1}(X_{\text{New York}})$ and $\mathbf{p}_{A_1}(X_{\text{New York}})$ give the logistic score and predicted probability for each tract in the New York City metropolitan area based on a training sample made up of all tracts in the Chicago, Dallas, and Boston metros.

The distribution of predicted probabilities across the tracts in all 34 labeled metros is shown in Figure 6. Tracts with an out-of-sample predicted probability below 0.05 account for 96 percent of all tracts, approximately matching the 97 percent of tracts we subjectively classified as not belonging to their metro's CEZ. As the probability threshold, $p^*$, increases from 0.50 to 0.95, type 1 error rates (false inclusions) steadily fall and type 2 error rates (false exclusions) steadily rise. Figure 7 illustrates this dependence. The lines for each size group are based on the stacked predicted probabilities from each excluded metro. Lines for each group show the type 1 and type 2 error rates, the number of tracts with the respective error type relative to the number of stacked tracts.

The implied biases in estimation matter more than the errors themselves. Figure 8 shows the mean and maximum misclassification of employment across the 34 labeled metros.[9] The mean share of employment falsely classified as belonging to the CEZ is very low at even the 0.50 threshold. But the maximum falsely included share across the metros at 0.50 threshold is moderately high, at just over 6 percent, implying that we risk significantly over-estimating

---

[9]Minimum shares were zero across all probability thresholds and error types.
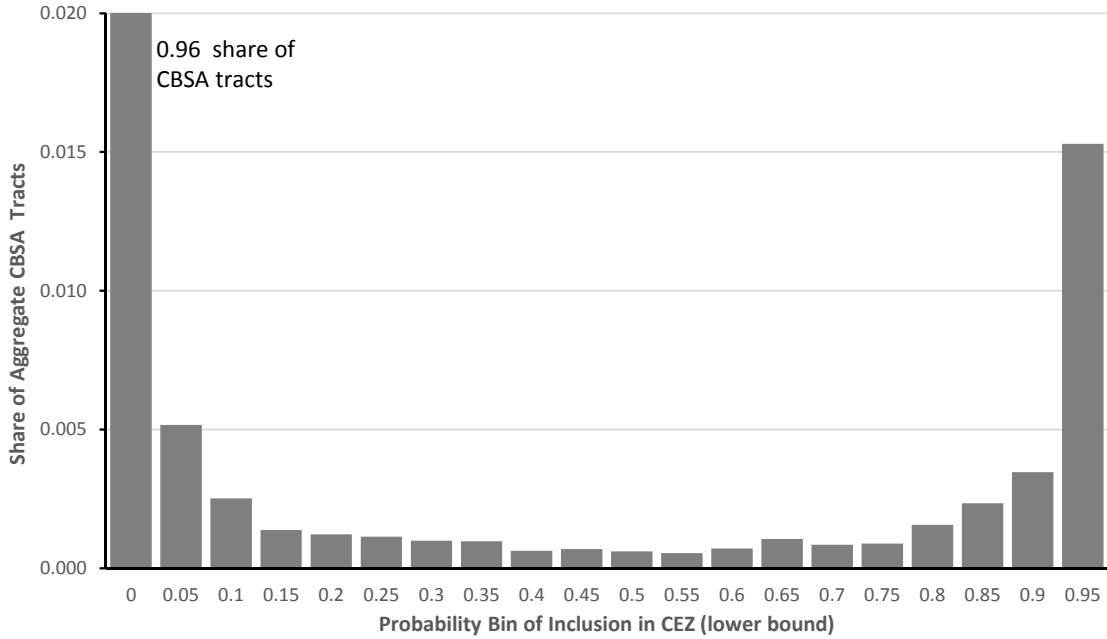
Figure 6: Predicted Probability of CEZ Inclusion, Out-of-Sample Labeled Metros Figure shows the distribution of the probability that a tract belongs to the CEZ of its metropolitan area.

employment centralization in some metros using this threshold. To mitigate this risk, we instead use $p^*$=0.80 for our baseline estimates. We also report robustness results for $p^*$=0.95, a threshold at which the maximum type 1 error is only 1 percent. Of course, these thresholds had large maximum type 2 errors for the labeled metros, approximately 12 percent at $p^*$=0.80 and 19 percent at $p^*$=0.95, implying that we are likely to be significantly underestimating centralization for some metros. We also report robustness results for $p^*$=0.50.

Lastly, the final step to delineating CEZs is to apply the learned prediction functions to the tracts in all 183 metropolitan areas. In the current version of the paper, we do so with just 3 prediction functions, one each for the small, medium, and large groups of metros. This implies that our predictions for the labeled metros are estimated in sample. LogitBoost attains extremely tight in-sample fits and so only a handful of tracts in the labeled metros are misclassified. An important concern is that any idiosyncratic judgments on our part get carried through to the predicted CEZs. In a revised version, we plan to use only the out-of-
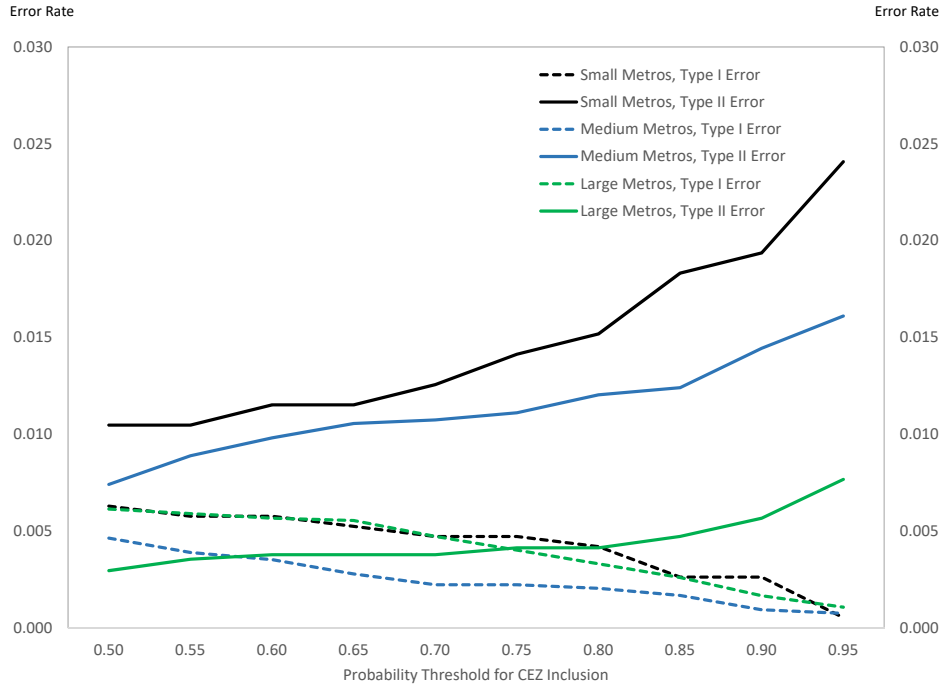
Figure 7: Type I and II Errors by Probability Threshold

sample predictions described above for the labeled metros. This out-of-sample methodology implies an important filtering role for learning even if we were to subjectively judge the CEZs for all metros.

# 4    Predicted Central Employment Zones

The learned prediction functions delineate central employment zones for 160 of the 183 metros to which they were applied. For each of the remaining 23 metros, all tracts were assigned an inclusion probability below the baseline 0.80 threshold.[10]  The first subsection

---

[10]The 23 metros with no baseline CEZs include two large ones, Los Angeles and Detroit, 1 medium one, Virginia Beach-Norfolk-Newport News, and 20 small ones (Table A.1). Some of the zero predictions, such as for Los Angeles and Virginia Beach-Norfolk-Newport News, appear to reflect multiple potential CEZ clusters. Some, including several tourist-oriented metros in Florida, appear to reflect a lack of sufficiently clustered employment anywhere in the metro. And some, including Detroit and Youngstown, may be related to industrial decline. In the next revision of this paper, we will include some of these metros in our training

26

Employment Share Falsely Included/Excluded
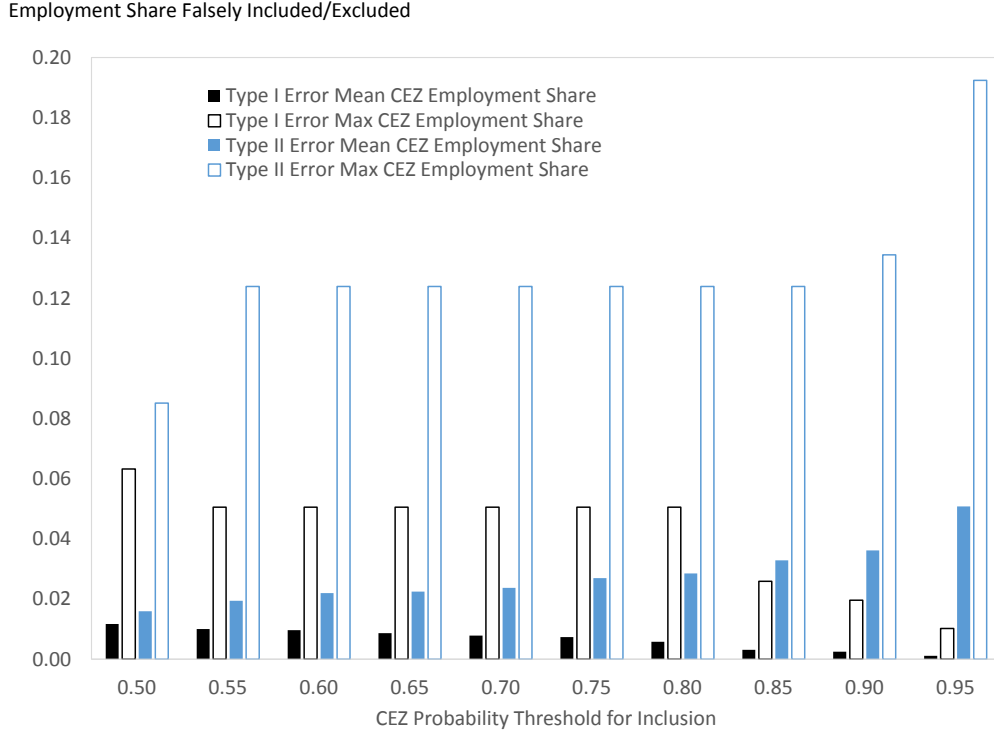


Figure 8: Metro Employment Falsely Classified by Probability Threshold

below summarizes some geographic characteristics of the predicted CEZs. The second sub-section describes how centralized metros' employment is by our baseline and alternative measures. The third section describes employment and population density in the predicted CEZs.

## 4.1 Geographic Size

The CEZs predicted by our baseline threshold vary considerably in size, encompassing from 1 to 60 census tracts and spanning from less than 1 square mile to 19 square miles (Table 3). Across metros, the mean distance of the CEZ census tract farthest from the CEZ employment centroid is 1.3 miles. In some metros, CEZ tracts extend to more than 4 miles away from the CEZ employment centroid. An alternative diagonal distance, measured

samples. For Los Angeles, Redfearn (2007) documents that downtown remains the largest employment center and, by far, the most dense.

| | Metro | Central Employment Zone | | | | Non-Rural Remainder of CBSA | | | |
|---|---|---|---|---|---|---|---|---|---|
| Attribute | Size | Mean | Std. Dev. | Min | Max | Mean | Std. Dev. | Min | Max |
| # Tracts | All | 6.7 | 7.9 | 1 | 61 | 206 | 408 | 18 | 4238 |
| | Small | 3.7 | 2.6 | 1 | 14 | 71 | 39 | 18 | 189 |
| | Medium | 14.1 | 8.3 | 3 | 30 | 336 | 143 | 130 | 610 |
| | Large | 15.1 | 18.7 | 3 | 61 | 1,304 | 1,098 | 546 | 4238 |
| Land Area (sqmi.) | All | 3.5 | 3.4 | 0.2 | 18.6 | 376 | 481 | 58 | 3,456 |
| | Small | 2.3 | 1.8 | 0.2 | 11.9 | 172 | 84 | 58 | 428 |
| | Medium | 7.3 | 3.9 | 1.9 | 16.1 | 610 | 214 | 320 | 1,047 |
| | Large | 4.2 | 5.4 | 0.7 | 18.6 | 1,895 | 714 | 983 | 3,456 |
| Max Dist to CEZ Emp Centroid (mi) | All | 1.3 | 1.0 | 0.0 | 4.2 | 31 | 17 | 6 | 90 |
| | Small | 0.9 | 0.6 | 0.0 | 3.5 | 25 | 12 | 6 | 78 |
| | Medium | 2.5 | 1.0 | 0.9 | 4.2 | 41 | 14 | 20 | 90 |
| | Large | 1.8 | 1.0 | 0.5 | 3.2 | 62 | 15 | 41 | 90 |
| Diagonal Distance (mi) | All | 2.3 | 1.8 | 0.0 | 7.4 | 53 | 31 | 10 | 192 |
| | Small | 1.5 | 1.1 | 0.0 | 6.1 | 40 | 18 | 10 | 120 |
| | Medium | 4.6 | 1.6 | 1.9 | 7.4 | 77 | 23 | 42 | 136 |
| | Large | 2.9 | 1.6 | 0.8 | 5.3 | 123 | 31 | 69 | 192 |

Table 3: **CEZ Geographic Size**. Table reports characteristics in 2000 for the predicted CEZ and remaining non-rural portion of 160 CBSAs (115 small, 35 medium, 10 large). The CEZ employment centroid is measured as the employment-weighted mean of each CEZ tract's geographic centroid. Maximum distance is measured from the tract centroid that is farthest from the employment centroid. Diagonal distance is measured from the maximum latitude and maximum longitude across CEZ tracts to the minimum latitude and minimum longitude across CEZ tracts.

from the maximum latitude and longitude of a CEZ's tracts to the minimum latitude and longitude of a CEZ's tracts, averages just over 2 miles but ranges above 7 miles.[11]

For comparison, Table 3 also shows analogous geographic sizes for the remainder of the non-rural portion of CBSAs, the combination of all non-CEZ tracts with either population

---

[11]The means and ranges for all three size variables are highest for the group of medium metros. Rappaport (2016) finds a similar non-monotonic pattern of CBD size and metro population. Increases in metro population from low levels put more upward pressure on prices for central commercial land than for central residential land, causing CBD land area to increase. But increases in metro population from higher levels put more upward pressure on prices for central residential land, causing CBD land area to contract.

density or employment density of at least 500 per square mile.[12] On average, the non-rural portions CBSAs excluding their CEZ encompass about 200 tracts and 400 square miles, include tracts as far as 31 miles from the CEZ employment centroid, and span 53 miles between their most distant tracts as measured by extreme latitudes and longitudes.

## 4.2   Employment Centralization

Consistent with skepticism about the empirical relevance of centralized employment, employment in most CBSAs was characterized by relatively low employment centralization in 2000. But in a significant number of CBSAs, employment centralization was somewhat higher. And for agglomerative occupations—those that arguably benefit most from face-to-face interaction—employment centralization was moderately high in numerous CBSAs.

We measure baseline centralization by the CEZ share of all CBSA employment rather than by its share of employment located in only the non-rural portion. We do so notwithstanding interpreting our motivating question on the centralization of employment as applying strictly to land used for metropolitan purposes. But some residents of the non-rural portion commute out to jobs in the rural portion, which arguably constitutes satellite metropolitan use. Our baseline measure thus again errs towards understating centralization.

Table 4 summarizes several measures of employment centralization for the 183 metros to which we applied the prediction functions.[13] The share of employment taking place in the baseline predicted CEZs, delineated using the 0.80 inclusion probability threshold, was typically low, especially for large metros. Across all metros, the mean CEZ share was 12

---

[12] Core-Based Statistical Areas are constructed as combinations of whole counties, large parts of which are agriculture or essentially unsettled. Our non-rural classification encompasses a larger area than the analogous urban classification used by the Census Bureau, for which having population density of at least 500 per square mile is a necessary but not sufficient criterion. We additionally classify tracts as non-rural if they have employment density above 500 per square mile in order to capture pockets with few residents but significant employment that are interspersed within the more settled portions of CBSAs. Our non-rural portions on average account for 73 percent of CBSA population, 84 percent of CBSA employment, but just 15 percent of CBSA land area (Appendix Table A.2). We interpret the residual rural portions of CBSAs as exemplifying non-metropolitan land use and so exclude it from our analysis.

[13] We report summary results for all 183 metros rather than only for the 160 metros for which a CEZ was identified in order not to upwardly bias results. Metros with no identified CEZ are treated as having a CEZ employment share of zero.

percent; across large metros, it was just 8 percent (first horizontal block).

| | Metro Characteristic | Metro Size | Mean | Std. Dev. | Min | 80th pctile | Max |
|---|---|---|---|---|---|---|---|
| 1. | All Employment (Baseline--0.80 inclusion thershold) | All | 0.12 | 0.08 | 0 | 0.18 | 0.37 |
| | | Small | 0.12 | 0.08 | 0 | 0.18 | 0.37 |
| | | Medium | 0.15 | 0.06 | 0 | 0.20 | 0.32 |
| | | Large | 0.08 | 0.06 | 0 | 0.12 | 0.18 |
| 2. | All Employment (0.95 inclusion threshold) | All | 0.09 | 0.08 | 0 | 0.16 | 0.32 |
| | | Small | 0.09 | 0.08 | 0 | 0.16 | 0.32 |
| | | Medium | 0.10 | 0.07 | 0 | 0.16 | 0.22 |
| | | Large | 0.05 | 0.06 | 0 | 0.09 | 0.17 |
| 3. | All Employment (0.50 inclusion threshold) | All | 0.14 | 0.08 | 0 | 0.19 | 0.39 |
| | | Small | 0.14 | 0.09 | 0 | 0.19 | 0.39 |
| | | Medium | 0.16 | 0.06 | 0 | 0.22 | 0.36 |
| | | Large | 0.10 | 0.05 | 0 | 0.13 | 0.18 |
| 4. | All Employment (share of non-rural portion of CBSA) | All | 0.15 | 0.10 | 0 | 0.23 | 0.50 |
| | | Small | 0.15 | 0.10 | 0 | 0.23 | 0.50 |
| | | Medium | 0.17 | 0.07 | 0 | 0.23 | 0.34 |
| | | Large | 0.09 | 0.06 | 0 | 0.13 | 0.19 |
| 5. | Agglomerative Employment | All | 0.20 | 0.12 | 0 | 0.31 | 0.50 |
| | | Small | 0.20 | 0.12 | 0 | 0.31 | 0.50 |
| | | Medium | 0.25 | 0.09 | 0 | 0.31 | 0.39 |
| | | Large | 0.15 | 0.11 | 0 | 0.22 | 0.35 |
| 6. | Legal Employment | All | 0.41 | 0.22 | 0 | 0.59 | 0.80 |
| | | Small | 0.41 | 0.23 | 0 | 0.60 | 0.80 |
| | | Medium | 0.47 | 0.16 | 0 | 0.59 | 0.63 |
| | | Large | 0.31 | 0.18 | 0 | 0.43 | 0.59 |
| 7. | Population | All | 0.02 | 0.02 | 0 | 0.03 | 0.10 |
| | | Small | 0.02 | 0.02 | 0 | 0.03 | 0.10 |
| | | Medium | 0.02 | 0.01 | 0 | 0.03 | 0.05 |
| | | Large | 0.00 | 0.00 | 0 | 0.01 | 0.01 |

Table 4: **Employment Centralization** Table reports the CEZ share of employment in 2000 for the 183 CBSAs (134 small, 37 medium, 12 large) to which we applied the learned prediction functions. Under the baseline 0.80 probability threshold for inclusion, 23 of these metros have no predicted CEZ tracts and so are considered to have a CEZ share of 0. Unless otherwise indicated, shares are based on the baseline predicted CEZs.
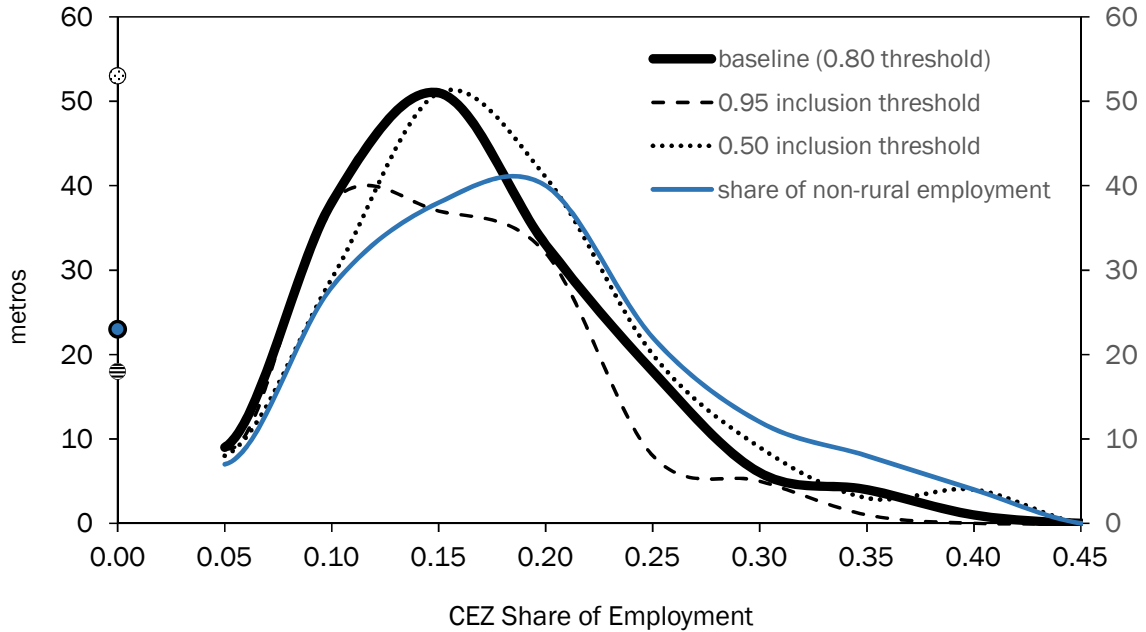
Figure 9: **Distribution of CEZ Employment Shares** Figure shows the distribution of the CEZ share of CBSA employment in 2000 under the respective baseline, tight, and standard inclusion thresholds of 0.80, 0.95, and 0.50. In addition, the blue line shows the distribution of the baseline CEZ's share of employment in the non-rural portion of CBSAs. Markers on the left vertical axis correspond to the number of metros with no predicted CEZ tracts at the corresponding inclusion threshold.

For some metros, centralization was less low. The distribution of baseline CEZ shares is positively skewed, with many metros having shares well above the mean (Figure 9, solid black line). For example, 29 metros had a baseline CEZ share of at least 20 percent and five metros had a baseline CEZ share of at least 30 percent.

Alternative measures of centralization using the 0.95 probability and 0.50 probability inclusion thresholds give a sense of the robustness of the baseline results. The high threshold significantly cuts the number of metros with at least a moderate share of employment in the predicted CEZ (dashed line). Even so, 14 metros had a predicted CEZ share of at least 20 percent. Using the 0.50 threshold, which is standard in binary inference, significantly fattens the right tail compared to the baseline (dotted line). In this case, 36 metros had a predicted CEZ share of at least 20 percent and seven had a predicted CEZ share of at least 30 percent.

Alternatively measuring centralization by the baseline CEZ share of non-rural metropolitan employment, which arguably better corresponds to the standard monocentric urban model, fattens the right tail of the distribution by even more (blue line). In this case, 47 metros had a baseline CEZ share of non-rural employment of at least 20 percent and 13 had a baseline CEZ share of non-rural employment of at least 30 percent.

Employment was considerably more centralized for agglomerative occupations—those that arguably benefit most from face-to-face contact and that are not strongly complementary to residential, retail, and manufacturing locations.[14] The mean CEZ share of agglomerative employment was 20 percent across all metros and 25 percent across medium metros (Table 4, fifth vertical block). Employment in legal occupations, which is one of the agglomerative categories, was especially centralized, presumably reflecting past and present ties to court houses located in traditional downtowns. The mean CEZ share of legal employment was 41 percent across all metros and 47 percent across the medium metros (sixth vertical block).[15]

The CEZ shares for both agglomerative and legal occupations are distributed diffusely, reflecting moderately high agglomerative centralization and significantly high legal centralization in a number of metros (Figure 10). For example, 24 metros had a baseline CEZ share of CBSA agglomerative employment of at least 33 percent (blue line). Seventy-nine metros had a baseline CEZ share of CBSA legal employment of at least 50 percent and 12 metros had one of at least 66 percent (green line).

Table 5 summarizes the various measures of employment centralization for the metros

---

[14]The 2000 Census Transportation Planning Package re-tabulates occupation by place of work for 22 categories. Appendix Table A.3 lists the baseline CEZ share of CBSA employment for each. We consider five of the categories to be agglomerative: legal occupations; computer and mathematical occupations; business and financial operations specialists; life, physical, and social science occupations; and arts, design, entertainment, sports, and media occupations. Two additional categories—management occupations, and architecture and engineering occupations—also include many jobs that significantly benefit from face-to-face contact. We nevertheless exclude them from our agglomerative category because these categories also include many jobs that are strongly complementary to residential locations (via the retail sector) and to manufacturing locations (e.g., production engineers).

[15]Elvery and Sveikauskas (2011) document that metropolitan subcenters, like our CEZ, are characterized by a high share of employment in agglomerative occupations.
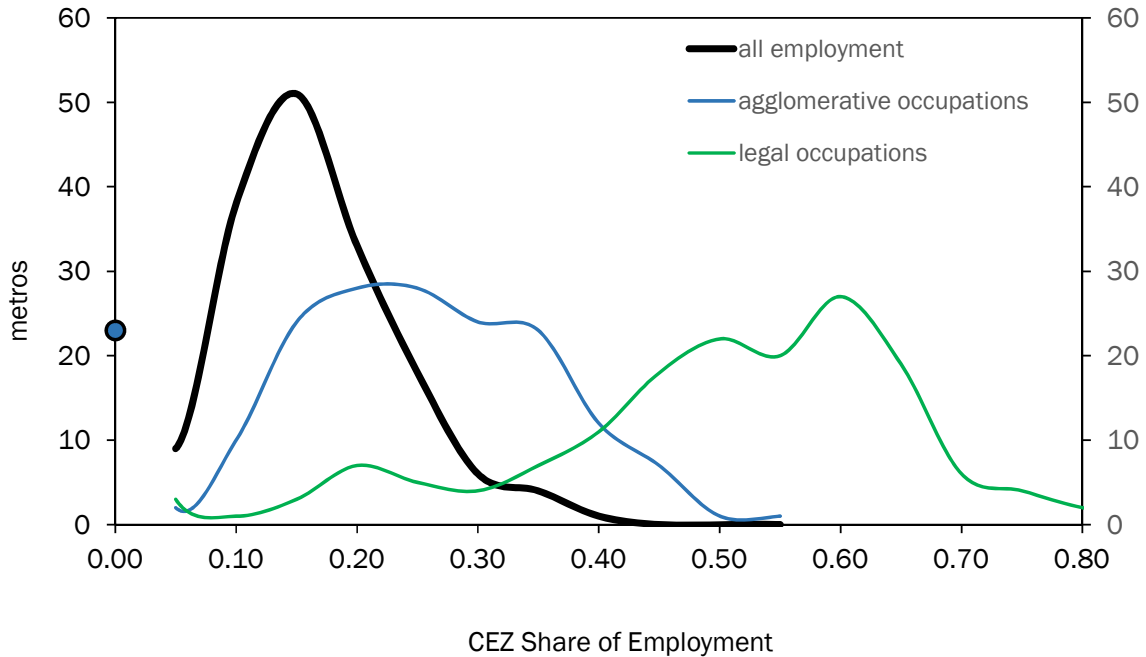
Figure 10: **Distribution of CEZ Employment Shares by Occupation** Figure shows the distribution of baseline CEZ shares for all CBSA employment and for CBSA employment in agglomerative and legal occupations.

with a baseline CEZ share of total CBSA employment of at least 20 percent. The total employment CEZ share is predicted with considerable precision for some metros, as measured by the spread between the tight 0.95 and standard 0.50 inclusion probabilities. This 95/50 spread is less than 2 percentage points for eight of the listed metros. For example, the predicted CEZ share for the Topeka CBSA is 32 percent at both the tight and standard thresholds and the predicted CEZ share for the Columbia South Carolina CBSA is 28 percent at the tight threshold and 30 percent at the standard threshold. Consistent with LogitBoost's ability to tightly fit sample observations, all eight of the listed metros that meet this precision criterion were among those we labeled. But numerous non-labeled metros not included in the table also have a 95/50 spread below 2 percentage points.

Conversely, a number of metros have considerably imprecise predictions. Six of the high centralization metros listed in the table and 20 more of the remaining metros have a 95/50 spread of at least 10 percentage points. The spread is especially egregious for Las Vegas,

| | | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | CEZ Employment Share | | | | non-rural |
| | | | | Total | | | | | | |
| | CBSA | pop | base | 0.95 thrsh | 0.50 thrsh | non-rural | agglm occup | legal occup | pop share | land share |
| | All metros (mean) | | 0.12 | 0.09 | 0.14 | 0.15 | 0.20 | 0.41 | 0.02 | 0.012 |
| 1 | Atlantic City, NJ | 253,000 | 0.37 | 0.26 | 0.38 | 0.50 | 0.29 | 0.38 | 0.06 | 0.026 |
| 2 | *Spokane, WA* | 418,000 | 0.33 | 0.26 | 0.33 | 0.37 | 0.42 | 0.71 | 0.04 | 0.039 |
| 3 | Las Vegas, NV | 1,375,000 | 0.32 | 0.00 | 0.36 | 0.34 | 0.32 | 0.50 | 0.04 | 0.039 |
| 4 | *Topeka, KS* | 225,000 | 0.32 | 0.32 | 0.32 | 0.38 | 0.50 | 0.70 | 0.04 | 0.045 |
| 5 | *Columbia, SC* | 647,000 | 0.30 | 0.28 | 0.30 | 0.39 | 0.45 | 0.73 | 0.05 | 0.041 |
| 6 | Honolulu, HI | 876,000 | 0.29 | 0.22 | 0.39 | 0.30 | 0.44 | 0.80 | 0.05 | 0.013 |
| 7 | Charleston, WV | 305,000 | 0.27 | 0.27 | 0.37 | 0.33 | 0.45 | 0.73 | 0.03 | 0.019 |
| 8 | Gainesville, FL | 232,000 | 0.27 | 0.19 | 0.27 | 0.34 | 0.32 | 0.60 | 0.07 | 0.041 |
| 9 | *Tallahassee, FL* | 320,000 | 0.27 | 0.27 | 0.27 | 0.31 | 0.38 | 0.58 | 0.04 | 0.039 |
| 10 | Anchorage, AK | 320,000 | 0.27 | 0.13 | 0.27 | 0.35 | 0.42 | 0.67 | 0.04 | 0.047 |
| 11 | *Lansing, MI* | 448,000 | 0.26 | 0.23 | 0.26 | 0.30 | 0.37 | 0.57 | 0.06 | 0.031 |
| 12 | Asheville, NC | 369,000 | 0.25 | 0.19 | 0.25 | 0.36 | 0.35 | 0.77 | 0.03 | 0.037 |
| 13 | *Ann Arbor, MI* | 323,000 | 0.25 | 0.22 | 0.25 | 0.30 | 0.31 | 0.48 | 0.10 | 0.031 |
| 14 | *Syracuse, NY* | 650,000 | 0.25 | 0.24 | 0.25 | 0.30 | 0.37 | 0.63 | 0.05 | 0.027 |
| 15 | *Des Moines, IA* | 481,000 | 0.24 | 0.24 | 0.24 | 0.29 | 0.44 | 0.62 | 0.02 | 0.020 |
| 16 | Austin, TX | 1,250,000 | 0.24 | 0.22 | 0.25 | 0.27 | 0.31 | 0.63 | 0.03 | 0.017 |
| 17 | Madison, WI | 502,000 | 0.23 | 0.20 | 0.24 | 0.30 | 0.32 | 0.60 | 0.10 | 0.021 |
| 18 | Jackson, MS | 497,000 | 0.23 | 0.16 | 0.25 | 0.31 | 0.36 | 0.63 | 0.03 | 0.037 |
| 19 | *Portland, OR* | 1,928,000 | 0.22 | 0.22 | 0.23 | 0.24 | 0.35 | 0.58 | 0.03 | 0.021 |
| 20 | Columbus, GA | 282,000 | 0.22 | 0.15 | 0.26 | 0.27 | 0.40 | 0.47 | 0.03 | 0.037 |
| 21 | Denver, CO | 2,158,000 | 0.22 | 0.20 | 0.22 | 0.24 | 0.31 | 0.60 | 0.05 | 0.026 |
| 22 | Birmingham, AL | 1,052,000 | 0.22 | 0.14 | 0.24 | 0.26 | 0.31 | 0.53 | 0.02 | 0.022 |
| 23 | Little Rock, AR | 611,000 | 0.21 | 0.15 | 0.21 | 0.28 | 0.39 | 0.63 | 0.03 | 0.030 |
| 24 | Cedar Rapids, IA | 237,000 | 0.21 | 0.19 | 0.21 | 0.26 | 0.33 | 0.53 | 0.03 | 0.020 |
| 25 | *Sacramento, CA* | 1,797,000 | 0.21 | 0.19 | 0.22 | 0.24 | 0.35 | 0.58 | 0.04 | 0.027 |
| 26 | New Orleans, LA | 1,316,000 | 0.21 | 0.17 | 0.22 | 0.23 | 0.35 | 0.52 | 0.02 | 0.010 |
| 27 | Albuquerque, NM | 729,000 | 0.21 | 0.13 | 0.22 | 0.23 | 0.25 | 0.51 | 0.03 | 0.026 |
| 28 | Macon, GA | 222,000 | 0.20 | 0.16 | 0.20 | 0.30 | 0.29 | 0.65 | 0.02 | 0.023 |
| 29 | *Pittsburgh, PA* | 2,431,000 | 0.20 | 0.20 | 0.20 | 0.24 | 0.39 | 0.62 | 0.03 | 0.009 |

Table 5: **Metros with High Baseline Employment Centralization** Table reports the CEZ share of employment in 2000 for metros with at least 20 percent of CBSA employment located in the baseline CEZ. Italicized metros are in the training samples

which has a predicted CEZ share of 0 under the 0.95 inclusion threshold (i.e., no predicted CEZ) and a predicted CEZ share of 0.36 under the 0.50 inclusion threshold. Las Vegas' baseline predicted employment share, 32 percent, suggests the zero predicted share under the tight threshold is largely a type-2 error. Similarly, the 18 percentage point maximum type-2 error observed in the out-of-sample fit for the labeled metros (Figure 8, rightmost bar) suggests that much of the Las Vegas' imprecision arises from a type-2 error. More generally, imprecise predictions are likely to arise from both type-1 and type-2 errors. The baseline predicted CEZ employment share sometimes falls at the bottom of large 95/50 spreads (e.g., Charleston WV), sometimes in the middle of the spread (e.g., Honolulu), and sometimes at the top of it (e.g., Atlantic City).[16]

## 4.3    CEZ Employment and Population Density

Employment density, the most important consideration in our subjective classification, was obviously much higher on average in CEZs than average employment density elsewhere in the same metro. But employment density also varied considerably within CEZs.

For each metro, we calculated a CEZ's mean employment density in 2000 as the employment-weighted mean of each CEZ tract's raw employment density, thereby capturing mean employment density as experienced by CEZ workers. Measured this way, CEZ mean employment density was distributed approximately log normally across metros with a bit of a right skew and elongated right tail (Figure 11). The mean across metros of CEZ mean employment density was 35,000 workers per square mile, more than ten times the mean across metros of mean employment density in the remaining non-rural portions (Table 6, first vertical block). The variation across metros in CEZ mean employment density was especially large, ranging more than 100-fold from less than 5,000 to more than 500,000 workers per square mile.

---

[16]In the next revision of the paper, metros with especially imprecise predictions may be good candidates to add to the training samples because they might help "clarify" how the learning algorithm should treat similar ambiguities in other metros. Doing so would not affect the imprecision of the prediction of the added
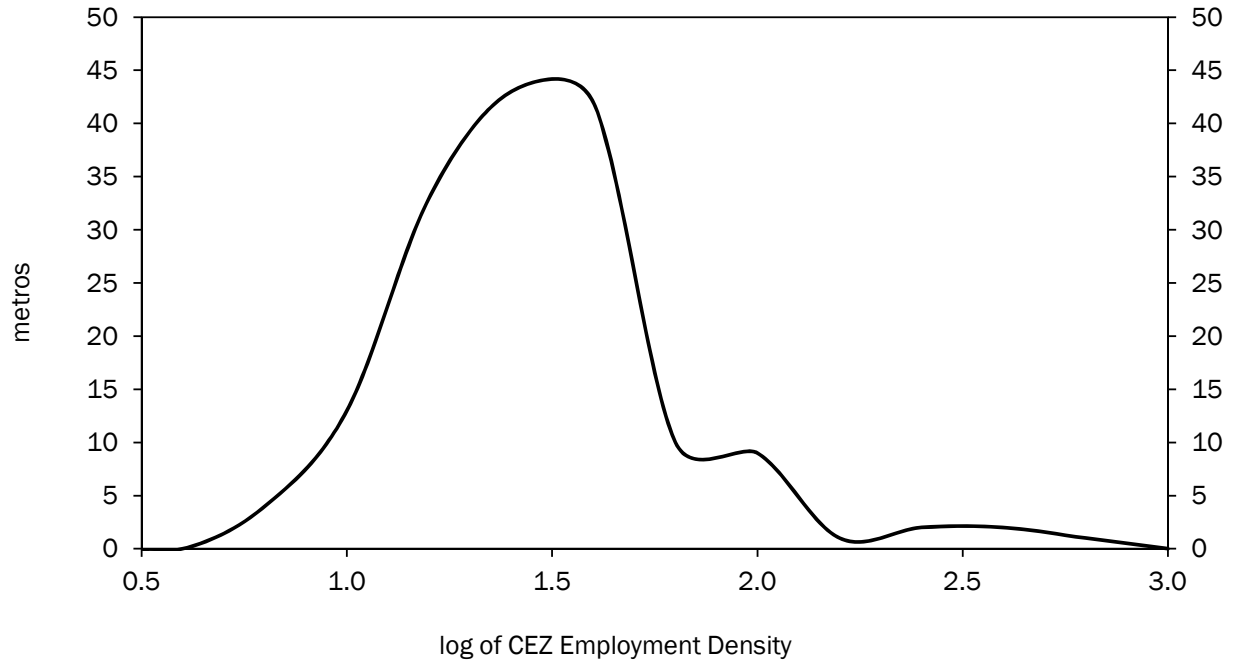
Figure 11: **Distribution of CEZ Mean Employment Density** Figure shows the distribution across metros of the log of CEZ mean employment density.

Most CEZs included census tracts with employment density considerably below the CEZ's weighted mean and other census tracts with employment density considerably above the CEZ's weighted mean. The respective means across metros of the minimum and maximum employment density within each CEZ were 10,000 and 60,000 workers per square mile (Table 6, second and third vertical blocks). Across metros, minimum CEZ employment density ranged from less than 1,000 to almost 190,000 workers per square mile. Maximum CEZ employment density ranged from 5,000 to almost 1,000,000 workers per square mile. On average across metros, employment density varied by a multiplicative factor of 8 within CEZs.

Population density was also higher on average in CEZs than in the remaining non-rural portions of metropolitan areas. The mean across metros of the population-weighted mean population density across CEZ tracts was 7,400 residents per square mile, almost twice

---

metro as we will be doing all predictions out of sample.

|  | | Central Employment Zone | | | | Non-Rural Remainder of CBSA | | | |
| Attribute | Metro Size | Mean | Std. Dev. | Min | Max | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|---|---|---|---|
| Mean (emp wghtd, ths per sqmi) | All | 35.0 | 55.9 | 4.9 | 509.5 | 2.8 | 2.5 | 0.8 | 23.6 |
|  | Small | 19.1 | 10.2 | 4.9 | 81.6 | 2.2 | 1.0 | 0.8 | 9.2 |
|  | Medium | 41.4 | 19.4 | 11.5 | 85.9 | 3.0 | 1.1 | 1.7 | 6.7 |
|  | Large | 195.4 | 143.0 | 43.1 | 509.5 | 9.2 | 6.6 | 3.9 | 23.6 |
| Minimum Tract (ths per sqmi) | All | 10.2 | 18.6 | 2.2 | 189.0 | 0.1 | 0.0 | 0.0 | 0.2 |
|  | Small | 7.5 | 4.0 | 2.2 | 23.8 | 0.1 | 0.0 | 0.0 | 0.2 |
|  | Medium | 6.8 | 2.7 | 2.3 | 15.1 | 0.0 | 0.0 | 0.0 | 0.1 |
|  | Large | 53.5 | 60.4 | 4.2 | 189.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Maximum Tract (ths per sqmi) | All | 60.4 | 110.9 | 5.0 | 984.0 | 21.8 | 69.0 | 1.8 | 790.7 |
|  | Small | 27.8 | 24.9 | 5.0 | 247.1 | 8.3 | 8.1 | 1.8 | 72.0 |
|  | Medium | 82.1 | 48.0 | 13.7 | 213.6 | 20.3 | 13.4 | 4.9 | 67.3 |
|  | Large | 359.5 | 292.9 | 100.6 | 984.0 | 181.5 | 228.1 | 35.9 | 790.7 |

Table 6: **CEZ Employment Density**. Table reports employment density in 2000 for the CEZ and remaining non-rural portion of 160 CBSAs (115 small, 35 medium, 10 large). The non-rural portion of a CBSA comprises all census tracts with either population density or employment density of at least 500 per square mile.

the similarly-calculated mean across remaining census tracts in the non-rural portion of the metro (Table 7, first vertical block). As with employment density, population density typically varied considerably within CEZs.

On average, all tracts within CEZs had at least moderate population density. For example, the mean across metros of the minimum population density within CEZs was 2,400 residents per square mile (second vertical block), more than half the mean across metros of population-weighted mean across tracts in the non-rural remainder of metros. Similarly, most CEZs included at least one tract with relatively high population density. The mean across metros of the maximum density within CEZs was 13,300 residents per square mile, almost three quarters the mean across metros of the maximum tract density within the non-rural remainder of metros (third vertical block).

|  | | Central Employment Zone | | | | Non-Rural Remainder of CBSA | | | |
| Attribute | Metro Size | Mean | Std. Dev. | Min | Max | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|---|---|---|---|
| Mean (pop wghtd, ths per sqmi) | All | 7.4 | 7.5 | 0.7 | 68.8 | 3.9 | 3.0 | 1.1 | 33.5 |
| | Small | 6.1 | 5.1 | 0.7 | 40.0 | 3.3 | 1.5 | 1.1 | 10.7 |
| | Medium | 8.0 | 5.2 | 1.7 | 29.0 | 4.4 | 1.5 | 2.1 | 8.6 |
| | Large | 19.8 | 19.4 | 5.1 | 68.8 | 9.8 | 8.8 | 2.6 | 33.5 |
| Minimum Tract (ths per sqmi) | All | 2.4 | 1.9 | 0.0 | 8.4 | 0.3 | 0.2 | 0.0 | 0.8 |
| | Small | 2.8 | 1.8 | 0.1 | 8.4 | 0.4 | 0.2 | 0.0 | 0.8 |
| | Medium | 1.1 | 1.0 | 0.0 | 4.6 | 0.1 | 0.1 | 0.0 | 0.5 |
| | Large | 2.8 | 2.5 | 0.0 | 6.6 | 0.0 | 0.1 | 0.0 | 0.2 |
| Maximum Tract (ths per sqmi) | All | 13.3 | 21.2 | 0.7 | 176.5 | 18.0 | 23.3 | 2.8 | 229.8 |
| | Small | 9.5 | 13.1 | 0.7 | 114.6 | 12.0 | 9.6 | 2.8 | 79.3 |
| | Medium | 18.9 | 28.7 | 2.1 | 176.5 | 20.3 | 9.5 | 7.6 | 45.3 |
| | Large | 38.5 | 40.0 | 5.9 | 123.4 | 79.6 | 58.4 | 32.6 | 229.8 |

Table 7: **CEZ Population Density** Table reports population density in 2000 for the CEZ and remaining non-rural portion of 160 CBSAs (115 small, 35 medium, 10 large). The non-rural portion of a CBSA comprises all census tracts with either population density or employment density of at least 500 per square mile.

# 5 Conclusion

Numerous papers have been written that investigate the importance of centralized employment for various socioeconomic outcomes in U.S. metropolitan areas. The Central Business District remains the main concept and approach researchers have used to test for these relationships. However, in comparison, less work has been done on actually defining the CBD in each metro. Most prior studies that have analyzed this issue only looked at few metros likely because of the difficult and time intensive nature of the task.

Our solution to this problem is to develop an algorithm using machine learning techniques that replicates the subjective process of identifying tracts that are in a metro's CBD and apply the process to a much larger number of metros. We propose a new and broader concept

of the CBD, which we call the centralized employment zone (CEZ). We define a CEZ to be the Central Business District together with nearby concentrated employment. This broadening makes sense in the context of distinguishing employment that is centralized from employment that is located in clusters further away from the CBD and from employment that is spread diffusely throughout a metropolitan area.

Our results reveal a consistent pattern of centralized employment across metros between 250 thousand and 2.7 million people. On average, the CEZ contains 12 percent of total employment and 20 percent of agglomerative occupations. Using our identification of tracts in CEZs, future research could investigate to what extent centrality affects other outcomes in metropolitan areas.

# References

**Alonso, William.** 1964. *Location and Land Use: Toward a General Theory of Land Rent.* Harvard University Press.

**Anderson, Nathan B., and William T. Bogart.** 2001. "The Structure of Sprawl: Identifying and Characterizing Employment Centers in Polycentric Metropolitan Areas." *American Journal of Economics and Sociology, Inc. , Special Issue: City and Country: An Interdisciplinary Collection*, 60(1): 147–169.

**Arzaghi, Mohammad, and Vernon J. Henderson.** 2008. "Networking off Madison Avenue." *The Review of Economic Studies*, 75(4): 1011–1038.

**Asabere, Paul, and Forrest Huffman.** 1991. "Historic Districts and Land Values." *Journal of Real Estate Research*, 6(1): 1–7.

**Atack, Jeremy, and Robert Margo.** 1998. "Location, Location, Location! The price gradient for vacant urban land: New York, 1835-1900." *Journal of Real Estate Finance and Economics*, 16(2): 151–172.

**Baum-Snow, Nathaniel.** 2014. "Urban Transport Expansions, Employment Decentralization, and the Spatial Scope of Agglomeration Economies."

**Baum-Snow, Nathaniel, and Daniel Hartley.** 2016. "Accounting for Central Neighborhood Change, 1980-2010." Federal Reserve Bank of Chicago, Research Working Paper, WP16-09.

**Bogart, William T., and William C. Ferry.** 1999. "Employment Centres in Greater Cleveland: Evidence of Evolution in a Fomerly Monocentric City." *Urban Studies*, 36(12): 2099–2110.

**Brinkman, Jeffrey.** 2016. "Congestion, Agglomeration, and the Structure of Cities." Federal Reserve Bank of Philadelphia, Working Paper 16-13.

**Brinkman, Jeffrey, Daniele Coen-Priani, and Holger Seig.** 2016. "The Political Economy of Underfunded Municipal Pension Plans." Federal Reserve Bank of Philadelphia, Working Paper 16-16.

**Burgess, Ernest.** 1925. "The growth of the city: an introduction to a research project." In *The Trend of Population.* Chapter 8, 85–97. American Sociological Society.

**Combes, Pierre-Philippe, Gilles Duranton, and Laurent Gobillion.** 2008. "Spatial Wage Disparities: Sorting Matters!" *Journal of Urban Economics*, 63: 723–742.

**Elvery, Joel A., and Leo Sveikauskas.** 2011. "Urban Centers and Subcenters: Their Traits and Specializations." Cleveland State University Working Paper.

**Friedman, Jerom, Trevor Hastie, and Robert Tibshirani.** 2000. "Additive Logistic Regression: A Statistical View of Boosting." *The Annals of Statistics*, 28(2): 337–374.

**Glaeser, Edward L., and David C. Maré.** 2001. "Cities and Skills." *Journal of Labor Economics*, 19(2): 316–342.

**Glaeser, Edward L., and Matthew E. Kahn.** 2004. "Sprawl and Urban Growth." In *Handbook of Regional and Urban Economics.* , ed. J. Vernon Henderson and Jacques Francoise Thisse, 2481–2527. Elsevier North-Holland.

**Gleaser, Edward.** 2011. *Triumph of the City.* The Penguin Press: New York.

**Holian, Matthew J., and Matthew E. Kahn.** 2012. *The Impact of Central City Economic and Cultural Vibrancy of Greenhouse Gas Emissions from Transportation.* San Jose, CA:MIT Publications. Dataset avaivle from http://mattholian.blogspot.com/2013/05/central-business-district-geocodes.html.

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

**Limehouse, Frank F., and Robert E. McCormick.** 2011. "Impacts of Central Business District Location: A Hedonic Analysis of Legal Service Establishments." U.S. Census Bureau Center for Economic Studies Working Paper 11-21.

**Marlay, Matthew, and Todd K. Gardner.** 2010. "Identifying Concentrations of Employment in Metropolitan Areas." Working Paper.

**McMillen, Daniel P.** 2001. "Nonparametric Employment Subcenter Identification." *Journal of Urban Economics*, 50: 448–473.

**McMillen, Daniel P., and Stefani C. Smith.** 2003. "The number of subcenters in large urban areas." *Journal of Urban Economics*, 53: 321–338.

**Mills, Edwin S.** 1967. "An Aggregation Model of Resource Allocation in a Metropolitan Area." *American Economic Review*, 57: 197–210.

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

**Rappaport, Jordan.** 2008. "A Productivity Model of City Crowdedness." *Journal of Urban Economics*, 63: 715–722.

**Rappaport, Jordan.** 2014. "Monocentric City Redux." Federal Reserve Bank of Kansas City, Working Paper 14-09.

**Rappaport, Jordan.** 2016. "Productivity, Congested Commuting, and Metro Size." Federal Reserve Bank of Kansas City Research Working Paper 16-03.

**Rappaport, Jordan.** 2017. "Crowdedness, Centralized Employment, and Multifamily Home Construction." Federal Reserve Bank of Kansas City *Economic Review*, 102(1): 41–83.

**Redfearn, Christian L.** 2007. "The topography of metropolitan employment: Identifying centers of employment in a polycentric urban area." *Journal of Urban Economics*, 61: 519–541.

**Rosenthal, Stuart S., and William C. Strange.** 2003. "Geography, Industrial Organization, and Agglomeration." *Review of Economics and Statistics*, 85(2): 377–393.

**Rosenthal, Stuart S., and William C. Strange.** 2008. "The Attenuation of Human Capital Spillovers." *Journal of Urban Economics*, 64(2): 373–389.

**Schuetz, Jenny, Arturo Gonzalez, Jeff Larrimore, Ellen A. Merry, and Barbara Robles.** 2017. "Are Central Cities Poor and Non-White?" Federal Reserve Board, Washington, D.C. Finance and Economics Discussion Series 2017-031.

**Small, Kenneth A., and Shunfeng Song.** 1993. "Population and Employment Densities: Structure and Change." *Journal of Urban Economics*, 292–313.

**U.S. Bureau of the Census.** 1987. *History of the 1982 Economic Censuses.* Washington D.C.:U.S. Government Printing Office.

# Appendix A: Supplemental Tables

| | metro | population | highest tract inclusion probability |
|---|---|---|---|
| 1 | Virginia Beach-Norfolk-Newport News, VA-NC | 1,576,000 | 0.76 |
| 2 | Detroit-Warren-Livonia, MI | 4,453,000 | 0.68 |
| 3 | Riverside-San Bernardino-Ontario, CA | 3,255,000 | 0.61 |
| 4 | Poughkeepsie-Newburgh-Middletown, NY | 622,000 | 0.60 |
| 5 | McAllen-Edinburg-Pharr, TX | 569,000 | 0.55 |
| 6 | Scranton--Wilkes-Barre, PA | 561,000 | 0.48 |
| 7 | Los Angeles-Long Beach-Santa Ana, CA | 12,365,000 | 0.45 |
| 8 | Fayetteville, NC | 337,000 | 0.43 |
| 9 | Gulfport-Biloxi, MS | 246,000 | 0.39 |
| 10 | Fayetteville-Springdale-Rogers, AR-MO | 347,000 | 0.36 |
| 11 | Barnstable Town, MA | 222,000 | 0.35 |
| 12 | Killeen-Temple-Fort Hood, TX | 331,000 | 0.34 |
| 13 | Port St. Lucie-Fort Pierce, FL | 319,000 | 0.33 |
| 14 | Youngstown-Warren-Boardman, OH-PA | 603,000 | 0.33 |
| 15 | Visalia-Porterville, CA | 368,000 | 0.31 |
| 16 | Clarksville, TN-KY | 232,000 | 0.29 |
| 17 | Sarasota-Bradenton-Venice, FL | 590,000 | 0.27 |
| 18 | Kingsport-Bristol, TN-VA | 230,000 | 0.22 |
| 19 | Naples-Marco Island, FL | 251,000 | 0.20 |
| 20 | Vallejo-Fairfield, CA | 395,000 | 0.09 |
| 21 | Holland-Grand Haven, MI | 238,000 | 0.06 |
| 22 | Oxnard-Thousand Oaks-Ventura, CA | 753,000 | 0.04 |
| 23 | Palm Bay-Melbourne-Titusville, FL | 476,000 | 0.03 |

Table A.1: **Metropolitan Areas with No Predicted CEZ**. Table lists the 23 metros that have no CEZ tracts using the baseline 0.80 inclusion threshold. The third column gives the highest inclusion frequency of a metro's tracts. For example, at least one tract in the Virginia Beach-Norfolk-Newport News metro is sorted into leaf nodes of the 100 decision trees with weighted frequency of CEZ inclusion equal to 0.76. Hence we would delineate a CEZ for it using an inclusion threshold of 0.76 or less.

| Attribute | Metro Size | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| Land Area | All | 0.15 | 0.13 | 0.00 | 0.74 |
| | Small | 0.12 | 0.12 | 0.00 | 0.74 |
| | Medium | 0.17 | 0.10 | 0.04 | 0.41 |
| | Large | 0.34 | 0.11 | 0.18 | 0.52 |
| Population | All | 0.73 | 0.14 | 0.35 | 0.99 |
| | Small | 0.69 | 0.13 | 0.35 | 0.95 |
| | Medium | 0.81 | 0.09 | 0.57 | 0.97 |
| | Large | 0.91 | 0.05 | 0.80 | 0.99 |
| Employment | All | 0.84 | 0.08 | 0.62 | 0.98 |
| | Small | 0.81 | 0.08 | 0.62 | 0.97 |
| | Medium | 0.90 | 0.04 | 0.82 | 0.96 |
| | Large | 0.95 | 0.02 | 0.92 | 0.98 |

Table A.2: **Non-Rural Portions of CBSAs** Table reports the share of land, population, and employment located in the non-rural portion of the 183 CBSAs (134 small, 37 medium, 12 large) to which we applied the learned decision trees. The non-rural portion of a CBSA comprises all census tracts with either population density or employment density of at least 500 per square mile.

| | Occupation | Mean | Std. Dev. | Min | 80th pctile | Max |
|---|---|---|---|---|---|---|
| 1 | *legal* | 0.41 | 0.22 | 0 | 0.59 | 0.80 |
| 2 | protective service | 0.21 | 0.13 | 0 | 0.32 | 0.56 |
| 3 | healthcare practitioners and technicians | 0.19 | 0.16 | 0 | 0.34 | 0.58 |
| 4 | *computer and mathematical* | 0.19 | 0.13 | 0 | 0.28 | 0.67 |
| 5 | *business and financial operations specialists* | 0.18 | 0.11 | 0 | 0.28 | 0.46 |
| 6 | *life, physical, and social science* | 0.18 | 0.13 | 0 | 0.29 | 0.51 |
| 7 | *arts, design, entertainment, sports, and media* | 0.17 | 0.11 | 0 | 0.27 | 0.42 |
| 8 | community and social service | 0.17 | 0.10 | 0 | 0.25 | 0.40 |
| 9 | office and administrative support | 0.15 | 0.09 | 0 | 0.22 | 0.43 |
| 10 | management (non-farm) | 0.14 | 0.09 | 0 | 0.21 | 0.40 |
| 11 | architecture and engineering | 0.12 | 0.09 | 0 | 0.19 | 0.45 |
| 12 | healthcare support | 0.12 | 0.09 | 0 | 0.20 | 0.40 |
| 13 | building and grounds cleaning and maintenance | 0.12 | 0.08 | 0 | 0.17 | 0.53 |
| 14 | food preparation and serving related | 0.11 | 0.09 | 0 | 0.16 | 0.61 |
| 15 | education, training, and library | 0.10 | 0.09 | 0 | 0.16 | 0.43 |
| 16 | sales and related | 0.09 | 0.06 | 0 | 0.13 | 0.35 |
| 17 | armed forces | 0.08 | 0.14 | 0 | 0.17 | 0.68 |
| 18 | construction and excavation | 0.08 | 0.06 | 0 | 0.13 | 0.31 |
| 19 | personal care and service | 0.08 | 0.08 | 0 | 0.11 | 0.80 |
| 20 | installation, maintenance, and repair | 0.08 | 0.06 | 0 | 0.12 | 0.33 |
| 21 | transportation and material moving | 0.07 | 0.06 | 0 | 0.11 | 0.32 |
| 22 | production | 0.07 | 0.06 | 0 | 0.10 | 0.43 |

Table A.3: **Employment Centralization by Occupation** Table summarizes the baseline CEZ share of employment in 2000 by occupation group for the 183 CBSAs to which we applied the learned prediction functions. Italics indicate occupations we categorize as agglomerative.

| rank | CBSA | Central Employment Zone | | | | Non-Rural Remainder of CBSA | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | emp-wghtd mean (ths per sqmi) | min | max | pop-wghtd mean | emp-wghtd mean | min | max | pop-wghtd mean |
| | All metros (mean) | 35.0 | 10.2 | 60.4 | 21.3 | 2.8 | 0.1 | 21.8 | 1.4 |
| 1 | Atlantic City, NJ | 30.2 | 5.3 | 39.4 | 20.2 | 1.3 | 0.1 | 5.9 | 1.1 |
| 2 | *Spokane, WA* | 18.1 | 4.5 | 32.1 | 12.8 | 1.8 | 0.0 | 7.0 | 1.1 |
| 3 | Las Vegas, NV | 24.0 | 6.6 | 53.6 | 15.3 | 3.2 | 0.0 | 11.0 | 1.5 |
| 4 | *Topeka, KS* | 11.0 | 3.6 | 12.2 | 7.7 | 1.3 | 0.1 | 2.9 | 1.0 |
| 5 | *Columbia, SC* | 22.6 | 2.5 | 59.8 | 6.5 | 1.0 | 0.1 | 3.0 | 0.7 |
| 6 | Honolulu, HI | 81.6 | 11.0 | 247.1 | 33.5 | 9.2 | 0.0 | 72.0 | 2.6 |
| 7 | Charleston, WV | 28.5 | 7.1 | 37.4 | 15.0 | 2.3 | 0.1 | 6.5 | 1.1 |
| 8 | Gainesville, FL | 13.8 | 5.7 | 20.3 | 11.5 | 2.0 | 0.1 | 3.7 | 1.2 |
| 9 | *Tallahassee, FL* | 12.5 | 5.1 | 16.6 | 11.5 | 2.1 | 0.1 | 3.5 | 1.2 |
| 10 | Anchorage, AK | 15.5 | 5.4 | 29.4 | 9.5 | 2.2 | 0.1 | 5.2 | 1.1 |
| 11 | *Lansing, MI* | 21.0 | 4.2 | 48.6 | 11.1 | 1.5 | 0.0 | 4.0 | 1.1 |
| 12 | Asheville, NC | 10.9 | 3.5 | 25.3 | 6.7 | 0.9 | 0.0 | 2.0 | 0.5 |
| 13 | *Ann Arbor, MI* | 36.8 | 4.2 | 54.6 | 18.2 | 3.0 | 0.1 | 11.1 | 1.6 |
| 14 | *Syracuse, NY* | 31.1 | 3.5 | 53.4 | 19.4 | 1.7 | 0.0 | 8.2 | 1.0 |
| 15 | *Des Moines, IA* | 21.0 | 4.9 | 23.3 | 13.4 | 2.1 | 0.1 | 7.2 | 1.2 |
| 16 | Austin, TX | 39.4 | 6.7 | 64.3 | 22.9 | 2.9 | 0.1 | 8.7 | 1.7 |
| 17 | Madison, WI | 26.6 | 7.2 | 46.8 | 20.0 | 2.0 | 0.1 | 5.8 | 1.4 |
| 18 | Jackson, MS | 6.7 | 3.0 | 8.4 | 6.5 | 1.2 | 0.0 | 3.4 | 0.8 |
| 19 | *Portland, OR* | 53.3 | 7.2 | 175.5 | 23.9 | 2.9 | 0.1 | 15.2 | 1.5 |
| 20 | Columbus, GA | 11.2 | 4.6 | 18.9 | 6.6 | 1.9 | 0.1 | 6.2 | 1.0 |
| 21 | Denver, CO | 52.8 | 4.3 | 135.3 | 17.2 | 3.3 | 0.1 | 14.1 | 1.5 |
| 22 | Birmingham, AL | 15.8 | 5.6 | 23.6 | 11.4 | 1.7 | 0.1 | 4.9 | 0.9 |
| 23 | Little Rock, AR | 16.9 | 3.6 | 28.2 | 6.4 | 1.6 | 0.1 | 3.7 | 1.0 |
| 24 | Cedar Rapids, IA | 19.7 | 5.6 | 26.6 | 14.6 | 1.2 | 0.1 | 3.8 | 1.0 |
| 25 | *Sacramento, CA* | 34.5 | 3.5 | 118.2 | 13.1 | 2.4 | 0.1 | 8.3 | 1.3 |
| 26 | New Orleans, LA | 62.4 | 7.8 | 108.9 | 17.5 | 4.1 | 0.0 | 25.2 | 1.7 |
| 27 | Albuquerque, NM | 15.6 | 6.4 | 34.1 | 11.4 | 2.5 | 0.0 | 8.2 | 1.3 |
| 28 | Macon, GA | 14.1 | 3.7 | 20.1 | 6.3 | 1.0 | 0.1 | 2.6 | 0.7 |
| 29 | *Pittsburgh, PA* | 76.9 | 6.8 | 144.1 | 29.1 | 2.6 | 0.0 | 23.9 | 1.3 |

Table A.4: **Employment Density in More-Centralized Metros** Table reports employment density in the CEZ and remaining non-rural portion of metros with high employment centralization. Labeled metros are italicized.

| rank | metro | Central Employment Zone | | | | Non-Rural Remainder of CBSA | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | pop-wghtd mean (ths per sqmi) | min | max | emp-wghtd mean | pop-wghtd mean (psqmi) | min | max | emp-wghtd mean |
| 1 | Atlantic City, NJ | 11.3 | 4.7 | 27.9 | 6.7 | 4.6 | 0.5 | 24.9 | 3.6 |
| 2 | *Spokane, WA* | 4.5 | 0.6 | 5.9 | 3.4 | 3.7 | 0.5 | 8.6 | 3.1 |
| 3 | Las Vegas, NV | 6.6 | 0.0 | 15.4 | 4.3 | 6.8 | 0.0 | 24.2 | 4.5 |
| 4 | *Topeka, KS* | 4.2 | 1.4 | 6.1 | 2.2 | 2.7 | 0.5 | 5.5 | 2.1 |
| 5 | *Columbia, SC* | 5.4 | 0.9 | 16.9 | 2.4 | 1.9 | 0.2 | 6.4 | 1.7 |
| 6 | Honolulu, HI | 26.2 | 2.3 | 67.7 | 12.7 | 10.7 | 0.0 | 79.3 | 7.2 |
| 7 | Charleston, WV | 6.3 | 2.5 | 8.9 | 4.0 | 2.2 | 0.5 | 6.6 | 2.1 |
| 8 | Gainesville, FL | 6.0 | 3.1 | 7.3 | 6.1 | 2.8 | 0.5 | 8.7 | 2.7 |
| 9 | *Tallahassee, FL* | 5.8 | 2.2 | 11.0 | 4.0 | 2.7 | 0.6 | 7.6 | 2.4 |
| 10 | Anchorage, AK | 4.4 | 2.4 | 6.1 | 3.5 | 4.1 | 0.8 | 8.8 | 3.3 |
| 11 | *Lansing, MI* | 8.5 | 0.2 | 14.5 | 5.3 | 2.8 | 0.6 | 7.6 | 2.3 |
| 12 | Asheville, NC | 2.2 | 1.5 | 3.6 | 2.4 | 1.2 | 0.5 | 3.5 | 1.4 |
| 13 | *Ann Arbor, MI* | 12.1 | 5.6 | 15.7 | 11.2 | 3.8 | 0.3 | 19.4 | 2.9 |
| 14 | *Syracuse, NY* | 9.1 | 0.2 | 13.3 | 7.7 | 3.7 | 0.4 | 16.3 | 2.3 |
| 15 | *Des Moines, IA* | 4.0 | 1.8 | 6.9 | 2.2 | 3.3 | 0.4 | 8.0 | 2.6 |
| 16 | Austin, TX | 9.0 | 1.6 | 19.7 | 4.8 | 4.1 | 0.2 | 17.6 | 3.4 |
| 17 | Madison, WI | 18.5 | 6.8 | 46.3 | 14.3 | 2.9 | 0.3 | 10.1 | 2.5 |
| 18 | Jackson, MS | 2.9 | 0.6 | 4.7 | 1.9 | 2.2 | 0.5 | 6.8 | 1.9 |
| 19 | *Portland, OR* | 9.1 | 0.1 | 23.6 | 6.8 | 4.5 | 0.1 | 14.1 | 3.4 |
| 20 | Columbus, GA | 3.3 | 0.7 | 4.2 | 2.1 | 2.7 | 0.7 | 7.6 | 2.4 |
| 21 | Denver, CO | 8.8 | 2.0 | 21.2 | 6.9 | 5.2 | 0.2 | 25.9 | 3.6 |
| 22 | Birmingham, AL | 4.1 | 1.3 | 8.8 | 2.7 | 2.4 | 0.3 | 9.3 | 2.1 |
| 23 | Little Rock, AR | 2.7 | 0.6 | 3.2 | 1.5 | 2.2 | 0.3 | 4.7 | 2.1 |
| 24 | Cedar Rapids, IA | 4.5 | 2.1 | 5.4 | 4.4 | 2.7 | 0.2 | 6.8 | 1.9 |
| 25 | *Sacramento, CA* | 6.7 | 0.7 | 14.6 | 5.4 | 5.1 | 0.1 | 16.4 | 3.4 |
| 26 | New Orleans, LA | 10.7 | 1.5 | 19.4 | 3.5 | 6.6 | 0.1 | 40.3 | 4.8 |
| 27 | Albuquerque, NM | 3.4 | 1.4 | 6.4 | 3.0 | 3.9 | 0.6 | 12.6 | 3.3 |
| 28 | Macon, GA | 3.8 | 0.9 | 6.4 | 1.6 | 2.1 | 0.6 | 6.0 | 1.8 |
| 29 | *Pittsburgh, PA* | 12.1 | 0.1 | 24.4 | 8.6 | 4.0 | 0.2 | 25.1 | 3.3 |

Table A.5: **Population Density in More-Centralized Metros** Table reports population density in the CEZ and remaining non-rural portion of metros with high employment centralization. Labeled metros are italicized.