Driving Past Commuting Zones: Re-examining Local Labor Market Definitions*

Andrew Foote[1], Mark Kutzbach[1], and Lars Vilhuber[1,2]

[1]Center for Economic Studies, U.S. Census Bureau
[2]Labor Dynamics Institute, Cornell University

December 21, 2016

[**Preliminary Draft - Do not cite without authors' permission**]

Abstract

This paper evaluates the suitability of local labor market definitions for representing the theoretical and empirical goal of assigning workers to distinct loci of commuting, wages, and employment opportunities. First, we revisit the commuting zone definitions from Tolbert and Sizer (1996). We discuss their methodology, highlighting two main weaknesses that affect the resulting labor market definitions. We then demonstrate how these weaknesses affect empirical estimates from a prominent paper using commuting zones. Finally, we define an objective function to measure the degree of integration in local labor markets and maximize it over an alternative clustering method. We conclude by comparing our method to other candidate local labor market definitions in the literature, and show that it is competitive with these definitions.

1 Introduction

Local labor markets are an important unit of analysis in labor economics, both in the theoretical and empirical literatures. Theoretical papers emphasize characteristics of a local labor market including common wage and rent levels (Roback, 1982; Moretti, 2011), as well as job-finding and unemployment rates (Head and Lloyd-Ellis, 2012; Schmutz and Sidibé, 2014).

*Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed, and no confidential data was used in this paper. This document is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

In empirical labor economics, researchers may be interested in estimating the effect of some local, exogenous shock on labor market outcomes, and an important decision in any research design is defining the area that is directly affected by the shock. In estimating this effect, researchers have a number of different definitions of local labor markets from which to choose. Blanchard and Katz (1992), Wozniak (2010), and Kennan and Walker (2011) use the state, while other researchers use metropolitan areas (Bound and Holzer, 2000; Card, 2001; Notowidigdo, 2011; Diamond, forthcoming), while still others use counties (Monte, Redding and Rossi-Hansberg, 2015; Foote, Grosz and Stevens, 2015).

One alternative labor market definition is commuting zones, which were originally defined by Tolbert and Sizer (1996) (henceforth, TS1996). Commuting zones have two distinct advantages over the above definitions. First, they span the entire United States, allowing researchers to measure effects for the entire country rather than just metropolitan areas. Second, commuting zones group together counties based on commuting flows, which implies some level of economic integration (metropolitan areas are also based on commuting flows). The definition acknowledges that labor markets are not constrained by county and state lines, but are based on relevant linkages between counties. While the strict assignment of counties to one commuting zone or another overstates the distinctiveness of labor markets, this simplifying assumption facilitates many empirical applications.

Given these advantages, commuting zones have been used in a number of influential papers in the labor economics literature, including Autor, Dorn and Hanson (2013), as well as Chetty et al. (2014), Amior and Manning (2015), Restrepo (2015), and Yagan (2016). Despite their widespread use, to the best of our knowledge, the methodology underlying commuting zone definitions and its impact on empirical estimates has not received much scrutiny.

Our paper makes three contributions to the literature on local labor markets. First, we outline a number of methodological issues that researchers should be aware of when they use the commuting zone definitions. Second, we show how these methodological issues impact empirical estimates. Our findings suggest that researchers should consider evaluating the sensitivity of results to local labor market definitions. We propose two main ways that researchers can test to see if their results are robust to the uncertainty induced by commuting zones. Finally, we propose an alternative method similar to the one employed by Tolbert and Sizer, which allows us to measure integration beyond just commuting flows. We define a metric for the quality of a regionalization definition, and show that our proposed method performs better than alternative methods.

The remainder of the paper proceeds as follows. We describe the data we use in Section 2, and then in Section 3 describes the commuting zone definition and methodology in detail. In Section 4, we discuss a number of issues with the commuting zone methodology. In Section 5, we demonstrate how these issues impact empirical estimates. Finally, in Sections 6 and 7, we discuss our measure of local labor market integration and methodology for defining local labor markets. Section 8 summarizes and discusses next steps.

2 Data

The primary data set that we use is the 1990 Journey to Work (JTW) data, which is derived from the 1990 Census long form. In the JTW questions, employed respondents are asked "at what location did this person work LAST WEEK (If this person worked at more than one location, print where he or she worked most last week." The 1990 JTW data geocodes these responses and publishes estimated home-to-work flows for all counties and county equivalents in the United States.[1] The Office of Management and Budget (OMB) uses JTW in its guidelines for defining Metropolitan Statistical Areas, a geographic unit used for research, statistical publications, and federal programs.[2] Tolbert and Sizer (1996) use the 1990 JTW data for their analysis. We use the 1990 JTW data to replicate TS1996's Commuting Zone definitions, as well as estimate our alternate labor market definitions.

[1] This data, as well as the ACS data described later, is available for download from the Census website, at http://www.census.gov/hhes/commuting/data/commutingflows.html

[2] See the Federal Register Part IX, Office of Management and Budget: Standards for Defining Metropolitan and Micropolitan Statistical Areas; Notice, December 27, 2000, available at http://www.census.gov/population/metro/files/00-32997.pdf

Table 1: Commuting Summary Statistics

Dataset	Counties	Total Residential Labor Force	Mean Residential Labor Force	Mean Flows Home-to-work	Mean Within-county Share of Flows
Journey-to-Work (JTW) Estimates					
1990 Census	3,105	115,004,732	460,966	392,408	0.76
2000 Census	3,107	128,170,381	444,149	369,102	0.74
2009-2013 ACS	3,143	139,679,674	505,765	419,784	0.73

Notes: Authors' calculations, using public-use county-to-county flows from Journey to Work estimates for the 1990 Census, 2000 Census, and the 2009-2013 5-Year ACS.

In addition to the 1990 JTW data, we also use the 2000 JTW data, and the 2009-2013 5-Year Pooled American Community Survey JTW data. The ACS data are similar to the Decennial Census data, but include fewer pairs of counties in the published flows data. The JTW from ACS also includes published margins of error for these flows, whereas the 1990 and 2000 JTW data do not include these measures.

Table 1 provides summary statistics on the worker flows data from the three main data sources. We summarize the number of residence counties in the file, the total count of jobs included in the flows, the mean residence labor force and home-to-work flow sizes, and the mean share of workers who work in the same county where they live. We note the rise in population, or total job count, from 1990 through 2009, as well as the declining share of within-county commuting during that period (a lower within-county share). The decline in the share of within-county commuting, which shows an increased distance of jobs from where workers live, is consistent both with studies of Census and ACS data showing increasing travel time in recent years (McKenzie and Rapino, 2011).

We use a number of other data sources to measure local labor market integration in Section 6: county-level unemployment rates from the Bureau of Labor Statistics Local Area Unemployment Statistics series; and annual earnings data from the Bureau of Economic Affairs Regional Economic Indicators Series (REIS), both from 1990-1995, which aligns most closely with the timing of our commuting data.

3 Commuting Zone Methodology

3.1 Background

As an alternative to metropolitan statistical areas (or Core Based Statistical Areas), many researchers use Commuting Zones because they cover the entire country and group counties based on commuting flows, which is a measure of labor market integration. However, few researchers are familiar with the underlying methodology. To that end, this section describes the commuting zone methodology, as developed by TS1996.

The methodology was originally developed in Tolbert and Killian (1987), but the 1996 paper is much more widely cited. The Economic Research Service (ERS), an agency under the Department of Agriculture for which this methodology was developed, distributes commuting zone definitions on its website. Commuting Zones are especially relevant for the economic analysis of rural areas because the include all counties, not just urban counties. ERS also distributes a version of Commuting Zones based on 2000 data. For a research into extending the time series of Commuting Zones and further background on TS1996, see Fowler, Rhubart and Jensen (2016).

To define an exhaustive classification of counties, TS1996 calculated a dissimilarity matrix D, where an entry D_{ij} is the dissimilarity of county i from county j, calculated as below:[3]

$$D_{ij} = 1 - P_{ij} = 1 - \frac{f_{ij} + f_{ji}}{\min(rlf_i, rlf_j)} \tag{1}$$

where f_{ij} is the number of commuters who live in i and work in j, and $rlf_i = \sum_j f_{ij}$ (including f_{ii}) is the resident workforce. The data used in TS1996 to estimate the flows is from the 1990 Journey to Work data, released from the 1990 Census. Other releases based on data from the 1980 and 2000 Censuses are also available.[4] After constructing this dissimilarity matrix,

[3]Due to computational constraints at the time, TS1996 broke the country into six overlapping regions: Northeast, Southeast, Midwest, Southwest, Central and West, calculating a separate distance matrix for each region.

[4]These CZ definitions are available at http://www.ers.usda.gov/data-products/

Table 2: Replication of Commuting Zones from TS1996: Summary Statistics

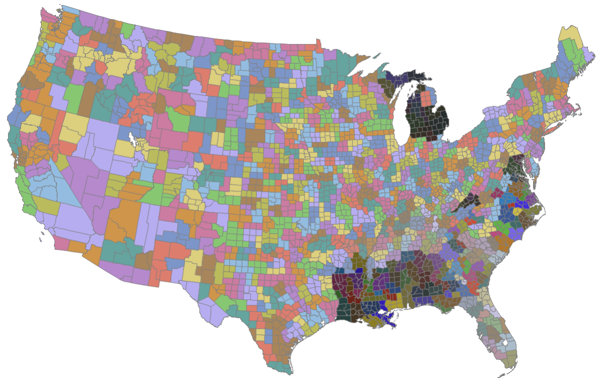
	TS1990	FKV1990
Mean Cluster Size	4.24	4.19
Median Cluster Size	4	4
Number of Clusters	741	741
Number of Singletons	62	10

Notes: Both TS1990 and FKV1990 are based on JTW tabulations from the 1990 Census. Summary statistics for TS1990 are from Table 8 of TS1996.

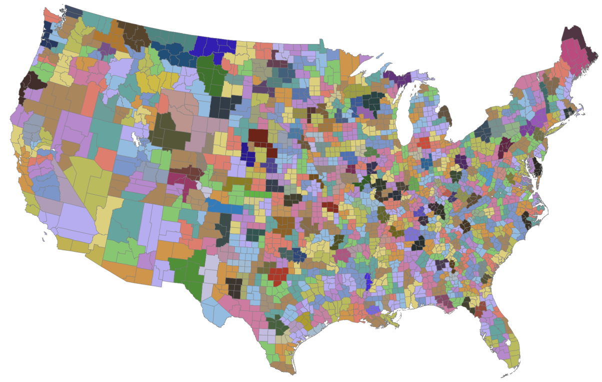
they use it as the input into an average-linkage hierarchical clustering algorithm (PROC CLUSTER in SAS). They allow clusters to form that are no higher than a 'height' of 0.98.[5] Additionally, because they cluster separately for region, and there are a number of states that are in multiple regions, they manually reconcile these competing cluster definitions, but do not provide their methodology for this reconciliation in the paper. Finally, they mention that they do perform a national analysis, but do not report the results.

We illustrate the process in Appendix Figure 8, which shows the heirarchical progression of how counties are clustered together for California. In the top left-hand corner, only a few counties have joined at a height of 0.8. As we increase the height of allowable clusters from 0.8 to 0.88, more counties are joined together to make clusters, and the same is true when increasing the height to 0.96. Finally, at a height of 1, almost all the counties have merged together and there is one large cluster and a few much smaller clusters.

Figure 1: Replication of Commuting Zones from TS1996: County Mapping



Commuting Zones from TS1996: TS1990



Replication of Commuting Zones: FKV1990

3.2 Replicating Tolbert and Sizer's Definitions

In order to replicate the clustering result in TS1996, which we refer to as TS1990, we use the 1990 Census JTW data and the methodology described above.[6] Our analysis showed that using a height cutoff of 0.9418 gives us results that are closest to the TS1990 mapping. We refer to the resulting clusters as FKV1990. Figure 1 shows a visual comparison of TS1990 and FKV1990 communting zones, while Table 2 shows summary statistics of TS1990 and FKV1990. There are 739 clusters in FKV1990, compared with 741 clusters in TS1990. In both sets, the average number of counties per cluster is 4.24, while the median is 4.

We also calculate two similarity measures between the clusters: for each county, the share of counties in our clusters that are in the same TS1990 commuting zone; and the share of counties in a TS1990 commuting zone are in the same cluster (unweighted). Both metrics

---

`commuting-zones-and-labor-market-areas.aspx`

[5] In the average-linkage clustering algorithm, the height (or distance) between clusters is given by the following formula:

$$D_{KL} = \frac{1}{N_K N_L} \sum_{i \in C_k} \sum_{j \in C_L} d(x_i, x_j)$$

That is, the distance between two clusters is the average of the pair-wise distances between each component of the two clusters. Therefore, that implies no clusters were merged such that $D_{KL} > 0.98$. (SAS Manual, version 9.2)

[6] Rather than splitting the country into separate regions, we perform our clustering at the national level, because it is computationally feasible to do so. We attempted to replicate their methodology at the regional level, but were unable to determine how they reconciled competing definitions across regions. Fowler, Rhubart and Jensen (2016) document the expert review process in the original Commuting Zone delineations.

are over 80%. Overall, we conclude that while not exactly identical, the FKV1990 replication of TS1990 is reasonably close.

4 Sensitivity Analysis of the Commuting Zone Methodology

While Commuting Zones are thought of as representing local labor markets, they have a number of shortcomings for empirical research that are not regularly discussed in the literature. In this section, we evaluate the sensitivity of Commuting Zone definitions for two aspects of the TS1996 methodology. First, the method is highly sensitive to variation in the input data, and if the input data are measured with error, the definitions may convey a false sense of certainty. Second, the resulting clusters are very sensitive to the decision of when to stop merging clusters, which implies that small changes in the chosen cutoff height drastically affect the number and size of clusters. Overall, this uncertainty in the definition of commuting zones contributes to conventional standard errors understating the true level of uncertainty in estimates, as well as non-classical measurement error that may bias empirical estimates. At the end of the section, we list other weaknesses that also undermine commuting zones as a definition of local labor markets.

4.1 Sensitivity of Clustering Results to Underlying Error

Given the reliance of TS1996 on the commuting flows data, we want to analyze the extent to which the outputs of the TS1996 methodology are sensitive to error in the input data. First, recall equation 1 for the dissimilarity matrix entries. If f_{ij} is measured without error, then D_{ij} is measured without error. However, if the flows are measure with some error, ϵ_{ij}, then we have an estimate of D_{ij}, $\hat{D}_{ij}$, which can be expressed as below:

$$\hat{D}_{ij} = 1 - \hat{P}_{ij} = 1 - \frac{f_{ij} + \epsilon_{ij} + f_{ji} + \epsilon_{ji}}{\min(r\hat{l}f_i, r\hat{l}f_j)} \tag{2}$$

Now suppose without loss of generality that county i is the smaller county; that means that

$$\hat{D}_{ij} = 1 - \frac{f_{ij} + \epsilon_{ij} + f_{ji} + \epsilon_{ji}}{rlf_i + \sum_j \epsilon_{ij}}$$

Or, rewritten:

$$\hat{D}_{ij} = 1 - \frac{f_{ij} + f_{ji}}{rlf_i + \sum_j \epsilon_{ij}} + \frac{\epsilon_{ij} + \epsilon_{ji}}{rlf_i + \sum_j \epsilon_{ij}}$$

However, even if $E[\sum_j \epsilon_{ij}] = 0$, we only have one realization of this draw (the published figures), which were calculated from survey responses. Additionally, we know two things about ϵ_{ij}. First, ϵ_{ij}/f_{ij} is larger for smaller counties. Second, this implies that rlf_i is more subject to measurement error for small counties than larger counties. This will increase D_{ij} for some small counties and decrease it for others. Because of the hierarchical nature of the clustering method, this will affect the formation of all other clusters in the data.[7]

To demonstrate how this measurement error affects the outcome of the clustering procedure, we use journey-to-work data from the pooled 2009-2013 ACS because it has published margins of error (hereafter, MOEs).[8] For reference, summary statistics on the ratio of the margins of error to the flows are in Table 4 in the appendix, and show that margins of error are much larger relative to smaller flows. Using these MOEs, we perform the following analysis:

1. For each origin-destination pair, we draw ϵ_{ij} from a normal distribution with mean 0 and standard deviation $MOE_{ij}/(2 \times 1.64)$, since the MOE is the 90% confidence interval.

[7] Additionally, because the heights are normalized in the procedure, it also affects where the effective cutoff is, even for counties unaffected by errors in flows.

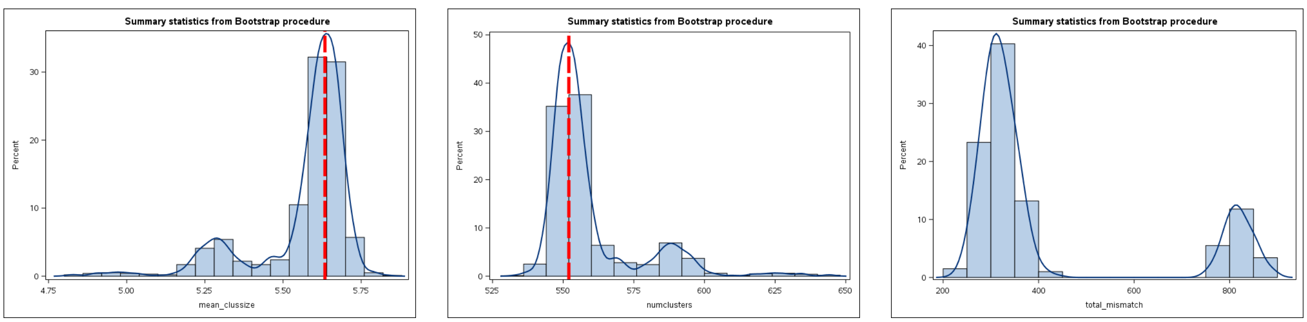
[8] In the 2009-2013 ACS, the average ratio of margin of error to reported commuting flow (unweighted) is 1.24. The Decennial Census numbers do not publish margins of error, but presumably they are somewhat smaller than ACS MOEs. Later in the paper, we use the 2009-2013 ACS MOEs to estimate the MOEs for the 1990 JTW flows.

2. Add the value from (1) to the reported flow ($\hat{f}_{ij} = f_{ij} + \epsilon_{ij}$).

3. Re-aggregate the flows, recalculate the dissimilarity matrix, and re-run the TS1996 clustering procedure.

4. Using the new clusters, calculate the following statistics: average number of counties in a cluster; the number of clusters; and the total number of counties in a different cluster than the one they would have been assigned, had the flows been assigned using only the published flow estimates (that is, not taking into account the MOE).

Figure 2: Sensitivity Analysis



(a) Mean Cluster Size (b) Number of Clusters (c)Mismatched Counties

Notes: Histograms plot the density of summary statistics from commuting zones produced using the TS1996 methodology from 1000 simulations of Margins of Error from the 2009-2013 ACS commuting flows data. Red lines provide the summary statistics from commuting zones produced from the published ACS estimates (mean cluster size of 5.63 for 552 clusters).

We iterate over this procedure 1000 times in order to obtain distributions for these statistics. These graphs are shown in Figure 2, where the red vertical dashed lines are the values that would be obtained using only the published figures (without MOE purturbations). Note that there are fewer clusters in the TS1996 implementation for JTW based on ACS, with 551 formed based on the published data compared with 741 in TS1990 and 709 in TS2000.[9] As one can see, the average cluster size has a wide range of values; while most values are around what the published numbers would give, there is still a substantial portion

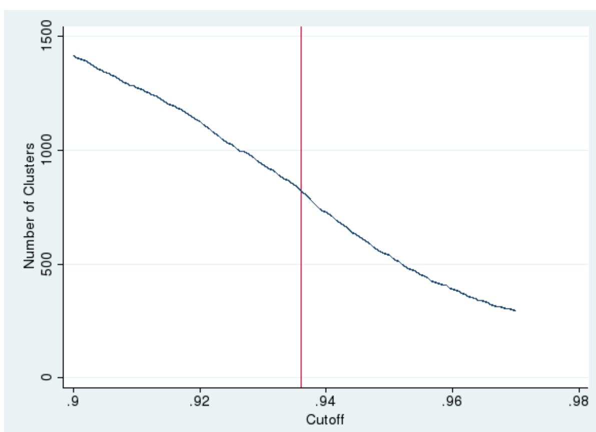
[9]We use the same cutoff here as in FKV1990, our replication of TS1990. While commuting flows have increased in distance since 1990 (see Table 1), it is not clear why the same methodology on later data results in so many fewer zones. These results hold irrespective of cluster height.

that have smaller than average clusters. Second, the number of clusters ranges widely, 330 to 450. Finally, the share of population that is mismatched is on average about 5% of the US population, a small but non-negligible number. Overall, the underlying measurement error in the data causes uncertainty in the cluster definitions, which is exacerbated by the sharp cutoff imposed in cluster analysis, which we discuss in the next subsection.

4.2 Distribution of Cluster Height

We turn to the sensitivity of the clustering to the chosen cutoff value. Tolbert and Killian (1987), describe the algorithm for choosing a cutoff value as follows: “As a rule of thumb, a normalized average distance of 0.98 was considered sufficient distance between sets of counties to treat them as separate [Labor Market Areas]” (Tolbert and Killian, 1987, page 15). The article does not provide an analysis of the sensitivity to changing the cutoff marginally up or down. In this subsection, we investigate how sensitive the resultant clusters are to the choice of the cutoff value.

Figure 3: Distribution of Cluster Consolidation, by Height



Notes: Histograms plot number of clusters that form at various heights, based on hierarchical clustering procedure. The vertical line is the value that most closely replicates TS1990.

Figure 3 shows the number of clusters by the chosen cutoff, with the vertical line indicating the cutoff we chose that most closely approximated TS1996 (0.936). The key takeaway from this figure is that it is unclear where a researcher would choose to stop the clustering process - but that any change in the cutoff will change the number of cluster.[10]

As we described above, the measurement error in commuting flows causes some uncertainty in terms of true dissimilarity, and hence true cluster height. Because of the presence of a strict cutoff, some clusters that would have formed (if there were no measurement error) did not form, and vice-versa. More broadly, TS1996 provide no empirical guidance for choosing the 'optimal' cutoff and cluster size other than referring to expert knowledge. Later, in Section 6, we consider measures of local labor market integration that may be informative of the optimality of various clustering definitions.

5 Robustness of Empirical Estimates to Commuting Zone Uncertainty

While we have shown that commuting zone definitions are sensitive to input data and parameters, this sensitivity only matters insofar as it impacts empirical estimates. We demonstrate the consequences of this sensitivity using a well know example of estimates measured at the commuting zone level.

In their paper, Autor, Dorn and Hanson (2013) estimate the impact that increased trade competition from China had on manufacturing employment in the United States. To estimate this effect empirically, they use variation in the initial distribution of manufacturing employment at the commuting zone level and national increases in imports from China by manufacturing subsector. Because they use commuting zones as their definition of the local labor market, their empirical analysis is exposed to the critiques we have discussed above.[11]

[10] Another consideration, not discussed here, is that TS1996 normalize the heights within each region before clustering. This causes the cutoff to be at different absolut heights depending on the region.

[11] We want to acknowledge that Autor, Dorn and Hanson have been incredibly helpful in the process of replicating their paper, both in providing data and helping troubleshoot, as well as being receptive to this exercise.

Table 3: China Syndrome Replication and Comparison, 1990-2000

	ADH Estimate	Our RHS	Our LHS and Weight	CZ Clustering
$\Delta IPW_{cz,t}$	-0.8875	-0.8871	-0.8748	-0.8748
	(0.1812)	(0.1811)	(0.1527)	(0.1243)

Notes: Table from author's calculations, using data from Autor, Dorn and Hanson (2013) and constructed data, based on equation 3. Column 1 is Table 2, Column 1 from ADH (2013). Column 2 replaces their measure of import exposure to ours. Column 3 replaces their measure of change in manufacturing employment and CZ-specific weights with ours. Column 4 does not cluster on state. Standard errors are in parentheses. All coefficients are significant with p-values less than 0.01.

Their main estimating equation in the paper is as follows:

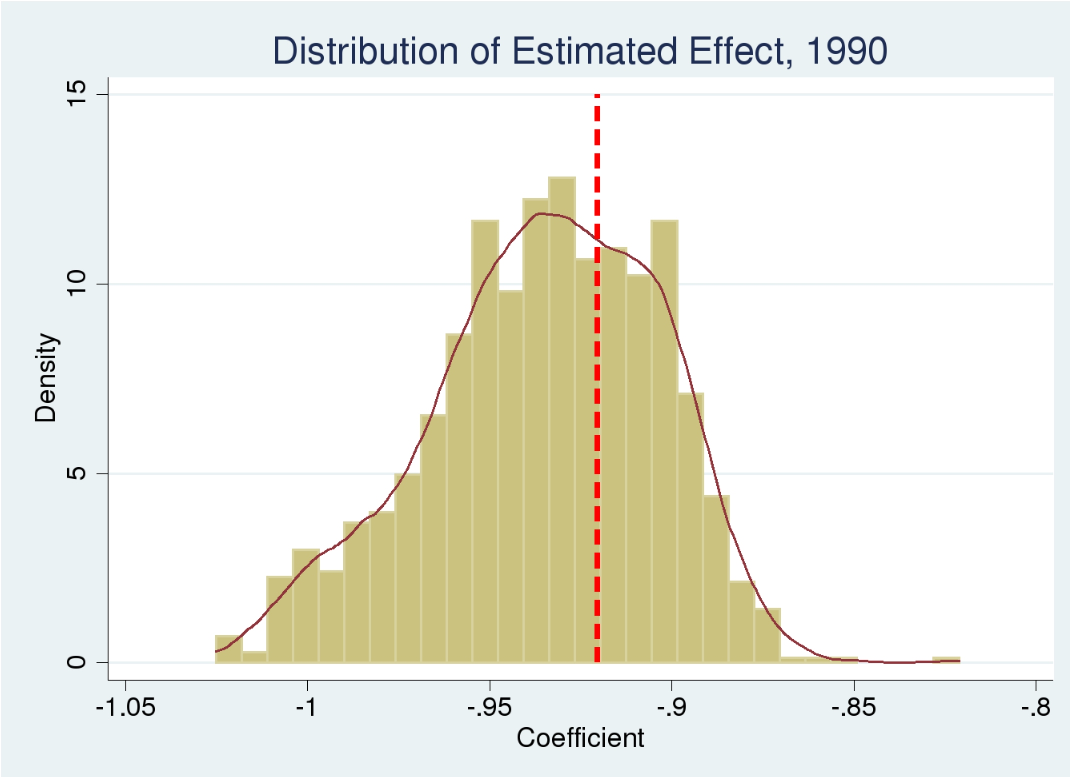
$$\Delta L_{it}^m = \gamma_t + \beta_1 \Delta IPW_{uit} + \beta_2 X_{it} + e_{it} \tag{3}$$

Where ΔL_{it}^m is the decadal change in manufacturing employment in Commuting Zone i following year t, ΔIPW_{uit} is the import exposure measure for the United States, and X_{it} are control variables. All regressions are weighted by population share of the commuting zone.

Since we use slightly different methods of aggregating data, we compare the main estimates from Autor, Dorn and Hanson (2013) (Table 1 in their paper) to our replication, which we show in Table 3. Each cell in the table is a coefficient from a different regression, and for simplicity we just estimate it for the time period 1990-2000 (other results available upon request). The first column shows the estimates from their paper, while the second column changes the import exposure measures to our replicated measure. In the third column, we use our estimate of the change in manufacturing employment amd weights, rather than the estimate and weights from their data. The final column clusters on commuting zone rather than state.

Overall, the estimates are considerably stable, giving us confidence that we are properly replicating their initial findings. We now turn to demonstrating how these estimates are affected by the concerns with the commuting zone definitions themselves.

Figure 4: Distribution of Effect, 1990-2000



Note: Histogram plots estimates of β_1 from equation 3, based on commuting zone realizations as outlined in Section 4.

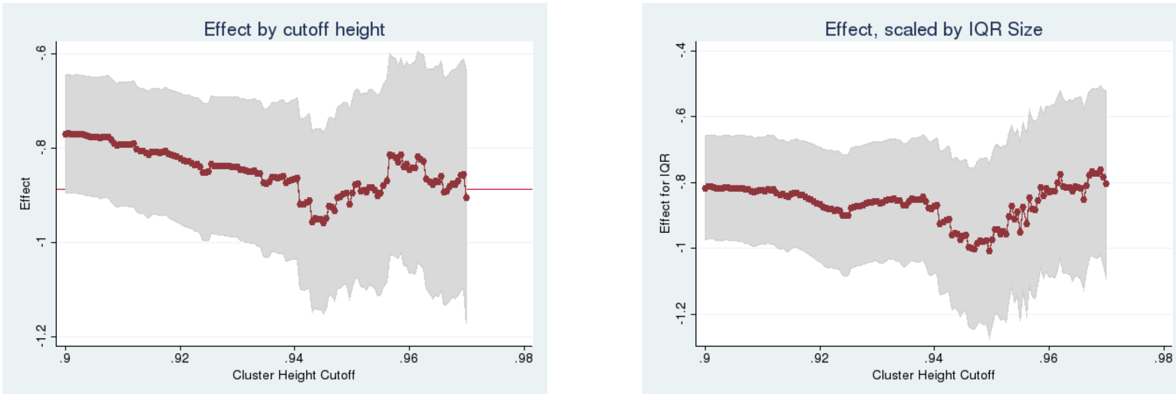
5.1 Effect of Errors in Flows Data

First, we show how the estimate is affected based on MOE in the commuting flows input data. For this exercise, we used the distribution of margins of error published in the 2009-2013 ACS flows, and matched the ratio based on the flow sizes of the 1990 JTW data.[12] We then draw from a normal distribution implied by the margins of error, and calculated a new dissimilarity matrix with the perturbed flows. Using the same cutoff that we used for our replication of TS1996, we produce a new realization of commuting zones, and then repeat the above procedure 1000 times.

We use these different commuting zone definitions to aggregate counties and then estimate equation 3. These estimates are graphed in Figure 4, which shows the distribution of the

[12]These flow size bins are the following percentile bins: 0-50; 50-90; 90-95; 95-99; and 99+.

Figure 5: Differences in Effect Based on Cluster Cutoff



(a) Effect by Cutoff Height (b) Effect by Cutoff Height, Scaled by IQR

Note: Author's calculations based on replication of Tolbert and Sizer's method. Panel (a) shows estimates of β_1 from equation 3 for different definitions of commuting zones based on height cutoff, while Panel (b) shows estimates of β_1 scaled by the difference in exposure between the 25th and 75th percentile commuting zone. The horizontal line in panel (a) is the main estimate from Autor, Dorn and Hanson (2013)

estimated effect for the 1990-2000 period, and the red vertical line shows the estimate under the standard commuting zone defintions. The estimates are somewhat dispersed, with the left tail of the distribution almost 10% lower than the estimate from ADH; however, the values are within two standard errors of the original estimate. Additionally, the distribution is skewed to the right.

This exercise demonstrates that there is additional uncertainty induced by the construction of the commuting zones that is not addressed in empirical estimates that use these definitions, which may overstate the precision of the results.

5.2 Effects on the Chosen Cutoff

In addition to the uncertainty that is induced by underlying error in the commuting flows, in Section 4.2 we showed that the decision of where to stop the clustering process was rather arbitrary, since there is no clear guidance on what cutoff is most appropriate. To demonstrate how the cutoff choice affects the estimate of β_1 from equation 3, we generate clusters based on cutoffs between 0.9 and 0.97 and estimate the model using the resulting clusters.

Figure 5 displays the results of this exercise, where panel (a) shows the raw coefficient, and panel (b) shows the coefficient scaled by the interquartile range of ΔIPW_{uit}, since this changes depending on the cutoff. In panel (a), the red horizontal line is the estimate from ADH.

Again, our results show that there is some variation in the estimate based on the cutoff value. Around the cutoff value that most closely replicates TS1996, the estimate is the most negative. However, cutoff values marginally higher or lower give different results, which reinforces the point we made with Figure 3 - the number of clusters that merge is incredibly dense near the cutoff for commuting zones, so that any change in the cutoff changes the commuting zone definitions non-neglibly. This density causes estimates using commuting zone observations to change near the cutoff. Given the sensitivity of estimates to the chosen cutoff, best practices would be to report estimates for a broad range of cutoffs.

6 Objective Function for Measuring Local Labor Market Integration

Up to this point, we have focused on the commuting zone methodology exclusively. However, in the next two sections we turn to two important issues: how to measure labor market integration, and an alternative clustering method that does not have the drawbacks of the hierarchical technique.

The theoretical literature consistently defines a local labor market as characterized by similar wages, unemployment rates, and commuting links within the area. However, there are no established methods for measuring how appropriate a set of local labor markets is in approximating the theoretical area.

To that end, we formulate an objective function to evaluate how well any given local labor market definition reflects this theoretical construct. To be specific, we measure four statistics that reflect the integration of an area in terms of unemployment and wage series as well as in- and out- commuting rates.[13]

First, for both unemployment rate and wages (using BLS data), we measure the weighted average pairwise correlations between counties in a cluster, which are $\bar{\rho}_i^{URATE}$ and $\bar{\rho}_i^{Wages}$, respectively.[14] The higher these values are, the more integrated the counties are.

Second, we measure commuting flows into the labor market from other counties, as a share of the local labor force, as well as commuting flows to outside areas ($InflowShare_i$ and $OutflowShare_i$). We expect that the commuting flows within a labor market ought to be much higher than commuting flows outside of the labor market, reflecting an integrated area, such that a higher value of these measure reflects lower labor market integration. We sum these measures across labor markets in the following manner:

$$Objfn(C) = \frac{1}{4N_c}\sum_{i \in C}(\gamma_1 \bar{\rho}_i^{URate} + \gamma_2 \bar{\rho}_i^{Wages} - \gamma_3 InflowShare_i - \gamma_4 OutflowShare_i) \quad (4)$$

Where $Objfn(C)$ is the objective function value for a definition of local labor markets, C. Given that a larger objective function value reflects a more integrated local labor market, when comparing two competing definitions of local labor markets, the definition with the higher objective function value is the better labor market definition. More formally, for two local labor market definitions, C and D, if $Objfn(C) > Objfn(D)$, then C is the more appropriate local labor market.

In the next section, we develop a methodology using this objective function, such that our resulting local labor market definitions maximizes the objective function.

[13] We have considered additional measures that correspond to the theoretical literature, including housing or rental price series and compactness, which could be measured by average pairwise distance.

[14] The pair-wise correlations between counties are calculated using six years of data (for this paper, that is 1990-1995).

$$\bar{\rho}_i^{URATE} = \frac{1}{2N}\sum_{i \in C}\sum_{j \in C}\omega_{ij}\rho_{i,j}$$

Where the weights $\omega_{ij} = \frac{reslf_i = reslf_j}{2 * \sum_{k \in C} reslf_k}$

7 Our Proposed Methodology

We propose an alternative method based on spectral clustering for obtaining local labor market definitions. We refer to definitions based on this methodology as Mobility Zones, but note that the methodology is still under development and that mappings presented here are preliminary. We first describe spectral clustering, then show the results of our implementation, and then compare our definitions with Tolbert and Sizer's commuting zones and other local labor market defintions using our objective function.

7.1 Introduction to Spectral Clustering

Spectral clustering classifies nodes by implementing k-means on the eigenvectors of a graph Laplacian of pairwise similarities (von Luxburg, 2007). Spectral clustering can be explained as a random walk on the similarity graph, where the probability of jumping from one node to another is given by edge weights. In the stationary distribution of these stochastic jumps, a clustering may be thought of as the classification that minimizes jumps between clusters, for a given number of clusters. This interpretive framework is a natural way to think of local labor markets and job search or migration models. For this work, we assume an undirected or symmetric graph of similarities. Spectral clustering is a popular methodology implemented in a wide range of fields.

We implement normalized spectral clustering as defined by Ng, Jordan and Weiss (2001), which re-normalizes the rows of the eigenvectors for the number of clusters specified. This re-normalization has the advantage that even fairly isolated nodes end up clustered. In our experience, both hierarchical clustering and standard versions of spectral clustering were highly sensitive to the intensity of commuting flows and often failed to differentiate clusters outside of major urban centers. These residual, mostly rural, areas would end up lumped into a "mega-cluster" with little geographic integrity. Only by saturating the model with a high number of clusters (or a high threshold in the case of hierarchical clustering) could we allocate all counties to similarly sized clusters. With the methodology of Ng, Jordan and Weiss (2001), we can reliably allocate all counties to clusters for any quantity of clusters that we specify. We rely on the objective function to determine the optimal quantity of clusters, as implemented by normalized spectral clustering.

Spectral clustering is also amenable to classifying nodes based on multiple views of a graph, as we have argued is appropriate for characterizing local labor markets. Nodes may be characterized both by the intensity of pairwise links (.e.g commuting flows, distance, correlation of unemployment rates) and by feature space similarity (e.g. share employed in manufacturing). Zhou and Burges (2007) generalize single-view spectral clustering, showing that multi-views can be summarized in an undirected similarity graph as a weighted, linear combination of each view. Considering again the random walk interpretation of spectral clustering, Zhou and Burges (2007) interpret the multi-view graph as a mixture of Markov transition chains for each view. The objective function, which is also a linear combination of weighted values for each cluster, is not necessarily helpful for selecting an optimal weights. In the current implementation, we assign an equal linear weight to each view in the multi-view graph, using the same views that enter into the objective function. As with the objective function, the linear weights would depend on the economic interpretation of the clusters.

The following, using notation from von Luxburg (2007), describes the specific steps of our implementation of multi-view spectral clustering, with the goal of classifying n counties into k clusters:

1. We construct an undirected similarity graph S, where the edge weights between nodes i and j are defined as w_{ij}, with the weight consisting of a linear combination of the set of views (e.g. commute flows, proximity, pairwise corrleation in unemployment rates).

2. We compute the Laplacian, a symmetric matrix, as $L = Deg(S)^{-1/2} \cdot S \cdot Deg(S)^{-1/2}$, where $Deg(S)$ is the Degree matrix giving the row sums of S on the diagonal.

3. We compute eigenvalues and eigenvectors of L and form an n by k matrix, U, from the

k eigenvectors with the largest eigenvalue.

4. We form an n by k matrix T by normalize the rows of U to norm 1, each element $t_{ij} = u_{ij}/(\sum_{j \in k} u_{ik}^2)^{1/2}$.

5. We classify the n rows of T (one for each county) into k clusters, $C_1, ..., C_k$ using the k-means algorithm.

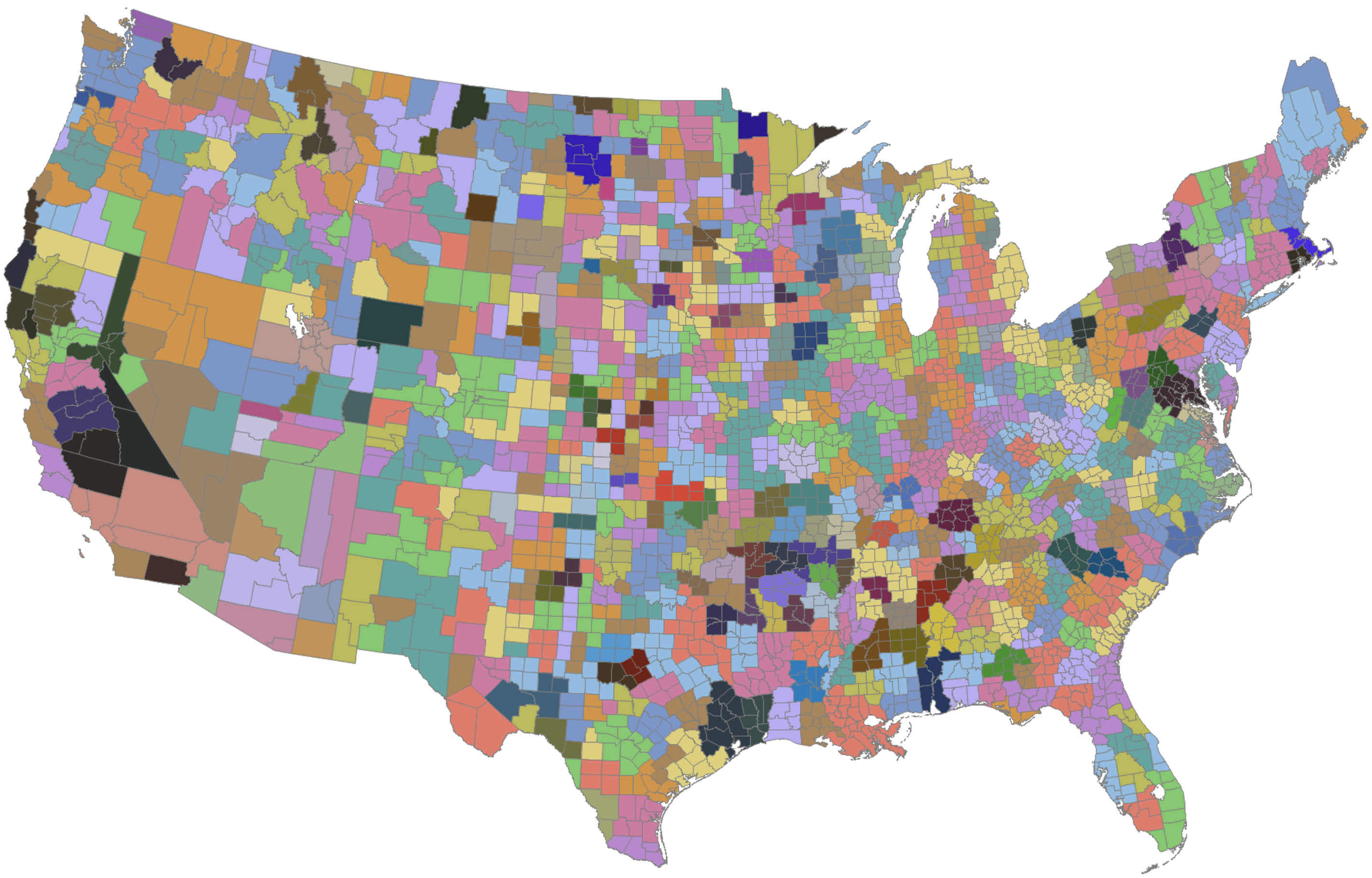
7.2 Implementation of Spectral Clustering

In implementing this methodology, we use the similarity matrix implied by TS1996 as S, over which the Laplacian is computed. Additionally, in spectral clustering, the choice variable is not the cutoff height (as in hierarchical clustering) but the number of clusters, k. Rather than arbitrarily choose a number of clusters, we use our objective function specificed in equation 4 to choose the set of clusters that is the most integrated, such that k maximizes our objective function. Formally, we choose k such that:

$$k* : ObjFn(C_{k*}) = \max_{k \in [400,800]} ObjFn(C_k)$$

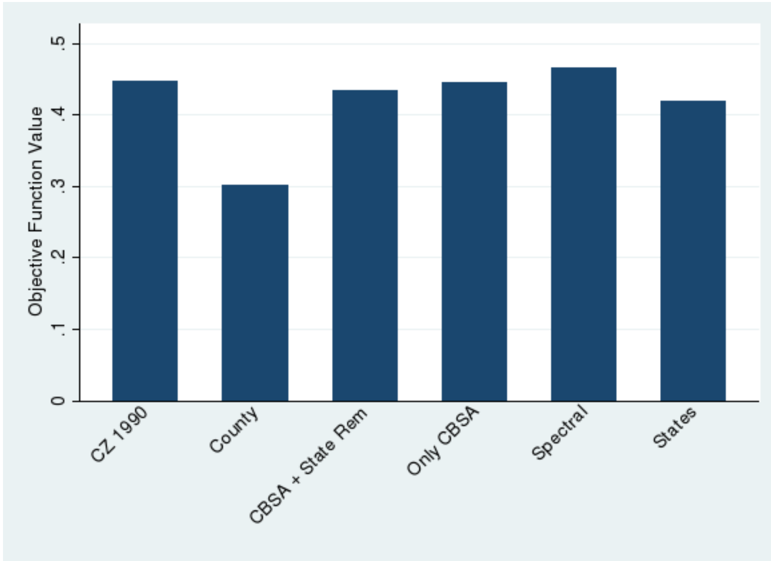
Our resulting clusters are mapped in Figure 6, and our calculations show that the optimal number of clusters is 755. Additionally, using our methodoogy clusters have an average of 4.12 counties each, which is smaller than Tolbert and Sizer's defintions (4.24). Comparing Figure 6 with Tolbert and Sizer's commuting zones, there are a number of differences. First, the clusters in the East are larger in area than the clusters in Tolbert and Sizer's definitions. This is particularly telling in Florida, which has fewer clusters using our definitions. The same is true of Texas; in our clusters, the main metropolitan areas are visible, while they are not in Tolbert and Sizer. One notable similarity are the high frequency of very small clusters in Kansas. It appears that irrespective of methodology, there is not a lot of long-distance commuting in Kansas, which may reflect the different industrial composition.

Figure 6: Optimal Spectral Output, 755 Clusters



Note: Clusters are a result of authors' calculations, using methodology outlined in text.

Figure 7: Comparing Different Local Labor Market Definitions



Notes: Author's calculations. CBSA definitions are 2010 current definitions. Data used in these calculations are discussed in Section 2.

Next, we turn to comparing our results from the spectral clustering method with other local labor market definitions, using our objective function.

7.3 Comparing Common Local Labor Market Definitions

To compare our definitions to other candidate local labor markets, we calculate the objective function for a number of candidate local labor markets: our maximized spectral clustering definintion, Tolbert and Sizer's commuting zones, states, CBSAs (with a non-metro state residual), CBSAs only, and counties. These results are displayed in Figure 7.

Spectral clustering outperforms commuting zones. Interestingly, if the components of the objective function are decomposed, we find that the commuting inflow and outflow share are similar, but that the correlations in unemployment rate and earnings are much higher for clusters using spectral clustering.

Spectral clustering also has a higher objective function value than both CBSA and CBSA plus rest of state. While states perform only slightly worse than commuting zones (because there is very little inter-state commuting), counties perform particularly badly because of inter-county commuting rates are relatively high.

In future work, we also want to compare these definitions over time, to see whether the decay in quality of local labor market definition is different across these definitions. This is especially relevant for current research which uses commuting zones, since these areas may not reflect current local labor market definitions.

8 Summary and Next Steps

Recent influential papers in labor economics have used commuting zones as an alternative definition to local labor markets. However, no one has carefully analyzed how elements of how these commuting zones are constructed may affect empirical findings. Our paper contributes to this literature by analyzing this methodology and its implications.

We document that the commuting zone methodology is sensitive to uncertainty in the input data and parameter choices and we demonstrate how these features affect the resulting labor market definitions. Furthermore, we demonstrate that uncertainty in local llabor market definitions also affects empirical estimates that use commuting zones as a unit of analysis. Finally, we develop a metric to compare competing local labor market definitions, as well as a method for obtaining alternative local labor market definitions. We show that our alternative definition out-performs commuting zones, as well as other local labor market definitions commonly used.

There are a number of questions and issues that we have not yet addressed in this research. First, we want to demonstrate that our approach is not as sensitive to the issues outlined for heirarchical clustering. Second, we want to compare how robust various local labor market definitions are over time. Finally, we want to allow our clustering procedure to use more information than commuting to form clusters.

References

Amior, Michael, and Alan Manning. 2015. "The Persistence of Local Joblessness." Centre for Economic Performance, LSE CEP Discussion Papers dp1357.

Autor, David H., David Dorn, and Gordon H. Hanson. 2013. "The China Syndrome: Local Labor Market Effects of Import Competition in the United States." *American Economic Review*, 103(6): 2121–68.

Blanchard, Olivier Jean, and Lawrence F. Katz. 1992. "Regional Evolutions." *Brookings Papers on Economic Activity*, 23(1): 1–76.

Bound, John, and Harry J Holzer. 2000. "Demand Shifts, Population Adjustments, and Labor Market Outcomes during the 1980s." *Journal of Labor Economics*, 18(1): 20–54.

Card, David. 2001. "Immigrant Inflows, Native Outflows, and the Local Labor Market Impacts of Higher Immigration." *Journal of Labor Economics*, 19(1): 22–64.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. 2014. "Where is the land of Opportunity? The Geography of Intergenerational Mobility in the United States." *Quarterly Journal of Economics*, 129(4): 1553–1623.

Diamond, Rebecca. forthcoming. "The Determinants and Welfare Implications of US Workers' Diverging Location Choices by Skill: 1980-2000." *American Economic Review.*

Foote, Andrew, Michel Grosz, and Ann Stevens. 2015. "Locate Your Nearest Exit: Mass Layoffs and Local Labor Market Response." Center for Economic Studies, U.S. Census Bureau Working Papers 15-25.

Fowler, Christopher S., Danielle C. Rhubart, and Leif Jensen. 2016. "Reassessing and Revising Commuting Zones for 2010: History, Assessment, and Updates for U.S. 'Labor-Sheds' 1990–2010." *Population Research and Policy Review*, 35: 263–286.

Head, Allen, and Huw Lloyd-Ellis. 2012. "Housing Liquidity, Mobility, and the Labour Market." *Review of Economic Studies*, 79(4): 1559–1589.

Kennan, John, and James R. Walker. 2011. "The Effect of Expected Income on Individual Migration Decisions." *Econometrica*, 79(1): 211–251.

McKenzie, Brian, and Melanie Rapino. 2011. "Commuting in the United States: 2009." U.S. Census Bureau American Community Survey Reports ACS-15.

Monte, Ferdinando, Stephen J. Redding, and Esteban Rossi-Hansberg. 2015. "Commuting, Migration and Local Employment Elasticities." National Bureau of Economic Research, Inc NBER Working Papers 21706.

Moretti, Enrico. 2011. "Local Labor Markets." , ed. O. Ashenfelter and D. Card Vol. 4 of *Handbook of Labor Economics*, Chapter 14, 1237–1313. Elsevier.

Ng, Andrew Y., Michael I. Jordan, and Yair Weiss. 2001. “On Spectral Clustering: Analysis and an algorithm.” *Advances in Neural Information Processing Systems*, 849–856.

Notowidigdo, Matthew J. 2011. “The Incidence of Local Labor Demand Shocks.” National Bureau of Economic Research, Inc NBER Working Papers 17167.

Restrepo, Pascual. 2015. “Skill Mismatch and Structural Unemployment.” Massachusetts Institute of Technology Job Market Paper.

Roback, Jennifer. 1982. “Wages, Rents, and the Quality of Life.” *Journal of Political Economy*, 90(6): 1257–78.

Schmutz, Benoit, and Modibo Sidibé. 2014. “Job Search and Migration in a System of Cities.” Centre de Recherche en Economie et Statistique Working Papers 2014-43.

Tolbert, Charles M, and Molly Sizer. 1996. “US commuting zones and labor market areas: A 1990 update.” Economic Research Service.

Tolbert, Charles M, and Molly Sizer Killian. 1987. “Labor market areas for the United States.” *ERS staff report*.

von Luxburg, Ulrike. 2007. “A Tutorial on Spectral Clustering.” *Statistics and Computing*, 17(4): 395–416.

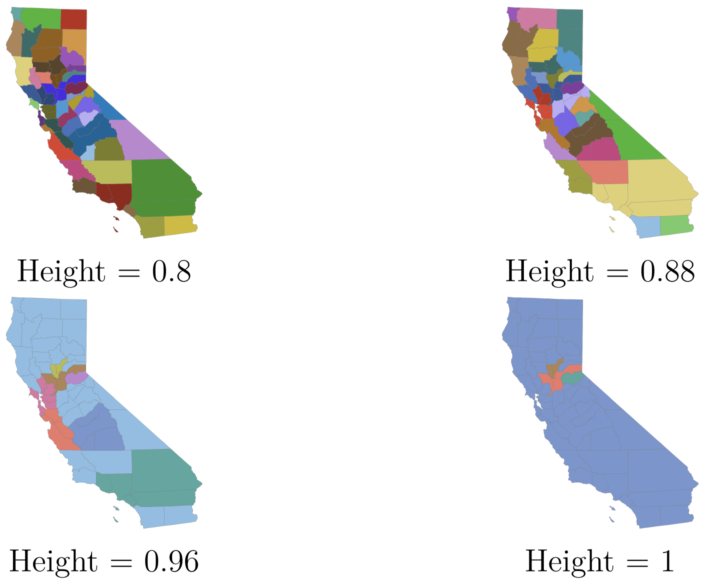
Wozniak, Abigail. 2010. “Are College Graduates More Responsive to Distant Labor Market Opportunities?” *Journal of Human Resources*, 45(4): 944–970.

Yagan, Danny. 2016. “The Enduring Employment Impact of Your Great Recession Location.” University of California, Berkeley Working Paper.

Zhou, Dengyong, and Christopher J. C. Burges. 2007. “Spectral Clustering and Transductive Learning with Multiple Views.” *Proceedings of the 24th International Conference on Machine Learning*, 1159–1166.

Tables and Figures Appendix

Figure 8: Various Height Cutoffs for California



Notes: The above graphs are generated using the methodology outlined in Section 3, using 1990 Census JTW data. More detail is in the text.

Table 4: Summary Statistics of Ratio of MOE to Flows

	Mean	25th Pctile	50th Pctile	75th Pctile
All counties	1.236	0.845	1.370	1.600
Flows <100	1.432	1.148	1.500	1.636
Flows 100-1000	0.444	0.301	0.414	0.549
Flows 1000-10000	0.131	0.087	0.124	0.169
Flows 10000+	0.037	0.024	0.036	0.049

Notes: Author's calculation using 2009-2013 ACS Journey-to-Work data.