# The Role of Data Sample Uncertainty in Delineations of Core Based Statistical Areas and Rural Urban Commuting Areas

Christopher S. Fowler[1] · John Cromartie[2]

## Abstract

Federal standards that assign counties to Core Based Statistical Areas (CBSAs) and designate tracts with a Rural Urban Commuting Area (RUCA) code rely on journey to work data from the American Community Survey (ACS). Because the ACS is based on a relatively small sample of the population commuting flows are represented as point estimates, frequently with high margins of error. In this paper we examine the impact of uncertainty in commuting flows on these two critical designations. We find that, for the vast majority of counties (85%) and tracts (87%), the official designation remains consistent and that, in over 1,000 replications of the data, 98% of county assignments and 97% of tract assignments were consistent with the official delineations. While these results are reassuring, a small number of counties and tracts do experience assignments that are different from the official delineation at very high rates. We also test the official delineations against an alternative data source, the Longitudinal Employer- Household Dynamics Survey's Origin–Destination Employment Statistics (LEHD-LODES). We find the LODES data result in designations that are largely consistent with those from the ACS but we lack a clear way to choose between the data products in places where they differ. Overall, our findings suggest that the current delineation method for assigning counties to CBSAs and tracts to RUCA codes is sufficiently robust to uncertainty to continue going forward. Nevertheless, the presence of significant uncertainty for some observations suggests the need for continued consideration of uncertainty in the context of future data releases.

✉ Christopher S. Fowler
csfowler@psu.edu

John Cromartie
john.cromartie@usda.gov

1 Penn State University, State College, PA, USA

2 USDA Economic Research Service, Kansas City, MO, USA

🙋 Springer

# 1 Introduction

Every year billions of dollars are allocated through programs where eligibility is based on a com- munity's status as rural or urban. How these funds are allocated depends crucially on federal standards that define geographic areas in particular ways. Key among the many standards are designations that assign counties to Core Based Statistical Areas (CBSA), the blanket term covering both metropolitan and micropolitan regions. Another important designation is the Rural Urban Commuting Area codes (RUCA codes) maintained by the U.S. Department of Agriculture's Economic Research Service (USDA ERS). Crucially, both the CBSA and RUCA designations rely on survey data produced by the Census Bureau that documents county to county or tract to tract commuting flows.

Both the CBSA and RUCA designations rely on commuting flow data from the American Community Survey (ACS) as a mechanism for indicating social and economic connection between geographic areas. Drawn from a rolling 2% sample of the population,[1] the five-year ACS commuting data is delivered with both a point estimate and a margin of error (MOE) covering the 90% confidence interval around the provided estimate. In counties where the population is either small or flows are divided among a large number of origins and destinations these MOEs can be quite high (the most extreme case in the 2006–2010 county data links El Paso County, Colorado with Otero County, New Mexico with a single commuter and a margin of error of 127 or 12,700%). In tracts, where the population counts are typically smaller, 68% of flows have a MOE that is larger than the flow itself. Despite the significance of the uncertainty in the commuting flow data, uncertainty does not enter into the delineation process for CBSA's or RUCA codes. Additionally, the uncertainty associated with the underlying data is lost in the delineation and, in the case of CBSA's, the end user is never made aware that the uncertainty existed.

Careful consideration of uncertainty in data is experiencing something of a renaissance as new data and new methods work to distinguish statistical or methodological uncertainty from data uncertainty (see Franklin, 2022 for a thorough review of this issue within human geography). Nevertheless, many practitioners who work with Census data do not have a clear sense of how to deal with uncertainty such as that presented by the ACS (Jurjevich et al., 2018). The default practice (and Jurjevich provides supporting evidence for this among planners in particular) is to ignore provided margin of error information. Lacking information on uncertainty allows delineations to appear definitive with no means to ascertain their reliability. Inclusion of uncertainty analysis is warranted in the face of research documenting that counties and tracts vary significantly in the degree to which they represent a good

---

[1]  Our analysis relies on the five-year combined sample release, the only release for smaller geographic areas like tracts. In theory, this five-year data would represent a 10% sample of the population in any given area, though this is not precisely correct because of entrance and exit of respondents, oversampling in some areas and for some populations, and other efforts designed to make the sample representative of the population.

fit for different measures (Plane, 1981; Fowler et al., 2019; Fowler & Jensen, 2020; Cromartie & Swanson 1996).

The purpose of this paper is to examine the impact of uncertainty in commuting flows on the delineations for CBSA's and RUCA codes. We find that only a small share of counties and tracts are impacted by uncertainty in terms of what CBSA or RUCA designation they receive (about 15% of counties and 13% of tracts). However, for the units where alternative assignment is possible, inconsistent assignments can happen quite frequently when commuting flows are allowed to vary within their provided margin of error; as much as 54% of the time for affected counties and as much as 100% of the time for affected tracts. The places at greatest risk of inconsistent assignment are not geographically concentrated and they are broadly distributed across urban and rural communities. Affected places are not the smallest or largest places in terms of population, but places in the middle of the distribution; counties between 10,000 and 100,000 persons and tracts of all sizes (within the relatively narrow population targets that the Census uses for tracts).

Additionally with this paper we consider the impact of employing an alternative to the ACS commuting data, the Longitudinal Employer-Household Dynamics Survey's Origin–Destination Employment Statistics (LEHD-LODES). LODES data are the primary competitor to the ACS commuting data and have national coverage, annual updates, and origins in a broad sample of the commuting population that all make LODES an attractive alternative. While LODES does not have an MOE, it is a synthetic data set whose precise deviations from the 'true' data are not disclosed to protect individual privacy. As a result it is hard to know whether any differences that emerge from using LODES are the result of a different data source or a different methodology. Our analysis finds that using LODES data produce a somewhat different set of delineations that differ from the official point-estimate delineations and from the alternatives incorporating uncertainty. One positive finding of our analysis is that the differences in delineations arrived at with ACS and LODES data are not biased in one direction or the other; that is LODES is not more likely to designate places as more urban or less urban. This suggests that the differences between the two data sets are a function of their joint uncertainty, not a specific structural difference.

Finally, we attempt to provide further value to the research community by providing scripts that replicate both the official delineations and our method for examining uncertainty as well as output files that associate the official delineations with our uncertainty measure for use in uncertainty-sensitive research using these geographic areas. Replication of the extant delineations turned out to be a non-trivial undertaking even with support from agency experts. As the standards come up for review and revision in 2023 based on the 2020 census we hope these delineation scripts can support careful examination of proposed modifications.

In the remainder of this paper we first situate our work within the broader literature on uncertainty. Next we describe the data and methods used to replicate the federal standards and then replicate them in the context of uncertain commuting flows. Subsequently we present the results of our analysis of 1,000 alternative flow matrices built from the point estimates and MOE's provided by the ACS and discuss the variation we observe in delineations built from these matrices in terms of geography

and observation type. Ultimately we argue for the inclusion of this type of uncertainty analysis in the next generation of CBSA and RUCA designations based on the 2020 census.

## 2  Literature Review

Uncertainty enters any form of analysis in myriad ways from the selection of subjects and the categories used to describe them to the structure of analytic techniques and the norms for reporting results (Franklin, 2022). Geographic data have additional mechanisms for introducing uncertainty based on practices of measurement and reporting (Robertson & Feick, 2018) but within the broad realm of geographic data, that which attempts to capture mobility has probably the greatest uncertainty of all (Folch et al., 2016; Franklin & Plane, 2006; Kwan, 2012). For the purposes of this discussion we will focus on three forms of uncertainty. First, we look at uncertainty in the data and its basis in a relatively small sample of individuals. Next we look at uncertainty in the meanings of the delineations we consider, particularly the significant debates around defining rural and urban or aggregating sub-units into 'functional regions' meant to capture shared socio-economic practices. Finally, we look at methods of delineation that address uncertainty in how specific places are designated given a set of possible categories. This last framing of uncertainty is the one that we address most directly in this paper, but it is contingent on the other two, so we lay them out first.

There is a significant literature that addresses questions of uncertainty in the use of data from the ACS (Folch et al., 2016; Jung et al., 2019; Spielman & Folch, 2015; Spielman et al., 2014; Sun & Wong, 2010). Much of this work is designed to help researchers interpret and work with the MOE's provided in the survey data and to point out the varia- tion in the quality of the data as represented by these MOE's. Jurjevich et al. (2018) has demonstrated that most planners who work with ACS data simply ignore the information on uncertainty that comes with the data. Anecdotally, the academic literature seems to do the same except in cases, like those listed above, where the authors are specifically addressing the topic of uncertainty. One theme that emerges from this work is that the quality of esti- mates from the ACS varies in non-random ways with lower quality estimates in urban cores, in poorer areas, and in the Southern United States (Folch et al., 2016). Non-random data quality is a significant problem, especially for national-level characterization (such as the CBSA and RUCA delineations covered here). Particularly concerning is that the uncertainty conveyed in the original data does not pass through to these subsequent delineations. One promising thread to emerge from this literature is that the directed aggregation of geographic subunits through regionalization can reduce the impact of uncertainty in larger units (Spielman & Folch, 2015). This technique holds promise for future iterations of the CBSA and RUCA delin- eation methodologies, which already utilize regionalization but without the focus on reducing uncertainty.

Another aspect of uncertainty that is relevant here has to do with uncertainty in the meanings we assign in the process of delineation. There is a significant debate about what constitutes rural and/or urban, with interest in thinking about

the 'continuum' of contexts that exist within this dichotomy (Isserman, 2005; Baer et al., 1997; Fowler et al., 2016; Ratcliffe et al., 2016; Schroeder & Pacas, 2021; Cromartie & Bucholtz, 2008). The problem of distinguishing between rural and urban places for the purposes of federal funding has a long history and comes to a head every decade or so when new Census data triggers new delineations (Morrill et al. 1999; Fitzsimmons & Ratcliffe, 2004; Porter & Howell, 2009). While these various interpretations differ on the specifics, there is generally agreement on roles for population density and connectivity as key measures of where places fall on this continuum. In a significant critique of extant definitions of rural and Urban Andrew Isserman drew attention to the way in which major federal agencies precisely identified 'urban' and 'metropolitan' but left 'rural' and 'nonmetropolitan' poorly defined as remainder categories (Isserman, 2005). Isserman argued that problems with delineations (by the Census Bureau and the Office of Management and Budget) were significant because they could lead to miss-allocation of federal funds among other issues (Isserman, 2005). In a similar strain of argument Porter and Howell (2009) argued for caution in assumptions about the internal homogeneity of places designated as nonmetropolitan or metropolitan. Their work found significant heterogeneity not only in counties on the fringes of metropolitan areas, but also in counties that were unequivocally metropolitan. More recently, Wright et al. have documented how updates to the OMB metropolitan delineations have shaped perceptions of changing diversity and segregation by continually adding peripheral counties to metropolitan areas that tend to be whiter than the counties included in previous delineations (Wright et al., 2022).

Further complicating this debate is the fact that what is commonly understood as rural or urban changes over time as there is no fixed understanding of how many people or how densely settled they must be to constitute an urban context. In general, as population and city sizes increase, what constitutes a city also changes in the public view. Experts in charge of determining delineation methods are thus placed in a quandary between maintaining consistent definitions over time and updating definitions to reflect changing meanings (Ratcliffe et al., 2016). Even small changes in these definitions will generate plenty of public debate (Vowell, 2021). Finally, on top of uncertainty of meaning and changing meanings, there is an issue of how well any given observation will fit some established meaning. Plane documented this phenomenon with respect to metropolitan areas over 40 years ago, noting how cross-metro commuting patterns in New England created a polycentric landscape that was quite different from what was generally observed elsewhere (Plane, 1981). Fowler and Jensen noted similar phenomena with regards to a broad range of labor market delineations in the U.S. (Fowler & Jensen, 2020). All this uncertainty around the definition and fit for delineations suggests that delineations should include resources to document the implications of definitonal choices as well as metrics for evaluating the fit of individual observations.

Building uncertainty into methods of delineation is one area where significant progress has been made in recent years. Recent efforts employ a range of techniques for dealing with the inevitable 'fuzziness' of data and meaning (Foote et al., 2021; Fowler & Jensen, 2020; Halás et al., 2019; Kropp & Schwengler, 2016; Wei et al., 2021; Tong & Plane, 2014; Dash Nelson & Rae 2016). 'Functional regions' are

typically understood as geographic areas with strong internal connections and weak connections with other regions (Halás et al., 2018). What constitutes a connection can vary considerably, but frequently includes commuting patterns, phone calls, correlation in wages, and housing moves (Fowler & Jensen, 2020; Karlsson & Olsson, 2006). Of these, commuting patterns are by far the most widely used, at least in part because of data availability. There is a long history of delineating functional regions with commuting data (for a partial history see Coombes, 2014). Numerous proposals have emerged that rely on different methodological tools to define functional regions in ways that ostensibly overcome limitations of extant delineations. (Fowler et al., 2016; Kropp & Schwengler, 2016; Wei et al., 2021; Tong & Plane, 2014; Dash Nelson & Rae 2016). All of these come up against a complex reality of human spatial organization that means no delineation will be perfect and every delineation will include geographic variations in quality (Fowler & Jensen, 2020; Plane, 1981). Whereas earlier contributions in this literature tended to focus on variations in the quality of delineations (e.g. Plane, 1981), later contributions tend to conceive of functional regions in terms of fuzzy membership that expressly deal with uncertainty of membership (Kropp & Schwengler, 2016; Flórez-Revuelta et al., 2008; Halás et al., 2019). The advent of new techniques for delineation comes with both challenges and opportunities. The new techniques leverage sophisticated tools for identifying relationships in data, but they tend to obscure the decision-making processes that underlie a given delineation. With additional complexity it becomes hard to distinguish quality since observations that fit well simply confirm our expectations and outliers could either reflect relationships that previously went undetected or else failures of the delineation mechanism.

The discussion above leads us to address uncertainty in our data as a key part of the delin- eation process, to incorporate that uncertainty and degree of 'fit' into any final delineation product, and to prefer simple methodologies for delineation over more complex ones so that we can better understand the mechanics that change our delineations over time. To the first point, we address uncertainty in our data by testing the consistency of our delineations using the MOE provided with the ACS data products. To the second point, we follow Fowler and Jensen (2020) in incorporating uncertainty into our delineation by providing a fit measure for observations within our delineations. To the third point, we argue for a retention of the 'simple' methodologies used to delineate urban and rural places over recent decades even as more complicated methods relying on machine learning and other algorithms improve and gain wide acceptance. While arguing for the retention of comparatively simple methodologies, we provide the tools for replicating those methodologies with explicit recognition of the size and connectivity thresholds that act as key methodological choices within the delineation so that users can adjust those thresholds to reflect different understandings of connectivity, urban density, and size thresholds for defining classes of urban agglomeration. All of this work has a very applied purpose. There needs to be an ongoing discussion of how federal standards for delineating counties and tracts within these standards can be updated. We hope that the information we provide here and the analytic tools that accompany our work, available at https://github.com/csfowler/uncertainCBSAandRUCA can support careful implemen- tation of the next generation of these delineations.

## 3 Data

The data for this analysis are chosen to conform to that used in the CBSA and RUCA delineations associated with the 2010 decennial Census.[2] This means using the journey to work flow data from the 2006–2010 ACS five year summary data (U.S. Census Bureau, 2010a). We modify these data to remove destinations abroad and outlying territories. We retain data for 3,221 counties or county equivalents (counties for short) covering all fifty states and Puerto Rico. We match this to Core-Based Statistical Area delineations from 2013 (U.S. Census Bureau 2013).[3] We also employ a county to urban area crosswalk (U.S. Census Bureau, 2010b) that provides the underlying criteria for designating counties as "central." These underlying data are crucial to an accurate replication of the OMB delineation because several counties have their designation changed in the delineation process from 'central' to 'outlying' so that the final status in the Census designation file is not reliable for replication. Finally, we use the county adjacency file produced by the Census to identify county neighbors, another key crite- rion for inclusion in a given CBSA (U.S. Census Bureau, 2018). For the purposes of conducting spatial overlays and visualizing our results spatially we access 2010 era county boundaries via the Tidycensus package (Walker & Herman, 2022).

For tract to tract commuting flows we use a special tabulation of this information made available by the Federal Highway Administration through the Census Transportation Planning Products (CTPP) program (U.S.D.T. Federal Highway Administration, 2013). These are drawn from ACS surveys over a 5 year period and are sample based estimates with very high published margins of error. As with the county to county flows we retain information from all fifty states and Puerto Rico. To ascertain the population in tracts that is in urban areas, urban clusters, or neither we employ tract level data from the National Historic Geographic Information, IPUMS system (Manson et al., 2022). We obtained the spatial information on urban agglomerations and water-clipped tract boundaries for the purpose of conducting overlays from the same source. For comparison of our results with the existing delineation we utilize both the published RUCA codes from the Economic Research Service (ERS) and the unpublished methodology provided by ERS staff which includes interim data steps that permit us to identify the basis for differences we observe.

As a further check on our analysis we use tract to tract commuting flows provided as part of the Longitudinal Employer-Household Dynamics, the LEHD

---

[2] RUCA's and CBSA's for 2020 had not been released at the time this article was published. A mid-decade delineation of CBSA's was available based on commuting flows from the 2011–2015 ACS. However, we chose the original delineation conducted after the 2010 census (with 2005–2009 ACS data). This is because urban areas, a key input to CBSA's, are only updated every 10 years meaning that the mid-decade delineations used new commuting flows overlaid on the old urban area definitions creating unknown inconsistencies.

[3] More recent, mid-decade, delineations of CBSA's do exist, but they rely on updated population estimates but 2010 delineations of urbanized areas. The static nature of the urbanized areas causes the determination of metropolitan and micropolitan status to be slightly out of line with the intention of the delineation, so we prefer the 2013 era delineations where the data are all in sync with one another.

Origin–Destinations Employment Statistics (LODES U.S. Census Bureau, 2010b). These are partially synthetic data built from administrative records provided by the Bureau of Labor Statistics (BLS) and the Office of Personnel Management (OPM) covering a broad range of employment in the U.S. The data are not sample based like the ACS, but are partially synthetic to protect privacy. These data have the advantage of being built from a much larger and more complete set of information (primarily unemployment insurance filings) but they differ from the 'real' data in ways that are, by definition, not knowable. The use of LODES data serves an additional purpose. LODES is frequently proposed as an improved data source for work on commuting, but the impact of using its partially synthetic data is not fully tested. While we cannot specifically know whether the ACS sample or the synthetic LODES data are more correct, it is helpful to at least characterize their differences. One important issue we face is that the 2010 LODES data that we employ appears to be missing roughly 1,000 tracts spread across the U.S. The tracts come from a wide range of states, and capture areas that are both urban and rural. A better understanding of where, why, and when tracts are missing will be necessary to fully assess the suitability of the LODES data for future use. For now we offer a comparison with the available data for the most comparable year.

## 4 Method

A key contribution of this paper is to programmatically replicate and validate the official methodologies for delineating metropolitan areas and RUCA codes. We base our methodology for delineating metropolitan/micropolitan/nonmetropolitan on the published methodology in the Federal Register (Office of Management and Budget, 2010). Our methodology for delin- eating RUCA codes is based on the published description of these codes (U.S.D.A. Economic Research Service 2023) as well as the participation of the delineation's author at ERS. In each case we first develop a function that takes population, urbanization, and commuting flow data as inputs and assigns counties and tracts according to the published methodology.

A contribution of this paper is that we provide detailed scripts for downloading the appropriate data as well as functions that conduct the precise steps necessary to replicate delineations given data from an arbitrary year. The publication of this code anticipates the need to apply the same methodology to new data from the 2020 Census. As far as we know, neither the CBSA nor RUCA delineation methodologies have ever been made publicly available in a way that permits replication across varied data sets and assumptions. The code is available at https://github.com/csfowler/uncertainCBSAandRUCA.

Subsequent to validating that our method replicates the official delineations, we apply the same delineation functions just described to alternative flow matrices generated based on random draws from the point estimates and MOE's provided with the ACS data. For every origin and destination pair in the ACS commuting data we first generate a standard deviation by dividing the MOE by 1.645 (the procedure recommended by the Census). Next, we randomly sample 1,000 values from the distribution centered on the point estimate and the standard deviation. This results

in 1,000 different versions of the complete flow matrix. We use each of these flow matrices as an input for the validated delineation methodologies and produce 1,000 alternative delineations. We extend our comparison by also using the LODES flow matrix to produce one additional delineation.

Our analysis is designed to concisely characterize uncertainty in federal delineations. We employ a simple measure of 'consistency' applied to individual counties and tracts that is simply the percentage of our 1,000 alternative delineations where the assigned value matched.the official designation. For example, a county that was assigned to its official CBSA in 600 of the 1,000 flow matrices would receive a consistency score of 60%. Divergence from the assigned delineation could happen when a county was assigned to a different CBSA or not assigned to any CBSA, or when a tract received a different RUCA designation. One notable limitation of this very simple metric is that, while we do consider counties that are assigned to the same category but a different CBSA as different we do not consider tracts that retain the same RUCA code, but are attached to a different urban area as having changed. Another limitation of our measure is that we do not offer much nuance in terms of what constitutes a failure to match. A tract that goes from a RUCA score of 1 (core metropolitan) to 2 (commuting to core metropolitan) is treated identically if it gets assigned a value of RUCA 10 (rural). Alternative assignments tend to mark relatively small shifts, so we do not lose much information with this decision, but detailed examination of changes is probably warranted as further exploration of the implication of threshold changes and new data continues.

Our method for stochastically generating alternative flow matrices in the ACS is not ideal since the random draws we enact do not account for correlations among flows. In constructing the point estimates and MOEs for the ACS the Census brings housing unit and population counts in line with other published data by assigning person and housing weights to individual survey responses. Each response is assigned an integer weight so that the sum of all weights for persons in a block group aligns with the total population for that block group in the Census population estimates program. The weight given to each specific survey is further refined to bring counts for specific sub-populations in line with population estimates as well. Since there is no objectively correct weight or combination of weights that will allow the sample to completely replicate the true population, the Census employs a procedure called successive differences replication (SDR, Fay & Train, 1995; cited in Spielman & Folch, 2015, p. 1006) to generate eighty distinct 'replicate weights' that offer different viable combinations of weights that all add up to the population estimate totals. Point estimates and MOEs are subsequently generated from these eighty replicate weights tables. The key advantage of these replicate weights is that as weights on one individual go up, weights on another necessarily go down. For our purposes this means that, in comparing replicate tables, increased flows between an origin and a destination would be balanced by decreased flows between that origin and another destination. In contrast, the random sampling strategy we employ treats every origin–destination pair as independent, so increases in one flow can accompany increases in another or vice versa. Replicate weights are available for many variables in the ACS, but not for commuting flows. This means that in our analysis every alternative estimate for flows from a given origin tract or county that we

generate is sampled independently. This likely increases the measured uncertainty for some edge cases (and explains the existence of tracts that are always assigned a different RUCA code than the one in the official delineation). With these exceptions aside, independent draws would generally tend to reduce our measure of uncertainty as increases in one flow can be matched by increases in another flow increasing the likelihood that their relative importance will remain in place. One concrete recommendation we have from undertaking this analysis is that the replicate weights used to produce the point estimates and MOEs for the ACS commuting flows should, in the future, be retained and made publicly.available as they are for so many other variables.

### 4.1 Replicating Metropolitan, Micropolitan, Nonmetropolitan and CBSA Delineations

Counties are assigned to one of three categories by OMB: metropolitan statistical area, microp- olitan statistical area, or 'outside core based statistical areas.' These latter two designations are often combined and referred to as 'nonmetropolitan.' County status is assigned based on criteria listed in the Federal Register (Office of Management and Budget, 2010). The delineation depends first on the identification of urbanized areas/urban clusters of sufficient size (urbanized areas of at least 50,000 people for metropolitan status, urban clusters of at least 10,000 people for micropolitan status).[4] A complicated process of identifying 'central' or 'outlying' status ensues based on the share of the resident population that works in adja- cent counties. A threshold of 25% of the workforce–either commuting to a central county or commuting from a central county is used to signal connection between an out- lying county and a given urban core. With a few extra rules about adjacency and combining central counties that are closely linked the county-level designations and CBSA codes emerge. Our analysis replicates the published methodology from the Federal Register to ensure that our function for assigning counties to a particular status is accurate (Office of Management and Budget, 2010). We construct our func- tion so that the thresholds for urbanized areas (50,000), urban clusters (10,000), and commuting connection (25%) can all be adjusted by the user for maximum flexi- bility anticipating that these thresholds may be up for debate as future delineation standards are revised.[5] We successfully match the assignment of all 3,221 counties in our data to the correct metropolitan/micropolitan/nonmetropolitan designation and the correct CBSA.

Having established the accuracy of our function for replicating the official OMB definitions, we subsequently apply the replicating method to all one-thousand

---

[4]    The distinction between urbanized areas and urban clusters was based solely on population size (above and below 50,000 people) and the labels were dropped in 2020 in favor of the all-inclusive 'urban areas'(U.S. Census Bureau, 2022).

[5]    In revising the CBSA standards for 2020, OMB proposed raising the threshold for an urbanized area to qualify as a metropolitan statistical area to 100, 000 people. The proposal was ultimately rejected fol- lowing a review of public comments (Office of Management and Budget, 2021).

alternative flow matrices and the LODES flow matrix. For each county in each delineation we record whether the assigned CBSA code (or nonmetropolitan status) is consistent with the official delineation. We generate our measure of consistency and use that measure in the visualizations that follow.

## 4.2 Replicating RUCA Codes

RUCA codes are meant to mirror the terminology of the OMB metropolitan delineation but at a finer geographic scale. The methodology allocates census tracts to one of ten primary categories based on the degree to which they are themselves urban and the degree to which they are connected to urban places via commuting patterns. The numeric designations 1, 4, and 7 indicate that a tract is part of the urban core of a metropolitan, micropolitan, or small town urban area/cluster respectively. Codes 2, 5, and 8 indicate tracts with high (> 30%) levels of commuting to a core (2 commutes to 1, 5 commutes to 4, etc.…), while codes 3, 6, and 9 indicate tracts with low (10%-30%) levels of commuting to a core. Code 10 designates rural areas dominated either by internal flows or commuting to other rural areas. The published RUCA codes also contain a secondary code that establishes the type of connection indicated by the second largest flow for a tract. These secondary codes are not replicated here, but because they are, by definition, smaller than the primary flows we would expect them to vary with somewhat higher frequency than the primary codes.

Because the RUCA codes rely on the same population density-based delineation of urbanized areas and urban clusters and use identical thresholds to assign metropolitan (50,000 people) and micropolitan (10,000 people) status to urban cores the two delineations are quite similar. The finer scale of the RUCA delineation is useful for understanding spatial variation within counties, a crucial concern given the heterogeneity of county sizes (Curtis et al., 2012).

Our approach to analyzing the RUCA codes is quite similar to that used for the metropoli- tan/CBSA delineation. We begin by testing our ability to replicate the official assignment and then move on to apply our replication script to one-thousand alternative flow matrices. Our ability to exactly replicate the official RUCA codes is complicated by some small elements of randomization in the original delineation.

The most significant difference occurs in the assignment of a tract to an urbanized area or urban cluster when the tract overlaps parts of more than one of these areas. Specifically, the original methodology incorporated a census table listing tract population in an urbanized area, in an urban cluster, or not in either. The intended use of this table is to assign the tract to the correct agglomeration type based on which type contains the larger population. When there is one UA and one UC there is no issue and the one with the larger portion of the tract population in it is assigned. The problem occurs if the tract overlaps more than one UA or more than one UC. In that case the census table does not tell us how the population is allocated between the two areas of the same type. Faced with this uncertainty the original methodology assigned a tract randomly to one of the overlapping areas. Our method brings in additional information on the land area of overlap and chooses the UA or UC with

**Table 1** RUCA Code Reconstruction

|  |  | Unmatched | Matched |
|---|---|---|---|
| *Urban Area Assignment* | | | |
| Chose UA with larger overlap | | 327 | 72,997 |
| *Code Assignment* | | | |
| | Chose Internal Flow | 33 | 33 |
| | Chose Lower Number | 16 | 9 |
| | Chose Randomly | 1 | 12 |
| | Zero Population Tract | 252 | 324 |
| Total | | — | 629 | 73,375 |

the largest overlap. This results in several hundred differences between our delineation and the original.

Randomness emerges again in choosing which commuting flow to treat as the major flow when two flows are tied in magnitude. The original method chose randomly. Our method first tries to choose a flow that is internal to the urban area (UA or UC). If an internal flow is not one of the tied flows our method chooses the flow that connects the observation to a larger urban type (metropolitan > micropolitan > small town). Finally, if the tied flows are external and to urban types of the same category, we choose randomly. In total, the random choice is only invoked thirteen times in 74,002 tracts.

Because our method reduces the randomness in selecting a RUCA designation for each tract in each of the cases just described we use our own delineation as the base case for comparison with the simulated flows rather than using the official delineation. Some differences emerge from the stochasticity in the assignment process, but these represent a very small portion of the overall comparison.

One other difference emerges between our replication and the official delineation. The RUCA methodology designates zero population tracts as category '99'. However, some zero population tracts are employment destinations. The original methodology ultimately assigns these tracts a RUCA code inconsistently. Specifically, zero population tracts show up with a mixture of '99' and other designations in cases where flows to the zero population tract represented the largest flow out of a populated tract. Our methodology enforces the intended '99' designation for these tracts creating one final source of difference between our delineation and the official delineation. Table 1 reports on the differences between our delineation and the official RUCA delineation.

With a functioning replication script in place we use the one-thousand alternative flow matrices as inputs to the script and generate a table containing one-thousand alternative delineations.

### 4.3 LEHD-LODES Data

Given the relatively small sample size of the ACS commuting flow data, the LEHD LODES data are often suggested as an alternative source of commuting information

**Table 2** County designation changes across 1,000 simulations

| | Simulation Outcome | | | |
|---|---|---|---|---|
| | Consistent% | Metropolitan% | Micropolitan% | Nonmetro% |
| Metropolitan | 98.86 | 0.08 | 0.00 | 1.06 |
| Micropolitan | 97.73 | 1.01 | 0.07 | 1.19 |
| Nonmetro | 97.90 | 1.07 | 1.03 | 0.00 |

that could be used to delineate CBSA's or RUCA codes. In the interest of exploring the utility of these data we apply our function for assigning counties to CBSA's and assigning tracts to RUCA's using the LODES data for 2010. We report the results of this matching process in our analysis below. As noted above, our comparison is complicated by 926 missing records for populated tracts.

## 5 Analysis

Our analysis of the alternative flow matrices focuses on the magnitude and characteristics of differences between the official delineations of metropolitan/CBSA and RUCA codes. In each case we measure the frequency of alternative assignments for particular spatial units and seek to describe the characteristics of the units where we do observe differences. In general, the results show a relatively small number of alternative assignments (relatively high levels of consistency), but some units do get assigned differently with high frequencies.

### 5.1 CBSA Assignment Consistency by Type of County and by Location

Overall, the assignment to CBSA and metropolitan/micropolitan/nonmetropolitan status is quite consistent. Across all 3,221,000 assignments (1,000 flow matrices applied to 3,221 countries) Table 2 documents that over 98% of the time a county is assigned to the same status as in the official delineation. Within that high level of overall consistency, however, 15% of counties have a different assignment at least once and there are a small number of places for which assignment is highly uncertain.

We can see where this uncertainty occurs in Fig. 1, a map showing the rate of inconsistent assignment for counties across the entire country. Inconsistency in a county's assignment does not occur with any spatial pattern,[6] though it is notable that some counties have inconsistent assignments at very high rates.

Figure 2 breaks down patterns of inconsistent assignment by county population and by share of the population that is Hispanic and Non-Hispanic Black.

---

[6] Moran's *I* test for clustering of similar results is 0.005, a value very close to the expected result of a random distribution.
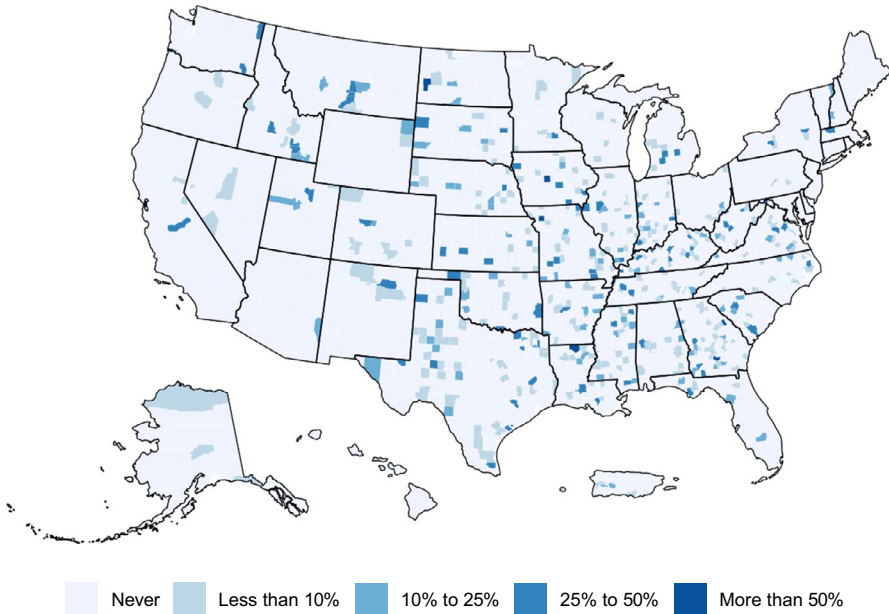
**Fig. 1** Percent of assignments that were inconsistent

Inconsistently assigned counties tend to have slightly smaller shares of Hispanic and Black populations than counties as a whole. Moreover, the population distribution shown in Fig. 2 demonstrate that the inconsistently assigned counties are not the smallest or largest counties, but are also generally from the middle of the distribution; counties of between 10,000 and 80,000 persons. The largest county that has an inconsistent assignment is Hall County, Georgia at 180,000 people. The smallest is Loving County, Texas at 83 people which gets changed from nonmetropolitan to metropolitan.

In 49% of our alternative flow matrices. In all there are 182 counties that are consistent less than 90% of the time. A small, but significant minority, perhaps acceptable when viewed on a national scale, but certainly of real importance to the 3.27 million people living in those counties.

## 5.2 LODES Data Metropolitan/Micropolitan/Nonmetropolitan Comparison

Examining the results of a delineation process conducted with the LEHD-LODES data gives us another glimpse into how the use of a different data source has the potential to change our understanding of metropolitan, micropolitan, and nonmetropolitan status. Figure 3 reveals two things about the LODES data. First, there are more consistent counties in the LODES delineation than are *always* consistent in the alternative flow data (90% in LODES compared with 86% using the alternative flow matrices). However, reflecting back on the results reported in Table 1, the alternative
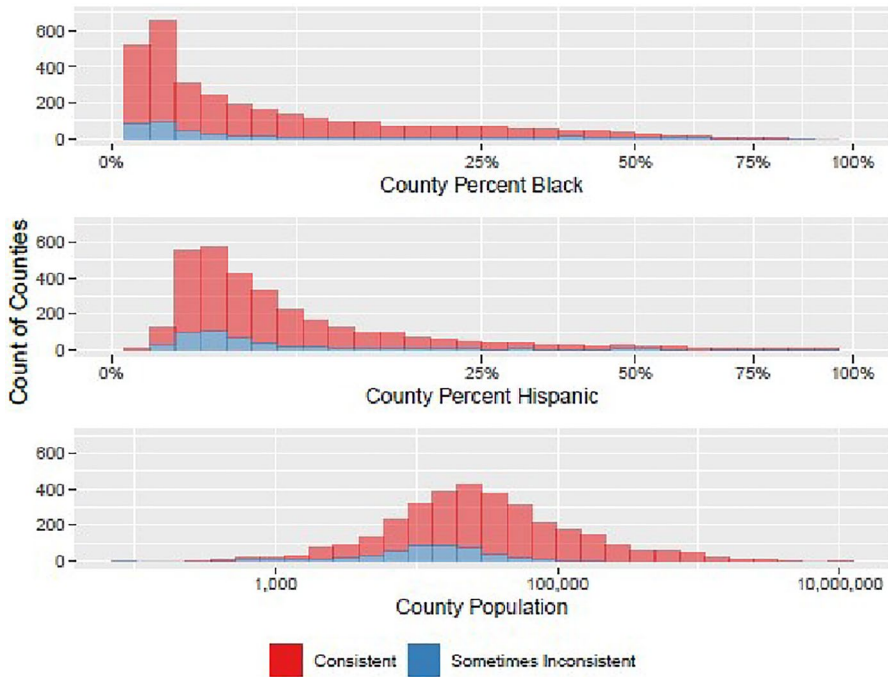
**Fig. 2** Distribution of consistent and inconsistent assignments

flow matrices are consistent about 98% of the time (because most counties are only inconsistent in a small number of alternatives). This indicates a complex structure for the uncertainty we are trying to measure and complicates any claims we might make about which data source is more reliable vis a vis the unknown 'true' designation. Second, the sankey diagram shows that the counties that were consistent or inconsistent were not the same in the LODES and the assignment through alternative flow matrices. Some of the counties that were consistent for 1,000 random draws were not consistent using LODES. This suggests that the LODES data are, at times, well outside the distribution indicated in the ACS data. Whether this is because LODES is capturing a different sample of commuters than the ACS, a function of the synthetic data, or inaccuracy in the ACS sample is unknown.

## 5.3  RUCA Assignment Consistency

Our analysis of RUCA codes follows roughly the same pattern as for metropolitan/micropolitan/nonmetropolitan county assignment.We first examine the rate of matching and find that almost 97% of the time alternative flow matrices produce assignments that match their assigned RUCA code (Table 3 based on our replication assignment function, not the official ERS assignment, see above). Looking at tracts individually, 87% if them are always assigned to the same RUCA code (e.g. 100%

**Table 3** Tract RUCA designation changes over 1,000 alternative flow matrices

| Status | Frequency | Percent(%) |
|---|---|---|
| Unmatched | 2, 263, 397 | 3.10 |
| Matched | 71, 738, 603 | 96.90 |

consistent). However, those tracts that do end up with a different assignment cover the entire range from unmatched 1 time in 1,000 to never matching.

The existence of tracts that never match their expected RUCA assignment suggests something more than simply stochastic processes and so we zoom in to a specific tract to see how this result occurs. A closer look at Tract 263,200 in Litchfield County, Connecticut, which never matches its official assignment, shows a rural tract dominated by internal flows (280 out of 884) with lots of small connections to tracts within the New York City metropolitan area. These small connections have almost uniformly high MOE's (123) and so the simulation process always ends up producing a result where these small flows add up to a number that exceeds the simulated internal flow. As a result, Tract 26,300 is always assigned a RUCA code 2 (Metropolitan area high commuting) as a tract dominated by flows into a large metropolitan area instead of a 10 (Rural area). While Tract 263,200 is a rather extreme case, it
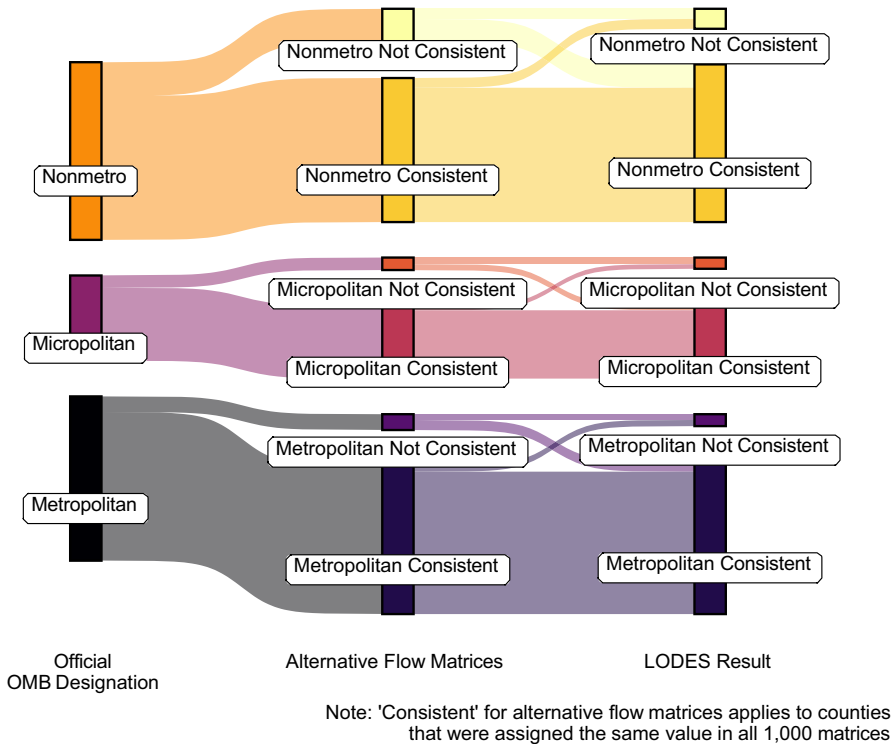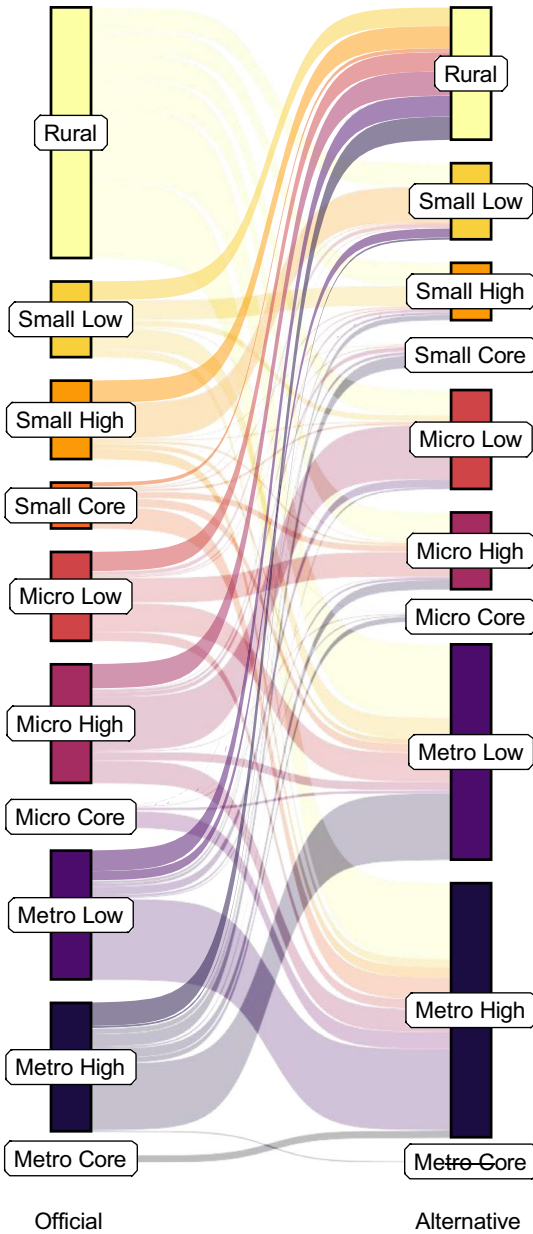


Note: 'Consistent' for alternative flow matrices applies to counties that were assigned the same value in all 1,000 matrices

**Fig. 3** Consistency by metropolitan/micropolitan/nonmetropolitan status

**Fig. 4** Transitions in RUCA codes

is worth noting that 7.3% of tracts are consistent with their official designation less than 90% of the time and 2% of tracts are consistent less than half of the time.

Examining how tracts changed assignment we observe that within the broader stability of the official delineation there is still the possibility for an alternative designation across the rural-urban continuum. Figure 4 demonstrates that rural tracts (RUCA 10) are the most likely to change assignment, but assignment changes are present in all 10 classes and inconsistent assignments move to almost every other class. While slight moves (up or down one on the RUCA continuum) dominate, 63% of the moves are more than one position on the continuum and significant numbers of tracts go up or down nine positions. While the patterns are complex, tracts that have strong commuting to a core (codes Metro High, Micro High, and Small High) tend to move to less intense/smaller core positions (up the continuum) while core and low commuting tracts tend to move to larger/more intense positions (down the continuum).

## 5.4  LODES Data RUCA Comparison

Finally, we turn to the comparison of the RUCA codes based on LODES data. Figure 5 shows the transition matrix for the 8,392 (11.3%) of tracts that were not consistent between the LODES assignment and the replication assignment. Notably, all the Metro High (RUCA 1).

Tracts matched the replication, but 57% of the 1,502 missing tracts were designated Metro High in the replication delineation and most of the rest (32%) were zero population tracts. Aside from this trend, the transition matrix shows a relatively broad distribution of transition outcomes with a roughly normal distribution of values around the expected match. Think- ing broadly about the implications of this result we cannot allocate uncertainty between the LODES and ACS data, but the mismatch of 11% clearly impacts a large number of tracts and, by extension, people impacted by an uncertain designation. The balance between over and under prediction in the transition matrix does suggest, however, that neither data source is biased either high or low relative to the other.

## 6  Discussion

The analysis described above indicates that, while there is considerable uncertainty in the point estimates for commuting flows for both counties and tracts, that uncertainty does not affect the metropolitan/micropolitan/nonmetropolitan or RUCA designation for most people or most places. With any delineation there will always be edge cases that do not neatly fit into one category or the other, but both the CBSA and RUCA delineations are robust to the uncertainty in the underlying data for the vast majority of counties and tracts. About 85% of counties and 87% of tracts are always consistent within our alternative flow matrices and 90% of counties and tracts are consistent when calculating with the LODES data. More broadly, of the
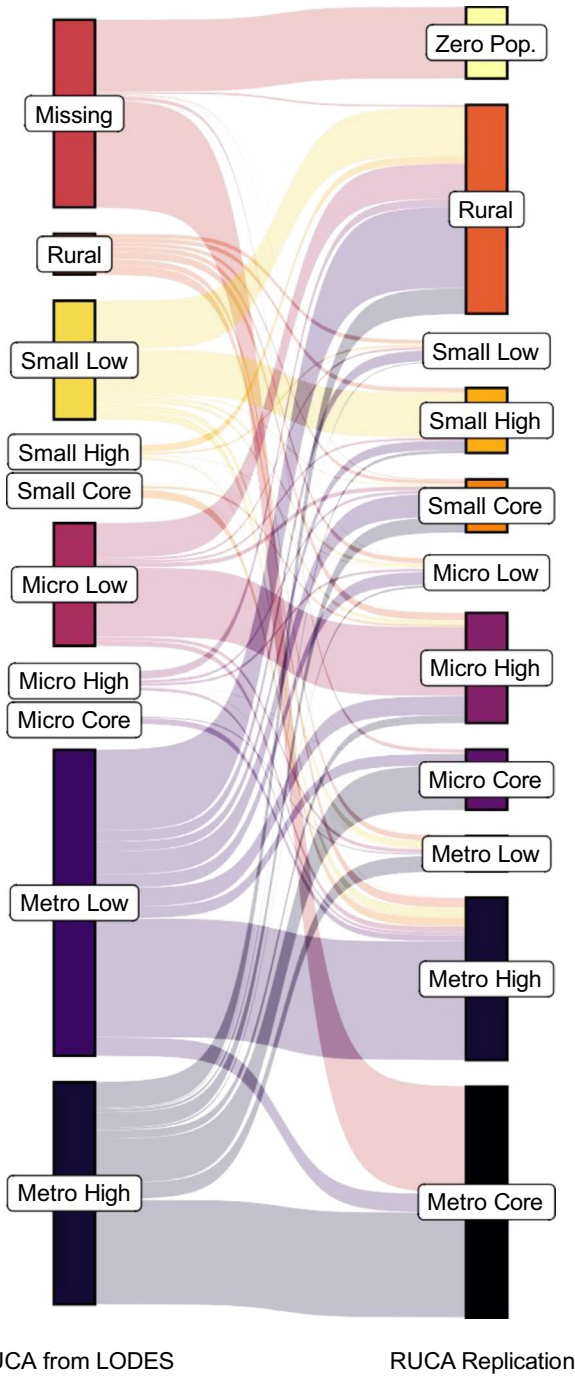
**Fig. 5** Transitions in RUCA codes comparing LODES and RUCA replications for unmatched tracts only

77,300,223 assignments we made in our analysis 98% of counties and 97% of tracts were consistent with the official delineation.

Employing an alternative data source like the LEHD LODES would change the designation for many places but it is not possible to know whether those changes increase the accuracy of the delineation or simply change the structure of the uncertainty. Notably, the delineations based on LODES data diverge from the official delineation in different ways than the alternative flow matrices. More importantly, since the alternative matrices give us a likelihood that another designation is possible, we can focus on the small number of places where this likelihood is high as opposed to the binary consistent/inconsistent produced with the LODES data.

This analysis could benefit significantly from access to the underlying data that produce the public facing versions of the ACS and LODES data products. Access to the replicate weights information for the ACS would allow us to better understand how uncertainty in flows is correlated for groups of origin–destination pairs. While the data to undertake this analysis is available within the Federal Statistical Research Data Centers (FSRDCs), current Census policy severely restricts the release of data for populations smaller than a Congressional District. Public LODES data are produced as a synthetic data product that introduces noise to protect individual privacy. Previous work by Fowler et al. (2016) has found the level of noise to be too high for some kinds of delineation processes to work. While, the analysis here found the distribution of LODES assignment to be unbiased with respect to the ACS, the unknown structure of the noise does suggest that an examination of the original data could be extremely useful.

The nature of federal designations is that for most applications the assignment process is understood as definitive and all of the outcomes that derive from that assignment process are contingent on the result. Our work here suggests that, for some places, these assignments are not as clear cut as they may seem. This has implications for funding allocated based on these designations, for research conducted with these designations and for the outcomes that derive from funding and research. It is important to recognize that we do not have a true picture of the actual commuting patterns of the population, our data have uncertainty built into them. Even if we did have perfect data, they would only be a snapshot subject to change and even shock (as in the tremendous shift to working from home during the Covid 19 pandemic). Taken together this understanding of uncertainty indicates that we should treat these designations as at least partially contingent and should be aware of the degree to which a set of simple categories hides degrees of fit and certainty within it.

The changes in working from home accelerated by the Covid 19 pandemic raise further ques- tions about the suitability of even using commuting patterns as the basis for delineating these types of patterns. In past decades the Bureau of Labor Statistics delineated 'Economic Areas' based on the extent of newspaper subscriptions, a method with significant appeal made almost completely irrelevant in recent decades by the decline of local journalism and the changing pat- terns of news subscription. Perhaps it is time to retire commuting patterns for the same reason. We know that rural counties often gain metropolitan status because their local economies de- cline, increasing the number of long commutes into nearby metropolitan

area.s It could be argued the decline of local employment should not trigger a move from nonmetropolitan to metropolitan, though such shifts are possible under current methodologies. Some alternatives have been proposed based on travel time, activity spaces, cellphone data, etc., but we are unaware of any systematic effort to compare the implications of these alternative data sources to solving the problem of delineation.

Furthermore, our methods for designating places are themselves imperfect. Even if we find that a place is consistently designated in the same way, this does not definitively indicate that the designation is correct. One of the authors' own home in State College Pennsylvania is designated as metropolitan by OMB, but starting from the center of town a person can walk three miles in any direction and be standing in farmland. This is clearly not consistent with what most people would understand as metropolitan yet peculiarities of settlement, county size, and the powerful commuting draw of Penn State University all combine to pull the designation into the metropolitan column. Part of what the analysis conducted here points to is that we need to carefully examine the assumptions that underlie federal designations. Measuring uncertainty is only the first step in how we should interrogate our designations. A key second step is to carefully examine the methodology to identify hidden assumptions or assumptions that are themselves subject to difference of opinion. Reproducing the official delineations allows us to highlight the thresholds for population size, commuting share, land area overlap and other variables that shape the delineations we ultimately observe. Future work will need to systematically examine the implications of adjusting these variables.

The replication scripts that accompany this paper provide the opportunity to adjust these thresholds as parameters to see how important they are to the final designations. In anticipa- tion of the release of 2020 decennial census data OMB and other federal agencies opened up discussions of how these methodologies might be revised (Office of Management and Budget, 2021). The largest proposed adjustment (ultimately rejected) was an increase in the population threshold used to qualify an urban area as metropolitan from 50,000 people to 100,000 people. This change would have significantly reshaped the landscape of what is defined as metropolitan. At the very least the scripts provided here can help generate alternative delineations based on proposed changes to thresholds or allow researchers to back calculate designations based on future changes to delineation methods.

Overall we find the designations for CBSAs and RUCA codes to be relatively robust to the uncertainty that accompanies the commuting data on which they are based. While we think that incorporating uncertainty improves the usefulness of these designations we were, frankly, surprised at how little they changed given the relatively high levels of uncertainty associated with the small sample. Our work here gives us more confidence in the usability of these designations going forward even if we will continue to test assumptions and consider alternative ways of defining urban and rural places.

## Declarations

**Conflict of interest** On behalf of all authors, the corresponding author states that there is no conflict of interest.

**Disclaimer** The findings and conclusions in this publication are those of the author(s) and should not be construed to represent any official USDA or U.S. Government determination or policy.

## References

Baer, L. D., Johnson-Webb, K. D., & Gesler, W. M. (1997). What is rural? A focus on urban influence codes. *Journal of Rural Health, 13*(4), 329–333. https://doi.org/10.1111/j.1748-0361.1997.tb00975.x

Coombes, M. (2014). From city-region concept to boundaries for governance: The English case. *Urban Studies, 51*(11), 2426–2443. https://doi.org/10.1177/0042098013493482

Cromartie, J., & Bucholtz, S. (2008). Defining the rural in rural America. *Amber Waves 6*(3).

Cromartie, J. B., & Swanson, L. L. (1996). Census tracts more precisely define rural populations and areas. *Rural America/Rural Development Perspectives, 11*, 31–39. https://doi.org/10.22004/ag.econ.289755

Curtis, K. J., Voss, P. R., & Long, D. D. (2012). Spatial variation in poverty-generating processes: Child poverty in the United States. *Social Science Research, 41*(1), 146–159. https://doi.org/10.1016/j.ssresearch.2011.07.007

Fay, R. E., & George F. T. (1995). Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties. In *Proceedings of the Government Statistics Section, American Statistical Association* (pp. 154–59).

Fitzsimmons, J. D., & Ratcliffe, M. R. (2004). Reflections on the review of metropolitan area standards in the United States, 1990–2000. In Tony Champion & Graeme Hugo (Eds.) *New forms of urbanization*. Routledge. https://doi.org/10.4324/9781315248073-18

Flórez-Revuelta, F., Casado-Díaz, J. M., & Martínez-Bernabeu, L. (2008). An evolutionary approach to the delineation of functional areas based on travel-to-work flows. *International Journal of Automation and Computing, 5*(1), 10–21. https://doi.org/10.1007/s11633-008-0010-6

Folch, D. C., Arribas-Bel, D., Koschinsky, J., & Spielman, S. E. (2016). Spatial variation in the quality of American community survey estimates. *Demography, 53*(5), 1535–1554.

Foote, A., Kutzbach, M. J., & Vilhuber, L. (2021). Recalculating: How uncertainty in local labour market definitions affects empirical findings. *Applied Economics, 53*(14), 1598–1612. https://doi.org/10.1080/00036846.2020.1841083

Fowler, C. S., & Jensen, L. (2020). Bridging the gap between geographic concept and the data we have: The case of labor markets in the USA. *Environment and Planning A: Economy and Space, 52*(7), 1395–1414. https://doi.org/10.1177/0308518X20906154

Fowler, C. S., Rhubart, D. C., & Jensen, L. (2016). Reassessing and revising commuting zones for 2010: History, assessment, and updates for US 'laborsheds' 1990-2010. *Population Research and Policy Review, 35*(2), 263–286.

Fowler, C. S., Frey, N., Folch, D. C., Nagle, N., & Spielman, S. (2020). Who are the people in my neighborhood?: The contextual fallacy of measuring individual context with census geographies. *Geographical Analysis, 52*(2), 155–168.

Franklin, R. (2022). Quantitative methods I: Reckoning with uncertainty. *Progress in Human Geography, 46*(2), 689–697. https://doi.org/10.1177/03091325211063635

Franklin, R. S., & Plane, D. A. (2006). Pandora's box: The potential and peril of migration data from the American community survey. *International Regional Science Review, 29*(3), 231–246. https://doi.org/10.1177/0160017606289895

Halás, M., Klapka, P., Hurbánek, P., Bleha, B., Pénzes, J., & Pálóczi, G. (2018). A definition of relevant functional regions for international comparisons: The case of central Europe. *Area, July.* https://doi.org/10.1111/area.12487

Halás, M., Klapka, P., & Erlebach, M. (2019). Unveiling spatial uncertainty: A method to evaluate the fuzzy nature of functional regions. *Regional Studies, 53*(7), 1029–1041. https://doi.org/10.1080/00343404.2018.1537483

Isserman, A. M. (2005). In the national interest: Defining rural and urban correctly in research and public policy. *International Regional Science Review, 28*(4), 465–499. https://doi.org/10.1177/0160017605279000

Jung, P. H., Thill, J.-C., & Issel, M. (2019). Spatial autocorrelation and data uncertainty in the American community survey: A critique. *International Journal of Geographical Information Science, 33*(6), 1155–1175. https://doi.org/10.1080/13658816.2018.1554811

Jurjevich, J. R., Griffin, A. L., Spielman, S. E., Folch, D. C., Merrick, M., & Nagle, N. N. (2018). Navigating statistical uncertainty: How urban and regional planners understand and work with American community survey (ACS) data for guiding policy. *Journal of the American Planning Association, 84*(2), 112–126. https://doi.org/10.1080/01944363.2018.1440182

Karlsson, C., & Olsson, M. (2006). The identification of functional regions: Theory, methods, and applications. *The Annals of Regional Science, 40*(1), 1–18. https://doi.org/10.1007/s00168-005-0019-5

Kropp, P., & Schwengler, B. (2016). Three-step method for delineating functional labour market regions. *Regional Studies, 50*(3), 429–445. https://doi.org/10.1080/00343404.2014.923093

Kwan, M.-P. (2012). The uncertain geographic context problem. *Annals of the Association of American Geographers, 102*(5), 958–968. https://doi.org/10.1080/00045608.2012.687349

Manson, S., Schroeder, J., Van Riper, D., Kugler, T., & Ruggles, S., (2022). National historical geographic information system: Version 17.0. Minneapolis, MN: IPUMS. https://doi.org/10.18128/D050.V17.0.

Morrill, R. L., Cromartie, J., & Gary Hart, L. (1999). Metropolitan, urban, and rural commuting areas: Toward a better depiction of the United States settlement system. *Urban Geography, 20*(8), 727–748. https://doi.org/10.2747/0272-3638.20.8.727

Nelson, G. D., & Rae, A. (2016). An economic geography of the United States: From commutes to megaregions. *PLoS ONE, 11*(11), e0166083. https://doi.org/10.1371/journal.pone.0166083

Office of Management and Budget. (2010). 2010 Standards for delineating metropolitan and micropolitan statistical areas. *Federal Register* (pp. 37245–37252).

Office of Management and Budget. (2021). 2020 Standards for delineating core based statistical areas. *Federal Register*. https://www.federalregister.gov/documents/2021/07/16/2021- 15159/2020-standards-for-delineating-core-based-statistical-areas.

Plane, D. A. (1981). The geography of urban commuting fields: Some empirical evidence from New England. *Professional Geographer, 33*(2), 182–188. https://doi.org/10.1111/j.0033-0124,1981.00182.x

Porter, J. R., & Howell, F. M. (2009). On the 'urbanness' of metropolitan areas: Testing the homogeneity assumption, 1970–2000. *Population Research and Policy Review, 28*(5), 589–613. https://doi.org/10.1007/s11113-008-9121-6

Ratcliffe, M., Burd, C., Holder, K., & Fields, A. (2016). Defining rural at the US Census Bureau. *American Community Survey and Geography Brief, 1*, 8.

Robertson, C., & Feick, R. (2018). Inference and analysis across spatial supports in the big data era: Uncertain point observations and geographic contexts. *Transactions in GIS, March.* https://doi.org/10.1111/tgis.12321

Schroeder, J. P., & Pacas, J. (2021). Across the rural-urban universe: Two continuous indices of urbanization for U.S. census microdata. *Spatial Demography, 9*(1), 131–154. https://doi.org/10.1007/s40980-021-00081-y

Spielman, S. E., & Folch, D. C. (2015). Reducing uncertainty in the American community survey through data-driven regionalization. *PlosOne, 10*(2), e0115626. https://doi.org/10.1371/journal.pone.0115626

Spielman, S. E., Folch, D., & Nagle, N. (2014). Patterns and causes of uncertainty in the American community survey. *Applied Geography, 46*, 147–157. https://doi.org/10.1016/j.apgeog.2013.11.002

Sun, M., & Wong, D. W. S. (2010). Incorporating data quality information in mapping American community survey data. *Cartography and Geographic Information Science, 37*(4), 285–299. https://doi.org/10.1559/152304010793454363

Tong, D., & Plane, D. A. (2014). A new spatial optimization perspective on the delineation of metropolitan and micropolitan statistical areas. *Geographical Analysis, 46*(3), 230–249. https://doi.org/10.1111/gean.12037

U.S. Census Bureau. (2010a). 2006–2010a 5-Year ACS commuting flows.

U.S. Census Bureau. (2010b). Longitudinal employment household dynamics (LEHD) origin-destination employment statistics (LODES). https://lehd.ces.census.gov/data/.

U.S. Census Bureau. (2010c). Urban area relationship files record layouts. *Census.gov*. https://www.census.gov/programs/surveys/geography/technical-documentation/records-layout/urban-area-record-layouts.html.

U.S. Census Bureau. (2013). Core based statistical areas, metropolitan divisions, and combined statistical areas. *Census.gov*. https://www.census.gov/geographies/reference-files/time- series/demo/metro-micro/delineation-files.html

U.S. Census Bureau. (2018). County adjacency files.

U.S. Census Bureau. (2022). Urban area criteria for the 2020 census-final criteria.

U.S.D.A. Economic Research Service. (2023). Rural urban commuting area codes for 2010.

U.S.D.T. Federal Highway Administration. (2013). Census transportation planning products 2006–2010 census tract to tract commuting flows. Federal highway administration.

Vowell S. (2021). Opinion | my beloved college town has a problem: It's too popular. *The New York Times*, May.

Walker, K., & Matt, H. (2022). Tidycensus: Load US census boundary and attribute data.

Wei, R., Rey, S. J., & Grubesic, T. H. (2021). A probabilistic approach to address data uncertainty in regionalization. *Geographical Analysis*. https://doi.org/10.1111/gean.12282

Wright, R., Ellis, M., & Tiao, N. (2022). Making metros white? The effects of US metropolitan reclassification on racial compositional change. *The Professional Geographer., 74*(4), 659–667.