



UNIVERSITY OF MINNESOTA

**Across the Rural-Urban Universe:
Two Continuous Indices of Urbanization for U.S.
Census Microdata**

Updated Version at <https://doi.org/10.18128/MPC2019-05.v2>

Jonathan Schroeder[†]
University of Minnesota

José Pacas
University of Minnesota

December 2019

Working Paper No. 2019-05
DOI: <https://doi.org/10.18128/MPC2019-05>

[†] Address correspondence to Jonathan Schroeder, University of Minnesota, Institute for Social Research and Data Innovation, 50 Willey Hall, Minneapolis, MN 55417 or email jps@umn.edu. Support for this work was provided by IPUMS USA (R01HD043392) and the Minnesota Population Center (P2C HD041023).

Abstract

To distinguish urban and rural populations in the U.S., a common practice is to use one of the official federal classifications of metropolitan/non-metropolitan (i.e., metro/nonmetro) or urban/rural areas. These classifications, however, are not identified in public use census microdata, which makes it difficult to investigate relationships between the urban/rural status of individuals and their demographic characteristics, socio-economic status, and living arrangements. The standard classifications are also coarse, masking potentially important variations, and they represent two distinct dimensions of settlement patterns, so it can be misleading to use only one classification (as is common) to generalize about urban-rural disparities. To address these limitations, we compute two continuous indices for public use microdata—average tract density and average metro/micro-area population—using population-weighted geometric means. We show how these indices correspond to two key dimensions of settlement patterns, and we demonstrate their utility through an examination of disparities in poverty throughout the rural-urban universe. Poverty rates vary across settlement patterns in nonlinear ways: poverty rates are lowest in moderately-dense parts of major metro areas, and rates are higher in both low- and high-density areas, as well as in smaller commuting systems. Using the two indices also reveals that correlations between poverty and demographic characteristics vary considerably across different settlement types. Both indices are now available for recent census microdata via IPUMS USA (<https://usa.ipums.org>).

Introduction

Public Use Microdata Sample (PUMS) files are one of the U.S. Census Bureau’s most valuable data products for social science and policy analysis, providing detailed questionnaire responses from the decennial censuses and the American Community Survey (ACS) for a large sample of the U.S. population. Using PUMS data, researchers can generate custom cross-tabulations with maximum flexibility and investigate relationships among all reported characteristics for individual respondents and their households. One limitation of PUMS files is that, in order to protect privacy, the Census Bureau restricts the detail of reported geographic information. The only sub-state geographic units identified are Public Use Microdata Areas (PUMAs), which are custom-designed agglomerations of other standard units (census tracts, counties, etc.), each required to have at least 100,000 total residents. This restriction makes it impossible to identify smaller communities and neighborhoods in PUMS data, and identifying larger regions is also often complicated by mismatches between PUMAs and other geographic units.

IPUMS USA (<https://usa.ipums.org>), a website that disseminates harmonized U.S. census microdata, has developed numerous tools and resources to facilitate microdata access and use. To expand on the limited geographic information provided in PUMS files, IPUMS USA supplies supplemental variables that identify several standard geographic units other than PUMAs. IPUMS can only identify units that correspond well to a set of PUMAs, but this approach has still enabled the identification of hundreds of counties, cities, and metropolitan areas for most decennial and ACS microdata samples.

In this paper, we introduce two PUMA-based indices that IPUMS USA recently added to its collection of supplemental variables in order to facilitate analysis of demographic variation across different levels of urbanization. The two indices—*average tract population density* and *average metro/micro-area population*—correspond to two distinct dimensions of settlement patterns: “concentration” (the local intensity of settlement) and “size” (the total population of the commuting system). For analysts seeking to distinguish rural and urban populations in microdata, IPUMS USA has also long provided a categorical variable named “METRO,” which identifies metropolitan status and central/principal city status based on PUMA information. We demonstrate here how the new indices offer

valuable advantages relative to METRO. Crucially, because they are continuous and represent two distinct dimensions, the new indices distinguish a much broader range of variation across the rural-urban universe of settlement patterns.

In succeeding sections, we first discuss the limitations of standard metropolitan and urban classifications and the relative value of multi-dimensional, continuous indices. We then provide the exact definitions of the new IPUMS USA indices and discuss how they correspond conceptually to two important dimensions of settlement patterns. Finally, we demonstrate the utility of the new indices in an examination of poverty across the rural-urban spectrum. We find that poverty rates are lowest in moderately dense parts of major metro areas, and they are high in both low-density and high-density areas, as well as in smaller commuting systems. We also find that correlations between poverty and demographic characteristics vary considerably across settlement types. More generally, our findings demonstrate the value of modeling urban/rural status as a continuously varying, multi-dimensional phenomenon, an approach that is directly facilitated by the new indices from IPUMS USA.

Limitations of Standard Classifications

To distinguish rural and urban populations, analysts commonly use one of two classification systems defined by federal agencies: the core-based statistical area (CBSA) definitions of the Office of Management and Budget (OMB), which delineate metropolitan and micropolitan statistical areas (i.e., metro and micro areas), or the official urban/rural classification of the Census Bureau.¹ The Bureau's criteria and guidelines for PUMA delineations² do not require any agreement with CBSAs or urban/rural

¹ The Census Bureau's urban/rural classifications have evolved over time (Ratcliffe 2015), but since 2000, the general procedure is to define "urban areas" as groups of relatively dense neighboring (or nearby) blocks with combined populations of at least 2,500 each (Ratcliffe *et al.* 2016). The Census then classifies all residents of urban areas as "urban" and all other population as "rural." OMB metropolitan area definitions have also evolved, but since 2003, the OMB has delineated "metropolitan statistical areas" as one of two types of CBSAs along with "micropolitan statistical areas." Each CBSA consists of a set of central counties, where a substantial population resides in the same core urban area(s), combined with any outlying counties, where a substantial proportion of workers commute to or from the central counties. To qualify as a metropolitan area, a CBSA must contain an urban area with at least 50,000 residents, while the largest urban area in a micropolitan area has between 10,000 and 50,000 residents (<https://www.census.gov/programs-surveys/metro-micro/about.html>).

² See <https://www.census.gov/programs-surveys/geography/guidance/geo-areas/pumas.html>.

delineations, so PUMAs do not consistently align with either of these systems (Figure 1). Additionally, to distinguish suburban populations, analysts sometimes use the OMB's central or principal city definitions,³ treating as “suburban” the population living in metro areas but outside of central/principal cities (*e.g.*, Mattingly and Bean 2010), but PUMAs need not align with city boundaries either.

As illustrated in Figure 1, PUMAs occasionally align with CBSAs but almost never with urban areas. The boundaries of urban areas are complex and idiosyncratic, and urban areas can also have relatively small populations (down to 2,500), so outside the cores of major urban areas, nearly all PUMAs encompass a mix of urban and rural areas. CBSA boundaries, on the other hand, always follow county boundaries, which often also form PUMA boundaries. Metro areas generally have populations larger than 100,000, enough for a single metro area to comprise one or more whole PUMAs. Likewise, central and principal cities are often large enough to comprise whole PUMAs.

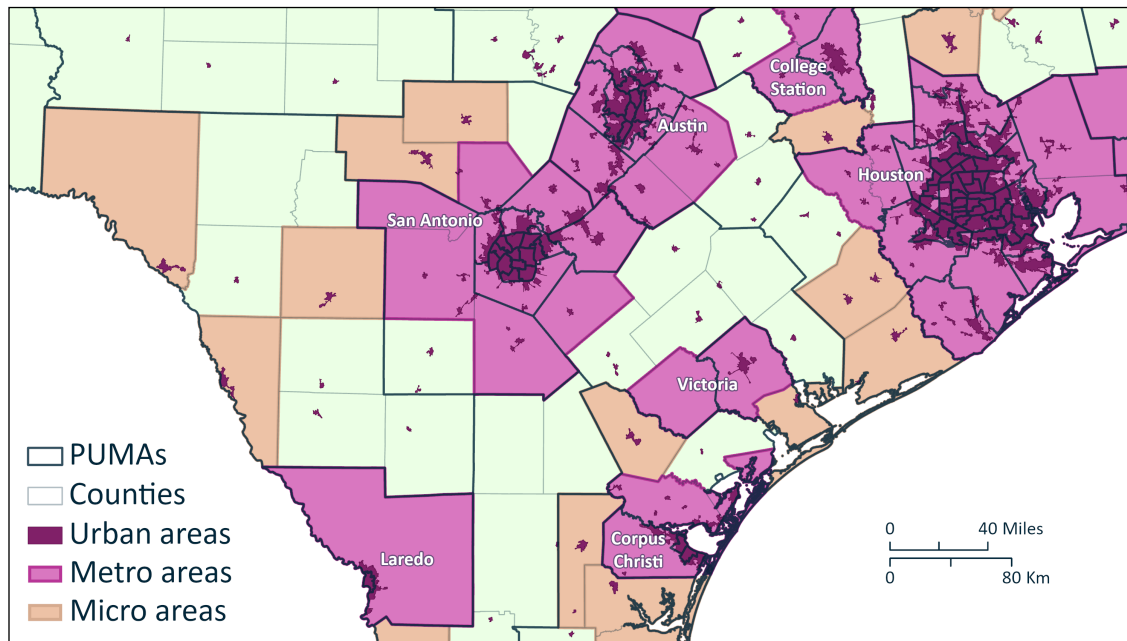


Figure 1. 2010 PUMAs, 2010 urban areas and 2013 CBSAs (metropolitan and micropolitan areas) in a section of south-central Texas.

³ Since 2003, the OMB has designated certain places within each CBSA as “principal cities,” typically the largest incorporated place within a CBSA along with other places of similar size. Prior to 2003, the OMB instead used the term “central city” to denote a similar concept.

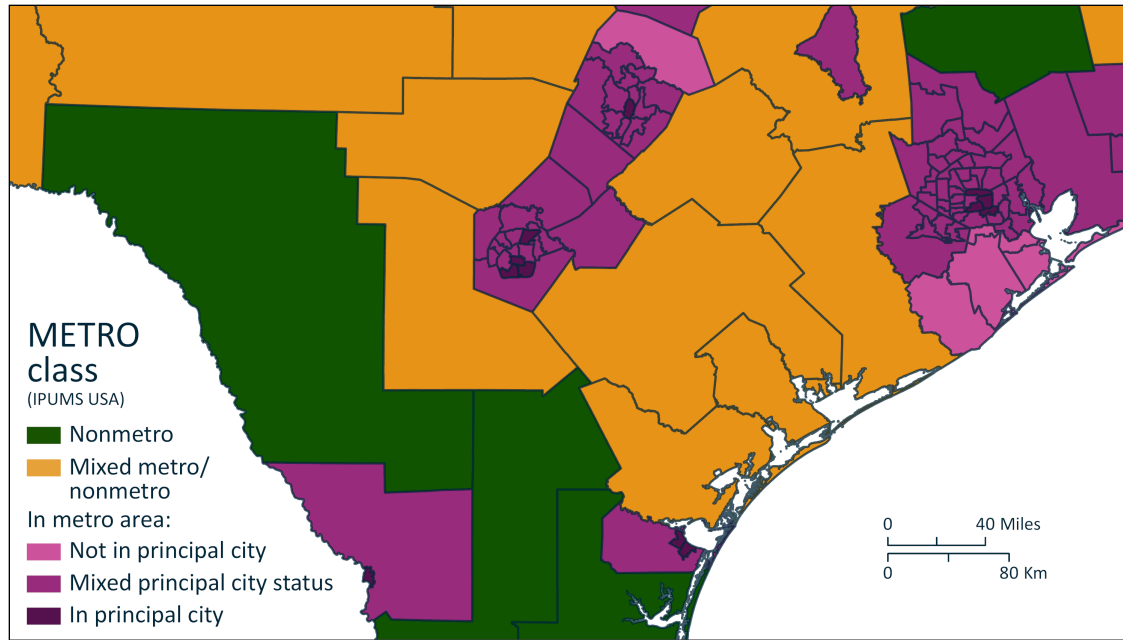


Figure 2. IPUMS-USA METRO classes for 2010 PUMAs in a section of south-central Texas.

The numerous correspondences between PUMAs and the OMB delineations make it possible to identify the exact metro/nonmetro status, and occasionally also the central/principal city status, for most PUMS records. Nevertheless, identifying a status for *all* PUMS records, as IPUMS USA does through its METRO variable, requires special handling for the many PUMAs with populations both within and outside of metro areas, or both within and outside of central/principal cities. In such cases, IPUMS USA assigns a “mixed” status, resulting in 5 distinct METRO classes, including 3 “pure” and 2 “mixed” classes (Figure 2). The Economic Research Service (ERS) of the U.S. Department of Agriculture has produced a similar classification that identifies all PUMAs as either metro or nonmetro, allocating each “mixed” PUMA to one of these two classes based on where the majority of PUMA residents live (U.S. Department of Agriculture 2019b). Translating the standard OMB classes into microdata, using the approach of either IPUMS USA or the ERS, offers the benefits of familiarity and conceptual consistency with many other applications that use the OMB definitions. This framework, however, unavoidably yields inexact class identifications because of the many discrepancies between PUMAs and OMB delineations.

An important related problem is that both the PUMA and OMB delineations change over time, as does the correspondence between them, which impairs studies of demographic change by metro status.

For example, when the Census changed the PUMAs identified in ACS PUMS files from the 2000 to 2010 definitions, IPUMS USA also changed which metropolitan definitions it used as the basis of the METRO variable, switching from the 1999 to the 2013 OMB definitions. In effect, the portion of population having a mixed metro/nonmetro status doubled from 7% for the 2011 ACS (using 2000 PUMAs) to 14% for the 2012 ACS (using 2010 PUMAs), and the portion with a mixed central/principal city status grew from 30% to 41%. These shifts are much larger than typical annual changes, indicating that they are mainly artifacts of definition changes, which calls into question any measured changes within metro classes over this period.

Even if the correspondence between PUMAs and OMB delineations were exact and persistent, there remain two major conceptual problems for any analyses that rely exclusively on the OMB classifications. First, subdividing the full range of U.S. settlement patterns into only a few classes is imprecise, potentially masking important variations within each class and separating similar locations into distinct classes (Waldorf 2006). For example, the largest U.S. metro areas have a hundred times more residents than the smallest, and socio-economic conditions may vary enormously across this spectrum. A single “metro” class nevertheless groups all these areas together.

A second limitation of the metro/nonmetro classification—especially when analysts use it alone to distinguish “urban” and “rural” populations—is that it indexes only one dimension of settlement patterns: the size of the commuting system. There are other important dimensions of difference between urban and rural places, such as population density and remoteness, which are correlated with but distinct from metro status (Coombes and Raybould 2001, Isserman 2005, Wang *et al.* 2011). Large portions of metro areas have rural densities, and some metro areas are relatively remote, etc. A phenomenon of interest could be associated with each of these dimensions in different ways, but that is impossible to determine using only the standard OMB classes.

The Census Bureau’s urban/rural definitions emphasize a second dimension of settlement patterns, subdividing the country based mainly on local population densities. (The Census also uses population size to distinguish urban areas, but the threshold of 2,500 is much lower than the core size of

50,000 required for metro areas.) These two standard ways of distinguishing “urban” and “rural” populations can lead to very different conclusions. For example, using the OMB classes, nonmetro areas had higher poverty rates than metro areas in 2017, but using the official urban/rural classification, the relationship is reversed; the rural areas had the lowest poverty rates overall (Figure 3). Clearly, drawing robust conclusions about demographic variations across

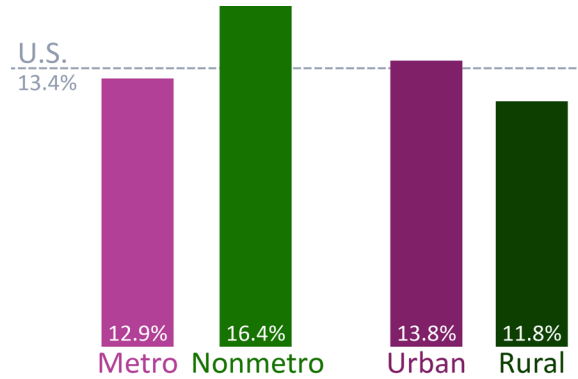


Figure 3. Poverty rates using standard metropolitan/non-metropolitan and urban/rural classifications. *2017 American Community Survey 1-Year Summary File, retrieved from IPUMS NHGIS (Manson et al. 2018).*

the rural-urban spectrum—and making sense of discrepancies like those depicted in Figure 3—demands a more sophisticated conceptual framework than is supported by either of these standard classifications alone.

Toward this end, the ERS has produced several alternative classifications (U.S. Department of Agriculture 2019b), including the Rural-Urban Continuum Codes (distinguishing 9 classes of counties), Urban-Influence Codes (distinguishing 12 classes of counties), Rural-Urban Continuum Areas (distinguishing 10 primary and 21 secondary classes of census tracts), and Frontier and Remote Area Codes (distinguishing 4 levels of remoteness among ZIP Code areas). These schemes offer more granularity than the standard OMB and Census classifications, and they distinguish more than one dimension of variation. However, in accord with ERS’s focus on agricultural economics, their classifications mainly differentiate types of rural areas and only minimally distinguish higher levels of urbanization. Also, aside from ERS’s metro/nonmetro classification of PUMAs, none of their classifications are PUMA-based, so they are not directly applicable for public-use microdata.

Two PUMA-Level Indices

When developing new rural-urban indices for IPUMS USA, we have sought out options that not only could be computed at the PUMA level consistently across time but that also varied continuously across a full spectrum of rural and urban settlement types. A secondary consideration was to select measures that are relatively easy to compute and to extend forward when integrating new PUMS files into IPUMS USA. The two newly added indices satisfy all these aims.

Two-Dimensional Conceptual Framework

Conceptually, the two new indices correspond to two basic dimensions of settlement patterns: concentration, ranging from sparse to dense, and size, ranging from small to large (Figure 4). In common usage, the terms “rural” and “urban” indicate variation in both dimensions: rural places are more sparsely settled and have smaller populations (bottom-left quadrant in Figure 4); urban places are more densely settled and have larger populations (upper-right quadrant). But places may be distinctly “rural” or “urban” along one dimension and not the other. An isolated town center (left-hand side) may be somewhat urban in its concentration level but decidedly rural in population size. Conversely, exurban large-lot developments (lower-right quadrant) may have rural levels of concentration but urban levels of access to amenities and services due to the large population of a nearby city. We expect the upper-left corner to be empty because the highest levels of concentration can occur only where there are ample populations.

Our choices for which two dimensions to emphasize and how to measure them are inspired by the two standard federal classification systems. To produce the official urban/rural classifications, the Census Bureau relies principally on local residential population density (Ratcliffe et al. 2016). This corresponds to classification

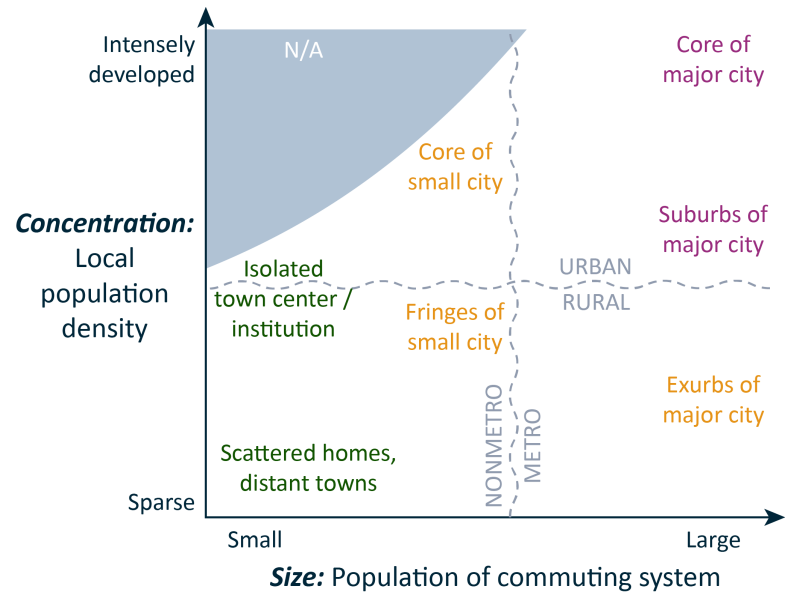


Figure 4. Conceptual model of two continuous dimensions of settlement patterns.

along the concentration dimension, as indicated by the horizontal dividing line in Figure 4. To delineate CBSAs, the OMB guidelines distinguish metro and micro areas based on the population of the urban core of each commuting system. This in turn corresponds (roughly) to classification along the size dimension, as indicated by the vertical dividing line in Figure 4.

The model also accords with other previously developed frameworks. Isserman (2005) and Wang et al. (2011) both argue—and demonstrate through case studies of poverty rates—that the two standard classifications can and should be treated as distinct, complementary indicators of urbanicity and rurality, though neither of these research efforts developed continuous indices. Coombes and Raybould (2001) suggest continuous measures for three dimensions of settlement patterns: settlement size (from hamlet to metropolitan), concentration (from sparse to dense), and accessibility (from remote to central). Our model re-uses their concentration dimension directly, and our second dimension of size corresponds with both the settlement size dimension and, to a lesser extent, the accessibility dimension.⁴

⁴ In Coombes and Raybould’s model, *settlement size* is associated with the size of an urban area (a concentrated settlement) and *accessibility* is associated with proximity to large settlements. In our model, “size” is associated with the size of an entire commuting system, which includes all areas with substantial commuting interactions with a central core, encompassing both urban areas and lower-density areas that are “accessible” to the core.

Index of Concentration

The name of the new IPUMS USA variable that indexes concentration is DENSITY, and the specific measure it reports is the *population-weighted geometric mean of census tract population densities in each PUMA*. In the initial release, DENSITY is available for 2000 census samples (using 2000 tract densities) and for the 2010 census and ACS samples (using 2010 tract densities).

We choose to use a population-weighted average density rather than the density of the whole PUMA because the latter (the PUMA's population divided by its area) is often a weak indicator of the *local* densities where PUMA residents live. Many PUMAs span across both densely and sparsely settled areas, as demonstrated by the varying densities of tracts within PUMAs (Figure 5a). The density of each whole PUMA (Figure 5b) is effectively an “area-weighted average” of these varying densities (Craig 1984), representing the typical density across all subdivisions of the PUMA rather than among the residents of the PUMA. Because PUMA residents are (by definition) more concentrated in the denser parts of a PUMA, their local densities are generally higher (and never lower) than the entire PUMA's density.

To summarize local concentrations throughout a large area, a better strategy is to compute densities in smaller “local” units, such as census tracts, and then compute the average of these densities, weighted by the local units' populations, so each resident's local density is given equal weight. The right-hand panels of Figure 5 illustrate the outcomes of measuring the population-weighted average of tract densities using the arithmetic mean (5c) and the geometric mean (5d). Following the notation of Craig (1984), the population-weighted arithmetic mean density is computed as:

$$d_{AM} = \frac{\sum P_i d_i}{\sum P_i} \quad (1)$$

where P_i and d_i are the population and density of subdivision i (in our case, a tract). The population-weighted geometric mean is computed as:

$$d_{GM} = \prod d_i^{\alpha_i} \quad (2)$$

where α_i is the proportion of the containing unit's (the PUMA's) population living in subdivision i . It can be helpful to think of the geometric mean density as measuring the average density *on a logarithmic scale*, which recasts Equation 2 into this form:

$$d_{GM} = \exp\left(\frac{\sum P_i \log d_i}{\sum P_i}\right) \quad (3)$$

In practice, our computations deviate somewhat from these equations where a PUMA boundary subdivides a tract.⁵ In such cases, we use the whole tract's density, but we limit the population weight to the portion that also resides in the PUMA.

⁵ Although 2010 census tracts nest exactly within 2010 PUMAs, not all 2000 census tracts nest within 2000 PUMAs. Also, the 2005-2011 ACS PUMS files use 2000 PUMA definitions, but DENSITY summarizes 2010 tract densities for those samples, so it is necessary to associate 2010 tracts with the 2000 PUMAs for those samples.

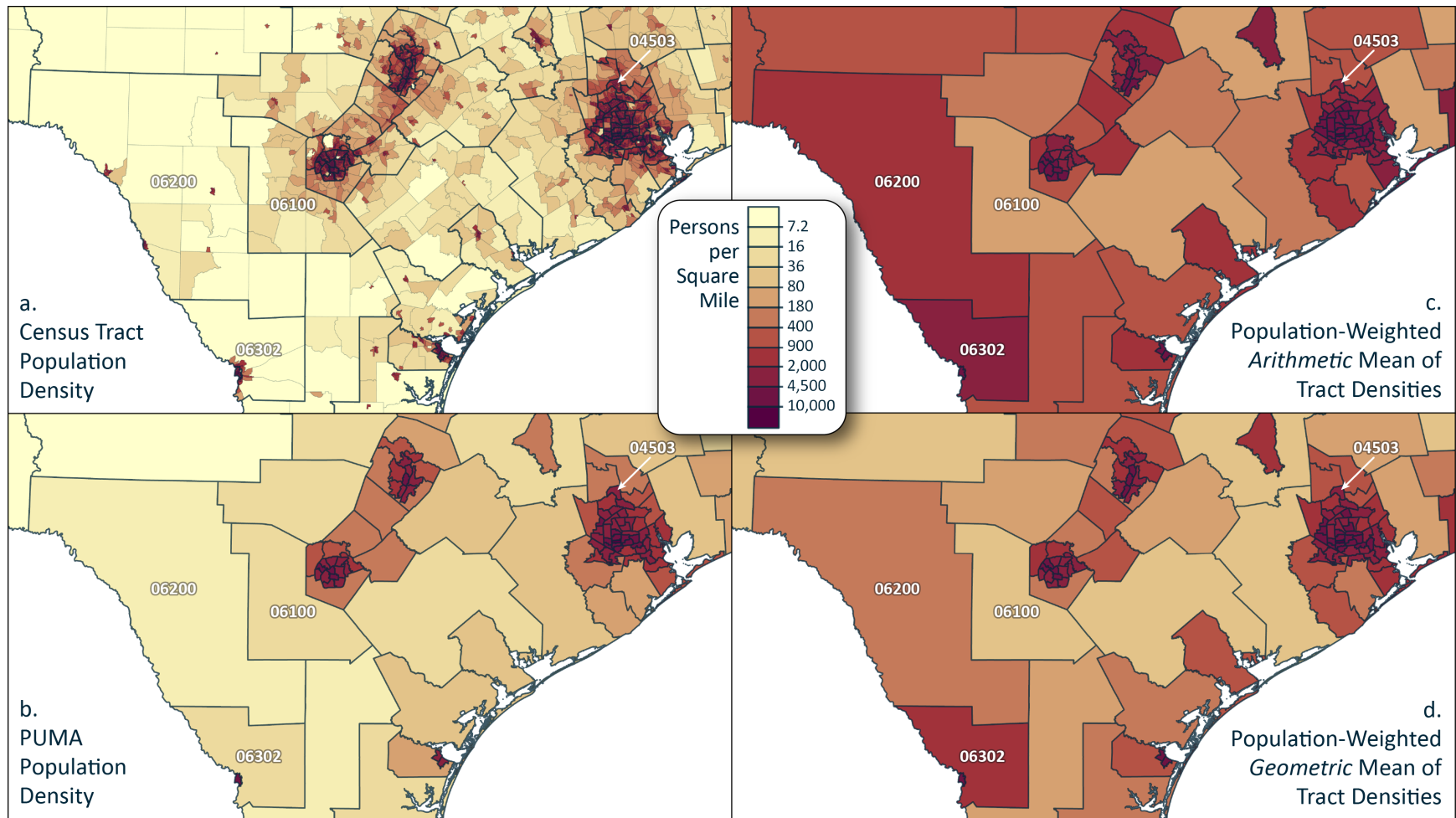


Figure 5. Four measures of 2010 population density within PUMAs in a section of south-central Texas.

Some prior applications of population-weighted average tract densities have used arithmetic means (e.g., Wilson *et al.* 2012, Kolko 2016), but we agree with others (Craig 1984, Dorling and Atkins 1995) that a geometric mean is more suitable. Densities generally have a log-normal distribution, heavily concentrated at the lower end of the distribution with a long positive tail. For such distributions, the geometric mean is appropriately less sensitive to large outliers, more sensitive to variations among small values, and typically closer to the median than is the arithmetic mean. In practical terms, a logarithmic scaling makes sense because a difference between densities of 10 and 100 is about as significant for the character of a place as any other factor-of-10 difference (e.g., 1,000 and 10,000), and it is clearly more significant than an equal absolute difference of 90 at high densities (e.g., 10,010 and 10,100).

Figure 6 illustrates how four PUMA-level density measures relate to the distribution of tract densities in the labeled PUMAs on Figure 5. The first PUMA, 04503, is roughly coincident with The Woodlands, a suburb of Houston. In this case, there is relatively little variation in densities among the tracts in the PUMA, so all four measures (the PUMA density, the population-weighted arithmetic mean, the population-weighted geometric mean, and the population-weighted median) are close to each other on both a linear and log scale, ranging only from 2,061 to 2,290 persons per square mile.

In PUMA 06100, which encompasses lower-density exurbs, small cities, and rangeland southwest of San Antonio, the tract densities vary less than PUMA 04503's on a linear scale but more than 04503's on a log scale. PUMA 06100's four density measures therefore bunch closely together on a linear scale but differ substantially on a log scale. As expected for a log-normal distribution, the median (51) and the geometric mean (48) are similar on either scale, but the arithmetic mean (84) is 65% greater than the median, and all three population-weighted densities are well above the whole PUMA's density (29).

In PUMA 06200, the tract densities have a relatively wide distribution on either scale, and on the log scale, the distribution is clearly bimodal, split between a set of large, sparse tracts and a set of small, dense tracts. The whole PUMA's density (12) lies within the lower cluster of tract densities, which is a good indication of the large expanses of sparsely populated rangeland in the PUMA, but it poorly

represents the much higher local densities of most PUMA residents. The arithmetic mean (1,436) and median

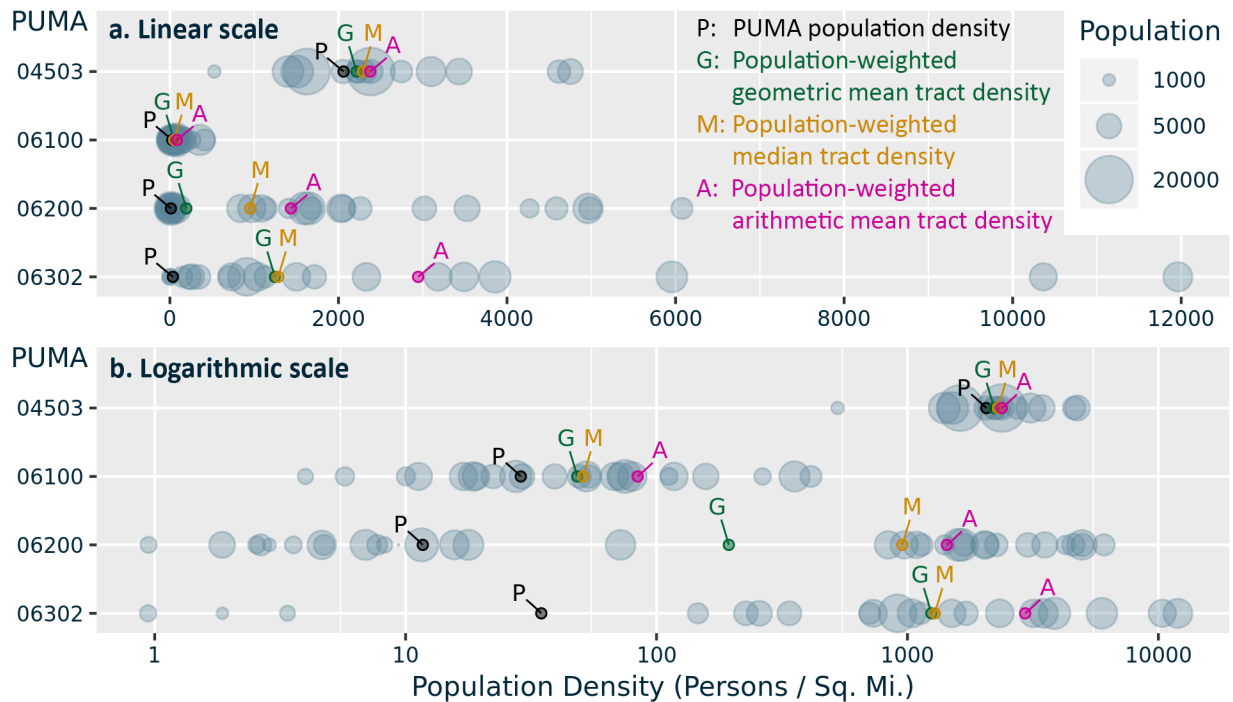


Figure 6. Distributions of 2010 census tract population densities in four Texas PUMAs, plotted on a linear scale and logarithmic scale, along with four PUMA-level density measures.

(953) are both much higher, lying in the upper cluster of tract densities, but this in turn poorly represents the sparse local densities of many PUMA residents. The geometric mean (194) is located between the two modes, suitably splitting the difference. (Of course, no single statistic can represent well the “typical” value of a bimodal distribution, but if a single statistic must be selected, something that lies between the two modal clusters seems most appropriate.)

The last of the four example PUMAs, 06302, has the widest distribution of tract densities. While the PUMA’s population is concentrated in dense tracts around Laredo, most of the PUMA’s area lies in three very sparse outlying tracts. This results in a PUMA density (35) that is much lower than the median (1,285) and geometric mean (1,245). At the other end, a few tract densities that are exceptionally high on a linear scale (but not on a log scale) result in an arithmetic mean (2,945) that is more than double the median and geometric mean.

These four example PUMAs indicate well the variety of density distributions across all PUMAs, and in each example, we find the population-weighted geometric mean of tract densities to be as good as or better than the other density measures as a general index of concentration. There are still more measures that could be considered, and there is one issue in particular that is of concern: a census tract is only one arbitrary approximation of a person's local context (Fowler et al. 2019), and averages of tract densities are subject to the Modifiable Area Unit Problem, or MAUP (Openshaw and Taylor 1981). For example, two PUMAs with identical population distributions could have very different mean tract densities depending on how the tract boundaries are drawn. One measure that would be less sensitive to the MAUP would be an inverse-distance-weighted average of block-level densities in a moving window around each census block. This is roughly the approach that Coombes and Raybould (2001) propose for an index of concentration. We have opted to rely on tract densities (for now) because it simplifies the computation and description of the index, because measures of population-weighted density are often based on tracts, and because we suspect its liabilities relative to a more robust measure are not important for most applications.

Index of Size

For an index of size, we use the *population-weighted geometric mean of the populations of CBSAs (metro and micro areas) in each PUMA*. The general aim is to summarize the typical population size of the commuting systems where PUMA residents live. Where a PUMA lies entirely within a single metro area, as is the case for 78% of 2010 PUMAs, this measure simply equates to the metro area's population. Elsewhere, the measure summarize the sizes of all CBSAs where PUMA residents live. For areas outside of any CBSA, i.e., “noncore” counties, we use the county's population as an approximation of the commuting system size.⁶

⁶ For Virginia “independent cities” that lie outside of CBSAs, we combine the populations of the independent cities with the populations of their neighboring counties.

We refer to this index as METPOP, and currently, IPUMS USA provides two versions of the index through two variables: METPOP10 summarizes the 2010 populations of 2013 CBSAs and noncore counties, and METPOP00 summarizes the 2000 populations of 2003 CBSAs and noncore counties.

Figure 7 illustrates how 2010 PUMAs correspond to the CBSA and county populations that METPOP10 summarizes, and Figure 8 illustrates METPOP10 values.

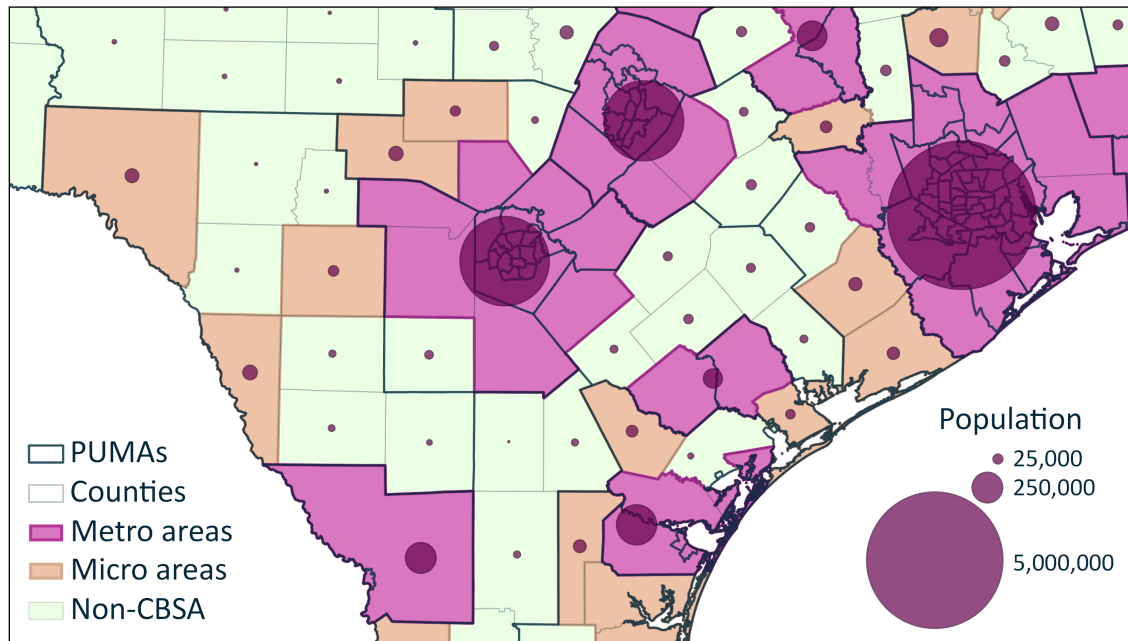


Figure 7. 2010 PUMAs, 2010 urban areas and 2013 CBSAs (metropolitan and micropolitan areas) in a section of south-central Texas.

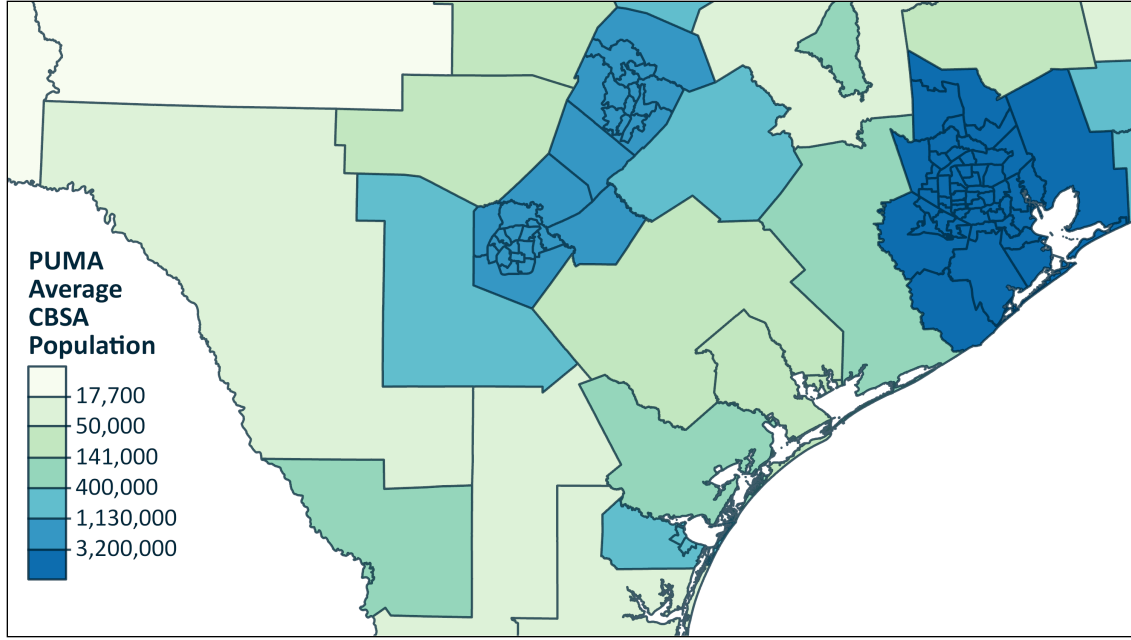


Figure 8. METPOP10 values for 2010 PUMAs in a section of south-central Texas.

The formula we use to compute the METPOP index mirrors Equation 3:

$$P_{GMj} = \exp\left(\frac{\sum P_{ij} \log P_i}{\sum P_{ij}}\right) \quad (4)$$

where P_{GMj} is the population-weighted geometric mean of the populations of CBSAs and noncore counties in PUMA j , P_i is the population of CBSA or noncore county i , and P_{ij} is the population in the area of intersection between i and j . We again use a geometric mean because commuting system populations, like tract densities, have a roughly log-normal distribution, and relative differences in populations are more important than absolute differences. For example, a difference between populations of 100,000 and 500,000 is about as significant for the character of a commuting system as any other factor-of-5 difference (e.g., 1 million and 5 million), and it is clearly more significant than an equal absolute difference of 400,000 in large commuting systems (e.g., 10.1 million and 10.5 million).

Like the DENSITY index, METPOP may also be impaired by an inexact spatial basis. The extents of “true” commuting systems need not correspond well with counties, and this a limitation not only where METPOP is based on noncore counties but even where it is based on CBSAs. For example, because of the great extents of its component counties, the Riverside–San Bernardino–Ontario CBSA in

California includes the small city of Needles, a 220-mile drive from the CBSA's largest city, Riverside. The PUMA that contains Needles is comprised mainly of desert, and the fraction of its residents who commute to the CBSA's urban core is likely small, but the PUMA nevertheless has a high METPOP value. A more effective index of size might delineate commuting systems based on tracts rather than counties, or it might use larger ERS commuting zones (U.S. Department of Agriculture 2019a, Fowler *et al.* 2018), a system which allocates every county to a zone, eliminating the problem of "noncore" counties. Alternatively, as with the index of concentration, the most effective approach may be to use a moving window, but instead of using a "local" moving-window average of densities, the index of size would use a larger "regional" moving-window summing populations within a typical commuting distance. We leave these possibilities for future research.

Pairing the Indices

Figure 9 illustrates the two-dimensional spread of average tract densities (DENSITY) and average CBSA populations (METPOP10) for all 2010 PUMAs. The point colors also indicate the metro/nonmetro class of each PUMA. The overall distribution mirrors closely the conceptual model in Figure 4: the upper right contains PUMAs with high densities in large metro areas; the lower right contains PUMAs with low densities in large metro areas; the lower left contains PUMAs with low densities and outside (or mostly outside) of any CBSA; and as expected, the upper left is empty, indicating that PUMAs with dense populations occur only in or around medium-to-large CBSAs.

The colors in Figure 9 indicate that most PUMAs that lie entirely within metro areas have relatively high average densities, but some have low average densities. Such low-density metro PUMAs may or may not fit our expectations for "rural" areas. They may or may not share characteristics with other

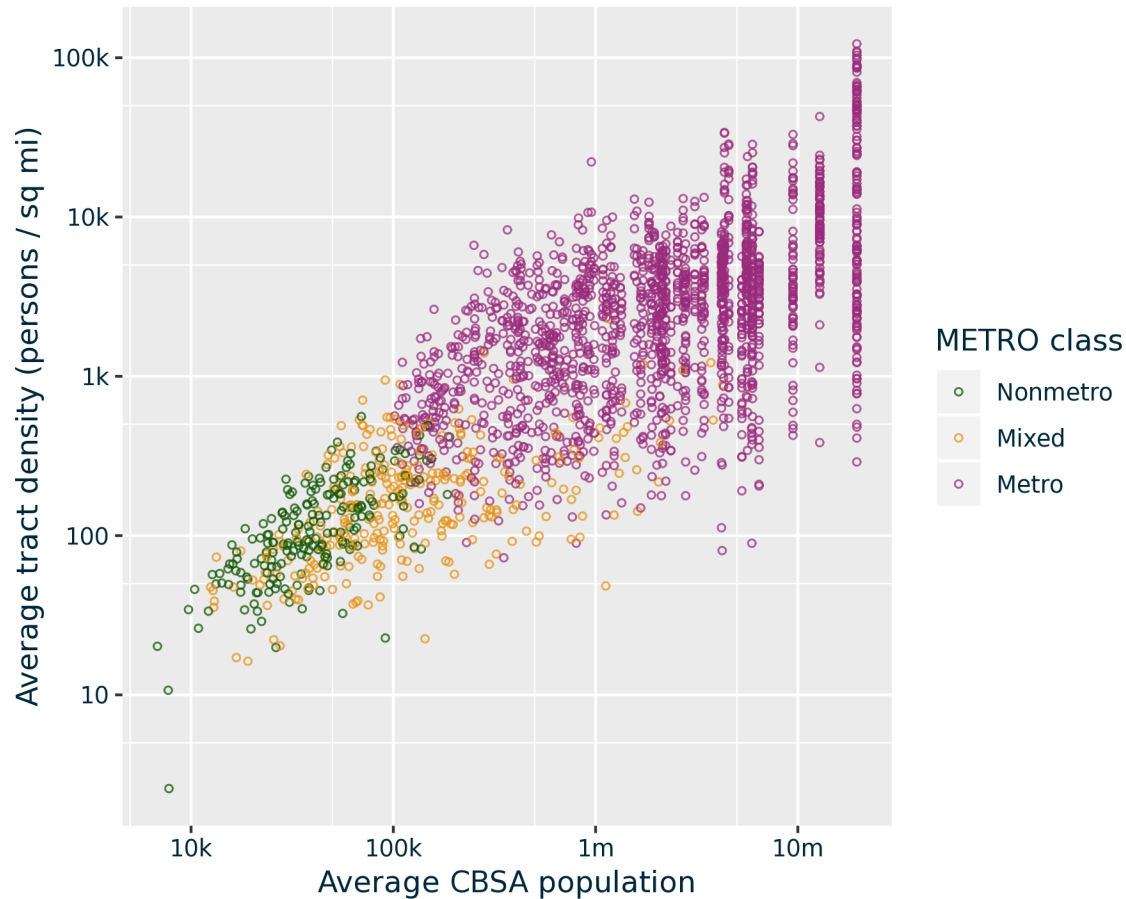


Figure 9. Relationships among three urbanization indices for 2010 PUMAs.

low-density PUMAs. Similarly, some nonmetro and mixed PUMAs have moderately high densities. It is possible that such areas have more in common with metro PUMAs at similar densities than with nonmetro and mixed PUMAs at lower densities. We believe that this two-dimensional framework offers great potential as a means to investigate such possibilities and to determine whether “concentration” or “size” are important factors, separately or together, in any study of urban-rural discrepancies. Because the indices are continuous measures, the framework should also be useful for distinguishing fine gradations of variation or for identifying inflection points across all densities or across all levels of the urban hierarchy.

Illustrative Results

As an illustration of the utility of these two measures, we analyze how poverty rates vary across settlement types in the U.S. We frame our analysis in the following context. To date, analyses of "rural" poverty often use the "metro/nonmetro" and "urban/rural" terminology interchangeably. This practice is problematic. As Figure 3 shows, under the metro/nonmetro classification, nonmetro areas have higher poverty rates than metro areas, but under the urban/rural classification, the relationship is reversed; the "rural" areas have the lowest poverty rates overall. Ultimately, this contradiction in poverty rates highlights the limitation our measures overcome: researchers are forced to adopt coarse dichotomous measures because of the absence of continuous measures. The availability of these two continuous measures allows researchers to better analyze differences across geographic regions beyond the coarse metro/nonmetro classification.

In this section, using 2012-2017 ACS microdata from IPUMS USA, including the new indices, we illustrate how poverty rates, and the likelihood of being in poverty relative to other demographic characteristics, varies across levels of urbanization as distinguished by both concentration (DENSITY) and by population size (METPOP10). Using microdata with the new indices enables us to fit regression models predicting poverty while controlling for other demographic factors. The use of our continuous measures shows that the correlation between poverty, rurality, and different demographic dimensions, varies in ways that cannot be captured by a simple metro/nonmetro distinction.

Table 1. Poverty rates by METRO category, 2012-2017

	Nonmetro	Mixed	Metro
Percent in poverty	17.6	15.9	14.5

IPUMS USA ACS samples, N = 18,120,063

We begin our analysis with the geographic categories currently available in IPUMS USA. Table 1 shows that PUMAs classified as wholly nonmetro (neither metro nor mixed) had poverty rates between

2012 and 2017 that were higher than in other PUMAs. The “mixed” PUMAs, those that straddle both metro and nonmetro areas, had a poverty rate between nonmetro and metro PUMAs’.

Figure 10 shows the heterogeneity in poverty rates across geographic regions that is uncovered by our two continuous indices. Each point represents a PUMA while the color represents levels of poverty: blue represents lower levels of poverty and red represents higher levels of poverty. Figure 10 shows a



Figure 10. Relationship between poverty and two urbanization indices of urbanization for 2010 PUMAs. IPUMS USA 2012-2017 ACS samples.

clear nonlinearity in poverty rates. We see that both the highest and lowest poverty rates are found in large commuting systems. Table 1 similarly showed that metro areas had lower rates of poverty relative to nonmetro areas, but by using continuous indices, we can see that the metro/nonmetro dichotomy masks significant differences in poverty within metro areas. Moreover, Figure 10 also shows the substantial dispersion of poverty rates among PUMAs at a given level of density or commuting system size.

For further analysis, we classify PUMAs along both dimensions with four levels of METPOP10 values (breaks at 50,000, 400,000, and 3.2 million) and five levels of DENSITY values (breaks at 80, 400, 2,000, and 10,000 persons per square mile). To avoid having a class represented by only one PUMA, we drop the lowest DENSITY break (at 80) for PUMAs with large commuting system sizes (above 400,000). This produces 14 classes of PUMAs, each with unique ranges of METPOP10 and DENSITY values. Table 2a shows how poverty rates vary among these classes, with the highest rate of poverty (22.8%) found in the densest PUMAs with moderately large CBSA populations (between 400,000 and 3.2

Table 2a. Poverty rates by DENSITY and METPOP10, 2012-2017

DENSITY (per sq mi)	METPOP10			
	0-50k	50k-400k	400k-3.2m	3.2m+
10,000+			22.8	20.5
2,000-10,000		20.7	16.6	12.6
400-2,000		15.6	11.6	9.2
80-400	17.7	16.8	13.6	12.5
0-80	17.5	18.3		

IPUMS USA ACS samples, N = 18,120,063

Table 2b. Increase in poverty rate by DENSITY and METPOP10 over lowest-poverty class (METPOP10 of 3.2m+, DENSITY of 400-2,000), 2012-2017

DENSITY (per sq mi)	METPOP10			
	0-50k	50k-400k	400k-3.2m	3.2m+
10,000+			13.6	11.3
2,000-10,000		11.5	7.4	3.4
400-2,000		6.4	2.4	0.0
80-400	8.5	7.6	4.3	3.3
0-80	8.2	9.1		

IPUMS USA ACS samples, N = 18,120,063

million people). The lowest poverty rates are found in areas with medium density in the largest CBSAs (populations of over 3.2 million). Table 2b converts the poverty rates from Table 2a to be poverty rates of a given category relative to the area of lowest poverty rates, which we use as a benchmark for analyzing poverty rates in a multiple regression framework.

Limiting our analysis to the wholly nonmetro PUMAs (Table 3) demonstrates that the new indices are useful for distinguishing different types of rural population within these more remote areas. The poverty rates for the least dense and most dense areas with larger commuting system populations (50k-400k) have the highest and lowest poverty rates (12.8% and 25.9%). This shows again that a coarse classification along only one dimension of urbanization can mask substantial disparities within classes or along other dimensions.

Table 3. Poverty rates by DENSITY and METPOP10 for residents of wholly nonmetro PUMAs, 2012-2017

DENSITY (per sq mi)	METPOP10	
	0-50k	50k-400k
400-2,000		12.8
80-400	17.8	17.6
0-80	17.5	25.9

IPUMS USA ACS samples, N = 1,910,022

A key motivation for producing PUMA-level indices is to enable microdata-based analyses that are impossible using existing census summary data. In Table 4, we provide summary statistics for the nonmetro areas as well as the more refined categories for nonmetro areas. In doing so, we illustrate how our measures can uncover important variation. Focusing on demographic variation first, we notice that nonmetro areas have an average age of about 40 years. However, within nonmetro areas, the largest and most dense PUMAs are, on average, about two years older, proportionally more White (80.4% v. 89.3%), and have higher incomes, wages and salaries, and Social Security income. Internet access is an important issues in rural areas. Indeed, we show that about 16.6% of people in nonmetro areas have no access to internet. But by looking at the least dense and least populated areas (in particular the 50k-400k

Table 4. Demographics of wholly nonmetro PUMAs by DENSITY and METPOP10, 2012-2017

DENSITY (per sq mi), METPOP10					
All Non-metro	0-80, 50k-400k		80-400, 50k-400k		400-2,000, 50k-400k
	0-80, 0-50k	50k-400k	80-400, 0-50k	50k-400k	

Demographics						
Age	40.2	40.3	39.4	40.0	40.3	41.9
Female (%)	50.6	50.3	51.1	50.7	50.8	50.5
Hispanic/Latino (%)	7.8	7.7	5.8	8.2	7.6	5.5
White (%)	80.4	80.6	55.3	80.7	80.3	89.3
African American (%)	6.7	5.6	8.5	8.1	6.7	2.5
Native American (%)	2.2	3.4	28.2	0.7	1.4	0.2
Asian (%)	1.0	0.8	0.5	0.6	1.6	1.0
Other Race (%)	0.1	0.1	0.1	0.1	0.1	0.1
Multiple Races (%)	1.9	1.7	1.7	1.7	2.3	1.5
Foreign-Born (%)	4.3	4.1	3.7	3.7	5.0	4.6
Noncitizen (%)	2.2	2.2	1.5	1.9	2.6	1.6
Income						
Total Personal Income	29,244	29,602	23,243	28,298	29,697	35,136
Wage and Salary Income	19,822	19,727	15,441	19,238	20,302	25,134
Social Security Income	3,295	3,292	3,059	3,299	3,291	3,479
Other						
No Access to Internet (%)	16.6	17.0	27.9	17.3	15.3	11.3
No Health Insurance (%)	11.7	12.0	16.1	12.0	11.2	6.8
N	1,910,022	626,153	36,383	619,333	583,797	44,356
IPUMS USA ACS samples						

area), we notice that this rate is substantially higher (27.9%). Meanwhile, the most dense and most populated nonmetro areas have much lower rates of internet inaccessibility (11.3%). This exercise highlights the variation that is noticeable within the nonmetro PUMAs.

These demographic differences highlight the need to control for demographic factors when analyzing poverty, a practice that is impossible without microdata. Table 5 makes the obvious point that controlling for demographics is necessary in a regression modeling poverty. The first linear probability model predicts poverty by controlling for residing in metro PUMAs or mixed PUMAs:

$$poverty = \beta_0 + \beta_1 metro + \beta_2 mixed. \quad (5)$$

The coefficient shows that people in metro PUMAs are less likely to be in poverty than those in nonmetro areas. Indeed, people residing in metro areas are about 3.1 percent points less likely to be in poverty than those living in nonmetro areas (this can be seen in Table 1 where the mean difference between nonmetro

and metro poverty rates is $17.6 - 14.5 = 3.1$). However, when we include a battery of demographic controls in the second regression, we notice a significant decreases in the coefficient on metro. We run the following linear probability model:

$$poverty = \beta_0 + \beta_1 metro + \beta_2 mixed + \gamma Demographics, \quad (6)$$

where γ is a vector of demographic controls. Specifically, γ includes controls on age, sex, race, Hispanic/Latino, citizenship, foreign-born, marital status, health care coverage, educational attainment, employment status & sector, year, and census division. After including these controls, the coefficient on metro decreases from -0.0311 to -0.0399, meaning that the expected probability of a person living in a metro area is about 4 percentage points lower than in nonmetro areas, all else constant. Clearly, in order to model predictors of poverty, it is necessary to use microdata, a hopefully obvious yet important point.

Table 5. Regressions coefficient for metro areas predicting likelihood of being in poverty, with and without controls, 2012-2017

METRO class	No controls	With demographic controls
Metro	-0.0311***	-0.0399***
Mixed	-0.0171***	-0.0137***

Omitted class is Nonmetro. IPUMS USA ACS samples, N = 18,120,063.

We next incorporate our continuous measures to our regression framework. Specifically, we estimate the following linear probability model:

$$poverty = \beta_0 + \delta DENSITYxMETPOP10 + \gamma Demographics, \quad (7)$$

where γ is the same vector of controls as in Equation 6 and δ are indicators for each DENSITY by METPOP10 category where the omitted geography is the area with the lowest poverty rate (400-2,000 density, 3.2 million plus population). Table 6 maintains is color-coded on the same scale as Table 2b and shows how the overall poverty rates difference are lower once controlling for demographics. As seen in Table 2b, the area with the highest poverty rate was 13.6%, relative to the areas with the lowest poverty rate. After controlling for demographics, these areas have poverty rates, on average, 8.3 percentage points

higher than the area of lowest poverty. In other words, about 60% of the raw difference (8.3/14.6) in poverty rates can be explained by demographic difference across geographic regions. This finding is evidence of how incorporating these measures into the microdata can result in more nuanced results regarding poverty.

Table 6. Coefficients for likelihood of poverty by DENSITY and METPOP10, 2012-2017

DENSITY (per sq mi)	METPOP10			
	0-50k	50k-400k	400k-3.2m	3.2m+
10,000+			8.3	5.8
2,000-10,000		8.9	4.6	1.1
400-2,000		5.5	2.2	n/a
80-400	7.2	6.5	3.8	2.3
0-80	7.1	6.9		

Omitted class is METPOP10 of 3.2m+ and DENSITY of 400-2,000. IPUMS USA ACS samples, N = 18,120,063.

A final angle we take is to focus on the correlation between Hispanic/Latino origin, citizenship, and poverty. We run the following model

$$poverty = \beta_0 + \gamma Demographics, \quad (8)$$

separately for each of the fourteen DENSITY by METPOP10 categories where γ is the same vector of controls as in Equation 6. In Table 7a, we report the coefficient on Hispanic/Latino for each category and maintain the same color-coding scheme as in Table 6. We notice that the coefficients do not display the same pattern as the regression results in Table 6. While the highest levels of poverty are found in the densest areas (8.3 in Table 6), the Hispanic/Latino population in these areas actually have the lowest levels of poverty relative to Hispanic/Latinos in all other areas. We report the coefficients on noncitizens in Table 7b and notice a different pattern. Noncitizens have lower levels of poverty in different densities and across different CBSA populations. That is, the pattern of poverty for noncitizens does not mimic that of the Hispanic/Latino population. In all, these results demonstrate the importance of having continuous measures of settlement patterns in microdata that allow researchers to control for different demographics.

Table 7a. Coefficient on Hispanic/Latino control by subsample of DENSITY and METPOP10, 2012-2017

DENSITY (per sq mi)	METPOP10			
	0-50k	50k-400k	400k-3.2m	3.2m+
10,000+			2.1	5.7
2,000-10,000		8.1	7.3	5.7
400-2,000		6.7	8.4	7.0
80-400	6.2	7.7	9.2	7.0
0-80	5.9	5.4		

IPUMS USA ACS samples, N = 18,120,063

Table 7b. Coefficient on noncitizen control by subsample of DENSITY and METPOP10, 2012-2017

DENSITY (per sq mi)	METPOP10			
	0-50k	50k-400k	400k-3.2m	3.2m+
10,000+			2.5	3.6
2,000-10,000		6.5	3.8	3.9
400-2,000		6.6	4.7	2.9
80-400	5.6	4.7	5.9	6.1
0-80	2.6	3.7		

IPUMS USA ACS samples, N = 18,120,063

Conclusion

We introduce two PUMA-based indices—average tract population density (IPUMS variable: DENSITY) and average metro/micro-area population (IPUMS variables: METPOP00 and METPOP10)—which correspond to two distinct dimensions of settlement patterns: “concentration” (the local intensity of settlement) and “size” (the total population of the commuting system). In this paper, we have detailed the construction of these variables, explained how to interpret them, and demonstrated the value of the indices by using them to distinguish a broad range of nonlinear variations in poverty rates and in demographic covariates of poverty across the rural-urban universe of settlement patterns.

References

- Coombes, M., and S. Raybould. (2001). Public policy and population distribution: developing appropriate indicators of settlement patterns. *Environment and Planning C: Government and Policy* 19(2), 223–248.
- Craig, J. (1984). Averaging population density. *Demography*, 21(3), 405-412.

- Dorling, D., and D. J. Atkins. (1995). Population density, change and concentration in Great Britain 1971, 1981 and 1991. Studies in Medical and Population Subjects No. 58. HMSO, London, UK.
- Fowler, C. S., N. Frey, D. C. Folch, N. Nagle, and S. Spielman. (2019). Who are the people in my neighborhood?: The “contextual fallacy” of measuring individual context with census geographies. *Geographical Analysis*. <https://doi.org/10.1111/gean.12192>
- Fowler, C. S., L. Jensen, and D. Rhubart. (2018). Assessing U.S. labor market delineations for containment, economic core, and wage correlation. <https://doi.org/10.17605/OSF.IO/T4HPU>
- 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.
- Kolko, J. (2016, March 30). Urban revival? Not for most Americans [Blog post]. Turner Center for Housing Innovation, UC Berkeley. <https://turnercenter.berkeley.edu/blog/urban-revival-not-for-most-americans> [Accessed 2019 December 12].
- Manson, S., J. Schroeder, D. Van Riper, and S. Ruggles. (2018). IPUMS National Historical Geographic Information System: Version 13.0 [Database]. Minneapolis, MN: IPUMS. <http://doi.org/10.18128/D050.V13.0>
- Mattingly, M.J. and J.A. Bean. (2010). The unequal distribution of child poverty: Highest rates among young Blacks and children of single mothers in rural America. Issue Brief No. 18, Carsey Institute, University of New Hampshire.
- Openshaw, S., and P.J. Taylor. (1981). The modifiable areal unit problem. In *Quantitative Geography: A British View*, 60–70, edited by N. Wrigley and R. J. Bennett. Routledge and Kegan Paul, London, UK.
- Ratcliffe M., C. Burd, K. Holder, and A. Fields. (2016). Defining rural at the U.S. Census Bureau. ACSGEO-1, U.S. Census Bureau, Washington, DC.
- U.S. Department of Agriculture. (2019a). Commuting Zones and Labor Market Areas. Economic Research Service, U.S. Department of Agriculture. Washington, DC. <https://www.ers.usda.gov/data-products/commuting-zones-and-labor-market-areas> [Accessed 2019 December 12].
- U.S. Department of Agriculture. (2019b). Rural Classifications Overview. Economic Research Service, U.S. Department of Agriculture. Washington, DC. <https://www.ers.usda.gov/topics/rural-economy-population/rural-classifications> [Accessed 2019 October 25].
- Wang, M., R. G. Kleit, J. Cover, and C. S. Fowler. (2012). Spatial variations in US poverty: Beyond metropolitan and non-metropolitan. *Urban Studies*, 49(3), 563-585.
- Waldorf, B. (2006, July). A continuous multi-dimensional measure of rurality: Moving beyond threshold measures. Paper presented at the annual meeting of the annual meeting of the American Agricultural Economics Association, Long Island, CA.
- Wilson, S. G., D. A. Plane, P. J. Mackun, T. R. Fischetti, J. Goworowska, D. Cohen, M. J. Perry, and G. W. Hatchard. (2012). Patterns of metropolitan and micropolitan population change: 2000 to 2010. 2010 Census Special Reports. U.S. Census Bureau, Washington, DC.

