



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

Jonathan P. Schroeder¹ · José D. Pacas¹

Accepted: 16 February 2021 / Published online: 15 March 2021
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2021

Abstract

Microdata from U.S. decennial censuses and the American Community Survey are a key resource for social science and policy analysis, enabling researchers to investigate relationships among all reported characteristics for individual respondents and their households. To protect privacy, the Census Bureau restricts the detail of geographic information in public use microdata, and this complicates how researchers can investigate and account for variations across levels of urbanization when analyzing microdata. One option is to focus on metropolitan status, which can be determined exactly for most microdata records and approximated for others, but a binary metro/nonmetro classification is still coarse and limited on its own, emphasizing one aspect of rural–urban variation and discounting others. To address these issues, 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 urbanization—concentration and size—and we demonstrate their utility through an examination of disparities in poverty throughout the rural–urban universe. Poverty rates vary across settlement types in nonlinear ways: 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 settlement types. Both indices are now available for recent census microdata via IPUMS USA (<https://usa.ipums.org>).

Keywords Census microdata · Urban/rural · Population density · Poverty · United States

Notice of Prior Versions: The authors have released a similar version of this manuscript through the Minnesota Population Center Working Paper Series: #2019-05, <https://doi.org/10.18128/MPC2019-05>. An earlier version appears in the 2019 conference proceedings of the Population Association of America under the title “Getting ‘Rural’ Right: Poverty Disparities Across Two Dimensions of Rurality,” including a third author, David Van Riper.

Extended author information available on the last page of the article

1 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 great 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 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 identify only 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.

We introduce here 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 levels of urbanization 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 limitations of the standard metropolitan and urban classifications, particularly for analyses of microdata, and the potential advantages of using two 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.

2 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 delineations, so PUMAs do not consistently align with either of these systems (Fig. 1). To distinguish suburban populations, analysts sometimes also 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 & Bean, 2010), but PUMAs need not align with city boundaries either.

PUMAs do occasionally align with CBSAs but almost never with urban areas (Fig. 1). The boundaries of urban areas are complex and idiosyncratic, and urban areas can also have relatively small populations (down to 2500), 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.

The numerous correspondences between PUMAs and the OMB delineations make it possible to identify the exact metro/nonmetro status for most PUMS records

¹ The Census Bureau's urban/rural classifications have evolved over time (Ratcliffe, 2015), but for the 2000 and 2010 censuses, the general procedure was 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 classified 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>.

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.

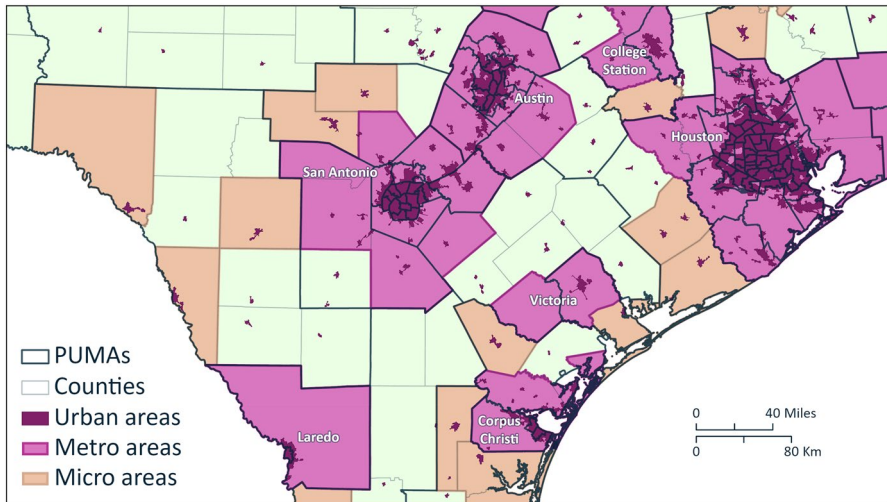


Fig. 1 2010 PUMAs, 2010 urban areas and 2013 CBSAs (metropolitan and micropolitan areas) in a section of south-central Texas

and occasionally also the central/principal city status. Nevertheless, identifying a status for all PUMAs 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 (Fig. 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).⁴ 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 as both the PUMA and OMB delineations change over time, so does the correspondence between them, which can impair 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 delineations. In

⁴ IPUMS USA recently added a variable, PCTMETRO, that gives the percentage of each PUMA’s population living in metro areas, which analysts can use to produce a binary metro/nonmetro classification like that of the ERS.

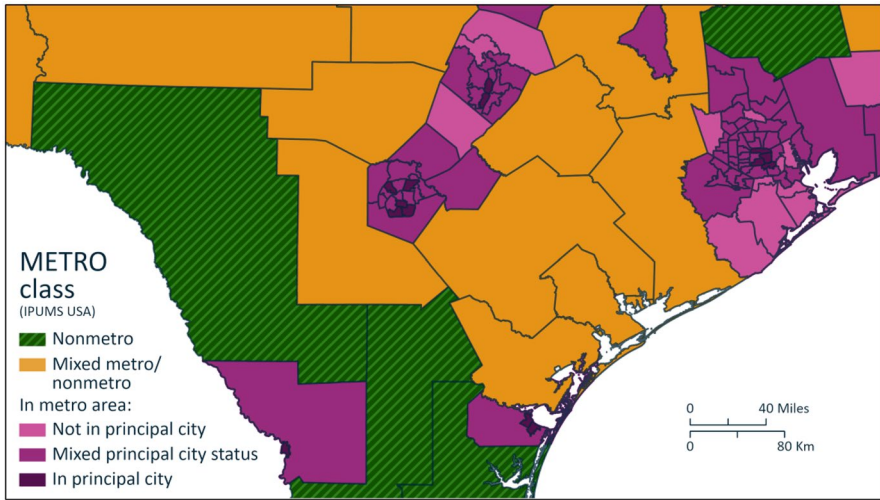


Fig. 2 IPUMS-USA METRO classes for 2010 PUMAs in a section of south-central Texas

effect, the portion of population having a mixed metro/nonmetro status grew from 7% for the 2011 ACS (using 2000 PUMAs) to 12% for the 2012 ACS (using 2010 PUMAs), and the portion with a mixed central/principal city status grew from 30 to 36%. These shifts are much larger than typical annual changes, and they both result in larger portions of population in “mixed” PUMAs, indicating that they are mainly artifacts of a weakened correspondence between PUMAs and OMB delineations. The degree to which any such changes in METRO status are meaningful is difficult to determine.

Even if the correspondence between PUMAs and OMB delineations were exact and persistent, there remain two important conceptual problems for analyses that rely exclusively on the metro/nonmetro classification. 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 emphasizes only one of the multiple dimensions of urbanization: *population size*. The delineation of metro areas does incorporate other factors; both population density and size are used to define the core urban areas of CBSAs, and commuting flows are used to determine which counties are associated with which urban cores, but after these associations are established, the single basic feature that distinguishes a metro county is that it is part of a commuting system where the urban core has a population of 50,000 or more. In other regards, a metro county may be very urban *or* very rural; it may have a high or low population density, and it may be at either the core or the

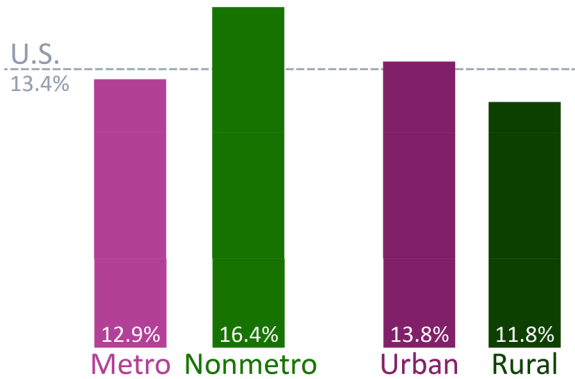
distant periphery of a commuting system. The metro/nonmetro classification alone tells us little about these other important dimensions of settlement patterns, which are correlated with but distinct from population size (Coombes & Raybould, 2001; Isserman, 2005; Wang et al., 2012). A phenomenon of interest could be associated with each of these dimensions in different ways, but that is impossible to determine using only the metro/nonmetro classes.⁵

Similarly, the Census Bureau's urban/rural classification also incorporates multiple aspects of settlement patterns while emphasizing a single dimension of urbanization—in this case, *concentration*. Unlike CBSAs, urban areas are designed to correspond closely to the extents of urbanized land use (Ratcliffe, 2015; Ratcliffe et al., 2016), and the Bureau is able to achieve this goal with a high degree of precision by building up from individual census blocks, units much smaller than the counties that form CBSAs. According to the 2000 and 2010 standards, the main criteria for a group of blocks to be identified as an urban area is that they surpass a density threshold (1000 persons per square mile at the core and 500 persons per square mile in neighboring areas, with other rules used to add some types of low-density blocks), and their combined population is at least 2500. The classification is therefore associated with both population size and density, but in combination, the two criteria serve as a singular indicator of concentration *at a certain resolution*. For example, if a few neighboring blocks in an isolated community have high densities, but their combined population is below 2500, then for a somewhat larger area, e.g., a space of 5 square miles, the density would be less than 500 persons per square mile, effectively a “rural” level of concentration at that resolution. From this perspective, each urban area represents a *substantial region* of concentrated settlement, and rural areas may contain some groups of dense blocks but never in a “large concentration.” Meanwhile, this distinction alone tells us little about other dimensions of urban/rural variation, such as the size of the commuting system or accessibility to a major city center.

The coarseness of the standard binary classifications and their distinct emphases are useful and appropriate in many contexts. Many regional disparities have a basic metro/nonmetro or urban/rural divide, and the standards used to delineate these classes have evolved through decades of refinement to correspond well with important, observable distinctions in settlement and commuting patterns. Problems can arise, however, when analysts draw general conclusions about urban/rural differences using only one of these classifications. For example, using the binary 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 (Fig. 3). In other words, poverty is higher outside of large commuting systems but also lower outside of concentrated settlements. How

⁵ The complete CBSA specifications include additional information distinguishing central and outlying counties as well as central/principal cities, but neither the IPUMS METRO variable nor the ERS metro/nonmetro classifications convey all this information, nor could they with great precision at the PUMA level. This additional CBSA information also pertains mainly to a second dimension of urbanization, *accessibility/remoteness*, and still reveals little about another key dimension, *concentration*.

Fig. 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)



could this be? Poverty may be especially high in concentrated settlements outside of large commuting systems (urban nonmetro population) or especially low in the exurbs of large commuting systems (rural metro population). Investigating such possibilities and accurately characterizing them requires a more robust analytical 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 (12 classes of counties), rural–urban continuum areas (10 primary and 21 secondary classes of census tracts), and frontier and remote area codes (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.

3 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 these aims.

3.1 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

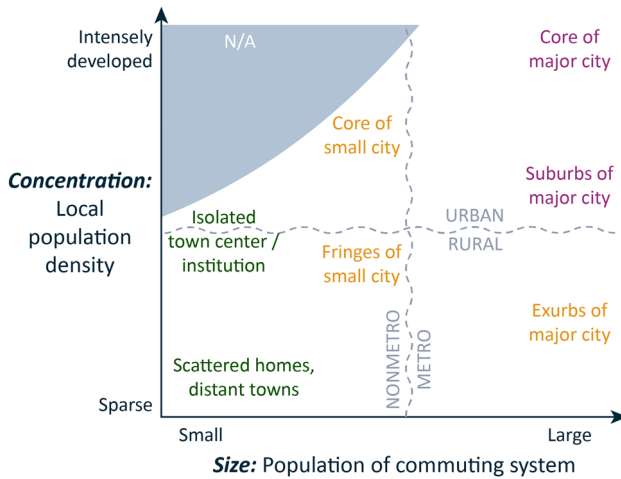


Fig. 4 Conceptual model of two continuous dimensions of settlement patterns

to large (Fig. 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 Fig. 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 (depending on the exact spatial resolution at which densities are measured).

Our choices for which two dimensions to emphasize and how to measure them are inspired by the two standard federal classification systems. As discussed in Sect. 2, the Census Bureau’s urban/rural classification emphasizes the concentration dimension, corresponding roughly to the horizontal dividing line in Fig. 4. The OMB’s metro/nonmetro classification in turn emphasizes the size dimension (specifically, the size of the urban core of a commuting system), corresponding roughly to the vertical dividing line in Fig. 4.

The model also accords with other previously developed frameworks. Isserman (2005) and Wang et al. (2012) both argue—and demonstrate through case studies—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. They instead use county-based classifications, translating the Census Bureau’s urban/rural classes to a county basis through a new system of “Rural Urban Density Codes,” which classify counties into four groups based on population density and the amount of urban population in each county. 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 corresponds with the settlement size dimension and, to a lesser extent, the accessibility dimension.⁶

3.2 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 (Fig. 5a). The density of each whole PUMA (Fig. 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, the average of their local densities is generally higher (and cannot be 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 Fig. 5 illustrate the outcomes of measuring the population-weighted average of tract densities using the arithmetic mean (c) and the geometric mean (d). Following the notation of Craig (1984), the population-weighted arithmetic mean density is computed as:

$$d_{AM} = \frac{\sum P_i d_i}{\sum P_i} \tag{1}$$

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:

$$d_{GM} = \prod d_i^{\alpha_i} \tag{2}$$

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 Eq. (2) into this form:

⁶ 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, encompassing both urban areas and lower-density areas that are “accessible” to the urban core as determined by commuting flows.

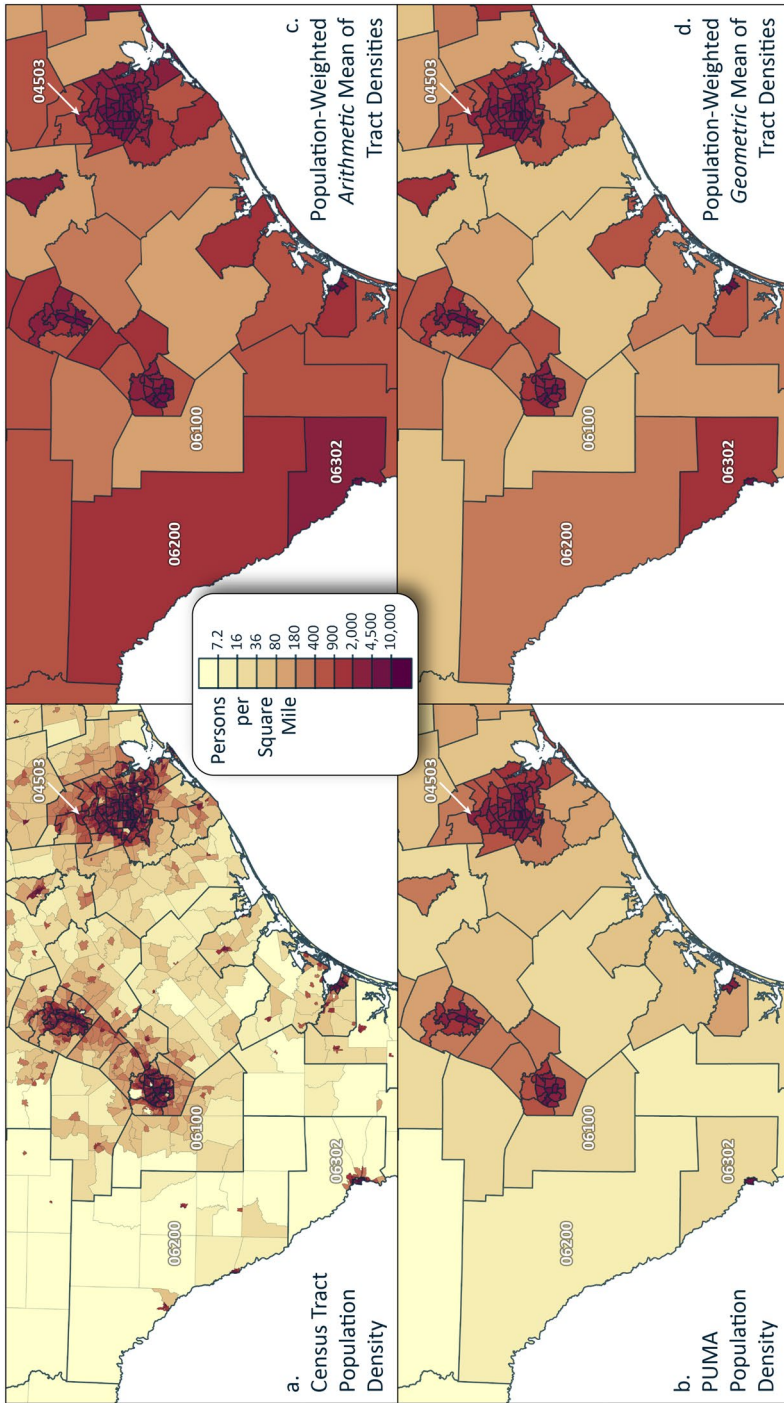


Fig. 5 Four measures of 2010 population density within PUMAs in a section of south-central Texas



Fig. 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

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

In practice, our computations deviate somewhat from these equations where a PUMA boundary subdivides a tract.⁷ 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 (determining this portion by summing the populations of the census blocks with centroids in each PUMA).

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 & 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).

⁷ 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 6 illustrates how four PUMA-level density measures relate to the distribution of tract densities in the PUMAs labeled on Fig. 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 2061 to 2290 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 (1436) and median (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 (1285) and geometric mean (1245). 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 (2945) 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 & 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

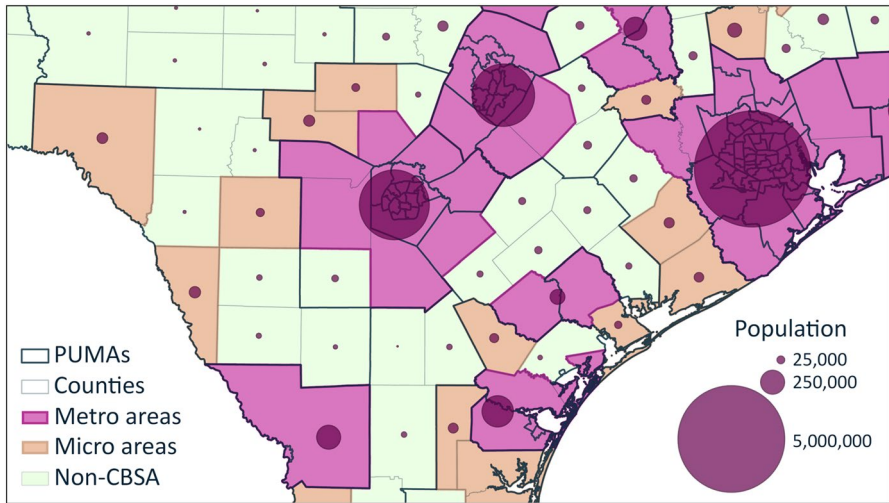


Fig. 7 2010 PUMAs and the 2010 populations of 2013 CBSAs and noncore counties, which form the basis for the METPOP10 variable, in a section of south-central Texas

window around each census block. This is similar to 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.

3.3 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 summarizes the sizes of all CBSAs where PUMA residents live. For the “noncore” counties located outside of any CBSA, we use the county’s population as an approximation of the commuting system size.⁸

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

⁸ For Virginia “independent cities” that lie outside of CBSAs, we combine the populations of the independent cities with the populations of their neighboring counties.

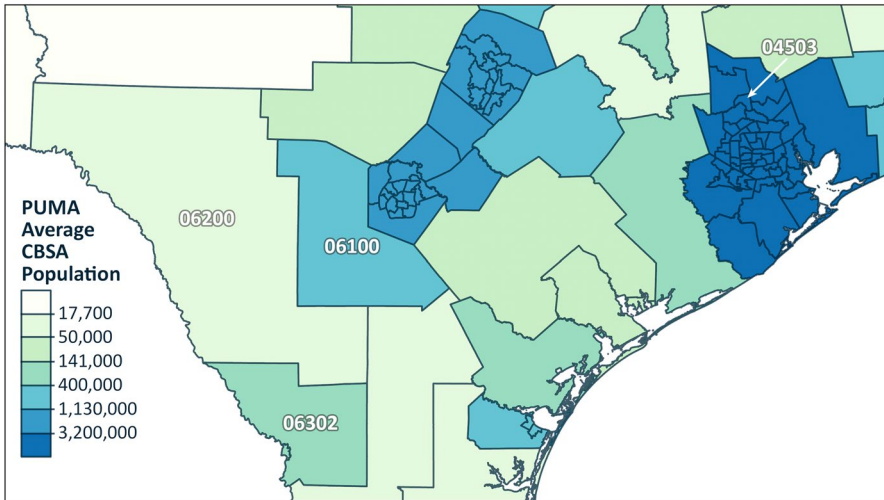


Fig. 8 METPOP10 values in a section of south-central Texas

PUMAs correspond to the CBSA and county populations that METPOP10 summarizes, and Fig. 8 illustrates METPOP10 values.

The formula we use to compute the METPOP index mirrors the DENSITY formula [Eq. (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 200,000 is about as significant for the character of a commuting system as any other factor-of-2 difference (e.g., 1 million and 2 million), and it is clearly more significant than an equal absolute difference of 100,000 in large commuting systems (e.g., 10.1 million and 10.2 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 is 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

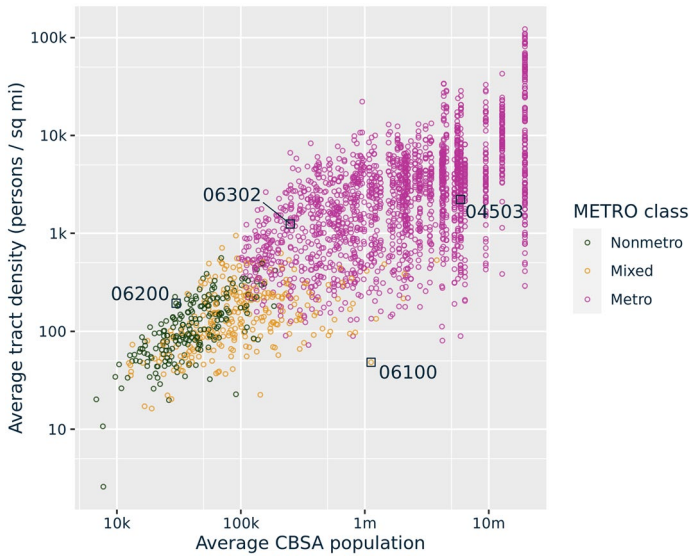


Fig. 9 Relationships among three urbanization indices for 2010 PUMAs. Labels identify the four Texas example PUMAs. (Color figure online)

size might delineate commuting systems based on tracts rather than counties, or it might use ERS commuting zones (Fowler & Jensen, 2020; U.S. Department of Agriculture, 2019a), 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.

3.4 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 indicate the METRO class of each PUMA. The overall distribution mirrors closely the conceptual model in Fig. 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 high average densities occur only in or around medium-to-large CBSAs.

The colors in Fig. 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.

Table 1 Poverty rates by METRO category, 2012–2017

	Nonmetro	Mixed	Metro
Percent in poverty	17.6	16.5	14.4
N = 18,120,063			

They may or may not share characteristics with other low-density PUMAs. Similarly, the nonmetro and mixed PUMAs with moderately high densities may 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 also makes it possible to distinguish fine gradations of variation and to identify inflection points across all densities or across all levels of the urban hierarchy.

4 Illustrative Results

To demonstrate the utility of the two indices, we analyze how poverty rates vary across settlement types in the U.S. Past analyses of “rural” poverty have often used the “metro/nonmetro” classification alone to distinguish urban and rural populations (e.g., Cromartie, 2017; Ziliak, 2018). This practice is problematic. As Fig. 3 shows, the basic question of whether rural or urban areas have higher poverty rates has distinctly different answers depending on how “rural” and “urban” areas are distinguished. The availability of two continuous measures, indexing two dimensions of urbanization, allows researchers to complete a more thorough and nuanced analysis of variations across geographic regions.

In this section, using the new indices with 2012–2017 ACS microdata from IPUMS USA (Ruggles et al., 2019), pooling six 1-year samples, we illustrate how both poverty rates and individuals’ likelihood of being in poverty vary across levels of urbanization as distinguished by both concentration (DENSITY) and population size (METPOPI0). 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 correlations between poverty, rurality, and other demographic characteristics vary in ways that cannot be captured by a simple metro/nonmetro distinction.

We begin our analysis with the basic METRO classification that had previously been (and still is) available in IPUMS USA. PUMAs classified as wholly nonmetro (neither metro nor mixed) have a higher poverty rate than other PUMAs, and the “mixed” PUMAs, those that straddle metro and nonmetro areas, have a poverty rate between nonmetro and metro PUMAs’ (Table 1).

Examining how poverty rates vary with the two new indices uncovers a more nuanced geographic pattern. Each point in Fig. 10 represents a PUMA while the

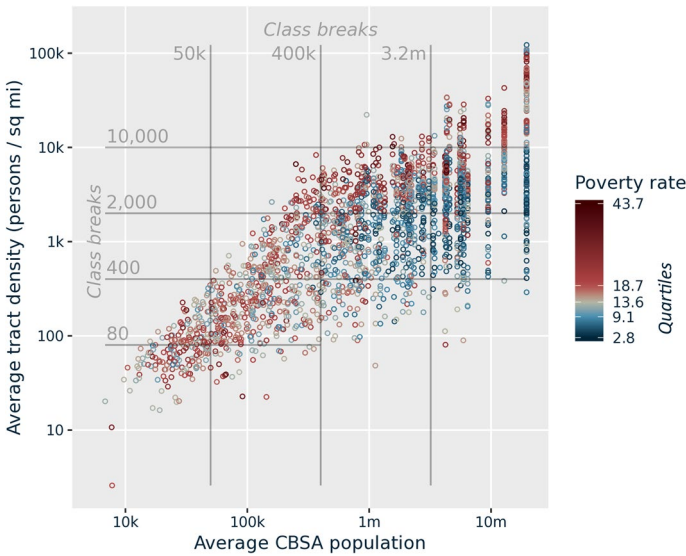


Fig. 10 Relationship between poverty and two urbanization indices of urbanization for 2010 PUMAs. IPUMS USA 2012–2017 ACS samples. (Color figure online)

color represents levels of poverty: blue represents lower poverty rates and red represents higher rates. As in Table 1, we see again that metro areas have generally lower rates of poverty than nonmetro areas, but by using two continuous indices, we can see how the metro/nonmetro dichotomy masks significant differences in poverty rates within metro areas. The high-density PUMAs in large metro areas generally have high poverty rates—similar to or even higher than the rates in the PUMAs of small commuting systems—while the lower-density PUMAs in large metro areas, encompassing mostly suburban and exurban communities, appear to have the lowest poverty rates overall.

To quantify how rates vary across this two-dimensional space, we first 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, 2000, 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 in large commuting systems (above 400,000). This produces 14 classes of PUMAs, each with unique ranges of METPOP10 and DENSITY values. Table 2 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 million residents). The lowest poverty rates are found in areas with medium density in the largest CBSAs (over 3.2 million residents). Table 3 shows how the poverty rate for each class differs from the rate for the lowest-poverty class, which we use later as a benchmark for analyzing poverty rates in a multiple regression framework. Both tables show that, within each of the four size classes, the highest poverty rates occur in the highest-density

Table 2 Poverty rates (%) by DENSITY and METPOP10, 2012–2017. (Color figure online)

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		

N = 18,120,063

Table 3 Difference in poverty rate (%) by DENSITY and METPOP10 from lowest-poverty class (METPOP10 of 3.2 m+, DENSITY of 400–2000), 2012–2017. (Color figure online)

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		

N = 18,120,063

classes. The PUMAs in small commuting systems also have relatively high rates, but not as high as in the high-density PUMA classes.

Of course, classifying populations according to PUMA-level averages, as in Tables 2 and 3, may obscure variations that are apparent only *within* PUMAs. Perhaps the highest rates of poverty actually occur within the lowest-density census tracts, but if all the PUMAs that contain these tracts also include a mix of higher-density tracts, then the distinct characteristics of the lowest-density tracts would be “averaged out.” Why then use PUMA-level indices? We reiterate that a key motivation is to enable microdata-based analyses that are impossible using existing census summary data. For example, the main reason for the higher poverty rates in high-density areas could be that those areas have disproportionately high concentrations of higher-poverty demographic groups, such as younger and/or minority populations, in which case the relationship between density and poverty rate might be insignificant after controlling for individuals’ demographic characteristics. Directly controlling for individuals’ characteristics is not possible with census tract summary data, but it is possible using microdata and PUMA-level indices.

Table 4 Coefficients for percent likelihood of poverty by METRO class, with and without controls, 2012–2017

METRO class	No controls	With demographic controls
Metro	– 3.17	– 4.01
Mixed	– 1.14	– 1.01

Omitted class is nonmetro. N = 18,120,063. All coefficients significant at $p < 0.01$

To demonstrate the value of the indices with an analysis that requires microdata, we begin with two models that associate poverty with metropolitan status. The first model predicts poverty status based only on the METRO status of a person’s PUMA:

$$poverty = \alpha + \beta METRO. \tag{5}$$

where β is a vector of coefficients for each METRO class (*mixed* and *metro*, with *nonmetro* omitted). The second model expands on the first by controlling for a large range of demographic characteristics:

$$poverty = \alpha + \beta METRO + \gamma D \tag{6}$$

where D is a vector of individual-level demographic controls, available only through microdata. Specifically, D includes age, sex, race, ethnicity (Hispanic/Latino), nativity, citizenship, marital status, health insurance coverage, educational attainment, employment status & sector, year, and geographic subregion (census division). Table 4 gives the *metro* and *mixed* coefficients after fitting these two models through linear regression on 2012–2017 ACS microdata.

Based only on metropolitan status, people residing in wholly metro PUMAs are about 3.2 percentage points less likely to be in poverty than those living in nonmetro PUMAs. (This is consistent with Table 1, which shows the difference between poverty rates in nonmetro and metro PUMAs to be $17.6 - 14.4 = 3.2$). However, when we include a battery of demographic controls, the metro coefficient decreases from -3.2 to -4.0, meaning that for individuals *with the same demographic characteristics*, those living in metro PUMAs are about 4 percentage points less likely to be in poverty than those in nonmetro PUMAs. This indicates that the demographics that predominate in nonmetro PUMAs would generally yield lower poverty rates than the demographics in metro PUMAs, but living in nonmetro PUMAs increases the likelihood of poverty enough to produce a higher poverty rate in those areas despite their demographics. Clearly, using microdata to control for demographic characteristics can help to reveal key dynamics in rural–urban poverty discrepancies.

Incorporating the two continuous urbanization indices into the analysis yields yet again more value. To demonstrate, we estimate the following linear probability model:

$$poverty = \alpha + \beta DENSITY \times METPOP10 + \gamma D \tag{7}$$

Table 5 Coefficients for percent likelihood of poverty by DENSITY and METPOP10, with controls, 2012–2017. (Color figure online)

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 3.2 m+ and DENSITY of 400–2000. N = 18,120,063. All coefficients significant at $p < 0.01$

Table 6 Coefficient for Hispanic/Latino likelihood of poverty by DENSITY and METPOP10, with controls, 2012–2017. (Color figure online)

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		

N = 18,120,063. All coefficients significant at $p < 0.01$

where D is the same vector of controls as in Eq. (6) and β represents coefficients for each DENSITY-by-METPOP10 class. Table 5 provides the results in a format that can be directly compared with the uncontrolled rate differences in Table 3, with both tables color-coded on the same scale. This comparison shows that the differences in the likelihood of poverty, relative to the lowest-poverty reference class, are consistently smaller after controlling for demographics. Without controls, the highest poverty rate among the classes is 13.6 percentage points above the lowest rate. After controlling for demographics, the same two classes differ by only 8.3 points. This indicates that 61% (8.3/13.6) of the difference between these classes is explained by the difference in densities and population sizes, and the remaining 39% is attributable to demographic differences between the populations in these areas.

A final angle we take is to examine how associations between poverty and specific demographic characteristics vary across settlement types. We run this model

$$poverty = \alpha + \gamma D, \quad (8)$$

Table 7 Coefficient for noncitizen likelihood of poverty by DENSITY and METPOP10, with controls, 2012–2017. (Color figure online)

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		

N = 18,120,063. All coefficients significant at $p < 0.01$

separately for each of the fourteen DENSITY-by-METPOP10 classes, where *D* is again the same vector of controls as in Eqs. (6) and (7). Table 6 reports the resulting Hispanic/Latino coefficient for each class. The pattern is altogether different from the pattern in previous tables. In the class with the highest overall poverty rate (22.8 in Table 3), the difference in poverty between Hispanic/Latino populations and other groups is at its smallest (2.1 in Table 6). The increase for Hispanic/Latino population is greatest (at +9.2) in the least dense PUMAs in mid-size metro areas. Table 7 similarly shows how coefficients for noncitizens vary among the classes. In this case, the likelihood that a noncitizen is in poverty differs most from citizens’ likelihood (+6.6) in medium-density PUMAs in moderately small commuting systems, and it is generally smallest in both the upper right (denser PUMAs in large metro areas) and lower left (low-density PUMAs in small commuting systems).

In all, these results demonstrate the utility and flexibility offered by continuous PUMA-level indices of urbanization, enabling researchers to distinguish a diverse range of settlement types and to quantify associations with robust demographic controls.

5 Conclusion

The limited precision of the geographic information in public use microdata from the U.S. Census and ACS makes it impossible to identify all urban areas and commuting systems exactly as they are defined by federal standards, but the spatial units that *are* identified in these microdata—PUMAs—are still small enough and diverse enough to enable analysts to investigate and account for demographic variations across multiple dimensions of urbanization. This ability, combined with the unique flexibility microdata offer for developing robust models of demographic relationships, make these data a potentially critical source for understanding and responding to the evolving differences among rural, suburban, and urban populations.

To support the effective analysis of multiple dimensions of urbanization using public use microdata, we have introduced 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). We have specified how the indices are computed, explained how to interpret them, and demonstrated their value 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.

We expect that these indices should be similarly useful for many other research applications on a range of subjects at least as wide as the subject coverage in ACS microdata, including migration, commuting, racial disparities, regional economics, housing markets, socio-economic sorting, etc. We would also emphasize that the analytical framework we use in our case study of poverty is only one of many approaches supported by the indices. For example, rather than classifying PUMAs into discrete groups, as we have, analysts could directly include the continuous measures in models, making it possible to quantify and plot formulaic relationships between population density, commuting-system size, and demographic variables, potentially identifying interactions or specific inflection points. We hope our framework will also serve as a basis for further innovation in indices of settlement patterns, including some exploration into the alternatives we proposed here such as a block-based (rather than tract-based) local density measure or a regional moving-window population total (rather than an average of CBSA populations).

Acknowledgements This work was supported by grants from the National Institutes of Health (R01HD043392, P2C HD041023). David Van Riper provided helpful feedback throughout the research process, and John Cromartie provided helpful comments as a conference discussant.

Authors' Contributions Jonathan Schroeder developed and computed the indices, produced the figures, and led the writing. José Pacas initiated the research, developed the illustrative models, produced the tables, and contributed to the writing.

Funding Support for this work was provided by IPUMS USA (NIH R01HD043392) and the Minnesota Population Center (NIH P2C HD041023).

Availability of Data and Materials Data for the featured indices are freely available at <https://usa.ipums.org>.

Code Availability The authors have not made code available but can upon request.

Declarations

Conflicts of Interest The authors declare that they have no conflict of interest.

Notice of Prior Versions The authors have released a similar version of this manuscript through the Minnesota Population Center Working Paper Series: #2019-05, <https://doi.org/10.18128/MPC2019-05>. An earlier version appears in the 2019 conference proceedings of the Population Association of America under the title

“Getting ‘Rural’ Right: Poverty Disparities Across Two Dimensions of Rurality,” including a third author, David Van Riper.

References

- Coombes, M., & Raybould, S. (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.
- Cromartie, J. (2017). Rural America at a glance, 2017 edition. *Economic Information Bulletin EIB-182*. Economic Research Service. U.S. Department of Agriculture.
- Dorling, D., & Atkins, D. J. (1995). Population density, change and concentration in Great Britain 1971, 1981 and 1991. *Studies in Medical and Population Subjects No. 58*. HMSO.
- Fowler, C. S., Frey, N., Folch, D. C., Nagle, N., & Spielman, S. (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., & 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>
- 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. Retrieved December 12, 2019, from <https://turnercenter.berkeley.edu/blog/urban-revival-not-for-most-americans>
- Manson, S., Schroeder, J., Van Riper, D., & Ruggles, S. (2018). IPUMS national historical geographic information system: Version 13.0 [Database]. IPUMS. <https://doi.org/10.18128/D050.V13.0>
- Mattingly, M. J., & Bean, J.A. (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., & Taylor, P. J. (1981). The modifiable areal unit problem. In N. Wrigley & R. J. Bennett (Eds.), *Quantitative geography: A British view* (pp. 60–70). London: Routledge and Kegan Paul.
- Ratcliffe, M. (2015). A century of delineating a changing landscape: The Census Bureau’s urban and rural classification, 1910 to 2010. Annual Meeting of the Social Science History Association, Baltimore, November 12–15, 2015. Retrieved November 11, 2020, from https://www2.census.gov/geo/pdfs/reference/ua/Century_of_Defining_Urban.pdf
- Ratcliffe, M., Burd, C., Holder, K., & Fields, A. (2016). *Defining rural at the U.S. Census Bureau*. ACS-GEO-1, U.S. Census Bureau.
- Ruggles, S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J., & Sobek, M. (2019). Integrated public use microdata series: Version 9.0. University of Minnesota. <https://doi.org/10.18128/D010.V9.0>
- U.S. Department of Agriculture. (2019a). Commuting zones and labor market areas. Economic Research Service, U.S. Department of Agriculture. Retrieved December 12, 2019, from <https://www.ers.usda.gov/data-products/commuting-zones-and-labor-market-areas>
- U.S. Department of Agriculture. (2019b). Rural classifications overview. Economic Research Service, U.S. Department of Agriculture. Retrieved October 25, 2019, from <https://www.ers.usda.gov/topics/rural-economy-population/rural-classifications>
- Wang, M., Kleit, R. G., Cover, J., & Fowler, C. S. (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., Plane, D. A., Mackun, P. J., Fischetti, T. R., Goworowska, J., Cohen, D., Perry, M. J., & Hatchard, G. W. (2012). Patterns of metropolitan and micropolitan population change: 2000 to 2010. 2010 Census Special Reports. U.S. Census Bureau.
- Ziliak, J. (2018). Are rural Americans still behind? *IRP Focus*, 34(2), 13–24.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Jonathan P. Schroeder¹  · José D. Pacas¹ 

✉ Jonathan P. Schroeder
jps@umn.edu

¹ Institute for Social Research and Data Innovation, University of Minnesota, 50 Willey Hall, 225 19th Ave S, Minneapolis, MN 55417, USA