Supplementary Material:

State Legislative Districts Urban-Rural Dataset

Objective

Scholars of American politics may be interested in measuring the urban-rural makeup of state legislative districts to answer a wide range of questions. However, existing data do not provide high-quality mapping of urban-rural indicators onto state legislative districts. The most common strategy to measure the geographic makeup of state legislative districts is using urban-rural population counts from the U.S. Census. The Census provides the urban and rural populations by population count and population ratio for all state legislative districts for both the 2000 and 2010 Census. The Census defines rural as "all population, housing, and territory not included within an urban area." Urban areas include Urbanized Areas (UAs) of 50,000 or more people and Urban Clusters (UCs) of at least 2,500 and less than 50,000 people. A significant downside of this measure is the lack of distinction between urban, suburban, exurban, and rural locations. This measure also does not account for proximity to other population centers or population density, which may be of interest to political scientists. To address these shortcomings, we created a dataset of Rural Urban Commuting Area Codes assigned to state legislative districts for state legislative boundaries in 2007, 2010, 2012, 2014, and 2016. The codebook is on the last page of this document.

Data Source

We created the *State Legislative Districts Urban-Rural Dataset* using the geographic relationship files from the Missouri Census Data Center's Geocorr program. Table 1 shows the availability of relationship files mapping ZCTA populations onto state legislative districts.[1] Table 2 shows the years of data that were selected to create the dataset.

[1]Note that 1990 Census data is available for only Missouri.

Table 1: Missouri Census Data Center Geocorr Relationship Files

Database Version	Years Available for District Boundaries	Years Available for Population Data
Geocorr 2000	2007	2000 2009 (estimate)
Geocorr 2014	2010 2012 2014	2000 2010 2014 (estimate)
Geocorr 2018	2010 2012 2014 2016	2000 2010 2016 (estimate)

Table 2: Relationship Files and Data used in Dataset

	Source: Missouri Census Data Center			Source: U.S. Census
Geocorr Version	**Census Tract Year**	**Legislative District Year**	**Population Year**	**RUCA Version**
2000	2000	2007	2000	2000
2014	2010	2010	2010	2010
2014	2010	2012	2010	2010
2014	2010	2014	2014	2010
2018	2010	2016	2016	2010

The Geocorr program allows us to match the census tract RUCA codes from the U.S. Census onto corresponding legislative districts. Ninety-three percent of the census tracts from the Geocorr relationship file were exactly matched with a RUCA code from the census data. All remaining RUCA codes were matched onto census tracts using a secondary match.[2].

Tables 3 and 4 show the summary statistics for the amount of census tracts and unique RUCA codes per state legislative district. Although most census tracts have populations between 1,200 and 8,000 people, the range of populations for census tracts is 0 to 37,452 people (Source: 2010 US Census RUCA Data). This explains the minimum values of 1 for census tracts per district (Table 4) because a legislative district may consist of a single, atypically-large tract. The distribution of

[2]Most, but not all, census tracts are assigned a RUCA code by the USDA. Select census tracts are not coded with a RUCA code for a variety of reasons, such as no commuting data available at unit of analysis or no population lives in the census tract. We wanted to retain as many census tracts as allowed by a reasonable matching procedure. A secondary match involves removing the smallest digit to allow for more flexibility in matching RUCA-coded census tracts to the census tracts in the Geocorr relationship file. For example, census tract 1130.1 may be divided into two tracts, 1130.11 and 1130.12, at the next Census due to population growth during the decade between censuses. A secondary match removes the smallest digit, 0.01 and 0.02 (respectively), so that 1130.1 and 1130.1 can be matched with the RUCA code assigned to 1130.1. Census tracts are such fine-grained geography that even merging tracts outdated by one Census or neighboring census tracts is more accurate than using Zip Code Tabulated Areas, the next smallest unit of geographic aggregation. For more information on how census tracts change, see this resource

census tracts per legislative district is displayed graphically in Figure 1.

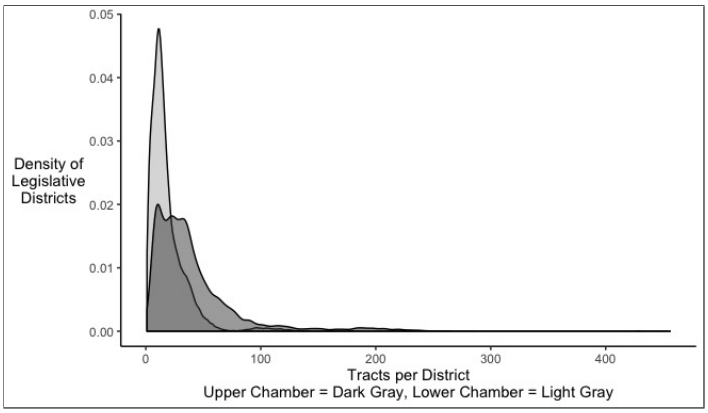
Table 3: Unique RUCA codes per State Legislative District

	Upper Chamber	Lower Chamber
Mean	3	2
Median	2	2
Standard Deviation	2	2
Minimum	1	1
Maximum	10	10

Table 4: Census Tracts per State Legislative District

	Upper Chamber	Lower Chamber
Mean	41	19
Median	31	14
Standard Deviation	38	18
Minimum	2	1
Maximum	281	456

Figure 1: Census Tracts per Legislative District for 2007, 2010, 2012, 2014, and 2016 district boundaries



Empirical Strategy

We use three methods for assigning RUCA codes to state legislative districts. The dataset includes state legislative district RUCA codes calculated by all three methods.

The first method probabilistically assigns RUCA codes to state legislative districts. This is similar to the method used by Tausanovitch and Warshaw (2013). They use respondents zip codes probabilistically assigning survey respondents to state legislative districts based on the proportion of people in their zip code that live in each district (Tausanovitch and Warshaw 2013). The down-side of this method is that it is possible that a state legislative district is assigned a RUCA code that represents a very small population of the district. While this procedure may create noise into the probabilistic assignments, this type of assignment occurs without systematic bias.

The second method is using the averages of the RUCA scores weighted by population. Unlike probabilistic assignment, this value represents that range of RUCA codes within a district by using the average of RUCA values within the district instead of selecting a single RUCA value within the district. A drawback of this method is that the RUCA codes are ordinal values that thus should not be averaged, although this is common practice among political scientists.

The third method is to assign the RUCA code that describes the plurality of the legislative district. Unlike the first method, it ensures that the state legislative district RUCA code is representative of a plurality of the district. This method is more accurate for districts in which a clear majority of the population belong to a specific RUCA code than it is for legislative districts evenly split among many different RUCA codes. This method is likely too imprecise to be of much use in statistical analyses, but would be a useful descriptive statistic of the most common geographic classification within a district.

We encourage researchers to use data produced by the method that best aligns with their theory, or to employ more than one measure in a series of robustness checks. The correlations between the measures are shown in Table 5. The weighted average and plurality methods are highly correlated. Using them interchangeably is unlikely to alter empirical results in a statistically or substantively meaningful way. As noted above, the plurality measure is not an appropriate method for drawing conclusions about a district overall.

Table 5: Correlations Between State Legislative District RUCA Classification Methods

	Probability	Plurality
Plurality	0.720 (0.715, 0.725)	
Weighted Average	0.846 (0.843, 0.849)	0.854 (0.851, 0.857)

Notes: Pearson's correlations between methods of assigning RUCA codes to state legislative districts. Ninety-five percent confidence intervals in parentheses below the correlation coefficients.

Diagnostics

There is a small amount of missingness in the dataset, the origins of which are unknown. The coverage rates are shown in Table 6. Researchers may also consider patching the missingness with data from an adjacent year. For example, a researcher could replace the NAs in the 2010 dataset with values from the 2012 dataset, the most proximate year.

Table 6: Data Coverage of State Legislative Districts in United States

	2007 Boundaries	2010 Boundaries	2012 Boundaries	2014 Boundaries	2016 Boundaries
Number of Districts in the Dataset	6,549	6,555	6,622	6,622	6,622
Coverge as Proportion of Total Districts	97%	97%	98%	98%	98%

Note: there are 6,764 total state legislative districts in the United States. This is the denominator for the coverage as proportion of total districts.

Table 7 shows the count of districts at varying levels of missingness across years. Ninety-five percent of the observations (districts) in the dataset have no data missing across the five years of legislative boundaries (2007, 2010, 2012, 2014, and 2016).

Table 7: Missingness at District Level

Years of Missing Data for a particular district (out of 5 total years)	Count of Districts
0	6,543
1	18
2	170
3	101
4	16

Table 8 shows the population coverage within districts. This calculation is the population from the RUCA census tract dataset pertaining to a particular district, divided by the overall district population (source: Ballotpedia). Coverage loss for district populations occurs if not every Census tract within a district received a RUCA code during the coding process (done by the United States Department of Agriculture Economic Research Service). Census tracts may not receive a RUCA

Table 8: Population Coverage by District

Percent of District Population Included in Coverage	Count of Districts
>0 - 9%	0
10 - 19%	0
20 - 29%	1
30 - 39%	3
40 - 49%	7
50 - 59%	36
60 - 69%	107
70 - 79%	410
80 - 89%	872
>90%	31,797

code if there is not data available for that census tract. The missing data are usually commuting estimates, which are harder to collect than population or geographic size. Table 8 groups districts into ten bins based on the proportion of district population used to calculate the district RUCA code. For example, 107 districts have population coverage of between 60% and 69%, meaning that the RUCA code assigned to those districts was based off of the RUCA codings of around two-thirds of the district's residents. Ideally, coverage would be 100%. Notably, the vast majority of districts have over 90% population coverage. Some district-year observations have greater than 100% population coverage because the district population has grown in the decade since redistricting. Districts with no (0%) population coverage are excluded from the dataset because there is no way to calculate a RUCA code for such districts. There are 298 district-years (out of 33,820[3] district-years) that have no population coverage and are excluded from the dataset and not included in Table 8.

[3]6,764 districts * 5 years = 33,820.

Codebook

abbreviation State postal abbreviation

state State FIPS code

district Legislative district number

year Year of the district boundary, *see Table 2 for more detail*

upper Dummy variable for upper chamber = 1, otherwise 0

weightedaverage_YEAR RUCA code assigned using weighted average

probability_YEAR RUCA code assigned using probability

plurality_YEAR RUCA code assigned using plurality (most common)