Evaluating the reliability of ACS data for transportation planning

Addison Larson

Delaware Valley Regional Planning Commission
April 25, 2019

Transportation and planning organizations rely on the Census Transportation Planning Products (CTPP) for information about people and their commuting behaviors. For this reason, the 2019 release of the CTPP tabulation derived from 2012-2016 American Community Survey (ACS) 5-Year Estimates was highly anticipated.

However, CTPP data present the same challenges as ACS estimates to planners and researchers: sample error can render the data unusable, especially when working with small geographies and detailed variables. Data users must think critically about the trade-offs between detail and reliability when working with CTPP data.

This paper addresses reliability issues associated with CTPP data in three sections. Section 1 introduces the CTPP special tabulation and its offerings regarding table type, geography, and variables. It briefly compares the CTPP to two other commuting data sources, the ACS and LODES (LEHD Origin-Destination Employment Statistics). Section 2 investigates CTPP reliability by table type, geography, variable detail, and local context. Section 3 lists recommendations for data users to improve the reliability of their analyses, including data selection, cartographic choices, and data aggregation methods.

1 Introduction to the CTPP special tabulation

The CTPP is a special tabulation of ACS 5-Year Estimates offering residence-based, workplace-based, and commuting flows data for specialized variables and a range of geographies. This section provides an overview of CTPP table types, geographies, and variables; and briefly compares it to two other commuting data sources, the ACS and LODES.

1.1 Table type

The CTPP provides data in residence-based, workplace-based, and commuting flows formats. To understand the difference between these three types of tables, consider Census Tract 4.02, a census tract in Center City Philadelphia. Census Tract 4.02 is part of the downtown employment hub immediately west of Philadelphia's City Hall (see Figure 1).

Data users interested in identifying the total number of workers in Census Tract 4.02 can find estimates in the CTPP's residence-based, workplace-based, and commuting flows tables, as shown in Table 1. Each estimate tells a different story. The residence-based table says that 1,810 ($\pm$327) workers live in Census Tract 4.02; workers who live in Census Tract 4.02 and work outside of it are included in this count. The workplace-based table says that 51,425 ($\pm 1,603$) workers come to Census Tract 4.02 for work; it says nothing of workers' origins. Lastly, the flows table gives a complete accounting of workers who live in Census Tract 4.02,

Fig. 1. Census Tract 4.02 is a Center City Philadelphia tract immediately west of City Hall.

work in Census Tract 4.02, or both. In this case, an estimated 260 people both live and work in Census Tract 4.02. Because the commuting flows data are structured as an origin-destination matrix, the tract estimate of 260 (±104) workers is one of 767 unique origin-destination cells containing information for Census Tract 4.02 in Philadelphia's tract-level commuting flows data.

Table 1. Number of workers with Census Tract 4.02 as origin, destination, and O-D pair. Sources: CTPP Tables A102101 (Residence-based), A202100 (Workplace-based), and A302100 (Commuting flows).

	Number of workers 16 years and over	
Table type	**Estimate**	**Margin of error**
Residence-Based	1,810	327
Workplace-Based	51,425	1,603
Commuting flows	260	104

Comparison to ACS and LODES ACS only offers residence-based tabulations. LODES offers Residence Area Characteristics (RAC), Workplace Area Characteristics (WAC), and Origin-Destination (OD) tables analogous to the CTPP residence-based, workplace-based, and commuting flows tables.

1.2 Geography

The CTPP is available at the state, county, Minor Civil Division (MCD), Place, Metropolitan Statistical Area (MSA), Public Use Microdata Area (PUMA), Census Tract, Transportation Analysis Zone (TAZ) and Traffic Analysis District (TAD) geographies. Figure 2 shows the county and sub-county geographies available in the current CTPP release for Philadelphia County, Pennsylvania. PUMAs and their workplace-based equivalent, Place of work PUMAs (POWPUMAs) are both shown in Figure 2 to highlight their different boundaries.

The 2012-2016 5-Year Estimates release will be the last CTPP tabulation available at the TAZ and TAD geographies. Block groups will replace TAZs in future releases, and because TADs are aggregations of TAZs, they will disappear as a consequence (Census Transportation Planning Products Program Oversight Board, 2018). See Figure 3 for a visual comparison of TAZ and 2010 block group geographies in Philadelphia.

Comparison to ACS and LODES The ACS offers more geographies than the CTPP, such as Zip Code Tabulation Areas and Combined Statistical Areas. ACS estimates at the PUMA geography can be calculated from the Public Use Microdata Sample.

LODES is only available at the census block level. Data users who wish to use LODES data at larger geographies can aggregate geographies based on the block GEOID or the geographic crosswalk file provided by the LEHD program.

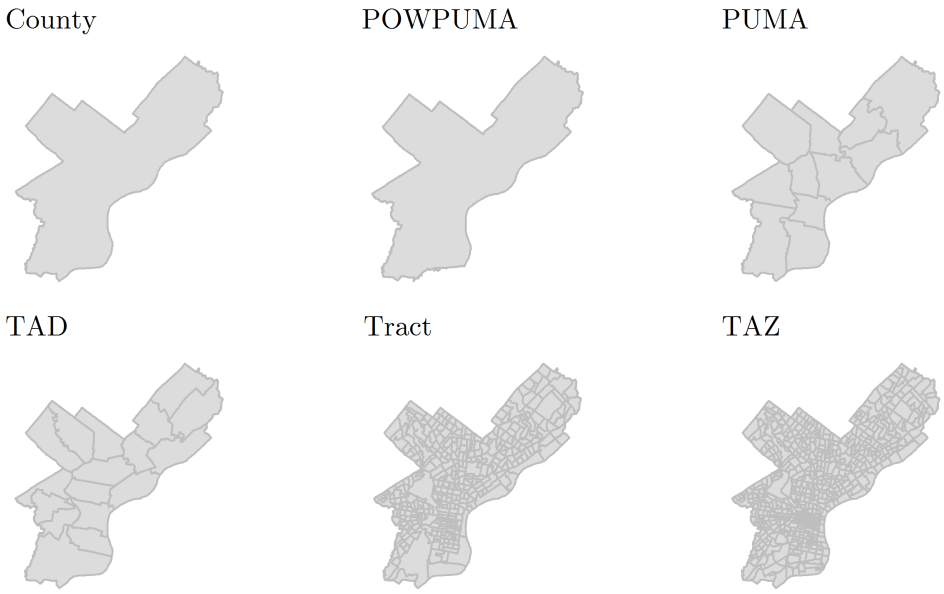

Fig. 2. A comparison of CTPP county and sub-county geographies.

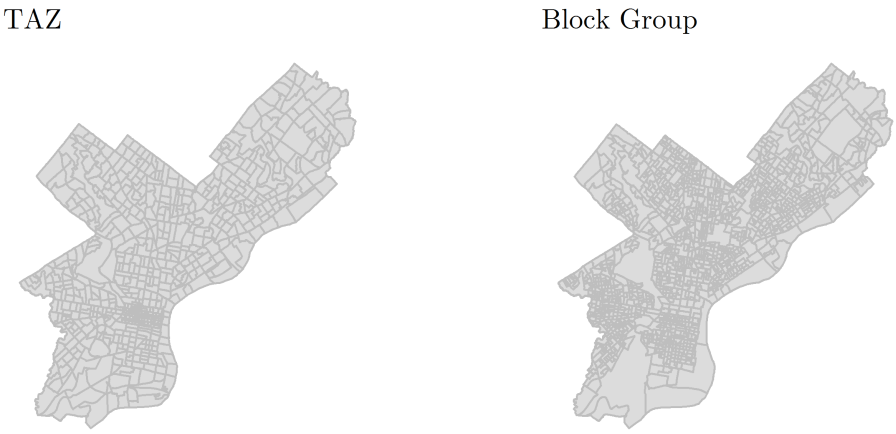

Fig. 3. A comparison of the CTPP TAZ geography and the 2010 block group geography. Note that block group boundaries are subject to change in the 2020 Census.

1.3 Variables

The CTPP provides demographic information, such as age, sex, income, language, race, ethnicity, and poverty; housing information including the number of housing units, vacancy status, and tenure; worker information such as industry and class; and commuting information such as time leaving for work, commute mode, and number of vehicles available. It also provides detailed commuting-related variables and cross-tabulations.

Comparison to ACS and LODES While the ACS has commuting-related estimates, the CTPP offers these variables at a unique level of detail. For example, ACS Table B08302 (Time leaving home to go to work) presents residents' departure times split into 30-minute increments from 5:00 a.m. to 9:00 a.m. and larger increments outside the AM peak. CTPP Table A102108 (Time leaving home) presents residents' departure times split into finer-grained 15-minute increments from 5:00 a.m. to 11:00 a.m. and larger increments outside the AM peak.

The CTPP also offers commuting-related cross-tabulations not available in ACS or LODES data. For example, CTPP Table A112310 (Number of workers in household by vehicles available by household income in the past 12 months) can give data users unique insight into households with few vehicles relative to the number of workers.

LODES data contain information on age, wage, sex, industry, race, ethnicity, and education. The level of variable detail in LODES is generally lower than what is offered in the CTPP.

1.4 Conclusion

The CTPP offers table types, geographies, and variables tailor-made for transportation and planning organizations. However, the choice to use CTPP, ACS, or LODES data depends on the use case and can vary based on additional considerations such as update frequency, the universe of workers considered, and the assignment of workers to establishments. See *Employment Data for Planning: A Resource Guide* for an overview on additional differences between CTPP, ACS, and LODES (Cambridge Systematics, 2017).

2 Reliability of CTPP data

Because CTPP data is derived from the ACS, it is subject to sampling error. Each observation in CTPP data comes with a margin of error (MOE). The magnitude of the sampling error relative to the estimate, also known as the coefficient of variation (CV), varies with table type, geography, variable detail, and local context.

A common approach to working with sample-based data like the CTPP or the ACS is to ignore MOEs and CVs entirely. While this is the simplest approach, data users risk making decisions based on data with uncertainty that far exceeds the values of the estimates.

2.1 Margins of error make a difference

Consider a Philadelphia transportation planner intending to prioritize investment in the city's bicycle facilities. The planner uses a tract-level CTPP commuting flows table on means of

transportation to justify selection of road segments for a proposed bicycle lane. Knowing that Census Tract 4.02 in Center City Philadelphia is a major workplace destination, the planner identifies the eight origin census tracts with the highest estimated number of cyclists to Census Tract 4.02. Figure 4 is a map of the eight census tracts' estimated bicycle commuters disregarding uncertainty about the estimates, and Table 2 shows the estimates, MOEs, and CVs for the eight origin tracts. If the planner looks only at the map of Philadelphia's bicycle commuters, then it appears that the two westernmost census tracts, shown in yellow, have more bicycle commuters to Census Tract 4.02 than any other origin census tracts.

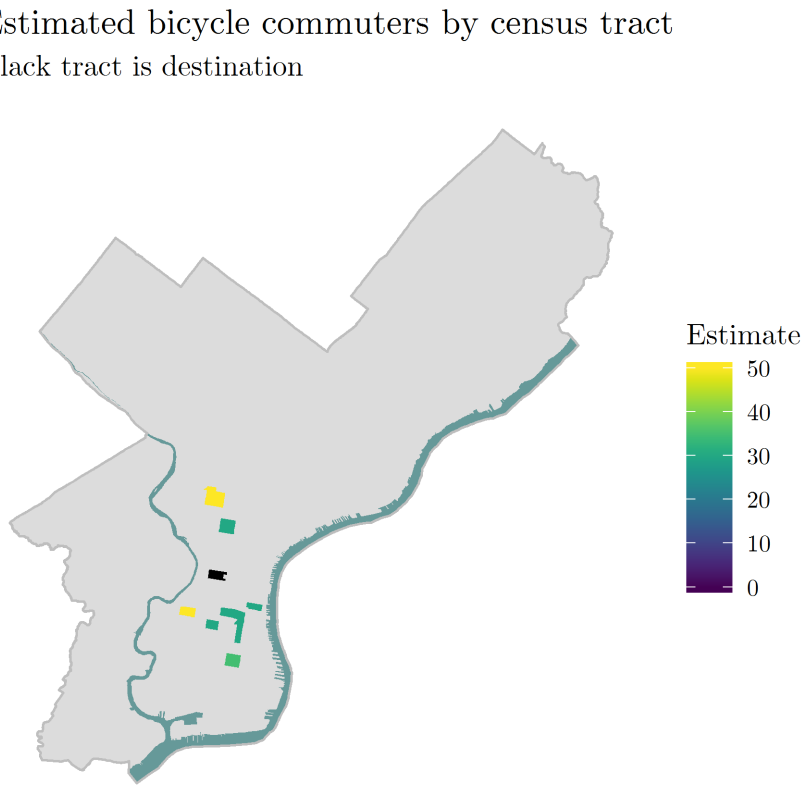

Fig. 4. Estimated bicycle commuters by census tract. Only the eight origin tracts with the highest estimated number of cyclists are shown. The black census tract is Census Tract 4.02. Source: CTPP Table A302103.

Considering the uncertainty about each estimate makes the planner's decision more difficult. In all but one case, the MOEs of the eight origin tracts are larger than the estimates. For

Table 2. Origins of bicycle commuters with destination of Census Tract 4.02. Source: CTPP Table A302103.

Origin	Estimate	MOE	CV
Census Tract 20	50	65	79.03
Census Tract 152	50	73	88.75
Census Tract 41.02	35	58	100.74
Census Tract 140	30	37	74.97
Census Tract 28.01	30	27	54.71
Census Tract 30.02	30	37	74.97
Census Tract 16	30	37	74.97
Census Tract 24	30	38	77.00

example, Census Tract 20 has an estimated 50 bicycle commuters, plus or minus 65 cyclists, meaning that the tract has somewhere between 0 and 115 cyclists.

While MOEs can change according to the scale of the underlying data, CVs transform MOEs into a consistent scale that enables comparison among datasets. An acceptable CV depends on the use case, but a general rule of thumb is that CVs in the range of 10-12% are acceptable, and CVs in excess of 30% indicate unreliable data (Francis, Vink, Tontisirin, Anantsuksomsri, & Zhong, 2012; Citro & Kalton, 2007, p.64). In the case of Philadelphia's cyclists, the CVs of all eight estimates are over 75% and high enough to warrant concern.

The uncertainty associated with each estimate makes it impossible to differentiate between census tracts regarding their relative numbers of cyclists. This is evident when plotting the possible ranges of cyclists originating in all eight census tracts simultaneously, as in Figure 5. The margin of error associated with each estimate represents the spread of the possible number of cyclists on the ground. For a given estimate, the range of possible values is normally distributed, with a mean equal to the estimate and a standard deviation equal to the standard error, or $\left(\frac{MOE}{1.645}\right)$. Given the estimate and its standard error, we are then able to plot the probability density of the number of bicycle commuters from each of the eight origin tracts. In the resulting plot, we see substantial overlap among the eight tracts, with the bulk of the probable values for all origin tracts somewhere between 0 and 60 cyclists.

Because the potential number of cyclists to Census Tract 4.02 is largely similar between the eight origin tracts, simply prioritizing tracts by their estimates draws distinctions where they may not exist.

2.2 Catalysts and inhibitors of data reliability

The example of prioritizing areas for bicycle facilities reveals the risks of ignoring sampling error. Data users' selections of table type, geography, and level of variable detail, in varying local contexts, can all serve as "catalysts" of CVs; in this instance, selecting a commuting flows *table type* and a small census tract *geography* in a *context* where bicycle commuters are rare compared to other modes of transportation causes CVs to explode.

This need not always be the case. Smart selections of table types, geographies, and variables can serve as "inhibitors" that keep CVs in check, and a data user with a keen eye for local context can find reliable data amid selections of table types, geographies, and variables that appear risky at face value.

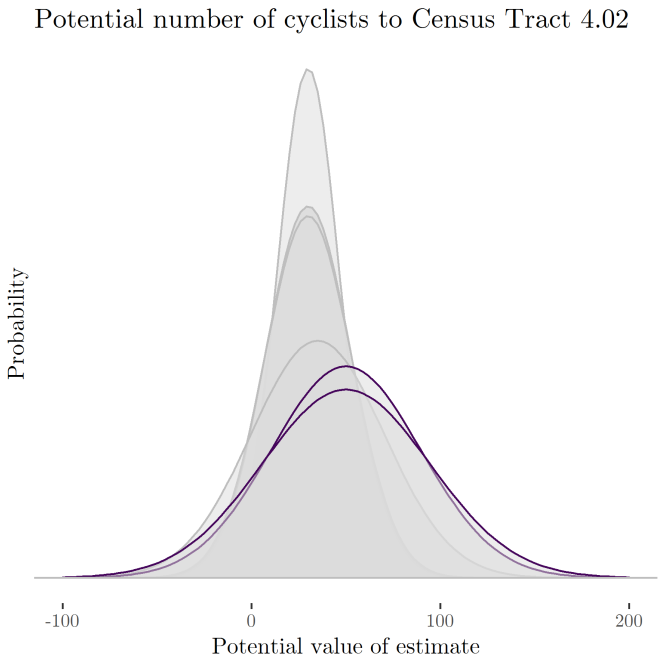

Fig. 5. Probability density of cyclists to Census Tract 4.02. Source: CTPP Table A302103.

This section analyzes the way CTPP reliability changes based on selections of table type, geography, and variable detail and gives two examples on the way CVs vary based on local context. It also evaluates changes in reliability along table type and geography as an example of the way data users' selections can simultaneously affect data reliability. Unless otherwise noted, CVs in this section are computed for the nine-county region served by the Delaware Valley Regional Planning Commission (DVRPC), the Metropolitan Planning Organization for Greater Philadelphia. The DVRPC region spans two states and includes the cities of Philadelphia, PA; Trenton, NJ; and Camden, NJ, as well as suburban and rural areas, serving as a useful microcosm of several geographic contexts.

All CVs in this section are computed using the formula below and are written as percentages for readability. Percent signs are not written in tables except when indicating the share a certain cell comprises among the universe of values.

$$CV = \frac{\left(\frac{MOE}{1.645}\right)}{Estimate} \cdot 100 \tag{1}$$

When an estimate is 0, it is not possible to compute the CV. In these instances, the CV is assigned a value of 100%. Tables where the median CV equals 100% indicate that the variable or variables under scrutiny had several zero estimates.

Lastly, the Online Appendix provides access to summary tables of all variables in all CTPP residence-based tables without cross-tabulations at the county, PUMA, TAD, Census Tract, and TAZ geographies for the DVRPC region. Observations gleaned from the Online Appendix informed the content of this paper.

Table type Workplace-based and commuting flows tables are less reliable than residence-based tables, holding universes and geographies constant. Table 3 shows the CVs for five CTPP tables that are available with identical universes in all three table types at the county level.

Even at the county level, the maximum CVs shown for industry and means of transportation tables can be very high. The industry table is comprised of 15 variables, including industries such as "Manufacturing" and "Other services (except public administration)". Similarly, the means of transportation table is comprised of 18 variables and has a category specifically for people who commute by ferryboat. Both the industry and means of transportation tables have variables that have low counts in one or more counties in the region, which is likely the reason their maximum CVs can be so high.

Geography Reliability decreases when using smaller geographic units. Table 4 shows the CVs for the total number of workers at five geographies. Switching from TAD to tract and from tract to TAZ geographies carry serious reliability penalties. In the case of the total number of workers, the median CV at the census tract level is more than quadruple the median CV of the TAD, and the median CV of the TAZ is twice as large as the tract CV.

Variable detail Reliability decreases when using more detailed variables or cross-tabulations. As an example, Table 5 shows the effect of simultaneously considering household size and vehicles available. For most geographies, the median CV roughly doubles when considering household size.

Table 3. CVs by table type for the nine-county DVRPC region. See Table 10 in the Appendix for a list of source tables by CTPP table type.

Description	Statistic	Residence	Workplace	Flows
Total workers	Minimum	0.31	0.4	0.46
	Median	0.42	0.65	6.67
	Mean	0.45	0.7	7.7
	IQR	0.19	0.17	8.02
	Maximum	0.64	1.23	34.57
Age of worker	Minimum	0.38	0.74	0.66
	Median	2.19	2.87	21.78
	Mean	3.29	3.55	36.58
	IQR	4.39	2.62	43.71
	Maximum	10.67	11.22	258.36
Industry	Minimum	0.85	0.78	0.96
	Median	3.51	3.69	18.56
	Mean	6.39	6.55	32.98
	IQR	2.27	2.7	30.74
	Maximum	85.11	58.76	2051.67
Means of transportation	Minimum	0.41	0.61	0.71
	Median	12.29	14.18	100
	Mean	19.54	22.43	75.11
	IQR	18.59	23.6	62.42
	Maximum	109.42	167.17	288.75
Travel time to work	Minimum	0.3	0.41	0.49
	Median	2.55	2.72	23.82
	Mean	3.26	3.62	40.34
	IQR	2.75	2.9	53.15
	Maximum	10.05	16.75	638.3
Overall	Minimum	0.3	0.4	0.46
	Median	3.97	4.19	36.47
	Mean	9.62	10.49	51.52
	IQR	7.24	8.03	88.33
	Maximum	109.42	167.17	2051.67

Table 4. CVs by geography for the nine-county DVRPC region. Source: CTPP Table A102101.

CV	County	PUMA	TAD	Tract	TAZ
Minimum	0.31	0.75	0.85	3.09	3.34
Median	0.42	1.01	1.51	6.93	14.1
Mean	0.45	1.19	1.63	8.68	23.62
Maximum	0.64	2.29	3.37	182.37	218.84

The TAZ geography is the exception. Even the zero-car households variable without cross-tabulations has a median CV of 70.33%, so the effect of considering household size only makes the CVs marginally worse. CVs at the TAZ geography point toward two trends. First, the median TAZ CV of zero-car households with one worker is 100%, indicating that several TAZs have an estimated 0 zero-car households with one worker. Second, the fact that the mean CV is less than the median CV indicates that some TAZs with nonzero estimates have more reliable data than one might expect. Despite the quality of a handful of observations, most observations at the TAZ geography have CVs that far exceed the common CV reliability threshold of 30%.

Table 5. Data reliability decreases when using cross-tabulations. Source: CTPP Table A112310.

Count of households with zero vehicles available

CV	**County**	**PUMA**	**TAD**	**Tract**	**TAZ**
Minimum	0.86	2.09	2.17	6.73	7.38
Median	3.34	6.76	9.29	33.77	70.33
Mean	3.38	5.97	10.23	40.73	69.12
Maximum	5.59	10.09	27.23	303.95	303.95
Count of households with zero vehicles available and *1 worker*					
Minimum	1.59	3	3.13	11.85	12.66
Median	7.29	13.84	20.56	70.52	100
Mean	6.98	12.96	22.75	73.49	88.05
Maximum	10.67	27.72	60.79	319.15	319.15

Local context The example of zero-car households and zero-car households with 1 worker shows that data reliability is context-dependent. While most census tracts and TAZs have excessively high CVs, some tracts have CVs as low as 11.85%, which is sufficient for most analyses.

Areas with higher counts tend to have more reliable CVs, and areas with lower counts tend to have less reliable CVs. Tables 6 and 7 evidence this trend. Along the x-axis of each table, estimates of zero-car households and zero-car households with 1 worker for the City of Philadelphia at the tract geography are grouped into standard deviation bins, where estimates above $1.5 \cdot SD$ are well above average. Along the y-axis of each table, the CVs of the estimates are grouped into reliability bins following Francis et al. (Francis et al., 2012). Observations with CVs below 15% have high reliability, and observations with CVs between 15% and 30% are sufficient but should be used with caution. Lastly, cells in Tables 6 and 7 are shaded according to the share each $estimate \times reliability$ bin comprises of Philadelphia's census tracts overall.

The first visual impression of Tables 6 and 7 are the streaks of shading tracing from top right to bottom left. The top right area of each table represents observations with above-average and well above average estimates and high reliability. The bottom left area of each table represents observations with below-average estimates and low reliability. The shading of both tables indicates that observations with higher estimates tend to have more reliable CVs.

Table 6 shows *estimate* × *reliability* bins for zero-car households. Of Philadelphia's 384 census tracts, 333 (86.72%) meet the 30% CV reliability threshold. Of those 152 census tracts with high reliability, 91 (59.87%) have above-average or well above average estimates of zero-car households. Only 5 of 152 (3.29%) census tracts with high reliability have below-average estimates.

Table 7 shows *estimate* × *reliability* bins for zero-car households with 1 worker. While the shading in Table 7 follows the same shape as Table 6, the trend generally shifts down by one row, indicating that many observations with CVs less than 15% in Table 6 have CVs up to 30% in Table 7. The shift in overall reliability likely reflects the decrease in the number of zero-car households with one worker compared to the number of zero-car households in general. Still, 151 (39.32%) of Philadelphia's 384 census tracts have estimates with CVs sufficient for analysis.

Table 6. Zero-car households by estimate and reliability for Philadelphia's census tracts. Source: CTPP Table A112310.

	Estimate, zero-car households				
CV	$(-\infty, -1.5\,SD]$	$(-1.5\,SD, -0.5\,SD]$	$(-0.5\,SD, 0.5\,SD]$	$(0.5\,SD, 1.5\,SD]$	$(1.5\,SD, \infty)$
0-15%	0 (0%)	5 (1.30%)	56 (14.58%)	61 (15.89%)	30 (7.81%)
15.1-30%	0 (0%)	79 (20.57%)	86 (22.40%)	14 (3.65%)	2 (0.52%)
30.1-60%	0 (0%)	34 (8.85%)	0 (0%)	0 (0%)	0 (0%)
60.1+%	14 (3.65%)	3 (0.78%)	0 (0%)	0 (0%)	0 (0%)

Table 7. Zero-car households with 1 worker by estimate and reliability for Philadelphia's census tracts. Source: CTPP Table A112310.

	Estimate, zero-car households with 1 worker				
CV	$(-\infty, -1.5\,SD]$	$(-1.5\,SD, -0.5\,SD]$	$(-0.5\,SD, 0.5\,SD]$	$(0.5\,SD, 1.5\,SD]$	$(1.5\,SD, \infty)$
0-15%	0 (0%)	0 (0%)	0 (0%)	1 (0.26%)	4 (1.04%)
15.1-30%	0 (0%)	4 (1.04%)	60 (15.62%)	55 (14.32%)	27 (7.03%)
30.1-60%	0 (0%)	72 (18.75%)	99 (25.78%)	6 (1.56%)	0 (0%)
60.1+%	18 (4.69%)	36 (9.38%)	2 (0.52%)	0 (0%)	0 (0%)

Tobler's First Law states that nearer things are more related than distant things, and the spatial patterning of zero-car households with 1 worker is no exception. Figure 6 shows the estimates and CVs for zero-car households with 1 worker. Census tracts with well above average estimates of zero-car households with 1 worker are located in a belt extending east to west from Old City through Center City and into West Philadelphia. These are the same census tracts where CVs consistently do not exceed 30%.

Depending on one's research purpose and area of interest, data that appears unreliable at face value may be sufficient, and even well-suited, for analysis. A data user with a particular interest in Center City Philadelphia would find reliable tract-level data even when using a cross-tabulation.

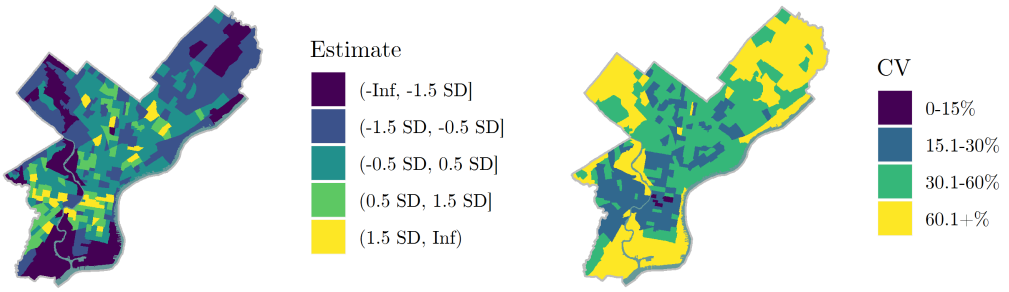

Fig. 6. Estimated zero-car households with 1 worker and their CVs for Philadelphia's census tracts. Source: CTPP Table A112310.

Local context: A practical example As the Metropolitan Planning Organization for Greater Philadelphia, DVRPC administers funds for transportation improvements throughout the region. To determine which projects receive funding, DVRPC scores each project using a series of quantitative evaluation criteria. One criterion is the Indicators of Potential Disadvantage (IPD) Score. The IPD score measures concentrations of nine population groups of interest under Title VI and environmental justice mandates and can range from 0 to 36, where 36 indicates population concentrations that are well above average in all nine groups.

Transit and transportation projects that have a higher IPD score receive higher priority according to DVRPC's project funding criteria.

Survey-based estimates are well-suited to IPD analysis. Census tracts with higher counts of populations of interest—for example, foreign-born individuals—tend have both higher concentrations and lower CVs. Because these areas have higher concentrations, they receive higher IPD scores, and because they receive higher IPD scores, projects in these areas are prioritized for transportation funding. In effect, the tracts that are prioritized for transportation funding are also the areas where data reliability is highest. Figure 7 provides evidence for the link between IPD scores and data reliability.

Simultaneous changes in table type and geography An analysis of 209 total tables and 1,966 total variables across five geographies and all CTPP table types indicates that residence-based tables have consistently smaller CVs at smaller geographies; and that flows tables have consistently larger CVs, even at large geographies, and should be used with caution.

CVs were computed for all CTPP residence-based, workplace-based, and flows tables without cross-tabulations and with data available at the county, PUMA or POWPUMA, TAD, Census Tract, and TAZ geographies. See Tables 11 through 13 in the Appendix for a list of tables included in analysis. Table 8 displays the distribution of CVs by table type and geography. Note that tract-to-tract and TAZ-to-TAZ flows were not computed.

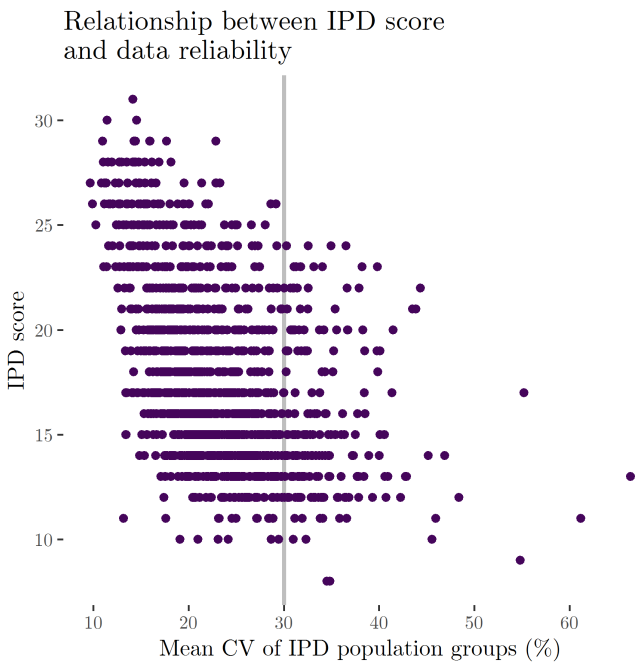

Fig. 7. Census tracts with lower CVs tend to have higher IPD scores. The gray vertical line indicates a CV threshold of 30%. Source: 2017 DVRPC Indicators of Potential Disadvantage.

Table 8. CVs by table type and geography for the nine-county DVRPC region. Tract-to-tract and TAZ-to-TAZ flows are not computed. Note that PUMAs and POWPUMAs are not spatially identical; see Figure 2.

CV	**Residence-based**	**Workplace-based**	**Flows**
County			
0-15%	1,496 (90.34%)	1,336 (89.42%)	2,224 (38.13%)
15.1-30%	104 (6.28%)	93 (6.23%)	984 (16.87%)
30.1-60%	38 (2.30%)	45 (3.01%)	931 (15.96%)
60.1+%	18 (1.09%)	20 (1.34%)	1,693 (29.03%)
Total obs.	1,656	1,494	5,832
PUMA/POWPUMA			
0-15%	5,932 (78.63%)	2,856 (86.02%)	5,876 (9.95%)
15.1-30%	886 (11.74%)	291 (8.77%)	5,480 (9.28%)
30.1-60%	432 (5.73%)	109 (3.28%)	7,780 (13.18%)
60.1+%	294 (3.90%)	64 (1.93%)	39,904 (67.59%)
Total obs.	7,544	3,320	59,040
TAD			
0-15%	9,537 (67.31%)	6,323 (49.47%)	9,401 (2.17%)
15.1-30%	2,528 (17.84%)	3,458 (27.05%)	21,507 (4.97%)
30.1-60%	1,154 (8.15%)	1,883 (14.73%)	43,616 (10.09%)
60.1+%	949 (6.70%)	1,118 (8.75%)	357,908 (82.77%)
Total obs.	14,168	12,782	432,432
Census Tract			
0-15%	55,863 (22.02%)	19,574 (8.21%)	—
15.1-30%	46,690 (18.40%)	29,906 (12.55%)	—
30.1-60%	64,808 (25.54%)	54,880 (23.03%)	—
60.1+%	86,375 (34.04%)	133,984 (56.22%)	—
Total obs.	253,736	238,524	
TAZ			
0-15%	70,665 (12.23%)	20,972 (4.02%)	—
15.1-30%	72,091 (12.48%)	46,631 (8.95%)	—
30.1-60%	121,142 (20.97%)	87,612 (16.81%)	—
60.1+%	313,862 (54.32%)	366,025 (70.22%)	—
Total obs.	577,760	521,060	

3 Recommendations

3.1 Data selection

Table type A data user's choice of table type is often driven by the research question; one often cannot default to a CTPP residence-based table simply because it has smaller CVs. Caution is advised when working with workplace-based and commuting flows tables since these tables have larger CVs in general.

Data from LODES or other sources can be used to verify conclusions derived from CTPP data. Consider the bicycle planning scenario discussed in Section 2.1. Given the CVs in the tract-to-tract commuting flows data by mode, it is impossible for the transportation planner to definitively say which origin tracts had the most bicyclists. Supplementing the CTPP data with existing bicycle counts along each tract-to-tract route could help to identify priority areas.

Geography When working with survey data like the CTPP or the ACS, select the largest useful geography to answer the research question. The reliability penalties of moving from TAD to tract and tract to TAZ geographies are substantial.

Variable detail Resist the appeal of detailed cross-tabulations unless they are needed to answer the research question. Adding even a single cross-tabulation can cause median CVs to double, regardless of the geography.

Local context Data reliability can vary across space with changes in sample counts, population densities, and urban-rural distinctions. For each variable, it is advisable to look at maps of its estimates and CVs side by side and to ask whether the data is reliable enough for one's research needs.

3.2 Cartographic choices

There are least two ways to consider CVs in presenting CTPP data visually. The first is to denote observations with CVs large enough to warrant cautious interpretation (Francis et al., 2012). Francis et al. begin with typical choropleth maps, where a higher value or concentration of a variable is indicated by a darker color on the map. The authors denote the magnitude of the CVs by adding a layer of varying dot and line patterns on top of the choropleth layer, where the dot and line patterns correspond to the data reliability. Observations with CVs above 60% are blacked out entirely.

Another way to consider CVs in mapping is to use the distribution of observations and their CVs to choose an optimal map classification scheme, as demonstrated by Alvarez and Salvo at the New York City Department of City Planning (DCP) (Alvarez & Salvo, 2017). DCP staff created a Map Reliability Calculator in MS Excel (New York City Department of City Planning, n.d.-b) to simplify the process for data users, and an interactive online reliability calculator is available at https://aplarson.shinyapps.io/MapClassificationAutoreporter/.

3.3 Data aggregation methods

Traditional methods Two ways to reduce CVs include aggregating observations to form larger geographies and collapsing subfields. Both methods tend to increase the sample size and the confidence about each estimate, thereby reducing CVs.

Arriving at new CVs from one of these two methods requires three steps: first, aggregate observations or collapse subfields; second, compute the new MOE from the Census Bureau's estimation formulas (United States Census Bureau, 2018) or Variance Replicate tables; and third, use the new estimate and MOE values to derive the new CVs.

Neighborhood Tabulation Areas Selecting which observations to aggregate into larger geographies can be difficult to execute and to justify. The New York City DCP simplified the process for data users in New York City and set an example for other cities by creating a custom geography aggregated from census tracts. The DCP initially created Neighborhood Projection Areas to forecast population change from 2000 to 2030 (New York City Department of City Planning, n.d.-a); since then, Neighborhood Projection Areas have been rebranded as Neighborhood Tabulation Areas (NTAs) and used more broadly in demographic analysis. NTAs are aggregations of census tracts that are smaller than PUMAs, do not cross PUMA boundaries, and reflect the general understanding of New York City neighborhoods (Donnelly, 2018; Tsao, Reilly, & Zhilkova, 2017). The NTA geography is small enough to address many research questions and large enough to greatly improve on census tract reliability.

Data-driven regionalization Advanced data users who are interested in creating a custom aggregate geography can harness the similarities between neighboring geographies to improve data reliability. Spielman and Folch implemented a spatial optimization algorithm in an open-source software environment that aggregates spatially contiguous observations. The algorithm maximizes intra-region similarity along one or several variables and meets a user-defined reliability threshold while also keeping regions as small as possible (Spielman & Folch, 2015). The optimal regions created by data-driven regionalization are a more modern analogue of Neighborhood Tabulation Areas.

References

Alvarez, J., & Salvo, J. (2017). Towards standards in mapping ACS data. 2017 ACS Data Users Conference.

American Association of State Highway and Transportation Officials. (2019a). *2012-2016 5-Year CTPP: Commuting flows tables.* Retrieved 2019-04-12, from https://ctpp.transportation.org/2012-2016-5-year-ctpp

American Association of State Highway and Transportation Officials. (2019b). *2012-2016 5-Year CTPP: Residence-based tables.* Retrieved 2019-04-08, from https://ctpp.transportation.org/2012-2016-5-year-ctpp

American Association of State Highway and Transportation Officials. (2019c). *2012-2016 5-Year CTPP: Workplace-based tables.* Retrieved 2019-04-11, from https://ctpp.transportation.org/2012-2016-5-year-ctpp

Cambridge Systematics. (2017). *Employment data for planning: A resource guide.* Retrieved 2019-04-19, from http://onlinepubs.trb.org/onlinepubs/nchrp/docs/NCHRP08-36(127)_EmployDataGuide.PDF (NCHRP Project 08-36, Task 127)

Census Transportation Planning Products Program Oversight Board. (2018). *Policy change announcement on small and custom geography in CTPP.* Retrieved 2019-04-19, from https://ctpp.transportation.org/policy-change-on-small-geography

Citro, C. F., & Kalton, G. (Eds.). (2007). *Using the American Community Survey: Benefits and challenges.* Washington, DC: National Academies Press.

Donnelly, F. (2018). *Finding NYC neighborhood census data.* Retrieved 2019-04-18, from http://faculty.baruch.cuny.edu/geoportal/resources/census/census_nbhd.pdf

Francis, J., Vink, J., Tontisirin, N., Anantsuksomsri, S., & Zhong, V. (2012). *Alternative strategies for mapping ACS Estimates and error of estimation.* Retrieved 2019-04-18, from https://pad.human.cornell.edu/papers/index.cfm

Hodges, K. (2014). Improving the accuracy of block group data from the American Community Survey. 2014 ACS Data Users Conference.

New York City Department of City Planning. (n.d.-a). *Appendix 16: Neighborhood Tabulation Areas and PUMAs.* Retrieved 2019-04-18, from https://nycplanning.github.io/Geosupport-UPG/appendices/appendix16

New York City Department of City Planning. (n.d.-b). *Map reliability calculator.* Retrieved 2019-04-18, from https://www1.nyc.gov/site/planning/data-maps/nyc-population/geographic-reference.page

Spielman, S. E., & Folch, D. C. (2015). Reducing uncertainty in the American Community Survey through data-driven regionalization. *PLOS One*, *10*(2), e0115626.

Tsao, T.-Y., Reilly, K., & Zhilkova, A. (2017). Neighborhood Tabulation Areas: Triangulating the ACS with other data at the small area level to enhance population health improvement capacity in New York City. 2017 ACS Data Users Conference.

United States Census Bureau. (2018). *Understanding and using American Community Survey data: What all data users need to know.* Retrieved 2019-04-20, from https://www.census.gov/programs-surveys/acs/guidance/handbooks/general.html

Appendix

Online Appendix

The Online Appendix (916 pp., 1.31 GB) is available for download at https://drive.google.com/open?id=17KtqVzbFoyWiEsJvffyoYXXGqFdqCdlA.

The Online Appendix summarizes the CVs for all CTPP residence-based tables without cross-tabulations at five geographies, including the county, PUMA, TAD, Census Tract, and TAZ. The study area is DVRPC's nine-county region, including Bucks, Chester, Delaware, Montgomery, and Philadelphia counties in Pennsylvania and Camden, Gloucester, Mercer, and Trenton counties in New Jersey. The document can be used to glean general trends in CTPP reliability along the dimensions of geography, variable detail, and local context.

Structure of the Online Appendix Each page of the Online Appendix represents a single variable in a CTPP table at a given geography. The top lines list the CTPP table number, variable name, and geography.

Beneath the title are four small tables.

- *Table 1* summarizes the estimates of the given variable and geography, including the minimum, median, mean, and maximum.
- *Table 2* summarizes the CVs for the given variable and geography.
- *Table 3* displays the total count of high CVs. Because a "high" CV value can vary based on geography and variable detail, the definition and computation of a high CV is discussed in more detail below.
- *Table 4* groups the observations into four data reliability brackets by CV: 0-15%, 15.1%-30%, 30.1%-60%, and more than 60% (Francis et al., 2012).

The bottom of the page includes two maps. For the given variable and geography, the map on the left-hand side shows the distribution of the CV across the DVRPC region. The ranges of CVs differ drastically, especially between geographies, so the user is encouraged to pay close attention to the map legend when comparing across geographies. The map on the right-hand side indicates observations with high CVs in purple.

When an estimate is 0, it is not possible to compute the CV. In these instances, the CV is assigned a value of 100%.

Defining a high CV The aim of defining high CVs in the Online Appendix is twofold: first, to identify observations with high CVs relative to the study area; and second, to identify observations with CVs that render the data practically unusable. Therefore, there are two possible criteria for an observation to be classified as having a high CV. First, the CV of each observation is compared to the mean CV of its neighbors, which include any observations that share a common side or vertex. This approach is analogous to the Nielsen "Touch Method" to smooth differences between neighboring census block groups (Hodges, 2014). If the percentage difference between the observation and its neighbors is equal to exceeds 40%, then the observation is flagged as an outlier. See the formula below, where i is an observation, j is its neighbor, and PD is the percentage difference.

$$PD = \frac{x_i - \frac{\sum_{j=1}^{n} x_j}{n_j}}{\frac{\left(\frac{\sum_{j=1}^{n} x_j}{n_j}\right)}{2}} \tag{2}$$

To put the percentage difference metric in context, a census tract with a CV of 30% or greater and a mean neighbor CV of 20% would be flagged as an outlier; because the tract has a high CV relative to its neighbors, it may merit further investigation. The percentage difference metric can flag observations that have high CVs compared to their neighbors even for variables with high overall reliability, such as total population.

By contrast, many variables in the Online Appendix have low overall reliability. Using only the percentage metric, if CVs of 80% were uniformly distributed across the study area, no observations would be flagged as outliers. The second criterion addresses these situations by flagging observations when their CVs exceed a geography-specific threshold as shown in Table 9.

Observations with CVs that meet either or both criteria are flagged as having high CVs.

Table 9. CV reliability thresholds by geography.

Geography	Threshold
County PUMA TAD	30%
Census Tract TAZ	60%

Source Tables

Table 10. Input tables analyzed in comparisons of CTPP table types. All share the same universe of workers 16 years and over.

Description	Residence	Workplace	Flows
Total workers	A102101	A202100	A302100
Age of worker	A102102	A202101	B302101
Industry	A102105	A202104	B302102
Means of transportation	A102106	A202105	A302103
Travel time to work	A102110	A202113	B302106

Table 11. Residence-based input tables analyzed at the county, PUMA, TAD, census tract, and TAZ geographies. The analysis includes 23 tables and 184 variables per geography.

Table ID	**Description**	**Universe**
A101100	Total population	All persons
A101102	Urban/rural residence	All persons
A101103	Hispanic origin	All persons
A101104	Length of US residence	All persons
A101108	Race	All persons
A101109	School enrollment	All persons
A102101	Total workers	Workers 16 years and over
A102104	Earnings in the past 12 months (2016$)	Workers 16 years and over
A102105	Industry	Workers 16 years and over
A102106	Means of transportation	Workers 16 years and over
A102107	Occupation	Workers 16 years and over
A102108	Time leaving home	Workers 16 years and over
A102109	Usual hours worked per week	Workers 16 years and over
A102110	Travel time to work	Workers 16 years and over
A103100	Total workers in households	Workers 16 years and over in households
A111102	Vehicles available	Occupied housing units
A112100	Total households	Households
A112101	Number of persons under 18 in household	Households
A112106	Household size	Households
A112109	Number of workers in household	Households
A113100	Poverty status of household	Households for which poverty status is determined
B111103	Aggregate number of vehicles available in households	Households
B112108	Mean number of persons per household	Households

Table 12. Workplace-based input tables analyzed at the county, PUMA, TAD, census tract, and TAZ geographies. The analysis includes 14 tables and 166 variables per geography.

Table ID	Description	Universe
A202100	Total workers	Workers 16 years and over
A202101	Age of worker	Workers 16 years and over
A202103	Earnings in the past 12 months (2016$)	Workers 16 years and over
A202104	Industry	Workers 16 years and over
A202105	Means of transportation	Workers 16 years and over
A202106	Occupation	Workers 16 years and over
A202107	Hispanic origin	Workers 16 years and over
A202110	Race	Workers 16 years and over
A202111	Sex	Workers 16 years and over
A202112	Time arriving	Workers 16 years and over
A203100	Total workers in households	Workers 16 years and over
B203101	Household income in the past 12 months (2016$)	Workers 16 years and over in households
B203102	Vehicles available	Workers 16 years and over in households
B207100	Workers per car, truck, or van	Workers 16 years and over in households

Table 13. Commuting flows input tables analyzed at the county, PUMA, and TAD geographies. The analysis includes 8 tables and 72 variables per geography.

Table ID	Description	Universe
A302100	Total workers	Workers 16 years and over
A302103	Means of transportation	Workers 16 years and over
B302102	Industry	Workers 16 years and over
B302104	Time leaving home to go to work	Workers 16 years and over
B302105	Minority status	Workers 16 years and over
B302106	Travel time to work	Workers 16 years and over
B303100	Household income in the past 12 months (2016$)	Workers 16 years and over in households
B304100	Poverty status	Workers 16 years and over in households