Disclosure and Utility of Census Journey-to-Work Flow Data from the American Community Survey

Is There a Right Balance?

by

Ed Christopher
FHWA Resource Center
19900 Governors Drive
Olympia Fields, Illinois 60461
708-283-3501 (fax)
708-574-8131 (cell)
708-283-3534 (voice)
edc@edthefed.com

Nandu Srinivasan
Cambridge Systematics Inc.
FHWA, HPPI (Room 3306)
400 7th Street SW
Washington DC 20590
202-366-7742 (fax)
202-366-5021 (voice)
nanda.srinivasan@fhwa.dot.gov

This paper was developed to augment the display poster prepared for the Conference on Census Data for Transportation Planning: Preparing for the Future. The opinions and views expressed in this document and subsequent poster represent those of the authors (and those who have influenced them) but should not be considered the views, policy positions or in no way be attributed to the organizations for which they work or have any affiliations.

Irvine, California
May 11 to 13, 2005

Abstract

Early in 2003 the transportation community contracted with the Census Bureau to produce the CTPP2000, a special tabulation. A special tabulation is made up of user defined tables and falls outside the "standard" products distributed by the Census Bureau like SF1, SF3, and PUMS. With the 2000 decennial data, the Census Bureau required all special tabulations to have disclosure avoidance techniques applied to them. For CTPP2000 this meant the institution of rounding and threshold techniques in addition to the already applied procedures of data swapping and imputation.

The specific disclosure rules for the American Community Survey after 5 years of data collection are likely to be similar, if not stricter than to those used for CTPP2000. In this paper the effects of rounding and thresholds on the CTPP will be exposed along with an examination of their effects under the American Community Survey. CTPP2000, ACS, 1990 CTPP and the NCHRP 8-48 data sets are used in this analysis.

We show how the rounding rules cause an undercount in the published datasets. The rounding rules for CTPP2000 could have worked better had the underlying data been more closely examined for the frequency of occurrence of cell values before the rounding decision was made. Finally, we show that a minor tweaking of the rules could have produced a more consistent dataset.

As for thresholds, they will always cause severe data loss even at a medium level of geographic aggregation, let alone for small geography. Compounding the severe data loss, consider that the number of observations in a 5 year accumulated ACS will be at least 25 percent smaller than those collected from the decennial census.

Acknowledgments

The authors would like to thank Elaine Murakami, Siim Soot and Mary Kay Christopher for their inspiration, patience and assistance in reviewing and discussing this document as it evolved.

1.0 Introduction

Journey-to-Work (JTW) data or the Census Transportation Planning Package (CTPP) has been around since the 1960 decennial census (1). The CTPP is a special tabulation with the States and Metropolitan Planning Organizations (MPOs) paying for the product (2).

Having worked with 4 previous JTW data sets, the transportation community was unprepared when its CTPP2000 table request was subjected to limitations imposed by the Census Bureau (CB) Disclosure Review Board (DRB). One main DRB objection was to the Part 3 or "flow" data. Initially the DRB said that only flows with 50 or more unweighted records could be released. After negotiation, the complexity of the tables requested were reduced, some tables eliminated, and the threshold requirement was reduced to 3 un-weighted records. Another concern of the DRB was having unique zones that did not fully nest within the existing census geography of Blocks or Block Groups. The DRB characterized this concern as "slivering" and required all the CTPP tables to be rounded regardless of geography. Believing these restrictions would not compromise the quality and use of the data, the American Association of State Highway Transportation Officials (AASHTO) entered into a contractual relationship with the CB for the provision of CTPP2000.

Two disclosure avoidance techniques were applied to CTPP2000. First, all the CTPP 2000 tables except for those containing means, medians, and standard deviation values were rounded. The rounding rules were simple.

- Values of zero would remain zero.
- Values between 1 and 7 would be rounded to 4.
- values of 8 or more would be rounded to the nearest multiple of 5.

The second disclosure avoidance technique was to apply a threshold rule to the Origin-Destination (OD) worker flows tables. The threshold rule stated that no data would be provided for any OD pair that had 3 or less records (worker flows) before weighting.

Exhibit 1.1: Disclosure Avoidance Rules for CTPP 2000

Part 1: at Residence (121 Tables)
All Tables Rounded
Zero = 0
1 through 7 = 4
8 though ∞ = Nearest Multiple of "5"

Part 2: at Workplace (68 Tables)
All Tables Rounded

Part 3: Worker Flows (14 Tables)
All Tables Rounded
Some Tables with Thresholds

Exhibit 1.1 summarizes the disclosure avoidance rules for CTPP2000. As can be seen, not all the Part 3 tables would be subject to thresholds. During the negotiations with the DRB a decision was made to release two tables without threshold suppression; Table 3-01, Total Worker Flows, and Table 3-02 or the Vehicles Available per Household (3) by Means of Transportation to Work (7). Exhibit 1.2 shows the Part 3 tables that were subject to thresholds and those that were not. Noteworthy is that Tables 3-08 to 3-14 were exempt from both rounding and thresholds since they fell under the CB "normal" process for reporting aggregates, means, medians and standard deviations.

Exhibit 1.2: Part 3 Worker Flow Tables

Table	Content
1	Total Workers (1)
2	Vehicles Available (3--zero,one or two+) by Means of Transportation (7 modes)
3	Poverty Status (3 categories)
4	Minority Status (2--white non-hispanic and all oth
5	Household Income (8 classifications)
6	Means of Transportation (17 modes)
7	Household Income (4 classifications) by Means of Transp
8	Mean Travel Time by Means of Transportation to Work (7 modes) and Time Leaving Home for Work (2--AM peak and all other times)
9	Median Travel Time by Means of Transportation to Work (7 modes) and Time Leaving Home for Work (2 groupings)
10	Aggregate Number of Vehicles by Time Leaving Home for Work (2, see table 8)
11	Number of Workers per Vehicle by Time Leaving Home for Work (2, see table 8)
12	Aggregate Number of Carpools by Time Leaving Home for Work (2, see table 8)
13	Number of Workers per Carpool by Time Leaving Home for Work (2, see table 8)
14	Aggregate Travel Time by Means of transportation to work (7 modes) and Time Leaving Home for Work (2, see table 8)

No record threshold

Must have 3 unweighted records

Now that the CTPP2000 data has been released, users are just beginning to analyze and understand the full effects of the DRB restrictions. The remainder of this paper will review and explore the impact of those restrictions

2.0 Rounding

All the CTPP2000 tables except for those containing means, medians, and standard deviation values were rounded. The method, rounding the values between 1 and 7 to 4 was first dubbed the "Rule of Four-Seven" but was later shortened to the "Rule of Seven" by the transportation community.

Mechanically, each cell of each table is rounded independently of the other cells. This means that the totals are rounded independently from the other values in the table. We call this "row rounding". The example in Exhibit 2.1 shows how the rounding would work using 1990 unrounded values and applying the 2000 rules. The thing to notice is that the 1990 total of 352 is rounded separately to 350 and not to the sum of the rounded values or 354 and then 355.

Exhibit 2.1 How Rounding Works

Mode to Work	Circa 1990	For 2000 (ROUNDED)
Total	352	350 (not 355!)
Drive Alone	212	210
Carpool	46	45
Transit	59	60
Walk	33	35
Bike	2	4
		True Total 354

To analyze the effect of the DRB rounding rules we took 1990 un-rounded data and applied the 2000 rounding rules. To see how Summary Levels may be affected, we looked at un-rounded and rounded data across Traffic Analysis Zones (TAZs), Tracts and Block Group (BGs). We were especially concerned because many MPOs were telling us about data losses while others were complaining that the "numbers don't add up".

The first step was to select a CTPP part and universe for analysis. Because of the importance of the worker (commuter) flows on transportation planning and a greater likelihood of values less than 7 occurring in the OD data, we chose to use the flow data or Part 3 from 1990. In terms of the universe we limited the analysis to those commuters (resident workers) who lived in each of the three regions while excluding those workers who worked at home. This universe was used to minimize computer processing time and to simplify the programming.

Exhibit 2.2 Study Areas Used for Rounding Analysis

Chicago **Traffic Analysis Zones**	**Los Angeles** **Census Tracts**	**Boston** **Block Groups**
9-Counties 1990 Population: 7,429,181 Area (sq. miles): 137 Number of zones: 14,127 People per zone: 526 Resident workers: 3,563,603 Work place workers: 3,635,769 Workers at home: 76,371 Total households: 2,675,257 Counties include: Cook, DuPage, Grundy, Kane, Kankakee, Kendall, Lake, McHenry, and Will	6-Counties 1990 Population:14,640,832 Area (sq. miles): 578 Number of Tracts: 3,934 People per Tract: 3,722 Resident workers: 6,844,948 Work place workers: 6,849,916 Workers at home: 187,091 Total households: 4,942,075 Counties include: Imperial, Los Angeles, Orange, Riverside, San Benardino and Ventura	Counties (see below) 1990 Population: 4,056,947 Area (sq. miles): 809 Number of BGs: 3,850 People per BG: 1,054 Resident workers: 2,073,508 Work place workers: 2,201,473 Workers at home: 50,989 Total households: 1,507,077 Counties include: All MCDs in 1990 Boston definition including parts of Middlesex, Essex Worcester, Suffolk, Norfolk, Bristol and Plymouth

The next task was to apply the 2000 rounding rules and examine its effect. Several preliminary studies with CTPP2000 data showed worker losses in the neighborhood of 3 to 5 percent associated with rounding. To identify the data loss in any region all one has to do is to sum the commuter trips from Table 3-01 at the county to county level and compare it to the number of commuter trips at lower levels of geography like Tracts, BGs or TAZs. For example, for the San Francisco region, Chuck Purvis reported a rounding data loss of 3.5 percent when moving from county to county to TAZ data (3). For many of those working with Part 3 data, examining the commuter trips lost is one of the first checks performed.

Exhibit 2.3 shows the number and percent of commuter trips without and with the "Rule of Seven". Note that the data loss in our 1990 example is in the neighborhood of two to four percent. This is very consistent with the data losses others around the country have been reporting.

Exhibit 2.3 Work Trip Commuters Lost due to Rounding

Area and Summary Level	Rounding Rule of Seven	Total Commuters	Lost Commuters	Percent Lost
Chicago, IL (TAZ)	Without	3,487,232	0	0.00
	With	3,342,963	144,269	4.14
Los Angeles, CA (Tracts)	Without	6,657,857	0	0.00
	With	6,505,471	152,386	2.29
Boston, MA (Block Groups)	Without	2,022,519	0	0.00
	With	1,941,612	80,907	4.00

Source: 1990 CTPP data for Commuters who lived in region, excludes workers who worked at home.

Following this preliminary analysis, the commuters were summed by the number of trips per OD pair (Exhibit 2.4). The distributions are rather consistent across summary levels. Well over 50 percent of the trips occur between OD pairs with less than 10 trips. Zonal pairs with 7 and less trips account for anywhere between 34 and 44 percent of all the trips and 4 trips per OD pair is obviously nowhere near the mid point of the distribution of commuters.

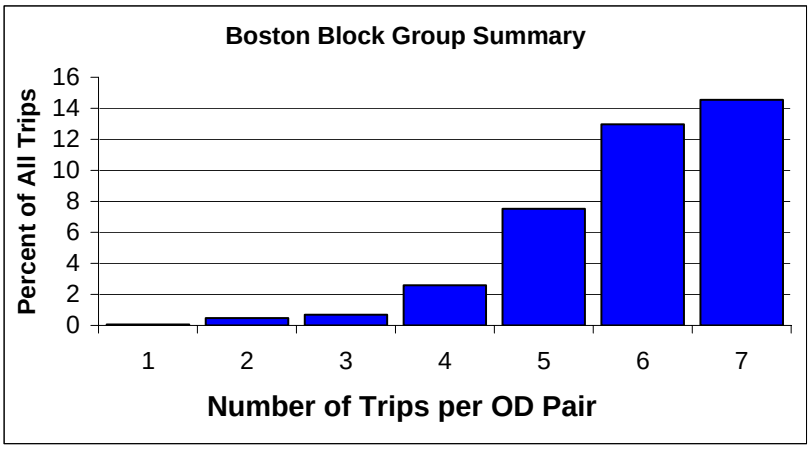
Exhibit 2.4 Number and Percent of Trips per OD Pair

Number of Trips per OD Pair	Chicago -- TAZ			Los Angeles -- Tract			Boston -- BG		
	Number of Workers	Percent	Cum Percent	Number of Workers	Percent	Cum Percent	Number of Workers	Percent	Cum Percent
1	1,075	0.3	0.3	377	0.1	0.1	110	0.1	0.1
2	9,227	2.7	3.0	1,727	0.4	0.5	863	0.5	0.5
3	6,372	1.9	4.9	4,278	1.0	1.5	1,248	0.7	1.2
4	10,825	3.2	8.0	13,161	3.2	4.7	4,711	2.6	3.8
5	29,259	8.5	16.6	30,138	7.3	12.0	13,696	7.5	11.3
6	47,016	13.7	30.3	44,779	10.8	22.8	23,588	13.0	24.3
7	**50,173**	**14.6**	**44.9**	**46,764**	**11.3**	**34.0**	**26,467**	**14.5**	**38.8**
8	43,484	12.7	57.6	42,957	10.3	44.4	24,073	13.2	52.1
9	33,459	9.8	67.4	34,907	8.4	52.8	18,914	10.4	62.5
10+	111,921	32.6	100.0	196,098	47.2	100.0	68,278	37.5	100.0
	342,811	100		415,186	100		181,948	100	

Source: 1990 CTPP data for Commuters who lived the region, excludes workers who worked at home

From this simple analysis it is clear that the DRB decision to round values between 1 and 7 to 4 caused an underestimate. This is because values of 5, 6 and 7 trips per OD pair are far more common than 1, 2, or 3 trips. Exhibit 2.5 clearly shows this using BG data from the Boston Area. It is at this juncture that some have wondered if the DRB ever took into consideration the weighting and expansion process used by the CB. This notion should be a topic for further study.

Exhibit 2.5 Percent of trips between OD pairs with 1 through 7 trips



Assuming that the CB had some statistical reason for choosing seven as the upper bound for rounding, we set out to determine if there was an optimum value to round to. To do this, we needed to determine what percent of trips would represent the midpoint

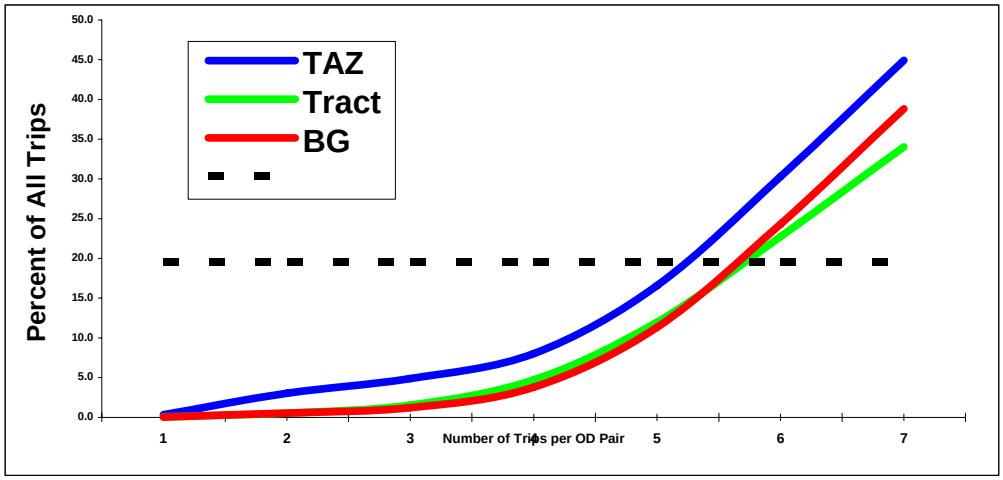
of all the trips occurring between OD pairs with 7 or less trips. To minimize the effect of summary levels we averaged the data from three areas together. The mathematics of the process was to take the cumulative percent values representing seven or less trips per OD pair, find the simple average and then its midpoint.

((44.9 +34.0+38.8) / 3) /2 = 19.6

We also calculated the weighted average across the three areas which incidentally, turned out to be 19.5 percent which was relatively close to our simple average of 19.6 percent.

What this told us is if 7 is our upper bound of trips per OD pair for rounding, we should be looking for a value to round to which represents approximately 19.6 percent of trips. Looking at Exhibit 2.4 and the cumulative percentage column it is easy to see that 19.6 consistently falls between OD pairs with 5 and 6 trips. Exhibit 2.6 shows this graphically. Can you find the midpoint? It is around 5.49 trips per OD pair.

Exhibit 2.6 Graphical Representation Depicting the Midpoint of the Number of Trips per OD Pairs with Seven or Less Trips by Geographical Summary Level



Given our analysis, the DRB could have minimized a systematic undercount in the data by rounding to 5. Not only would 5 have helped eliminate the undercount bias, it is also a rounded number that people are used to seeing.

One big reason for the concern about the undercount is because of a tendency in the transportation field to aggregate zonal data together depending upon the analysis or study at hand. While there is “no fix” for this, it is instructive for users to be aware of this undercount when working with the data. In Appendix A is a more detailed discussion of the impacts of rounding which occurred on the CTPP list serve along with some tips for users working with this data.

3.0 Thresholds

The second major area we examined was the potential effect of the ”Threshold” rule on the ACS data. Specifically, we compared commuter flows from CTPP2000 and the ACS without and with threshold suppression. The ACS test sites used for this analysis include; Pima County in Arizona (Tucson), Douglas County in Nebraska (Omaha) and Franklin County in Ohio (Columbus).

The effects of thresholds were first reported by Wende Mix in a report commissioned by the Federal Highway Administration in 2003 (4) and later by Elaine Murakami in a CTPP2000 Status Report newsletter article in 2004 (5). Both authors alerted users to the potential of lost trips and OD pairs with ACS data due to thresholds.

Our threshold analysis compared data from CTPP2000 with ACS data taken from the three-year ACS test site data prepared for NCHRP 8-48. The NCHRP 8-48 data base consisted of a special tabulation of the ACS data. The special tabulation contained 1999-2001 ACS along side Census 2000 long form data for 9 of the 36 test ACS counties. It includes a small subset of the CTPP 2000 tables. The intent of the special tabulation was to allow for some side by side comparisons ACS and CTPP data. Exhibit 3.1 depicts the counties included in the NCHRP 8-48 special tabulation along with their sampling and response percentages.

Originally, we believed that the counties in the ACS 3-year test data were sampled at rates that approximate the same number of observations that would be available from accumulating 5 years of ACS data. However, as Exhibit 3.1 shows not only were the sampling rates slightly different between areas but there is a rather large difference in the percent of the population who completed the ACS as compared to the decennial (CTPP) forms. As will be seen, the difference in completed survey responses compounds the negative effects of thresholds for small area data.

Exhibit 3.1 ACS Urban Test Counties in NCHRP 8-48 Data Base

County and State Name	Most Detailed Geography	Percent of Housing Units Sampled		Percent of Population Responding	
		CTPP	ACS	CTPP	ACS
Pima AZ	**TAZ**	**12.5**	**13.4.**	**12.7**	**8.6**
San Fran. CA	TAZ	11.7	9.6	11.8	5.5
Broward FL	TAZ	11.7	9.5	11.5	5.9
Lake IL	TAZ	14.3	10.3	14.4	6.6
Hampden MA	Tract	13.4	14.6	13.5	9.4
Douglas NE	**TAZ**	**13.9**	**15.2**	**13.9**	**10.5**
Bronx NY	Tract	11.3	10.2	11.6	4.4
Franklin OH	**Tract**	**14.1**	**9.4**	**14.1**	**6.2**
Multnomah OR	Tract	14.1	15.0	14.0	10.0

Note: Study areas for Threshold analysis are bolded.
Source: NCHRP 8-48 test data set tables.

The counties used for the NCHRP tabulation were selected because their population exceeds 400,000 so that small area geography, TAZs and Tract data, were available.

To make the CTPP and ACS tables within the NCHRP data base somewhat comparable, Group Quarters (GQ) data were removed from the CTPP tables. This was done because the original ACS test sites did not include GQ. Also, because the ACS sample was restricted to residents of a particular test county, the workplace and flow tabulations were similarly restricted. That is, unlike CTPP2000 where the Part 2 tables include all workers who work in a county, no matter where they live; the ACS tables were limited to only those people who both live and work in the selected county. Unfortunately, the rounding rules applied to the ACS test county data and the CTPP decennial data were different. The decennial data was rounded to the nearest 10 while the ACS data used the "rule of seven".

Another small difference in the CTPP and ACS data is due to something called "Extended Allocation" (EA). When geocoding a workplace location, not all responses can be coded to a TAZ or BG. Many times the individual completing the questionnaire gives an incomplete address and legally, the CB is only required to code workplaces to the place level. However, because of the importance of individual trips at the smallest geography possible, TAZs or BGs, a process of imputing or allocating place level data was implemented for CTPP. Ed Limoges, retired from the Detroit MPO, was contracted by the CB with a portion of the AASHTO pooled fund money to develop the process. EA is more fully discussed in (6).

When considering the effect of EA on thresholds, many believe that it helped to add OD pairs in the decennial CTPP data because it helped to increase the number of zonal pairs with less than 3 trips. There are others however, who suggest that by the very nature of the process it only increased the number of trips for existing OD pairs which more than likely met the threshold criteria therefore minimizing any effect on thresholds. EA was applied only to the CTPP data and not the ACS data.

Although we fully intended to use both CTPP and ACS data from the NCHRP data set we had to use "regular" CTPP data with the NCHRP ACS data. The main reason was to ensure that the rules of rounding were consistent. In using the "regular" CTPP data meant that we would have a slight difference in our universes. The ACS data did not contain workers in Group Quarters while the CTPP did. Given that we are comparing the data loss in CTPP against CTPP total workers, and the data loss for ACS against ACS totals, the methodology is valid. Exhibit 3.2 shows a side by side comparison of the key differences in the two data sets used.

Exhibit 3.2 Comparison of Key Data Issues in the Analysis Data Sets

Key Data Issues	ACS	CTPP
Rounding Rules	Same	Same
Group Quarters	No	Yes
Threshold Rules	Same	Same
Extended Allocation	No	Yes
Housing Units Sampled	12.7%	13.5%
Population Responding	8.4%	13.6%

Note: 'Housing units sampled' and 'population responding' represent simple un-weighted averages of the three areas used in the analysis.

Exhibit 3.3 shows the results of the comparison of the three areas without and with thresholds. To compare the CTPP and ACS without thresholds we used Table 3-01, "Commuter Flows for Total Workers" from CTPP2000 and the corollary ACS table from the NCHRP data base. For thresholds we used table 3-06 from the CTPP along with table 3-03 from the NCHRP data set. Although the table numbers are different, both tables measured the same commuter flows except for the small difference in GQ discussed above. To make the CTPP data consistent with the ACS data we also restricted the universe to those workers who both lived and worked in the county under study.

Exhibit 3.3 Detailed Examination of Workers Lost due to Rounding and Thresholds

Franklin County (Columbus, OH)

Total Workers Living and Working in the County (Census 2000) = 508,393

		County-County	Place-Place	Tract-Tract	Zone-Zone
CTPP2000	Table 3-01 (No Thresholds)	508,395	508,361	500,426	487,979
	Percent Loss	0.00%	0.01%	1.57%	4.02%
	Table 3-06 (Thresholds)	508,395	507,604	358,170	177,643
	Percent Loss	0.00%	0.16%	29.55%	65.06%
Total Workers Living and Working in the County (ACS, 3-yr) = 498,220					
ACS (1999, 2000 and 2001)	Table 3-01 (No Thresholds)	498,220	498,168	447,446	na
	Percent Loss	0.00%	0.01%	10.19%	
	Table 3-03 (Thresholds)	498,220	495,840	233,920	na
	Percent Loss	0.00%	0.48%	53.05%	

Douglas County (Omaha, NE)

Total Workers Living and Working in the County (Census 2000) = 213,642

		County-County	Place-Place	Tract-Tract	Zone-Zone
CTPP2000	Table 3-01 (No Thresholds)	213,640	213,655	211,565	209,315
	Percent Loss	0.00%	-0.01%	0.97%	2.03%
	Table 3-06 (Thresholds)	213,640	213,640	157,334	109,247
	Percent Loss	0.00%	0.00%	26.36%	48.86%
Total Workers Living and Working in the County (ACS, 3-yr) = 209,970					
ACS (1999, 2000 and 2001)	Table 3-01 (No Thresholds)	209970	209,970	190,287	190,145
	Percent Loss	0.00%	0.00%	9.37%	9.44%
	Table 3-03 (Thresholds)	209,970	209,960	124,103	79,665
	Percent Loss	0.00%	0.00%	40.89%	62.06%

Pima County (Tucson, AZ)

Total Workers Living and Working in the County (Census 2000) = 359,296

		County-County	Place-Place	Tract-Tract	Zone-Zone
CTPP2000	Table 3-01 (No Thresholds)	359,295	359,281	357,695	354,566
	Percent Loss	0.00%	0.00%	0.45%	1.32%
	Table 3-06 (Thresholds)	359,295	358,204	264,146	126,218
	Percent Loss	0.00%	0.30%	26.48%	64.87%
Total Workers Living and Working in the County (ACS, 3-yr) = 354,130					
ACS (1999, 2000 and 2001)	Table 3-01 (No Thresholds)	354,130	354,164	314,781	316,878
	Percent Loss	0.00%	-0.01%	11.11%	10.52%
	Table 3-03 (Thresholds)	354,130	352,635	197,924	87,319
	Percent Loss	0.00%	0.42%	44.11%	75.34%

Notes: CTPP 2000 includes All Workers Who Lived and Worked in the County, including worked at home
ACS includes Workers in households Who Lived and Worked in the County, including worked at home

Source: CTPP 2000 Part 3 and NCHRP 8-48 ACS test data.

Exhibit 3.3 shows that the application of thresholds causes severe loss of commuters in both decennial CTPP and ACS data. As one would expect, the effects of thresholds are more pronounced with ACS because fewer people responded compared to the decennial data. The more people you have responding to your survey, the greater the likelihood you will have more OD pairs and travelers.

When reviewing Exhibit 3.3 keep in mind that data loss in Table 3-01 (without thresholds) is due primarily to rounding. Knowing this, makes it possible to subtract the effect of rounding from Tables 3-06 and 3-03 and see the general impact of thresholds.

Exhibit 3.4 shows the losses attributed primarily to thresholds for. Surprisingly, what we thought were considerable large losses when working with just CTPP data turned out to be even worse when we looked at ACS.

Exhibit 3.4 Percent of Lost Workers Due to Thresholds Netting Out the Effects of Rounding

Franklin County		
	CTPP	ACS
Place-Place	0.15%	0.47%
Tract-Tract	27.98%	42.88%
BG-BG	61.04%	-----

Douglas County		
Place-Place	-0.01%	0.00%
Tract-Tract	25.38%	31.52%
Taz-Taz	46.84%	52.62%

Pima County		
Place-Place	0.30%	0.43%
Tract-Tract	26.04%	33.00%
Taz-Taz	63.55%	64.82%

Another, more relevant way for transportation planners to look at the data loss is to examine it from the perspective of lost OD pairs. Exhibit 3.5 shows how thresholds can affect the number of OD pairs. Where the threshold data loss looked extreme in terms of commuters, the number of lost OD pairs is even more startling.

At the smallest level of geography upwards of 85 percent of CTPP data OD pairs with data are lost while the ACS losses top out at 90 Percent. The lost OD pairs are solely the result of thresholds. Therefore, we have to conclude that in terms of OD pairing, the decision to apply thresholds has rendered both the CTPP and ACS data useless at TAZ, Tract and even Place levels.

Exhibit 3.5 Summary of OD Pairs Lost Due to Thresholds

Franklin County (Columbus, OH)

	CTPP OD Pairs w/Trips			ACS OD Pairs w/Trips		
	Without Thresholds	With Thresholds	Percent Lost	Without Thresholds	With Thresholds	Percent Lost
Place-Place	384	306	20%	334	229	31%
Tract-Tract	23,289	6,794	71%	13,380	2,459	82%
BG-BG	44,266	5,045	89%	-----	-----	-----

Douglas County (Omaha, NE)

	CTPP OD Pairs w/Trips			ACS OD Pairs w/Trips		
	Without Thresholds	With Thresholds	Percent Lost	Without Thresholds	With Thresholds	Percent Lost
Palce-Place	15	14	7%	15	14	7%
Tract-Tract	8,830	3,044	66%	7,485	2,089	72%
Taz-Taz	14,389	3,081	79%	11,269	1,809	84%

Pima County (Tucson, AZ)

	CTPP OD Pairs w/Trips			ACS OD Pairs w/Trips		
	Without Thresholds	With Thresholds	Percent Lost	Without Thresholds	With Thresholds	Percent Lost
Palce-Place	318	209	34%	270	175	35%
Tract-Tract	13,320	4,644	65%	10,573	2,911	72%
Taz-Taz	26,781	3,179	88%	18,168	1,675	91%

Source: CTPP 2000 Part 3 and NCHRP 8-48 ACS test data.

4.0 Findings and Conclusions

General Effects of Rounding

1. **Produces Inconsistencies Among CTPP Table Values**
2. **Caused a Systematic Undercount of Workers**
3. **Did not Show a Significant Noticeable Difference on Summary Levels**
4. **Rounding to 5 Would Have Been Better**
5. **Was Not Well Received by Users**

The CB decision to subject CTPP2000 to rounding and thresholds has had negative effects on the data. From an analysts perspective it caused inconsistencies between different tables and created the possibility for getting different answers to the same question. While the rounding differences are small, they still exist. Probably the most important effect of rounding was that it produced a systematic undercount of workers that is not easily corrected. More importantly, one is still left wondering if rounding truly met the confidentiality goals of the CB or was it even necessary. To date we have been unable to find any documented cases of disclosure.

General Effects of Thresholds

1. **Eliminates Most OD Pairs and Commuters**
2. **Renders the Flow Data Useless**
3. **Undermines the Utility of Small Area Data**
4. **Was Not Well Received by Users**

Regarding the ACS and thresholds there are two things going on. First it is clear that the notion of applying a threshold to the flow data will undoubtedly cause data to be suppressed. Second, the fact that the number of completed ACS survey forms is lower than the traditional long form compounds the threshold effect. As a result, even more records are lost with the ACS as was seen with the CTPP2000 data.

As ACS data comes online, the transportation planning community will need to decide whether or not to contract with the CB for any small area flow data. Serious questions must be asked about the utility of the data when so many pairs are removed from the data set. One option is to aggregate areas and to use a much courser zone system for tabulating flows. How big should zones be? What would the DRB accept? Are thresholds even solving the perceived problems? These are all topics for further research and discussion. Another topic for research and consideration is the notion of variable zone sizes. For OD pairs can the origin and destination zones different sizes?

When we began this exercise there was some uncertainty about what the ACS disclosure rules would be. However, posted on the CB ACS web site and shown in Attachment B are the current disclosure rules that apply to ACS special tabulations (http://www.census.gov/acs/www/Products/spec_tabs/drb_rules.htm While these rules do raise some questions, it is clear that rounding and thresholds are here to stay.

References

1. Christopher, Ed, “The CTPP: Historical Perspective”, December 2002, as found at http://www.TRBcensus.com/articles/ctpphistory.pdf (March 21, 2005).

2. Christopher, Ed, Elaine Murakami, Sherry Riklin and Nanda Srinivasan, The Long Form and American Community Survey Questions: Their Relevance to Transportation, Report to Submitted to OMB, March 9, 2001. http://www.TRBcensus.com/articles/030901acs.pdf (March 1, 2005)

3. Chuck Purvis, Metropolitan Transportation Commission, Oakland, San Francisco, e-mail to the CTPP list serve on 02/19/04, See attachment A.

4. Mix, Wende A, Ph. D., A Comparison of JTW Data in the 2000 Decennial Census and the ACS. Research sponsored by the U.S DOT Federal Highway Administration, December 2003, paper found at ftp://ftp.abag.ca.gov/pub/mtc/census2000/ACS/acs_pow.pdf (April 1, 2005)

5. Murakami, Elaine, CTPP Status Report newsletter article “ACS and Decennial Census SAS files available for research purposes”, Federal Highway Administration, August 2004 as found at http://www.trbcensus.com/newsltr/sr0804.pdf (March 31, 2005)

6. Limoges, Ed, CTPP Status Report newsletter article “Allocation of Missing Place of Work Data in Decennial Censuses and CTPP 2000”, Sabre Systems Inc., January 2004 as found at http://www.trbcensus.com/newsltr/sr0104.pdf (March 31, 2005)

Attachment A

Taken from the CTPP ListServe, Feb 2004

Rounding and CTPP

Ed Christopher: edc@berwyned.com
02/12/04 12:01PM CST
The rounding within the CTPP data can play heck with doing any data analysis. In the Chicago Central Area there are 155 individual TAZs. If you take a simple table from Part 2, say mode to work by sex, some interesting things happen. If you sum the total workers using the "total" field you get 631,999. This becomes an important number because people like to know the total. However, when you sum all the modes by zone you get 631,883. This is not a big deal except if you want to show drive alone, carpool, transit and other with their modal share percents. In this region, some of us like to see the actual numbers along with the percents. Logic would say to use the 631,883 when calculating the percentages but then that means the sum of the totals (which we know to be the better number because row rounding was applied to the tables) 631,999 gets tossed aside. One could get creative and distribute the 116 workers in some weighted fashion which would not likely affect any percentages but then the next guy who comes along using the CTPP data and software would get different numbers and we are back splitting hairs over who got what number from where.

Are others finding the issue of rounded numbers a bit frustrating, especially when it comes to aggregating TAZs?

Patty Becker; pbecker@umich.edu
02/12/04 14:20 PM CST
The most important thing to note here is that there is in fact no difference between 631999 and 631883. The 116 difference is well within the sampling error for these numbers.

Personally, I would percentage by the 613833 and then, if necessary, present the totals as is. It DOES NOT MATTER that you "tossed aside" the 631999 when calculating the percentages! Better yet, present the only the grand total (i.e. 631999) and not the total number in each mode; if people really want to know they can multiply it out and get approximations which are as close to the truth as the numbers on the file. I would also leave decimals off the percentages (i.e. 3%, not 3.1%) because the decimals are not statistically significant.

Chuck Purvis: Cpurvis@mtc.ca.gov
02/19/04 12:09 PM CST
The rounding of values inside the CTPP is, right now, a modest, annoying data processing issue. As professional data analysts, we are always on the lookout to make sure our numbers "add up" so that we're not missing anything. Rounding should be a privilege of the data analyst, AFTER all of the precise number-crunching has been performed. So, I want to make sure in my data analysis that the year 2000 total

population of my region is ALWAYS 6,783,760. IF IT'S DIFFERENT, THEN I MADE A MISTAKE THAT I HAVE TO CORRECT. After I get the precise number, then I can do the rounding off to my heart's desire, that is, 6.8 million persons, or 7 million, whatever. It is annoying, frustrating, an inconvenience, and a pain to NOT have the numbers add up!

The Census Bureau's use of rounding is an attempt at "disclosure avoidance" that is, to foil attempts of the data analyst to "reverse engineer" the precise name, address, and characteristics of individuals and their households. I frankly do not believe that rounding is the best method for ensuring disclosure avoidance. I believe other mathematical techniques to "dither" or randomize the reported data would be more useful, in terms of disclosure avoidance, and useful to the analyst, in terms of removing all of the rounding errors inherent in the current CTPP. My recommendation to the Census Bureau: do the right thing and hire mathematicians to find best methods to a) protect the identity of respondents; and b) to make things easier for the data user.

Frankly, you can use American FactFinder to enter your home address, and get the block-level population of persons on your block by race, by sex and by age. So then how is the Census Bureau providing "disclosure avoidance" for standard products like Summary File #1? If the Census Bureau had implemented rounding on standard census products such as SF1, SF2, SF3, and SF4 then there would have been a riot among the data users, Congress would have intervened, and the Census Bureau would be backtracking as fast as you could say Appropriations Committee.

Right now we have two classes of Census Bureau products: "first class" products such as the summary files and the Public Use Microdata Sample where there is (thank goodness!) NO rounding at all. (There are data thresholds in SF2 and SF4, but that's another matter.) The "second class" products are the CTPP and the EEO files, where there is rounding of data to the nearest, 10, 15, 20, etc. Perhaps it is the intent of the Census Bureau to implement rounding in future releases of "regular" Census Bureau products, such as American Community Survey and 2010 Census short form data. That would be a big mistake.

The rounding of data in the CTPP guarantees loss of productivity: the data analyst will lose productivity in terms of always second -guessing the data processing steps (is a tract or zone missing? are there problems in my computer code?); and the data analyst will lose time in explaining to data users: WHY THE NUMBERS DO NOT ADD UP!

Try explaining why: 10 + 10 + 10 + 10 = 50 !!!

I have spent too much time over the past 20+ years explaining the difference between commuters and "home-based work" trips; and "workers at work" and "total employment." Now, we can be guaranteed to spend a heck of a lot more time explaining "why don't the numbers add up?" (Does anybody have the home phone numbers for Census Bureau management?)

Here's a real life example using the CTPP Part 2 data. Let's say my boss asks as simple question: "How many transit commuters are at work in the Bay Area?" Using the Part 2 data, I am able to provide my boss 15 different answers!

The short answer is "320 thousand."

The long answer:
In Table 2-2 (Means18) there are five categories of "transit" that need to be summed to derive "total transit. In table 2-12 (Means11) there are three categories of "transit" that need to be summed; and in Table 2-27 (Means8) there are two categories of transit that need to be summed to get total transit. (There are no "Means5" tables in CTPP2 where "transit" is one, and only one category.)

And there are multiple summary levels where one can derive a regional total count of transit commuters, including TAZ, block group, tract, county and the "MPO Summary Level". (Also, the county-place-remainder, the place-remainder-tract, and MSA/CMSA summary levels can be used to extract more "different answers")

So, the following table illustrates the range of "regional transit commuters" using the three available means-of-transportation tables, and five of the different summary levels available in CTPP:

N	SUMLEV	Table 2 -2 (Transit=5 cats)	Table 2-12 (Transit=3 cats)	Table 2 -27 (Transit=2 cats)
4,031	TAZ	319,435	319,553	319,600
4,384	Blk Grp	319,433	319,521	319,541
1,403	Tract	319,717	319,780	319,836
9	County	320,116	320,129	320,125
1	MPO	320,125	320,120	320,120

What this tells us is that the number of "regional transit commuters" working in the Bay Area is somewhere between 320,118 and 320,122, and it's rounded to 320,120. All of the other numbers are subject to a modest degree of rounding error.

AND THERE IS A PATTERN!!! There is data "leakage" the more one aggregates from lower levels of geography, and from greater number of subcategories (e.g., aggregating from the five transit sub -groups versus the two transit sub-groups.) This data leakage is hardly statistically significant. It is, however, annoying.

My recommendation to users of CTPP data (Part 1 and Part 2)

1. Obtain your "regional control totals" or "state control totals" from the most geographically aggregate summary levels, e.g., SUMLEV=040 for states, and SUMLEV=930 for MPOs.
2. . Avoid aggregating (summing together) your geographies whenever and wherever possible.

3. . Avoid aggregating categories (e.g., detailed household income versus grouped household income; means of transportation) whenever and wherever possible. For example, to get the least affected count of 3-plus carpools, use tables based on Means of Transportation (8 categories.)
4. . Sum as few categories as possible to derive aggregated measures such as "total transit." For "total transit" use CTPP Part 2, Table 27, where you are only summing bus/trolleybus to streetcar/subway/railroad/ferry.
5. . Adjust (de-round, un-round) as you see fit. Use SF3 or PUMS to provide control totals to adjust the CTPP Part 1 data.
6. . Develop a sense of humor. As I see it, this data rounding is a real joke. Don't take these data issues too seriously. And it's kind of funny that the numbers don't add up. Or, as they say: "close enough for government work."

cheers and good luck,
Chuck Purvis, MTC
****
Pop Quiz:
Question: Using the CTPP, 10 + 10 + 10 + 10 = ?
a) 30
b) 35
c) 40
d) 45
e) 50
f) Any of the Above

Attachment B

The following information was taken from the Census Bureau web site at http://www.census.gov/acs/www/Products/spec_tabs/drb_rules.htm on March 22, 2005

Disclosure Review Board Rules/Requirements (October 25, 2004)

1. All American Community Survey special tabulations must be reviewed by the Disclosure Review Board. After the tabulation has been created, if the program area identifies any potential disclosure problems, they will refer them back to the DRB.
2. All cells in any American Community Survey special tabulation must be rounded.

   The rounding schematic for all tables is:

   0 remains 0
   1-7 rounds to 4
   8 or greater rounds to nearest multiple of 5 (i.e., 864 rounds to 865, 982 rounds to 980)
   Any number that already ends in 5 or 0 stays as is.

   Any totals or subtotals needed should be constructed before rounding. This assures that universes remain the same from table to table, and it is recognized that cells in a table will no longer be additive after rounding.

3. Medians or other quantiles may be calculated as
   A. an interpolation from a frequency distribution of unrounded data (these are not subject to additional rounding), or
   B. as a point quantile. These must be rounded to two significant digits: 12,345 would round to 12,000; 167,452 would round to 170,000. There must be at least 5 cases on either side of the quantile point.

   It is recognized that a quantile may indeed be some individual's response, but it is coincidental, not by design.

4. Thresholds on universes will normally be applied to avoid showing data for very small geographic areas or for very small population groups (often 3 or 50 unweighted cases). Tables may normally not have more than 3 or 4 dimensions, and mean cell size lower limits may also be required (mean cell size of each table is 3 unweighted cases).
5. Percents, rates, etc., should be calculated after rounding, but the DRB has granted exceptions to this rule when the numerator and/or denominator of the percent or rate is not shown.
6. Means and aggregates must be based on at least 3 values.
7. The finest level of detail shown for Group Quarters data will be Institutional/ Noninstitutional.
8. For Demographic Profiles from user-defined geographic areas (neighborhoods), all areas must have at least 300 (weighted) people in them. Using a computer program, the user-defined areas will be compared with standard Census Bureau areas to make sure users cannot obtain data from very small geographic areas by subtraction. If such small areas

are found, the boundaries of the user-defined areas must be changed.

Source: U.S. Census Bureau
American Community Survey Office
Last revised: Tuesday November 02, 2004