The U.S. Census Bureau's Recommendations Concerning the Census 2000 Public Use Microdata Sample (PUMS) Files

Population Division
U.S. Census Bureau
November 3, 2000

I. Introduction

In response to concerns over confidentiality, and, at the same time, in an effort to meet users' needs, we are recommending two sets of PUMS files: 5-percent state files and a 1-percent national characteristics file. We also describe, but do not recommend, a 1-percent national metro file. We have attached background documentation to describe the proposed files in detail.

II. Concerns over confidentiality and suggested solutions

Several factors have led to strengthening the confidentiality protection that we provide for public use microdata files. Rapid advances in technology, including more powerful computers, greater data storage capacity, increased access to the Internet, and advances in data linking and data mining make it more difficult for the Census Bureau to protect the confidentiality of microdata through disclosure-limitation techniques.

The Census Bureau's task in designing the PUMS is to balance the needs of our users with our responsibility to protect the privacy of our respondents. In response to the possibility of disclosure problems, we had already planned to implement swapping and top-coding.[1] In addition to these measures, implementation of some combination of the following options was also discussed: (1) raising the minimum population threshold for Public Use Microdata Areas (PUMAs) above 100,000 (possibly increasing the minimum population threshold to 250,000 as the Census Bureau did for the 1970 PUMS files) and (2) collapsing variables.

III. File types

After consideration of the issues, two types of PUMS files are being proposed: 5-percent state-level files and a 1-percent national characteristics file.

A. State-level PUMS files

We recommend the development of 5-percent state-level files. These files would provide information for most metropolitan areas and the more populous counties and central cities. We do not intend to raise the minimum population threshold for the state-level PUMAs above 100,000. Instead, we recommend increasing the degree of variable collapsing to the level deemed necessary to maintain confidentiality while retaining the current threshold.

First, from a user's standpoint, raising the minimum population threshold for PUMAs

[1] Data swapping is a method of disclosure limitation designed to protect confidentiality in tables of frequency data (the number or percent of the population with certain characteristics). Data swapping is done by editing the source data or exchanging records for a sample of cases. (Swapping is applied to individual records and therefore also protects microdata.) Top-coding is a method of disclosure limitation in which all cases in or above a certain percentage of the distribution are placed into a single category.

above 100,000 would greatly restrict a wide variety of local-level geographic analyses (such as studies of non-metropolitan, metropolitan, and intrametropolitan areas) conducted by public agencies, academic researchers, and the private sector. These analyses are the very reasons for the inclusion of a PUMA variable. A minimum population threshold of 100,000, while respecting state boundaries, will permit recognition of 231 of the 276 metropolitan areas (84 percent). If the minimum population threshold were to be increased to 250,000, only 139 metropolitan areas (50 percent) could be recognized.

Secondly, users are satisfied with the 100,000 minimum population threshold – the threshold set for both the 1980 and 1990 PUMS files – and we have heard clearly from users their displeasure at the possibility of an increase in the threshold for Census 2000. This is particularly true for those interested in time-series analysis; raising the population threshold makes comparability of results from different decades more difficult. Criticism of the Census Bureau's use of 250,000 as the minimum threshold for PUMAs in 1970 was an important reason for the decision to lower the minimum threshold to 100,000 people for the 1980 PUMS files and maintain it in the 1990 PUMS files.

Minimum population threshold for categorical variables

To maintain confidentiality while retaining as much characteristic detail as possible, we are proposing setting a minimum threshold of 10,000 in the national population for the identification of groups within categorical variables in the state-level PUMS files. At the May 22 PUMS Users Conference in Alexandria, VA, in response to concerns about confidentiality, some users had suggested a minimum population threshold of 25,000. The Disclosure Review Board has determined that a minimum threshold of 10,000 will maintain the confidentiality of responses while providing greater detail to the user.

We are asking for user opinion on what criteria should be used to collapse categories which do not meet the 10,000 threshold.

Post-processing

We recommend post-processing for the state-level files. Instead of identifying variable categories based upon pre-tabulation assumptions about the composition of the population, this approach develops variable collapsing requirements after the microdata samples have been drawn and the PUMA boundaries have been delineated. Each variable will be analyzed and only those values that do not meet the 10,000 minimum national population threshold will be recoded. This method will allow the Census Bureau to have user input to find the most meaningful combinations of variable categories, given the minimum population threshold for identifying categories.

Post-processing will improve the PUMS products by offering a more precise means of ensuring confidentiality. However, it will increase the processing and analytic workload and delay the release of the PUMS products to the public by approximately six months. Users appear to understand the need for that delay – especially if a national characteristics file (discussed in section B) is available before the state-level files.

B. National characteristics PUMS file

While we place the greatest emphasis on maintaining the current minimum population threshold to facilitate geographic analysis, we also recognize that the amount of variable collapsing necessary to maintain this threshold may result in the loss of social, economic, and housing detail in the state-level files. We recommend an additional PUMS file – a national characteristics file – to accompany the state-level PUMS files. In this national file, the amount of variable collapsing would be minimized in order to maximize the amount of social, economic, and housing information available. The goal of this file is to provide as close to the amount of detail in the 1990 PUMS files as possible (and in some cases, more detail). No national minimum population threshold for the identification of variable categories is planned. Limits on certain variables, deemed necessary to protect confidentiality, are covered in section C.

Geographic detail: minimum population thresholds for super-PUMAs

Users have indicated a strong preference for the identification of states in the national characteristics file. To ensure the identification of all the states, as well as the identification of super-PUMAs (geographic units in the national characteristics file that are combinations of three to four PUMAs delineated in the state-level files), we are proposing a minimum population threshold of 400,000 for the super-PUMAs.

C. Specifications for the state-level and the national characteristics files

The Disclosure Review Board has set the following specifications for the following variables in the state-level and national characteristics files.

1. Dollar amounts

We are planning to round dollar amounts before all summations, ratio calculations, or presentations of amounts. The dollar amounts will be rounded to the nearest $1,000 or $100 or $10 or $5 (including negative amounts) as follows:

$1-$7:	$5
$8-$999:	round to the nearest $10
$1,000-$49,999:	round to the nearest $100
$50,000 or more:	round to the nearest $1000

This rule would be applied to income types, utility costs, mortgage costs, rent, condominium fees, hazard insurance costs, and mobile home fees.

We recommend the following procedure for implementing income top-coding. An individual person's income would be rounded on a graduated scale and would be independently top-coded by variable type. The value inserted for observations above the top-code would be the state mean of all cases greater than the top-code minimum value. Incomes would then be summed across household members to obtain household totals, without any additional top-coding. The bottom-

coding for all income types which can have negative dollar values would be set at a maximum negative value of $10,000.

The housing-related dollar amount variables would be treated in the following way. Property taxes would be categorized in a similar way to 1990, with the exception of the higher tax categories (see Appendix A). All other housing-related dollar amounts would be treated similarly to income. That is, the variables would use the same rounding scale that income is using and they would use the state mean of the top-coded cases as the top-code value. For the items that are aggregated to create selected monthly owner costs (SMOC) and gross rent, each item would be rounded independently and top-coded before summing to the SMOC or gross rent total. No further rounding will be performed on the aggregated amount.

2. Race recodes and Hispanic recode

Appendix B shows our proposal for race, Hispanic origin, and American Indian and Alaska Native tribe categories in the state and national characteristics files. All categories would appear on the 1-percent file. Only categories with at least 10,000 people nationally would appear on the 5-percent files.

3. Age detail

For both the state-level and national characteristics files, we are proposing single-year age categories for ages 0 through 89 and a top-code of 90.

4. Reporting of household size

Due to concerns about confidentiality, large households, defined as households containing 10 people or more, will require a masking technique in the state-level files. We recommend that these large household records contain a state identifier, but not a PUMA identifier. We are asking for advice on how users should perform analysis on total population or total households in 100,000 PUMAs on state files when the households with 10 or more members will have state identifiers only.

In the national characteristics file, we are recommending that all household records would contain both a state identifier and a PUMA identifier (or, more precisely, a super-PUMA identifier). Because the minimum population threshold is 400,000 – just slightly less than the population of the smallest state – a super-PUMA identifier is equivalent to the state identifier for all states with populations of less than 800,000.

5. Other related issues:

Travel time will be treated as a continuous variable with standard top-coding, but no other collapsing or rounding.[2]

2 Given the level of travel time in the PUMS files, an additional level of disclosure avoidance is need in summary file matrices where aggregate travel time to work is tabulated. If a summary file aggregate travel matrix for any geographic level contains responses for only one or two sample persons, then the value of that aggregate should be rounded to the nearest multiple of five minutes (if the aggregate is not already evenly divisible by five).

Departure time will be rounded, as illustrated in Appendix C.

Year of entry into the country will be bottom-coded in direct relation to the top-coding for age.

D. No national metro file

We have discussed the possibility of creating a national metro file. This file would contain information on variables for all metropolitan areas with populations of 100,000 or more as was done in 1990, including the identification of multi-state metropolitan areas where one or more of the state parts contain less than 100,000 people. The file would use the same minimum population threshold and the same social, economic, and housing content as the 5-percent state-level files.

Although this option has been discussed, we are not recommending its implementation -- each of the three potential combinations of files which would include the national metro file (described below) has a significant drawback. The first combination -- 5-percent state-level files, a 1-percent national characteristics file, and a 1-percent national metro file -- is precluded by the Census Bureau's Disclosure Review Board's limitation on the total sample density of the files to 6 percent. The second combination -- 5-percent state-level files and a 1 percent national metro file – does not include a 1-percent national characteristics file, which many users deem necessary because of the reduction in characteristic detail in the state-level files. The third combination – 4-percent state-level files, a 1-percent national characteristics file, and a 1-percent national metro file -- is opposed by many users because it includes a decrease in the sample density of the state-level files from 5 percent to 4 percent. This would result in the loss of approximately 20 percent of cases, making the study of some small subpopulations potentially more difficult.

IV. Timetables for PUMS files

The 1-percent national characteristics file would be the first file released to the public in mid-2002. The 5-percent state-level files, requiring more time for post-processing, would be released to the public in 2003.

V. Issues for Consideration and Advice

1) We are asking for advice on how users should perform analysis on total population or total households in the state file PUMAs when the households with 10 or more members will have state identifiers only, and not PUMA identifiers. One possibility is to randomly assign each household a PUMA identifier in proportion to the total population of the state.

2) We are also asking for advice on what criteria should be used to collapse categories which do not meet the 10,000 threshold in the state-level files.

VI. Comments

Any comments on this proposal should be forwarded to: Paul Mackun
(e-mail: Paul.J.Mackun@census.gov or phone at: 301-457-2419.)

Prepared by: Paul J. Mackun
Population Division
U.S. Census Bureau

Appendix A: Proposed Categories for the Property Tax Variable[1]:

Category	Property Tax Ranges
0	N/A
1	None
2	$2 - $49
3	$50 - $99
4	$100 - $149
5	$150 - $199
6	$200 - $249
7	$250- $299
8	$300- $349
9	$350- $399
10	$400- $449
11	$450- $499
12	$500- $549
13	$550- $599
14	$600- $649
15	$650- $699
16	$700- $749
17	$750- $799
18	$800- $849
19	$850- $899
20	$900- $949
21	$950- $999
22	$1,000 - $1,099
23	$1,100 - $1,199
24	$1,200- $1,299
25	$1,300- $1,399
26	$1,400- $1,499
27	$1,500- $1,599
28	$1,600- $1,699
29	$1,700- $1,799
30	$1,800- $1,899
31	$1,900- $1,999
32	$2,000- $2,099
33	$2,100- $2,199
34	$2,200- $2,299
35	$2,300- $2,399
36	$2,400- $2,499
37	$2,500- $2,599
38	$2,600- $2,699
39	$2,700- $2,799
40	$2,800- $2,899
41	$2,900- $2,999

42	**$3,000- $3,099**
43	**$3,100- $3,199**
44	**$3,200- $3,299**
45	**$3,300- $3,399**
46	**$3,400- $3,499**
47	**$3,500- $3,599**
48	**$3,600- $3,699**
49	**$3,700- $3,799**
50	**$3,800- $3,899**
51	**$3,900- $3,999**
52	**$4,000- $4,099**
53	**$4,100- $4,199**
54	**$4,200- $4,299**
55	**$4,300- $4,399**
56	**$4,400- $4,499**
57	**$4,500 - $4,599**
58	**$4,600 - $4,699**
59	**$4,700 - $4,799**
60	**$4,800 - $4,899**
61	**$4,900 - $4,999**
62	**$5,000 - $5,499**
63	**$5,500 - $5,999**
64	**$6,000 - $6,999**
65	**$7,000 - $7,999**
66	**$8,000 - $8,999**
67	**$9,000 - $9,999**
68[2]	**$10,000 or more**

[1]Categories 0-56 are the same as in 1990.

[2]There is one nationwide top-code lower limit and each state receives the mean of the cases in the state above the top-code minimum value.

Appendix B. Race Recodes and Hispanic Recode for the PUMS

The following is the Population Division=s proposal for race recodes for the 1 percent and 5 percent PUMS files. All categories would appear on the 1 percent file. Only categories with at least 10,000 people nationally would appear on the 5 percent files.

Race Indicator for PUMS of race alone or in combination with one or more other races:

RACEW	White recode (defines White alone or in combination with one or more other races) 0 = No 1 = Yes
RACEB	Black or African American recode (defines Black or African American alone or in combination with one or more other races) 0 = No 1 = Yes
RACEAIAN	American Indian and Alaska Native recode (defines American Indian and Alaska Native alone or in combination with one or more other races) 0 = No 1 = Yes
RACASIAN	Asian recode (defines Asian alone or in combination with one or more other races) 0 = No 1 = Yes
RACENHPI	Native Hawaiian and Other Pacific Islander recode (defines Native Hawaiian and Other Pacific Islander alone or in combination with one or more other races) 0 = No 1 = Yes

RACESOR — Some other race recode (defines Some other race alone or in combination with one or more other races)

0 = No
1 = Yes

Race Version 1 for PUMS -- RACPUMS1

1 = White alone
2 = Black or African American alone
American Indian and Alaska Native alone:
3 = American Indian alone
4 = Alaska Native alone
5 = Both American Indian and Alaska Native
6 = American Indian or Alaska Native, not specified
7 = Asian alone
8 = Native Hawaiian and Other Pacific Islander alone
9 = Some other race alone
10=Two or more major race groups

Race Version 2 for PUMS -- RACPUMS2

One major race group:
1 = White alone
2 = Black or African American alone
American Indian and Alaska Native alone:
American Indian alone:
3 = Apache alone
4 = Blackfeet alone
5 = Cherokee alone
6 = Cheyenne alone
7 = Chickasaw alone
8 = Chippewa alone
9 = Choctaw alone
10= Colville alone
11= Comanche alone
12= Cree alone
13= Creek alone

RACPUMS2 (continued)
14= Crow alone
15= Delaware alone
16= Houma alone
17= Iroquois alone
18= Kiowa alone
19= Latin American Indian alone
20= Lumbee alone
21= Menominee alone
22= Navajo alone
23= Osage alone
24= Ottawa alone
25= Paiute alone
26= Pima alone
27= Potawatomi alone
28= Pueblo alone
29= Puget Sound Salish alone
30= Seminole alone
31= Shoshone alone
32= Sioux alone
33= Tohono O'odham alone
34= Ute alone
35= Yakama alone
36= Yaqui alone
37= Yuman alone
38= Other specified American Indian tribes alone
39= All other specified American Indian tribe combinations
Alaska Native alone:
40= Alaska Athabaskan alone
41= Aleut alone
42= Eskimo alone
43= Tlingit-Haida alone
44= Other specified Alaska Native tribes alone
45= All other specified Alaska Native tribe combinations
46= All combinations of specified American Indian and Alaska Native tribes
47= American Indian and Alaska Native, not specified
Asian alone:
One Asian race:
48= Asian Indian alone
49= Bangladeshi alone
50= Cambodian alone
Chinese alone:
51= Chinese, except Taiwanese, alone

RACPUMS2 (continued)
52= Taiwanese alone
53= Filipino alone
54= Hmong alone
55= Indonesian alone
56= Japanese alone
57= Korean alone
58= Laotian alone
59= Malaysian alone
60= Pakistani alone
61= Sri Lankan alone
62= Thai alone
63= Vietnamese alone
64= Other specified Asian alone
65= Asian, not specified alone
66= All combinations of Asian races only
Native Hawaiian and Other Pacific Islander alone:
One Native Hawaiian and Other Pacific Islander race:
Polynesian alone:
67= Native Hawaiian alone
68= Samoan alone
69= Tongan alone
70= Other Polynesian alone
71= All combinations of Polynesian races only
Micronesian alone:
72= Guamanian or Chamorro alone
73= Other Micronesian alone
74= All combinations of Micronesian races only
Melanesian:
75= Fijian alone
76= Other Melanesian alone
77= All combinations of Melanesian races only
78= Other Pacific Islander, specified, alone
79= Pacific Islander, not specified, alone
80= Two or more Native Hawaiian and Other Pacific Islander races only
81= Some other race alone
82= Two or more major races

Hispanic or Latino for PUMS -- SPANLONG

1 = Not Spanish/Hispanic/Latino
2 = Mexican
3 = Puerto Rican
4 = Cuban
5 = Dominican
6 = Costa Rican
7 = Guatemalan
8 = Honduran
9 = Nicaraguan
10= Panamanian
11= Salvadoran
12= Other Central American
13= Argentinean
14= Bolivian
15= Chilean
16= Colombian
17= Ecuadorian
18= Paraguayan
19= Peruvian
20= Uruguayan
21= Venezuelan
22= Other South American
23= Spaniard
24= All other Spanish/Hispanic/Latino

Appendix C. Reporting of Travel Time and Departure Time in the PUMS

Travel time:	Travel time will be treated as a continuous variable with standard top-coding, but no other collapsing or rounding. Given the level of travel-time detail in the PUMS files, an additional level of disclosure avoidance is needed in summary file matrices where aggregate travel time to work is tabulated. If a summary file aggregate travel time matrix for any geographic level contains responses for only one or two sample persons, then the value of that aggregate should be rounded to the nearest multiple of five minutes (if the aggregate is not already evenly divisible by five).
Departure time:	2400-0259 in 30-minute intervals, i.e. 2400-0029, 0030-0059
	0300-0459 in 10-minute intervals, i.e. 0300-0309, 0310-0319
	0500-1059 in 5-minute intervals, i.e. 0500-0504, 0505-0509
	1100-2359 in 10-minute intervals, i.e. 1100-1109, 1110-1119

Appendix D. List of 1990 Ancestry Categories with 1990 values of 10,000 or More[1]

1	German	57,947,374
2	Irish	38,735,539
3	English	32,651,788
4	Afro-American	23,777,098
5	Italian	14,664,550
6	American	12,395,999
7	Mexican	11,586,983
8	French	10,320,935
9	Polish	9,366,106
10	American Indian	8,708,220
11	Dutch	6,227,089
12	Scotch-Irish	5,617,773
13	Scottish	5,393,581
14	Swedish	4,680,863
15	Norwegian	3,869,395
16	Russian	2,952,987
17	French Canadian	2,167,127
18	Welsh	2,033,893
19	Spanish	2,024,004
20	Puerto Rican	1,955,323
21	Slovak	1,882,897
22	White/Caucasian	1,799,711
23	Danish	1,634,669
24	Hungarian	1,582,302
25	Chinese	1,505,245
26	Filipino	1,450,512
27	Czech	1,296,411
28	Portuguese, n.e.c .	1,148,857
29	British	1,119,154
30	Hispanic	1,113,259
31	Greek	1,110,373
32	Swiss	1,045,495
33	Japanese	1,004,645
34	Austrian	864,783
35	Cuban	859,739
36	Korean	836,987
37	Lithuanian	811,865
38	Ukrainian	740,803
39	Scandinavian	678,880
40	Acadian/Cajun	668,271
41	Finnish	658,870
42	United States	643,561
43	Asian Indian	570,322
44	Canadian	549,990
45	Croatian	544,270
46	Vietnamese	535,825
47	Dominican	505,690
48	Salvadoran	499,153
49	European, nec	466,718

50	Jamaican	435,024
51	Lebanese	394,180
52	Belgian	380,498
53	Romanian	365,544
54	Spaniard	360,935
55	Colombian	351,717
56	Czechoslovakian	315,285
57	Armenian	308,096
58	Pennsylvania German	305,841
59	Haitian	289,521
60	Yugoslavian	257,994
61	Hawaiian	256,081
62	African	245,845
63	Guatemalan	241,559
64	Iranian	235,521
65	Ecuadorian	197,374
66	Taiwanese	192,973
67	Nicaraguan	177,077
68	Peruvian	161,866
69	West Indian	159,167
70	Laotian	146,930
71	Cambodian	134,955
72	Other Eastern European and Soviet Union, n.e.c.	132,332
73	Syrian	129,606
74	Arab	127,364
75	Slovene	124,437
76	Serbian	116,795
77	Honduran	116,635
78	Thai	112,117
79	Asian	107,172
80	Latvian	100,331
81	Pakistani	99,974
82	Nigerian	91,688
83	Panamanian	88,649
84	Hmong	84,823
85	Turkish	83,850
86	Israeli	81,677
87	Guyanese	81,665
88	Egyptian	78,574
89	Slavic	76,931
90	Trinidadian/Tobagonian	76,270
91	Northern European	65,993
92	Brazilian	65,875
93	Argentinean	63,176
94	Dutch West Indian	61,530
95	Chilean	61,465
96	Samoan	55,419
97	Eskimo	52,920
98	Australian	52,133
99	Costa Rican	51,771
100	Assyrian/Chaldean/Syriac	51,765
101	Cape Verdean	50,772

102	Sicilian	50,389
103	Luxemburger	49,061
104	Palestinian	48,019
105	Albanian	47,710
106	Indonesian	43,969
107	Latin American	43,521
108	Western European, n.e.c.	42,409
109	Icelander	40,529
110	Venezuelan	40,331
111	Maltese	39,600
112	Guamanian	39,237
113	British West Indian	37,819
114	Barbadian	35,455
115	Basque, n.e.c.	34,335
116	Bolivian	33,738
117	Afghanistan	31,301
118	Ethiopian	30,581
119	Celtic	29,652
120	Bulgarian	29,595
121	Malaysian	27,800
122	Estonian	26,762
123	Prussian	25,469
124	Cantonese	25,020
125	Iraqi	23,212
126	Belizean	22,922
127	Bahamian	21,081
128	Jordanian	20,656
129	Other Subsaharan African, n.e.c.	20,607
130	Macedonian	20,365
131	Ghanian	20,066
132	Moroccan	19,089
133	South African	17,992
134	Alsatian	16,465
135	Tongan	16,019
136	Aleut	15,816
137	Amerasian	15,523
138	Uruguayan	14,641
139	Sri Lankan	14,448
140	Eurasian	14,177
141	Flemish	14,157
142	North American	12,618
143	Bangladeshi	12,486
144	Pacific Islander	11,330
145	Grenadian	11,188
146	South American	10,867
147	Polynesian	10,854
148	Other North African and Southwest Asian, n.e.c.	10,670
149	Okinawan	10,554
150	Central American	10,310
151	German Russian/Volga	10,153

[1]Values are based on 1990 CP-S-1-2 “Detailed Ancestry Groups for States” and may be different from the PUMS files.
Note: n.e.c. stands for “not elsewhere classified”