Enhancing the "100,000 rule" On The Variation Of The Per Cent Of Uniques In A Microdata Sample And The Geographic Area Size Identified On The File

Sam Hawala, Statistical Research Division, US Census Bureau
4700 Silver Hill Road 3209/4, Washington, DC 20233

Abstract: The Bureau of the Census releases microdata files, that is, data sets containing for each respondent the values of a number of characteristics. Data release is made under Title 13 of the U.S. Code, which prohibits wrongful disclosure of information on individuals. To make the identification of individuals highly unlikely, the Bureau of the Census does not identify geographic areas with less than 100,000 people in the microdata files. Using the concept of uniques, we find some evidence supporting this rule.

Key Words: Statistical Disclosure Control; Reidentification; Uniques.

1. INTRODUCTION

Through its various surveys, the Bureau of the Census develops public use microdata files, which contain information on individuals, households, businesses, or other units. The Bureau can only produce a small fraction of the potentially very valuable scientific analyses these extensive files make possible. Fortunately there are many researchers in the universities, foundations, and research firms, working independently, who are eager and able to study various social phenomena through the use of these microdata files.

The release of microdata files inevitably reveals some information about individual data subjects. Identity disclosure or re-identification occurs when a data subject is identified from a released microdata file. All data released, in print or electronically, by the Bureau are subject to confidentiality measures imposed by the legislation code under which the data were collected: Data are collected under Title 13 U.S. Code which protects the confidentiality of the individual respondents. Responses to the questionnaire can be used only for statistical purposes, and Census Bureau employees are sworn to protect respondents' identities.

The Bureau takes several measures to minimize the risk of re-identification. These measures include *anonymizing* the files by removing direct identifiers such as names and addresses. The Bureau also deletes the code identifying smaller geographic areas - that is places smaller than 100,000 in population - because anyone trying to identify a respondent will have his task greatly simplified if he knows the respondent's local area. (Mugge 1983)

The proportion of records, which might be uniquely identified in a microdata file, is related to the geographic detail on the file. The size of the geographic area, the number of characteristic variables on the file, and the detail of the characteristics provided determine uniqueness. We investigate the relationship between size of geographical area and percent of uniques based on several sets of variables and we attempt to provide methodological support for the 100,000-population threshold used for public use microdata files.

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

2. METHODOLOGY

2.1. Public-use Microdata Samples

The Bureau of the Census produced three separate public-use microdata samples (PUMS) from the 1990 decennial census. These are files that contain records for a sample of housing units with information on the characteristics of each housing unit and the people in it. Each of these separate PUMS represents a given percent of the population and housing of the United States:

(a) 5% Sample, identifying all States and various subdivisions within them, including most counties with 100,000 or more inhabitants;
(b) 1% Sample, identifying all metropolitan territories and most metropolitan areas (MA) with 100,000 or more inhabitants individually, and groups of MAs elsewhere.
(c) Elderly subset of 3% sample, identifying all States and various subdivisions within them, and containing only households with at least one person age 60 or older.

Each microdata file is a subsample of the population sample (≈17% of all housing units) that received the census long-form questionnaires. To prevent disclosure of identifiable information about any individual, the geographic units of reporting used in the PUMS are the Public-Use Microdata Areas (PUMAs). The PUMAs are non-overlapping. The 1-percent, 5-percent, and 3-percent PUMAs comprise areas that contain at least 100,000 people.

In the field of Disclosure Limitation some of the users of the data are considered 'intruders' or 'attackers'. When using microdata files, 'attackers' may, by chance or intentionally, successfully re-identify respondents. Microdata records are edited for confidentiality. They contain no names, addresses, or telephone numbers. The Bureau of the Census (BOC) also limits the detail (topcodes, recodes) on income, age, occupation, and other selected items to further protect the confidentiality of the records. The concern is that some 'attackers' however may be able to use detailed combinations of certain characteristics to link microdata records to outside files, containing identifiers, allowing them to identify respondents.

2.2. Uniqueness

Some respondents may possess characteristics or combinations of characteristics that make them stand out from other respondents on a microdata sample. They are called *Sample Uniques* with respect to the set of characteristics considered. The combination of characteristics may be different from all the other combinations in the population, in which case the unit is called a *Population Uniques*. A young child with an unusually large income or a college degree, or a working elderly person is likely to be a population unique within their local area. The area and the specific set of characteristics are the defining ingredients of population uniques and sample uniques.

A record can be a sample unique without being a population unique. In fact, there may be other people with the same characteristics, but they are not included in the sample. It can also be observed that records that are not sample uniques cannot be population uniques. From the sample one can estimate the percent of population uniques, as was shown in (Zayatz '91).

A population uniques is at risk of disclosure if it is represented on the microdata file. An 'attacker' potentially could link the record on the microdata file to outside files containing names or other identifiers to re-identify a respondent. If the 'attacker' knows that a particular respondent, who is a population uniques, participated in the survey then the respondent's record, which is then a sample uniques, is at risk of disclosure. The survey in this case discloses more information about the respondent than the attacker already knows.

A logical course of action for an attacker is to first identify the area in which a respondent resides, then within the area attempt to re-identify him or her. The present work is an attempt to address the question of how does the percent of population uniques, with respect to a given set of characteristics, in a microdata file vary as a function of the size of the geographical area identified on the file.

It is worth noting that, in demographic microdata containing person as well as household information, if a household can be identified, based on some combination of household characteristics and members' characteristics, then each of its members' risk of re-identification increases dramatically. For example, very large households of size eight or more are relatively rare. According to the 1990 census, only 0.627% of the households had 8 or more persons. When the size of the household is combined with information on age-sex-race of the members, the household becomes even more rare. This paper does not consider the re-identification risk due to re-identification of the entire household, but rather looks at individuals separately.

2.3. Data preparation

We selected stratified random samples from the 1% PUMS files, with geographic area sizes ranging from 20,000 to 500,000 person records. These samples were to simulate average geographical areas of various sizes. These may be thought of as communities or as geo-political areas with well defined boundaries. The samples were drawn without replacement from each state with probability proportional to the size of the state. For each sample we calculated the proportion of uniques records. We experimented with nine different models each offering a different combination of characteristics or a different recoding scheme for the same characteristics. The number of categories in each model ranged from 2.5×10^{18} in model I to 324,000 in model IX. The characteristics used are varied details on Age (up to 90 categories); Race (up to 64 categories); Sex (2 categories); Hispanic Origin (up to 64 categories); Ancestry (up to 143 categories); Birth Place (up to 167 categories); Occupation (up to 443 categories); Industry (up to 245 categories); Language (up to 74 categories).

3. RESULTS

The proportions of population uniques are given in Table1 in the Appendix. We mention several findings that can be seen from this table:

(a) The results from model IX show that even with a community that is larger than 500,000 persons there could be more than 5,000 (1%) persons that are uniques.
(b) The proportion of uniques decreases monotonically as the geographical area population size increases
(c) For a fixed sample size the relationship between the proportion of uniques and the number of possibilities defined by the characteristic set is not monotonic. although it clearly decreases rapidly from Model II to Model VI.

We ran a regression model relating the proportion of uniques (RATE) to population size (SIZE) for Model VII. See the appendix for the graph of the data, the regression line and the regression equation. (Y stands for RATE and X stands for SIZE) The Adjusted-R^2 = .9976..

According to the model there is a local minimum near 100,000. As was expected the rate of uniques decreases monotonically as the population size increases and as is shown in the graph. The decrease in the proportion of uniques is not significant as the population size varies from 100,000 to 500,000. This result was also seen for the other Models (not shown here due to lack of room but available from the author.)

4. CONCLUSION

The chances that a unit responding to a survey can be re-identified from a microdata file, can be limited if the number of uniques elements with respect to a set of characteristics measured by the survey is kept to a minimum. We studied the variation of the percent of uniques elements with respect to nine given sets of variables on a microdata file as a function of the size of the geographical areas identified. This study supports the validity of the use of the 100,000 population limit for most demographic microdata files.

5. REFERENCES

Horn John, (?), *A Simulation Study of the Identifiability of Survey Respondents when their Community of Residence is Known*, National Center for Health Statistics, Memorandum not dated.

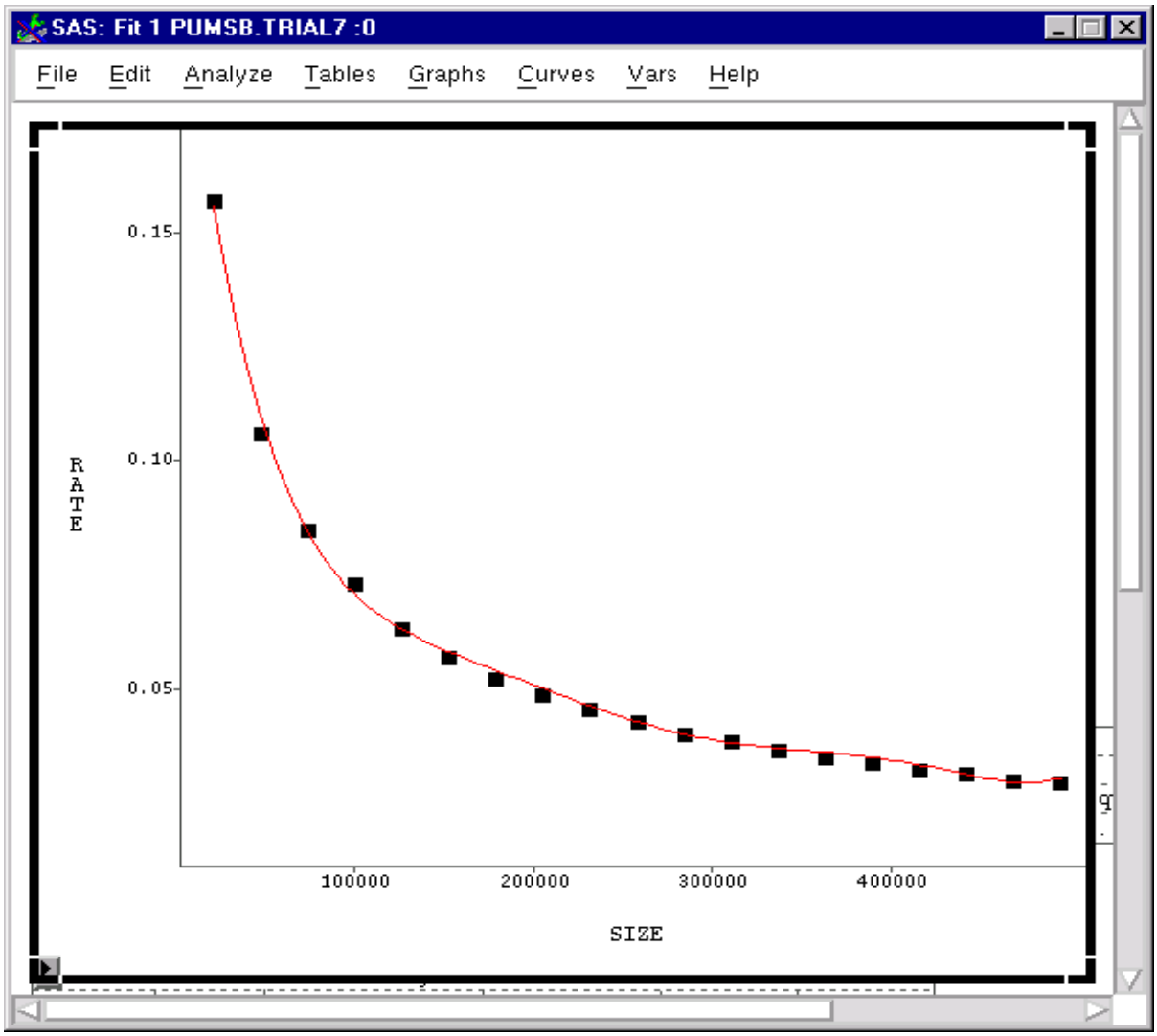
Mugge, Robert H. (1983), *Issues in Protecting Confidentiality in National Health Statistics*, in Proceedings of the Social Statistics Section, American Statistical Association, pp. 592-594

Zayatz L. (1991), *Estimation of the Percent of Uniques Population Elements on a Microdata File Using the Sample*, Bureau of the Census, Statistical Research Division, Report Number: Census/SRD/RR-91/08.

Appendix
Table 1 : Percent of Uniques by population size for nine models

Population Size	Model I 2.5×10^{18}	Model II 4.6×10^{16}	Model III 1.1×10^{14}	Model IV 4.5×10^{11}	Model V 3.9×10^{6}	Model VI 4.5×10^{6}	Model VII 647,680	Model VIII 434,140	Model IX 324,000
21034	90.7%	91.1%	90.5%	72.6%	32.6%	22.7%	12.7%	11.0%	15.7%
47421	84.7%	84.7%	83.5%	56.9%	23.1%	14.6%	8.1%	6.4%	10.6%
73813	81.3%	81.3%	79.7%	48.3%	18.6%	12.0%	6.0%	4.9%	8.5%
100199	79.1%	78.7%	77.0%	43.0%	16.0%	9.8%	5.1%	3.8%	7.3%
126587	77.0%	77.1%	74.3%	38.6%	14.0%	8.7%	4.4%	3.2%	6.4%
152974	75.6%	75.6%	72.3%	35.8%	12.6%	7.7%	4.0%	2.8%	5.7%
179364	74.4%	74.2%	70.7%	33.3%	11.6%	6.9%	3.6%	2.5%	5.2%
205753	73.2%	73.1%	69.1%	31.3%	10.7%	6.4%	3.4%	2.2%	4.9%
232139	72.0%	72.2%	68.1%	29.6%	9.9%	5.9%	3.2%	2.0%	4.6%
258531	71.3%	71.4%	67.0%	28.2%	9.3%	5.5%	3.0%	1.8%	4.3%
284919	70.6%	70.6%	65.8%	27.0%	8.8%	5.2%	2.8%	1.7%	4.0%
311309	69.9%	69.6%	64.9%	26.0%	8.4%	4.8%	2.7%	1.6%	3.9%
337697	69.2%	69.2%	64.1%	25.0%	8.0%	4.5%	2.6%	1.5%	3.7%
364086	68.8%	68.7%	63.2%	24.3%	7.7%	4.3%	2.5%	1.4%	3.5%
390475	68.1%	68.0%	62.4%	23.5%	7.4%	4.1%	2.3%	1.3%	3.4%
416865	67.7%	67.5%	61.9%	22.8%	7.1%	3.8%	2.3%	1.2%	3.2%
443250	67.0%	67.2%	61.0%	22.2%	6.8%	3.7%	2.2%	1.2%	3.2%
469645	66.7%	66.7%	60.5%	21.6%	6.5%	3.6%	2.2%	1.1%	3.0%
496028	66.2%	66.2%	59.8%	21.0%	6.4%	3.4%	2.1%	1.1%	3.0%
522417	65.8%	65.8%	59.2%	20.7%	6.1%	3.3%	2.0%	1.0%	2.9%

Graph for the data from Model VII



Regression equation

$$Rate = .2 - 3.5\times10^{-6}X + 3.4\times10^{-11}X^2 - 1.7\times10^{-16}X^3 + 4.9\times10^{-22}X^4 - 7\times10^{-28}X^5 + 4\times10^{-34}X^6$$

$$where \quad X = Size$$