RELATING RISK OF DISCLOSURE FOR MICRODATA AND GEOGRAPHIC AREA SIZE

Brian Greenberg and Laura Voshell*
Bureau of the Census

KEY WORDS: Microdata, Disclosure Avoidance, Geographic Detail, Unique

1. INTRODUCTION

National statistical agencies and offices collect information about a nation's population and institutions and make the information available to the public. Statistical agencies have the responsibility of designing data release strategies which will not violate pledges of confidentiality either through intent or neglect. When a statistical agency releases microdata products, one of the important considerations is the geographic detail on the file. The finer the geographic breakout, the greater the risk that a respondent may be identified based on individual or household characteristics. In this paper, we regard the number of population uniques present on the microdata file as one of the components of a measure of disclosure risk and then relate this component of risk to identifiable geographic area size. One objective of this work is to contribute to the development of geographic area cut-offs when designing microdata release strategies.

Microdata files consist of records at the respondent level which contain characteristics of a sample of the individuals or households in a certain population. All obvious identifiers of respondents such as name or address have been removed. These records also contain geographic identifiers such as state or metropolitan area in which each respondent is located. The Census Bureau currently employs a general rule stating that no geographic region containing less than 100,000 people in the sampled area may be identified on a microdata file. However, for microdata from some surveys or censuses, the minimum number of people required per identified region may be larger than 100,000 if it is thought that the disclosure risk would be too great at that level. For example, for microdata from the Survey of Income and Program Participation (SIPP), no geographic region containing less than 250,000 people may be identified. One can reasonably assume that the smaller the identifiable geographic region on the file, the greater the disclosure risk.

We define the key variables on a set of microdata to be those variables which taken together may contribute to the linking of a record to its respondent (Bethlehem, Keller, and Pannekoek 1990; Greenberg 1990). In each identified geographic region, there may be records on the microdata file that represent individuals or households in that region which are unlike any other individuals or households in that region for the set of key variables. These records will be called population uniques. The population uniques on a microdata file possess a high disclosure risk. A user of the microdata may know that an individual or a household in a given region has a unique combination of key variables, and if that combination of variables is represented in the microdata, the user would be able to link that respondent to its record. Also, if a user has access to a set of data records with individual identifiers and the same key variables as on the proposed public use microdata file, the user could match all records appearing in both sets of microdata that are unique with respect to the key variables. Under this scenario, unique individuals or households could be linked to their records and confidential information would be disclosed.

This paper attempts to describe the relationship between the percent of population uniques on a microdata file from a specific geographic region and the size of that geographic region. In most cases, when a geographic region is enlarged, the percent of individuals or households in that region which are unique decreases. This is because some of the individuals or households which are added to the region when the region is enlarged have the same combination of key variables as those individuals or households which were unique in the smaller region.

We are particularly interested in answering two questions. When increasing the size of a geographic region in order to reduce the percent of unique individuals or households in that region, do we reach a point at which a further increase in size has no appreciable affect upon the percent of unique individuals or households in the region? And secondly, how does the similarity or dissimilarity of the individuals or households in a region affect the percent of population uniques as a function of the size of the region? Knowing the answers to these questions may help in the designing of microdata release strategies.

In all of the work described below, we took simple random samples of a data set to model the change in the size of a geographic region and noted the difference between the percent of uniques in the original data set and in the subsets. We chose to use the procedure of taking random subsets from a data set rather than removing geographic areas from a specified geographic region in order to ensure that our work was controlled and replicatable and that our results would not be relevant solely to the region with which we were working. We have examined the change in the percent of population uniques as geographic sub-areas are removed from actual geographic regions, and we have found only negligible differences in the effect on percent of uniques between the two procedures. For example, suppose that an actual geographic region contains 50000 households, 10% of which are unique, and the region is reduced in size to a smaller region containing 30000 households. The percent of uniques may increase to say 25%. We have found that if a simple random sample containing 30000 of the original 50000 households was considered, the percent of unique households would be approximately 25%. This finding is explained and supported in detail in Greenberg and Voshell (1990).

In Section 2, we discuss how the size of a geographic region affects the percent of unique individuals or households in that region for different sets of key variables and various categorical breakdowns of the key variables. We introduce the concept of equivalence classes in Section 3 and describe how the distribution of equivalence classes in a region affects the change in the percent of unique individuals or households brought about by a change in the size of the region. In the conclusion, we summarize our findings and offer recommendations concerning the development of geographic area cut-offs for microdata files. Further observations and appendices omitted from this paper due to space limitations are contained in the report Greenberg and Voshell (1990), from which this paper is an extract.

2. SIZE OF THE GEOGRAPHIC REGION VERSUS PERCENT OF UNIQUE HOUSEHOLDS

In this section, we discuss how the size of a geographic region affects the percent of households that are unique in that region for different sets of key variables and for various categorical breakdowns of one of those key variables. To model the effects of reducing the size of a geographic region, we conducted the following experiment. Starting with a "population" data set, we took simple random samples of the data set and noted the difference between the percent of uniques in the original data set and in the subsets.

2.1 Varying the Number of Key Variables

We began with a data set of 87959 household records from the 1980 Decennial Census. The 87959 households contained a total of approximately 220000 individuals, and record variables were recoded to resemble possible key variables on SIPP microdata. A decription of these variables may be found in Greenberg and Voshell (1990). Starting with the full 87959 records and using Poisson sampling, we randomly removed approximately 4398 records from the file to obtain a subset containing about 95% of the original records. We continued to randomly remove this number of records until we had obtained 19 random, nested subsets containing approximately 95%, 90%, 85%, ..., 5% of the records in the original data set. Using 6, 10, and 15 record variables, we then counted the number of unique households in each of these data sets.

We plotted the percent of unique households versus the size of the data set for the data sets using the 6, 10, and 15 variables, as shown in Figure 1. These plots were decreasing and concave up. In each case, the percent of unique households leveled off considerably as the size of the data set increased. Note that we did reach a point where a further increase in the size of the data set had no appreciable affect upon the percent of unique households in that data set. Consider, for example, the data with 6 variables, as shown in Figure 1. When the size of the data set reached about 30000, a further increase in size offered almost no decrease in the percent of unique households in the data set.

In Figure 1, we also see that the more variables used in the analysis, the larger the percent of unique households for data sets of corresponding size. This is to be expected because the larger the number of variables used, the more likely to find differences between those variables for different households. The more variables used, the more dissimilar the households can be. Also note that the greater the number of variables, the larger the decrease in the percent of unique households brought about by an increase in the size of the data set. Thus the more dissimilar the households in a data set, the greater the decrease in the percent of unique households brought about by an increase in the size of the data set. A more detailed discussion on the method we used for quantifying the similarity or dissimilarity of households in a data set through the use of the entropy function will be presented in Section 3.

2.2 Varying the Categorical Breakdown of a Key Variable

Using the same 87959 household records with 15 variables and Poisson sampling, we randomly removed approximately 3159 records to obtain a subset containing about 96.4% of the records in the original data set. We continued to randomly remove this number of records until we had obtained 10 random, nested subsets of this data set. Our smallest subset contained 56372 records. We calculated the percent of unique households in each data set six times using different categorical breakdowns of the variable "payment" in order to see how geographic detail and the categorical breakdown of a key variable interact to affect the percent of unique households. In the SIPP context, the variable "payment" is the sum of utility costs and rent or mortgage payment, property taxes, and insurance. See Greenberg and Voshell (1990) for the six different categorical breakdowns of the variable "payment".

We plotted the percent of unique households in the data set versus the size of the data set for the various breakdowns of the variable "payment". These plots, shown in Figure 2, were decreasing and slightly concave up.

Entropy was used to measure the dispersion of the households over the categories of the variable "payment" for the original data sets of 87959 household records. If there were M categories of the variable "payment", and p_i was the probability that a household's "payment" was in category i, then

$$ENTROPY = - \sum_{i=1}^{M} p_i \times \ln(p_i)$$

Both the number of categories of the variable "payment" and the dispersion of the households over those categories affect the entropy value. For a fixed number of categories, the more evenly spread the household "payment" values over the categories, the higher the entropy. Entropy also increases as the number of categories increases given an even spread over the categories. We wanted to see whether this measure of dispersion was indicative of the percent of unique households in the data set. As seen in Figure 2, the larger this entropy value of the variable "payment", the larger the percent of households that were unique. Also note that the larger the entropy value, the larger (slightly) the rate of decrease in the percent of uniques as the size of the data set became larger. Thus the more disperse the households in the data set as measured by the entropy of one variable holding all others constant, the greater the decrease in the percent of population uniques brought about by an increase in the size of the data set. We extend the use of entropy to incorporate several variables jointly in Section 3 through the use of the equivalence class structure of the data set.

When examining Figures 1 and 2, it is interesting to note that, in this study, no matter how many variables are used in the analysis and no matter how the variable "payment" is broken into categories, the difference between the percent of unique households in a data set of 87959 household records and the percent of unique households in a data set of 56372 household records is never more than five percent.

3. EQUIVALENCE CLASSES, GEOGRAPHIC DETAIL, AND PERCENT OF POPULATION UNIQUES

Decreasing the size of a geographic region will cause some of the households which were not unique in the larger region to become unique in the smaller region. The number of households which become unique because of the reduction in the size of the region depends on the size of the reduction and on the similarity or dissimilarity of the households in the original region. This similarity is reflected in the distribution of the sizes of the equivalence classes (in a geographic region). An equivalence class consists of all households which have the same combination of key variables. All households within a region can be grouped with all other households exactly like them, and each group is an equivalence class. The number of households in each equivalence class is the size of that equivalence class. Unique households are equivalence classes of size 1.

3.1 New Uniques

When a subset of a data set is considered, there will be some records in the subset which are unique with respect to all other records in the subset but which were not unique in the original data set. We will use the term new uniques for all such records. We will use the term original uniques for the records that were unique in the original data set. The expected number of new uniques in a random subset taken from a data set with a given equivalence class structure is calculated as follows.

Let N = number of records in the original data set
n = number of records in the subset
L = the size of the largest equivalence class in the original data set
t_k = the number of equivalence classes of size k in the original data set

Then the expected number of new uniques in a random subset of size n is

$$\sum_{k=2}^{L} t_k \times \frac{\binom{k}{1}\binom{N-k}{n-1}}{\binom{N}{n}}$$

The expected number of original uniques in a random subset of size n is

$$\frac{t_1 \times n}{N}$$

Thus, the expected percent of uniques in a random subset of size n is

$$\frac{t_1}{N} + \sum_{k=2}^{L} \frac{t_k}{n} \times \frac{\binom{k}{1}\binom{N-k}{n-1}}{\binom{N}{n}}$$

which is greater than or equal to t_1 / N, the percent of uniques in the original data set. As noted previously, we have found that any difference between the change in the percent of population uniques brought about by the reduction in size of a geographic region and the change in the percent of population uniques brought about by removing a simple random sample of the households in that region is negligible. Therefore, when the size of a geographic region is reduced, it is expected that the percent of unique households in that region will increase. This formula also shows that the percent of household records that are unique with respect to other household records in a sample of a population is larger than the percent of households which are unique with respect to all other households in the entire population.

3.2 Equivalence Class Structure and Overall Entropy

We have shown that the expected increase in the percent of unique households brought about by a reduction in the size of a geographic region depends upon the equivalence class structure of the households in the original region. We now attempt to quantify the dispersion or dissimilarity of the households in a region using a measure based upon the equivalence class structure of the households. We define this measure of dispersion as overall entropy which may be calculated as follows:

Let N = number of households in the original region
L = size of the largest equivalence class in the original region
t_k = number of equivalence classes of size k in the original region

We Define

$$OVERALL\ ENTROPY = -\sum_{k=1}^{L} t_k \times [(k/N) \times \ln(k/N)]$$

The greater the dispersion of the households, the larger the value of overall entropy. Using the same 87959 household records and Poisson sampling, we created 9 random, nested subsets. We calculated the overall entropy of the original data set, and we calculated the percent of unique households in each subset ten times using sets of 6, 7, 8, ... , and 15 variables. For a description of these variables, see Greenberg and Voshell (1990). As one would assume, the larger the number of variables, the larger the overall entropy. The results are plotted in Figure 3. Note that the greater the dispersion as measured by overall entropy, the larger the percent of unique households for corresponding subset sizes and the larger the increase in the percent of unique households brought about by a decrease in subset sizes. So again, the more dissimilar the households in a data set, the greater the change in the percent of unique households brought about by a change in the size of the data set.

4. CONCLUSIONS

As was stated earlier, we desired to answer two main questions. When increasing the size of a geographic region in order to reduce the percent of unique individuals or households in that region, do we reach a point at which a further increase in size has no appreciable affect upon the percent of unique individuals or households in the region? And secondly, how does the dispersion of the individuals or households in a region affect the change in the percent of unique individuals or households brought about by a change in the size of the region? In our research, we have discovered that one does reach a point at which a further increase in the size of a region has almost no affect upon the percent of unique households in that region. The size at which this point occurs, however, varies for different data sets with different key variables. We have also noted that the more disperse the households from a region, the greater the increase in the percent of unique households brought about by a decrease in the size of the region.

Because different data sets contain different key variables, different numbers of key variables, and different categorical breakdowns of key variables, geographic detail has a different impact on each one. Each data set must be examined individually for possible disclosure risk. One may wish to use the percent of records in the data set which represent unique individuals or households as a way of quantifying disclosure risk. Although most sets of microdata records represent only a sample of a population, the percent of population uniques appearing on the file may be estimated using information from the sample (Willenborg, Mokken, and Pannekoek 1990; Voshell 1990). Also, the percent of sample uniques may be used as an over-estimate of the percent of population uniques appearing on the file.

If a statistical agency chooses a certain maximum acceptable percent of either sample uniques or estimated population uniques required prior to the releasing of a set of microdata, it can change the number of key variables, the categorical breakdowns of those key variables, and the geographic detail on the microdata file until it has fulfilled that requirement. Dropping some key variables from the file, collapsing some of the key variable categories, and decreasing geographic detail are all ways of decreasing the percent of uniques on a file. The potential users of the microdata may express interest in some variables more than others or perhaps accept a decrease in the detail of all variables for an increase in geographic detail. In this way, the users may assist the statistical agency in arriving at a file providing as much data utility as possible with an acceptable disclosure risk. This type of interaction between agency and users and this type of trade-off between key variable detail and geographic detail will be incorporated in the design of the release strategies of the Public Use Microdata Samples (PUMS) from the 1990 Decennial Census and may be used in the future to develop a new geographic area cut-off for SIPP microdata.

Acknowledgment

The authors wish to thank some of their colleagues at the Census Bureau for their help during the preparation of this paper. They thank Jay Kim for assistance with obtaining and recoding the data used for this work and Jim Fagan for his valuable discussions on the topics in this paper. They also thank Cynthia Clark, Colleen Sullivan, and Eric Schindler for their comments on an earlier version.

* This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

Figure 1. Percent of Unique Households Versus Size of Data Set. The symbols used in this figure represent the number of variables in the data set. a: 6 Variables, b: 10 Variables, and c: 15 Variables.

```
75     +
Percent]   c
Unique ]
70     +
       ]
       ]
65     +
       ]
       ]     c
60     +
       ]
       ]         c
55     +
       ]
       ]           c
50     +                c
       ]
       ]                  c
45     +                       c
       ]                        c
       ]                            c  c
40     +                                    c   c
       ]                                         c  c
       ]                                                c  c
35     +                                                       c   c
       ]
       ]   b
30     +
       ]
       ]
25     +       b
       ]
       ]         b
20     +
       ]            b
       ]              b
15     +                   b   b
       ]                       b  b
       ]                            b  b   b  b
10     +                                         b   b  b   b  b  b
       ]                                                              b
       ]
5      +
       ]   a
       ]      a  a  a  a   a
0      +                     a  a  a  a  a   a  a  a  a  a   a  a  a  a
     --+---------+---------+---------+---------+---------+---------+--
        0       15000                45000                 75000

                           Size of Data Set
```

Figure 2. Percent of Unique Households Versus Size of Data Set. The symbols used in this figure represent the entropy of the variable "payment". a: Lowest Entropy of the Variable "Payment", ..., f: Highest Entropy of the Variable "Payment"

```
Percent]           f
Unique ]                 f     f
42.5   +                           f
       ]                                f
       ]
       ]                                       f  f   f
40.0   +                                                   f  f
       ]         e                                               f
       ]               e
       ]                    e
37.5   +                            e
       ]                                e     e
       ]                                          e   e
       ]                                                 e   e
35.0   +           d                                              e
       ]               d     d
       ]                          d
       ]                                 d     d
32.5   +                                            d   d
       ]                                                 d   d   d
       ]
       ]
30.0   +
       ]
       ]         c
       ]               c
27.5   +                      c
       ]                          c      c
       ]                                      c   c
       ]                                              c   c
25.0   +                                                      c   c
       ]
       ]           b
       ]               b     b
22.5   +                            b      b
       ]                                      b   b
       ]                                              b   b   b
       ]                                                         b
20.0   +
       ]           a    a
       ]                     a    a
       ]                                 a     a
17.5   +                                            a   a   a   a
       ]                                                         a
       ]
       ]
---+-------+-------+-------+-------+-------+-------+-------+-----
      50000            60000            70000            80000

                       Size of Data Set
```

Figure 3. Percent of Unique Households Versus Size of Data Set. The symbols used in this figure represent the overall entropy of the original data set. a: Lowest Value of Overall Entropy, ..., j: Highest Value of Overall Entropy

```
Percent]
Unique ]
   ]
60 +       j
   ]
   ]
55 +
   ]
   ]      i      j
50 +
   ]      h              j
   ]
45 +                              j
   ]             i                    j
40 +        g       h                            j
   ]                    i                            j
   ]                                                        j      j
35 +                      h       i
   ]
   ]             g              h       i
30 +                                            i
   ]                     g              h             i
   ]                                          h              i      i
25 +        f                     g                     h      h
   ]                                   g                            h
   ]                                          g
20 +                                                   g
   ]             f                                           g      g
   ]
15 +                      f
   ]                            f
   ]                                   f      f
10 +                                                   f      f      f
   ]      e
   ]      d       e
 5 +               d      e       e
   ]      c              d       d      e       e      e       e      e
   ]      b       c      c       c      c       d       d      d       d
 0 +              a       a      b       b      b       b      b       c
  -+-------+-------+-------+-------+-------+-------+-------+-------
   0              20000              40000             60000    70000

                    Size of Data Set
```

NOTE: 14 OBSERVATIONS HIDDEN

REFERENCES

Bethlehem, J. G., Keller, W. J., and Pannekoek, J. (1990), "Disclosure Control of Microdata," Journal of the American Statistical Association, Vol. 85, pp. 38-45. (An earlier version of this paper appeared in (1988), "Disclosure Control of Micro Data," Proceedings of the Bureau of the Census Fourth Annual Research Conference, pp. 181-192.)

Greenberg, B. (1990), "Disclosure Avoidance Research at the Census Bureau," Proceedings of the Bureau of the Census Sixth Annual Research Conference, Bureau of the Census, Washington, D.C., pp. 144-166.

Greenberg, B., and Voshell, L. (1990), "The Geographic Component of Disclosure Risk for Microdata", Statistical Research Division Report Series, Census/SRD/RR-90/12, Bureau of the Census, Statistical Research Division, Washington, D.C.

Voshell, L. (1990), "Estimating the Number of Population Uniques Using Information from a Sample", Statistical Research Division Report Series forthcoming.

Willenborg, L.C.R.J., Mokken, R. J., and Pannekoek, J. (1990), "Microdata and Disclosure Risks," Proceedings of the Bureau of the Census Sixth Annual Research Conference, Bureau of the Census, Washington, D.C., pp. 167-180.