DETERMINISTIC MODIFICATIONS OF MICRODATA: BALANCING DISCLOSURE RISK VS. INFORMATION LOSS

Paul B. Massell, U.S. Census Bureau
SRD, Room 3209-4, Washington, D.C. 20233

Key Words: Microdata, Disclosure Risk, Information Content, Entropy

Introduction[1]

In recent years, with the rapid increase of data sources on the Internet, there has been increasing concern about the possibility of matching records from publicly released survey microdata files with records from other sources that are readily available to a data intruder. The concern is that the data intruder would be able to use certain variables found on a typical survey microdata record that does not contain explicit identifiers (such as name, address, telephone number, or SSN) and match these variables to those found on some record in a different data source that does contain explicit identifiers. The data intruder would then be able to identify the respondent associated with the matched microdata record. This identification would cause various types of problems. If the survey were conducted with a promise of confidentiality to the respondents, it would violate that promise. If the identification were publicized, it would likely decrease the response rates on future surveys conducted by the organization conducting the survey and probably also by other survey organizations as well. The seriousness of such disclosures is also related to which sensitive items are on the survey record and thus can be acquired by a data intruder. Income, living arrangement information, and health history, are often considered to be among the more sensitive items that appear on demographic surveys.

With disclosure risk reduction as the motivation, we explore information-reducing methods that are simple to implement and to fine-tune. We confine ourselves to an often used method, coarsening or broadening (combining) of categories. We also mention how measurement error affects the choice of categories. Several other methods for information-reduction have been explored in the literature with the goal of determining their effect on disclosure risk. What we believe is new here, is the application of specific measures of disclosure risk and information loss to a specific survey file. Information loss is measured using Shannon entropy.

In general, there are several reasonable definitions of disclosure risk and for each there is often more than one way to estimate the associated measure. We chose one commonly used measure, namely the fraction of records in the microdata file that are unique in the population associated with the file. Two points should be made here. First, the estimation of this fraction of uniques is not done with the set of all file variables but only with the largest subset of the file variables that we suspect are accessible to a data intruder. Such a subset is called a 'key' for the disclosure problem for the given file. Second, since the survey is typically not a census, we need to estimate the fraction of unique records in the population from the fraction of records that are unique in the sample.

The last goal of the paper is to express the tradeoff of maximizing information while insuring that disclosure risk is below an acceptable threshold. We express this in various ways as an optimization problem. There is similar work in the literature (ref: Z, DE).

I. Assessing information associated with individual variables

a. Data quality; its effect on information

Data quality, specifically the response error of a given variable, is clearly related to the information content of the variable. Response error may be viewed as a data modification performed prior to data collection. It's a complicated type of data modification having both deterministic and stochastic components that are reflected in the response bias and variance. Response error is a complex topic which has been the subject of entire books (ref: BGLMS, F, LK). We are including a brief discussion of this topic because it affects decisions about optimal interval sizes for continuous variables and the optimal number of categories for discrete variables. Normally one strives in a survey for the highest quality data possible. However, there is at least one positive aspect to poor data quality; it often makes the data intruder's effort to match a survey record with an external

[1]This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau Publications. This report is released to inform interested parties of research and to encourage discussion

database record more difficult (ref: WdW, p.vii). (One exception to this is the case of biased data with the same bias on many files.)
A rough estimate of the measurement error for each of the key variables is needed to determine the set of 'reasonable categorizations'. This idea is discussed below and then is used in the formulation of the optimization problem. For example, using a very narrow interval for categorizing annual income (say, $10) is not reasonable due to the ambiguity of the term 'income' and the various levels of precision with which it is reported on surveys.

b. Coarsening microdata to reduce disclosure risk
Coarsening of data can be applied to both continuous and discrete variables. We will give an example of coarsening for each variable type and then discuss ways to measure the associated information loss. Coarsening of data has been a common method for data reduction for many years. It is known by many other names depending on the type of variable being coarsened, e.g., collapsing (or rolling up or combining) of categories, reduction of detail, or downscaling. For a general discussion of coarsening, including top and bottom coding, see EURO.

i. Continuous variables
For continuous variables, coarsening means decreasing the precision at which observed values are reported. For example, income is, for practical purposes, a continuous variable. As mentioned above the data quality of responses on these items may be low for various reasons. To the extent that the respondent tries to report an accurate response but fails to, due to exclusion or inclusion of some minor source of income, or some minor misunderstanding in the time interval for the income, it makes sense to conclude that income is reported only approximately. If this assumption is made, it can be expressed to some extent by using income intervals rather than a precise value. In so doing one is essentially discretizing the income variable, i.e., converting a continuous one to a discrete one. Discretizing is one way of expressing one's uncertainty of the true value. One could argue that there are better ways to express the uncertainty in the variable due to measurement error. In particular, one could try to construct an interval about each response that reflects a 95% probability that the true value lies in the interval. If this were done, we would have, in general, a number of intervals equal to the number of responses, and the intervals would overlap. One could derive a measure for the uncertainty of a given response based on the width of its 95% interval.
However, for simplicity, we restrict ourselves in this paper to the traditional way of constructing uncertainty intervals, i.e., a set of intervals that form a **partition** of the value space from the bottom code or value (often zero) to the topcode or value.
Once the decision to discretize is made, one must decide on the interval widths. One may take uniform intervals, e.g. for income one may take $1000 intervals, k * $1000 to (k+1) * $1000 for k=0,1,2,3,...,(topcode/1000)-1. The upper bound of such intervals may be chosen large enough to include all the responses, or it may be topcoded. Whether to topcode (or bottom code), and if so, at what level, are important decisions affecting disclosure risk. Non-uniform intervals are acceptable although they may be somewhat difficult to implement.

ii. Discrete variables
For discrete variables, coarsening means combining categories. For example, a question may have several categories (possible responses) but we may wish to combine some of them prior to analysis. Similar to the decision for interval construction for continuous variables, it is necessary to decide how many categories to form and how small a proportion to allow for a single category. Similar to including thin tails for continuous variables, maintaining certain small proportion categories can greatly increase disclosure risk.

iii. A potential increase in population information
The practical information about a variable may decrease only slightly under a discretization. Even in cases in which the practical information does decrease significantly after a discretization, further computations involving the (now) discrete variable may lead to results that are insignificantly different from the result that would be produced based on the original (continuous) variable. On rare occasions, discretization may even have a positive effect on the information gained about the variable as a result of some process. Suppose, for example, the computational goal is to form a histogram for a continuous variable that approximates closely the true density function. Suppose the sample file contains records for only a small fraction of the population. Then moderate coarsening may have only a negligible effect on the histograms. Indeed, because overfitting a sample can lead to misleading results, there will be situations in which the coarsened data will actually lead to a histogram that better approximates the population density than one based on a very fine partition (ref: LZ).

iv. The entropy measure of information content for a single variable
There are various definitions of entropy. Shannon entropy $H = -\sum p_i \log(p_i)$ is commonly used in books on mathematical (statistical) information theory

(ref: A, K). The key ideas about entropy as a measure of information content are: (1) entropy represents the average uncertainty removed by a sample of the variable (2) for a given discretization of a variable, the more uniform the distribution of the probability function, the greater the entropy (3) for a given sample of a variable, the finer the discretization used the larger the entropy. With the more familiar term "bin width" as used for histograms, we can say the smaller the bin widths, the larger the entropy. Other researchers have mentioned using entropy as a measure of information content or loss (ref: WdW, p.138, DE, Z)

c. Other deterministic methods for decreasing information (local suppression)

Local suppression means blanking of data in microdata variable fields. This is sometimes done when the values are thought to be extreme enough to pose a disclosure risk. The effect of this information reduction depends on the categorization that has preceded the local suppression. If, for example, an extreme value is in a 1-cell (i.e. a cell with a value derived from only one respondent), the information reduction may be significant. Entropy may be used to measure this.

II. Measuring combined information from two or more variables

We first note that the general statements about entropy for a single variable stated above generalize easily to the multivariate case. The main change is that discretizations here produce table cells rather than one-dimensional bins as mentioned above. Thus, H measures the information content of a multidimensional table. The joint uncertainty of variables $X_1, X_2, \ldots, X_n$,

$$H(X_1, X_2, \ldots, X_n) =$$

$$- \sum_{X1, X2, \ldots Xn} p(x_1, x_2, .., x_n) \cdot \log(p(x_1, x_2, .., x_n))$$

eq.(1)

over the table formed by a given discretization of the X_i's. Note that H > 0. A basic theorem states that:

$$H(X_1, X_2, \ldots, X_n) \leq H(X_1) + H(X_2) + \ldots + H(X_n)$$

eq.(2)

with equality if and only if the X_i's are independent (ref. A, p.19).

Using (2) with the knowledge or assumption that a key variable X that is to be coarsened is independent of the other key variables, the informational effect of coarsening X can be gotten simply by calculating one-dimensional H(X) before and after the coarsening. Typically, however, the key variables are dependent and then we need to calculate the **joint** H before and after the coarsening of each variable to measure the informational impact.

Now we present an example of change in information content when two variables in a 4-d table are coarsened. Let's say we have four key variables that are categorical (possibly including ones that have been discretized). Label them X1,X2,X3,X4. Suppose the first two are being considered for coarsening. Let Y1,Y2 be the coarsened variables corresponding to X1,X2. Then, unless we know that Y1 and Y2 are each independent of the other variables, we need to compute H(X1,X2,X3,X4), H(Y1,X2,X3,X4), H(X1,Y2,X3,X4) and H(Y1,Y2,X3,X4). The maximal information content, as measured by H, occurs for the most finely discretized (i.e. least coarsened) set of variables, viz. X1, X2, X3, X4.

III. The disclosure risk and information content tradeoff

a. The effect of coarsening on disclosure risk

i. Estimating the fraction of uniques in the population

One commonly used measure of disclosure risk is the fraction of records in a survey file that are unique in the population (i.e. survey frame). A mathematical statistical problem of great interest over the past decade is how best to estimate this fraction. In our calculations below, we used a formula that uses only the fraction of uniques in the **sample**. (Other, perhaps more accurate 'extension' formulas, use more equivalence class data from the sample, e.g., the number of k cells for k small). There is limited numerical evidence that a simple formula based only on uniques in the sample provides a good estimate of the fraction of uniques in the population if the population is no greater than 10 times the size of the sample(ref: GZ). Thus, with a given microdata file that constitutes a sample of some population, one may compute the number of uniques for a given sample file and key, and then one may use that number to estimate the fraction of uniques in the population. For simplicity, however, in our examples below we will consider our samples to equal our population so that use of a 'extension' formula will not be needed. This simplification does not detract from the results below; they require only some (reasonable) measure of disclosure risk for the given microdata file. How well this risk can be estimated is a separate question. (For examples of the 'extension' formulas see ref GZ or M).

ii. Viewing uniques in a table

It is useful to view the records as distributed in a (frequency) count table whose dimensions are determined by the key variables. This assumes all the variables are discrete; i.e., the continuous ones have already been discretized prior to table formation. With this "table-view" of the file, the unique records lie in 1-cells, i.e., they are the only records with a key vector that matches the cell's categories.

In terms of this "table-view," in order to minimize disclosure risk, we should coarsen the variables to minimize 1-cells and probably also 0-cells. Suppose we have a table of dimension two or more formed by crossing discrete variables each with a finite number of categories. In order to reduce sample uniques, we need to determine whether there is a subset of the set of table variables that seem to be most responsible for creating small cell sizes. Most likely these are variables that have very non-uniform frequency distributions. In particular they probably have some categories with very low proportions. We can only say 'probably' because it is possible to get a one count or even a zero count in a cell which is the intersection of two variables with at least moderate size frequencies for each category, if the variables are correlated. This can occur in a table of any dimension; i.e., a cell can represent a small frequency even though none of its corresponding marginal frequencies is small. Nevertheless, it is often the case that combining categories in some of the table's variables will eliminate most of the small cells, in particular the 0 and 1- cells.

iii. Disclosure risk associated with 0-cells

The presence of 0- cells may be quite informative; e.g. it may show (or suggest) functional relationships among the variables . However, from a disclosure point of view, zero cells may pose a risk. In particular, zero cells at the upper and lower ranges of a given variable indicate upper and lower bounds on the values of the given variable for the sample data. Of course, this may not extend to the population but it is suggestive if the sample is a large fraction of the population. Knowing that any person with a given combination of demographic-geographic variables has an income (or other sensitive variable) in some narrow range may represent a disclosure if it is easy to determine if a person is a member of the group defined by the given combination of demographic-geographic variables. This situation is called an **attribute disclosure**. Note that this type of disclosure does not require a matching file (ref: WdW, p.92).

b. The dilemma and the need for a tradeoff

In general, finer discretizations yield more informative views of the data. Thus any proposed measure of information content should have the property that it increases as the discretization becomes finer, or equivalently, it should decrease as coarsening is increased. The entropy H, that we are using here as our measure of information content, has this property. However the number of uniques increases as the discretization is made finer. This leads to the dilemma: if there a way of balancing the dual requirements of high information content and low disclosure risk? If so, can it be found as the solution to a mathematical optimization problem? What are possible formulations of this problem?

The tradeoff

The goal is to have the optimal balance of an informative and low disclosure risk microdata file. This implies an optimal balance of information content as measured by an information function applied to the table generated by the key variables, and disclosure risk as measured by 1-cells in the same table. Before formulating this tradeoff using functional notation, we present results from an American Housing Survey (AHS) national microdata file. This file consists of an approximately 1 in 2000 sample of all housing units in the United States.

c. Examples from the AHS public use microdata file

Consider the 1997 AHS national microdata file (ref: HUD-WEB, HUD) Let the key consist of the variables: SEX, RACE, AGE, SALARY. There are many other variables one could select for the key but our ideas can be illustrated with these four. The full sample file has 102,761 records. All records have data for SEX, RACE, and AGE. However, 21,194 records have missing data for SALARY. Most of these are for persons under 16. For simplicity we will eliminate all the records with missing data; this leaves us a file with 81,567 records.

The frequency tables for SEX and RACE are:

SEX	Frequency	Percent
Male	38830	47.6
Female	42737	52.4

RACE	Frequency	Percent
White	67500	82.8
Black	8845	10.8
Amer Ind., Aleut	511	0.6
Asian / Pacific .Is.	2954	3.6
Other	1757	2.2

AGE is an integer from 0 to 90 (i.e. 90 is a topcode).

We will consider two coarsenings of AGE.

AGE3: 0, 3, 6, 9,...., 90
AGE5: 0, 5, 10, 15,..., 90

We present the AGE5 distribution below. The data for SALARY were missing for all persons under 14 so there are no age groups below 15. A code of 3*i for AGE3 represents those in the sample with ages 3*i, (3*i)-1, and (3*i)-2. Similarly for AGE5. The 0 category includes only 0.
We consider four discretizations of the continuous variable SALARY. We use 0 as a category, and then equal width intervals of sizes $1000, $2000, $5000, $10000 up to the topcoded value of $100,000. We present the SAL10000 distribution below.

AGE5	Freq.	%	SAL10000	Freq.	%
15	3349	4.1	0	30615	37.5
20	7213	8.8	10000	13772	16.9
25	6347	7.8	20000	11360	13.9
30	7225	8.9	30000	9715	11.9
35	8063	9.9	40000	6386	7.8
40	8476	10.4	50000	3754	4.6
45	7804	9.6	60000	2155	2.6
50	7201	8.8	70000	1200	1.5
55	5397	6.6	80000	825	1.0
60	4437	5.4	90000	404	0.5
65	3832	4.7	100000	1381	1.7
70	3664	4.5			
75	3461	4.2			
80	2550	3.1			
85	1570	1.9			
90	978	1.2			

Now we describe eight tables each of which is generated by SEX, RACE, AGE3 or AGE5, and one of the SAL variables.

Variables:	#(cells)
1. SEX, RACE, AGE5, SAL10000	2090
2. SEX, RACE, AGE5, SAL5000	3990
3. SEX, RACE, AGE5, SAL2000	9690
4. SEX, RACE, AGE5, SAL1000	19,190
5. SEX, RACE, AGE3, SAL10000	3410
6. SEX, RACE, AGE3, SAL5000	6510
7. SEX, RACE, AGE3, SAL2000	15,810
8. SEX, RACE, AGE3, SAL1000	31,310

Let NZ denote the number of non-zero cells.

The units of the entropy H are bits per record since we used logarithms to the base 2 for calculating it. Since H is the expected information revealed by a sample record, the units are bits 'per record'.

TABLE: INFO-DRM points

	NZ	#(1-cells)	%uniques	H: (bits/record)
1.	(1730)	352	0.43	8.56
2.	(3157)	828	1.02	9.29
3.	(4765)	1490	1.83	9.75
4.	(1627)	307	0.37	8.66
5.	(2557)	607	0.74	9.24
6.	(4513)	1353	1.66	9.95
7.	(6630)	2301	2.82	10.40
8.	(1078)	173	0.21	7.97

d. Tradeoff Thresholds and The Tradeoff Rectangle
We have a multiple criteria optimization problem. Perhaps the simplest way to treat such a problem is to establish a threshold for each criterion. For this tradeoff problem we need to establish an information threshold, T-info, and a disclosure risk measure threshold, T-drm. A data modification and the key formed by the associated modified variables are 'acceptable' only if both :

(i) Info > T-info and (ii) Drm < T-drm

A data providing agency might establish a rule that a survey file can be released only if all keys formed from intruder accessible variables, either original or modified, are acceptable. In this paper, we have confined our discussion to the modification of coarsening, the disclosure risk measure to the fraction of sample records that are unique in the population, and the information measure to entropy.

If two or more modifications have been applied to the sample data, e.g the eight discretizations listed in the tables above. we can plot the values for each of the criteria in a graph that has one dimension for each criterion (see graph below). Clearly the conditions Info> T-info and 0 < Drm < T-drm imply the acceptance region is an infinite rectangular strip in the plane. It is possible that for a given value of one criterion there will be more than value of the other criterion. One simple way of selecting the "optimal" value is simply to take one of the acceptable points with the largest value of T-info. In other words, select a modification which has the largest information content among all modification that have an acceptably small value of disclosure risk. Of course, one could instead select the point with the minimum disclosure risk with an acceptably high value of information. In many cases, we would suppose, these two strategies lead to nearly the same result if a large of points in the upper right region of the Info-Drm rectangle are calculated. In the graph below, all points satisfy the T-info condition, and all but one satisfy the T-Drm condition. Therefore point 6 in the INFO-DRM table for which (Info, Drm) = (9.95, 1.66) is optimal.

It may be possible to determine by curve fitting or by theory, some functional relationship between Drm and Info for (a 1-parameter) continuously varying set of modifications (ref: ZE). In that case, we would expect that Drm would be an increasing function of Info. This property expresses the basic tradeoff between disclosure risk and information in functional form. When two or more sets of modifications are considered, as in this example, one can expect only a roughly monotonic relationship (see fitted line in graph).

Our values for T-drm and T-info in the graph were arbitrarily chosen. Obviously there must be an objective way of determining these values. One way to determine the latter is to survey data users on whether a modified microdata file has sufficient information to be useful.

IV. Conclusions

We have discussed some general ideas for deciding which modification of the raw microdata has the optimal balance of disclosure risk and information content. There are several aspects of this problems that need to be explored in order to convert these ideas to a practical procedure; (1) need to decide on one, or at most a few, measures of disclosure risk; (2) need to determine ways of calculating information loss for modifications other than coarsening, e.g., noise addition and swapping; (3) need to find a justifiable way of establishing thresholds for disclosure risk and information; (4) need to determine if use of a multiple criterion threshold function is a better way to determine optimal balance; if so, need to construct an optimal such function.

REFERENCES:

A: Ash, R., Information Theory, Wiley 1965.

BGLMS: Biemer,Groves, Lyberg, Mathiowetz, Sudman, Measurement Errors in Surveys, Wiley, 1991

C: Cox, L. Statistical Disclosure and Disclosure Limitation, JPSM course notes, Dec. 1998

DE: DeWaal, AG and Willenborg, LCRJ, Information loss through global recoding and local suppression, Netherlands Official Statistics, Spring, 1999

EURO: Manual on Disclosure Control Methods, EUROSTAT, Brussels, 1996.

F: Fuller,W.A.,Measurement Error Models, Wiley, 1987

GZ: Greenberg, BV, and Zayatz, LV, Strategies for measuring risk in public use microdata files Statistica Neerlandica , 1992, vol 46, nr. 1, pp. 33-48

HUD: Codebook for the American Housing Survey: Vol. 3:1997 SAS Files & Questionnaire (HUDWEB) www.huduser.org/datasets/ahs/ahsdata97.html

K: Kullback, S. Information Theory and Statistics, Dover, 1997.

LK: Lessler, J.T., Kalsbeek, W.D., Nonsampling Errors in Surveys, Wiley, 1992

LZ : Linhart, H., Zucchini, W. Model Selection, Wiley, 1986

M: Massell, P B, Assessing the Statistical Disclosure Risk of a Demographic Microdata File, Conference Proceedings 1999 FCSM Mtg.

WdW: Willenborg, L, de Waal, T., Statistical Disclosure Control in Practice, Springer, 1996

Z: Zaslavsky, AM, Horton, N J, Balancing Disclosure Risk against the Loss of Nonpublication, JOS, Dec, 1998

ZE: Zeleny, M, Multiple Criteria Decision Making, McGraw Hill, 1982.