Producing a Public Use File: A Case Study

Kenneth Rasinski, Jeffrey Timberlake, Lisa Lee, and Javier Porras National Opinion Research Center, Chicago, IL, 60637
and
Jeri Mulrow, Ernst & Young

October 17, 1997

I. INTRODUCTION

Government agencies face a challenge in balancing the need to provide microdata for public policy and research with the need to protect confidentiality. The Substance Abuse and Mental Health Services Administration (SAMHSA) is facing that challenge today. Recently SAMHSA has undertaken a large scale project to release microdata public use files. The statistical literature on disclosure avoidance provides many broad recommendations and several choices of techniques, but no overall consensus on the "best" method to produce an analytically useful, but low disclosure risk, microdata file. This paper provides a case study of the process taken by the National Opinion Research Center (NORC) in creating public use files for SAMHSA. In this paper we discuss the practical and statistical problems that we faced while producing the files. We explain, in a detailed manner, the process and measures used to create the public use file, or PUF, and we end with our views on the advantages and disadvantages of using this particular approach. Because of disclosure concerns, we will not mention which SAMSHA drug data set is being referred to in this paper. From hereafter, we will refer to it as the PUF.

In an attempt to develop an appropriate technique to produce the PUF, each step or decision in the process was weighed, in order of importance, in terms of: (1) disclosure risk, (2) analytic utility, and (3) level of effort/resources required to implement. Definitions for each of these terms are provided below:

Disclosure risk. Disclosure risk is the likelihood or probability that the identity of a respondent will be discovered. This is of prime importance when releasing a file for public use, and must be as low as possible. Disclosure risk has been defined in the literature in a direct sense as simply the probability of "identifying individuals in released statistical information,"[1] and in an inferential sense as the probability that "publication of statistical data makes it possible to determine characteristics of specified individuals more accurately than is possible without access to this statistical information."[2]

Analytic utility. This term refers to the ability of researchers to "get what they want" out of the data. To a large degree, the interests of disclosure protection and the interests of analytic utility are at odds. Researchers would like the greatest amount of detail in the data as possible but the demands of disclosure protection require some degree of shielding of the data. The goal was to provide a PUF dataset that preserve as much analytic utility as possible while keeping disclosure risk low.

Level of effort/resources required. As is always the case, neither time nor money were unlimited in this project. It would not be beneficial to devise an extraordinarily complicated sampling design or a complex scheme for collapsing or recoding variables or an intricate method of imputing values to problematic observations if the process would have taken a very long time to develop or exhaust resources. Rather, the purpose was to find the practical compromises that lead to high analytic utility, low disclosure risk, and a feasible schedule for the public use files.

II. OTHER SOURCES OF DATA

Identification or disclosure generally occurs in one of two ways. The first way is by "spontaneous recognition" whereby the characteristics of an individual are unique in the file and allow recognition. The second way is by "matching". Records on the released file are matched to records on other public files and an identification is made possible. Both types of disclosure should be prevented to as great of an extent as feasible. This paper mainly focuses on the steps NORC took to prevent spontaneous recognition from occurring. However, it should be noted that NORC did spend extensive time insuring that file matching across data sets would be a difficult task for any intruder. The remainder of the paper will discuss the procedures implemented to combat spontaneous identification from taking place.

III. CLASSIFICATION OF PUF VARIABLES

The variables were classified into two categories: key and non-key variables. The key variables were identified as those carrying high disclosure risk and high analytic utility. The remainder of the variables were classified as non-key variables.

In classifying the variables, there was a fair amount of subjectivity involved. For example, it can be argued that if a variable did not have a high degree of analytic usefulness, then it would not be in the file to begin with. Although a valid statement, the classification of the variables was done with respect to each other.

The significance of this classification scheme will soon become apparent.

IV. DISCLOSURE TECHNIQUES

The literature presents several techniques that can be used to reduce the risk of disclosure in a microdata file. This section discusses various methods and their applicability to the creation of public use files.

Dropping Variables

The first step taken to reduce the disclosure risk was the dropping of certain variables. In dropping any variable, it followed immediately that the analytic utility of the file would also be reduced. The challenge was to carefully chose variables to eliminate from the file so as to minimize the reduction in analytic utility as much as possible. Among the non-key variables, those that offered low analytic utility were immediately dropped from the file. In particular, those variables with a lot of missing data proved to be of little use. Also, those non-key variables that carried a high disclosure risk were also dropped. The remainder of the non-key variables were considered individually, Some were dropped from the PUF, while some were kept. The file began with 38 non-key variables. After examining each of them, 27 non-key variables remained.

The key variables were given special attention. Because of their high analytic utility, there was a strong preference to avoid having to drop any of them. However, this preference was not allowed to interfere with NORC's responsibility to produce a disclosure-safe file; consequently, some of the key variables were dropped. Like some of the non-key variables, key variables with a high proportion of missing data were eliminated. The raw file began with 15 key variables, but only eleven were to be included in the PUF.

Recoding/collapsing schemes

The second step involved collapsing variables into broader categories: some of the continuous variables were grouped into categorical variables; some categorical variables were collapsed into even broader categories. Particular values of the variables were recoded into "unknown" or "missing" categories, or top- or bottom-coded to obscure the rare values on the file.

In recoding/collapsing, detail was being lost and, consequently, the file was once again losing analytic utility. It was, however, the only way that would allow some of the variables to be kept.

Small Cell Collapsing

The third step involved the elimination of small cells. After this step was completed, each non-empty cell had at least three records in it, a direct result of the small cell elimination procedure. This will be discussed at length in the following section.

Sampling

The final step was to sample records from the population file for the PUF. Sampling is a simple and easy method to reduce the risk of disclosure in a microdata file while preserving the integrity of the data [3]. Three schemes were considered for the creation of the PUF: simple random sampling, cluster sampling within providers, and cluster sampling within clients. Based on the need to pick a scheme that would be easy to implement, we chose a simple random sampling scheme.

The sampling fraction, *f*, was set at 1/i, where i was less than ten. After all small cells had been eliminated, the file was sorted be state, PMSA and a random number assigned to each record. After which, the first record and every i-th record thereafter were sampled into the PUF.

V. SMALL CELL ANALYSIS

As was briefly described above, the elimination of small cells was the third step in producing the PUF. Because this step is the most complex of the four, an entire section is dedicated to explaining the procedure.

Definitions

A *cell* is defined as a set of records with matching data values along the key variables, and a small cell is defined as a cell of size one or two. We chose to follow the general rule used by several federal statistical agencies [4] that any cell in a multivariate crosstabulation have zero or at least three observations in the full file in order for data to be released publicly. Thus, our goal was to produce a population data file with cells no less than size three from which to sample.

Treatment of Small Cells

Crosstabulations on the 11 key variables indicated that only about 11.5% of records in the preliminary data were in small cells. *That is, for 88.5% of records, the recoding scheme would provide adequate protection from disclosure. Only the 11.5% of records in small cells would require further treatment to reduce the risk of identification.*

Small cells were collapsed into non-small cells by suppressing the data values on two adjacent small cells which differed. For example, if two records --that is, two singleton cells-- differed only in their values on the age and gender variables, then setting the values on these variables to missing would make the two records identical, or create a single cell of size two. If cells that are as alike as possible with respect to the key variables are combined in this manner, then much of the original data would be preserved.

Algorithm to Eliminate Small Cells

The algorithm that was used to eliminate small cells is described below. An example will be included that will include only four key variables, which will be named VAR1, VAR2, VAR3, and VAR4.

Step 1. *Extract small cells from the PUF and sort by the key variables.* Sorting by the key variables made a lot of sense; it served a dual purpose. The first reason centered around minimizing data suppression. Sorting the small cells by the key variables assured the placement of *similar* cells next to each other. As the following steps will show, the collapsing algorithm compared adjacent cells to each other and collapsed them if they were *similar*, that is, if they matched on a high number of data values along the key variables. Thus, sorting by the eleven key variables in effect "minimized" the loss of information due to data suppression as similar cells were collapsed into each other. The second reason for sorting by the eleven key variables was controlling *where* the data suppression would take place: data values for variables low in the sorting order would be prone to loss of data due to suppression, while the variables receiving high priority were placed high in the sorting order were they were protected, but not immune, to suppression. In our example, we will begin with a file that is sorted in the following order: VAR1, VAR2, VAR3, VAR4.

Step 2. *Set maximum distance criterion for adjacent cells to be collapsed.* The distance measure is a count of the number of key variables that two adjacent cells differ from each other. For the PUF this measure is an integer ranging between 1 and 11 (the number of key variables). In the example being shown, the distance criterion would require a value between 1 and 4. A distance criterion of 1, which is adopted here, means that the two adjacent cells must match on three of the four key variables.

Step 3. *Compute a distance measure for two adjacent cells on the file, beginning at the top of the file and working down the file.* If the distance measure does not meet the criterion, then one must compute the distance for the next set of cells. If the measure meets the criterion, then one proceeds to step 4. Table 1 shows the distance being computed for cells 2 and 3. Cells 1 and 2 are too distant (differing on both VAR2 and VAR3); thus, cell 1 will remain unaltered in this iteration. The following set of cells (2 and 3), however, meet the criterion, differing only on the VAR4, thus they are collapsed into each other.

Table 1: Cells 2 and 3 meet Distance Criterion

Cell Num	VAR1	VAR2	VAR3	VAR4	Cell Size	Dist
1	1	1	1	1	1	*
2	1	2	2	1	2	2
3	1	2	2	2	1	1
4	1	3	3	3	1	
5	1	3	4	1	1	
6	1	3	5	2	1	

Step 4. *Collapse the two adjacent small cells by setting data values to missing for all variables that do not have matching values.* In Table 2, collapsing cells 2 and 3 produces a non-small cell with three observations. Note the change in the distance for cell 3, indicating that cell 2 and 3 have collapsed into each other. The process begins again at Step 3 with the next set of cells and continues until the end of the file is reached. The remaining cells are not collapsed since none of them meet the distance criterion, as Table 3 shows.

Table 2: Collapse Cells 2 and 3 into each other

Cell Num	VAR1	VAR2	VAR3	VAR4	Cell Size	Dist
1	1	1	1	1	1	*
2	1	2	2	*	2	2
3	1	2	2	*	1	0
4	1	3	3	3	1	
5	1	3	4	1	1	
6	1	3	5	2	1	

Table 3: Cells 4,5 and 6 are left unaltered

Cell Num	VAR1	VAR2	VAR3	VAR4	Cell Size	Dist
1	1	1	1	1	1	*
2	1	2	2	*	2	2
3	1	2	2	*	1	0
4	1	3	3	3	1	3
5	1	3	4	1	1	2
6	1	3	5	2	1	2

Step 5. *Extract remaining small cells and begin again at Step 1.* Table 4 shows the reduced small cells, with cell 2 and 3 removed. Distances, although unaltered, actually are recomputed, since cell 4 has a new "neighbor" in cell 1 after the removal of cells 2 and 3.

Table 4: Remaining Small Cells after First Iteration

Cell Num	VAR1	VAR2	VAR3	VAR4	Cell Size	Dist
1	1	1	1	1	1	.
4	1	3	3	3	1	3
5	1	3	4	1	1	2
6	1	3	5	2	1	2

Table 4 also shows that with each iteration of the algorithm, the distance criterion must be increased, as fewer and more disparate small cells remain in the small cell file.

Comparison of File Before and After Small Cell Treatment

Our file began with a total of 1,433,544 records, of which 164,949 were members of small cells. Table 5 shows the pre-collapsed distances of the small cells. It can be seen that by setting the distance criterion to one, the first iteration would result in a cell collapsing percent of about 33%.

Table 5: Distribution of Distances of Small Cells

Distance	Percent Frequenc	Cumulative Frequency
1	33.10	33.10
2	33.90	67.00
3	20.00	87.10
4	8.70	95.80
5	3.00	98.80
6	0.90	99.60
7	0.30	99.90
8	0.10	100.00
9	0.00	100.00
10	0.00	100.00
11	0.00	100.00

By the fourth iteration, only 3627 small cells containing 5240 records remained, about three percent of the initial number of records that were originally members of small cells. Table 6 illustrates the reduction of records that are members of small cells after each iteration.

Table 6: Summary of Collapsing Iterations (x10,000)

	Iteration	0	1	2	3	4	5
Distance	Criterion	*	1	2	3	4	11
Records	*Small*	16.5	10.5	4.9	1.8	0.5	0.0
	Big	0.0	6.0	11.6	14.7	16.0	16.5
	Total	16.5	16.5	16.5	16.5	16.5	16.5
Cells	*Small*	13.3	7.7	3.4	1.2	0.4	0.0

Now, suppression rates for two of the key variables are presented. Masking their real names, we will refer to these variables as VAR3 and VAR10. They were third and tenth in the sorting of the key variables, respectively. Table 7 summarizes these results.

Table 7: Suppression/Missing Percent Frequencies

	Missing/Suppression Rates	
	BEFORE	AFTER
VAR3	0.6	0.7
VAR10	49.8	52.3

It can be seen that the variable VAR10, which was considered the second lowest in importance, lost more data than variable VAR3, which was placed high in the sorting order. As had been predicted, variables placed high in the sorting order were protected from suppression.

How different was the distribution of the variable before and after the small cell suppression procedure. With the suppression of data taking place, questions about the bias that is introduced naturally arise. Table 8 suggests that the percent distributions are only slightly altered for VAR3 and VAR10. It should be noted that in computing the percent frequencies, missing and suppressed values are not included in the computations.

Table 8: Percent Frequencies Before and After Collapsing

		Percent Frequencies	
		BEFORE	AFTER
VAR3	category 1	71.5	71.5
	category 2	28.5	28.5
VAR10	category 1	11.9	10.5
	category 2	88.1	89.5

Again, it appears that suppression had little effect on VAR3, in large part because it had little data suppressed. However, for VAR10 the data suppression appears to have had a surprisingly low effect on the distribution of the variable.

VI. CONCLUSIONS

After running the four steps outlined in section III, a PUF was created with chances of identification of records being significantly lowered. The process for creating a PUF that we outlined offers a few advantages over other methods. One, it is easy to implement. Complicated statistical procedures, such as blurring, are not required. Two, the data loss due to suppression was minimal. The small cell elimination procedure collapsed cells that were similar and thus preserved much of the information. In the end, NORC produced a PUF with a relatively high level of utility and a low disclosure risk.

References

[1] Bethlehem, J.G., Keller, W.J. and Pannekoek, J. (1990). Disclosure Control of Microdata, *Journal of the American Statistical Association*, Vol. 85, No. 409, pp. 38-45.

[2] Dalanius, T. (1988). *Controlling Invasion of Privacy in Surveys*. Department of Development and Research, Statistics Sweden.

[3] Skinner, C.J., Marsh, C., Openshaw, S., and Wymer, C. (1990). Disclosure Avoidance for Census Microdata in Great Britain. *Proceedings of the Annual Research Conference of the Census Bureau*.

[4] Jabine, T.B. (1993). Statistical Disclosure Limitation Practices of United States Statistical Agencies, *Journal of Official Statistics*, Vol. 9, No. 2, pp. 427-454.