The Microdata Analysis System at the U.S. Census Bureau*

Jason Lucero and Laura Zayatz

U.S. Census Bureau, Statistical Research Division,
4600 Silver Hill Road, Washington, DC 20233-9100, United States
jason.lucero@census.gov

Abstract. The U.S. Census Bureau collects its survey and census data under Title 13 of the U. S. code, which promises to protect the confidentiality of our respondents. The agency has the responsibility to release high quality data products without violating the confidentiality of our respondents. This paper discuses a Microdata Analysis System (MAS) that is currently under development at the Census Bureau. We begin by discussing the reason for developing a MAS, and answer some questions about the MAS. We next give a brief overview of the MAS and the confidentiality rules within the system. The rest of this paper gives an overview of the evaluation of the universe subsampling routine in the MAS known as the *Drop Q Rule*. We conclude with some remarks on future research.

Keywords: Data Confidentiality, Remote Access Servers, Universe Uubsampling, Schur-Convexity.

1 Introduction

The U.S. Census Bureau collects its survey and census data under Title 13 of the U.S. Code. This prevents the Census Bureau from releasing any data "…whereby the data furnished by any particular establishment or individual under this title can be identified." In addition to Title 13, the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) requires the protection of information collected or acquired for exclusively statistical purposes under a pledge of confidentiality. In addition, the agency has the responsibility of releasing data for the purpose of statistical analysis. In common with most national statistical institutes, our goal is to release as much high quality data as possible without violating the pledge of confidentiality [1], [2].

This paper discusses a Microdata Analysis System (MAS) that is under development at the Census Bureau. The system is designed to allow data users to perform various statistical analyses (for example, regressions, cross-tabulations,

* This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 234–248, 2010.

generation of correlation coefficients, etc.) of confidential survey and census microdata without seeing or downloading the underlying microdata. We begin by answering some frequently asked questions about the MAS. We then discuss the current state of the system, including an overview of the types of data sets and statistical analyses that will be available in the system, as well as a brief outline of the confidentiality rules used to protect the data products generated from the MAS. We next give a brief overview of a recent evaluation of a particular confidentiality rule called the *Drop q Rule*. We end with remarks on future work.

2 Frequently Asked Questions about the MAS

2.1 Why Do We Need a MAS?

The Census Bureau conducts reindentification studies on our public use microdata files. In these studies, we attempt to link outside files that have identifiers on them to our public use files. We have found and fixed a few problems, but there is a growing concern that more problems will arise in the future because more and more data is becoming publicly available on the internet, and more people are using record linkage software and data mining in an effort to increase the amount of information they can work with. We are worried that we might have to cut back on the detail in our files and use more data perturbation techniques to protect them.

Another reason for developing a MAS is to allow data users to access more detailed and accurate information than what is currently available in our public use microdata files. For example, the data that can be accessed through the MAS could identify smaller geographic areas or show more detail in variable categories that are normally not shown in our public use files. Our goal for the MAS is to allow access to as much high quality data as possible [3], [4].

2.2 What Data Sets and What Types of Statistical Analyses Will Be Available on the MAS?

We will begin with data from demographic surveys and decennial censuses. Eventually, we would like to add establishment survey and census data as well as linked data sets. We will initially begin with regression analyses, cross-tabulations and correlation matrices. We will add other analyses within the MAS in the future.

2.3 Who Will Use the MAS and Will It Cost Anything?

The MAS will be used by people with needs for fairly simple statistical analyses, for example: news media, some policy makers, teachers, and students. Some users may feel the need to use the underlying confidential microdata for more exploratory data analysis. For that, they will have to continue to use our public use files. A final decision on cost has not yet been made. However, the current plan is to offer this as a free service through the Census Bureau's DataFERRETT service [5].

2.4 Will the Census Bureau Keep Track of Who Uses the MAS and What Queries Have Been Submitted?

We must mention that we are not sure if we are allowed to legally do this. We are investigating this. There are two possible reasons why we would want to do this. First, we would like to see how people are using the system, so we can make modifications and enhancements to improve the user's experience. The second reason would be for disclosure avoidance purposes, as these data may be used to help identify disclosure risks arising from multiple queries to the system.

3 A Brief Overview of the MAS and the Confidentiality Rules within the System

In 2005, the Census Bureau contracted with Synectics to develop an alpha prototype of the MAS, which was written in SAS. We also contracted with Jerry Reiter of Duke University to help us to develop the confidentiality rules within the system, and Steve Roehrig of Carnegie Mellon University to help us test these confidentiality rules. Some rules were developed and modified as a result of the testing. The alpha prototype used the Current Population Survey (CPS) March 2000 Demographic Supplement and the 2005 American Community Survey (ACS) as initial test data sets. The alpha prototype allowed users to perform cross-tabulations, ordinary least squares regression, logistic regression, and generate matrices of correlation coefficients. A beta prototype of the MAS is now being developed as part of DataFERRET [5]. Unlike the alpha prototype, this prototype is being written in R.

The MAS software is programmed with several confidentiality rules and procedures that uphold disclosure avoidance standards. The purpose of these rules and procedures is to prevent data intruders from reconstructing the microdata records of individuals within the underlying confidential data through submitting multiple queries. The confidentiality rules discussed within the next few sections are quite complex. This paper gives a brief overview of them. Much more detail can be found in [6] and [7]. In addition, we will only discuss the confidentiality rules for universe formations and for regression models. The confidentiality rules for cross-tabulations and correlation coefficients are still under development.

3.1 Confidentiality Rules for Universe Formation

On the MAS, users are allowed to limit their statistical analysis on a universe, or subpopulation, of interest. To form a universe on the MAS, users first select conditions on a subset of recoded variables, presented to the user in the form of metadata. These recoded variables within the metadata are categorical recodes of the raw categorical and numerical variables found in the microdata.

The category levels of the raw categorical variables within the original microdata set are coded directly into the metadata. To define a universe using a categorical variable, a user simply selects the categorical variable name and observed category level bin they see in the metadata. For example, if the user selects *sex* = 2 (female) from the metadata, they have defined their universe to be the subpopulation of all females.

Raw numerical variables are presented to the user as categorical recodes based on output from a separate cutpoint program. This cutpoint program generates buckets or bins of numerical values, and ensures that there is a pre-specified minimum number of observations between any two given cutpoint values [8]. To define universe using a numerical variable, users must a range of numerical variable values from the pre-specified bins they see from the metadata. For example, if the user selects *income* = 4 ($45,000 to $53,000) from the metadata, they have defined their universe to be the subpopulation of all individuals whose income is between $45,000 and $53,000. This furthers the confidentiality protection by preventing users from forming universes bases on a single raw numerical value. That is, users cannot define their universe to be *income* = $45,000, they must choose a range of values.

To define a universe on the MAS, users would first select *m* recoded variables from the metadata, then select up to *j* bin levels for each of the *m* recoded variables. Universe formation on the MAS is performed using an implicit table server. For example, suppose a data user defines their universe as:

$$[\text{gender} = \text{female and } \$45{,}000 < \text{income} \le \$53{,}000] \qquad (1)$$

OR

$$[\text{gender} = \text{male and } \$28{,}000 < \text{income} \le \$45{,}000] \qquad (2)$$

This universe is represented as a two-way table of counts for *sex* by *income*, as shown in Table 1. Piece (1) is represented by the outlined cell in Table 1, while piece (2) is represented by the set of shaded cells. Note that there $n_{24} + n_{12} + n_{13}$ total observations in this universe. For convenience, we will use the notation U(n) to indicate a universe that contains n total observations. For example, the universe defined from pieces (1) and (2) above will be referred to as U($n_{24} + n_{12} + n_{13}$).

Table 1. Table representation of the universe defined from (1) and (2)

	income				
gender	$0 to $28,000	$28,000 to $39,000	$39,000 to $45,000	$45,000 to $53,000	Total
male	n_{11}	n_{12}	n_{13}	n_{14}	$n_{1\cdot}$
female	n_{21}	n_{22}	n_{23}	n_{24}	$n_{2\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	$n_{\cdot 4}$	$n_{\cdot\cdot}$

All universes formed on the MAS must past both of the following two confidentiality rules. If at least one of these two rules fail, the MAS will reject the universe query and prompt the universe to modify his or her selections. Note that these rules are tested prior to performing the user's selected statistical analysis on their defined universe.

The first rule is the *No Marginal 1 or 2 Rule*. No universe defined with exactly *m* variables on the MAS may be defined from an *m*-way table that contains at least one *m*-1 dimensional marginal total equal to 1 or 2. For example, to check the No

Marginal 1 or 2 Rule for U(n_{24} + n_{12} + n_{13}) defined above, the following equations must be satisfied:

$$n_{i\cdot} \neq 1 \text{ or } 2, \text{ for i} = 1,2 \quad \text{and} \quad n_{\cdot j} \neq 1 \text{ or } 2, \text{ for j} = 1,\ldots,4$$

The second rule is the Minimum Number of Observations Rule. In general, a universe must contain at least Γ observations before a user can perform a statistical analysis on this universe. The value of Γ is not given here since it is Census Confidential. Cutpoint bins are always combined to check this rule. In addition, the way this rule is checked is dependent on whether or not the universe pieces are disjoint or joint. A universe is classified as *disjoint* if its individual pieces do not share cell counts in common. For example, pieces (1) and (2) for the universe U(n_{24} + n_{12} + n_{13}) are disjoint since they do not share any cell counts in common. Since U(n_{24} + n_{12} + n_{13}) is a disjoint universe, the MAS would check that both piece (1) and piece (2) contain at least Γ observations. That is, both of the following equations must be satisfied. Note that the cutpoint bins of *income* are combined within piece (2) prior to performing the test.

$$n_{24} \geq \Gamma \quad \text{and} \quad (n_{12} + n_{13}) \geq \Gamma$$

A universe is classified as *joint* if as least one of its individual pieces shares cell counts in common with at least one other piece. For example, suppose the user defines the following universe, U($n_{2\cdot}$ + $n_{\cdot 3}$ + $n_{\cdot 4}$) = (3) OR (4), where pieces (3) and (4) are defined as:

$$[\text{gender} = \text{female}] \tag{3}$$

$$[\$39{,}000 < \text{income} \leq \$53{,}000] \tag{4}$$

U($n_{2\cdot}$ + $n_{\cdot 3}$ + $n_{\cdot 4}$) is derived from the set of outlined and shaded cells in Table 2, where the outlined cells represent piece (3) and the shaded cells represent piece (4). Note that the cell counts n_{23} and n_{24} are shared among pieces (3) and (4).

Table 2. Table representation of the universe defined from (3) and (4)

	income				
gender	\$0 to \$28,000	\$28,000 to \$39,000	\$39,000 to \$45,000	\$45,000 to \$53,000	Total
male	n_{11}	n_{12}	n_{13}	n_{14}	$n_{1\cdot}$
female	n_{21}	n_{22}	n_{23}	n_{24}	$n_{2\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	$n_{\cdot 4}$	$n_{\cdot\cdot}$

Since U($n_{2\cdot}$ + $n_{\cdot 3}$ + $n_{\cdot 4}$) is a joint universe, to test the *Minimum Number of Observations Rule*, the MAS would first check that both pieces (3) and (4) contain at least Γ observations each, then check that the non-empty intersection I = (3) ∩ (4)

contains at least Γ* observation, where $\Gamma^* < \Gamma$. That is, the following three equations must be satistified:

$$n_{2\cdot} \geq \Gamma \text{ and } (n_{\cdot 3} + n_{\cdot 4}) \geq \Gamma \text{ and } (n_{23} + n_{24}) \geq \Gamma^*$$

If at least one piece (3) or (4), or if the intersection *I*, fails to pass the tests above, then the MAS will reject the entire universe. Once again, the cutpoint bins of income are first combined within piece (4) and within $I = (3) \cap (4)$ prior to the testing of the *Minimum Number of Observations* Rule for (4) or *I*.

While the above rules ensures that a universe U(*n*) meets a minimum size requirement, it does not protect against differencing attack disclosures. A *differencing attack disclosure* occurs when a data intruder attempts to rebuild a confidential microdata record by subtracting the statistical analysis results obtained through two separate and similar queries. For example, suppose a data intruder first creates two universes on the MAS: U(*n*) and U(*n*-1), where U(*n*-1) contains the exact same *n* observations as U(*n*), less one unique observation. The difference U(*n*) – U(*n*-1) = U(1), where U(1) is a manipulated universe that contains the one unique observation. Suppose further that the data intruder then requests two separate cross-tabulations for *sex* by *employment status*, T[U(*n*)] and T[U(*n*-1)], fitted on U(*n*) and U(*n*-1), respectively, as shown in Figure 1. Since U(*n*) and U(*n*-1) only differ by one unique observation, T[U(*n*-1)] will be exactly the same as T[U(*n*)], less one unique cell count.

The matrix subtraction T[U(*n*)] – T[U(*n*-1)] = T[U(1)], where T[U(1)] is a two-way table of *sex* by *employment status* built upon the one unique observation contained in U(1) = U(*n*) – U(*n*-1). As shown in Figure 1, T[U(1)] contains a cell count in the male non-employed cell with zeros in the remaining cell, which tells the data intruder that the one unique observation contained in U(1) is a non-employed male. By performing similar differencing attacks like the shown above, a data intruder can successfully rebuild the confidential microdata record for the one unique observation contained in U(1).

T[U(*n*)]	*employment status*	
sex	employed	non-employed
Male	n_1	n_2
female	n_3	n_4

-

T[U(*n*-1)]	*employment status*	
sex	employed	non-employed
male	n_1	n_2-1
female	n_3	n_4

=

T[U(1)]	*employment status*	
sex	employed	non-employed
male	0	1
female	0	0

Fig. 1. An Example of a Differencing Attack Disclosure

To help protect against differencing attack disclosures, the MAS implements a universe subsampling routine called the *Drop Q Rule*. Once a universe data set passes the universe formation rules, the MAS will first draw a random value of $Q_v = q_v \in \{2,\ldots,k\}$ from a Discrete Uniform distribution with probability mass function

P($Q_v = q_v$) = 1/(k-1). Then, given $Q_v = q_v$, the MAS will subsample the universe data set U(n) by removing q_v records at random from U(n) to yield a new subsampled universe data set, U(n-q_v).

On the MAS, all statistical analyses are performed on the subsampled U(n-q_v) data set and not on the original U(n) data set. Each unique universe U(n) that is defined on the MAS will be subsampled independently according to the *Drop Q Rule*. In addition, to prevent an "averaging of results" attack, the MAS will produce only one subsampled U(n-q_v) data set for each unique U(n) data set, and will fix the subsampled U(n-q_v) data set to each unique U(n) data set for the lifetime of the system. That is, if the same user, or a different user, selects the same unique U(n) data set as before, then the MAS would use the exact same subsampled U(n-q_v) data set as before for the statistical analysis.

Therefore, if the data intruder attempts the differencing attack T[U(n)] – T[U(n-1)] = T[U(1)] as shown in Figure 1, he would actually be performing the differencing attack T[U(n-q_1)] – T[U(n-1-q_2)] = T[U(1)] as shown in Figure 2, where T[U(n-q_1)] and T[U(n-1-q_2)] are two-way tables of *sex* by *employment status* based on the two independently subsampled universes, U(n-q_1) and U(n-1-q_2), where the random vectors $\mathbf{X} = \mathbf{x} = (x_1,\ldots,x_4)$ and $\mathbf{Y} = \mathbf{y} = (y_1,\ldots,y_4)$ give the number of counts that were specifically removed from each cell in T[U(n-q_1)] and T[U(n-1-q_2)], respectively, and $\Sigma_j x_j = q_1$,$\Sigma_j y_j = q_2$, $0 \le x_j \le q_1$, and $0 \le y_j \le q_2$, for j = 1,...,4.

Since each $Q_1 = q_1$ and $Q_2 = q_2$ are independently drawn from a Discrete Uniform distribution, q_1 will not necessarily be equal to q_2, and the resulting table T[U(1)] = T[U(n-q_1)] – T[U(n-1-q_2)] may or may not yield a successful disclosure of *sex* = male and *employment status* = non-employed for the one unique observation contained in U(1). A brief overview of the effectiveness of the *Drop Q Rule* against differencing attack disclosures will be discussed in Section 4.

T[U(n-q_1)]	employment status	
Sex	employed	non-employed
Male	n_1-x$_1$	n_2-x$_2$
Female	n_3-x$_3$	n_4-x$_4$

-

T[U(n-1-q_2)]	*employment status*	
sex	employed	non-employed
male	n_1-y$_1$	n_2-1-y$_2$
female	n_3-y$_3$	n_4-y$_4$

$\overset{?}{=}$

T[U(1)]	*employment status*	
sex	employed	non-employed
male	0	1
female	0	0

Fig. 2.

3.2 Confidentiality Rules for Regression Models

The MAS also implements a series of confidentiality rules for regression models. For example, users may only select up to 20 independent variables for any single regression equation. Users are allowed to transform numerical variables only, and they must select their transformations from a pre-approved list. This prevents the user from using transformations that deliberately over emphasize outliers.

Any fully interacted regression model that contains only dummy variables as predictors can pose a potential disclosure risk [9], [10]. Therefore, users are allowed to include only two-way and three-way interaction terms within any specified regression model. No regression model that contains more than three variables can be fully interacted. Predictor dummy variables must each contain a minimum of α observations each. Any dummy variable that fails this requirement gets dropped from the regression model equation, along with the dummy variable that represents the reference category level for that particular categorical predictor variable. That is, dummy variables (or equivalently, category levels) that contain less than α observations are combined with the reference category level to prevent data intruders from fitting regression models with categorical predictors that contain sparse categories. The value of α is not given here since it is Census confidential.

Prior to passing back any regression output back to the user, the MAS checks and ensures that R^2 is not too close to 1. If R^2 is too close to 1, then the MAS will withhold from outputting any regression analysis results back to the user. If R^2 is not too close to 1, then the MAS will pass the estimated regression coefficients and the Analysis of Variance (or Deviance) table to the user without restrictions.

Actual residual values can pose a potential disclosure risk, since a data intruder can obtain the actual real values of the dependent variable by simply adding the residual to the fitted values obtained from the regression model. Therefore, the MAS never passes back real residual values back to the user. To help data users assess the fit of their Ordinary Least Squares regression models, all diagnostic plots on the MAS are based on synthetic residuals and synthetic real values. These plots are designed to mimic the actual patterns seen in the scatter plots of the real residuals vs. the real fitted values [11].

4 Evaluation of the Effectiveness of the Drop Q Rule

We will only present a brief overview of this evaluation here. Full details about this evaluation can be found in [12]. Given a pair of similar universes that only differ by one unique observation, U(n) and U(n-1), we investigated the effectiveness of the *Drop Q Rule* in preventing contingency table differencing attack disclosures of the form T[U(1)] = T[U(n-q_1)] – T[U(n-1-q_2)], as was shown in Figure 2.

Using the same example as was shown in Section 3.1, since U(n-q_1) and U(n-1-q_2) are two independently subsampled universes, the resulting table T[U(1)] = T[U(n-q_1)] – T[U(n-1-q_2)] (in Figure 2) may or may not a count of 1 in the shaded cell that represents the *sex* and *employment status* categories of the one unique observation contained in U(1), with zeros within the remaining three cells.

We wanted to find the probability of obtaining such a table T[U(1)] that contains a 1 in the shaded cell for s*ex* = male and *employment status* = non-employed, with zeros within the remaining three cells, from the differencing attack T[U(n-q_1)] – T[U(n-1-q_2)] = T[U(1)]. That is, we wanted to find the probability that the resulting table T[U(1)] = T[U(n-q_1)] – T[U(n-1-q_2)] yielded a successful disclosure of *sex* and *employment status* for the one unique observation contained in U(1).

For a given value $Q_1 = q_1$ the observed vector ($X_1 = x_1,\ldots,X_4 = x_4$) of counts that are actually removed from each cell in T[U(n-q_1)] are dependent on the distribution of cell proportions $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_4)$ within the original two-way table, T[U(n)]. Similarly, for a given value $Q_2 = q_2$ the observed vector ($Y_1 = y_1,\ldots,Y_4 = y_4$) of counts that are actually removed from each cell in T[U(n-1-q_2)] are dependent on the distribution of cell proportions $\boldsymbol{\psi} = (\psi_1,\ldots,\psi_4)$ within the original two-way table T[U(n-1)]. However, if n is large, $\boldsymbol{\pi} \approx \boldsymbol{\psi}$. Therefore, for large values of n, $\mathbf{X} \mid \boldsymbol{\pi}, Q = q_1 \overset{\cdot}{\sim}$ Multinomial($\boldsymbol{\pi}$, q_1) and $\mathbf{Y} \mid \boldsymbol{\pi}, Q = q_2 \overset{\cdot}{\sim}$ Multinomial($\boldsymbol{\pi}$, q_2). Since U(n-q_1) and U(n-1-q_2) are subsampled independently, the random tables T[U(n-q_1)] and T[U(n-1-q_2)] are also subsampled independently and the random vectors **X** and **Y** are independent. Therefore, the approximate joint probability of $\mathbf{X}\mid \boldsymbol{\pi}, Q_1 = q_1$ and $\mathbf{Y}\mid\boldsymbol{\pi}, Q = q_2$ is:

$$\begin{aligned}
&P(\mathbf{X} = \mathbf{x} \mid \boldsymbol{\pi}, Q = q_1 \cap \mathbf{Y} = \mathbf{y} \mid \boldsymbol{\pi}, Q = q_2) = \\
&P(\mathbf{X} = \mathbf{x} \mid \boldsymbol{\pi}, Q = q_1)\, P(\mathbf{Y} = \mathbf{y} \mid \boldsymbol{\pi}, Q = q_2) = \\
&\left(\frac{q_1!}{x_1!\cdots x_4!}\right)\left(\frac{q_2!}{y_1!\cdots y_4!}\right)\pi_1^{x_1+y_1}\cdots\pi_4^{x_4+y_4}
\end{aligned} \tag{5}$$

However, $P(Q_v = q_v) = 1/(k\text{-}1)$, therefore

$$\begin{aligned}
&P([\mathbf{X} = \mathbf{x} \cap Q_1 = q_1 \mid \boldsymbol{\pi}\,] \cap [\mathbf{Y} = \mathbf{y} \cap Q_2 = q_2 \mid \boldsymbol{\pi}]) = \\
&P(\mathbf{X} = \mathbf{x} \cap \boldsymbol{\pi}, Q_1 = q_1)\, P(\mathbf{Y} = \mathbf{y} \cap \boldsymbol{\pi}, Q_2 = q_2) = \\
&P(\mathbf{X} = \mathbf{x} \mid \boldsymbol{\pi}, Q_1 = q_1)\, P(Q = q_1) P(\mathbf{Y} = \mathbf{y} \mid \boldsymbol{\pi}, Q_2 = q_2)\, P(Q_2 = q_2) = \\
&\left(\frac{1}{k-1}\right)^2\left(\frac{q_1!}{x_1!\cdots x_4!}\right)\left(\frac{q_2!}{y_1!\cdots y_4!}\right)\pi_1^{x_1+y_1}\cdots\pi_4^{x_4+y_4}
\end{aligned} \tag{6}$$

Equation (6) gives us the approximate joint probability of observing a T[U(n-q_1)] with exactly ($x_1,\ldots,x_4$) counts removed from each cell, given $Q = q_1$, and observing a T[U(n-1-q_2)] with exactly ($y_1,\ldots,y_4$) counts removed from each cell, given $Q = q_2$, where $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_4)$ are the observed cell proportions of counts within the original two-way table of *sex* by *employment status*, T[U(n)], and $\Sigma_j\, \pi_j = 1$.

T[U(n-q_1)]	employment status	
sex	employed	non-employed
male	n_1-x_1	n_2-x_2
female	n_3-x_3	n_4-x_4

-

T[U(n-1-q_1)]	*employment status*	
sex	employed	non-employed
male	n_1-x_1	n_2-1-x_2
female	n_3-x_3	n_4-x_4

=

T[U(1)]	*employment status*	
sex	employed	non-employed
male	0	1
female	0	0

Fig. 3.

We found that the resulting table, T[U(1)], would yield a successful disclosure if and only if $q_1 = q_2$ and if and only if ($x_1,\ldots,x_4$) = ($y_1,\ldots,y_4$), as shown in Figure 3.Therefore, equation (9) becomes:

$$\left(\frac{1}{k-1}\right)^2\left(\frac{q_1!}{x_1!\cdots x_4!}\right)^2\pi_1^{2x_1}\cdots\pi_4^{2x_4} \tag{7}$$

Equation (7) gives us the approximate joint probability of obtaining any one pair of subsampled two-way tables, T[U(n-q_1)] and T[U(n-1-q_1)] (as shown in Figure 3), such that the same value $Q_1 = q_1$ was drawn for both subsampled universes U(n-q_1) and U(n-1-q_1), and the exact same observed vector of counts ($x_1,\ldots,x_4$) were removed at random among the four cells of both T[U(n-q_1)] and T[U(n-1-q_1)], where $\Sigma_j\ x_j = q_1$. For any given value of $Q = q_1$, there are exactly C(4+q_1-1, q_1) sequences of vectors ($x_1,\ldots,x_4$), such that $\Sigma_j\ x_j = q_1$. Therefore, if we sum (7) over all possible C(4+q_1-1, q_1) sequences of ($x_1,\ldots,x_4$), we obtain:

$$\sum_{x_1,\ldots,x_4 \geq 0}^{x_1+\cdots+x_4=q_1} \left(\tfrac{1}{k-1}\right)^2 \left(\tfrac{q_1!}{x_1!\cdots x_4!}\right)^2 \pi_1^{2x_1}\cdots\pi_4^{2x_4} \tag{8}$$

(8) gives us the approximate joint probability of obtaining all possible pairs of subsampled tables, T[U(n-q_1)] and T[U(n-1-q_1)] (from Figure 3), for a single given observed value of $Q_1= q_1$. However, since $Q_1 = q_1 \in \{2,\ldots,k\}$, if we sum (8) over all possible observed values q_1 can take, we obtain (9), the total approximate joint probability of observing all pairs of subsampled two-way tables, T[U(n-q_1)] and T[U(n-1-q_1)], , for all possible values of Q = $q_1 \in \{2,\ldots,k\}$. As a result, (9) gives us the approximate total probability of obtaining a successful disclosure of gender and employment status, for the one observation contained in U(1) = U(n) – U(n-1), from the differencing attack T[U(n-q_1)] – T[U(n-1-q_1)] = T[U(1)]. Proposition 1 summarizes these results to differencing attacks performed on m-way tables.

$$\sum_{q_1=2}^{k} \sum_{x_1,\ldots,x_4 \geq 0}^{x_1+\cdots+x_4=q_1} \left(\tfrac{1}{k-1}\right)^2 \left(\tfrac{q_1!}{x_1!\cdots x_4!}\right)^2 \pi_1^{2x_1}\cdots\pi_4^{2x_4} \tag{9}$$

Proposition 1: Suppose T[U(n-q_1)] and T[U(n-1-q_2)] are any two pairs of similar m-way tables that both contain J total cells each, fitted on two independently subsampled universe data sets, U(n-q_1) and U(n-1-q_2), where both U(n-q_1) and U(n-1-q_2) were subsampled according to the *Drop Q Rule*. Let ($x_1,\ldots,x_J$) and ($y_1,\ldots,y_J$) be the observed vector of counts that were randomly removed from each cell in T[U(n-q_1)] and T[U(n-1-q_2)], where $\Sigma_j\ x_j = q_1$, $\Sigma_j\ y_j = q_2$, $0 \leq x_j \leq q_1$, and $0 \leq y_j \leq q_2$. Then the differencing attack of T[U(n-q_1)] – T[U(n-1-q_2)] will yield an m-way table, T[U(1)], that will successfully discloses all observed category levels for all m variables for the one unique observation contained in U(1) if and only if $q_1 = q_2$, and if and only if ($x_1,\ldots,x_J$) = ($y_1,\ldots,y_J$) in T[U(n-q_1)] and T[U(n-1-q_1)], respectively. Let $\boldsymbol{\pi}$ = ($\pi_1,\ldots,\pi_J$) be the observed cell proportions of the original m-way table T[U(n)], fitted on the full universe U(n). Then, if n is large, the approximate probability of obtaining a successful disclosure from T[U(1)] = T[U(n-q_1)] - T[U(n-1-q_2)] is:

$$\xi_{J,k}(\pi_1,\ldots,\pi_J) = \sum_{q_1=2}^{k} \sum_{x_1,\ldots,x_J \geq 0}^{x_1+\cdots+x_J=q_1} \left(\tfrac{1}{k-1}\right)^2 \left(\tfrac{q_1!}{x_1!\cdots x_J!}\right)^2 \pi_1^{2x_1}\cdots\pi_J^{2x_J} \tag{10}$$

where $\Sigma_j\ \pi_j = 1$. Note that if at least one $\pi_j = 0$, then we define $0^{2x_j} = 1$ ■

Theorem 1: The approximate probability function (10) is a Schur-convex function of $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_J)$, where $\Sigma_j\, \pi_j = 1$ ■

The proof of Theorem 1 relies on the concepts of majorization and Schur-convexity ([11],[12] and [13]) and can be found in the appendix. Using a series of simulated differencing attacks on one-way, two-way, and three-way tables, we found that, on the average, (10) approximates the probability of obtaining a successful disclosure from the resulting table T[U(1)] = T[U(n-q_1)] - T[U(n-1-q_2)], within two decimal places.

Theorem 2: The approximate probability function $\xi_{J,k}(\boldsymbol{\pi})$ (13), subject to the linear constraint $\Sigma_j\, \pi_j = 1$, achieves a minimum value when $\pi_1 = \cdots = \pi_J = 1/J$ and achieves a maximum when at one $\pi_j = 1$ with the remaining $\pi_i = 0$, for $i \neq j$. Futhermore, (10) satisfies the following bounds:

$$\left(\tfrac{1}{k-1}\right)^2 \sum_{q_1=2}^{k} \sum_{x_1,\ldots,x_J \geq 0}^{x_1+\cdots+x_J=q_1} \left(\tfrac{q_1!}{x_1!\cdots x_J!}\right)^2 \left(\tfrac{1}{J}\right)^{2q_1} \leq \xi_{J,k}(\pi_1,\ldots,\pi_J) \leq \tfrac{1}{k-1}. \quad ■ \quad (11)$$

The proof of Theorem 2 relies on the fact that (11) is a Schur-convex function, can be found in the appendix.

5 Future Work

The MAS will continue to be developed within Data FERRET. We will soon be testing the software itself and the confidentiality rules within the MAS beta prototype to ensure they properly uphold disclosure avoidance standards. We will draft up a set of confidentiality rules for cross-tabulations, and add different types of statistical analyses within the system. We will explore other types of differencing attack disclosures, and explore ways to prevent such differencing attacks.

References

[1] Duncan, G.T., Keller-McNulty, S., Stokes, S.L.: Disclosure Risk vs. Data Utility: The R-U Confidentiality Map. Technical Report 2003-6, Heinz School of Public Policy and Management, Carnegie Mellon University (2003)

[2] Kaufman, S., Seastrom, M., Roey, S.: Do Disclosure Controls to Protect Confidentiality Degrade the Quality of the Data? In: Proceedings of the Section on Survey Research, American Statistical Association (2005)

[3] Weinberg, D., Abowd, J., Rowland, S., Steel, P., Zayatz, L.: Access Methods for United States Microdata. In: Proceedings of the Workshop on Data Access to Microdata, Nurembourg, Germany, August 20-21 (2007); Also found on the Social Science Research Network `http://hq.ssrn.com` and US Census Bureau Center for Economic Studies Paper No. CES-WP-07-25

[4] Rowland, S., Zayatz, L.: Automating Access with Confidentiality Protection: The American FactFinder. In: Proceedings of the Section on Government Statistics, American Statistical Association (2001)

[5] Chaudhry, M.: Overview of the Microdata Analysis System. Statistical Research Division Internal Report, Census Bureau (2007)

[6] Lucero, J.: Confidentiality Rules for Universe Formation and Geographies for the Microdata Analysis System. Statistical Research Division Confidential Research Report Series No. CCRR-2009/01, Census Bureau (2009)

[7] Lucero, J.: Confidentiality Rule Specifications for Performing Regression Analysis on the Microdata Analysis System. Statistical Research Division Confidential Research Report Series #????, U.S. Census Bureau (2010) (in Progress)

[8] Lucero, J., Zayatz, L., Singh, L.: The Current State of the Microdata Analysis System at the Census Bureau. In: Proceedings of the American Statistical Association, Government Statistics Section (2009)

[9] Reznek, A.P.: Disclosure Risks in Cross Section Regression Models. In: Proceedings of the American Statistical Association, Government Statistics Section (CD-ROM). American Statistical Association, Alexandria (2003)

[10] Reznek, A.P., Riggs, T.L.: Disclosure Risks in Regression Models: Some Further Results. In: Proceedings of the American Statistical Association, Government Statistics Section (CD-ROM). American Statistical Association, Alexandria (2004)

[11] Reiter, J.P.: Model Diagnostics for Remote-Access Regression Servers. Statistics and Computing 13, 371–380 (2003)

[12] Lucero, J.: Evaluation of the Effectiveness of the Drop q Rule Against Differencing Attack Disclosures. Statistical Research Division Confidential Research Report Series No.CCRR-2010/03, U.S. Census Bureau (2010)

[13] Marshall, A.W., Olkin, I.: Inequalities: Theory of Majorization and Its Applications. Mathematics in Science and Engineering Series, vol. 143. Academic Press, London (1979)

[14] Zhang, X.: Schur-Convex Functions and Isoperimetric Inequalities. Proceedings of the American Mathematical Society 126(2), 461–470 (1998)

[15] Ben-Haim, Z., Dvorkind, T.: Majorization and Applications to Optimization (2004), unknown publication found from the web at
`http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.4.9813&rep=rep1&type=pdf`

Appendix A

We will use concepts from the theory of majorization and Schur-convexity to prove Theorems 1 and 2. Majorization and Schur-Convexity are useful tools that can be sometimes used to prove certain properties of functions, such as the solution to an optimization problem. The following definitions and theorems were taken from [13], [14] and [15].

Definition 1: For a given vector $\mathbf{z} = (z_1,\ldots,z_J)$, let $z_{(1)}$ denote the smallest element of **z**, let $z_{(2)}$ denote the second smallest element of z, and so on. A vector z is said to majorize a vector y (denoted $z \succ y$) if

$$\sum_{i=1}^{I} z_{(i)} \geq \sum_{i=1}^{I} y_{(i)}, \text{ for I = 1,..., J – 1, and } \sum_{j=1}^{J} z_j = \sum_{j=1}^{J} y_j .$$

Lemma 1: Let **z** = ($z_1,\ldots,z_J$) be any arbitrary vector with $\Sigma_j\, z_j = s$ where $z_j \geq 0$ for j = 1,…,J. Define the uniform vector **u** = $\left(\frac{s}{n},\ldots,\frac{s}{n}\right)$. Then for any given vector **z**, **z** $\succ$ **u**.

Majorization is a partial ordering among any two vectors of equal dimensions, and applies only to vectors having the same sum. It is a measure of the degree to which the vector elements differ. Intuitively, the uniform vector is the vector with the minimal difference between its elements. Therefore, all other vectors z, whose elements add up to the same sum s, will majorize u.

Definition 2: For $J > 1$, a function g: $R^J \longrightarrow R$ is called a symmetric function if for every permutation matrix Π, g(Π($z_1,\ldots,z_J$)) = g($z_1,\ldots,z_J$).

Definition 3: For $J > 1$, a function g(**z**): $R^J \longrightarrow R$ is called a Schur-convex function if **z** $\succ$ **y** implies g(**z**) $\succ$ g(**y**).

Schur-convex functions are functions that preserve the ordering of majorization. That is, a Schur-convex function translates the ordering of vectors to a standard scaler ordering. Symmetry is a necessary condition for a function to be Schur-convex. In addition, any function that is both symmetric and convex is a Schur-convex function. We will use the following useful lemmas to prove Theorem 1.

Lemma 2: The complete homogenous symmetric function

$$H(\pi_1,\ldots,\pi_J) = \sum_{x_1,\ldots,x_J}^{x_1+\cdots x_J=q} \pi_1^{x_1}\cdots\pi_J^{x_J} \tag{12}$$

Is a Schur-convex function of $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_J)$.

Lemma 3: Let φ(**z**): $R^J \longrightarrow R$ be a Schur-convex function of **z** = ($z_1,\ldots,z_J$) and let g(z_j): R ⟶ R be a convex function of z_j, for j = 1,…,J. Then the composition Φ(**z**) = φ(g(z_1),…,g(z_J)), Φ(**z**): $R^J \longrightarrow R$ is a Schur-Convex function of **z**.

Lemma 4: Let the function $\psi(z_1,\ldots,z_J)$ be an increasing, real valued function defined on R^J, and let $\varphi_1,\ldots,\varphi_k$ be real valued Schur-Convex functions, each with common domain $A \subset R^J$. Then the composition: $\Phi(z_1,\ldots,z_J) = \psi(\varphi_1(z_1,\ldots,z_J),\ldots, \varphi_k(z_1,\ldots,z_J))$ is a Schur-convex function of **z** = ($z_1,\ldots,z_J$).

Lemma 5: Let c > 0 be any constant, and let $z_k \geq 0$ for all k = 1,…,K. Then the function $\psi(z_1,\ldots,z_K) = c(\Sigma_k\, z_k)$ is an increasing function on R^K.

The following Lemma will be useful to prove Theorem 1:

Lemma 6: The function

$$\mathrm{F}_{J,q}(\pi_1,\ldots,\pi_J) = \sum_{x_1,\ldots,x_J}^{x_1+\cdots x_J=q} \left(\tfrac{q!}{x_1!,\ldots,x_J!}\right)^2 \pi_1^{2x_1}\cdots\pi_J^{2x_J} \tag{13}$$

is a Schur-convex function of $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_J)$, where $\Sigma_j\, \pi_j = 1$, $0 \leq \pi_j \leq 1$ for all j, and the summation in (13) is taken over all possible sequences of $\Sigma_j\, x_j = q$, for any given value of q. (Note: if one $\pi_j = 0$, then we define $0^{2x_j} = 1$)

Proof: (13) is a symmetric function on the set R^J. To apply Lemma 3 to (13), define $g(\pi_j) = a_j\pi_j^2$, where $0 \leq \pi_j \leq 1$ for all j, and

$$a_j = \begin{cases} \left(\dfrac{q!^{\frac{1}{J}}}{x_j!}\right)^{\frac{2}{x_j}} & if\ x_j \geq 1 \\ 1 & if\ x_j = 0 \end{cases}$$

and set $\varphi(\pi_1,\ldots,\pi_J) = H(\pi_1,\ldots,\pi_J)$ (12) from Lemma 2. Since (12) is a Schur-convex function of $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_J)$, and $g(\pi_j) = a_j\pi_j^2$ is convex for each π_j. Then, by Lemma 3 the compostition

$$\Phi(\boldsymbol{\pi}) = H(g(\pi_1),\ldots,g(\pi_J)) = \sum_{x_1,\ldots,x_J\geq 0}^{x_1+\cdots+x_J=q} \left[\left(\frac{q!^{\frac{1}{J}}}{x_1!}\right)^{\frac{2}{x_1}} \pi_1^2\right]^{x_1} \cdots \left[\left(\frac{q!^{\frac{1}{J}}}{x_J!}\right)^{\frac{2}{x_J}} \pi_J^2\right]^{x_J}$$

$$= \sum_{x_1,\ldots,x_J\geq 0}^{x_1+\cdots+x_J=q} \left(\frac{q!^{\frac{1}{J}}}{x_1!}\right)^2 \pi_1^{2x_1} \cdots \left(\frac{q!^{\frac{1}{J}}}{x_J!}\right)^2 \pi_J^{2x_J} = \sum_{x_1,\ldots,x_J\geq 0}^{x_1+\cdots+x_J=q} \left(\frac{q!^{\frac{1}{J}}\cdots q!^{\frac{1}{J}}}{x_1!\cdots x_J!}\right)^2 \pi_1^{2x_1}\cdots\pi_J^{2x_J}$$

$$= \sum_{x_1,\ldots,x_J\geq 0}^{x_1+\cdots+x_J=q} \left(\tfrac{q!}{x_1!\cdots x_J!}\right)^2 \pi_1^{2x_1}\cdots\pi_J^{2x_J}$$

is a Schur-convex function of $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_J)$. ■

Proof of Theorem 1: (10) is a symmetric function of $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_J)$ on the set R^J. To apply Lemma (4) to (10), set

$$\psi(z_2,\ldots,z_k) = \left(\tfrac{1}{k-1}\right)^2 \sum_{q1=2}^{k} z_{q_1} \text{ for } z_2,\ldots,z_k \geq 0 \tag{14}$$

$$\text{and set } \varphi_{q_1} = \sum_{x_1,\dots,x_J}^{x_1+\cdots x_J=q_1} \left(\tfrac{q_1!}{x_1!,\dots,x_J!}\right)^2 \pi_1^{2x_1}\cdots\pi_J^{2x_J} \tag{15}$$

By Lemma 5, (14) is an increasing function of $z_2,\dots,z_k$ on the set R^{k-1}, and (15) is a Schur-convex function by Lemma 6, for a given value of $Q_1 = q_1$. Therefore, by Lemma (4), the composition:

$$\xi_{J,k}(\pi_1,\dots,\pi_J) = \psi(\varphi_2(\pi_1,\dots,\pi_J),\dots,\varphi_k(\pi_1,\dots,\pi_J)) = \left(\tfrac{1}{k-1}\right)^2 \sum_{q_1=2}^{k} \varphi_{q_1}(\pi_1,\dots,\pi_J)$$

$$= \left(\tfrac{1}{k-1}\right)^2 \sum_{q_1=2}^{k} \sum_{x_1,\dots,x_J\ge 0}^{x_1+\cdots+x_J=q_1} \left(\tfrac{q_1!}{x_1!\cdots x_J!}\right)^2 \pi_1^{2x_1}\cdots\pi_J^{2x_J}$$

is a Schur-convex function of $\boldsymbol{\pi} = (\pi_1,\dots,\pi_J)$. ■

Lemma 7 will be used to help prove Theorem 2.

Lemma 7: Let $\varphi(z_1,\dots,z_n)$ be a Schur-convex function. Suppose we wish to find the minimum of $\varphi(z_1,\dots,z_n)$ given the linear constraint $\Sigma_j\, z_j = s$. Then, since $\varphi(z_1,\dots,z_n)$ is Schur-convex, the minimum of φ is achieved when $\mathbf{z} = \mathbf{u} = \left(\tfrac{s}{n},\dots,\tfrac{s}{n}\right)$.

Proof: By Lemma 1, the uniform vector $\mathbf{u}$ is majorized by any other vector $\mathbf{z}$ that has the same sum s. Since $\varphi(\mathbf{z})$ is a Schur-convex function, by Definition 3, $\mathbf{z} \succ \mathbf{u}$ implies $\varphi(\mathbf{z}) \ge \varphi(\mathbf{u})$. Therefore, the minimum of $\varphi(z_1,\dots,z_n)$, subject to the constraint $\Sigma_j\, z_j = s$, is achieved at $\mathbf{z} = \mathbf{u}$. ■

Proof of Theorem 2: Since (10) is a Schur-convex function, by Lemma 7, the minimum of (10) subject to the constraint $\Sigma_j\, \pi_j = 1$ is achieved when $\boldsymbol{\pi} = \mathbf{u} = \left(\tfrac{1}{J},\dots,\tfrac{1}{J}\right)$, and the minimum of (10) is:

$$\left(\tfrac{1}{k-1}\right)^2 \sum_{q_1=2}^{k} \sum_{x_1,\dots,x_J\ge 0}^{x_1+\cdots+x_J=q_1} \left(\tfrac{q_1!}{x_1!\cdots x_J!}\right)^2 \left(\tfrac{1}{J}\right)^{2q_1} \tag{16}$$

In addition, since (10) is a Schur-convex function, it is symmetric. Therefore, for $j = 1,\dots,J$, for any given permutation matrix Π, $\xi_{J,k}(\Pi\,(\pi_1,\dots,\pi_\mathrm{J})) = \xi_{J,k}(\pi_1,\dots,\pi_\mathrm{J})$. Furthermore, for any given vector $\boldsymbol{\pi} = (\pi_1,\dots,\pi_\mathrm{J})$ whose elements sum to 1 and $0 \le \pi_j \le 1$ for all J, it is easy to check that $(1,0,\dots,0) \succ \boldsymbol{\pi} \succ \mathbf{u}$, which implies

$$\tfrac{1}{k-1} = \xi_{J,k}(\pi_1,\dots,\pi_\mathrm{J}) \succ \xi_{J,k}(\boldsymbol{\pi}) \succ \xi_{J,k}(\mathbf{u}). \tag{17}$$

Combining (16) and (17) gives (11). ■