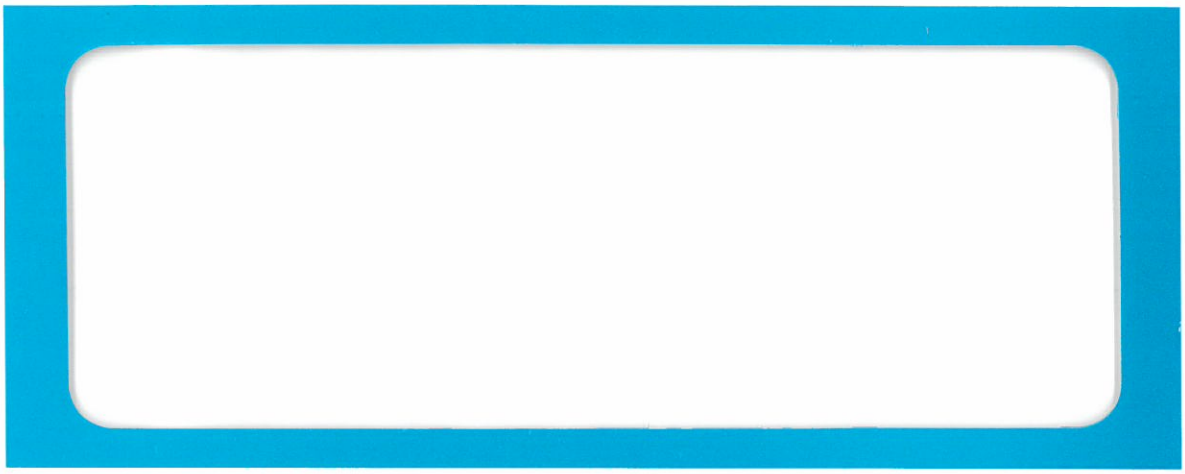
Methodology Branch

Direction de la méthodologie



Household Survey Methods Division

Division des méthodes d'enquêtes auprès des ménages

Statistics Canada Statistique Canada

Canada

WORKING PAPER
METHODOLOGY BRANCH

Statistical Disclosure Control For Public Use Microdata Files – A Guide

HSMD - 2016 - 002E

Jean-Louis Tambay

Household Survey Methods Division
Statistics Canada

May 2016

STATISTICAL DISCLOSURE CONTROL FOR PUBLIC USE MICRODATA FILES – A GUIDE

JEAN-LOUIS TAMBAY

ABSTRACT

Statistics Canada has been producing Public Use Microdata Files (PUMFs) from survey master files since the 1980s. Two key features in the production of PUMFs are the assessment of risks of disclosure at the unit and datafile levels and the implementation of measures to bring this risk to an acceptable level. This brief guide aims to provide the reader with a meaningful appreciation of the risks and issues surrounding PUMF data and to serve as a starting point for further study. The first half gives a general overview of concepts and issues, and the second half presents current techniques for the estimation of disclosure risk. While the guide was primarily aimed at Survey Methodologists, the first half is addressed to a more general audience.

Key words: Confidentiality, Disclosure Risks, Public Use Microdata Files (PUMFs)

CONTRÔLE DE LA DIVULGATION STATISTIQUE POUR LES FICHIERS DE MICRODONNÉES À GRANDE DIFFUSION – UN GUIDE

JEAN-LOUIS TAMBAY

RÉSUMÉ

Statistique Canada diffuse des Fichiers de microdonnées à grande diffusion (FMGD) à partir de fichiers maîtres depuis les années 80. Deux composantes principales de la production de FMGD sont l'évaluation du risque de divulgation aux niveaux de l'unité et du fichier et la mise-en-œuvre de mesures pour maintenir ces risques à des niveaux acceptables. Ce guide condensé vise à offrir au lecteur une bonne appréciation des risques et des enjeux concernant les données des FMGD et de servir de point de départ pour une étude approfondie. La première partie donne un aperçu général des concepts et des enjeux, et la deuxième présente des méthodes courantes pour l'estimation du risque de divulgation. Ce guide a été rédigé principalement pour les méthodologistes d'enquête, mais la première partie s'adresse à un auditoire plus large.

Mots clé : Confidentialité, risques de divulgation, Fichiers de microdonnées à grande diffusion (FMGD)

Introduction

This paper gives an overview of Statistical Disclosure Control concepts and methods for anonymized microdata files, called Public Use Microdata Files (PUMFs) at Statistics Canada. It has two parts. Part I, a general summary, introduces a few key concepts, presents factors affecting the disclosure risk and outlines common strategies to protect PUMF data. Part II discusses methods to measure the disclosure risk and identify units at risk and proposes approaches for individual and household level data. The report was primarily written to familiarize Survey Methodologists with issues and tools concerning the protection of PUMF data confidentiality and to serve as supporting reference material for other Statistics Canada documents. In its treatment and discussion of subjects, particularly in Part II, the report reflects the author's points of view. For the sake of completeness, a few concepts and techniques not covered in the main text are presented in Appendix A. While the paper includes references to protected internal documents its contents are not confidential.

I. General Concepts and Methods

1.1 Types of disclosure

Microdata files are files of records pertaining to individual respondent units, where respondents can be persons, households or businesses. In ***hierarchical microdata*** files the respondents are linked into higher order units, e.g., individuals by household or students by school. All microdata must be regarded as confidential and protected against identity and attribute disclosure. ***Identity disclosure*** occurs when a particular data record is correctly associated with a particular unit in the population. ***Attribute disclosure*** occurs when it is possible to associate a particular attribute with a particular population unit. Identity disclosure leads to attribute disclosure but attribute disclosure can occur without identity disclosure. When an individual is known to be in a microdata file, what is called ***response knowledge***, then attribute disclosure can occur if every respondent sharing certain characteristics known about that individual also share another attribute (e.g., a dentist is known to be on a file where every dentist reported taking antidepressants). Response knowledge increases the risks of both identity and attribute disclosure. Since it is unavoidable with census-like data, a designation that includes administrative data, PUMFs typically hold sample data. With sample data the focus is on preventing identity disclosure, but one should keep attribute disclosure in mind – especially when dealing with highly sensitive personal information.

A related concept is that of ***residual disclosure***, which is disclosure that occurs by combining released information with previously released or publicly available information – including aggregate results obtained through venues such as remote access or Research Data Centres (RDCs).

1.2 Factors affecting disclosure risk

The risk of re-identification is affected by factors such as: (a) characteristics of the population and sample, (b) attributes of the data, (c) availability of related outputs, and (d) motivation and

opportunity of intruders. Characteristics of businesses are different from those of households or people. Many large businesses are recognizable and difficult to "mask", which is why Statistics Canada has not released a business PUMF[1]. Units with a hierarchical structure are also hard to protect because the amount of identifying information provided by linked units surpasses that provided by the units individually (e.g., rare combinations of spouses' ages/occupations, presence of twins) and because linked respondents can more easily find each other in a microdata file. For those reasons surveys with geographically clustered samples often suppress cluster membership information on their PUMFs.

Other aspects of the sample design can affect disclosure risk. Higher sampling rates are associated with higher re-identification risks. The selection of the sample from a source such as an administrative file or another survey adds a level of risk. Administrative data providers will have a list of the population, which could help them ascertain sample membership and identify surveyed units. Supplemental surveys, whose sample comes from another survey, have to deal with additional risks from linking respondents from both survey PUMFs and pooling their data. Likewise, the linking of longitudinal or panel surveys respondent data across time increases the likelihood of a re-identification; which is why it is difficult to produce "safe" yet analytically useful longitudinal PUMFs.

Disclosure risk is also affected by the attributes of the data, such as whether they are qualitative or quantitative; whether they are available on other, e.g., administrative, sources; if they come from such sources; and how accurately or consistently they are recorded on various sources. For disclosure control purposes data variables are classified as direct identifiers, indirect identifiers and sensitive variables. ***Direct identifiers*** are those variables, such as names, addresses and SINs, that can directly identify respondents. They are removed from PUMFs. ***Indirect identifiers*** are variables that may be known about respondents and that, taken together, may serve to identify some of them. Examples are province of residence, sex, age, marital status, place of birth, ethnicity, education, occupation, household size and dwelling type. Persons with unique combinations of such variables (e.g., female 76 year old professor with a degree in architecture) may be identified if on a PUMF. Disclosure control measures often target indirect identifiers, which are also called ***key variables***. ***Sensitive variables*** represent characteristics that are not considered to be known about respondents and are generally not useful to identify them. Some sensitive variables can possess attributes of indirect identifiers. While exact income is rarely known, except perhaps by some holders of administrative databases, approximate income can serve as an indirect identifier. Note that the inclusion of administrative data on survey PUMFs makes it easier for database holders to link respondents to their database; which is why it is preferable to coarsen or perturb these data on PUMFs. Some quantitative variables, while not particularly identifying, may benefit from having extreme values masked (e.g., height/weight).

Survey data include two other types of variables. ***Design variables***, such as strata and cluster identifiers, and survey weights, and ***paradata***, which are data about the process by which the data were collected. Except for survey weights, which are essential for analysis, these variables are considered to be risky and/or for internal use and are usually not on Statistics Canada PUMFs.

1 A dummy (perturbed) file of small and medium enterprise data was produced so that researchers with researcher data centre access could test program code at their workplace.

Survey weights can become problematic when they reveal withheld sensitive information such as detailed geography (e.g., if low weights relate to a small region), links between units (if household members share the same weight) or identifying characteristics (e.g., a released estimate of 137 blind musicians is tied to one musician on the PUMF with a survey weight of 137). Other issues with survey weights are given in de Waal and Willenborg (1997). Replicate weights, such as jackknife and bootstrap weights, can reveal cluster membership (Mayda, *et al.*, 1996) and are usually not released for clustered surveys at Statistics Canada. The generalized bootstrap technique can be used to generate replicate weights without revealing sample design information (Beaumont and Patak, 2012).

The example with the musician above points to another factor contributing to risk: the availability of related outputs through other venues such as remote access or RDCs. Those outputs may undermine PUMF disclosure control measures such as suppression and perturbation (e.g., top-coding) – possibly leading to a residual disclosure. Options to reduce this risk include imposing restrictions on related outputs (e.g., disallow results based on few respondents, round totals, suppress maximums) and modifying the PUMF data (e.g., round or perturb values and weights, keep a subsample of respondents). The subsampling of respondents also mitigates the risks from response knowledge, and is a useful way to control the risks for areas or subgroups that are oversampled (at minimum, by imposing a maximum sampling rate for PUMF units).

Microdata threat and risk assessments usually involve three intruder scenarios. The first is an attack by an ***intruder*** or ***hacker*** trying to re-identify PUMF respondents. The intruder's aim may be to gain notoriety, to discredit the statistical agency or simply to acquire confidential information. The attack can be opportunistic, where one is interested in identifying anyone and is seeking out vulnerable respondents, or targeted, where one is trying to find the identity of specific respondents. Except for response knowledge situations, addressing the former usually takes care of the latter. The second threat is an attempt by a database holder to link PUMF records to his datafile. The aim may be to enrich the content of his database or to single out certain individuals (e.g., a health insurer trying to link his client database to health survey microdata). A linkage attempt would require the two files to have variables in common; preferably with no data discrepancy. A third threat is that of ***spontaneous recognition***, which occurs when a PUMF user accidentally or otherwise recognizes a respondent, for example, a public figure, a relative or an acquaintance. This is more likely to happen if the respondent has unique characteristics that makes him stand out, and the PUMF user can be quite certain of having identified the right person.

Marsh, *et al.* (1991) considered identification from database holders, journalists and hackers. They presented four conditions for a user to know a re-identification was successful: (*a*) identifying key variables to be recorded the same way on both datasets, (*b*) presence of the individual in the PUMF, (*c*) population uniqueness of the combination of key values and (*d*) verification of population uniqueness. Specifically, $\Pr(\text{identification}|\text{attempt}) = \Pr(a)\Pr(b|a)\Pr(c|a,b)\Pr(d|a,b,c)$. To assess the risk for a 2% PUMF from the UK Census they estimated these probabilities, respectively, as $(0.6)(0.02)(0.02)(0.001) = 2.4*10^{-7}$. Dale and Elliot (2001) updated the values to $(0.18)(0.02)(0.048)(0.001) = 1.73*10^{-7}$.

1.3 Protecting PUMF data

Statistical agencies can take statistical and non-statistical measures to protect the confidentiality of their PUMFs. Non-statistical measures include regulating PUMF access by technical and other means (e.g., restrict access to researchers from accredited institutions); stipulating how the data can, and cannot, be used in PUMF license agreements (e.g., prohibition from linkage); and instituting administrative, financial or legal penalties for failure to respect those agreements[2]. In general, making access to PUMF data easy and anonymous increases the risk of disclosure by spontaneous recognition or response knowledge by a family member or acquaintance.

Other protection measures target the PUMF content. The first step consists of limiting the amount of identifying information present. This means variables and their level of detail, but also information such as cluster (household) membership. Geographic detail is particularly singled out and some agencies have rules or practices governing the geographical detail on their PUMFs. The second step usually involves applying protective measures at the global (file) level. Coarsening and perturbing the data (e.g., round values, add noise, swap data between records) hampers record linkage and other re-identification attempts. Some measures may target records from small areas or subpopulations only, while others (e.g., top-coding) may target extreme values for quantitative variables. Next, data suppression or perturbation is applied at the local (record) level to units at greatest risk of identification. These units may be identified by the application of a population-based rule (e.g., bivariate categories with less than 1000 people), a disclosure risk measure (see Part II), or simply based on subject matter knowledge and experience (e.g., rules for age-dependent characteristics). Additional disclosure control measures may be taken, for example internal attempts at record linkage, which may lead to more data suppression or perturbation to bring down the match rate. Rules for PUMFs at the U.S. Census Bureau (Lauger, *et al.*, 2014) include geographical area population thresholds (usually 100,000), rounding for dollar amounts, a minimum population size of 10,000 for values of categorical variables, and top-coding for continuous variables using the half-percent/three-percent rule (topcodes must include ≥0.5% of all cases and, for variables that apply to subpopulations, include either 3% of nonzero cases or 0.5% of all cases).

In finalizing the PUMF and related documentation decisions need to be made on the amount of detail to provide. Some aspects of the survey design may be deemed sensitive, for example the relation between low weights and subregions, or how neighbouring units are selected within a cluster. To confound intruders it may be decided not to flag imputed values on the PUMF. Information about the perturbation strategy that may be useful to hackers, such as the scope and levels of perturbation and swapping, may be provided in general terms only – although maximum perturbation rates may be given to discourage re-identification attempts. Conversely, there is support in the statistical community for more openness on the disclosure control strategy based on the notion that a good data protection strategy should not rely on the withholding of information about its particulars. Information loss measures such as the impacts of perturbation on aggregates or on relationships between variables may be provided to users (for examples of loss measures see Domingo-Ferrer and Torra (2001)).

---

2 At the Australian Bureau of Statistics (2009) consequences of failing to comply with the Confidentialised Unit Record File (i.e., PUMF) undertaking include fines of up to $AUS 21,600 and/or imprisonment of up to 2 years.

II. Disclosure Risk Measures and their Application

As noted above, the protection of microdata includes the application of data suppression or perturbation to individual units identified as being at risk based on different criteria. Here we focus on criteria that relate to the application of disclosure risk measures. We present three risk measures, introduce the concepts of multiplicity and special uniques, and show how these can be combined to protect individual and household level data. We also give an example of how Census data were analyzed to develop population size risk thresholds for health microdata.

2.1 Estimating the disclosure risk

Disclosure risk measures for microdata typically focus on the successful re-identification of respondents using a set of identifying key variables whose values are known to an intruder. Attention is mainly directed at units whose combination of values for those key variables is unique in the sample (***sample uniques***). Some measures estimate the probability of being unique in the population given that one is unique in the sample (***union uniques***), others the probability of a match to a PUMF respondent being correct. For notational purposes the set of key variables, which are categorical or made to be, may be concatenated into a single key that takes K values (e.g., with 2 sexes, 10 age groups, 6 marital statuses and 20 occupations K=2400). The number of units with the k^{th} combination of values in the sample is f_k and the corresponding number of units in the population is F_k. Sample uniques have f_k=1, union uniques have f_k=1 and F_k=1. Three approaches to measuring risk are discussed.

a) Poisson Negative Binomial Approach

In the individual risk methodology of Benedetti and Franconi (1998) disclosure scenario assumptions are that (a) a register of the population is available to an intruder, (b) data are from a sample and weights are available, (c) the register contains a set of key variables that are also in the sample, (d) the intruder tries to link units on both files using those variables, (e) the intruder has no extra information than what is in the register, and (f) re-identification occurs when a link between a unit in the sample and the register is established and the link is correct. In the worst case, it is assumed that a re-identification attempt is made (Pr(attempt)=1) and the key variables are recorded the same way on both files (no data discrepancy). In this scenario each of the f_k sample units can be linked to each of the F_k register units sharing the key value k.

Conditional on a re-identification attempt the probability of a link being correct is $1/F_k$. For a sample unit i with key value k=$k(i)$ the estimated re-identification risk is $\hat{r}_k$ =E($1/F_k|f_k$). Following work by Bethlehem, Keller and Pannekoek (1990) a superpopulation approach is proposed where the F_k follow a Poisson distribution and the f_k given F_k follow a Binomial (F_k,p_k) distribution. This gives a Negative Binomial posterior distribution for F_k given f_k. Polettini (2003) obtains an expression for the risk in terms of the Hypergeometric function with parameters involving f_k and p_k. The p_k are estimated by $f_k/\hat{F}_k = \bar{w}_k^{-1}$, the inverse of the average sample weight for units with key value k. When f_k=1, $\hat{p}_k$=w_i^{-1}, the inverse weight for sample unique i, and $\hat{r}_k$=ln(w_i)/(w_i–1); but as f_k increases calculating $\hat{r}_k$ becomes very cumbersome.

Unlike other approaches, this measure does not only cover sample uniques. A critique is given in Rinott (2003). He notes that under full information (F_k are known) and $f_k=1$ the $\hat{r}_k$ overestimate the true $r_k=1/F_k=p_k/f_k$, with severe overestimation for small p_k (around 0.01). Conversely, the risk can also be underestimated because $\hat{F}_k$ overestimates F_k, especially for small f_k. Note also that, given f_k, $\hat{r}_k$ depends entirely on the average cell weight $\bar{w}_k$ and ignores all other information from the table. Thus for sample uniques a maximum risk threshold r^* can be converted to a minimum weight threshold w^*, so that with a low enough sampling fraction uniques will never be deemed to be at risk. The measure is incorporated in the μ-ARGUS microdata protection program, with an approximation used for large f_k (Hundepool, *et al.*, 2014).

Global risk measures can be produced at the domain or file levels. The expected number of re-identifications is $\Sigma_k f_k \hat{r}_k$ and the expected re-identification rate is $\Sigma_k f_k \hat{r}_k / \Sigma_k f_k$. In μ-ARGUS, setting a file-level maximum re-identification rate of ξ^* leads to treating all records with $\hat{r}_k > r^*$, where r^* is set so that $\Sigma_k f_k \min(\hat{r}_k, r^*)/\Sigma_k f_k < \xi^*$.

b) Poisson Log-linear Model Approach

An approach that gives better results was developed by Skinner and Holmes (1998). They focus on two record-level risk measures for the worst case situation of sample uniqueness (and assuming no data discrepancy). The measures are $r_{1k}=\Pr(F_k=1|f_k=1)$, the probability of also being unique in the population, and $r_{2k}=E(1/F_k|f_k=1)$, the probability of a correct match. Summing these measures over the set of sample uniques give file-level measures $\tau_1^*=\Sigma_{SU} r_{1k}$, the expected number of union uniques in the sample, and $\tau_2^*=\Sigma_{SU} r_{2k}$, the expected number of correct matches for the sample uniques. With K large enough these two values will closely approximate the actual number of union uniques, $\tau_1=\Sigma_k I(f_k=1,F_k=1)$, and $\tau_2=\Sigma_k I(f_k=1)/F_k$.

The model assumes $F_k \sim \text{Poisson}(\lambda_k)$ and Bernoulli sampling with inclusion probability π_k for units in cell k, such that $f_k \sim \text{Poisson}(\pi_k \lambda_k)$ and $F_k|f_k \sim \text{Poisson}(\lambda_k(1-\pi_k)) + f_k$. This gives record-level measures $r_{1k} = e^{-(1-\pi_k)\lambda_k}$ and $r_{2k} = (1 - e^{-(1-\pi_k)\lambda_k})/(1 - \pi_k)\lambda_k$, with corresponding file level measures. As in Small Area Estimation, one borrows strength with log-linear model $\log(\lambda_k)=\boldsymbol{x}'_k\boldsymbol{\beta}$ where $\boldsymbol{x}_k$ is a design vector depending on the key variables in cell k and $\boldsymbol{\beta}$ is a parameter vector. Using maximum likelihood estimates of $\boldsymbol{\beta}$ risk measures are generated by replacing λ_k by $e^{x'_k\hat{\beta}}$ in previous expressions. Skinner and Shlomo (2008) developed goodness-of-fit criteria that minimize the bias of the $\hat{\tau}_i$.

This approach gives reasonably good estimates of risk. Furthermore, the parameters λ_k can be estimated under complex sampling schemes using pseudo-maximum likelihood estimation (Rao and Thomas, 2003). Shlomo and Skinner (2009) have shown how to adjust the log-linear model for misclassification. However, the model generation aspect is intensive, making it less attractive in an environment where a large number of scenarios (sets of key variables) are contemplated. A global risk measure using this log-linear model is freely available in the R-Package sdcMicro (Templ, Kowarik and Meindl, 2015). The sdcMicro program covers several disclosure risk measures and techniques, including those in μ-ARGUS, and the SUDA and DIS-SUDA scores (see section 2.2).

c) Data Intrusion Simulation (DIS)

Skinner and Elliot (2002) estimate the probability that, for a given set of key variables with no data discrepancy, an intruder who matches an arbitrary unit in the population against a sample unique in the PUMF is correct. Among unique matches, the probability of a correct match is given by $\theta = \Pr(CM|UM)=\Sigma_k I(f_k=1)/\Sigma_k F_k I(f_k=1)$. For Bernoulli sampling with probability π, θ can be estimated by simulating a scenario where one successively removes each unit from the sample, copies it back to the sample with probability π, and registers if the unit would be a true or false unique match to the sample. The resulting estimate is $\hat{\theta} = n_1/[n_1+2n_2(\pi^{-1}-1)]$, where $n_j = \Sigma_k I(f_k=j)$. Skinner and Carter (2003) adapted the estimator to Poisson sampling with probabilities $\pi_i (=1/w_i)$, giving $\hat{\theta} = n_1/[n_1+2n_2(\bar{w}_2-1)]$, where $\bar{w}_2$ is the average weight of units in the n_2 sample pairs. This risk value, calculated for a file or domain, can be assigned to every sample unique herein. Note that when $n_1>0$ and $n_2=0$ the estimated risk for uniques becomes 100%.

This approach, like the previous one, makes use of the distribution of key values to calculate the risk – although to a much lesser degree. And it has also been adapted for misclassification (Elamir and Skinner, 2004). However, assigning the risk value to sample uniques can present some peculiarities. The risk for, say, a dentist who is a sample unique may be affected by whether civil and electrical engineers are placed in the same or different occupation categories. Sometimes, increasing the detail for key variables can decrease the risk. And the risk's variance can be quite high. However, the measure generally behaves as expected, and its simplicity makes it a very attractive tool when considering a large number of scenarios or when comparing strategies, like different levels of geographical detail.

d) Studies on re-identification risk using 2001 Canadian Census data

Two studies tried to establish geographic area population size cut-offs for public health microdata using 2001 Census data. In El Emam, Brown and AbdelMalik (2009) Census PUMF data were used to simulate region sizes in 5,000 increments. For different sets of (up to 5) key variables, population size cut-offs were set at where the relationship between the geographic population size and the proportion of uniques flattened out. These cut-offs were modelled on the total number of combinations for the values of the key variables (K), giving cut-offs of $1588K^{0.42}$, $1436K^{0.43}$ and $1978K^{0.304}$ for Western, Central and Eastern Canada, respectively. In El Emam, *et al.* (2010) urban Forward Sortation Area data were used to predict when the percentage of uniques exceeded 0%, 5% and 20% (representing different levels of security). For example, the 5% model was defined as $\text{logit}(\pi_{05})=b_0POP+b_1K+b_2(POP*K)$, where π_{05} is the probability that the area of size POP has >5% uniques. The model could be used to determine if $\pi_{05}>0.5$, in which case the area is too small and must be suppressed/aggregated. With this framework, data custodians could determine the amount of geographic suppression or aggregation in relation to the risks of disclosing a particular dataset.

2.2 Multiplicity, Fingerprints and Special Uniques

The above measures relate to the risk of re-identification for a given set of identifying variables (or *key*). PUMFs usually hold several indirect identifiers and it is not realistic to assume that an intruder will know more than a few of them (and without discrepancy). Risk values can be calculated for different scenarios involving different keys. Boudreau (1995) hypothesized that most population uniques are also unique for a subset of variables. Given a set of m identifying variables, he defined a unit's ***multiplicity*** as the number of 3-way combinations (i.e., 3-way tables) among those variables for which the unit is sample unique. The maximum multiplicity is $\binom{m}{3}$, which is 120, 455 and 1140 for m=10, 15 and 20, respectively. A simulation gave a good relationship between multiplicity and uniques, which led to the suggestion of treating units whose multiplicity is above some threshold value.

A method for the determination of multiplicity thresholds is needed. Some surveys use arbitrary thresholds, for example they focus on the x% of records with the highest multiplicities, ignoring key aspects of the design such as the sample rates. For the Census long form sample, with population size N and sample size $n \approx 0.2N$, the following approach was proposed in Kanagarajah, *et al.* (2009). For a sample unique with key value k, the estimated proportion of units in the population with this key value is $p=1/n$. The number of units not in sample with this key value is assumed to follow the Binomial($N–n$,p) distribution. Since $N–n$ is large and $(N–n)p \approx 4$, the Binomial distribution can be approximated by a Poisson distribution with $\lambda=4$, giving probability of uniqueness close to $e^{-4}=0.0183$. For a sample unique, an overestimate of the probability of being unique in the population is 0.0183 times its multiplicity. So a unit can be declared to be population unique if its multiplicity is above $q/0.0183$ for some user-specified q (≤ 1). Note that this approach does not take into consideration the number of identifying variables present which, as seen above, has a great impact on the multiplicity range.

A concept similar to multiplicity was proposed by Willenborg and Kardaun (1999). ***Fingerprints*** are combinations of values for identifying variables that are unique in the sample, and for which no proper subset of variables is unique. Records with "many short" fingerprints are deemed to be risky.

In examining the percentage of sample uniques that are union uniques Elliot, Skinner and Dale (1998) noted that the expectation that an increase in (geographical) detail increased the risk was not supported by empirical data. They distinguished ***special uniques*** from ***random uniques*** (i.e., uniques generated by the sampling process). Special uniques are unique at more aggregate geographical levels. As the detail increases, an increase in random uniques can mask the effects of increasing population uniques, possibly leading to a smaller than expected increase in risk. They determined that special uniques tend to be less sensitive to changes in sampling fraction or geographical detail, and appear with a smaller number of variables (shorter fingerprints), than random uniques. This concept was used in the development of the Special Uniques Detection Algorithm (SUDA) program (Elliot, Manning and Ford, 2002). SUDA generates all possible key subsets from K variables looking for Minimal Sample Uniques (MSUs – similar to fingerprints). From each MSU a score is generated which relates to the number of higher dimensional tables, up to a maximum of M dimensions, in which the unit will also be sample unique (e.g., a sample unique

based on 3 variables will also be sample unique in all 4-way and 5-way tables involving those 3 variables). The scores are summed at the unit level. The scores are *ad hoc*, but they open up the possibility for differential treatment of records based on their 'risk' level. Elliot and Manning (2004) developed a method that runs DIS to calibrate SUDA. Their empirical results show a strong relation between a unit's DIS-SUDA score and its probability of being population unique.

Without resorting to multiplicity the μ–ARGUS program (Hundepool, *et al.*, 2014) offers three ways of processing multiple risk scenarios from a set of identifying key variables. In the risk model approach the risk $\hat{r}_k$ from the Poisson Negative Binomial model is calculated for cells k in tables that are specified by the user, and risk thresholds are applied in each table as explained in section 2.1a. Specified tables cannot have variables in common, which is a severe limitation. In the two other approaches multiple tables are generated by μ–ARGUS and user-specified thresholds for sample frequencies f_k are used to identify cells, and thus units, at risk. In the first case key variables can be assigned identification levels 1 to 5 (or less). For example, with three identification levels used, all 3-way tables containing at most one level-3 variable, at most two level-2 variables and at most three level-1 variables are generated. A single, user-specified, threshold for the f_k is applied in every table. Alternatively, users can specify a maximum number of table dimensions and all tables containing up to that number of dimensions are generated. For each number of dimensions a different threshold for the f_k can be specified by the user.

Statistics Canada's internal program CoMicDIS (Tambay, 2016) combines DIS and multiplicity. From the user-supplied list of key variables all 3-way and, if requested, 2-way tables are generated. For each table the DIS value $\hat{\theta}$ is assigned to sample uniques; other units get $\hat{\theta}$=0. A unit's risk is defined based on its five "worst" tables as $\hat{\theta}_5 = 1-(1-\hat{\theta}_{[1]})(1-\hat{\theta}_{[2]})(1-\hat{\theta}_{[3]})(1-\hat{\theta}_{[4]})(1-\hat{\theta}_{[5]})$. If tables are independent $\hat{\theta}_5$ measures the probability of a correct match among the five riskiest attempts. For units with $\hat{\theta}_5$ above user-specified thresholds variables to suppress are identified which will make the risk acceptable. Users can also identify domain variables which, unlike key variables, are included in every table (increasing their number of dimensions). Users have two ways of incorporating domain variables in the calculation of $\hat{\theta}$. For example, with domain variable province, 3-way tables can be produced separately for each province or the province variable can be used to turn every 3-way table into a 4-way table. The first case results in province-specific values for $\hat{\theta}$ while the latter generates a single (national) $\hat{\theta}$ for 4-way uniques. As an option, domain-specific multiplicity thresholds can be generated based on the expected number of units at risk (sum of $\hat{\theta}_5$), so that units failing either their risk threshold or their multiplicity threshold are treated.

A comparison of the SUDA, Negative Binomial (μ-ARGUS) and Poisson Log-linear Model approaches was done at the Office for National Statistics (Shlomo and Barton, 2006). The Poisson Log-linear model approach, although more complex, gave the best estimates for disclosure risk compared to true risk. With the μ-ARGUS approach the disclosure risk was underestimated, and the per-record risk measure was found not to have enough variability to be correlated with the true per-record risk. SUDA provided a relative per-record disclosure risk measure that was correlated with the true risk, and presented an ordering of units according to their disclosure risk.

2.3 Discussion

"Although there is no shortage of [Statistical Disclosure Limitation] methods which have found application by government statistical agencies, a common scientific methodology for assessing disclosure risk and making decisions based upon these assessments has found less ready adoption in agency practice." (Skinner, 2012)

"Although there is abundant theoretical and empirical research, our review reveals lack of consensus on fundamental questions for empirical practice: how to assess disclosure risk, how to choose among disclosure methods, how to assess reidentification risk, and how to measure utility loss." (Prada, *et al.*, 2011)

We presented three approaches to measuring disclosure risk for microdata and showed how concepts such as multiplicity can be used to identify units at risk over a range of scenarios corresponding to different subsets of identifying variables. As noted above, there is no consensus on methods to assess the risk. This could in part be because risk measures tend to focus on small aggregates, mostly sample uniques, where sampling theory performs poorly and modeling becomes unavoidable. In many ways, protecting PUMF data seems to be more an art than a science. Even with a good measure, the risk will depend on how scenarios are created: how key variables are defined, how they are combined, at what (domain) level risk is evaluated. Empirical investigations with CoMicDIS showed that scenarios can affect risk outcomes substantially when dealing with hierarchical data. There is also the issue of what constitutes an acceptable risk at the unit or file level. Is it 10%, 1%, 0.1%, ...? And how should the risk account for factors such as the likelihood of an attempt at disclosure, the likelihood of a unique individual being in a register or in an intruder's *circle of acquaintances*, the likelihood that a particular key is known to the intruder, the likelihood – and the magnitude – of a data discrepancy, the likelihood of verifying a match? Estimates for some of those values were given at the end of section 1.2, which resulted in risks near $2*10^{-7}$. The presence of a data discrepancy does not eliminate the risk of a correct match; Winkler (1997) asserts that re-identification with PUMFs is far easier than people may think when powerful record linkage methods are used. Finally, risk measures assume that intruders only have the set of key variables at their disposal. But once an intruder has identified units of interest he may try to use other PUMF variables, even sensitive variables, to improve his re-identification results.

Thus, while disclosure risk measures play an important part in the identification of units at risk, they have limitations and should not be relied on exclusively. Rather than focus on absolutes (e.g., is the calculated risk above x%) the measures should be considered as a tool, to examine the relative impact of different dissemination strategies, to identify units and domains at greatest risk of re-identification and requiring the most attention, or even to see how a PUMF's overall level of risk compares to those of recently released PUMFs. A comprehensive strategy for the treatment of units at risk should include other practices such as the application of agency rules, the analysis of quantitative variables, the simulation of re-identification attempts and the use of subject matter knowledge and experience to identify potentially problematic cases. Attempts at re-identification or record linkage may be warranted in situations such as when related files are publicly available. Some agency rules are given in Lauger, *et al.* (2014) and in Schulte Nordholt (2001). Rules and

guidelines at Statistics Canada are outlined in its Handbook for creating PUMFs (Statistics Canada, 2016). Finally, in vetting the dissemination of PUMFs, Disclosure Review Boards can bring their expertise into the PUMF creation process.

2.4 Special case: Treatment of hierarchical household data

The protection of hierarchical household data is particularly difficult. The pooling of members' data significantly increases the amount of identifying information provided. The analysis of members' data individually is not a satisfactory strategy. Approaches for dealing with the problem are given.

In μ-ARGUS (Hundepool, *et al.*, 2014) the household risk is defined as the probability that at least one individual in the household is re-identified. Assuming independence of re-identification attempts within a household h, the household risk is derived from the individual member risks $\hat{r}_{hi}$ as $\hat{r}_h = 1 - \prod_{i \in h} (1 - \hat{r}_{hi})$. As with individual-level data (section 2.1a) a global measure of the expected re-identification rate is obtained as $\sum_h |h| \hat{r}_h / \sum_h |h|$, where $|h|$ is the size of household h. This re-identification rate can be used to define a threshold r^* for the household risk, so that unsafe households are those with $\hat{r}_h > r^*$. In what they consider a strongly prudent approach, this threshold is converted into an individual level threshold by dividing it by the household size. Thus, a threshold of $r^*/|h|$ is applied to individual level risks for all persons in households of size $|h|$. A major shortcoming of this approach is that it does not consider within household relationships. For example, a household with a 25 year-old male married to a 75 year old female may not show up as identifiable if these individuals are otherwise unexceptional.

Greenberg and Voshell (1990) created household level key variables for their risk analyses. Some variables were created by combining household class with other characteristics (e.g., *Class One, White Husband, Indian Wife*). The list of variables and categories used is given in Appendix B.

An approach tried by Boudreau and Manríquez (2006) consisted of creating hierarchical versions of key variables by concatenating the individual values of household members following a certain order. For example, the hierarchical age variable may be a concatenation of every member's age group, starting with the household head, his/her spouse, their children by decreasing age, etc. Risk analysis would be done by domains defined by region and household size. This approach incorporates the household structure, but in a limited sense. For example, in a three person household the concatenated values could be for a couple and their child, a parent with two children, three unrelated individuals… Using those key variables the authors examined the conditional probability of population uniqueness and the conditional probability of exact matches. They concluded that for households of size 4 and more the possibility of re-identification was nearly certain.

Tambay, Carrillo-Garcia and Kanagarajah (2015) used data perturbation to create a hierarchical PUMF from the National Household Survey. Households and individuals at risk were identified using both the CoMicDIS and rules-based approaches. For example, bounds were imposed on household size, on differences in spouses' ages, on the numbers of different places-of-birth, mother tongues, and visible minorities per household. Additionally, CoMicDIS was run on individual and

household key variables. For the household analyses, households were separated into four types: one-person households, one-couple households, multi-couple households and other households. For each type a different set of key and domain variables was created and used. In one-couple households domains were defined by province, rural indicator, sexes of the couple and a household class variable. There were 24 key variables including household size, household income group, the couple's joint education, occupation or ethno-cultural characteristics, combined characteristics of their children or other members, dwelling characteristics, etc. Some variables, like the age-sex distribution of children, had hundreds of categories. As expected, the risk was much, much higher when working at the household level. Nearly 40% of one-couple households and all multi-couple households exceeded their risk threshold. Another lesson from this endeavour was that the possibilities for coming up with a set of household key variables were nearly endless, and these significantly affected the risk outcomes. When dealing with person level data decisions do have to be made in coming up with risk scenarios (e.g., how to create categories for quantitative key variables, whether to consider chronic health conditions as key variables, what domains to use…) but the amount of flexibility pales in relation to that when dealing with household level data.

When dealing with household level data the vast possibilities for risk scenarios, and their impact, strongly favour the use of multiple approaches and methods. Furthermore, while global recoding and local suppression may be acceptable strategies for protecting individual microdata, the use of data perturbation is almost unavoidable when protecting hierarchical household microdata.

References

Australian Bureau of Statistics. (2009). *1406.0.55.003 - User Manual: Responsible Use of ABS CURFs*, Sept. 2009. (http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/1406.0.55.003Main+Features1Sep%202009?OpenDocument).

Beaumont, J.F. and Patak, Z. (2012). On the Generalized Bootstrap for Sample Surveys with Special Attention to Poisson Sampling. *International Statistical Review*, 80, 127-148.

Benedetti, R. and Franconi, L. (1998). Statistical and Technological Solutions for Controlled Data Dissemination. Pre-proceedings of *New Techniques and Technologies for Statistics, Vol. 1*. Sorrento, 225-232.

Bethlehem, J., Keller, W. and Pannekoek, J. (1990). Disclosure Control of Statistical Microdata. *JASA*, 85, 38-45.

Boudreau, J.R. (1995). Assessment and Reduction of Disclosure Risk in Microdata Files Containing Discrete Data. *Proceedings of Statistics Canada Symposium 95*. Statistics Canada, No. 11-522-XPE.

Boudreau, J.R. and Manríquez, R. (2006). *Research into the Possibility of Releasing Hierarchical Public Use Microdata Files for the Census of Population*. Methodology Branch Working Paper SSMD-2006-008E/F.

Dale, A. and Elliot, M. (2001). Proposals for 2001 Samples of Anonymized Records: An Assessment of Disclosure Risk. *JRSS(A)*, 164, 427-447.

De Waal, A.G. and Willenborg, L.C.R.J. (1997). Statistical Disclosure Control and Sampling Weights. *Journal of Official Statistics*, 13, 417-434.

Domingo-Ferrer, J. and Torra, V. (2001). Disclosure Control Methods and Information Loss for Microdata. In *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, North Holland, 91-110.

Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control – Theory and Implementation*. Lecture Notes in Statistics 201, Springer.

El Emam, K., Brown, A. and AbdelMalik, P. (2009). Evaluating Predictors of Geographic Area Population Size Cut-offs to manage Re-Identification Risk. *Journal of the American Medical Informatics Association*, 16, 256-266.

El Emam, K., Brown, A., AbdelMalik, P., Neisa, A., Walker, M., Bottomley, J. and Roffey, T. (2010). A Method for Managing Re-Identification Risk from Small Geographic Areas in Canada. *BMC Medical Informatics and Decision Making*, 10:18.

Elamir, E. and Skinner, C. (2004). *Record-level Measures of Disclosure Risk for Survey Microdata*. Technical Paper, Southampton Statistical Sciences Research Institute, University of Southampton.

Elliot, M.J., Manning, A.M. and Ford, R.W. (2002). A Computational Algorithm for Handling the Special Uniques Problem. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5), 493-509.

Elliot, M.J. and Manning, A. (2004). *The Methodology Used for the 2001 SARS Special Uniques Analysis*. Mimeo. University of Manchester.

Elliot, M.J., Skinner, C.J. and Dale, A. (1998). Special uniques, random uniques and sticky populations: some counterintuitive effects of geographical detail on disclosure risk. *Research in Official Statistics*, 1, 53-67.

Greenberg, B. and Voshell, L. (1990). Relating Risk of Disclosure for Microdata and Geographic Area Size. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 450-455.

Hundepool, A., de Wolf, P.P., Bakker, J., Reedijk, A., Franconi, L., Polettini, S., Capobianchi, A., Domingo, J. (2014). *μ-ARGUS, User's Manual, Version 5.1*. The Hague, The Netherlands: Statistics Netherlands.

Kanagarajah, S., Mbuluyo, M., Beck, J., Boudreau, J.R. and Coleman, K. (2009). *Fichier de microdonnées à grande diffusion de recensement 2006 – Fichier des particuliers*. Internal report prepared for the Microdata Release Committee, Sept., 2009.

Lauger, A., Wisniewski, B. and McKenna, L. (2014). *Disclosure Avoidance Techniques at the U.S. Census Bureau: Current Practices and Research*. Research Report Series (Disclosure Avoidance #2014-02). U.S. Census Bureau, Washington, DC.

Marsh, C., Skinner, C., Arber, S., Penhale, B, Openshaw, S., Hobcraft, J., Lievesley, D. and Walford, N. (1991) The Case for Samples of Anonymized Records from the 1991 Census. *JRSS (A)*, 154, 305-340.

Mayda, J., Mohl, C. and Tambay, J.L. (1996). Variance Estimation and Confidentiality: They Are Related! *Proceedings of the Survey Methods Section*, SSC Annual Meeting, June 1996.

Polettini, S. (2003). Some Remarks on the Individual Risk Methodology. *Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Luxembourg, April 7-9, 2003.

Prada, S.I., González-Martínez, C., Borton, J., Fernandes-Huessy, J., Holden, C., Hair, E. and Mulcahy, T. (2011). *Avoiding Disclosure of Individually Identifiable Health Information: A Literature Review*. SAGE Open.

Rao, J.N.K. and Thomas, D.R. (2003). Analysis of categorical data from complex surveys: an appraisal and update. Chambers, R.L. and Skinner, C.J. (eds.), *Analysis of Survey Data*, Chichester: Wiley.

Rinott, Y. (2003). On Models for Statistical Disclosure Risk Estimation. *Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Luxembourg, April 7-9, 2003.

Schulte Nordholt, E. (2001). Statistical Disclosure Control (SDC) in Practice: Some Examples in Official Statistics of Statistics Netherlands. *Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Skopje, Macedonia, March 14-16, 2001.

Shlomo, N. and Barton, J. (2006). Comparison of Methods for Estimating Disclosure Risks in Survey Micro-data at the UK Office for National Statistics. *Privacy in Statistical Databases, CENEX-SDC Project International Conference, proceedings*.

Shlomo, N. and Skinner, C. (2009). *Assessing the disclosure protection provided by misclassification for survey microdata*. Southampton, UK, Southampton Statistical Sciences Research Institute, 25pp. (S3RI Methodology Working Papers, M09/14).

Skinner, C.J. (2012). Statistical Disclosure Risk: Separating Potential and Harm, *International Statistical Review*, 80, 349–368.

Skinner, C.J. and Carter, R.G. (2003). Estimation of a Measure of Disclosure Risk for Survey Microdata Under Unequal Probability Sampling. *Survey Methodology*, 29, 177-180.

Skinner, C.J. and Elliot, M.J. (2002). A Measure of Disclosure Risk for Microdata. *JRSS (B)*, 64, 855-867.

Skinner, C.J. and Holmes, D.J. (1998). Estimating the Re-Identification Risk per Record in Microdata. *JOS*, 10, 31-51.

Skinner, C. and Shlomo, N. (2008). Assessing Identification Risk in Survey Microdata Using Log-Linear Models. *JASA*, 103, 989-1001.

Statistics Canada. (2002). *Guidelines for the Creation of Synthetic Data Files*. Internal document, July 24, 2002.

Statistics Canada. (2016). *Handbook for Creating Public Use Microdata Files (Revised)*. Internal working document – not for release outside Statistics Canada *(forthcoming)*.

Tambay, J.L. (2016). *CoMicDIS v1.09 Program Documentation*. Internal document, Feb. 23, 2016.

Tambay, J.L., Carrillo-Garcia, I. and Kanagarajah, S. (2015). *A New Approach for the Development of a Public Use Microdata File for Canada's 2011 National Household Survey*. Statistics Canada, Catalogue no. 99-137-X.

Templ, M., Kowarik, A. and Meindl, B. (2015). Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro. *Journal of Statistical Software*, vol. 67, issue 4.

Willenborg, L. and Kardaun, J. (1999).Fingerprints in Microdata Sets. *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Thessaloniki, March 8-10, 1999.

Winkler, W.E. (1997). *Views on the Production and Use of Confidential Microdata*. U.S. Census Bureau Research Report no. RR97/01.

Appendix A. Definitions

Inferential Disclosure: Inferential disclosure can occur when sensitive information can be inferred with high confidence from statistical properties of the data. For example, a model that can provide very accurate estimates for a sensitive variable given externally available covariates. Inferential disclosure relates more to aggregate or model outputs than to microdata.

Dummy files: Heavily perturbed microdata files created from survey Master Files for program testing purposes. Dummy files allow researchers with access to Master Files (at RDCs or indirectly through remote access) to validate their programs externally. Also called *synthetic files* (preferably *synthetic (dummy) files*) at Statistics Canada, where their creation is subject to the Guidelines for the creation of synthetic data files (Statistics Canada, 2002).

Information Loss Measures: Numerical assessment of the impact of disclosure control measures such as data suppression or data perturbation on information content. E.g., percentage of values suppressed, or impacts of perturbation on aggregate results. See Domingo-Ferrer and Torra (2001) for examples.

K-anonymity: A dataset satisfies k-anonymity for $k>1$ if, for each combination of key variables, there are at least k records in the dataset sharing that combination. This concept is not used in protecting PUMF data at Statistics Canada as sample uniques are allowed. The concept is also criticised for not protecting against attribute disclosure if all k individuals share the same value for another characteristic.

Microaggregation: A method for protecting quantitative microdata that replaces data for groups of k records by an average value for the group (e.g., $k = 3$ or 5 or 10). Protection is offered if no member's data dominates the group. Groups are formed using criteria of maximal similarity. Microaggregation can be univariate (different groupings for different variables) or multivariate (one grouping for several variables). The former is simpler and entails less data loss, but the latter offers better data protection.

Post-randomization (PRAM): A perturbation of categorical microdata using a probability mechanism. For a variable ξ with K categories, perturbed values X are generated using transition probabilities $p_{kl} = \Pr(X = l | \xi = k)$, e.g., Pr(race=black|race=white). PRAM is fully described by the transition matrix $\boldsymbol{P}$ with elements p_{kl}. $\boldsymbol{P}$ is known, so characteristics of true data can be estimated from the perturbed file.

R-U Confidentiality Map: A R-U confidentiality map is the set of paired values (R,U) of disclosure risk and data utility that correspond to various strategies for data release. R is a numerical assessment of the risk of unintended disclosures following the dissemination of the data. U is a numerical assessment of the usefulness of the released data for legitimate purposes (the opposite of Information Loss). E.g., after applying additive noise to quantitative data one can set R as the expected percentage of records that can still be correctly re-identified and U as the reciprocal of the variance-covariance inflation.

Synthetic datasets: Instead of releasing original data on PUMFs, multiply imputed synthetic datasets are created by replacing sensitive values with repeated draws from a model fit to the original data. The resulting microdata file is non confidential yet analytically useful. At Statistics Canada the term synthetic data is still used to describe dummy files. See Drechsler (2011).

Appendix B. Categorical breakdowns of household variables used in analysis (Greenberg and Voshell, 1990)

11. Household Class
    a. Householder has Spouse Present (Class One)
    b. Householder has No Spouse Present, Living with One or More Other Persons (Class Two)
    c. Single Person Household (Class Three)

1. Tenure
    a. NA
    b. Owner Occupied
    c. Renter with Cash Rent
    d. Renter with No Cash Rent

2. Household Type
    a. Everyone in Household Related
    b. At Least Two but Not All Persons in Household Related
    c. Single Person Household
    d. Otherwise

3. Race
    a. Class One, White Husband, White Wife
    b. Class One, White Husband, Black Wife
    c. Class One, White Husband, Indian Wife
    d. Class One, White Husband, Asian / Pacific Islander Wife
    e. Class One, Black Husband, White Wife
    f. Class One, Black Husband, Black Wife
    g. Class One, Black Husband, Indian Wife
    h. Class One, Black Husband, Asian / Pacific Islander Wife
    i. Class One, Indian Husband, White Wife
    j. Class One, Indian Husband, Black Wife
    k. Class One, Indian Husband, Indian Wife
    l. Class One, Indian Husband, Asian / Pacific Islander Wife
    m. Class One, Asian / Pacific Islander Husband, White Wife
    n. Class One, Asian / Pacific Islander Husband, Black Wife
    o. Class One, Asian / Pacific Islander Husband, Indian Wife
    p. Class One, Asian / Pacific Islander Husband, Asian / Pacific Islander Wife
    q. Class Two, Male Householder, White
    r. Class Two, Female Householder, White
    s. Class Two, Male Householder, Black
    t. Class Two, Female Householder, Black
    u. Class Two, Male Householder, Indian
    v. Class Two, Female Householder, Indian
    w. Class Two, Male Householder, Asian / Pacific Islander
    x. Class Two, Female Householder, Asian / Pacific Islander
    y. Class Three, White
    z. Class Three, Black
    aa. Class Three, Indian
    bb. Class Three, Asian / Pacific Islander
    cc. Otherwise

4. Ethnicity
    a. Class One, Both Spouses Spanish
    b. Class One, Male Spouse Spanish
    c. Class One, Female Spouse Spanish
    d. Class Two, Male Householder Spanish
    e. Class Two, Female Householder Spanish
    f. Class Three, Spanish
    g. Otherwise

5. Children
    a. NA
    b. Householder with Own Children Under 6
    c. Householder with Own Children Ages 6 - 17

d. Householder with Own Children, Some Under 6 and Some 6 - 17
e. Householder without children

6. Marital Status
   a. Now Married
   b. Widowed
   c. Divorced
   d. Separated
   e. Never Married

7. Payment (Rent or Mortgage Plus Utilities, Tax, Insurance, Etc.) 0,[1-50),[50-75),[75-100),[100-125),[125-150),[150-175),[175-200),[200-250),[250-300),[300-400),[400-500),[500-600),[600-700),[700-800),[800-900),[900-1000),[1000-∞)

8. Employment / Unemployment
   a. Class One, Both Spouses Unemployed
   b. Class One, Husband unemployed, Wife Employed
   c. Class One, Husband Unemployed, Wife Not in Labor Force
   d. Class One, Husband Employed, Wife Unemployed
   e. Class One, Husband Not in Labor Force, Wife Unemployed
   f. Class One, Both Spouses Not in Labor Force
   g. Class One, Husband Not in Labor Force, Wife Employed
   h. Class One, Husband Employed, Wife Not in Labor Force
   i. Class One, Both Spouses Employed
   j. Class Two, Male Householder Unemployed
   k. Class Two, Male Householder Not in Labor Force
   l. Class Two, Male Householder Employed
   m. Class Two, Female Householder Unemployed
   n. Class Two, Female Householder Not in Labor Force
   o. Class Two, Female Householder Employed
   p. Class Three, Unemployed
   q. Class Three, Not in Labor Force
   r. Class Three, Employed
   s. Other

9. Veteran Status
   a. Class One, Husband Veteran
   b. Class One, Wife Veteran
   c. Class One, Both Spouses Veterans
   d. Class Two, at Least One Male in Household is Veteran
   e. Class Two, at Least One Female in Household is Veteran
   f. Class Two, at Least One Male and at Least One Female are Veterans
   g. Class Three, Veteran
   h. Otherwise

10. Disability
   a. Class One, Husband Disabled
   b. Class One, Wife Disabled
   c. Class One, Both Spouses Disabled
   d. Class Two, Male Householder Disabled
   e. Class Two, Female Householder Disabled
   f. Class Three, Disabled
   g. Otherwise

12. Household Income (-∞,0],[1-1K),[1K-3K),[3K-5K),[5K-7K),[7K-9K),[9K-11K),[11K-13K),[13K-15K),[15K,∞)

13. Social Security 0,[1-500),[500-1000),[1000-1500),[1500-2000),[2000-2500),[2500,∞)

14. Public Assistance 0,[1-500),[500-1000),[1000-1500),[1500-2000),[2000-2500),[2500,∞)

15. Other Income 0,[1-500),[500-1000),[1000-1500),[1500-2000),[2000-2500),[2500-5000),[5K-10K),[10K-15K),[15K,∞)