Toward Quantifying Disclosure Risk for Area-Level Tables When Public Microdata Exists

Tom Krenzke[1] and David Hubble[1]
[1]Westat, 1600 Research Boulevard, Rockville, MD 20850

Abstract

There are several disclosure risk factors to consider prior to publishing tables or releasing an on-line data tool with underlying restricted use data. For example, secondary analysts may want to produce tables at a geographic level lower than allowed for public use microdata. Rules are needed from the data owner or licensor that balance the risk with data utility; preferably based on data driven analyses. A framework is presented for an initial assessment of disclosure risk when releasing a limited number of data tables. Considerations for risk factors include the matchability to existing files, level of detail in variables, impact of data masking, among others. As an illustration, risk factors are quantified using public use data from the American Community Survey to help measure the disclosure risk associated with the planned Census Transportation Planning Products.

Key Words: Data matching, confidentiality

1. Introduction

There are several disclosure risk factors to consider when generating tables from restricted use data, perhaps for a report for particular geographic areas when public use data also exists. Elliot (2001) discusses several factors, including sampling fraction, level of detail on variables, level of geographic detail, and the size of the matching key. A matching key is a concatenation of the known variables and can serve as an ID when matching to other files. Elliot also discusses the reduction of risk through data divergence, which includes response error, data coding error, data entry error, data aging, variable constructs, imputed data, and effects of disclosure control techniques (e.g., swapping). Further, he discusses concerns about table linkage, where a data snooper can link together published tables to arrive at a pseudo-microdata record for an individual. As an illustration of how to account for these factors, we use a scenario relating to the 3-year American Community Survey (ACS) Census Transportation Planning Products (CTPP) tables where recently considered disclosure rules have become more strict. The overall impact of the disclosure rules would be a fairly substantial amount of data loss, and questions have been raised about the "true" level of disclosure risk. Efforts here work toward a model to quantify disclosure risk in an attempt to determine if the rules can be relaxed, or determine the level of variable reduction needed to sufficiently decrease disclosure risk, or some alternative solution. Using the CTPP scenario, a framework is presented for an initial assessment of disclosure risk. Risk factors are identified, and the disclosure risk is measured for each as we work toward an overall disclosure risk measure.

2. CTPP Background

The 3-year special transportation planning tabulations are generated from ACS restricted use data by the Census Bureau for the American Association of State Highway Transportation Officials (AASHTO). The ACS is a large rolling sample survey that effectively has a sampling rate of 1.5% each year (after nonresponse). The Census Bureau produces a 1-year and 3-year Public Use Microdata Samples (PUMS) for a two-thirds subset of ACS sample for geographic areas that have a population of 100,000 or more. The ACS also produces a set of 1-year CTPP tabulations for populations of 65,000 or more and a set of 3-year CTPP tabulations for populations of 20,000 or more. The planned tables are comprised of Means of Transportation (MOT) (e.g., drove alone, 2-person carpool, etc) crossed pairwise with over 15 variables. There are three general sets of tables produced, including residence, workplace and flow tables. In the past decade, the disclosure rules being considered have become more strict with the "Rule of 3". That is, cell suppression would occur if there exists unweighted counts of one or two persons in the marginals of a pairwise crosstabulation with MOT. Suppressing cells that contain an unweighted cell count with less than or equal to two cases, as well as subsequent complimentary suppressions, is a common rule applied in tabular releases. The only reason to do so for counts of two is if a person in the sample finds his/her own cell in each table, and then that respondent can identify his/her cell partner as a pseudo-microdata record. Such a rule can be viewed as additional protection beyond what is provided through a microdata file. The overall impact of the disclosure rule would be a fairly substantial amount of data suppression, and questions have been raised about the "true" level of disclosure risk. The "Rule of 3" reduces the disclosure risk, however, it results in suppressed data in an estimated 80% or more of places in the nation using a 10-level *MOT* variable (Miller 2008). Ironically, there were no constraints on the CTPP generated from the 2000 Census Long Form, which had a higher sampling rate than the ACS, and was for one-year only.

3. Motivating Scenario

Here is a motivating scenario that is simplified for illustration purposes. Suppose the set of crosstabs, as shown in Table 1, are created from restricted use data for the transportation tables for county 1 of PUMA 1 – using MOT for example: MOT * gender, MOT* travel time, and MOT* occupation. Suppose County 1 has a population of 20,000, and there is a count of 1 for a category of MOT, that is, 1 biker/walker responded in County 1. A data snooper can link together tables and generate a pseudo-microdata record for an individual from PUMA 1, County 1, who is a male mathematician who bikes or walks over 30 minutes to work.

Table 1: Unweighted Cell Count Illustration, County 1 of PUMA 1, Population of 20,000

	Means of Transportation	
	Biker/Walker	*Other*
Total	1	399
Gender		
Male	1	199
Female	0	200
Occupation		
Mathematician	1	20
Other	0	379
Length of commute		
30 minutes or less	0	300
More than 30 minutes	1	99

Once the tables are linked together to arrive at a pseudo-microdata record, the snooper can match it to the PUMS for PUMA 1 using the PUMA, MOT, gender, travel time and occupation as the matching key, and obtain many more variables for the "pseudo-microdata" record. Since there is more than one place or county in PUMA 1, there may be many matches. But if there is only one record on the PUMS that matches, then the snooper now not only has MOT, gender, travel time and occupation, but he also has over 100 ACS variables, as well as the county of residence with population 20,000, which is a clear violation of disclosure rules of public microdata. There is also the chance that no records will match since the PUMS is a two-thirds subset of the full sample. With the additional information and lower level of geography, the snooper can then match to external sources to gain one's name, address, phone number, etc from a real estate information file, disabled veterans file, or ancestry list, for example.

4. Framework for an Initial Assessment of Disclosure Risk

Under the motivating scenario described in the previous section, there are some fundamental questions of disclosure risk:

- What is the matchability rate with the PUMS?
- What is inherent in the data that provides protection from disclosure?
- How much protection from disclosure has been added from masking procedures?
- How can the overall disclosure risk be measured in this scenario?

To help answer these questions, we identified several sources of data protection in the transportation table scenario, identified as *Safe1* through *Safe6*. The sources are grouped into 2 major categories. We refer to the first major category as *Initial protection*, in which, for the CTPP context, there is data protection from disclosure due to the extent of non-matchability of the transportation data to ACS PUMS and the amount of protection due to it being part of a sample, rather than a Census. We refer to the second major category as *Additional protection,* in which there is additional protection due to data swapping, people moving or changing workplaces, imputation, as well as data divergence, that is, the uncertainty of a variable due to it being a subjective assessment of the factual (re-identifiable) nature of the variable and how sensitive it is to change over time. The initial evaluation measure for a safe file due to initial protection (InitSafe) can be expressed as:

P(InitSafe) = 1-(1-P(Safe1))(1-P(Safe2)),

where,

P(Safe1) = chance of the data being protected through non-matchability of the CTPP data to ACS PUMS, and
P(Safe2) = chance of the data being protected through being part of a sample, rather than a census.

The chance that additional measures protect data from disclosure can be expressed as:

P(AddtSafe) = 1-(1-P(Safe3))(1-P(Safe4))(1-P(Safe5)(1-P(Safe6)),

where,

P(Safe3) = chance of the data being protected due to perturbation or data swapping,
P(Safe4) = chance of the data being protected due to moving or changing job locations over a 3-year period,
P(Safe5) = chance of the data being protected due to imputation, and
P(Safe6) = chance of the data being protected due to the uncertainty or divergence of the variable.

The impact of the additional measures on the overall value of a safe record is captured as follows:

P(Safe) = P(InitSafe) + (1-P(InitSafe))*P(AddtSafe)).

In the CTPP context, we are making the following assumptions:

1. The snooper does not know if any particular person is in the sample,
2. If a respondent finds data about himself, it is not a disclosure,
3. The sources of disclosure protection are independent of each other,
4. The snooper needs to find an exact true match
5. The snooper has nobody in particular in mind (individual risk), as opposed to the snooper having a specific person in mind (file risk).

5. Initial Protection

Safe1. We conducted a data-driven analysis to estimate the first risk factor, which is the amount of data protection due to the non-matchability to the ACS PUMS. To do so, we created a data snooper emulation using the ACS 2006 PUMS data for Maryland (emulating a pseudo-restricted use file). Using the data, we created a pseudo-place of size 20,000 for each of 44 PUMAs in Maryland. Then we created a pseudo-ACS PUMS file by selecting 2/3 of the individual records. The scenario attempts to simulate a data snooper using a pseudo-microdata record of CTPP coarsened variables as a key, and then matching to the ACS PUMS to gain much more information. The matching keys were set up as shown in Table 2. KEY1 has a reduced set of variables chosen fairly arbitrarily with identifiable characteristics. KEY2 contains all CTPP variables available on the ACS PUMS using a six category MOT(6). KEY3 contains all CTPP variables available on the ACS PUMS using a ten category MOT(10). We note that the file does not include the CTPP variables’ relating to the number of vehicles, and the number of workers in the household, so they are excluded from this analysis. We also note that the levels of the age of youngest child are slightly different than the CTPP variable. These idiosyncrasies are expected to have minimal or negligible impact on our analysis.

To illustrate, suppose there are 150 ACS records, then under the analysis setup, there are about 30 CTPP records (1/5 the size of the PUMA) among the 150 records, and there are about 100 ACS PUMS records (2/3 subset of the ACS full sample). When matching between the 30 CTPP records (treating them as pseudo-microdata records) to the 100 ACS records using the key, if 20 ACS records match exactly, then there is a disclosure problem. If 50 records match, then the intruder has only 20/50 chance of a correct match. An exact matching criteria is used, however, other probabilitistic matching approaches can be applied as well.

Table 2: List of Matching Keys

Key	*List of variables (levels)*
KEY1	MOT (6), Age (7), Disability (2), Minority (2), Occupation (7), Sex (2)
KEY2	MOT (6), Age (7), Class of Worker (8), Disability (2), Earnings (4), Industry (15), Years in US (3), Minority (2), Occupation (7), Sex (2), Time Leave (8), Travel Time (10), Age Youngest (4), Income (8), Poverty (3)
KEY3	MOT (10), Age (7), Class of Worker (8), Disability (2), Earnings (4), Industry (15), Years in US (3), Minority (2), Occupation (7), Sex (2), Time Leave (8), Travel Time (10), Age Youngest (4), Income (8), Poverty (3)

Note: The number of levels is shown in parenthesis

Let chance of the data being protected through non-matchability of the CTPP data to ACS PUMS, adjusted for the level of geography, be as follows:

P(Safe1) = (1-P(ExactMatch))*(CTPP place population/100,000).

Where, P(ExactMatch) = Expected number of matches / Actual number of ACS records that matched.

The P(Safe1) measure accounts for benefits of coarsening and suppression of CTPP variables, and also accounts for disclosure risk due to lower levels of geography. Table 3 provides the average match rate across the 44 PUMAs, and the values of P(Safe1) by matching key. The above analysis shows that even the highly coarsened variables (e.g., earnings, industry, occupation) give a high probability of an exact match (0.99) and low chance of protection (P(Safe1) = 0.002) when the full set of CTPP variables are available as a key (KEY3). However, a reduced set of key variables (KEY1) lends itself to less chance of an exact match (0.238) and more chance for data protection (P(Safe1) = 0.152). The difference between the match rates for KEY2 that uses MOT(6) and KEY3 that uses MOT(10) is negligible.

Table 3: Values of P(ExactMatch) and P(Safe1)

Measure	*KEY1*	*KEY2*	*KEY3*
P(ExactMatch)	0.238	0.988	0.990
P(Safe1)	0.152	0.002	0.002

Safe2. Sampling reduces the risk of disclosure as compared to a census of individuals. The Census Bureau provided a weight distribution for workers 16 and older, residing in Maryland, from the 2006 ACS. The average weight was about 79. For 3-year estimates, the weights can be estimated through division by 3, and therefore the average is about 26. The weights take into account differential sampling rates, nonresponse, and a calibration adjustment. Taking the inverse, the average proportion represented by the ACS participants over the course of 3 years for places with 20,000 or more in Maryland is about 3.8% (compared to about 17% from the 2000 Census Long Form). The maximum rate is about 10% -- however, this is highly unlikely since sub-areas would have had to have the highest sampling rates, which is only applied to very small areas.

The relationship between disclosure risk and sampling fractions has been well documented. The *mu-Argus* 4.1 manual provides a discussion of how disclosure risk is measured by an approximation to the hypergeometric function in the software for their microdata using the sampling fraction when an intruder knows the unweighted cell count is 1 (sample unique). Under the assumptions that the intruder has a full, high quality registry of individuals to match pseudo-microdata variables for CTPP, the probability of a correct match, given a cell count equal to 1 for cell *k*, is expressed as $-\log(p_k)(p_k/(1-p_k)$, where p_k = sampling fraction = 0.038. Then, $P(\text{Safe2}) =1-(-\log(p_k)(p_k/(1-p_k))$. The sampling fraction is estimated as the number of respondents in cell k, divided by the estimated population size in cell k (sum of weights across respondents). For the CTPP, P(Safe2) = 1- 0.056 = 0.944. The value of P(Safe2) is larger than the sampling fraction, and provides an upper bound of the risk. This is because of the uncertainty in the denominator of the sampling fraction due to the estimate of the cell population size by the sum of the weights. Skinner and Shlomo (2008) discuss alternative risk measures using log-linear models.

6. Additional Protection

External files are generally considered the largest threat to disclosure. Winkler (2004) describes the highest standard for estimating the proportion of records that can be re-identified. He describes that record linkage can be used to determine the level of confidentiality of a file, by matching the masked file to the original file. Winkler also discusses a scenario where it may be possible to match 0.5-2.0% of the records, which is stated as being at a “non-confidential” level, and that additional protection would need to be applied. Although we consider this criterion to be strict, nonetheless, we use it to gauge possible criteria that the Census Bureau uses to determine the riskiness of data releases. An upper bound on the value of P(InitSafe) for the CTPP scenario is about 0.942, which translates to 5.8% risk. Using Winkler’s criteria, additional sources of protection need consideration. This section discusses additional sources of data protection, whether it is through swapping, and the realization of other sources inherent in the data, such as moving and workplace changes over time, imputation, or other sources.

Safe3. Swapping is used to reduce the risk of disclosure in ACS data products. The swapping rate is kept confidential within the Disclosure Review Board (DRB). Without any other information available, we assume that about 5% of the records have been swapped (changed) and that it is constant across variables, although we realize that swapping is likely applied to higher risk variables. With a 5% perturbation rate, we set a safe value as P(Safe3) = 0.05, which is a conservative measure of the impact of swapping. With data swapping there is an immeasurable, but not negligible psychological impact, so that the intruder, knowing that the data has been masked, can never be certain what values in a given record have been changed (Winkler 2004). Others recognize that further protection is gained through perturbation approaches and can be recognized as providing adequate protection allowing cell counts less than 3. For example, the National Center for Education Statistics (NCES) Standards[1] 4-10-2 address the “Rule of 3” by applying this rule to tabulations on restricted use files only if confidentiality edits (e.g., perturbation methods such as data swapping) are not used in masking the restricted use files. Regardless of the perturbations, NCES requires matching against an external file (Standard 4-8-2) with the “Rule of 3”, if such a file can be used for identification.

Safe4. For the CTPP residence tables, we estimate P(Safe4) = 0.34, that is about 34% of householders have moved within the past three years. This is an interpolation of the movement within 1 year (20%) and within 5 year (49%) change in residence, as shown in Table 4. For the Workplace tables, we assign P(Safe4) = 0.42, as a conservative estimate, based on McWethy (2008). That is, 42% of persons changed employers during a 3-year period according to data from the Survey of Income and Program Participation. This is a conservative estimate since it is recognized that changing locations under the same employer is not included in the estimate.

Table 4: Year householder moved into unit: 2000

Year Moved	%
Moved in 1999 to March 2000	20
Moved in 1995 to 1998	29
Moved prior to 1995	51

Source: Decennial Census 2000. US Census Bureau

Safe5. Imputation flags will be available on the ACS PUMS for larger geographic areas. Since imputation flags will not be available with the CTPP tables, these values can be considered masked. Table 5 shows imputation rates from the Maryland PUMS and for the nation where available from the Census Bureau website for the full ACS file. Imputation rates on a crosstab of each item with MOT for the nation were estimated by adjusting the Maryland PUMS imputation rate on the crosstab by the ratio of the national univariate imputation rate to the Maryland PUMS univariate imputation rate. The imputation rate becomes P(Safe5).

Safe6. The uncertainty or divergence of a variable is a subjective assessment of the factual (re-identifiable) nature of the variable and how sensitive it is to change over time. As an upper bound on disclosure risk, we set P(Safe6)=0. As a lower bound on disclosure risk, a small additional protection could be added, varying by variable.

[1] Nces.ed.gov/statprog/2002/std4_2.asp

Table 5: Imputation Rates: 2006

Variable(s)	*Maryland PUMS Imputation Rate*	*National Imputation Rate*	*National Adjusted Imputation Rate for Pairwise CrossTabs with MOT*
Age	0.01	0.01	0.03
Age of youngest child	0.01	0.01	0.03
Class of worker	0.03	0.05	0.07
Disability status	0.05	0.03-0.04	0.04
Earnings, income, poverty	0.10	0.13	0.16
Industry	0.04	0.06	0.07
Occupation	0.04	0.06	0.07
Year of arrival	0.01		0.03
Minority status	0.03	0.01 (race) 0.02 (Ethnicity)	0.03
Time leaving home, travel time	0.05	0.09 (leaving home), 0.07 travel time)	0.09
Vehicle availability	0.02		0.02
Workers in hh	0.02		0.03
Sex	#	#	0.02
MOT	0.02	0.04	NA
Place of Work - State	0.02	0.05 (state) 0.06 (place)	0.06

# Rounds to zero

Source: 2006 American Community Survey PUMS

7. Overall Risk

For the Residence tables, individual-level values of risk range from 1.29% to 3.43% for the set of CTPP variables in the analysis, as shown in Table 6, For the Workplace tables, the individual-level values of risk range from 0.97% to 3.02%. It is interesting to note that without the matchability (Safe1), moving/changing (Safe4) and imputation effects (Safe5), the individual risk for the ACS annual PUMS is about 1.90%, which is at about the same level as the 3-year CTPP tabulations. For Flow tables, risk increases since the flow from residence to workplace basically adds an additional variable to the set of variables crossed with MOT. However some reduction is seen from the changes to both residence and workplace, and therefore increasing P(Safe4). The ACS PUMS file that we obtained did not have the Workplace Place, therefore in effect, the unavailability of the Workplace Place on the ACS PUMS file reduces the match rate to the ACS PUMS.

Table 6: Safe and Risk Values for Residence and Workplace Tables
(Place Size = 20,000)

			Chance of Disclosure Protection or Risk (in %)								
Type of Risk	*Bound*	*Key*	*Safe1*	*Safe2*	*Safe3*	*Safe4*	*Safe5*	*Safe6*	*InitSafe*	*Safe*	*Risk*
Residence	Upper	KEY1	15.2	94.4	5	34	2	0	95.3	97.1	2.92
Residence	Upper	KEY2	0.2	94.4	5	34	2	0	94.4	96.6	3.43
Residence	Upper	KEY3	0.2	94.4	5	34	2	0	94.4	96.6	3.43
Residence	Lower	KEY1	15.2	96.2	10	34	16	20	96.8	98.7	1.29
Residence	Lower	KEY2	0.2	96.2	10	34	16	20	96.2	98.5	1.51
Residence	Lower	KEY3	0.2	96.2	10	34	16	20	96.2	98.5	1.51
Workplace	Upper	KEY1	15.2	94.4	5	42	2	0	95.3	97.4	2.56
Workplace	Upper	KEY2	0.2	94.4	5	42	2	0	94.4	97.0	3.02
Workplace	Upper	KEY3	0.2	94.4	5	42	2	0	94.4	97.0	3.02
Workplace	Lower	KEY1	15.2	96.2	10	50	16	20	96.8	99.0	0.97
Workplace	Lower	KEY2	0.2	96.2	10	50	16	20	96.2	98.9	1.15
Workplace	Lower	KEY3	0.2	96.2	10	50	16	20	96.2	98.9	1.15

8. Last Leap of the Data Snooper

Now we turn to the last leap of the data snooper. Some additional risk occurs when a public registry exists. When measuring this risk, we assume that the survey and registry have the same question wording, same year of data collection, and same number of questions. The coverage of the registry with respect to these aspects has a big impact. As a rough approximation, consider this model for P′(Risk), which is a rough approximation of the overall risk measure and the coverage rate of the registry.

P′(Risk) = 1 / (W - (W-1)* F), where, W = 1/ P(Risk), and F = coverage rate of the registry.

At the extremes, if the coverage of the registry approaches full coverage, then P′(Risk) approaches 1 leading the snooper to greater certainty of disclosure. If F approaches 0 (that is, essentially no registry data), then P′(Risk) approaches P(Risk), which was as large as 3.4% in the previous section. Figure 1 shows the values of risk when a public registry exists under the model, whose coverage of the population varies from no coverage to full coverage. This model shows when there is partial coverage, say even up to 60 or 80%, then the registry data does not help the snooper much since there is an entire segment of the population missing. The shape of the curve depends on the exact application; how many variables are known, what types of variables are on the registry, for example. So this is just an illustration and emphasizes that the existence of registries needs to be brought into the discussion.

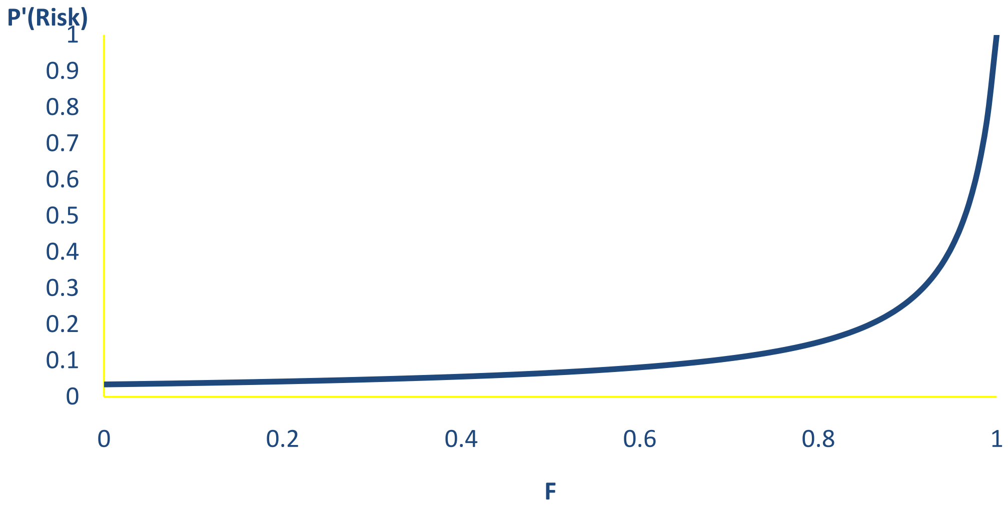

Figure 1: Values of Risk When a Public Registry Exists (Illustrative Model)

9. Concluding Remarks

Our approach to the CTPP risk assessment toward a framework for disclosure risk measure included identifying disclosure risk factors, arriving at disclosure risk estimates for each risk factor by emulating a snooper scenario through a data driven analysis, by using existing disclosure risk formulas where appropriate, and by using existing data on other characteristics, such as moving rates and imputation rates. The last step was to put all the components together to arrive at an overall risk measure.

In our analysis, we find low levels of disclosure risk in the CTPP tabulations of 3-year ACS data. The largest impact among six components of data protection is from sampling. Individual-level values of risk, i.e., assuming that the intruder has the pseudo-microdata record in hand, range from 0.97% to 3.43%. That is, in the worst case, snooper can only be 3.4% confident that he has data about who he thinks it is. While these levels are mentioned as non-confidential in Winkler (2004), the criterion is considered very strict by the authors for this context considering the original file is held within the Census Bureau, and external registry data is sparse.

The analysis results show the need to balance the low risk of the CTPP tabulations based on the 3-year ACS, with the large amount of data suppressed due to the “Rule of 3”. We conclude that the CTPP tabulations are at a low level of risk and the “Rule of 3” is not warranted. Data producers are challenged by the dual objective of maximizing data utility with reducing disclosure risks. Recently a National Science Foundation decision on data suppression for the 2006 Survey of Earned Doctorates was reversed in 2008 due to concerns expressed about reduced data utility. A recommendation is to further investigate the collapsing rules or reducing the number of CTPP demographic variables (non-MOT variables), and determine the effect on matchability to the ACS PUMS.

In summary, secondary analysts that publish tables from restricted use data when microdata is available to public should be aware of table-linking, matching to the public use microdata file, and the existence of registry data.

References

Elliot, M. 2001. Disclosure Risk Assessment. Chapter 4. *Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies*. Pat Doyle, Julia Lane, Jules Theeuwes and Laura Zayatz.

McWethy, L. 2008. SIPP Employment Data Analysis. Memo from Laura McWethy to Elaine Murakami (Federal Highway Administration), dated August 12, 2008.

Miller, D. 2008. Critical Need for Data from the American Community Survey (ACS). Memo from Deb Miller (AASHTO Standing committee on planning) to Christa Jones (Census Bureau). http://trbcensus.com/drb/08052008.pdf

Skinner, C. and Shlomo, N. 2008. Assessing the Identification Risk in Survey Micro-data Using Log-Linear Models. Southampton Statistical Sciences Research Institute Methodology Working Paper M06/14. University of Southampton.

Winkler, W. 2004. Masking and Re-Identification Methods for Public-Use Microdata: Overview and Research Problems. U.S. Bureau of the Census. Statistical Research Division Research Report Series (Statistics #2004-06).