Is an On-Line Microdata Tool That Uses American Community Survey Transportation Data Feasible?

Tom Krenzke,[1] Penelope Weinberger,[2] Jianzhu Li,[1] Elaine Murakami[3]

[1]Westat, 1600 Research Blvd., Rockville MD, 20850
[2] American Association of State Highway and Transportation Officials, 444 N. Capitol St. NW – Suite 249, Washington, DC 20001
[3] Federal Highway Administration, 1200 New Jersey Ave., SE, Washington, DC 20590

Abstract

This feasibility study investigates the value added of testing and developing an online microdata tool using the American Community Survey (ACS) for disseminating Census Transportation Planning Products (CTPP). The current approach for disseminating the CTPP data is through prespecified tabulations. The tables are generated using key transportation variables which include means of transportation, commute time, and time leaving home for work. Several aspects of feasibility are explored. Issues relating to system implementation and hosting are addressed. The progress and timing of the developing systems are provided. Since confidentiality screening is a must in a real-time system that is processed from ACS restricted-use microdata, its tradeoffs with user needs and data utility are discussed.

1. Introduction

Sponsored by the American Association of State Highway and Transportation Officials (AASHTO), this article reports on an investigation into the feasibility and value added of testing and developing an online analytic system (OAS) using ACS microdata for disseminating transportation data related to the CTPP. The current approach for disseminating the CTPP data is through prespecified tabulations. The tabulations have been released periodically after each decennial census, with the most recent released in 2013. The tables are generated using key transportation variables which include means of transportation, commute time, and time leaving home for work for three main parts:

Part 1. Residence-based tabulations;
Part 2. Workplace-based tabulations; and
Part 3. Tabulations for residence-to-workplace flows.

The types of statistics produced are mainly weighted counts, means, medians, and aggregates with associated standard errors for the worker population 16 years and older. A key point of interest to CTPP users is the small geography called traffic analysis zones (TAZs). The TAZs are formed to be small to allow data users the flexibility to combine them to form different geographic areas of interest. However, the small TAZ sizes yield sparse sample sizes in table cells, which create confidentiality concerns for the U.S. Census Bureau, the data producer and data collector for the ACS.

A major challenge to releasing the CTPP prespecified tables to the public is to reduce the disclosure risk such that the risk levels are acceptable to the Census Bureau's Disclosure Review Board (DRB), while retaining the utility of the data. The approach to reducing the disclosure risk for CTPP 2000 resulted in the suppression of a substantial amount of the prespecified tables. Through research conducted under National Cooperative Highway Research Program (NCHRP) 08-79 (NCHRP 2011), Westat worked with the National Academy of Sciences' Transportation Research Board (TRB), ACS operations staff, and the DRB to arrive at an approach to apply statistical disclosure control (SDC) treatments to the microdata before generating the CTPP tables. The result of the research allowed almost all estimates for all table cells to be shown to the data users, and the standard errors incorporated the impact of the treatments on the table estimates.

A limitation of the prespecified tables, however, is that the data user is limited by the information provided and, therefore, after specification of the tables, new analytical interests may arise and the foreseen set of tables becomes insufficient. That is, other tables, variables, and categories of the variables may be desired. In addition, the fixed set of variables and categories may be good for some areas of the country, but not good for others. For some analysts the set of tables is sufficient, while others may want to see different combinations of variables in cross-tabs, or different recodes/categories. Consequently, there is a need to make the present process of producing disclosure-controlled prespecified tables more flexible to better accommodate data user requirements.

The study objectives for the OAS for CTPP data were to conduct an initial investigation into the feasibility of an OAS, and to investigate the value added, that is, value added beyond the current CTPP static tables, in terms of important features to transportation planners. Examples of value added to be achieved may include flexibility in selecting variables and categories of variables, and making updated information available (e.g., fill the gap between periodic CTPP releases). Flows between home and work and small geography also need to be offered; otherwise the utility of such a system is greatly reduced.

During the feasibility study, the project team met several times with Census Bureau staff and identified possible challenges in the implementation of an OAS. Taking the challenges into account, this article provides some formulated scenarios involving an OAS, with some limited input from transportation planners (more input is desirable). A conclusion of the investigation into the feasibility of an OAS for CTPP-related data was not fully realized. However, we remain optimistic about the value added of such a vehicle to data dissemination, in terms of both its usefulness and its cost efficiency. First, the challenges of developing an OAS are described in Section 2. Possible scenarios and approaches for an OAS are discussed in Section 3. Finally, a comparison of the options is given in Section 4.

2. Challenges in building an OAS to meet transportation needs

Before discussing the challenges of developing an OAS, we discuss the challenges related to the CTPP that were generated from the 2006-2010 ACS microdata. First, there were very small geographies (TAZs), where the localities were about the size of block groups. Therefore, the CTPP tables were susceptible to disclosure risk, and the end-result of balancing the risk and utility objectives was that Set A tables and Set B tables were specified. Set A included tables where Census Bureau DRB suppression rules were defined for a small proportion of the Set A tables. These tables could be generated directly from original ACS microdata and could be published as is with very few limitations. Set B contained tables where Census Bureau DRB rules would have come into play for all Set B tables if the underlying data were not perturbed. For Set B, tables were generated from perturbed microdata and then the rules were lifted. If not perturbed, a large proportion of the Set B tables would not be published.

Second, this is a large undertaking. Before generating tables, perturbation was done to the underlying microdata. Then millions of tables were generated, and variance estimation processing was applied to account for the perturbation component of the total variance.

This procedure leads to an option to supplement the CTPP releases through an OAS. In an OAS, there are three main types. The first type requires submitting queries to a canned set of underlying tables, such as the American Factfinder, developed by the Census Bureau. There are no disclosure issues once the canned tables are checked. A second type of OAS is to generate tables from public microdata. This step has no disclosure issues as long as the public use microdata were approved at an acceptable level of risk. Finally, an OAS can produce results from restricted microdata, which involves more complexities. In the following sections, we list several challenges in developing an OAS from restricted-use microdata.

2.1 Maintaining confidentiality

With the underlying data being restricted information, it is extremely important that proper confidentiality screening be conducted in the OAS. This is the most critical and the most challenging requirement. For this purpose, the Census Bureau is developing a Microdata Analysis System (MAS) (Lucero et al., 2011) with disclosure controls. In addition to the work at the Census Bureau, another system under development by the National Center for Health Statistics (NCHS) is the National Health Interview Survey (NHIS) On-line Analytic Real-time System (OARS) (Gentleman, 2011) using Westat's WesDaX system. Other agencies and organizations have attempted, or are

attempting, to address the disclosure risk elements of a dynamic OAS that uses restricted data. Examples include the University of California Los Angeles Center for Health Policy Research's *Ask*CHIS (UCLA, 2012) with limited suppression and rounding; the Australian Bureau of Statistics (Tam, 2011), which applies a posttabular perturbation approach; and Statistics Canada (Simard, 2011), which applies random rounding.

It should be mentioned that the development of an OAS is not the answer to all data needs, since complete access to microdata would be prohibited; however, it would offer more flexibility to the user. In a flexible OAS, the available analyses in the OAS would be performed in real time from microdata with no prespecified results. The system would focus on providing local area estimates because of the importance of such estimates to transportation planners. Developers of disclosure control methods for an OAS need to realize that users can compare the results of the online system's analyses with the prespecified tables that would be available from the CTPP. The following paragraphs provide a brief overview of the concerns and statistical disclosure control treatments relating to an OAS processing from restricted-use microdata.

SDC treatments in data preparation

The amount of data preparation required would depend on the DRB rules for the system. Some Traffic Analysis District[1] (TAD)-to-TAD flows could have few sample records and cause concern about disclosure risk. Therefore, the alternatives are suppression, perturbation, and/or system-level SDC treatments (discussed below). There are concerns about spatial and nonspatial outliers which could be identified in the microdata through an initial risk analysis (Li and Krenzke (2013), Krenzke et al. (2013a)). The spatial outliers refer to a few data records that are geographically distinct from the majority of the records. The nonspatial outliers are the data records that are unique or almost unique based on a set of key characteristics (e.g., factorial characteristics such as age, gender, race, ethnicity, education, marital status). The identified outliers would be targeted for treatment in the perturbation process. Popular SDC treatments include coarsening, suppression, data swapping, random perturbation, etc. Krenzke, Li, and Zayatz (2013) discuss the model-assisted constrained hot-deck approach that was used to perturb ACS data for the CTPP. This approach provides the flexibility of achieving the balance between reducing disclosure risk and maintaining data utility.

Query restrictions

The system can be set up to allow a predetermined set of queries--for instance, the 2006-2010 CTPP set of defined tables--to further ensure confidentiality protection. This could be achieved by restricting the pool of variables in the underlying data files, limiting the combination of table or subsetting variables, or limiting the number of table dimensions. If a user specifies a query that is not allowed by the system, the system could display a warning message to help the user refine the query specification.

Confidentiality concerns in a real-time OAS

The risk scenario is that information from the OAS can potentially be gathered into the form of a pseudo-microdata record and attach restricted data, such as small geography, to a public use file. The intruder's attack would conceivably include table differencing where, for example, two slightly different universes are queried and both pass the filter rules that require a certain sample size. The difference between the acceptable tables reveals a "sliver," that is, an implicit table with a very small sample size. Then linking those implicit tables can be done with small cell counts to gather characteristics for the sliver, especially by using a marginal count of 1 for a variable repeated for a set of tables. Record linkage (Winkler, 1993) then can be used with those characteristics to match to other files to attach small geography gathered from OAS.

System SDC treatments

Protections and treatments can be applied to reduce the risk from such an attack in a real-time OAS (summarized in Krenzke et al., 2013b). For one, while preparing the data for the system, one can perturb underlying microdata as mentioned above. In the system itself, there can be real-time system approaches, such as threshold or filter rules,

[1] TADs were formed with a population of about 20,000 persons. They were defined to facilitate the CTPP results at a larger geographical area than the TAZ.

where unweighted counts need to be greater than some number at the universe level, for table marginals, and for table cells. Posttabular adjustments or dynamic subsampling adds noise to the estimates to protect against the attack described above. Rounding values, either randomly or deterministically, can also help reduce the risk. System SDC treatments could be implemented as guided by the DRB.

2.2 System development

The Census Bureau has been working through disclosure-related issues relating to an OAS, which has been very helpful to others in their own developments. An initial paper was written by Rowland and Zayatz (2001) to identify query restrictions in the context of an advanced query system for the American FactFinder. Soon after that, focus was given to disclosure risk in regression models, as discussed in Reznek (2003) and Reznek and Riggs (2004). A presentation was given in Steel and Reznek (2005) toward the design of the MAS. Several papers have been written since then, such as Freiman et al. (2011), giving status updates on the MAS. The more recent papers describe a subsampling approach, called the Drop Q rule, which was created to address differencing attacks. Other topics are described, such as a marginal threshold rule, cutpoints for continuous variables, testing residuals for normality, and graphs, including histograms and scatterplots.

Several meetings occurred with various staff at the Census Bureau to gauge the current status of the MAS – as of early 2013. To note, test data for the MAS were switched from the Current Population Survey to the ACS, which is beneficial for our purposes since it uses the ACS. The work has refocused toward generating tables and setting the regression analysis work to the side for the time being, which also is good for our purpose. A beta version was released internally in 2013. Given the meetings, our own experiences, and a review of the literature, in terms of this option being feasible, we provide the following thoughts on meeting transportation planners' needs.

Continued progress

The feasibility of an OAS for CTPP data is highly dependent upon the development of the MAS as it relates to the needs for transportation planners. A large support team is needed, including the transportation planners, MAS developers, DRB members, ACS operations staff, web staff, graphics, security staff, systems staff, senior managers, and statistical programmers. Therefore, it is complex in that it cuts across groups both within and outside of the Census Bureau. There have been some steps toward creating a work group, but more is needed. Among all the groups, the ownership of the system needs to be identified.

System usability and flexibility

The usability and flexibility of the system requires some basic functionality that may include the following:

- Ability to combine geographic areas or combine any category of the available variables
- Ability to choose the type of analysis (weighted frequency, percentage, mean, percentile)
- Ability to subset the population to a desired subgroup
- Ability to conduct statistical tests comparing one estimate with another between subgroups
- Ability to generate graphical reports

Best vehicle

What's the best vehicle for the MAS? The DataFerrett is planned to be the user interface for the MAS. The DataFerrett is a table generator system that currently uses public use microdata for real-time table generation. One critical element to satisfy data users is the response time from the moment the user submits the query to the moment the results are returned to the screen. Although it seems desirable if results are returned within 10 seconds, it would be an interesting study to see when users become anxious, or even frustrated.

Weighted data

As with other issues, much of the literature on the MAS concerned unweighted data. With the existence of weighted data, are current disclosure protections sufficient? In this context, Krenzke et al. (2013b) describes a cell-based subsampling approach which provides protection in extreme situations. Random seeds are used for subsample

selection to attain the same results between users, while addressing extreme table differencing attacks. Relating to weighted and clustered data as in the ACS, are valid variance estimates being produced by the MAS?

2.3 Policies

At this time, the following questions relating to policies are asked.

- Security. In terms of system security, is an OAS that runs off restricted-use microdata acceptable in the DataFerrett environment? If an alternative to the MAS is considered, it likely needs to be packaged within Census Bureau walls near the restricted data. Would Census Bureau security policy allow that? We note that any disclosure protections from an alternative system would need to be approved by the DRB.
- Consistency. Some confidentiality protections, such as Drop Q, may add noise. Is it acceptable if results are not the same as those already published? Other functions of the OAS will need to be evaluated relating to data consistency, which ensures that multiple users can obtain the same result for the same query.

2.4 Sustained level of effort

Depending upon the setup for the OAS, the following items could substantially increase the level of effort under a periodic OAS release. For the sake of this discussion, suppose O_ACS denotes original ACS microdata, used to support the Set A tables. Suppose P_ACS denotes the perturbed ACS microdata, used to support the Set B tables.

SDC treatments applied during data preparation

While existing programs can be processed each time for conducting perturbation, it takes time to review quality control checks and the impact analysis of the perturbations. The setup for the OAS affects whether or not perturbation is needed for the microdata. For example, it may be possible that the DRB would not require perturbation of microdata if a very conservative OAS setup exists in terms of query restrictions and system-level SDC treatments.

Variance estimation

As with SDC treatments during data preparation, the setup for the OAS affects whether or not perturbation is needed during data preparation, which then affects whether or not a special variance estimation approach is necessary. If necessary, variance estimation for the table estimates from the P_ACS would need to appropriately account for the error due to perturbation (if the amount of perturbation is significant). In Li et al. (2011, 2013), a variance estimator was developed for this purpose. A term of squared difference between the ACS and perturbed estimates is added to the original ACS variance as follows:

$$\mathrm{var}\left(\tilde{\theta}_0\right) = \mathrm{var}(\hat{\theta}_0) + \left(\tilde{\theta}_0 - \hat{\theta}_0\right)^2.$$

In the formula $\hat{\theta}_0$ represents the ACS estimate of θ derived from O_ACS and $\tilde{\theta}_0$ represents the perturbed estimate of θ derived from P_ACS. The first term $\mathrm{var}(\hat{\theta}_0)$ is the ACS variance that can be computed using the successive difference replication approach (see Fay and Train, 1995) as

$$\mathrm{var}\left(\tilde{\theta}_0\right) = \frac{4}{80}\sum_{k=1}^{80}(\hat{\theta}_k - \hat{\theta}_0)^2,$$

where $\hat{\theta}_k$ is the ACS estimate of θ for replicate $k = 1,..,80$. The Census Bureau DRB and the Census Bureau ACS Sample Design group approved the use of this formula for variance estimation in the production process of the CTPP Set B tables.

The application of this variance estimation approach is not straightforward in an OAS system. If a user submits a restricted query, the system would have to run this query twice, once from the P_ACS and once from the O_ACS. The ACS table estimates and ACS variances would be obtained from the former, whereas the perturbed table estimates would be obtained from the latter. The functionality needs to be present in the system to combine both

sources of results and then compute the valid variance for perturbed estimates. In a similar manner, the developing OAS for NCHS can process two files to display a result.

Workplace allocation

Imputation of the workplace, known as workplace allocation, was necessary for about 23 percent of records missing workplace geography in the most recent ACS microdata used for CTPP. In Part 2 (workplace) and Part 3 (worker flow), the tables include on average nonmissing block- and TAZ-level values of workplace allocation for about 77 percent of all the worker records. The benefits of the imputation would need to be weighed against the amount of time and effort required to conduct the task.

3. OAS options

Options are introduced in this section for the purpose of laying the groundwork for plausible scenarios that may be useful to transportation planners, and may be helpful to the Census Bureau to focus on a particular application while developing the MAS. To begin, we consider some preliminary confidentiality rules. Tentatively, we assume that the smallest geography allowed would be at the TAD level, although tract level could be considered with the realization that tradeoffs with confidentiality treatments would likely exist. If areas are combined in the system, each area would have to meet the disclosure threshold rules. Flows would probably also include a threshold rule for minimum number of unweighted records.

Each option discussed below adheres to the following general setup. The smallest geography would be the TAD, which is much larger than the smallest geography in the CTPP and larger than a typical tract. Therefore, TAD-to-TAD flows would be offered. The types of analysis and tables would be the same as for CTPP 2006-2010. The microdata would be the 5-year combined ACS with annual updates. Given the preliminary rules, for the sake of discussion, we initially suppose the following:

Smallest geography:	Traffic Analysis District (TAD)
Smallest geography for flows:	TAD to TAD
Types of analysis:	Same as CTPP 2006-2010
Tables allowed:	Same as CTPP 2006-2010
Microdata:	5-year combined ACS
Frequency of updated microdata:	Annual updates or every 3 years

Deviations from the above setup are explicitly mentioned below.

Option 1: Static tables

In Option 1, the system would be driven from underlying CTPP static tables (not real-time – just lookup). The confidentiality protections would be the same as for CTPP 2006-2010.

Option 2: Real-time OAS

In Option 2, the system would be driven from two underlying microdata files: one is the original ACS (O_ACS) which is used to generate the CTPP Set A tables; the other is the perturbed ACS (P_ACS) which is used to generate the CTPP Set B tables. This setup is analogous to the Subsystem P (runs from public use data with no restrictions) and Subsystem R (runs from restricted-use data with restrictions) in the developing NHIS OARS at NCHS. The system would intelligently determine whether the tables should be generated from the underlying O_ACS or P_ACS according to the query specification. It mainly would examine whether a query involves any Set B tables, which would invoke the use of P_ACS. Adding new tables to the system would not cause extra data processing work as long as the DRB approves of the new tables generated from the existing perturbation approach to generate P_ACS. Under this option, the tables are generated from the underlying data files in real time. In other words, the system would conduct the computation in real time and return the output to the users a short time (usually a few seconds) after a query is submitted.

Option 3: More flexibility

Option 3 considers that a predetermined set of variables would be allowed in the system. Consider the set of all variables allowed in any Set A and Set B table. Suppose the allowed queries are as follows:

- Part 1 Residence based. Any 2-way tables within the geographic area (e.g., TAD)
- Part 2 Workplace based. Any 2-way tables within the geographic area
- Part 3 Flows. Any 2-way tables within the flow

Flexibility in terms of variables is also desired. For example, more detailed versions of variables could be offered so that variation among regions can be accommodated. For example, the income cutoff of $75,000 may be too low for some areas. Also, commuting patterns across regions are sometimes not compatible with the rush hour defined in the current CTPP arrival time and departure time categories. This option raises several confidentiality concerns. However, similar concerns have been addressed in the NCHS' OARS development of the disclosure treatment methodology as mentioned above. It is not clear if those same treatments would be satisfactory to the Census Bureau's DRB.

4 Summary and comparison of options

Option 1 would be expensive and have little value added beyond the CTPP. The advantages, if any, are minor, as follows: user familiarity with the set of tables, established processing by the Census Bureau, established DRB rules, less perturbation since TADs would be much larger than TAZs. The challenges and disadvantages include the time and enormous cost to generate tables, some of which are necessary, some of which may not be necessary. Some tables are useful, but others are used by only a few analysts. Other disadvantages include a fixed set of variables and a fixed set of categories for variables. In general, Option 1 consists of very limited additional value to the user community at a high cost.

Option 2 is for an OAS from DRB already-approved CTPP table shells. Query restrictions would be set up to produce only the 2006-2010 CTPP tables on demand – so there would be no pregenerated tables, therefore reducing costs. It would be driven by the "Set A" microdata file and "Set B" perturbed microdata file framework, if perturbation is deemed necessary. The use of the appropriate file would be triggered by the query submitted. Compared to Option 1, this setup saves the time and cost of preprocessing since there is no need to generate tables ahead of time. The value added is more timely and updated data.

Option 3 is more flexible in that it would allow all 2-way tables within geographic areas to be queried, and it would allow more detailed categories for variables. It would also allow threshold rules to determine what is accepted -- that is, what tables get shown and what tables get denied.

Tables 1 and 2 provide an illustrative comparison of the three options if perturbation of microdata in data preparation is required, or not, respectively. The CTPP 2006-2010 is also provided in the tables for comparison. In the tables, the more "+" signs there are, the more is needed, or the more it would exist. The number of "+" signs is conceptual, somewhat arbitrary, and is a rough indicator of relative proportions of the amounts (e.g., five "+" signs do not necessarily equate to 20 percent more than four "+" signs).

Options 1 and 2 are similar, but Option 1 requires all tables to be produced ahead of time. Therefore, Option 1 is more costly than Option 2. As shown in Table 1, relative to CTPP 2006-2010 production, while Option 2 would require less perturbation, cost less, and require fewer meetings with the DRB, it would not provide any added flexibility. As a summary comparison of Options 2 and 3, as shown in Table 1, if perturbation is necessary in both options, the cost would be about the same for the two options but considerably less than the full blown CTPP. As shown in Table 2, if perturbation is not necessary for Option 2, it greatly simplifies the system and substantially reduces the cost and time for Option 2's system development. In terms of DRB involvement, Option 2 would have less involvement. Option 3 would have more DRB involvement due to 1) offering tables that were not previously considered to be queried, 2) offering more detailed categorical variables, and 3) requiring a review of the system protections as well as a risk assessment. The perturbation would be less, if any, under Option 2, while Option 3 would likely need perturbation since prior programs were tailored to a fixed set of tables and this option would include several new tables.

As represented in Table 1, if perturbation is required, the level of complexity in developing the system would be a bit more for Option 2, since it would probably need to deal with handling the Set A and Set B tables. As represented in Table 2, if perturbation is not required, some of the system development tasks are simplified since there is no need for two separate underlying microdata files (perturbed and original). For Option 3, we assume that the necessary threshold rules have already been set in place in the MAS. In terms of value added to usability, Option 2 is inflexible since it would be limited to a prespecified set of tables, while Option 3 would be much more flexible.

Table 1. Illustrative comparison of options – if perturbation of microdata in data preparation is required

		Option		
	CTPP 2006-2010	1	2	3
Perturbation amount	+++++	++	++	++++
Cost (staff time)	+++++	++++	++	++
DRB involvement	+++++	++	++	+++++
System complexity	NA	++++	++++	++
Flexibility	+	+	+	++++

Note: The more "+" signs there are, the more is needed, or the more it would exist. The number of "+" signs is conceptual, somewhat arbitrary, and a rough indicator of relative proportions of the amounts (e.g., five "+" signs is not necessarily 20 percent more than four "+" signs).

Table 2. Illustrative comparison of options – if perturbation of microdata in data preparation is not required

		Option		
	CTPP 2006-2010	1	2	3
Cost (staff time)	+++++	+++	+	+
DRB involvement	+++++	++	++	+++++
System complexity	NA	++	++	++
Flexibility	+	+	+	++++

Note: The more "+" signs there are, the more is needed, or the more it would exist. The number of "'+" signs is conceptual, somewhat arbitrary, and a rough indicator of relative proportions of the amounts (e.g., five "+" signs is not necessarily 20 percent more than four "+" signs).

In summary, transportation planners desire more flexibility from the CTPP, and an OAS may be a solution. An outline of scenarios has been provided and compared for a possible example or potential application of the MAS. However, the feasibility of a flexible OAS could become limited due to 1) confidentiality and security issues, 2) hosting, if an alternative to the MAS is considered, and 3) the fact that, if an alternative to the MAS is not considered, its feasibility is also very dependent upon the progress of the MAS's development. The transportation community is waiting to see what progress the Census Bureau is making on the MAS before doing more work toward it. Data preparation activities also could be a limitation if annual updates are needed. However, we are optimistic about the potential value added through an OAS. Its potential solution makes such a system appear to be useful and could be cost effective after the initial setup, as time goes on.

References

Fay, R., and Train, G. (1995). Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties. *Proceedings of the Joint Statistical Meetings, American Statistical Association, Section on Survey Research Methods*.

Freiman, M., Lucero, J., Singh, L., You, J., DePersio, M., and Zayatz, L. (2011). The Microdata Analysis System at the U.S. Census Bureau. In *Proceedings of the Joint Statistical Meetings of the American Statistical Association, Section on Survey Research Methods.*

Gentleman, J. F. (2011). A real-time online system for analyzing restricted data from the U.S. National Center for Health Statistics' National Health Interview Survey. In *Proceedings of the 58th World Statistics Congress of the International Statistical Institute.* Retrieved from: http://isi2011.congressplanner.eu/pdfs/650208.pdf.

Krenzke, T., Chen, S., Shlomo, N., and Soler, J.M. (2013a). Comparison of perturbation approaches for spatial outliers. Census Statistical Disclosure Control Research Project 4. Final Report. March 20, 2013. U.S. Bureau of the Census, Washington, DC.

Krenzke, T., Gentleman, J., Li, J., and Moriarity, C. (2013b). Addressing disclosure concerns and analysis demands in a real-time online analytic system. *Journal of Official Statistics*, 29(1), 99-134.

Krenzke, T., Li, J., and Zayatz (2013). Balancing use of weights, predictions, and locality effects in a model-assisted constrained hot deck approach for random perturbation. *Proceedings of the Joint Statistical Meetings, American Statistical Association.*

Li, J., and Krenzke, T. (2013). Comparing approaches that are used to identify high risk values in microdata Census Statistical Disclosure Control Research Project 3. Final Report. July 29, 2013. U.S. Bureau of the Census, Washington, DC.

Li, J., Krenzke, T., Brick, M., Judkins, D., and Larsen, M. (2011). Variance estimation for Census transportation planning products with perturbed American Community Survey data. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*.

Li, J., Larsen, M., Krenzke, T., and Zayatz, L. (2013). Alternative variance estimators for data perturbed for confidentiality protection. *Proceedings of the Joint Statistical Meetings, American Statistical Association.*

Lucero, J., Freiman, M, Singh, L., You, J. DePersio, M. and Zayatz, L. (2011). The Microdata Analysis System at the Census Bureau. SORT-Statistics and Operations Research Transactions 35, 77-98. Special issue on privacy in statistical databases. Privacy in Statistical Databases.

NCHRP (2011). Producing transportation data products from the American Community Survey that comply with disclosure rules. Washington, DC: National Cooperative Highway Research Program, Transportation Research Board, National Academy of Sciences. Written by Krenzke, T., Li, J., Freedman, M., Judkins, D., Hubble, D., Roisman, R., and Larsen, M.

Reznek, A. (2003). Disclosure risks in cross-section regression models. In *Proceedings of the Joint Statistical Meetings of the American Statistical Association, Section on Government Statistics*.

Reznek, A., and Riggs, T. (2004). Disclosure risks in regression models: Some further results. In *Proceedings of the Joint Statistical Meetings of the American Statistical Association, Section on Government Statistics*.

Rowland, S., and Zayatz, L. (2001). Automating access with confidentiality protection: The American FactFinder. In *Proceedings of the Section on Government Statistics, American Statistical Association*.

Simard, M. (2011). Real Time remote access at Statistics Canada: Development, challenges, and issues. *Proceedings of the 58th World Statistics Congress of the International Statistical Institute*. Available at: isi2011.congressplanner.eu/pdfs/650096.pdf (accessed February 1, 2013).

Steel, P., and Reznek, A. (2005). Issues in designing a confidentiality preserving model server. *Monographs of Official Statistics*, 9:29.

Tam, S. (2011). On-line access of micro-data at the Australian Bureau of Statistics: Challenges and future directions. In *Proceedings of the 58th World Statistics Congress of the International Statistical Institute*. Retrieved from: isi2011.congressplanner.eu/pdfs/650030.pdf (accessed February 1, 2013).

UCLA (2012). AskCHIS website: http://ask.chis.ucla.edu/main/default.asp (accessed February 1, 2013). University of California, Los Angeles Center for Health Policy Research.

Winkler, W. (1993). Matching and Record Linkage. U.S. Census Bureau.