Confidentiality Issues Related To Transportation Use Of Census Data for Transportation Planning: Preparing for the Future

David Banks and Jerry Reiter
Institute of Statistics and Decision Sciences
Duke University, Durham, NC 27708
{banks,jerry}@stat.duke.edu

1 Introduction

Government agencies spend public money on data collection in part to create information resources that serve the general welfare. Wide and public access to these data facilitates advances in transportation planning and other important areas. However, agencies must weigh the user's need for accurate detail against the legal requirement to protect the confidentiality of survey respondents' identities and attribute values. Agencies that fail to prevent disclosures of individuals' identities or sensitive attributes can be in violation of laws and therefore subject to legal actions; they may lose the trust of the public, so that respondents are less willing to participate in their studies; or, they may end up collecting data of dubious quality, since respondents may not give accurate answers when they believe their privacy is threatened.

For this reason, agencies typically alter data before releasing it to the public. For example, the Census Bureau aggregates and rounds cells in origin-destination matrices (O-D matrices) before releasing them to the public. The price to pay for such disclosure limitation is a reduction the accuracy or utility of the data. Different disclosure limitation strategies carry different reductions in data utility, and generally there is an inverse relationship between risk and utility. A sensible approach is to identify disclosure limitation strategies on the frontier of a risk-utility curve in an R-U confidentiality map (Duncan *et al.*, 2001), formed by quantifying the risks and utilities of candidate strategies. Issues in and strategies for achieving this balance are described by Duncan *et al.* (1993), Federal Committee on Statistical Methodology (1994), Fienberg (1994), and Willenborg and de Waal (2001), to name just a few.

The Office of Management and Budget requires all agencies to protect data confidentiality. And the United States Census Bureau is specifically enjoined to prevent identification of individuals, however innocuous. Title 13, Section 9, part (a) of the United States Code directs that:

> Neither the Secretary, nor any other officer or employee of the Department of Commerce or bureau or agency thereof, or local government census liaison, may, except as provided in section 8 or 16 or chapter 10 of this title or section 210 of the Departments of Commerce, Justice, and State, the Judiciary, and Related

Agencies Appropriations Act, 1998 or section 2(f) of the Census of Agriculture Act of 1997—

(1) use the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or

(2) make any publication whereby the data furnished by any particular establishment or individual under this title can be identified; or

(3) permit anyone other than the sworn officers and employees of the Department or bureau or agency thereof to examine the individual reports.

The language of Title 13 is not precise. There are open questions about the kind of protection that are required, and what is meant by identification.

- Suppose a Bayesian knows that her husband participated in a public health survey about sexual activity, and that the results of that study show that 20% of married men have had an extramarital affair. She updates her prior belief about her husband with the results of the study and concludes (correctly) that he has been unfaithful. Is this a disclosure?

- An agency releases a table of survey data in which a cell contains three people (the U.S. Census requires suppression all cells with fewer than three respondents). One of those three eliminates himself and then has a 50-50 chance of identifying the others. If he guesses correctly, is this a disclosure?

- Suppose a federal survey releases anonymized data in which respondents are not identifiable, but it is possible for someone to use record linkage with a commercial database that leads to identification. Is this a disclosure?

So far, the first and second cases are not considered to be disclosures, although the third case is a disclosure. But these decisions could change if new statistical methods or expanded commercial databases lead to increased probabilities of identification.

A further aspect of this concern is the distinction between identity disclosure and attribute disclosure. Identity disclosure occurs when a specific person's record can be found in the release. Attribute disclosure occurs when a specific value of a sensitive variable can be linked to a specific person, perhaps by use of additional knowledge. Attribute disclosure usually requires identification disclosure, although this is not necessary. For an illustrative if extreme example, one might have a file in which, as a measure of disclosure limitation, personal income is randomly permuted among all the anonymized respondents. If someone knows that the richest man in town is among the respondents, it is possible to figure out how much income that person has, even without identifying that person's record. In this paper, we are concerned primarily with identification disclosures, although we occasionally discuss attribute disclosures as well.

It is natural to wonder about the seriousness of the threats of disclosures. Most agencies keep any disclosures they learn about secret—for obvious reasons—so that it is difficult to point to genuine examples. However, a study by Sweeney (1997) illustrates the risks. Sweeney (1997) showed that 97% of the records in a medical database for Cambridge, MA, could be identified using only birth date and 9-digit ZIP code by linking them to a publicly available voter registration list. Given the ever-increasing resources available to those seeking to achieve disclosures—both in data and record linkage technologies—this example illustrates that it is prudent (and legally necessary) for federal agencies to treat disclosure risks seriously.

The conflict between privacy protection and public value is especially acute in the context of transportation data. These data are needed by metropolitan planning offices for siting bus-stops, building mass transit, and developing long-range zoning plans. And the data are expensive to collect; often respondents keep trip diaries or carry GPS units, and must answer long questionnaires about each member of their household and each trip that is taken. The data also can be exceptionally easy to identify. If one knows the origin block-group and the destination block-group, that often can lead to unique identification, especially when combined with other demographic data. And if one aggregates the block groups to prevent this, then one must coarsen the data to the point at which they could lose much of their value.

2 Privacy Issues in Transportation Data

In order to protect confidentiality in surveys, the Census Bureau screens their releases to minimize the chance of disclosure. If there appears to be a risk, then the Census Bureau uses some kind of statistical strategy to disguise the data. For example, if the publication consists of a table, then the Census Bureau requires that each cell of the table contain at least three respondents. To achieve this, they aggregate categories until the "rule of three" is satisfied. But the application of this rule may not be optimal from the standpoint of users—it could be that one aggregation preserves much useful structure, while another does not. Similar issues arise in other kinds of privacy protection, as discussed in Section 3.

The U.S. Census Bureau runs two major surveys that involve transportation data; the Federal Highway Administration runs the NHTS through contractors:

1. The American Community Survey (ACS) is scheduled to replace the Census long-form in 2010, and is regularly administered to a smaller set of households now. It asks for standard demographic information, and also how many vehicles the household owns, how each adult gets to work, whether people carpool to work, when people leave for work, and how long it takes to get home.

2. The Census Transportation Planning Package (CTPP) is a compilation of data from the 2000 census that summarizes long-form information by place of residence, by place

of work, and for worker-flows between home and work; it is designed for use by MPOs and State DOTs.

3. The 2001 National Household Travel Survey (NHTS) focused on transportation and collected data on daily and long-distance travel. It combined demographic data with information on short and long trips, trip purpose, modes of travel, household vehicles, and a large number of related variables.

Besides these surveys, many communities collect additional information on local travel; some of these surveys are done commercially, but others are run by the Census Bureau, as "add-ons" to the NHTS. It is conceivable that one can use the data obtained in these and other surveys in combination with Public Use MicroSample (PUMS) data, which are edited (by coarsening and cell suppression) to de-identify the records. This would not be easy, even in the case of some individuals with unusual attributes, but the possibility is real.

Transportation data has special features that interact with both privacy concerns and standard methods for conducting surveys. These include:

- *Origin-Destination Balance.* In most applications, such as studies of daily commuting, there is near-perfect equivalence between the travel from a given address and the travel to a given address. Many analytic tools, such as most versions of the gravity model, make strong use of this equivalence. But privacy protection methods that alter the data could either destroy the balance or fabricate wholly spurious origin-destination pairs; both of these avenues create obstacles for use of altered data.

- *Longitudinal Collection.* Often transportation data are collected over time, and this can happen at multiple scales. At one level, a trip diary may cover a month (for commuting) or a year (for air travel). At another level, the same people may be re-interviewed, as part of a panel survey, to discover changes in behavior regarding, say, use of public transportation. At a third level, the same community might be surveyed on an annual basis. The first two cases pose special threats of disclosure beyond the conventional privacy protection scenario of the last case.

- *Small Area Inference.* In transportation one often needs to make estimates about traffic volume for small areas, or flows between pairs of small areas, for which data are extremely sparse. Slight alterations in the data that protect privacy can have large effects on such estimates, especially if they are derived from models that do not borrow strength across areas.

- *Hierarchical Cluster Sampling With Some Fixed Locations.* Simple random samples are impractically expensive; quota sampling and related methods preclude estimates of uncertainty. The standard solution in many surveys is hierarchical cluster sampling, in which the some large geographic units are selected at random, and then within

those smaller units are selected at random (this nested selection may be repeated many times). The result is that one has a sample that is obtained by a random process (allowing uncertainty statements), but which has geographically concentrated subgroups (reducing survey cost). For various practical reasons, it is often necessary to depart from random sampling to ensure that some key geographic units are included. When some geographic units are forced into the sample, this information could be exploited in efforts to breach confidentiality.

In general, raw transportation data tend to be so unique that identification risks are greater than for the kinds of data collected in most other surveys—knowing an O-D pair to within a few blocks can be enough to uniquely identify an individual, and this risk increases further when demographic data are released. Efforts to de-identify the records produce anonymized data that are coarse or altered. There is real concern that relative to traditional survey data, the value of transportation data is more diminished by privacy protection measures.

The remainder of this paper discusses two issues. The first is a review of confidentiality protection methods in the context of transportation data. The second lays out methods for estimating the risk of disclosure. The last section attempts to draw the discussion together in a decision-theoretic framework.

3 Statistical Methods for Privacy Protection

There are five methods which are regularly used or considered when developing statistical privacy protection:

- Recoding variables into coarse categories, through rounding, topcoding, or use of thresholds.
- Swapping some respondent's data values with those of other respondents.
- Suppressing cell entries in tables or values in microdata.
- Perturbing data values.
- Creating synthetic data.

Ideally, data altered in one of these ways will be sufficiently disguised that it is impossible to identify individuals, but still retain sufficient accuracy for legitimate applications.

As a sixth strategy, very different from the five listed above, one could build a query system that tracks the questions that have been asked (Duncan and Mukherjee, 2000). If a question enables identification of an individual record, either by itself or in conjunction with all previous questions ever asked of the system, then the system refuses to answer.

All of the first five methods will affect the covariance structure of the data, and most will affect the expectations. For some analyses—for example those not involving small area estimation—these effects could be sufficiently small that they do not pose a major problem for transportation uses. On the other hand, for other analyses—such as small area estimations—they could be quite large. A one-size-fits-all quantification of data utility therefore may be difficult to obtain. The users in the transportation community are often uncomfortable because they do not know, in the specific cases for which they must make decisions, how much signal is being lost when the data are altered. This problem is exacerbated by the high cost of surveys, the problematic quality of much of the data that are ultimately obtained, and cultural clashes between data gatekeepers and data users.

One approach to measuring the magnitude of this problem is for transportation researchers to perform and compare typical analyses on the original and altered data. Such research may have to be done in restricted access data centers, which are spread thinly across the U.S. and require users to obtain special sworn status. Alternatively, transportation researchers could supply a list of analyses to Census employees, who then could perform, compare, and report the analyses to the transportation research community. Given the logistical difficulties of accessing and using Census RDCs, this could be the most effective strategy for reassuring the transportation research community and helping agencies gauge the risk-utility tradeoffs.

3.1 Rounding and Thresholding

Rounding and thresholding refer to methods that aggregate data. For example, one might disguise financial data by rounding someone's reported income to the nearest $1000. This would work well for large samples in the lower and middle income brackets, but fails at the upper end since it is likely that the few wealthy millionaires in the sample have incomes that are separated by more than $1000. This leads to thresholding, or topcoding, in which everyone with an income greater than, say, $1 million is grouped together.

There are several problems with rounding. One is that different tables produced using rounding can be inconsistent—with substantial effort, one can develop rounding schemes that preserve marginal totals, but this is not trivial. Another problem is that rounding can create bias. If the distribution within an interval is not uniform, then reporting the midpoint (as is frequently done) is misleading. It would be better to report the average, but for intervals that contain few respondents, this reveals a mathematical constraint that might be used for identification. (If an interval contains three people, and if the Census reports the mean and standard deviation for that interval, then any of those three could calculate the exact values for the others.) This bias problem is exacerbated in the case of topcoding, since the distribution will surely be asymmetric and the respondents are probably well-separated.

The CTPP 2000 required counts of respondents that are between 1 and 7 to be rounded to "4" and counts of 8 and above to be rounded to the nearest multiple of 5. Because tables

are rounded independently, it is possible to get different answers to the same question. For example, a total obtained by summing cells in one set of tables may not equal the same total obtained by summing cells from different sets of tables. (This phenomenon has been studied by N. Srinivasan, http://www.fhwa.dot.gov/ctpp/sr0404.htm). Ideally, one would like to construct a rounding system that produces values that are unbiased for the mean or the median, but in the case of CTPP several users have reported systematic rounding effects. There is emerging evidence that as one aggregates flows from Transportation Area Zones (TAZs) to tracts to counties, one loses between 2% and 5% of the trips (personal communication from Ed Christopher).

Another problem with rounding is that it is hard to balance the interests of all of the users in selecting the cut-points. Although it is possible to set up the problem of determining cut-points that maximize some joint utility function, this is hard to solve and generally ignored.

In the context of transportation data on location, the analogue of rounding is block aggregation. For most applications one has to aggregate many blocks together, since (using the Census's rule of three) one must ensure that at least three people in the sample travel from the origin aggregate to the destination aggregate. There also have been suggestions that at least 50 people must be in each cell, which could result in tables with inadequate utility.

If the block aggregates are larger than the distance someone is willing to walk to reach a bus stop, then the rounding prevents use of the survey data for planning bus routes. If the block aggregate is larger than the typical distance between highway exits, then the survey is not useful for road construction. Regrettably, affordable sample sizes can require high levels of aggregation to reduce the risks of re-identifications.

3.2 Data Swapping

Data swapping, proposed by Dalenius and Reiss (1982), is used by several federal agencies. This method exchanges a fraction of data at random between random pairs of respondents. Thus the resulting records cannot be definitively linked to a respondent. In practice, agencies tend to swap a very small fraction of their data, since swapping destroys or reduces correlation structure that is often of primary interest. If the precise fraction of swapped data became widely known, this could undermine the security of the system.

Data swapping is probably not a realistic option for origin-destination data. It is hard to imagine easy and automatic procedures for swapping origins and destinations that do not create inconsistencies or misleading results. For example, swapping respondents into different cells of the O-D matrix would require concomitant changes in travel mode (no respondent would bicycle 50 miles to work, especially if the reported commute time was an hour). If one relaxes the requirement that swapped records be physically possible, then analysts are left with the awkward task of interpreting relationships between fictional variable values, such as how distance to work affects mode choice. It may be possible to develop algorithms

that ensure only realistic swaps get made; whether such swaps can be made while protecting confidentiality for people with eccentric transportation patterns is an area for future research.

More concretely, the Census Bureau uses swapping for decennial census releases. Its application to the longform data has led to such things as commuters using mass transit being reassigned to areas in which mass transit is unavailable or extremely unlikely. The full ramifications of distortions of this type are hard to quantify, and depend upon both the decisions being made from the data and unavailable details about the swapping protocol. In practice, swapping may be beneficial if used on traditional variables such as age, race, and gender, where the relationships are not as strong as in origin-destination data. But it can break down for the kinds of transportation data patterns found in previous surveys.

3.3 Cell Suppression

Cell suppression is used in high-dimensional contingency tables, where some cells contain small numbers of respondents (e.g., a Native-American, Female, Ph.D., Durham resident easily could be unique). The presence of singletons means that privacy is automatically violated if the full table is published. Thus administrators suppress cells or collapse factors to prevent identification.

Cell suppression in sparse tables can have large effects on utility, as complementary cells must be suppressed if marginal totals are to be maintained. If a cell in a two-way table is suppressed, then one must also suppress at least three other cells (one in the same row and one in the same column, and one in the same row and column as these other two), else the marginal totals reveal the suppressed value. There also is no guarantee that cell suppression can be effective. As shown by Fienberg and Slavkovic (2004), sometimes the marginal constraints imply sharp bounds for suppressed entries. Adding holes to the data also impacts the data utility, as it introduces nonignorable missing data, the correct analysis for which poses difficult statistical problems.

For transportation data, cell suppression faces the same kind of difficulties that arise with rounding. There may need to be so much suppression to achieve de-identification that the released data have inadequate utility. For example, with large numbers of suppressions, all one can track are flows between the most common origin-destination block groups by respondents using the most common transportation modes. This could lead researchers to miss trends in transportation choice until after they have become culturally institutionalized.

3.4 Perturbation

Data perturbation generically means altering the values of cells before release. To distinguish perturbation from rounding, we consider perturbation as adding noise to data values. Typically, noise is added to data values that are considered at high risk, such as cells in tables with low counts. Adding noise "blurs" the actual values, which reduces disclosure risks. In

general, higher levels of noise increase confidentiality protection and lower data utility. It is possible that some tables with cells that have few observations and high weights will require enormous perturbation; in that case, use of synthetic data or suppression may be better.

Adding noise is done more commonly to numerical data than to count data. However, it is possible to add noise to counts. The distributions of the noise should be constrained to keep perturbed tables relatively consistent with the original ones. Simply adding random integers to cells could result in totals and subtotals that differ greatly from the original values, especially for high-order tables and small areas.

One promising approach is cyclic perturbation, developed recently by Roehrig and Duncan (http://www.niss.org/dgii/presentations/roehrig dctechday200311.pdf). This approach involves adding and subtracting observations from cells in tables to generate perturbed tables. Roehrig and Duncan describe how users can compute the posterior distribution of the true counts in each cell. One potential extension of the cyclic perturbation approach is to release multiple copies of the tables, in the spirit of multiple imputation, to simplify computations for the user. For example, with multiple perturbed tables, the user could perform analyses on each table and combine the results to obtain approximate posterior inferences. The validity of this multiple imputation extension has not yet been evaluated theoretically. This strategy should be investigated, but it is too early to recommend adoption.

3.5 Synthetic Data

A last method for privacy protection is based on the construction of synthetic data. Of all the available approaches, we believe this has the most promise as a way of releasing small area O-D matrices that provide consistent tables and sufficiently fine detail for transportation purposes. Several synthetic data approaches are described in detail in Section 4.

3.6 Determining Disclosure Risk

As mentioned previously, in this paper we are concerned primarily with identification disclosures. We therefore confine our discussion of disclosure risks to identification disclosures.

Many authors have proposed disclosure risk measures that are some function of the number of population uniques. Different approaches are described by Bethlehem *et al.* (1990), Greenberg and Zayatz (1992), Skinner (1992), Skinner *et al.* (1994), Chen and Keller-McNulty (1998), Fienberg and Makov (1998), Samuels (1998), Pannekoek (1999), and Dale and Elliot (2001).

Although useful, population uniqueness has limitations as a measure of disclosure risk. First, it does not account for the nature of the information possessed by the agent attempting to compromise security. For example, if the agent knows a particular person is in the sample and that person is a sample unique, the intruder can identify that person—this does not depend upon the person being unique in the population, only the sample. Second, the

number of uniques does not provide much information when there exist a large number of sample and population uniques, as happens when the database contains variables that can be treated as continuous or in very sparse tables. Third, using the number of population uniques may not allow the agency to gauge accurately the effect of some statistical disclosure limitation procedures. For example, records perturbed with random noise may contain just as many estimated population uniques as the original sample contains. Lastly, estimating the number of population uniques accurately is difficult to do in studies where the sampling fraction is small, so that the measures could be misleading.

Other authors have proposed that agencies attempt to link released records with raw records, either through direct matching using external databases (Paass, 1988; Blien *et al.*, 1992; Federal Committee on Statistical Methodology, 1994; Yancey *et al.*, 2002) or indirect matching using the existing database (Spruill, 1982; Duncan and Lambert, 1986, 1989; Lambert, 1993; Fienberg *et al.*, 1997; Skinner and Elliot, 2002). In both approaches, the agency essentially mimics the behavior of an agent trying to breach confidentiality. These approaches permit agencies to account for varying degrees of agent knowledge, are equally appropriate for continuous and categorical data, and can be applied to assess the effects of statistical disclosure limitation techniques. This kind of "red-teaming" probably provides the best protection against compromise—this is an adversarial situation, and the agency should try to think like their opponent.

3.6.1 Risks for Transportation Data

Transportation surveys sometimes use rotating panels, in which respondents participate for a fixed period of time (perhaps even just once) and then are replaced. It is tempting to conclude that as the length of time spanned in the survey increases, or as the sample size increases, the risk of disclosure should diminish. In general and in practice, this may not be true, as we discuss below with illustrative examples.

For the case of multiple waves of data on some of the same individuals, their historical origin-destination values could be unique, even if in some particular months the values are not. Releasing more than the current month for those individuals increases their risks of identification disclosure when similar historical information is possessed by agents trying to breach confidentiality.

When only one wave of data on each individual is released, a key issue is the scope of what constitutes a disclosure. For example, it arguably remains an identity disclosure when a record in previous waves of the survey is linked to some current individual, even if the attributes for that record are out-of-date; indeed, the Census Bureau waits decades before releasing census files in part for this reason. The situation is more muddied for attribute disclosures: is learning the value of an out-of-date attribute a disclosure? This question, and the question of what constitutes an identity disclosure in panel data, are best answered by lawyers and the federal agencies.

Increasing the sample size in a survey does not necessarily improve individuals' protection. Counterintuitively, one can show that for a high-dimensional contingency tables, larger samples generally do not decrease the number of sample uniques. To see this, suppose one has the most favorable case, in which all cells are equally likely, and that n people are distributed uniformly into the r cells. Then the expected number of uniques is:

$$\mathbb{E}[\text{ singleton cells }] = r - \mathbb{E}[\text{ cells with 2 or more }] - \mathbb{E}[\text{ empty cells }].$$

For $n << r$, Barbour *et al.* (1992) show that asymptotically,

$$\mathbb{E}[\text{ cells with 2 or more }] \approx r[1 - (1 - \frac{1}{r})^n - \frac{n}{r}(1 - \frac{1}{r})^{n-1}]$$

and from Aldous (1989),

$$\mathbb{E}[\text{ empty cells }] \approx r \exp(-\frac{n}{r}).$$

Thus the expected number of cells with a single entry is approximately $n - \frac{n(n-1)}{r}$. Because of the combinatorial explosion in high-dimensional tables when many categories are considered, this means that the number of singleton cells scales with the sample size. The larger the sample, the more singletons and thus the less security.

For longitudinal transportation data, we recommend that agencies compute probabilities of identification under a range of different assumptions about agent knowledge and behavior. This is in keeping with the recommendations of the Federal Committee on Statistical Methodology (1994) that agencies should conduct "re-identification experiments." It is conservative to assume the agent has a file consisting of data for each sampled person, including such variables as O-D values, race, sex, and others that are readily available to users. A measure of risk is then obtained by trying to match the (possibly altered) released data to the raw data using record linkage techniques as described in Section 3.6.2.

In the context of the American Community Survey (ACS), we note that it is not a rotating panel survey, but it does cover a substantial fraction of time and thus inherits the concerns mentioned in this section. The longer the period of time covered and the more people included, the greater the opportunity for re-identifications. Even if the data for small areas are only released after long periods of aggregation (say five years), it still seems likely that fine detail on O-D pairs, perhaps in conjunction with additional demographic information, could produce unique records and hence enable identification disclosures, especially since most people do not move in such time spans (e.g., just the current O of the the O-D pair could be sufficient to identify an individual when combined with demographics). If identification in any time period is a disclosure, even people who move during the aggregation period face disclosure risks, assuming that historical information about those people is available to agents trying to breach confidentiality.

3.6.2 Record Linkage

The effort to match data planned for release with the original respondent data is an exercise in record linkage. Most current methodologies (e.g., see Yancey *et al.*, 2002) use probabilistic techniques rooted in the Fellegi-Sunter model (Fellegi and Sunter, 1969). Here one has two lists of records, A and B, and tries to determine a set M of true matches and a set F of false matches. To do this, one assumes a statistical model for the kinds of agreements and disagreements between them (e.g., values of continuous variables are perturbed by random normal noise). Then one declares a record $a_i \in A$ to match $b_j \in B$ if the ratio

$$R = \frac{\Pr[a_i = b_j \mid M]}{\Pr[a_i = b_j \mid F]}$$

is large, where the probabilities are calculated from the agent's model of data deidentification and relevant prior information. Key components of the record linkage model include the assumptions about the characteristics of the population in the small area, the privacy-protection steps used before release, and the kind of knowledge that the de-identifier can bring to bear.

Given the results of a Fellegi-Sunter analysis, the agency must make some hard decisions. Some of the matches will be wrong; some will be correct. Must the agency disguise the data so thoroughly that no probabilistic guess can be correct? Is an 80% chance of a match too high? And if the agency determines that the data must be disguised to the extent that the most probable match on the list is never correct, then the well-informed criminal mastermind will confidently pick the second-most probable match on the list and be correct most of the time.

We believe that disclosure risk should be weighed against the data utility in R-U confidentiality maps (Duncan *et al.*, 2001). One strategy for gauging utility is to run typical application analyses on the proposed data release to see whether the results from the released data agree closely with those from the original data. This enables one to frame the problem of disclosure-limitation in the context of decision theory, as first suggested by Duncan and Lambert (1986) and Lambert (1993). We shall return to this perspective in the conclusion of this paper.

4 The Potential of Synthetic Data

Transportation analysts need fine-scale origin-destination data. Associated with those, researchers want access to person-level variables like age, race, sex, and income. Releasing exact survey values of these variables is a disclosure risk, even if the geographical variables are aggregated to rather high levels. This is because there may not be very many people with particular characteristics in the cell defined by an origin and destination, even when these cells determine broad geographies.

Standard techniques of data alteration decrease the utility of the data. When releasing small area data, the amount of data alteration required to protect confidentiality can be so large as to compromise severely analyses based on the released data. Standard alteration techniques also complicate matters for users. For example, to analyze data that include additive random noise, users should apply non-trivial measurement error models (Fuller, 1993).

Rubin (1993) proposed an alternative approach: release multiply-imputed, synthetic data sets. In this approach, the agency

(i) draw a simple random sample from the sampling frame;

(ii) for each person in the sample, impute fake data values using statistical models fit with the original survey data; and

(iii) releases multiple versions of these datasets to the public.

These are called *fully synthetic* data sets. Releasing fully synthetic data preserves confidentiality since identification of individual data is generally impossible when the released data are not actual, collected values. And, with appropriate imputation and estimation methods developed by Raghunathan *et al.* (2003) and Reiter (2005b), the approach allows data users to make valid inferences for a variety of estimands with standard, complete-data statistical methods and software. Other attractive features of fully synthetic data are described by Rubin (1993), Little (1993), Fienberg *et al.* (1998), Raghunathan *et al.* (2003), and Reiter (2002, 2005a).

Little (1993) proposed a variant of the synthetic data approach: release datasets comprising the units originally surveyed with some collected values, such as sensitive values at high risk of disclosure or values of key identifiers, replaced with multiple imputations. These are called *partially synthetic* datasets. This has been implemented by the Federal Reserve Board (Kennickell, 1997) and by the Census Bureau (Abowd and Woodcock, 2001). Partially synthetic approaches promise to maintain the primary benefits of fully synthetic data–protecting confidentiality while allowing users to make inferences without learning complicated statistical methods or software–with decreased sensitivity to the specification of imputation models (Reiter, 2003). Valid inferences from partially synthetic datasets can be obtained using the methods developed by Reiter (2003, 2004, 2005b).

When fully synthetic data are released, re-identification risk is small. Almost all of the released, synthetic units are not in the original sample, having been randomly selected from the sampling frame (not the original sample). Their values of survey data are simulated, so that no genuine sensitive values are disclosed for these units. Furthermore, the synthetic records cannot be matched meaningfully to records in other data sets, such as administrative records, because the values of released survey variables are simulated rather than actual and, therefore, not identical to those in administrative databases. When compared to fully

synthetic data, the risks of disclosure for partially synthetic data are greater, since the original units and some of their actual values are purposefully released and available for matching to administrative records.

A primary advantage of the fully and partially synthetic approaches is that, when data are simulated from posterior predictive distributions that reflect the distributions of the observed data, frequency-valid inferences can be obtained from the multiple synthetic data sets. Microdata can be released for small areas, and users do not have to deal with the effects of rounding, thresholding, or other alterations. Of course, the validity of inferences from they synthetic data depend on the model used to generate the data. When that model gives implausible imputations, users will not obtain valid inferences.

4.1 Implementing Synthetic Data For Transportation Data

To implement either synthetic data approach, the agency first needs to decide what variables to release. Will the release will include only tabular data and not microdata? Or will data include characteristics like race, age, sex, and income? For the rest of the paper, we assume that analysts want all variables to be released, as this is the more general case. We therefore propose to build synthetic data at the individual level. Such data can be aggregated to form synthetic tables that will have consistent totals across different analyses.

Synthetic data use the original sample to generate data whose statistical properties closely match those of the original sample. To illustrate how this might work in practice for fully synthetic data, let us suppose an agency has collected data on a random sample of 10,000 people in a city. The data comprise each person's race, sex, income, and origin-destination data. We assume the agency has a list containing all people in the city, including their race and sex. This list could be the one used when selecting the random sample of 10,000, or it could be manufactured from census tabulations of the race-sex joint distribution. We assume the agency knows the income and O-D matrix only for the people who respond to the survey.

In the first step to generating synthetic data, the agency randomly samples some number of people, say 20,000, from the population list. In the second step, the agency uses the collected data to estimate the joint distribution of income and origin-destination for each race-sex combination. The agency then generates values of income and origin-destinations for the 20,000 synthetic people by randomly simulating values from these joint distributions. The result is one synthetic data set. The agency repeats the process say ten times, each time using different random samples of 20,000 people, to generate ten synthetic data sets. These ten data sets are then released to the public. If race and sex are not known for the entire population, there is no first step to the algorithm: all four variables are simulated from their joint posterior distribution.

For fully synthetic data, a sequential conditional modeling approach can be used to generate the data. In the example above, first a random sample of people is selected from the sampling frame, with block locations intact. If the races and sexes of these people are known,

they can be released on the file. If these variables are not known, they can be simulated from the posterior distribution of race and sex in each block. Second, synthetic values of the origin-destination values are sampled for each person from the posterior predictive distribution of the origin-destination matrix, conditional on the block location and any other known information. It may be helpful to use informative prior distributions when computing the posterior distribution of the transition probabilities of the O-D matrix. For example, data from previous years can feed into the prior distributions. Sampling from large tables is a computationally tricky process. However, techniques based on abstract algebra (e.g., see Slavkovic, 2004; Chen *et al.*, 2004) have been developed for this sampling. Models that borrow strength across similar blocks also are likely to be useful for O-D simulation.

Once the origin-destination data are generated for each synthetic person, data for the remaining variables are generated from posterior predictive distributions computed from the data. That is, the agency fits the distribution of income–conditional on the cell membership in the origin destination matrix and other characteristics–and draws new incomes using those distributions. Computing these distributions will be challenging: the data are likely to be sparse. It likely will be necessary to borrow strength across regions and time periods using hierarchical Bayesian modeling.

For partially synthetic data, the challenge is similar but slightly easier. In partially synthetic data, the original units remain on the released file, albeit with some data replaced with imputed values. There are many possible approaches here. One approach is to hold constant individuals' microdata and replace their origin-destination values with draws from the conditional posterior distribution of the O-D matrix. Another approach is to generate replacement values for both the O-D variables and for selected values of key identifiers and sensitive attributes. All these approaches have different risk-utility profiles, which can help generate a risk-utility frontier for selecting a strategy.

The primary challenge to implementing the synthetic data approach is determining the posterior distributions for generating the data. This is a major research effort that likely requires substantial resources, both in people and dollars, in the statistical research community. It will involve computationally intensive approaches, including possibly hierarchical models, spatial statistics, and nonparametric regressions. The researchers will need access to the unaltered data, which may require them to work in a Census Research Data Center.

One approach to beginning this research is to form a team of statisticians familiar with transportation modeling and synthetic data. This team would build synthetic datasets, which they would then install in the Research Data Centers. Users of the data would try their models on the synthetic and genuine data. The knowledge gained from comparisons of results then would be used to improve the synthetic data generation process.

4.2 Smoothed Parametric Bootstrapping

Besides drawing from posterior distributions for constructing synthetic data, it is worth mentioning a bootstrap-based approach (Efron, 1979). The three advantages of this strategy are that

- the bootstrap is simple to implement,
- it is possible to inject relevant population information about trip flows and demographic characteristics,
- the analysis is exactly the same as one would perform with the original sample—thus users can do appropriate analyses without the expertise required for the other kinds of privacy-protected data.

The disadvantage is that this approach has not been widely used, and that if too little smoothing is done then it is conceivable that disclosure could occur.

To motivate this, suppose one had done a survey of heights, measuring these to an infinite number of decimal places. If any of the actual heights were reported, this would lead to re-identification. But if the analyst believed that heights were normally distributed, then it would be simple to take a new random sample from a normal distribution with the same mean and variance as the survey sample. This new sample would be almost as good as the original survey data for all practical purposes. The only drawback is that the new sample could not be used to test the assumption of normally-distributed heights.

In the context of multivariate survey data, suppose one formed the empirical multivariate distribution $\hat{F}_n$ of the sample data. This puts mass $1/n$ on each of the observed data points:

$$\hat{F}_n(\boldsymbol{x}) = \sum_{i=1}^{n} \frac{1}{n} I_{\boldsymbol{X}_i}(\boldsymbol{x})$$

where $I_{\boldsymbol{X}_i}(\boldsymbol{x})$ is an indicator function that takes the value 1 if each component of $\boldsymbol{x}$ is less than or equal to the corresponding component of the ith observation $\boldsymbol{X}_i$, and is zero otherwise.

Then one smooths $\hat{F}_n$ a bit by convolving the empirical distribution with, say, a normal distribution having small variance. This gives the smoothed empirical distribution $\tilde{F}_n$, which has a very spiky density function with large bumps at the data points and small but non-zero values elsewhere. But $\tilde{F}_n$ will be very close to the original distribution, whose values cannot be revealed without re-identification.

Then the agency can draw a sample of size n (or larger) from the smoothed empirical distribution and release it to the public. It will be a close proxy for the original sample, but will be hard to re-identify if a reasonable amount of smoothing has been done.

In the context of transportation surveys, this approach has some attractive features. First, it is widely known that an affordable survey overlooks small flows between rare origin-destination pairs. To handle this, the smoothing could have two steps: the first puts a small amount of mass on all cells in the O-D matrix, perhaps proportional to the population, and the second step convolves this with a small-variance normal distribution. Also, the agency could smooth the non-location data so that, say, age/gender/race proportions lined up better with known demographic breakdowns from other survey collections. Finally, the smoothed bootstrap resample can be compared through goodness-of-fit tests with the original sample to check whether there are large discrepancies that suggest problems with the smoothing procedure.

The main danger in using the smoothed empirical bootstrap is that once the new sample has been drawn and released to the analysts, they must trust the agency to have used a reasonable model. The procedure will overlook or downweight data that do not conform to the kinds of patterns implicit in the model for smoothing. But this may be an acceptable balance for the ability to work high-quality data with fine levels of geographic detail.

This strategy can be viewed as a principled kind of perturbation, or a simplified version of synthetic data. To implement this in a practical setting would require the agency to have some domain knowledge and familiarity with the gravity model and related tools.

5 Conclusion

Traditional methods for privacy protection are generally inadequate for transportation analyses. The two disclosure limitation procedures that we recommend for further investigation are cyclic perturbation and synthetic data, and the latter includes smoothed parametric bootstrapping. Of these, we are especially enthusiastic about the potential of synthetic data, since this potentially provide analysts with microdata that can be examined using standard statistical techniques.

A non-statistical solution is to have one or more transportation analysts employed at the Census Bureau, empowered to work directly with the raw data. Then metropolitan planning offices and other legitimate users can address their questions to the Census expert, who produces the analyses they need. Given the costs of large surveys and the advantages to the Census Bureau of having relevant expertise on staff, this seems affordable. Alternatively, each MPO can get one of their own employees cleared, sworn, and authorized to access the data as needed, but this seems duplicative and there can be long delays as the Census Bureau processes background checks.

A decision-theoretic solution is to weigh the expected costs of disclosure against the value of the data, as has been suggested, for example, by Lambert (1993) and Trottini and Fienberg (2002). For transportation applications, it may be that few if any agents would want to spend the effort needed to crack lightly disguised data, and the damages caused

by such a breach of confidentiality could be negligible. If that supposition is supported by reasonable cost-benefit calculations, it would be in the public interest to make the survey available after only minor and inexpensive modifications. In fact, one could make a larger case that it is a waste of public monies to spend more to protect confidential information than it would cost to hire a private detective to uncover it. But, at least for now, this kind of approach is obviated by the legal requirements that Congress has imposed.

Obviously, it is important for the government to protect the confidentiality of citizens who participate in surveys. People have a right to privacy, and our federal agencies have properly invested significant resources to protect that. The intelligent integration of multiple protection procedures, while maintaining maximum usability and supporting inter-agency cooperation to the extent allowed by law is a triumph of dedicated public servants.

But it is not clear that the right to privacy should be absolute—it may be better to compare this to property rights in the context of eminent domain. Some kinds of information are so innocuous that no sensible person would care to protect them, and it is unreasonable to expect that legitimate public interests should be held hostage to paranoia. Some kinds of information are legitimately private but so publicly available that it is unreasonable for the government to invest resources to protect data available from Google or a telephone book. And some kinds of information are legitimately private and not widely available, but perhaps those protections need not be as perfect as the law currently requires. Few areas of the law permit prior restraint of conduct; rather, civil suits allow injured parties to recover damages. Perhaps it would make sense to generously compensate victims if, despite reasonable efforts to protect privacy, some person with massive computing resources, advanced statistical training, access to ancillary databases such as ChoicePoint, and an obsessive wish to embarrass someone were able to devise a way to hack the protections.

References

Abowd, J. M. and Woodcock, S. D. (2001). Disclosure limitation in longitudinal linked data. In P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes, eds., *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, 215–277. Amsterdam: North-Holland.

Aldous, D. (1989). *Probability Approximations Via the Poisson Clumping Heuristic*. Springer-Verlag, New York.

Barbour, A. D., Holst, L., and Janson, S. (1992). *Poisson Approximation*. Clarendon Press, Oxford.

Bethlehem, J. G., Keller, W. J., and Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association* **85**, 38–45.

Blien, U., Wirth, H., and Muller, M. (1992). Disclosure risk for microdata stemming from official statistics. *Statistica Neerlandica* **46**, 69–82.

Chen, G. and Keller-McNulty, S. (1998). Estimation of identification disclosure risk in microdata. *Journal of Official Statistics* **14**, 79–95.

Chen, Y., Dinwoodie, I. H., and Sullivant, S. (2004). Sequential importance sampling for multi-way tables: Discussion paper 04-20. Tech. rep., Institute of Statistics and Decision Sciences, Duke University.

Dale, A. and Elliot, M. (2001). Proposals for 2001 samples of anonymized records: an assessment of disclosure risk. *Journal of the Royal Statistical Society, Series A* **164**, 427–447.

Dalenius, T. and Reiss, S. P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference* **6**, 73–85.

Duncan, G. T., de Wolf, V. A., Jabine, T. B., and Straf, M. L. (1993). Report of the panel on confidentiality and data access. *Journal of Official Statistics* **9**, 271–274.

Duncan, G. T., Keller-McNulty, S. A., and Stokes, S. L. (2001). Disclosure risk vs. data utility: The R-U confidentiality map. Tech. rep., U.S. National Institute of Statistical Sciences.

Duncan, G. T. and Lambert, D. (1986). Disclosure-limited data dissemination. *Journal of the American Statistical Association* **81**, 10–28.

Duncan, G. T. and Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Business and Economic Statistics* **7**, 207–217.

Duncan, G. T. and Mukherjee, S. (2000). Optimal disclosure limitation strategy in statistical databases: Deterring tracker attacks through additive noise. *Journal of the American Statistical Association* **95**, 720–729.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics* **7**, 1–26.

Federal Committee on Statistical Methodology (1994). Statistical policy working paper 22: Report on statistical disclosure limitation methodology.

Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association* **64**, 1183–1210.

Fienberg, S. E. (1994). Conflicts between the need for access to statistical information and demands for confidentiality. *Journal of Official Statistics* **9**, 115–132.

Fienberg, S. E. and Makov, U. E. (1998). Confidentiality, uniqueness, and disclosure limitation for categorical data. *Journal of Official Statistics* **14**, 361–372.

Fienberg, S. E., Makov, U. E., and Sanil, A. P. (1997). A Bayesian approach to data disclosure: Optimal intruder behavior for continuous data. *Journal of Official Statistics* **13**, 75–89.

Fienberg, S. E., Makov, U. E., and Steele, R. J. (1998). Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics* **14**, 485–502.

Fienberg, S. E. and Slavkovic, A. B. (2004). Making the release of confidential data from multi-way tables count. *Chance* **17**, 3, 5–10.

Fuller, W. A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics* **9**, 383–406.

Greenberg, B. V. and Zayatz, L. V. (1992). Strategies for measuring risk in public use microdata files. *Statistica Neerlandica* **46**, 33–48.

Kennickell, A. B. (1997). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. In W. Alvey and B. Jamerson, eds., *Record Linkage Techniques, 1997*, 248–267. Washington, D.C.: National Academy Press.

Lambert, D. (1993). Measures of disclosure risk and harm. *Journal of Official Statistics* **9**, 313–331.

Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics* **9**, 407–426.

Paass, G. (1988). Disclosure risk and disclosure avoidance for microdata. *Journal of Business and Economic Statistics* **6**, 487–500.

Pannekoek, J. (1999). Statistical methods for some simple disclosure limitation rules. *Statistica Neerlandica* **53**, 55–67.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics* **19**, 1–16.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics* **18**, 531–544.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology* 181–189.

Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology* **30**, 235–242.

Reiter, J. P. (2005a). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A* **168**, 185–205.

Reiter, J. P. (2005b). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference* **131**, 365–377.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics* **9**, 462–468.

Samuels, S. M. (1998). A Bayesian species-sampling-inspired approach to the uniques problem in microdata. *Journal of Official Statistics* **14**, 373–384.

Skinner, C., Marsh, C., Openshaw, S., and Wymer, C. (1994). Disclosure control for census microdata. *Journal of Official Statistics* **10**, 31–51.

Skinner, C. J. (1992). On identification disclosure and prediction disclosure for microdata. *Statistica Neerlandica* **46**, 21–32.

Skinner, C. J. and Elliot, M. J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society, Series B* **64**, 855–867.

Slavkovic, A. B. (2004). *Statistical Disclosure Limitation Beyond the Margins*. Ph.D. thesis, Carnegie Mellon University, Dept. of Statistics.

Spruill, N. L. (1982). Measures of confidentiality. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 260–265.

Sweeney, L. (1997). Computational disclosure control for medical microdata: the Datafly system. In *Proceedings of an International Workshop and Exposition*, 442–453.

Trottini, M. and Fienberg, S. E. (2002). Modelling user uncertainty for disclosure risk and data utility. *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems* **10**, 511–527.

Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer-Verlag.

Yancey, W. E., Winkler, W. E., and Creecy, R. H. (2002). Disclosure risk assessment in perturbative microdata protection. In J. Domingo-Ferrer, ed., *Inference Control in Statistical Databases*, 135–152. Berlin: Springer-Verlag.