Combined-year State-level and Single-year Nation-level Public Use Files from the National Household Survey on Drug Use and Health Data

A.C. Singh[1], D. Wright[2], and F. Yu[1]
[1]RTI International, NC, and [2]SAMHSA/HHS, DC

Abstract

Since 1999, the Substance Abuse and Mental Health Services Administration (SAMHSA) has provided yearly National PUFs for NSDUH data using a procedure based on the MASSC (Micro Agglomeration with Substitution, Subsampling, and Calibration) system for statistical disclosure limitation. There is a growing demand for state-level data, and SAMHSA is considering providing state-level PUFs based on combining several years of NSDUH data. In this paper, we explore various concerns and approaches to State-level PUFs and indicate how MASSC could address some of them. Releasing combined-year state-level PUFs alongside single-year national PUFs poses several challenges. The most important one is that confidentiality of an individual could be compromised if an intruder is able to match the state-level PUFs with the national PUFs on the basis of various sensitive/outcome variables which are typically not perturbed, and thus may succeed in attaching state identifiers to the national PUFs. This problem can be reduced by taking advantage of the randomness in perturbation and suppression used in MASSC.

Key Words: MASSC method; random substitution; random subsampling

1. Introduction

In 1999, SAMHSA introduced a new design for the National Household Survey on Drug Abuse and Health (NSDUH, previously known as NHSDA) that included samples for the 50 States and the District of Columbia that were representative of those areas. The eight largest states were allocated a total sample size of 3600 each while each of the remaining 42 States and DC were assigned annual sample sizes of approximately 900, divided equally among three age groups, 12-17, 18-25, and 26 or older. The total national sample size has varied slightly from 1999 until the present, but has typically been approximately 68,000. The samples for the 8 largest states have generally been 3000 or more, with the larger samples designated to the larger population states. Each state is divided into either 12 (for 'small' states) or 48 (for 'large' states) field interview regions that cover the state. Within these FI regions, block groups are selected first, then households, and then individuals within households. The total sample is allocated equally across the 4 quarters. Survey responses are collected using audio computer-assisted self-interviews in which the respondent completes the answers on a laptop computer.

The NSDUH data are used for a variety of purposes such as tracking use of various drugs and other correlates of drug use. Given the devolvement of responsibilities from the federal to the state and local levels, there has been an increasing need for State and substate level information in order to target programs in an effective manner. NSDUH data, like any other database containing personal confidential information, requires protection from disclosure before being released as a public use file (PUF) for researchers at large. This is, in fact, obligatory in view of the confidentiality pledge given to the respondent before conducting the interview. Also, in view of regulations such as HIPAA (Health Insurance Portability and Accountability Act of 1996 for protected health information) and CIPSEA (Confidential Information Protection and Statistical Efficiency Act of 2002 for various kinds of government data), data producers are increasingly more concerned about protecting confidentiality.

Following the approach of other federal agencies, the SAMHSA has produced public use files (PUFs) in order to provide researchers and policy officials the information they need. These PUFs and those for a variety of other federal and non-federal surveys that address substance use, mental health, and related areas are available at the Substance abuse and Mental Health Data Archive (SAMHDA) maintained by ICPSR at the University of Michigan. Also, at that website is a Data Analysis System that can be used to analyze data from any of data sets in the archive.

While state-level data have been in significant demand, the small annual sample sizes for most states often lead to estimates with large standard errors. For this reason, SAMHSA developed a small area estimation modeling procedure that provides estimates for (now) a total of 20 measures related to substance use or mental health. Even for those measures, it is preferable to model two consecutive years' data simultaneously in order to better reflect specific state epidemiology relative to the national model. Recently, NSDUH data have been combined across 3 years in an attempt to produce more stable design-based estimates for measures that go beyond those estimated by small area estimation techniques.

Therefore, there is a growing need to provide the user community with greater access to state-level and other data for which the national PUF is not deemed adequate. One option being pursued is a licensing procedure whereby researchers who sign a pledge of confidentiality and agree to use the data only for statistical purposes can obtain a license

to use such data for a specific purpose as long as they agree to protect the confidentiality of individual records. Another option being pursued is to develop state-level PUFs by combining files across a number of years.

2. Intrusion Scenarios and Review of MASSC for NSDUH National PUF

It is known that an individual's confidential information collected in NSDUH could be at risk of disclosure. The direct identifiers such as date of birth, address, and telephone number do not pose any problems because the details can be suppressed, e.g., only broad age categories and geographic information are generally of analytic interest. An important disclosure scenario known as 'disclosure by response knowledge' (see Bethlehem et al., 1990) is of great concern to the subjects in the database. Here, the subject finds his own record in the released but untreated database. (This is fairly easy for the subject as the combination of values of various demographic and outcome variables corresponding to the subject can invariably make him unique in the database.) Then the subject is concerned about protecting his confidential information because someone close to him such as a family member, friend or coworker who knows about his presence in the database could possibly find values of sensitive/secret variables (SVs) such as past month cigarette, alcohol, marijuana, and cocaine use for his record by using a combination of values of several indirect identifying/intrusion variables (IVs) such as age, gender, race, and marital status. In the following, such an intruder will be termed as an inside intruder as opposed to an outside intruder who doesn't know the presence of his target in the database.

There are at least two ways the inside intruder might find values of SVs: first, the combination of IVs make the target unique in the database, and second, the combination of IVs may make the target nonunique but all records in the nonunique group may have common values for at least one SV. The above scenario of inside intrusion provides a practical way of finding out what records could possibly be at risk and then find ways of introducing uncertainty about their presence and identity in the database. The fact that some records could be at risk is of great concern to the data producer because it is the respondent's trust that the data producer wants. Therefore, it is not possible for the producer to release the original untreated NSDUH data to users for analysis purposes unless some form of a licensing agreement is in place for the user to have access to the original database. This option of a signed licensing agreement may be viable for some users and the details are currently being worked out by SAMHSA. However, there is a need of PUF for general users.

The problem of protecting confidentiality for NSDUH data turns out to be more challenging than the usual databases. Due to the nature of the sampling design of NSDUH, there are some other scenarios that may jeopardize the confidentiality of an individual's information. For instance, when a parent/child pair is selected from a household, the parent acting as an inside intruder could find his record first and find his child's record by using the household identifier. This implies that the household link cannot be included in the PUF. Now even if this link is deleted, since members of the same household belong to the same PSU (a primary sampling unit or a cluster of households whose identifier is needed for proper variance calculation even though the identity is scrambled in the interest of confidentiality), the parent only needs to search within a small subset of records corresponding to his own PSU to locate his child's record, and thus rendering his child's record more at risk. For this reason, the PSU indicators would also have to be treated to introduce uncertainty. Thus, for NSDUH, more protection is needed for individuals selected in pairs from households than those in singles.

In general it is harder to protect an individual's confidential information when the individual is known to belong to a subset of the database, e.g., belonging to a particular state than the situation when the state identifier is removed. For NSDUH, this would essentially imply creating 51 PUFs (for 50 states and DC) and would be a time-consuming task requiring further treatment of the individual states in order to achieve similar levels of disclosure protection. These considerations led to the decision to produce only a nation-level PUF from each year of NSDUH data, and to remove information about census region, state, and other geographic identifiers. However, due to state and substate-level demands for drug use, we need to find alternative ways to create state-level PUFs that take advantage of the treatment performed for the existing national PUFs. This is the purpose of this paper.

For disclosure treatment of any micro-data such as the NSDUH, most of the available methods consist of either some form of perturbation of IV values (such as swapping, recoding or adding noise,) or suppression of IV or SV values (such as deleting a field or the whole record) or both. Now any form of disclosure treatment introduces some loss of information. Under the inside intrusion test mentioned above, which provides a reasonable way of finding records at risk, typically there could be a large number of records at risk depending on the intruder's knowledge of IVs. Treatment of records found to be at risk via perturbation or suppression may introduce high information loss in that unacceptably large bias could be present in the analysis of resulting data. Therefore, a method that simultaneously protects both confidentiality and analytical quality of data is desirable such that a suitable balance between disclosure risk and information loss can be achieved.

RTI's MASSC method offers such an option and a version of it was used for NSDUH data for the years 1999-2002. Alternative methods in the class of nonsynthetic disclosure treatment methods use a deterministic selection of records for treatment, i.e., all the records identified to be at risk with respect to a set of IVs are treated. However, MASSC uses a stochastic selection of records for treatment in which all records are subject to treatment but only a small random subset is actually treated; this leads to low information loss and protection against new IVs that an intruder might know. Thus, under a probabilistic framework, MASSC introduces sufficient uncertainty about the presence and identity of a record, and provides measures of disclosure risk and information loss without any modeling assumptions.

The MASSC acronym signifies the four steps of Micro Agglomeration, Substitution, Subsampling, and Calibration. Its foundation rests on the theory of survey sampling as it uses a subtle analogy between releasing an untreated database and conducting a census. It is designed to minimize disclosure cost while controlling loss of information. **M**icro **A**gglomeration creates risk strata and checks for records at risk. This step controls the number of records initially at risk by determining the level of details of the IVs to be released. **S**ubstitution uses optimal sampling rates for selecting records at random for perturbation subject to substitution bias constraints. This step introduces uncertainty about the identity of a target. **S**ub-sampling uses optimal sampling rates to select records from the substituted database at random for non-suppression, subject to precision constraints. This step introduces uncertainty about the presence of a target. **C**alibration uses optimal weight calibration to adjust subsampling weights subject to preserving key estimates from the original database. This step reduces bias due to substitution and variance due to subsampling. The MASSC treatment adds a second phase to the NSDUH data. By making the selection of records independent from PSU to PSU for substitution and subsampling, under general conditions the commonly used single phase survey data analysis methods such as the ones in the SUDAAN software can be used to analyze the MASSC treated data. For more details on MASSC, see Singh (2002), Singh, Yu, and Dunteman (2003), and Singh and Yu (2004).

In applying MASSC to NSDUH data, we need to introduce distortion via substitution and subsampling, and then fix it via calibration. First, the risk strata are created in the Micro Agglomeration step to calculate the proportion of records that are uniques. If there are too many uniques, then some further recoding of IVs into broader categories is performed to reduce the number of uniques.

Next for each record, a substitution partner is assigned such that a suitable distance (Rao's Quadratic Entropy, in particular) based on IVs between the recipient and the donor record is minimized. Then risk strata are subdivided into substrata such that each substratum has relatively low contribution to bias due to substitution for a given set of study variables defined by drug use for various demographic domains. Next using nonlinear optimization techniques, selection probabilities for substitution are obtained such that a disclosure cost function is minimized subject to a set of constraints on bias relative to the original estimate. For PUFs from NSDUH, the overall substitution rates has been around 15% over the years, the substitution rates for individual substrata could be more or less than the overall rate depending on over/under treatment needed under optimization. When a record is selected for substitution, all variables related to the IVs are also substituted from the donor in order to maintain internal consistency.

Similarly, for the subsampling step of MASSC, risk strata are subdivided into substrata such that each stratum has relatively low contribution to variance for a given set of study variables, and then using nonlinear optimization, selection probabilities for subsampling are determined such that a disclosure cost function is minimized subject to a set of constraints on sampling variance relative to the squared original estimate. For NSDUH, the overall subsampling rates has been around 80% while rates for individual substrata reflect over/under treatment as needed.

Based on unique and nonunique occurrence rates, substitution and subsampling rates obtained from the above steps of MASSC, it is possible to compute measures of risk as part of confidentiality diagnostics and check for their adequacy. For example, for a record appearing unique, this risk (δ_u) is computed as the sum of two parts: first part is essentially the product of the probability that the record actually came from the unique risk stratum, with the probability that it survived substitution, with the probability that it was not sampled out, with the probability that it was not misclassified as a nonunique because other records after substitution may have assumed its profile, and finally with the probability that at least one of the SVs takes a sensitive value; while the second part is essentially the product of the probability that it came from a nonunique risk stratum, with the probability that it survived substitution, with the probability that it did not get sampled out, with the probability that it didn't get misclassified as a unique, and finally with the probability that for at least one of the SVs, it has a common value and that it is sensitive. The above risk can similarly be computed for subgroups or broad profiles defined by demographics such as age, gender, and race. Similarly, we can define for a record appearing nonunique double, the disclosure risk ($\delta_{nu(d)}$); for a record appearing nonunique triple, the disclosure risk ($\delta_{nu(t)}$), and for a record appearing nonunique other (i.e., four plus), the disclosure risk ($\delta_{nu(o)}$).

Now, for a record appearing unique in the MASSC-treated NSDUH national PUFs, the chance that it survived the treatment of substitution and subsampling and is at the risk of disclosure from an inside intruder is about 25% (or 1 in 4); this may be deemed reasonably safe in analogy with tabular data where a minimum cell size of 3 to 5 is often considered safe. It may be noted that this measure is quite conservative as it is based on the assumption that the inside intruder knows the presence of the target in the database as well as quite a few IVs. Similarly, for NSDUH PUFs, for a record appearing nonunique double, the chance that it survived treatment and is at the risk of disclosure is under 5%, for a record appearing nonunique triple, the risk is under 1%, and for a record appearing nonunique other (i.e., 4+), the risk is also under 1%. In addition to checking the adequacy of these quantitative measures, it is ensured that no variable is included PUF that provides more detailed information than the IVs considered in the treatment. Also the PUF is checked for records with possible rare values of SVs, if any, which are then treated by top/bottom coding.

Finally, the last step of calibration in MASSC is applied to the substituted and subsampled data by adjusting the subsampling weights so that subpopulation counts for various demographic domains are preserved in the treated data. The method of generalized exponential model of Folsom and Singh (2000) is used for this purpose.

Before the PUF is released, several diagnostics for analytical quality are performed in order to determine the impact of disclosure treatment on bias and precision of estimates compared to the original estimates. In particular, for a number of study variables, point estimates and standard errors are computed before and after the treatment. The difference in point estimates is typically 1 or 2 percent while the decrease in precision is around 7% on average. Ultimately what really matters is the extent to which inferential results of interest to the user could change from the original database. The bias and variance constraints in MASSC in terms of point estimates for various domains provide basic quality protection for key study variables, but it is not possible to guarantee that inference for certain parameters won't change from the original database. Since MASSC treatment depends on the choice of design parameters related to disclosure risk and information loss, it may be possible to repeat MASSC treatment on the same dataset under a different set of parameters so that the treated dataset has more desirable analytical properties.

3. State-level PUFs: Problems and Solutions

If one were to assign state variable to the national PUF created by MASSC, and compute disclosure risk with the state variable included in the set of IVs, the risk can be quite high. The reason for this is that the number of unique records goes up substantially, and hence the number of records at risk.

One could use broader categories of the IVs, thus reducing the available detail for each of those variables until the probability of disclosure was similar to that for the national PUF. However, that would result in a significant loss of information to the researcher. In addition, if the State PUFs were constructed from exactly those records from the national PUF, then there would be a concern that the record in the State PUF could be matched to the national PUF, thereby obtaining the finer detail on the IVs available on the national file.

Since the national PUF was constructed by "randomly" eliminating about one-fifth of the full national sample, that one-fifth-sample remains as a possible source for single-year State PUFs. However, that option is not appealing because then the two files can be combined and knowing that a targeted individual was in sample, a snooper then can also know that the individual is in one of the two PUFs – erasing the uncertainty of whether the target was in the national PUF in the first place.

Another option is to create a State PUF by combining files across multiple survey years. Assume that the desire is to combine data across 4 years. In that case, there is 4 times the sample available for each State PUF from the combined data. Here, MASSC can be applied to the combined four years of the original data - independently of the MASSC used for existing national PUFs. In other words, the random substitution and subsampling steps can be done independently of those used for the national PUFs. This way, matching a record in the State PUF to the national PUF using SVs is a more daunting task. Moreover, a record in the State PUF is not necessarily in the national PUF, nor does a substitute in a State PUF have to be a substitute in the national PUF. Further, the 'same' record could be substituted in both PUFs; however, the substitution variables may differ as well. Note that to assess the risk due to matching of a record in a state PUF to the national PUF amounts to being able to attach the State ID to that record in the national PUF and then the disclosure risk for the national PUF with state as an IV can be computed. However, this has to be pre-multiplied by the probability that a common record survives both the state PUF and the national PUF. It is given by the product of the individual disclosure risks for each of the two data sets which would be smaller than either one, and hence considered safe.

A practical problem with creating 51 PUFs using MASSC is that separately optimizing across 50 States and DC is a computer-intensive application. A compromise might be to use the same substitution and subsampling rates as those used for the national PUFs, but do the selection of records for substitution and subsampling independently from the

national PUF in order to provide an impediment to matching to the national file. The resulting disclosure risks for each state can be computed to see if they are deemed safe enough.

4. An illustration

While various methods of inducing uncertainty in the mind of the intruder have been discussed above, use of the MASSC method requires that the resulting probabilities of individual record identification remain *small* on average and that the resulting size of bias and precision remain acceptable to researchers as embodied in the resulting MSE.

The above idea of combining years for state level PUF was tested with the existing 1999-2002 national PUFs to see if the measure of risk can be reduced reasonably well after state is included as an IV. In this example, the parameters from the national sample were used without optimizing on each State separately. The results are shown in Tables 1(a,b), 2(a,b) and 3(a,b). The first table corresponds to about 80% subsampling similar to what was done for the national PUF, the second corresponds to about 64% subsampling, and the third about 54%. In tables 1(a), 2(a), and 3(a), the first four rows, labeled 1999, 2000, 2001 and 2002, display the result of including the State as an additional IV on the file. The last three rows show the results of combining two, three and four years of data and ignoring the year identifier in the MASSC procedure. The individual deltas are the probabilities of disclosure for 4 categories: all records that appear unique based on the combination of IVs, all records that appear as a nonunique double, all records that appear as a nonunique triple, and all records that appear four plus, i.e., cells having 4 or more records with the same pattern. (They are not the original risk strata because intruder cannot find the true risk stratum the target belongs from the treated data.) These risk probabilities are based on an inside intruder knowing the pattern of IVs of a targeted person on the file.

The delta values shown in the tables show an interesting pattern. Even after combining over four years, the risk does not decrease substantially when other things are the same as for the national PUF such as substitution rates and the categories of IVs. However, as the subsampling rates goes down from 80% to 64% to 54%, the delta values as expected do decrease considerably. In fact, for the 64% subsampling rate, i.e., about 2/3 subsample, the risk seems tolerable. This implies that for the four-year combined PUF, the sample size for the smaller state would be around 2400, which may not be sufficient for some state-level measures. One way out is to do some further recoding of IVs to reduce delta values while keeping subsampling at a higher rate. The above analysis shows that with a suitable choice of treatment parameters the proposed method holds promise. In particular, notice that the probabilities of disclosure for doubles, triples, and cells with 4 or more identical records (with respect to the IVs) have very small probabilities of disclosure whether the MASSC procedure is used on single-year files or multiple-year files. It is mainly the risk for unique appearing records that we need to worry about.

The above risk probabilities represent averages across all states. The probabilities for some of the states were somewhat higher than these overall averages as shown in tables 1(a), 2(a), and 3(a). One goal would be to set a maximum probability of disclosure for every state, and a sufficiently low maximum average across all of the states. Actually optimizing the results for each state rather than using the national settings would be expected to reduce the probabilities of disclosure even more.

5. Concluding Remarks

In general it is harder to protect an individual's confidential information when the individual is known to belong to a subset of the database, e.g., belonging to a particular state than the situation when the state identifier is removed. For NSDUH, this would imply creating 51 PUFs (for 50 states and DC) representing a difficult task as well as requiring more treatment for state-level data. These considerations led to the decision in the past to produce only a national-level PUF from each year of NSDUH data, and to remove any information about census region, state, and other geographic identifiers. However, in this paper we suggested a way in which state-level PUFs could be created if one is willing to combine data over several years.

Clearly, the provision of State PUFs as described above will not answer all needs of researchers. Those who require more detailed geographic identification at the block, tract, or county levels will not be served by this file; nor will those who are interested in a time series by year of their state data because the year variable is not provided. However, the State PUFs will provide a wealth of information for each state for the calculation of point estimates and for analyzing relationships within each state.

We have chosen to combine over four years of data in a single State PUF in order to improve the precision of estimates for the small states. For the large states, the sample sizes are larger, and we are considering the option of whether PUFs for those States should be based on a single year of data or more.

A problem that has not been fully addressed with this method is the possibility of matching a record in the State PUF to one in the national PUF using the full record of SVs. Given that there are over 1,000 variables in the survey questionnaire, matching a record in the State PUF to one in the national PUF based on the SVs could be accomplished with a high probability of a true match even if a number of the IVs have been substituted from another record. While there would be no match for those records that were subsampled out of

either the State PUF or the National PUF, there would be a high probability of a match for any record in both PUFs. With a match, the intruder could assign a State identifier to the same record on the national PUF; thereby making it easier to uniquely identify a person. It may be noted that a high probability of a match does not necessarily lead to disclosure because it depends on the match probability adjusted delta which is simply the probability of a record being at risk in a state PUF multiplied by the probability of a successful match of records from a given state PUF and the national PUF. In addition, further subsampling and substitution could be used to lower the risk of disclosure even more. However, there could still be a significant number of records remaining in common to both the State and national PUF, and it would seem to be advisable to avoid this situation altogether.

There are two methods to address this problem. One is to base the development of the State PUF only on (future) years for which the national PUF has not been released. Then, the MASSC procedure could be used by treating the State identifier as another IV. This would result in a single PUF that could be used for both national and state analysis. We have recently analyzed this approach, and it appears to be fairly satisfactory except that it necessarily results in more subsampling and substitution than were present before in the national PUF. The primary effect of those changes is that the sampling error is increased somewhat.

Another method is to place the "public use" file behind a Data Analysis System that limits the number of variables used in a single analysis. In this way, the entire record of IVs and SVs would never be available for an intruder to use in matching. SAMHSA currently has such an analysis system at its SAMHDA website, managed under contract by the ICPSR. The system permits simple crosstabulations and regression analyses and provides the correct variances that reflect the complex survey design and sample weights. SAMHSA is currently evaluating both approaches.

Because there will always be researcher needs that go beyond what can be provided in a PUF, the OAS is pursuing a licensing process for the NHSDA and other analytic survey files. The license would permit us to share the data with researchers who sign a pledge of confidentiality to protect the data from disclosure and who agree to use the data only for research purposes. In signing the license, a licensee would become an agent of the federal government subject to penalties of up to 5 years in jail and $250,000 fine for any disclosure of individually identifiable information.

References

Bethlehem, J.G, Keller, W.J., and Pannekoek, J. (1990), "Disclosure Control of Micro data", *Journal of the American Statistical association*, 85, 38-45.

Folsom, R.E., Jr., and Singh, A.C. (2000). A generalized exponential model for sampling weight calibration for a unified approach to nonresponse, post-stratification, and extreme weight adjustments. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 598-603.

Singh, A. C. (2002). Method for Statistical Disclosure Limitation. United States Patent Application Pub. No. US 2004/0049517A1 (patent pending).

Singh, A.C., Yu, F., and Dunteman, G.H. (2003). MASSC: A new data mask for limiting statistical information loss and disclosure. *Proceedings of the Joint UNECE/EUROSTAT Work Session on Statistical Data Confidentiality*, Luxembourg, April 7-9 (www.unece.org)

Singh, A.C., and Yu, F. (2004), " Protecting Quality and Confidentiality of Micro Data by MASSC: Review with application", *European Conference on Quality and Methodology in Official Statistics*, Mainz, Germany, May 24-26, (http://q2004.destatis.de)

Table 1(a): Overall Delta after adding state as a core IV (80% subsampling)

Year	delta (uniques)	delta (doubles)	delta (triples)	delta (others)
1999	0.3409	0.0066	0.0004	0.0002
2000	0.3277	0.0082	0.0009	0.0003
2001	0.3786	0.0091	0.0015	0.0006
2002	0.3586	0.0078	0.0010	0.0007
01-02 (2 years)	0.3594	0.0121	0.0017	0.0013
00-02 (3 years)	0.3390	0.0139	0.0022	0.0015
99-02 (4years)	0.3336	0.0151	0.0025	0.0017

Table 2(a) Overall Delta after adding State as a core IV (64% subsampling)

Year	delta (uniques)	delta (doubles)	delta (triples)	delta (others)
1999	0.2753	0.0042	0.0002	0.0001
2000	0.2646	0.0056	0.0005	0.0001
2001	0.3047	0.0063	0.0010	0.0004
2002	0.2893	0.0053	0.0008	0.0003
01-02 (2 years)	0.2904	0.0087	0.0011	0.0009
00-02 (3 years)	0.2748	0.0099	0.0015	0.0010
99-02 (4years)	0.2709	0.0109	0.0018	0.0010

Table 3(a) Overall Delta after adding state as a core IV (54% subsampling)

Year	delta (uniques)	delta (doubles)	delta (triples)	delta (others)
1999	0.2452	0.0039	0.0003	0.0000
2000	0.2378	0.0046	0.0006	0.0001
2001	0.2659	0.0054	0.0008	0.0002
2002	0.2409	0.0042	0.0005	0.0003
01-02 (2 years)	0.2484	0.0068	0.0007	0.0007
00-02 (3 years)	0.2390	0.0082	0.0011	0.0009
99-02 (4years)	0.2367	0.0092	0.0014	0.0010

Table 1(b): Delta for Records Appearing Uniques Calculated by State
(Single-year and Combined-Year: 80% subsampling, only states with 4-year delta more than 0.35 are shown)

State	1999	2000	2001	2002	01-02 (2 years)	00-02 (3 years)	99-02 (4 years)
AZ	0.3878	0.4093	0.4006	0.4032	0.3968	0.3966	0.3952
CO	0.4306	0.4636	0.4301	0.4258	0.4239	0.4302	0.4246
CT	0.3536	0.4422	0.4132	0.4035	0.3988	0.4013	0.3931
GA	0.3193	0.3492	0.3793	0.3965	0.3827	0.3669	0.3621
IA	0.3944	0.3863	0.3553	0.4123	0.3697	0.3637	0.3598
IN	0.3798	0.3876	0.4163	0.3726	0.3890	0.3789	0.3739
KY	0.3625	0.3685	0.4365	0.4406	0.4328	0.4037	0.3933
LA	0.3650	0.3975	0.4423	0.3905	0.4140	0.4039	0.3984
MA	0.4087	0.4238	0.4862	0.3915	0.4292	0.4186	0.4123
MD	0.3559	0.3775	0.3852	0.3723	0.3698	0.3627	0.3583
MN	0.3992	0.4624	0.5035	0.4126	0.4480	0.4408	0.4247
MO	0.3894	0.4034	0.4477	0.4203	0.4267	0.4087	0.3977
NC	0.3549	0.3845	0.4132	0.3878	0.3958	0.3812	0.3680
NJ	0.3732	0.3423	0.3962	0.3778	0.3798	0.3588	0.3559
OR	0.3595	0.3718	0.3990	0.4016	0.3964	0.3847	0.3769
VA	0.3725	0.3565	0.4241	0.3908	0.4029	0.3780	0.3729
WA	0.3703	0.3526	0.4415	0.4315	0.4360	0.4013	0.3878
WI	0.3997	0.4481	0.4836	0.3932	0.4329	0.4301	0.4184

Table 2(b) Delta for Records appearing Uniques Calculated by State
(Single-year and Combined-Year: 64% subsampling; only a subset of states shown as in Table 1(b))

State	1999	2000	2001	2002	01-02 (2 years)	00-02 (3 years)	99-02 (4 years)
AZ	0.3107	0.3319	0.3261	0.3278	0.3240	0.3236	0.3226
CO	0.3503	0.3654	0.3432	0.3530	0.3433	0.3456	0.3415
CT	0.2784	0.3535	0.3286	0.3248	0.3188	0.3223	0.3166
GA	0.2584	0.2847	0.3064	0.3173	0.3087	0.2967	0.2929
IA	0.3212	0.3103	0.2894	0.3272	0.2983	0.2953	0.2916
IN	0.3043	0.3095	0.3242	0.3056	0.3107	0.3026	0.3000
KY	0.2878	0.2964	0.3494	0.3525	0.3469	0.3255	0.3173
LA	0.2984	0.3253	0.3502	0.3041	0.3254	0.3216	0.3174
MA	0.3260	0.3420	0.3886	0.3108	0.3455	0.3383	0.3331
MD	0.2776	0.3142	0.3078	0.2977	0.2971	0.2962	0.2929
MN	0.3212	0.3849	0.4083	0.3232	0.3583	0.3584	0.3466
MO	0.3127	0.3138	0.3619	0.3387	0.3455	0.3285	0.3196
NC	0.2878	0.3114	0.3318	0.3046	0.3139	0.3070	0.2967
NJ	0.2999	0.2842	0.3199	0.3034	0.3045	0.2923	0.2915
OR	0.2908	0.3060	0.3267	0.3143	0.3177	0.3111	0.3053
VA	0.2963	0.2823	0.3428	0.3168	0.3267	0.3052	0.3011
WA	0.3027	0.2877	0.3585	0.3529	0.3553	0.3282	0.3173
WI	0.3197	0.3712	0.3880	0.3173	0.3487	0.3509	0.3422

Table 3(b) Delta for records Appearing Uniques Calculated by State
(Single-year and Combined-Year: 54% subsampling; only a subset of states shown as in Table 1(b))

State	1999	2000	2001	2002	01-02 (2 years)	00-02 (3 years)	99-02 (4 years)
AZ	0.2745	0.2915	0.2868	0.2786	0.2794	0.2810	0.2811
CO	0.3104	0.3424	0.3057	0.2894	0.2961	0.3068	0.3042
CT	0.2583	0.3115	0.2861	0.2684	0.2726	0.2765	0.2746
GA	0.2353	0.2495	0.2822	0.2806	0.2788	0.2655	0.2624
IA	0.2895	0.2686	0.2508	0.2829	0.2584	0.2541	0.2534
IN	0.2663	0.2881	0.3039	0.2713	0.2825	0.2794	0.2766
KY	0.2624	0.2663	0.3127	0.3053	0.3061	0.2871	0.2809
LA	0.2575	0.2785	0.3104	0.2629	0.2858	0.2803	0.2780
MA	0.2952	0.3054	0.3526	0.2677	0.3043	0.2998	0.2955
MD	0.2524	0.2728	0.2816	0.2598	0.2640	0.2597	0.2571
MN	0.2907	0.3410	0.3632	0.2900	0.3192	0.3193	0.3076
MO	0.2856	0.3057	0.3245	0.2983	0.3069	0.2989	0.2894
NC	0.2626	0.2864	0.2952	0.2892	0.2885	0.2817	0.2711
NJ	0.2653	0.2513	0.2892	0.2690	0.2748	0.2602	0.2585
OR	0.2543	0.2640	0.2814	0.2553	0.2652	0.2625	0.2587
VA	0.2686	0.2583	0.3021	0.2778	0.2859	0.2712	0.2683
WA	0.2700	0.2549	0.3227	0.3044	0.3140	0.2901	0.2799
WI	0.2887	0.3287	0.3498	0.2773	0.3094	0.3125	0.3028