NATIONAL ACADEMIES *Sciences Engineering Medicine*

A Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation

Committee on National Statistics

National Academies of Sciences, Engineering, and Medicine

Agenda

1

About the Panel

2

Assessing the Level of Disclosure Risk in SIPP—Joseph Hotz

3

Promoting Usability and Accessibility in SIPP While Respecting Respondent Confidentiality—Jennifer Van Hook

4

Virtual Data Enclaves and Secure Data Access—Heeju Sohn

5

The Potential of a Remote Analysis Platform as a Tool for Protecting Confidentiality—Thomas Krenzke

6

Geography and Small Area Estimation in SIPP—Scott Holan

7

Timeline—Brad Chaney

8

Discussant—Caleb Floyd

NATIONAL ACADEMIES Sciences Engineering Medicine

About the Panel

Panel Members

TRIVELLORE RAGHUNATHAN (*Chair*), University of Michigan
SCOTT H. HOLAN, University of Missouri
V. JOSEPH HOTZ, Duke University
THOMAS KRENZKE, Westat
FANG LIU, University of Notre Dame
ROBERT A. MOFFITT, Johns Hopkins University
AMY PIENTA, Inter-university Consortium for Political and Social Research
NATALIE SHLOMO, University of Manchester
ALEKSANDRA (SEŠA) SLAVKOVIĆ, Pennsylvania State University
HEEJU SOHN, Emory University
SALIL VADHAN, Harvard School of Engineering and Applied Sciences
JENNIFER VAN HOOK, Pennsylvania State University

Panel Members	Statisticians	Survey Methodologists	Economists	Computer and Data Scientists	Technical and Policy Experts Evaluating Government Programs	Sociology
Trivellore Raghunathan (Chair)	■	■				
Heeju Sohn					■	■
Thomas R. Krenzke	■	■				
Fang Liu	■					
Scott H. Holan	■					
Robert A. Moffitt			■			
Amy Pienta					■	■
Natalie Shlomo	■	■				
Jennifer Van Hook					■	■
V. Joseph Hotz			■		■	
Salil Vadhan				■		
Aleksandra (Seša) Slavković	■			■		
Total=12	6	3	2	2	4	3

Peer Reviewers of the Report

- **CLAIRE MCKAY BOWEN,** Urban Institute
- **SAKI KINNEY,** RTI International
- **DAVID VAN RIPER,** University of Minnesota
- **KENNETH W. WACHTER,** University of California, Berkeley
- **LANCE A. WALLER,** Emory University
- **EMILY E. WIEMERS,** Syracuse University

- **WILLIAM W. STEAD,** Vanderbilt University Medical Center (Monitor)
- **JOHN L. CZAJKA,** Independent consultant (Coordinator)

Report Sponsor

- **Census Bureau**

Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project.

SUMMARY STATEMENT OF TASK

The National Academies of Sciences, Engineering, and Medicine will appoint an ad hoc panel to consider the challenges facing the protection and release of public use data from the Survey of Income and Program Participation (SIPP). As part of its fact gathering, the panel will consider:

- the evolving privacy risks to releasing survey data;
- developments in methods for protecting privacy and reducing risks of disclosure, including formal privacy methods being implemented at the Census Bureau;
- the dimensionality and longitudinal nature of the SIPP data;
- the linking of SIPP data with administrative data;
- existing SIPP data products and the utility of detailed public use microdata that enable scientific discovery;
- selected other SIPP data products, such as a small area estimates program for key SIPP measures; and
- the need for protecting the confidentiality of the SIPP data, potentially across multiple data releases, while providing timely access for the many research uses of SIPP.

The panel will produce a report with conclusions and recommendations for disclosure protection and data provision from the SIPP program.

NATIONAL ACADEMIES *Sciences Engineering Medicine*

Assessing the Level of Risk in SIPP

NASEM Consensus Report: A Roadmap for Disclosure Avoidance in the SIPP

V. Joseph Hotz

Arts & Sciences Distinguished Professor of Economics & Public Policy

Duke University

Measures of Disclosure Risk

- ***Absolute Disclosure Risk***: Probabilistic measure of risk of identifying individuals & their information from released data, auxiliary data and intruder’s prior information.
  - This form of risk presumably is what an individual & data providers may care about in assessing disclosure avoidance ways/mechanisms
  - But assessing absolute risks in designing mechanisms is challenging and difficult to quantify, i.e., requires extensive knowledge & information on part of intruders.

- ***Relative Disclosure Risk***: Probabilistic measure to assess the relative risk of detecting individuals’ identities & information based on any individual being included or not included in released data and statistics.
  - Relative disclosure risks are easier to formally characterize & control.
  - This measure is focus of formal disclosure risk criteria, e.g., the ***Differential Privacy (DP) Criterion.*** (More on DP and associated mechanisms in other presentations.)

Forms of Detecting Confidential Data/Information from Released Data

- ***Reconstruction***: A Intruder/Attacker uses published statistics & solves large system of equations or optimization problem to reconstruct underlying confidential microdata.

- ***Reidentification***: Intruder/Attacker uses released data to identifies individuals/entities in released data.
  - Typically done by matching elements of released data to auxiliary dataset or by direct knowledge by matching on characteristics that are unique to individuals/entities.
  - Reidentification attacks can allow intruder to determine individuals' identities & information about individuals not available in auxiliary data, e.g., personal income.

- ***Membership Inference***: Intruder/Attacker uses released data & auxiliary data to make inference that individuals are in released data.
  - Unlike reidentification, this form of risk is ***inferential***, i.e., probabilistic reidentifiation, possibly with high degrees of precision.

- Disclosure avoidance mechanisms seek to address all 3 forms, especially reidentification and membership inference.

Strategies for Assessing Disclosure Risk from Released Data & Its Disclosure Avoidance

- Empirical &/or statistical modeling to quantify disclosure risks by simulating intruders' reidentification strategies, but with knowledge of confidential data to verify actual risks, especially for "uniques."
- Data providers (Census) employ:
  - External data sources (e.g., commercial databases) to determine ***exact matches via record linkages***.
  - Methods that determine & ***focus determining "population uniques"*** – individuals with unique characteristics in certain contexts), e.g., African Americans who are highly educated in contexts (locations) – to quantify risks.
  - Related is ***Data Intrusion Simulation***, i.e., quantifies probability that individual(s) in dataset by assessing ratios of unique characteristics to relevant populations.
- Latter methods may not be sufficient to determine exact reidentification but can be powerful in quantifying risks for certain uniques.

A Priori Disclosure Risks faced by SIPP

- SIPP has complex sample design, e.g., over samples, etc.
- Released SIPP contains microdata on a large number of characteristics of households & their members, e.g., family size & composition, marital status, program participation & income.
- SIPP is longitudinal, so contains information on stability &/or changes in characteristics that may increase reidentification risks.
- Data elements in SIPP are imputed.
  – Imputation may reduce disclosure risks, but (partial or complete) knowledge of methods of imputation may increase them disclosure.
- ***Summary***: Disclosure risk assessment is inherently more complicated for SIPP & less is known about how combination of features affects disclosure risks.

Census Bureau’s Initial Reidentification Study (Re-id Study)

- Census Bureau conducted an initial reidentification study for the SIPP which we refer to as the “Re-id Study” in our Report.
- Used data from the SIPP 2014 Panel, including 4 years of data for same households. Focused on primary taxpayers.
- Combined 36 indirect identifying variables in their data & combined several different external data sources: SSA & IRS data and SSA Numident file.
- Used statistical matching with two matching criteria to perform probabilistic matches.
- Using their Internal Use Files they are able determine true matches and compute 2 matching rates: ***Confirmed Rate*** and ***Conditional Rate*** (denominator for latter based on suspected matched cases)

Findings from Re-id Study

- Risks of reidentification where “better than expected” (Census assessment).
- 72.2% & 13.9% of state x metro/non-metro area strata had zero confirmed reidentifications, depending on metric used for probabilistic matching.
- Households which moved had similar confirmed matches compared to non-movers.
- Census found no clear patterns across identifying variables used in matching to their external databases.
- Did find that age was the most disclosive variable among those used.

Panel's Reactions to Re-id Study

- Panel was impressed with Re-id Study!
- But did note several limitations
  - ***Census used imputed values***, e.g., income, ***in Re-id Study***.
    - Disregarding this may underestimate disclosure risks as it assumes that imputations are best guess of an intruder. Alternatively, may overstate disclosure risks if imputations are less predictive of true income. Imputation needs more attention.
  - Released findings from Re-id Study limited, making it somewhat difficult to assess Study's findings.
  - Additional analyses would be informative
    - Alternative thresholds for putative matches.
    - Use more than household moves to assess disclosiveness of longitudinal features of data, e.g., changes in marital status.
    - Use of information about all household members & changes in household composition.
    - Use of other databases for matching, e.g., commercial databases, that are more likely to be available to intruders.
    - Further assessment of impact of differences in structure/configuration of databases Census did use.

Panel's Own Assessment of Potential Disclosure Risks in SIPP

- Analysis of the SIPP 2020 data shows considerable potential for disclosure risk, though the exact risk cannot be quantified without a reidentification study.
- A combination of six variables (sex, race, ethnicity, state, birth year, and highest level of education) is sufficient to uniquely characterize 59 percent of the respondents, even when looking only at the primary household member.
- Adding in data about a second person in the household will uniquely characterize more than 90 percent of the respondents.
- Adding two more variables (the number of children, and change over time in the number of children) is enough to uniquely characterize more than 90 percent of respondents even when looking only at the primary household member.
- Though such sample uniques by themselves may not pose disclosure risk, the rarity of those combinations of characteristics in the general population needs to be assessed, requiring a reidentification study using a variety of data sources.

Panel’s Recommendations for On-Going Disclosure Risk Assessments

- Important to develop and maintain an on-going program of disclosure risk assessment for SIPP (Rec. 3-1)
  - Important given evolving nature of algorithms available for reidentification attacks and increasing availability of external databases for such attacks.
  - As Census considers/develops new products and modes of dissemination (more in other presentations), disclosure risk assessment will be needed for these products & modes
- Census needs to communicate findings to user community so the latter better understands need for disclosure avoidance techniques (Rec. 3-2)
- Census should partner with & involve external researchers/experts in its risk assessment research program (Rec. 3-3)

Thank You

For more information, please contact:

Joseph Hotz

v.joseph.hotz@duke.edu

NATIONAL ACADEMIES Sciences Engineering Medicine

Maintaining usability while preserving confidentiality

Jennifer Van Hook (Penn State),

Robert Moffitt (Johns Hopkins)

Heeju Sohn (Emory)

Thursday, October 26, 2023, College Park, MD

The views presented are those of the author(s) and do not represent the views of any Government Agency/Department or Westat.

OCTOBER 2023

“Full usability is achievable by releasing all of the original data, and full confidentiality is achievable by suppressing all of the data. The difficulty is in finding an appropriate balance between the two.”

--*A Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation (SIPP)*

Three Dimensions of Usability

Accuracy: Ability to obtain valid and reliable inferences from data

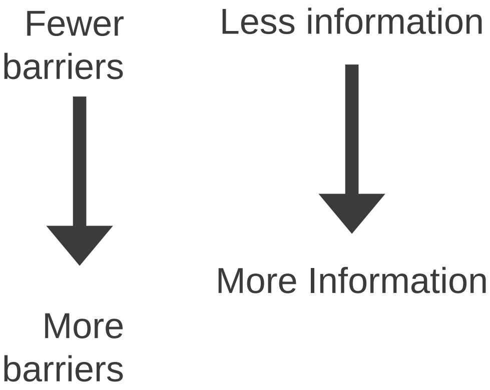
Feasibility: Are variables & software available to conduct analysis?

Accessibility: Can data be accessed; who can access it?

→ How should CB design modes of access that maximize these dimensions of usability?

- → Principles CB might follow
- → Communications with SIPP users
- → Interpretation of Title 13 in context of new legislation

Modes of Access



Fewer barriers

More barriers

Less information

More Information

Online tabular/analysis builder
Public-use microdata
Synthetic data
Secure online data access
Federal Statistical Research Data Centers

Accuracy

It is important to understand how privacy protections affect accuracy and communicate findings with SIPP users

Can Valid and Reliable Inferences be Made?

- Answer depends on mode of access and type of analysis
  - Jordan Stanley and Evan Totty (2023) compared SIPP Synthetic Beta with Gold Standard File
  - Simple statistics held up well
  - Less so for analyses that relied on data features that were not explicitly modeled

- Important for CB to clearly communicate such findings to the SIPP user community

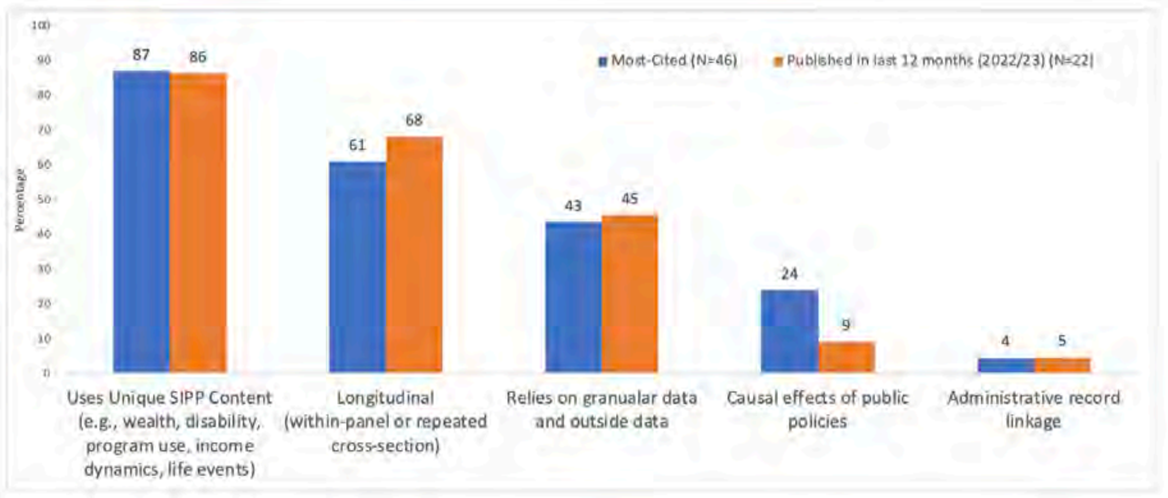
- Training needed for SIPP users if special methods/software required to generate correct estimates/SEs

Feasibility

CB should prioritize uses that build on the key strengths of SIPP data

Five Use Categories

1. Analyses Relying on Unique SIPP Content (e.g., disability, poverty, program use, life events). ~20% descriptive only.
2. Longitudinal Analysis (60-70%)
3. Analyses relying on granular or outside data (40-45%)
4. Causal effects of public policies (10-25%)
5. Analyses relying on administrative record linkages (4-5%)

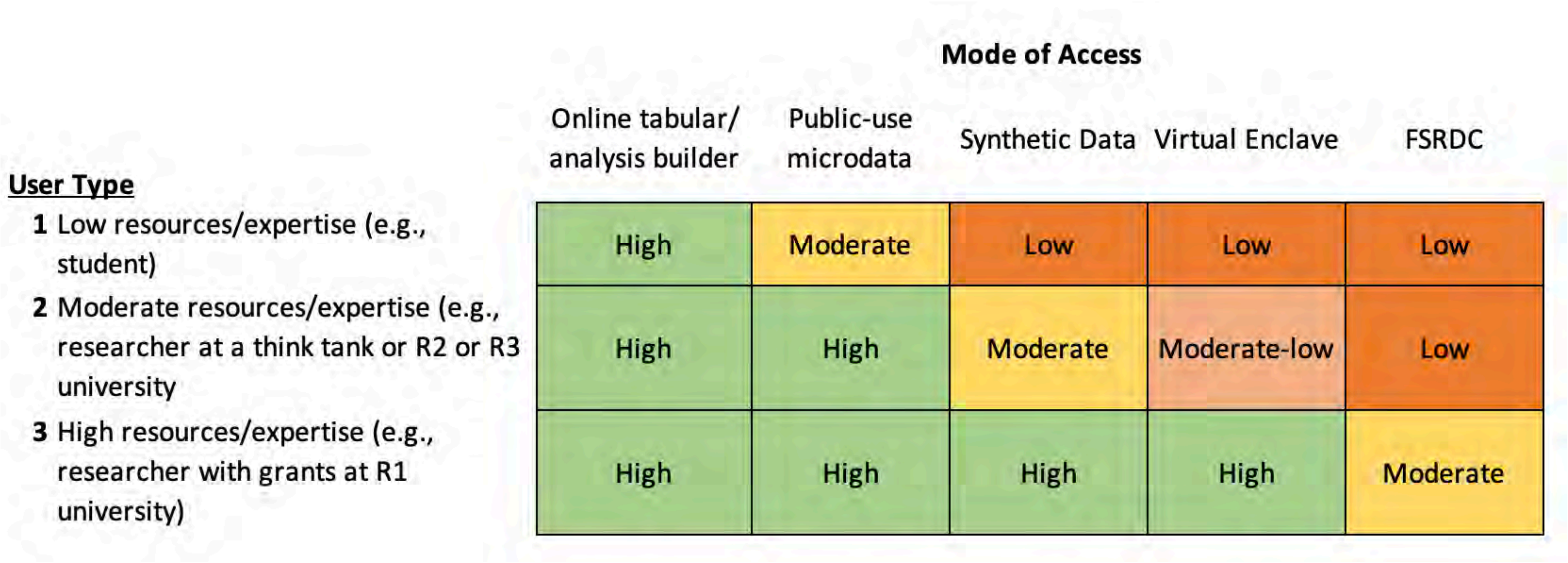


Accessibility

It is important to consider barriers and equity in access

Who benefits most from…

Addition of an on-line analysis tool?
Addition of a virtual enclave?
Preservation of public-use data file?



	Mode of Access				
User Type	Online tabular/ analysis builder	Public-use microdata	Synthetic Data	Virtual Enclave	FSRDC
1 Low resources/expertise (e.g., student)	High	Moderate	Low	Low	Low
2 Moderate resources/expertise (e.g., researcher at a think tank or R2 or R3 university	High	High	Moderate	Moderate-low	Low
3 High resources/expertise (e.g., researcher with grants at R1 university)	High	High	High	High	Moderate

Title 13

- Title 13 does not require benefit to CB specifically, only accomplishing the *work* of CB
- Expanding data access for evidence building is also a benefit to CB (Title 44, Evidence Act)
- Title 13 requires users be sworn to protect confidentiality but does not impose conditions on who may be sworn and how. Can process be streamlined?
- Rather than offering a largely dichotomous choice between public-use data and FSRDCs, the Census Bureau should consider allowing for multiple levels of security depending on the sensitivity of the data.

Recommendations

9-1: CB should conduct regular **assessments of validity and reliability** and **communicate** results to SIPP users.

9-2: When considering which access modes to prioritize, CB should evaluate feasibility for the **most common SIPP uses** and those that exploit the unique characteristics of SIPP

9-3: CB should seek to **continue providing public-use files** for SIPP users, assuming that appropriate disclosure avoidance techniques can be adopted

9-4: Given the differences in user needs and approaches, CB should offer **multiple tiers of access**

9-5: CB should **modernize its interpretation of Title 13** consistent with changes in technology, policy guidance, and legislation (i.e., the Evidence Act and the Information Quality Act).

Thank You

For more information, please contact:

Jennifer Van Hook

jxv21@psu.edu

Virtual Data Enclaves and Secure Online Data Access

Disclosure Avoidance in the Survey of Income and Program Participation

Heeju Sohn, Emory University

FCSM OCTOBER 2023

Agenda

1

Current access modes and their limitations

2

What is Secure Online Data Access (SODA)?

3

SODA Capabilities

4

Enhancing Research Equity

Current ways to work with the SIPP

PUBLIC USE FILES

- Available online without registration
- Protected using traditional statistical disclosure limitations

FSRDC

- Applied with fewer disclosure protections
- Restricted to users with Special Sworn Status

SIPP SYNTHETIC BETA

- Synthetized data of 9 SIPP panels linked to IRS, SSA records
- Access granted through an online server (currently unavailable)
- Output validated against the "Gold Standard"

Secure Online Data Access (SODA)

SODA offers greater protection than the public use file

1. Users undergo an approval process to gain access to the data
   - Project proposal can require a description of analysis and a justification for data access
   - Applicants can be vetted based on their affiliations and credentials
   - Requiring Institutional Review Board (IRB) approvals/exemptions and Data Use Agreements (DUA)s can enhance security

Secure Online Data Access (SODA)

SODA offers greater protection than the public use file

2. SODA environments resemble remote desktop connections
   - Virtual environments can be configured to meet security standards for SIPP
   - Remote monitoring can identify risk and terminate access in minutes
   - User-specific restrictions and controls can enhance security
   - Centralized administration provides efficient management of applications

Secure Online Data Access (SODA)

SODA offers greater protection than the public use file

3. SODA environments can augment social and legal controls for data access
   - Provide notifications for expiring agreements and manage compliance
   - Implement disclosure reviews for analysis output
   - Automatically terminate access at the end of the agreement period

Secure Online Data Access (SODA)

SODA can expand access and facilitate collaboration

- Simplified approval process can reach more users than FSRDCs
- Computing resources can adjust to user demand
- Multi-site research teams can apply for one access mode
- Virtual environment can offer specialized statistical software to users

Enhancing Research Equity through SODA

Expand access to researchers with less research support

Preserve research of people with uncommon characteristics

Use of SODA

NCES
National Center for Education Statistics

U.S. BUREAU OF LABOR STATISTICS

BJS
Bureau of Justice Statistics

USDA

NIH
National Institute on Drug Abuse

MILLENNIUM
CHALLENGE CORPORATION
UNITED STATES OF AMERICA

FDA

Recommendation

- Recommendation 5-1: If disclosure risk assessment studies find that the current public-use file does not provide adequate disclosure avoidance, the Census Bureau should consider secure online data access as a mode likely to support both access and security.

Thank You

For more information, please contact:

Heeju Sohn

heeju.sohn@emory.edu

NATIONAL ACADEMIES Sciences Engineering Medicine

The Potential of a Remote Analysis Platform as a Tool for Protecting Confidentiality in SIPP

Tom Krenzke (Westat),

Natalie Shlomo (University of Manchester)

Thursday, October 26, 2023, College Park, MD

The views presented are those of the author(s) and do not represent the views of any Government Agency/Department or Westat.

OCTOBER 2023

Outline

- Introduction to remote analysis server with table builder
- Disclosure limitation and risk mitigation
- Output perturbation
- Roadmap

Introduction

What is a Remote Analysis Platform?

- A query-based system (UI)
- Performs statistical analysis (appropriate for complex samples and censuses)
  - Tabular (Flexible table generator)
  - Exploratory data analyses
  - Regression models
- Uses underlying microdata which may have undergone some protection
- Displays safe outputs of summary or aggregate data and statistical analysis

Remote Analysis Platform Examples

- National Center for Education Statistics (NCES) DataLab
- European Union Census Hub: https://ec.europa.eu/CensusHub2
- Australian Bureau of Statistics Tablebuilder: https://www.abs.gov.au/statistics/microdata-tablebuilder/tablebuilder (includes means, medians, sums, confidence intervals)
- Westat’s WesDaX® www.wesdax.com

Why Consider a Remote Access Server for SIPP?

- Allows for open access
  - Public demands for more data at higher levels of granularity
  - Safe use of more restricted data
    - Potentially through perturbation of outputs
- Expand the user base
  - Data already loaded and ready to go
  - No coding necessary
  - May include policy analysts, grant writers and other stakeholders who otherwise would not be able to use SIPP data

Disclosure Limitation and Risk Mitigation

Risks in Summary and Aggregate Outputs

- Tabular outputs can be differenced and manipulated to reconstruct microdata
- There is less risk of attribute disclosure (rows/columns have 1 or 2 non-zero cells) for survey data where zeros can be random
- Medians and percentiles are disclosive on skewed data
- Regression coefficients, p-values
    - See Ritchie (2019) for further information regarding disclosure risk in regression coefficients
- Remote analysis servers on (weighted) survey data ‘less risky’ than census or business data

Reduce Risks from Input Data

- Underlying data likely needs to be coarsened (recoded)
- Some variables may need to be dropped and categories combined – due to sensitivity
- Limit what variables can be used for filtering in the table builder to reduce risks due to table differencing

Reduce Risks in Outputs

- System rules, e.g., limit dimensions of tables, number of categories in regression modelling, etc.
- Any analysis that does not meet a threshold should trigger a warning requesting users to redefine their analysis
- Different thresholds (T) depending on type of analysis, e.g., base populations in percentage distributions, average cell size for tables, number of records in X-variables for regression modelling, etc.

Reduce Risks in Outputs (cont.)

- For survey data, provide standard errors/confidence intervals with low-quality statistics flagged
- Consider using robust regression to down-weight outliers
- Round regression coefficients and other statistics
- Exact p-values can reveal z-scores and therefore exact data points so consider rounding or binning the p-values
- Ensure no single data points are disseminated:
  - Do not disseminate maximum, minimum
  - Only allow percentiles (median) if the data points have multiple values
  - Use sequential box plots for scatter plots, residual plots (O'Keefe and Shlomo, 2012)

Output Perturbation

Noise Infusion on Outputs

- In some cases it may be necessary for perturbative methods on outputs, e.g., census data or highly skewed target populations
  - Typically carried out by noise addition to outputs
- Perturbation matrix $\boldsymbol{P}$ where

$$P_{ij} = p(\text{perturbed cell value is } j | \text{original cell value is } i)$$

- For each cell count, change (or no change) the value according to probability P_{ij} and the outcome of a random draw
- Let $\boldsymbol{R}$ be the invariant matrix of $\boldsymbol{P}$ dependent on vector of frequencies so that $tR = t$ (Shlomo and Young 2008) and additivity preserved in expectation (IPF can also be performed)

Noise Infusion on Outputs

- Properties relating to achieving the same results (Krenzke, et al. 2013)
  - Cell consistency – Across multiple users, if the same set of records contributes to a table cell, the same results are attained
    - This is attained by a summing a microdata-level random key among cell members
  - Query consistency – Across multiple users using the same query path (e.g., same specification for universe definition and requested table), the same results are attained
    - This is attained by a function of the sum of the microdata key for the cell, marginals associated with the cell, and the table universe
- Attaining cell consistency has less protection than attaining query consistency due to table differencing for tables that differ by one case
- This is referred to as a “non-interactive” perturbation mechanism

Differential Privacy

- List space $a = (a_1, \ldots, a_k)$, e.g., internal cells and margins (overlapping individuals)
- Consider $M(.)$ such that $M(a) = b = (b_1, \ldots, b_k)$ where $M(.)$ set of discrete conditional probabilities $p(b_k|a_k)$ and cells perturbed independently
- Definition: $M(.)$ satisfies ε - differential privacy if for all neighbouring lists a, a' differing by one individual:

  $P(M(a) = b) \leq e^{\varepsilon} P(M(a') = b)$ and this is true for all potential lists and all possible outcomes
- Relaxation: $M(.)$ satisfies (ε, δ) - differential privacy if

  $$P(M(a) = b) \leq e^{\varepsilon} P(M(a') = b) + \delta$$
- This is used to put a cap on the perturbations

Differential Privacy

- Exponential Mechanism (McSherry and Talwar, 2007)
  defined with respect to utility function u which assigns a utility score to possible perturbed values and the mechanism is selected that produces values with high utility
- Define a loss function:

  $l_1 = \sum_{i=1}^{K} |a_k - b_k|$ (motivated by discretized Laplace)
- Then define $u_1 = -l_1$

Exponential Mechanism

- Exponential mechanism defined by: given a, choose $b \in B$ (B: range of b) with probability proportional to: $e^{(\varepsilon/2)u/\Delta u}$ where
$$\Delta u = \max_{b \in B} \max_{a \sim a' \in A} |u(\mathrm{a}, \mathrm{b}) - u(a', \mathrm{b})|$$
- Bound the perturbations $|a_k - b_k| \leq m, \forall k$, then for all $a \sim a' \in A$, if $P(M(a') = b)$ = 0 implies $P(M(a) = b) < \delta$ then $M(.)$ satisfies (ε, δ) - differential privacy
- Examples of perturbation vectors:

ε =1.5 δ=0.00002

-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
0.00002	0.00008	0.00035	0.00157	0.00706	0.03162	0.14172	0.63516	0.14172	0.03162	0.00706	0.00157	0.00035	0.00008	0.00002

ε = 0.5 δ = 0.008

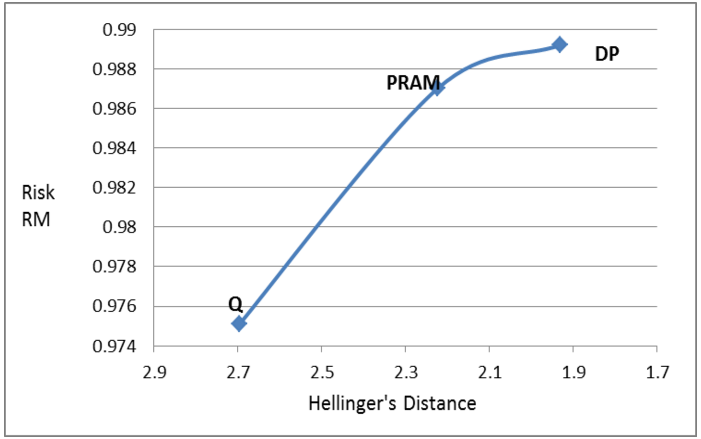
-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
0.0076	0.0125	0.0206	0.0339	0.0559	0.0922	0.1520	0.2506	0.1520	0.0922	0.0559	0.0339	0.0206	0.0125	0.0076

Exponential Mechanism Implications

- DP leads to negative values, setting to zero still ensures DP but bias perturbations
- All zeroes must be perturbed
- If list-space internal cells only $\Delta u = 1$, margins summed from internal cells DP but low utility
- In a t-way table all margins, $\Delta u = 2^t - 1$ (not including total) much larger perturbations
- Margins can be perturbed (with appropriate sensitivity) and IPF to ensure additivity would not negate DP
- Use microdata keys, filter variables and ensure thresholds that do not allow tables that differ by a single individual

Flexible Table Builders for Survey Data

- Perturbation carried out on sample counts and weighted count is adjusted accordingly
- For example, a perturbation: ‘add 3 to the sample count’ then we adjust weighted count by:
  - Add 3 × overall average weight to the original weighted count; or
  - Add 3 × average weight in the cell to the original weighted count
- See:
  Shlomo, Krenzke and Li (2019)
  for a comparison of methods for
  table builder of survey data:
  Post-randomization
  Drop/Add-up-to-q (Li and Krenzke, 2016)
  Noise infusion under Differential Privacy



Roadmap

Recommendation

- Recommendation 7-1: The Census Bureau should assess the demand for an initial flexible table generator as a simple tool, with a specific purpose, that is designed to gauge value and provide direction for further development, and proceed with the development of one if there appears to be sufficient demand.

Summary: Roadmap

- Release of flexible table generator as a simple tool and initial product to gauge value and interest, using the public-use SIPP, with no noise introduced
- Survey of tool users to determine needs, audience for tool, common queries, comments
- Decisions on future development and next steps, for example in the following sequence:
  - Flexible table generator, using the restricted data under the differential privacy framework
  - Remote analysis platform, using the public-use file data, with no noise introduced
  - Remote analysis platform, using the restricted data under the differential privacy framework

References

References

- Krenzke, T., Gentleman, J., Li, J., and Moriarity, C., (2013). Addressing disclosure concerns and analysis demands in a real-time online analytic system. Journal of Official Statistics, 29(1), 99-134.
- Li, J., and Krenzke, T. (2016) Confidentiality Approaches for Real-time Systems Generating Aggregated Results. Proceedings of the Survey Research Methods Section of the American Statistical Association. Available at: http://www.asasrms.org/Proceedings/y2016/files/389621.pdf
- McSherry, F. and Talwar, K. (2007) Mechanism design via differential privacy. In Foundations of Computer Science, 2007 (FOCS'07), 48th Annual IEEE Symposium. IEEE 94-103.
- O'Keefe, C.M. and Shlomo, N. (2012) Comparison of Remote Analysis with Statistical Disclosure Control for Protecting the Confidentiality of Business Data. Transactions on Data Privacy, 5(2), 403-432.
- Ritchie, F. (2019) Analyzing the disclosure risk of regression coefficients. Transactions on data privacy, 12(2), 145-173.
- Shlomo, N., Krenzke, T. and Li, J. (2019) Confidentiality Protection Approaches for Survey Weighted Frequency Tables. Transactions on Data Privacy, 12(3), 145 – 168.
- Shlomo, N. and Young, C. (2008) Invariant Post-tabular Protection of Census Frequency Counts. In PSD'2008 Privacy in Statistical Databases, (Eds. J. Domingo-Ferrer and Y. Saygin), Springer LNCS 5261, 77-89.

Thank You

For more information, please contact:

Thomas Krenzke

TomKrenzke@westat.com

NATIONAL ACADEMIES *Sciences Engineering Medicine*

Geography and Small Area Estimation in the SIPP

National Academies: A Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation (SIPP)

Scott H. Holan

Professor of Statistics – University of Missouri

FCSM - 2023

> “Measures of geography are a notable source of risk because they are highly identifiable, and they narrow down the sample and the overall population to much smaller subgroups.”

--*A Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation (SIPP)*

Challenges with Geographic Variables

Given the notable source of risk that measures of geography present, the Census Bureau limits the amount of geographic data available in the SIPP

- As is standard for confidential surveys, home addresses are automatically suppressed, as are the names of counties and specific metropolitan areas.
- While SIPP includes a measure indicating whether the respondent is in a metropolitan area, that measure is used only when a state has multiple metropolitan areas.
- SIPP does include state identifiers for all the states. However, the inclusion of state identifiers can still be highly disclosive.
- The challenge from a data usability perspective is that geography variables can be highly valuable for research.
- The panel considered three use-case scenarios in developing potential strategies for providing geographical variables and limiting disclosure.

Identifying Specific Geographies is Sometimes Unnecessary

Although geography is a natural variable used for SAE, it is not strictly needed to produce the national-level regression analyses that are often sought by data users (i.e., when the inferential task does not involve the geography).

- For some analyses it is sufficient to release public-use microdata with a “mask-id” (i.e., a “group-level” identifier) rather than a state-level identifier.
- Releasing a mask-id would allow researchers to incorporate this variable for analysis as either a fixed or random effect in a regression model and would provide the same benefit as including a state-level (or other geographic-level) identifier but with the benefit of not necessarily being disclosive.
- Still, some states may be identifiable even with a mask-id.
  - For example, the 10 largest states can each be identified by assuming that a number of respondents (or the sum of the sampling weights) in SIPP have the same rank ordering as the state population sizes.
    - Level five is Arial 9pt, line spacing 1.1, paragraph spacing 6pt. Color is dark gray.

Identifying Specific Geographies is Sometimes Unnecessary

- If the state-level id is determined to be a disclosure risk, then it may not be viable to simply associate a mask-id with each respondent.
- In this case, another potential path forward would be to consider partially synthetic data, where the state-level id (or mask-id) is synthesized for each record in the public-use file.
- Although there has been some research on generating synthetic geographies, this is typically conducted to produce a synthetic location for each respondent.
- Instead, generating a synthetic state-level id (or mask-id) for each respondent would proceed from a latent class model and require additional research.
- In principle, this type of partial synthesis could also proceed using formal privacy methods.

Making Subnational Estimates

There is an increasing demand for reliable estimates at more granular levels of geography and for smaller subpopulations

- In many cases, small geographies or subpopulations have insufficient sample sizes (or no sample) to produce reliable estimates.
- In these cases, model-based approaches allow for the “borrowing of strength” by leveraging different sources of dependence and by incorporating auxiliary information.
- The Census Bureau has a rich history of running various SAE programs (e.g., SAIPE and SAHIE).
- Both programs use Bayesian model-based methodology to provide estimates of increased precision at under-sampled geographies.
- Two types of model-based approaches
  - Area-level models
  - Unit-level models

Making Subnational Estimates: Area-Level Models

Area-level model-based approaches often proceed using a Bayesian hierarchical framework. The model usually begin with the response variable consisting of the direct estimate. Along with this estimate, the statistical agency typically publishes an estimate of the sampling error variance.

- Bayesian hierarchical model can accommodate a broad range of modelling tasks and provide meaningful measures of uncertainty.
- Fay-Herriot model is a special case of the multivariate spatial mixed effects model, where the latent "process" model's random effects are iid.
- The main considerations whether
  - to use multivariate or univariate modelling,
  - to incorporate geographical (spatial) or temporal (longitudinal) dependence (MSTM model),
  - to incorporate auxiliary information (e.g., administrative records and/or data from other surveys)
  - to deal with potentially non-Gaussian responses.

Making Subnational Estimates: Unit-level Models

These models rely directly on answers provided by individual survey respondents as the dependent variable.

- The method can be considered a “bottom-up” approach.
- In other words, in principle, predictions can be made for any level of geography (or subpopulation) and aggregated up to any desired tabulation level, such as county, state, or national (i.e., no need for benchmarking).
- Directly leverage the entire dataset, rather than summary-level statistics (e.g., direct estimates) – potentially more precise than estimates coming from an area-level model.
- Challenges:
  - Informative sampling
  - Non-Gaussian
  - High-dimensionality
- Several approaches, including pseudo-likelihood and regressing on the survey weights, among others.

Making Subnational Estimates

- Incorporating spatial dependence and correlated random effects: area-level versus unit-level
  - Spatial dependence CAR/ICAR model, spatial basis functions, or area-level random effect
  - Temporal/longitudinal dependence: (e.g., CPS, HPS, SIPP)
- Leveraging auxiliary data
  - Administrative records
  - Other surveys (e.g., American Community Survey)
- Formal privacy and SAE
  - Area-level (e.g., spatial change-of-support)
  - Unit-level (e.g., DP partially synthetic data)

Identifying Specific States or Localities

A third situation occurs when there is a need to identify specific states or localities, such as when the data user wishes to measure the impact of a policy or program by making use of differences across states or localities

- Data users will wish to access the raw data on geography.
  - Secure Online Data Access (SODA)
  - Federal Statistics Research Data Center (FSRDC)
- These provide another path forward for producing small area estimates.
- In conjunction with the public-use files, SODA and FSRDCs can be thought of as a tiered system of access, where the specific “tier” depends on the desired analysis.

Recommendation

Recommendation: The Census Bureau should continue to pursue the development of a small area estimation program to meet the need of SIPP users for geography-based analysis that preserve confidentiality and limits disclosure risk.

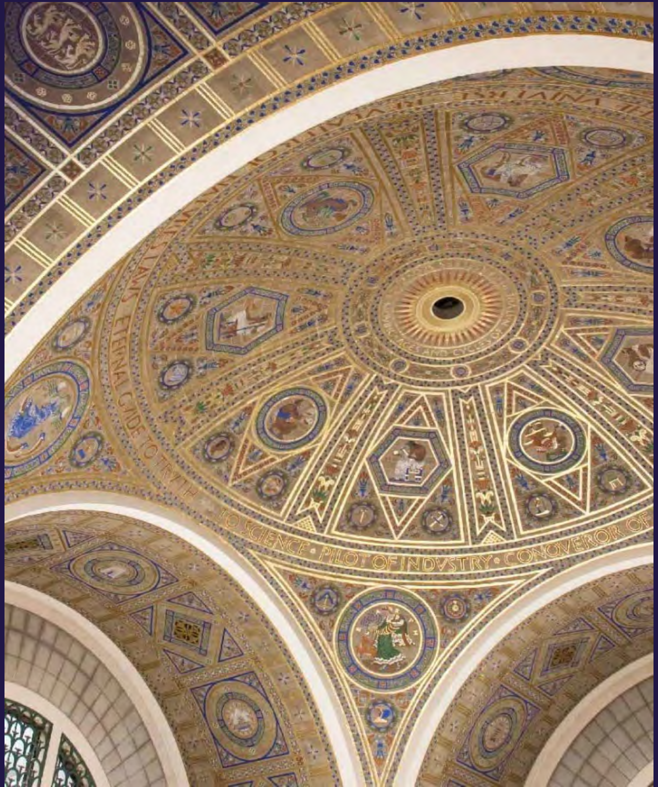
Thank You

For more information, please contact:

Scott Holan

holans@missouri.edu

Timeline


ETERNAL GUIDE TO TRUTH
TO SCIENCE • PILOT OF INDUSTRY • CONQUEROR OF

06

The Timeline

Implementing some of these recommendations will take time, while there is a need to continue providing access to SIPP data. Thus, the changes should be phased in, with preparatory work started early.

- In the immediate term, the Census Bureau might take traditional steps such as coarsening the data, and/or add a registration process and user agreement for those who download the public use data. Some data might be provided in synthetic form. However, these steps should not affect usability of the data.
- Secure online data access is probably the easiest to implement quickly, though decisions will need to be made on whether to use a private contractor, what application process will be needed, and what disclosure review will be conducted.
- A table generator will be easier to construct than a complete remote analysis platform and may provide preliminary data on the demand for such a tool.
- The report provides 6 milestones, with tasks to be accomplished within each milestone.

Milestones

- **Milestone 1**
  - Risk assessment plan
  - Communication plan
  - Assess use of table generator
  - Plan for SODA

- **Milestone 2**
  - Implement risk assessment
  - Continue public-use file
  - Design flexible table generator
  - Begin developing SODA
  - Plan for external partners
  - Coordinate with other federal agencies

- **Milestone 3**
  - Continue refining public-use file
  - Implement flexible table generator
  - Explore partial synthetic file
  - Implement SODA

- **Milestone 4**
  - Obtain feedback from data users
  - Expand table generator to remote analysis platform

- **Milestone 5**
  - Add formal privacy to remote analysis platform

- **Milestone 6**
  - Partially synthetic file with automated verification and validation server
  - Decide whether to adjust remote analysis platform to use partially synthetic file

Thank You

For more information, please contact:

Brad Chaney

Bchaney@nas.edu

Report available at:

https://doi.org/10.17226/27169