Privacy Protection at Statistics of Income, IRS

Barry W. Johnson

Statistics of Income Director &

Acting Chief Research and Analytics Officer

Current State

- Official IRS Publication 1075 specifies minimum cell size rules for tabulations
  - Cell sizes vary with granularity of the tables.
  - Additional rules apply to small geographic units.
- Individual Income Tax Public-Use File uses a combination of more sophisticated techniques to protect data.
- Partnering with Urban Institute to produce fully synthetic public-use files.
- Developing a pilot to introduce formal privacy techniques.
- We must continue to evolve.

The content of this presentation is the opinion of the writer(s) and does not necessarily represent the opinion of the Internal Revenue Service

Formal Privacy – Data users

- Users may fear data will not be as good as those produced using traditional methods.
- But are they fully aware of the impact of current practices on data?

Micro data:

- Suppression of key variables
- Removing extreme records
- Imputing ‘sensitive’ records or data items
- Adding random noise
- Collapsing categorical responses
- Swapping ‘similar’ records
- Top coding/recoding into intervals
- Blurring

Tabular Data:

- Rounding
- Cell suppression and complementary suppressions
- Dominance rules
- Application of micro data methods prior to tabulation

The content of this presentation is the opinion of the writer(s) and does not necessarily represent the opinion of the Internal Revenue Service

Simple Cell Suppression

Original table

Industry	Number of returns	Total assets
All industries	32	276
Mining	**4**	**120**
Utilities	6	45
Construction	7	32
Manufacturing	15	79

Disclosure adjusted

Industry	Number of returns	Total assets
All industries	32	276
Mining	**	**
Utilities	**10	**165
Construction	7	32
Manufacturing	15	79

- Requiring a minimum of 5 observations per cell, 20 percent of data are sensitive, but 40 percent of the data are lost due to complimentary suppressions.
- Should we do more to raise public awareness of data loss in products produced using legacy disclosure limitation methods?

The content of this presentation is the opinion of the writer(s) and does not necessarily represent the opinion of the Internal Revenue Service

Formal Privacy – Practical Considerations

Formal privacy tools are still evolving:

- Tools are not well developed for data that are not uniformly distributed.
- Methods for dealing with data that have complex accounting relationships are particularly challenging:
  - Numbers need to add up.
  - Tax data are subject to qualification thresholds and phaseouts that depend on multiple characteristics (for example married/single, size of adjusted gross income).

Apportioning the Privacy Budget:

- Tax data are used across the IRS to support research.
- External Researchers also use tax data: Census, Treasury, Congress, academics.

The content of this presentation is the opinion of the writer(s) and does not necessarily represent the opinion of the Internal Revenue Service

Formal Privacy – Staffing and Resource Impacts

- Implementing formal privacy methods will require infrastructure updates:
  - Staff need to learn new skills.
  - External partners with the necessary skills to help train staff and lead research will be essential.
  - Legacy disclosure limitation process will require retooling.
  - For some products, validation servers and/or tiered access models may be desirable.
- Tight budgets and limited ability to hire will require agencies to work together to make progress.

The content of this presentation is the opinion of the writer(s) and does not necessarily represent the opinion of the Internal Revenue Service

Formal Privacy – Legal Framework

Many privacy laws treat privacy protection as binary

Internal Revenue Code

- Section 6103: “The term “disclosure” means the making known to any person, in any manner whatever, a [tax] return or [tax] return information*”
- Section 6108: “No publication or other disclosure of statistics or other information ……shall in any manner permit the statistics, study, or any information so published, furnished, or otherwise disclosed to be associated with, or otherwise identify, directly or indirectly, a particular taxpayer.”

*Return information means a taxpayer’s identity (name, mailing address, identifying number); the nature, source, or amount of income, deductions, exemptions, credits, assets, liabilities, net worth, tax liability, tax withheld, deficiencies, overassessments, or tax payments; whether or not the return is being examined or subject to other investigation or processing; etc.

The content of this presentation is the opinion of the writer(s) and does not necessarily represent the opinion of the Internal Revenue Service

Formal Privacy – Legal Framework (continued)

Foundations for Evidence-Based Policymaking Act of 2018

- [Agencies must] ensure that individuals or organizations who supply information under a pledge of confidentiality to agencies for statistical purposes will [not] have that information disclosed in identifiable form.
- The term 'identifiable form' means any representation of information that permits the identity of the respondent to whom the information applies to be reasonably inferred by either direct or indirect means.
- Even if legal standards are met, perceptions matter.

The content of this presentation is the opinion of the writer(s) and does not necessarily represent the opinion of the Internal Revenue Service

Formal Privacy – Additional Considerations

- Under Differential Privacy, the trade off between disclosure risk, ϵ, and utility is explicit and ideally revealed to the public.
- Absent an accepted model of data utility, how will ϵ be set?
- How will legislators react?
  - Will they be willing to update laws to recognize this trade off?
  - Will they be willing to increase appropriations to support tiered access or other secure access models?
- How will potential survey respondents and those in the public whose administrative data are used for statistics react to being told that their data are not 100 percent protected?
- How will agencies balance finite "privacy budgets" and initiatives, such as the Federal Data Strategy and the Evidence Act that seek to INCREASE the use of data for evidence building?
- Will data users accept new products?
- Like all significant change, we have much work to do.

The content of this presentation is the opinion of the writer(s) and does not necessarily represent the opinion of the Internal Revenue Service

Research, Analysis & Statistics

STATISTICS OF INCOME

Barry.W.Johnson@irs.gov

Thank you!

The content of this presentation is the opinion of the writer(s) and does not necessarily represent the opinion of the Internal Revenue Service