Statistical Journal of the IAOS 34 (2018) 99–103
DOI 10.3233/SJI-170418
IOS Press

The SOI Databank: A case study in leveraging administrative data in support of evidence-based policymaking

Raj Chetty[a], John N. Friedman[b], Emmanuel Saez[c] and Danny Yagan[c,*]
[a]*Department of Economics, Stanford University, Stanford, CA 94305, USA*
[b]*Department of Economis, Robinson Hall, Brown University, Providence, RI 02912, USA*
[c]*Department of Economics, University of California-Berkeley, Berkeley, CA 94720, USA*

Abstract. Full-population administrative tax data collected for program administration holds potential for facilitating statistical work that previously had been infeasible. This article describes a framework – the SOI databank – for that facilitation. The Databank is a statistical database that comprises income and tax information from income tax returns and several information returns, along with links to children and employers. The Databank structure and process described in this article may provide a useful model for other government agencies facing similar challenges and opportunities with administrative data.

Keywords: Administrative data, statistical databases

1. Introduction

This article describes a framework – the SOI Databank – for transforming full-population administrative tax data collected for program administration into a statistical database for approved statistical analyses on a secure server. The Databank is a de-identified balanced panel of all U.S. individuals 1996–2015. It comprises income and tax information from income tax returns and several information returns, along with links to children and employers. The Databank has facilitated statistical work that previously had been infeasible. This article describes the opportunities presented by digitized full-population tax records, four impediments to realizing those opportunities, how the Databank addresses those problems, examples of the Databank's use, and how the Databank may be a useful template for other government agencies.

2. The opportunity

The United States Congress has charged the Internal Revenue Service (IRS), Department of the Treasury, Bureau of Economic Analysis, National Agricultural Statistical Service and Census Bureau to use administrative tax records for various statistical purposes.[1] The IRS Statistics of Income Division (SOI) created structured mechanisms for painstakingly transforming administrative tax records into edited statistical files (traditional SOI files) that are ready for analysis. Those traditional SOI files have facilitated pathbreaking and essential statistical work for decades [8].

Traditional SOI files on individuals (rather than businesses) have been based exclusively on stratified random samples of U.S. Individual Income Tax Returns, IRS Form 1040. Using Forms 1040 as the source for these samples has implied that the traditional files have omitted low-income individuals who were not re-

*Corresponding author: Danny Yagan, Department of Economics, University of California-Berkeley, Berkeley, CA 94720, USA. E-mail: yagan@berkeley.edu.

[1] U.S. Internal Revenue Code Sections 6103(j) and 6108.

quired to file income tax returns [5] and have had limited ability to link individuals to employers. Because the files are based on random samples, the traditional files have had a limited ability to link individuals across years and within families and also have left too few rows to analyze narrow geographies.[2]

The collection of digitized tax records on the full population makes it possible to address those limitations with limited data infidelities. In order to administer the tax system, the IRS digitizes data from the universe of Forms 1040, which comprised more than 150 million returns in 2016 [6], along with thirty types of information returns like U.S. Wage and Tax Statement, Form W-2, which is filed by employers to report wages, tips, and other compensation paid to employees as well as withheld income taxes.[3] Other information returns report income earned on investments, pension income and social insurance payments, as well as interest paid by taxpayers on mortgages and other types of debt. In 2016, the IRS received almost 3 billion information returns, including more than 250 million Forms W-2 [7]. Importantly, the information returns provide information for non-filers, primarily individuals whose total income was below the Federal income tax filing requirement for a given year, and can be used to link individuals and employers. The full-population digitization makes it possible to link individuals across years and within families and provide enough rows to analyze narrow geographies. Their collection for administering the tax system – including almost 80 percent of returns submitted via electronic filing and therefore subject to sophisticated validation procedures – means that the full-population records have limited and somewhat predictable data infidelities.

3. Four impediments

However, digitized full-population tax records were collected for program administration, not for statistical analysis. As a result, there are four impediments to using those records for statistical analysis:

1. *Inconsistent units of observation*. Form 1040 returns are at the level of the tax filing unit, which can comprise either a single individual or a married couple. In 2016, approximately 54 million, or (36 percent of Forms 1040) were for married couples filing jointly [6]. In contrast, information returns are generally at the individual level. Hence, decisions must be made when one wants to use data from both types of returns.
2. *Duplicate records and other complications of information returns*. Many types of information returns can have duplicate records (e.g., multiple W-2 records for a given individual-employer-year), often representing amended returns. Some records have erroneous values (e.g., an invalid masked taxpayer identification number). Finally, information returns are often most useful when aggregated to the individual-year level (e.g., an individual's annual W-2 wages aggregated across multiple employers), which then makes it difficult to preserve linkages (e.g., to an individual's employer or employers). Hence, decisions must be made before utilizing information returns.
3. *Choice of linkages*. Numerous types of linkages exist in the full-population data. For example, tax returns include information on spouses and dependents, which include children and qualified relatives; information documents include information on employers as well as financial and educational institutions. Hence, decisions must be made regarding which specific linkages are most useful.
4. *Processing burden*. The IRS is the repository for a large volume of information on U.S. taxpayers, drawn from more than 30 different administrative data sources. Currently, these encompass almost 3,000 relational data tables requiring approximately 2,500 terabytes of storage. Researchers access the data primarily using SAS, SQL, R, Stata, Hyperion, and ArcGIS. It is inefficient for each end user to process billions of records for each new statistical use due to the enormous processing time and systems processing load. Hence, it is efficient to prepare a statistical database that incurs the bulk of the processing burden once, while permitting maximum flexibility to end users for customized uses.

[2] SOI has produced several prospective individual income tax panel data sets, but due to the sample size, attrition, and changes in filing status for panel members over time, may not be suitable for some types of research [1].

[3] Every employer engaged in a trade or business who pays remuneration, including noncash payments of $600 or more for the year (all amounts if any income, social security, or Medicare tax was withheld) for services performed by an employee must file a Form W-2 for each employee (even if the employee is related to the employer) from whom income, Social Security, or Medicare tax was withheld.

4. A solution

The Databank was created to address the four impediments to realizing the statistical opportunities of digitized full-population tax records.

The Databank is a de-identified balanced panel of all U.S. individuals 1996–2015 drawn from billions of tax returns.[4] Specifically, there are twenty rows – one for each year 1996–2015 – for each U.S. individual who was issued a Social Security Number and who has not been recorded as deceased before 1996 in Social Security Administration records. There are over one hundred columns, each containing an income, tax, link, or similar value pertaining to the individual-year row. In total, the Databank consists of more than 9 billion data rows.

The Databank addresses the four impediments above as follows:

1. *Inconsistent units of observation*. The Databank is organized at the individual-year level, following the organization of information returns. Variables from Forms 1040 are included on the row of the primary filer and, in the case of married-filing-jointly returns, on the row of the secondary filer as well. For example, consider two individuals A and B filing married-filing-jointly in 2015 with $100,000 in adjusted gross income (AGI). A's 2015 Databank row will have an AGI value of $100,000, and B's 2015 row will also have an AGI value of $100,000. Each row contains the filing status corresponding to the Form 1040 used to populate the row's 1040 variables. These pieces of information allow end users the flexibility to apportion the income amounts reported on jointly filed Forms 1040 to each individual spouse based on research needs. Non-filers have missing values for Form 1040 values in that year.
2. *Duplicate records and other complications of information returns*. Each type of return is processed in a customized way before inclusion in the Databank. When duplicate records are found, the most recent one typically is chosen. For example, when multiple W-2 records for a given individual-employer-year are encountered, the most recently posted one is chosen. Records with invalid taxpayer identification numbers are excluded. Relevant dollar values are aggregated to the individual-year level. For example, for individuals with Forms W-2 from multiple employers in a given year, W-2 wages are summed across those Forms W-2 before inclusion in the Databank. However, in order to facilitate linkages between individuals and employers, the employer's masked TIN from each individual's highest- and second-highest-paying W-2 are included in the Databank as well. It is important to note that data errors remain, so end users must apply safeguards in their analyses as appropriate. Users requiring official aggregates must continue to use the traditional edited SOI stratified random samples.
3. *Choice of linkages*. The Databank includes variables that permit end users to make three types of linkages: spouses, parents and children, and (as already mentioned) firms and workers. Each individual who filed a married-filing-jointly or married-filing-separately return in a given year contains the masked TIN of the individual's spouse as listed on that individual's Form 1040 in that year. Pooling claimed dependent children across all years 1996–2015, time-invariant parent-children linkages are made – assigning each individual to up to two parents according to the first Form 1040 1996–2015 on which the individual was claimed as a dependent child. The Databank contains those parent-children linkages: the masked TIN of up to six children are listed on each parent's Databank rows.
4. *Processing burden*. The Databank is updated once annually and hosted on a secure server that can be accessed simultaneously by approved users. Content is carefully considered to balance utility against additional storage requirements and processing burden. Priority is given to items with the potential to benefit a wide range of research questions or that permit linkages to more detailed microdata stored in other relational data tables. The Databank includes a random number that can be used to draw samples as small as 0.01 percent of the filing population in any given year, to minimize processing time and system load. This feature allows users to draw cross-sectional or *ad hoc* panel samples for intensive analysis of records sharing desired characteristics. Often sample records are then linked to other IRS data tables to gather additional detailed tax and information document data required to support specific research projects.

[4] Databank records do not include personally identifiable information. Each taxpayer in the database is assigned a unique identifier, known as a masked taxpayer identification number (masked TIN), to facilitate record linkage.

Moreover, the Databank is not a set-in-stone static framework. Instead, the Databank employs a user-driven data governance model for improving the framework. Approved users across government agencies are solicited annually for feedback on how to improve the existing Databank framework and best to expand the content. For example, an ongoing effort attempts to incorporate new information return fields in order to implement SOI, Treasury, and Joint Committee on Taxation methods for imputing Forms 1040 for non-filers. The Databank also includes access to extensive metadata.

5. Usage examples

The Databank has been used for statistical analysis in several contexts that require linkages, information on non-filers, and the ability to isolate narrow geographies – including to utilize "natural experiments" to estimate causal impacts of interest. Here are two examples.

First, an important question in tax policy and administration is how much does the Earned Income Tax Credit (EITC) affect individuals' labor market earnings. Chetty et al. [2] used the first version of the Databank to answer this question. Their analysis required three statistical features that the Databank could provide. First, they utilized variation across local areas in knowledge of the EITC, so they required the Databank's large sample size and narrow geographies (based on the ZIP code reported on the Form 1040). Second, they required measurement of individual wages, as reflected in W-2 wage earnings. Third, and because collusion between employers and employees could in principle lead to misreported W-2 earnings, the authors had to replicate their results in the subsample of workers at large employers where collusion is least likely, which required the individual-employer linkages. They also used links between parents and children.

With these data, the authors were able to analyze a compelling natural experiment to identify the causal impact of the EITC on U.S. wage earnings. EITC refunds are a function of a household's number of children. In particular, the refunds rise substantially after the birth of a first child and a second child but not a third child. They then compared wage earnings changes of parents in areas with high EITC knowledge to those in other areas. Individuals in high-knowledge areas changed their wage earnings substantially after the births of first and second children and received larger EITC refunds than those in other areas, but not after third children. Thus, the authors uncovered substantial impacts of the EITC on U.S. wage earnings. This fine-grained analysis of a natural experiment – naturally occurring variation in EITC knowledge and eligibility, based on birth order and geographical location – was enabled by the richness and size of the Databank.

Second, a central priority of SOI and the Treasury Department is the measurement of income distribution. A key question of income distribution is to what degree is unequal income distribution persistent across generations. For example, what percentage of children born to parents in the bottom 20 percent of the income distribution end up reaching the top 20 percent of the income distribution as adults? Chetty et al. [3,4] used the Databank to link parent income rank and child income rank and found substantial persistence. In particular, 7.5 percent of children born to parents in the bottom 20 percent reach the top 20 percent as adults – which is closer to the 0 percent full-persistence benchmark than to the 20 percent no-persistence benchmark. Moreover, the Databank's ZIP code information allowed the authors to identify that some local areas exhibit much less inequality persistence across generations than others. Again, the richness and size of the Databank facilitated this work.

6. Conclusion

The advent of digitized full-population tax data provided an unprecedented opportunity for statistical work on individuals who do not file Form 1040 income tax returns and on individuals across time, across generations, with certain types of employers, and at narrow geographical levels with limited data infidelities. However, data collected for tax administration do not come ready-made for statistical analysis. They require a framework.

The SOI Databank is such a framework. It addresses four impediments to utilizing digitized full-population tax records for statistical analysis. It does so in a way that makes careful choices to prepare those data for analysis – saving end users the processing burden of doing so themselves while preserving maximal flexibility for end users' analyses. The Databank enjoys a data governance structure for soliciting feedback from users and improving the framework in annual updates.

Governments around the world are seeking to increase the use of data collected to administer government programs for statistical and research purposes. These efforts are driven by increased data collection costs associated with traditional sources, such as surveys and censuses, the desire to reduce respondent burden, and the growth of computing technology and software tools associated with the "Big Data" revolution. The Databank process and structure described here may be a useful model for other government agencies facing similar challenges and opportunities.

Acknowledgments

We thank Barry Johnson for helpful comments.

References

[1] Bryant V. 2008. Attrition in the Tax Years 1999–2005 Individual Income Tax Return Panel. *Proceedings of the American Statistical Association Section on Government Statistics Section.* http://ww2.amstat.org/sections/SRMS/Proceedings/y2008/Files/300929.pdf.

[2] Chetty R, Friedman JN, Saez E. 2013. Using Differences in Knowledge Across Neighborhoods to Uncover the Impacts of the EITC on Earnings. *American Economic Review* 103(7): 2683-2721.

[3] Chetty R, Hendren N, Kline P, Saez E. 2014. Where Is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States. *Quarterly Journal of Economics* 129(4): 1553-1623.

[4] Chetty R, Hendren N, Kline P, Saez E, Turner N. 2014. Is the United States Still a Land of Opportunity? Recent Trends in Intergenerational Mobility. *American Economic Review Papers and Proceedings* 104(5): 141-147.

[5] Johnson B, Moore K. 2004. Consider the Source: Differences in Estimates of Income and Wealth from Survey and Tax Data. *Statistics of Income Working Papers*. https://www.irs.gov/pub/irs-soi/johnsmoore.pdf.

[6] *Statisics of Income – 2015 Individual Income Tax Returns* (2017) Internal Revenue Service Publication 1304, Washington, D.C. 20224.

[7] *Statistics of Income – Calendar Year Projections of Information and Withholding Documents for the United States and IRS Campuses* (2017) Internal Revenue Service Publication 6961, Washington, D.C. 20224.

[8] Wilson R. 1988. Statistics of Income: A By-Product of the U.S. Tax System, *Statistics of Income Bulletin*, Internal Revenue Service, Washington, D.C.