Synthetic Micro-data Based on Integrating the Survey of Income and Program Participation with Lifetime Earnings and Retirement Benefits Data

Martha Stinson, John Abowd, Gary Benedetto

Outline

- Introduction and history
- A little background
- Sources for the “gold standard” data
- Synthetic data methods
- Assessing analytical validity
- Assessing disclosure avoidance
- The program for user feedback and improvement
- Conclusion

A Little History

- Motivated by a desire to allow research access to long administrative record histories of earnings and benefits with household demographic data integrated
- These data allow detailed statistical and simulation study of retirement and disability programs
- Census Bureau, Social Security Administration, Internal Revenue Service, and Congressional Budget Office all participated in development

SIPP/SSA/IRS Synthetic Beta

- Links 1990-1993; 1996 SIPP panels with complete lifetime earnings and benefit histories from IRS and SSA
- Only a few unsynthesized variables: sex, marital status, benefit status (initial and 2000)
- Biggest analytical validity challenge: preserving multivariate relationships across 616 variables
- Biggest disclosure avoidance challenge: requirement that users not be able to re-identify source records in existing SIPP public use files

Create Gold Standard Data

- Create a data extract from the SIPP panels conducted in the 1990s
  - Five panels: 1990, 1991, 1992, 1993, 1996
  - Data from core and topical module survey questions
- Standardize variables across panels
- Link to earnings and benefits administrative data
- These data are the “gold standard”
- Any synthetic data must preserve the characteristics of and relationships among the variables on this file

Synthetic Data Creation

- Purpose of synthetic data is to create micro data that can be used by researchers in the same manner as the original data while preserving the confidentiality of respondents' identities
- Fundamental trade-off: usefulness and analytical validity of data versus disclosure avoidance
- Our goal: not be able to re-identify anyone in the already released SIPP public use files while still preserving regression results

Multiple Imputation Confidentiality Protection

- Denote confidential data by Y and non-sensitive data by X.
- Both Y and X may contain missing data, so that $Y=(Y_{obs}, Y_{mis})$ and $X=(X_{obs}, X_{mis})$.
- Assume database can be represented by joint density $p(Y,X,\theta)$.
- Construct multiple samples from posterior predictive distribution given the observed data
- Bayesian bootstrap (Rubin 1981) and Sequential Regression Multivariate Imputation (Raghunathan et al. 2003) used as primary methods for sampling from PPD

Sequential Regression Multivariate Imputation Method

- Synthetic data values *Y* are draws from the posterior predictive density:

$$p\left(\tilde{Y} \mid Y_{obs}, X_{obs}\right) = \int p\left(\tilde{Y} \mid Y_{obs}, X_{obs}, \theta\right) p\left(\theta \mid Y_{obs}, X_{obs}\right) d\theta$$

- In practice, use a two-step procedure:
  1) draw m completed datasets using SRMI (imputes values for all missing data)
  2) draw r synthetic datasets for each completed dataset from predictive density given the completed data

Creating Synthetic Data

- Exact relationships preserved by using a nine-level parent-child tree to describe all conditions that must hold exactly among the variables
- Initialize data completion with Bayesian bootstrap to complete missing administrative data
- Use SRMI or BB at each level of the parent-child tree to complete other variables
- Use each of the four completed data implicates to create a four synthetic implicates
- Total of 16 implicates
- Combining rules from Reiter (2004)

Testing Analytical Validity

- Univariate distributions overall and within sub-domains
- Up to four-way interactions overall and within sub-domains
- Covariance structures overall and within sub-domains
- Regression analyses of important variables

Log Initial Monthly Benefit Amount

Table 49: Log of initial MBA for retired individuals (TOB_initial=1)

	Coefficient		Confidence Interval				Standard Error	
Explanatory Variables	Synthetic	Completed	Synthetic		Completed		Synthetic	Completed
Intercept	-61.534	-67.501	-66.344	-56.725	-71.471	-63.531	2.501	2.192
age_initial_entitle	0.033	0.038	0.028	0.038	0.033	0.044	0.003	0.003
blackfemale	-0.360	-0.329	-0.435	-0.285	-0.386	-0.272	0.036	0.030
blackmale	-0.110	-0.120	-0.150	-0.071	-0.150	-0.089	0.015	0.018
whitefemale	-0.301	-0.297	-0.364	-0.238	-0.354	-0.240	0.027	0.026
highschool_only	0.070	0.061	0.042	0.097	0.026	0.096	0.014	0.017
somecollege	0.121	0.089	0.078	0.163	0.054	0.124	0.020	0.018
college_only	0.164	0.124	0.143	0.184	0.080	0.168	0.011	0.022
graduate	0.191	0.147	0.150	0.232	0.119	0.175	0.020	0.016
disab	-0.048	-0.039	-0.076	-0.021	-0.067	-0.011	0.013	0.015
hispanic	-0.098	-0.058	-0.161	-0.035	-0.124	0.009	0.029	0.032
divorced	0.114	0.132	0.069	0.159	0.098	0.166	0.023	0.020
married	0.078	0.052	0.052	0.104	0.019	0.085	0.015	0.019
widowed	0.179	0.162	0.146	0.213	0.126	0.197	0.020	0.021
log_totnetworth	0.015	0.046	0.005	0.025	0.037	0.055	0.005	0.005
ser_pct_yrs_wrked	1.052	1.044	0.609	1.496	0.689	1.398	0.187	0.151
year_initial_entitle	0.033	0.035	0.030	0.035	0.033	0.037	0.001	0.001

Log Total Earnings (Uncapped)

Table 42: Log of Total DER Earnings in year 2000 for white males

Explanatory Variables	Coefficient		Confidence Interval				Standard Error	
	Synthetic	Completed	Synthetic		Completed		Synthetic	Completed
Intercept	8.373	7.717	8.295	8.451	7.658	7.777	0.046	0.036
highschool_only	0.180	0.248	0.157	0.203	0.221	0.274	0.013	0.016
somecollege	0.405	0.491	0.284	0.527	0.463	0.520	0.051	0.017
college_only	0.719	0.834	0.561	0.876	0.802	0.865	0.063	0.019
graduate	0.790	1.043	0.628	0.952	1.008	1.078	0.064	0.021
disab	-0.259	-0.470	-0.296	-0.222	-0.511	-0.428	0.022	0.024
foreign_born	0.075	0.097	0.032	0.117	0.062	0.132	0.025	0.020
hispanic	-0.008	0.043	-0.049	0.033	0.007	0.079	0.022	0.021
ser_totyrs_2000	0.104	0.216	0.039	0.169	0.201	0.230	0.025	0.009
ser_totyrs_2000_2	-0.038	-0.119	-0.091	0.014	-0.131	-0.107	0.020	0.007
ser_totyrs_2000_3	0.009	0.032	-0.007	0.025	0.029	0.036	0.006	0.002
ser_totyrs_2000_4	-0.001	-0.003	-0.003	0.001	-0.004	-0.003	0.001	0.000

Chart 3:
Comparison of Synthetic and Completed Earnings
Retired White Males and Females

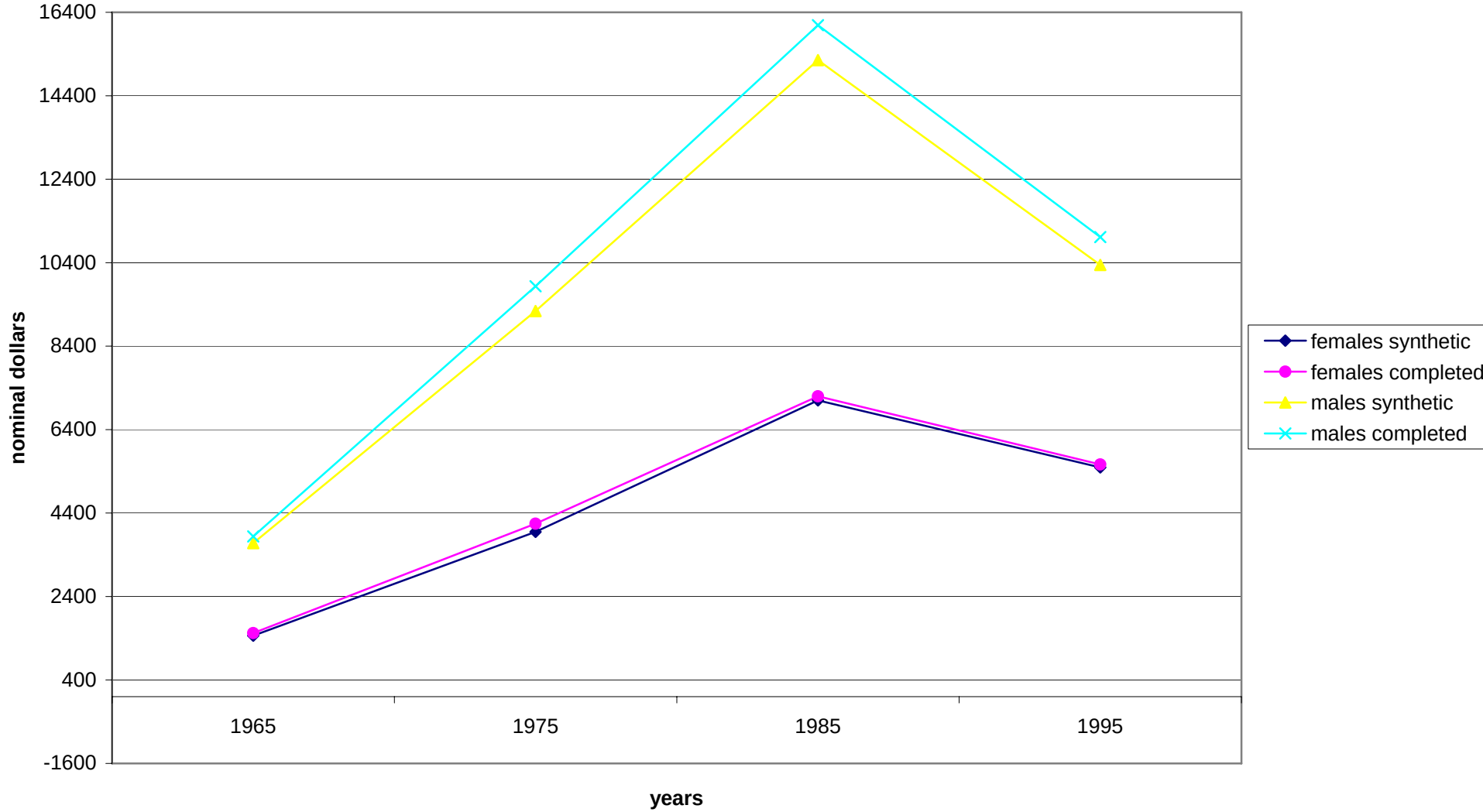
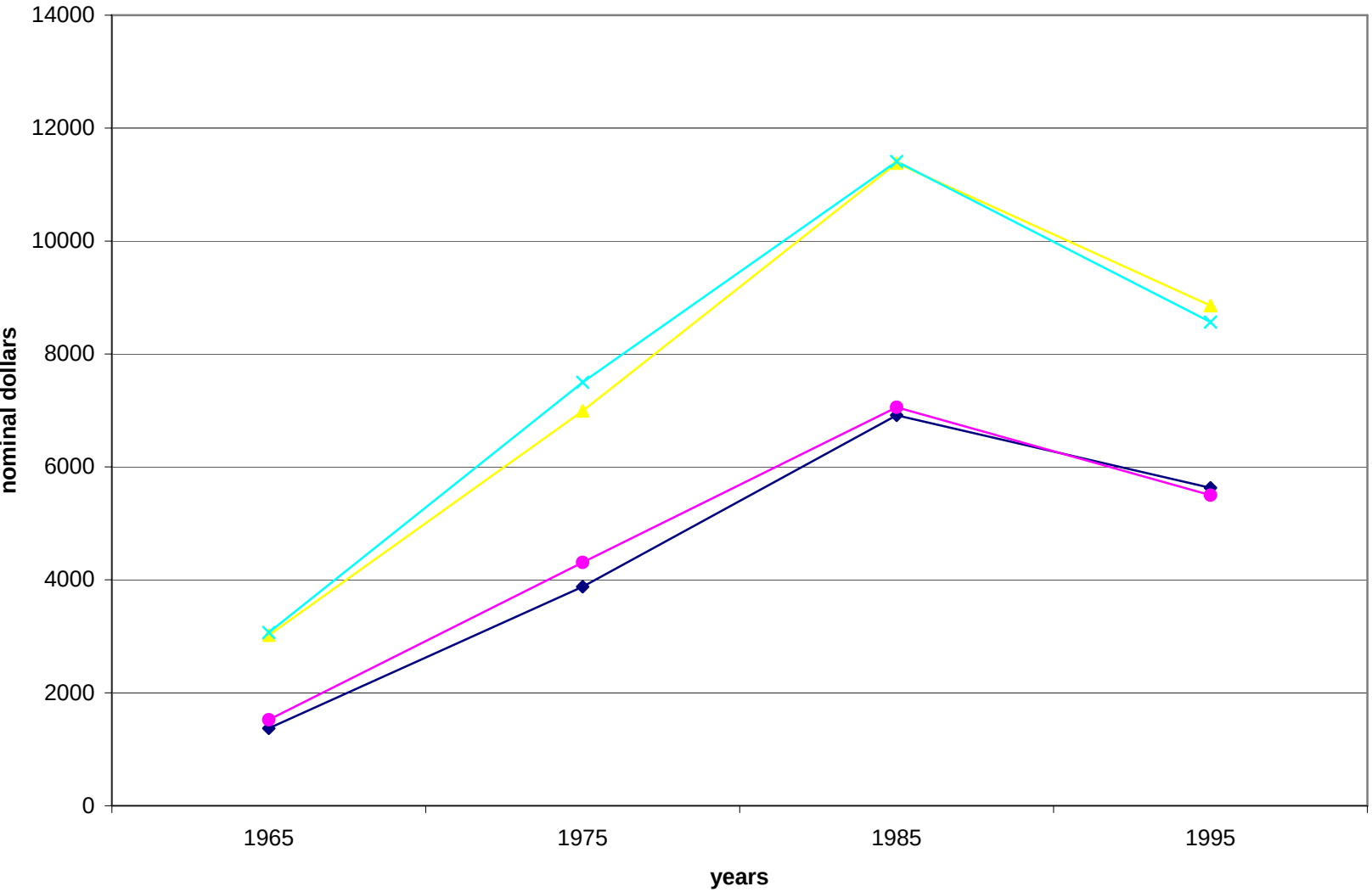


Chart 4:
Comparison of Synthetic and Completed Earnings
Retired Black Males and Females



Confidentiality Protection

- Protection is based on the inability of synthetic data users to re-identify the SIPP record upon which the synthetic record is based
- This prevents wholesale addition of SIPP data to the IRS and SSA data in the synthetic files
- Goal: best record linkage of SIPP records from the SIPP Synthetic Beta to existing public use files should produce mostly false matches

Disclosure Avoidance Analysis

- Uses probabilistic and distance-based record linking
- Each synthetic implicate is matched back to the gold standard
- All unsynthesized variables are used as blocking variables
- Different matching variable sets are used, including using all synthesized variables in the match list

Table 3: Match rates for Married and Single Individuals using Probablistic Record Linking

Married Individuals

Type of Match	Total True	Total Records	Match Rate	Ratio of 2 to 1	Ratio of 3,2 to 1
Highest scoring	4418	134662	0.0328	0.4033	0.7558
Second highest scoring	1782	134662	0.0132		
Third highest scoring	1557	134662	0.0116		

Single Individuals

Type of Match	Total True	Total Records	Match Rate	Ratio of 2 to 1	Ratio of 3,2 to 1
Highest scoring	4147	129132	0.0321	0.3053	0.5889
Second highest scoring	1266	129132	0.0098		
Third highest scoring	1176	129132	0.0091		

Using the SIPP Synthetic Beta

- Full version (16 implicates) released by Census Bureau, SSA and IRS to the Cornell Virtual RDC
- Any researcher may use these data
- During the testing phase, all analyses must be performed on the Virtual RDC
- Census Bureau research team will run the same analysis on the completed confidential data
- Results of the comparison will be released to the researcher, Census Bureau, SSA, and IRS (after traditional disclosure avoidance analysis of the runs on the confidential data)

Survey of Income and Program Participation (SIPP) - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.sipp.census.gov/sipp/

Cornell University We... 7-Day Forecast for L... Main Page - CISERWiki Main Page - ISS Netw... LEHD Staff Only www.cyclingnews.co...

7-Day Forecast ... Codian MCU - C... Videoconferenci... VirtualRDC Proj... City Creek (Salt... OpenTable Survey of I...

U.S. Census Bureau

Survey of Income and Program Participation

SIPP

- Introduction to SIPP
- SIPP Survey Content
- Technical Information
- Using & Linking Files
- SIPP Publications
- Access SIPP Data
- Access SIPP Synthetic Data

- User Notes/ ListServe/News
- SIPP Users' Guide
- SIPP Tutorial
- Technical Documentation
- SIPP Help

Dynamics of Economic Well-being System

Contact DEWS

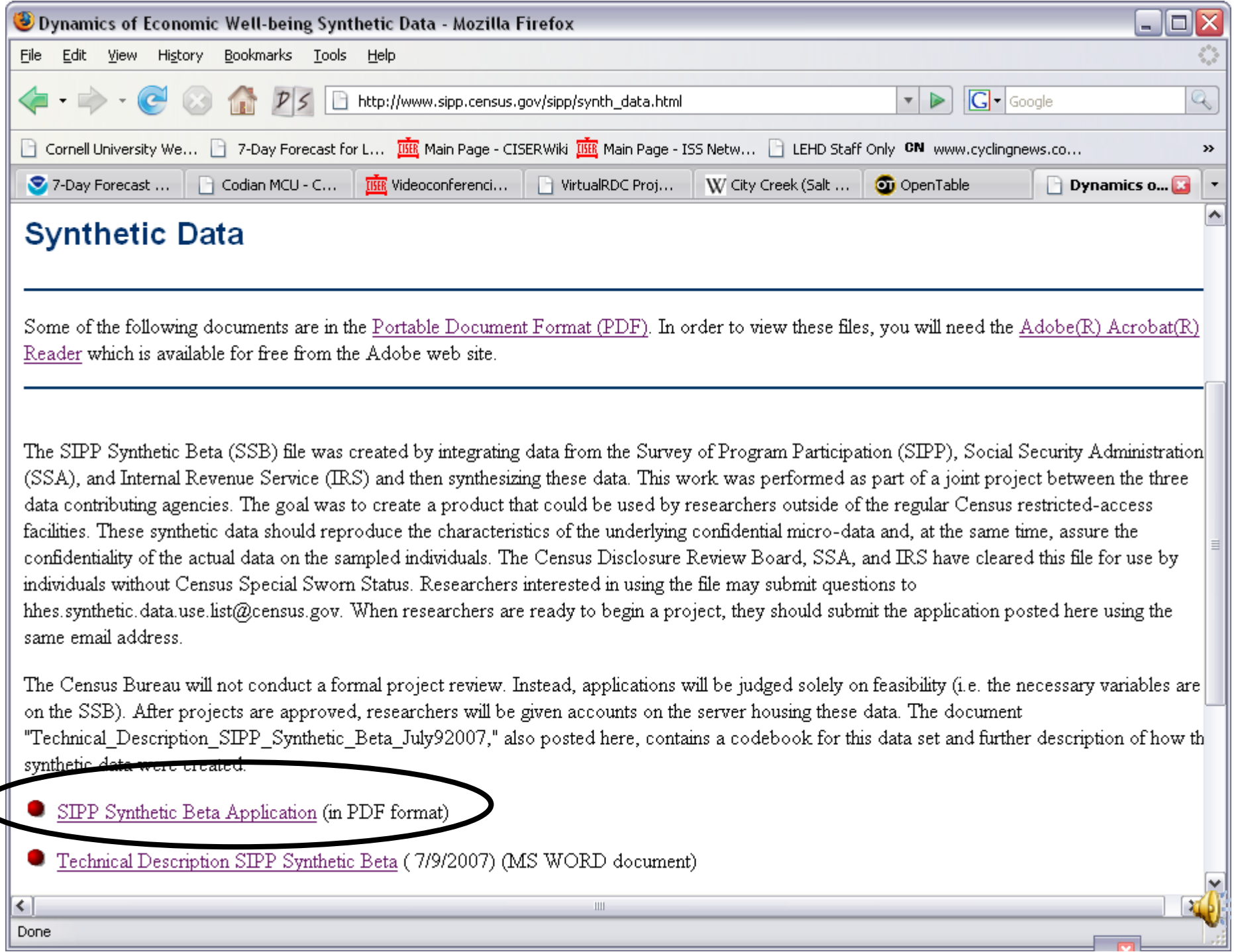
URL: http://www.sipp.census.gov/sipp/

Source: U.S. Census Bureau, Demographics Survey Division, Survey of Income and Program Participation branch
Created: February 14, 2002
Last revised: July 16, 2007

Census Bureau Links: Home • Search • Subjects A-Z • FAQs • Data Tools • Catalog • Census 2000 • Quality • Privacy Policy • Contact Us

USCENSUSBUREAU

Done


Dynamics of Economic Well-being Synthetic Data - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.sipp.census.gov/sipp/synth_data.html
Google
Cornell University We...
7-Day Forecast for L...
Main Page - CISERWiki
Main Page - ISS Netw...
LEHD Staff Only
www.cyclingnews.co...
7-Day Forecast ...
Codian MCU - C...
Videoconferenci...
VirtualRDC Proj...
City Creek (Salt ...
OpenTable
Dynamics o...
Synthetic Data
Some of the following documents are in the Portable Document Format (PDF). In order to view these files, you will need the Adobe(R) Acrobat(R) Reader which is available for free from the Adobe web site.
The SIPP Synthetic Beta (SSB) file was created by integrating data from the Survey of Program Participation (SIPP), Social Security Administration (SSA), and Internal Revenue Service (IRS) and then synthesizing these data. This work was performed as part of a joint project between the three data contributing agencies. The goal was to create a product that could be used by researchers outside of the regular Census restricted-access facilities. These synthetic data should reproduce the characteristics of the underlying confidential micro-data and, at the same time, assure the confidentiality of the actual data on the sampled individuals. The Census Disclosure Review Board, SSA, and IRS have cleared this file for use by individuals without Census Special Sworn Status. Researchers interested in using the file may submit questions to hhes.synthetic.data.use.list@census.gov. When researchers are ready to begin a project, they should submit the application posted here using the same email address.
The Census Bureau will not conduct a formal project review. Instead, applications will be judged solely on feasibility (i.e. the necessary variables are on the SSB). After projects are approved, researchers will be given accounts on the server housing these data. The document "Technical_Description_SIPP_Synthetic_Beta_July92007," also posted here, contains a codebook for this data set and further description of how th synthetic data were created.
SIPP Synthetic Beta Application (in PDF format)
Technical Description SIPP Synthetic Beta (7/9/2007) (MS WORD document)
Done

Cornell University
Cornell Institute for Social and Economic Research

Search Cornell

VirtualRDC News @ CISER

2007 Joint Statistical Meetings

July 29th, 2007

2007 Joint Statistical Meetings to be held at the Salt Palace Convention Center.
JSM (the Joint Statistical Meetings) is the largest gathering of statisticians held in North America. It is held jointly with the American Statistical Association, the International Biometric Society (ENAR and WNAR), the Institute of Mathematical Statistics, and the Statistical Society of Canada. Attended by over 5000 people, activities of the meeting include oral presentations, panel sessions, poster presentations, continuing education courses, exhibit hall (with state-of-the-art statistical products and opportunities), career placement service, society and section business meetings, committee meetings, social activities, and networking opportunities. Salt Lake City is the host city for JSM 2007 and offers a wide range of possibilities for sharing time with friends and colleagues. For information, contact jsm@amstat.org or phone toll-free (866) 421-7169.

Posted in Events | No Comments »

Corrected OTM data for IL posted on July 13

July 24th, 2007

Site search

Search for this text:

Search

Site navigation

Information about the VirtualRDC
Available resources
Data @ VirtualRDC
Classes and Tutorials
Help for RDC proposal writers

Recent articles

Events
General
Grants
Hardware
Library
Software

Related sites

CISER
NYCRDC

Conclusions

- Linked longitudinal worker-employer data will be crucial to the public policy debate on retirement and disability issues
- Research community needs to participate in the testing
- This will allow a more complete characterization of the quality of the synthetic data
- Future versions can be improved using this feedback