Recent Advances In Confidentiality Protection – Synthetic Data

John M. Abowd

April 2007

Overview

- The NSF-ITR Grant Synthetic Data Projects
- SIPP-SSA-IRS Synthetic Beta file
- OnTheMap

NSF-ITR Synthetic Data Projects

- Longitudinal Business Database
  - First release data due this week
  - First establishment-level micro-data file of its type
- Survey of Income and Program Participation/ Social Security Administration/ Internal Revenue Service Synthetic Beta File
  - Details later, beta file to be accessible on the Virtual RDC very soon
- OnTheMap
  - First release February 2006 (on Census.gov)
  - Second release April 2007 (on Census.gov)
  - All synthetic data (10 implicates on Virtual RDC)
- American Community Survey
  - Next official PUMS uses synthetic data as part of the confidentiality protection
- LEHD Infrastructure Files
  - Employer, individual, job data all synthetic
- Quarterly Workforce Indicators
  - All suppressions and variables related by identities replaced with synthetic values
  - Protype completed

SIPP-SSA-IRS Synthetic Beta File

- Links IRS detailed earnings records and Social Security benefit data to public use SIPP data
- Basic confidential data: SIPP (1990-1993, 1996); W-2 earnings data; SSA benefit data
- Gold standard: completely linked, edited version of the data with variables drawn from all of the sources
- Partially-synthetic data: created using the record structure of the existing SIPP panels with all data elements synthesized using Bayesian bootstrap and sequential regression multivariate imputation methods

Multiple Imputation Confidentiality Protection

- Denote confidential data by Y and disclosable 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 $p(Y,X,\theta)$.

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=4 completed datasets using SRMI (imputes values for all missing data)
  2) draw r=4 synthetic datasets for each completed dataset from predictive density given the completed data.

Confidentiality Protection

- Protection is based on the inability of PUF users to re-identify the SIPP record upon which the PUF record is based.
- This prevents wholesale addition of SIPP data to the IRS and SSA data in the PUF
- Goal: re-identification of SIPP records from the PUF should result in true matches and false matches with equal probability

Disclosure Analysis

- Uses probabilistic record linking and distance-based record linking
- Each synthetic implicate is matched to the gold standard
- All unsynthesized variables are used as blocking variables
- Different sets of unsynthesized variables are used as matching in the probabilistic record linking
- All synthesized variables are used in the distance-based record linking

Testing Analytic Validity

- Verify all marginal distributions
- Verify second, third, and fourth order interactions for stratifying variables
- Run analyses on each synthetic implicate
  - Average coefficients
  - Combine standard errors using formulae that take account of average variance of estimates (within implicate variance) and differences in variance across estimates (between implicate variance)
- Run analyses on gold standard data
- Compare average synthetic coefficient and standard error to the same quantities for the gold standard
- Analytic validity is measured by the overlap in the coverage of the synthetic and gold standard confidence intervals for a parameter

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

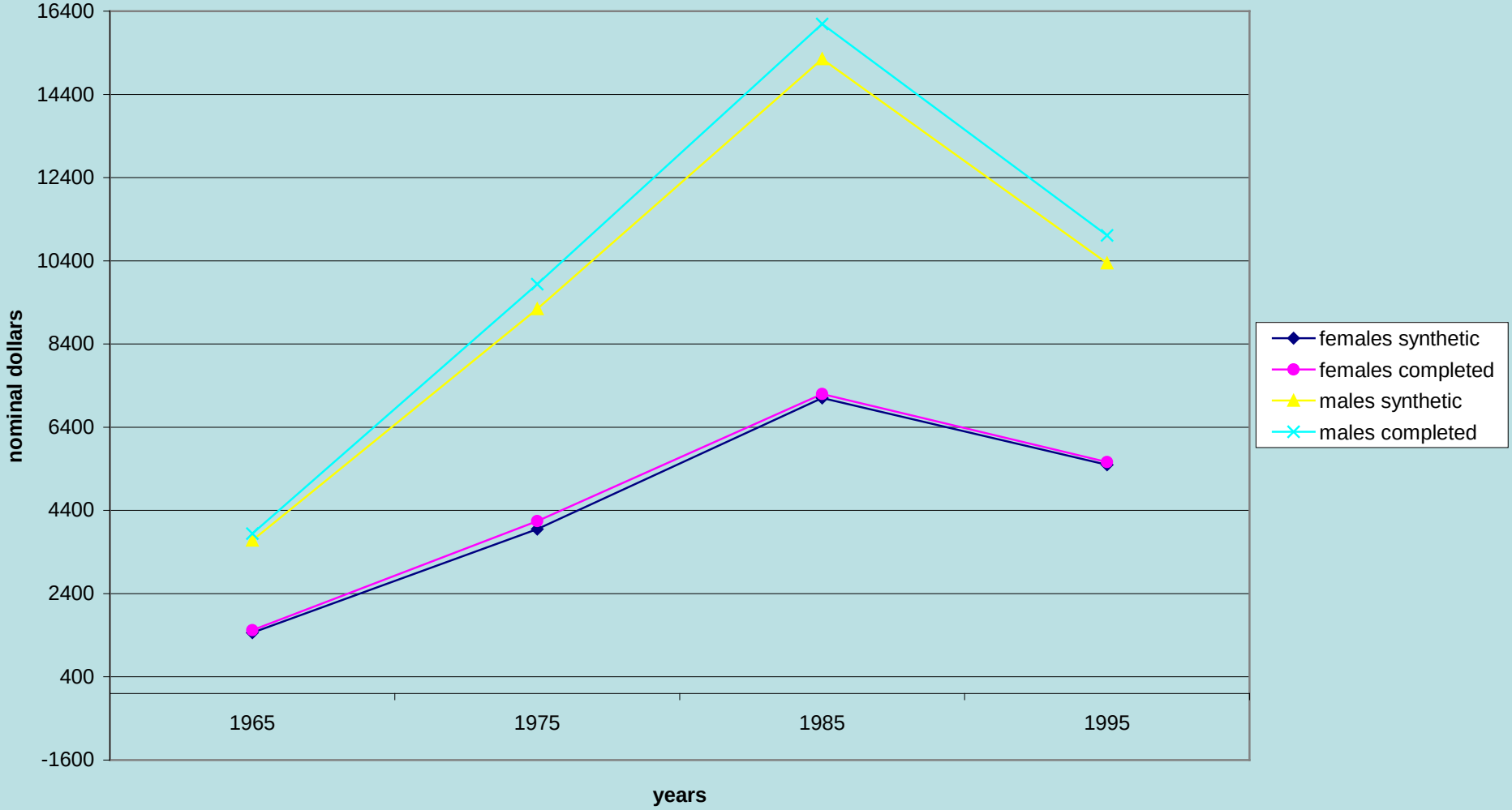
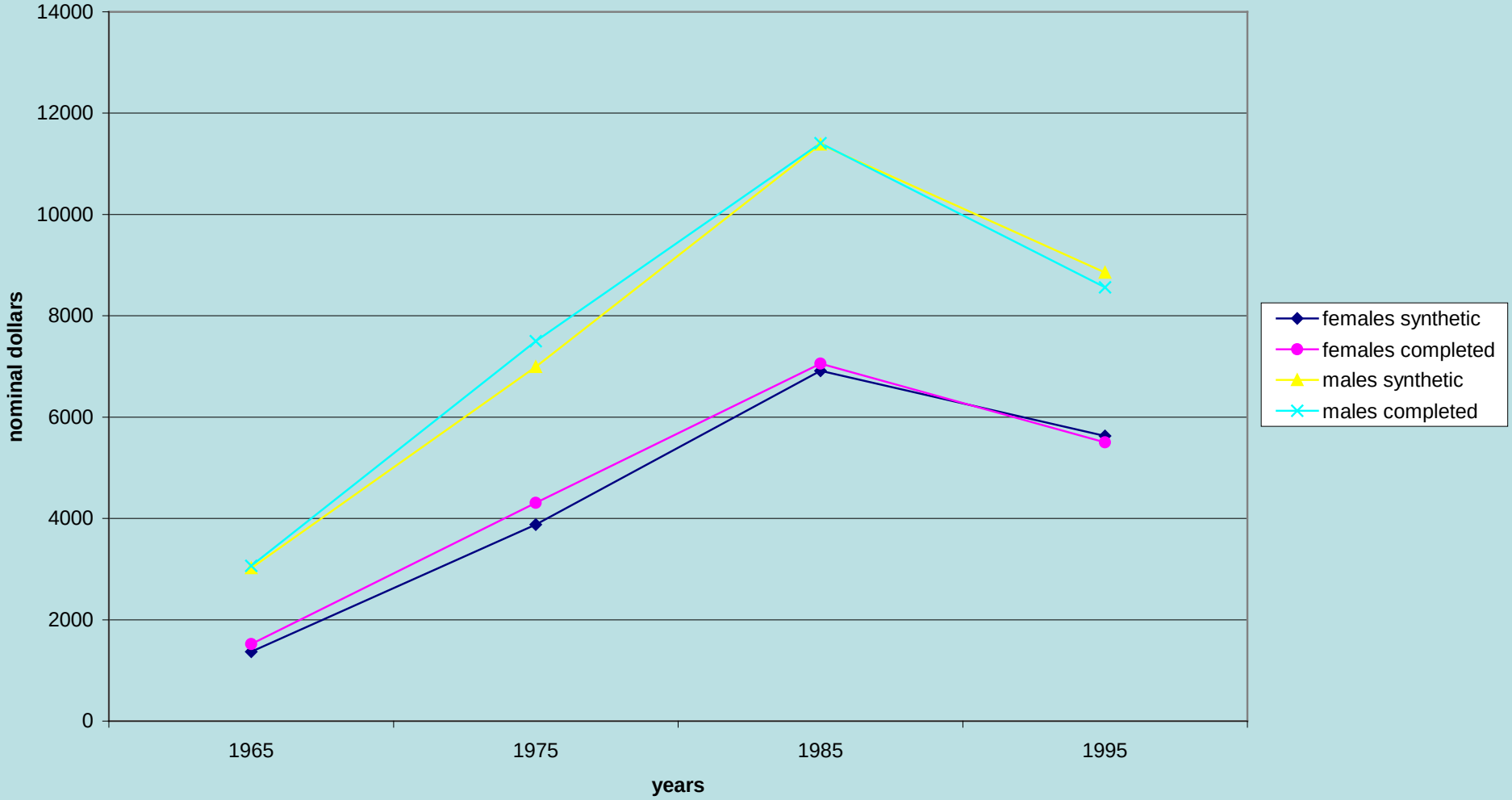


Chart 4:
Comparison of Synthetic and Completed Earnings
Retired Black Males and Females



Log Total Annual Labor Earnings (white males)

Table 40: Log of Total DER Earnings in year 2000 for white males

	Coefficient		Confidence Interval				Standard Error	
Explanatory Variables	Synthetic	Completed	Synthetic		Completed		Synthetic	Completed
Intercept	8.377	7.855	8.266	8.487	7.793	7.917	0.065	0.037
highschool_only	0.214	0.230	0.133	0.294	0.205	0.255	0.036	0.015
somecollege	0.400	0.431	0.263	0.537	0.404	0.457	0.059	0.016
college_only	0.738	0.880	0.530	0.947	0.851	0.909	0.086	0.017
graduate	0.830	1.110	0.632	1.028	1.080	1.140	0.085	0.018
disab	-0.354	-0.610	-0.380	-0.328	-0.657	-0.562	0.014	0.026
foreign_born	0.064	0.042	-0.029	0.157	0.013	0.070	0.042	0.017
hispanic	-0.072	-0.013	-0.113	-0.031	-0.040	0.013	0.021	0.016
ser_totyrs_2000	0.179	0.275	0.142	0.216	0.259	0.292	0.014	0.010
ser_totyrs_2000_2	-0.073	-0.140	-0.085	-0.062	-0.153	-0.128	0.007	0.007
ser_totyrs_2000_3	0.016	0.034	0.013	0.018	0.030	0.038	0.001	0.002
ser_totyrs_2000_4	-0.001	-0.003	-0.002	-0.001	-0.004	-0.003	0.000	0.000

Log Total Annual Labor Earnings (black males)

Table 41: Log of Total DER Earnings in year 2000 for black males

	Coefficient		Confidence Interval				Standard Error	
Explanatory Variables	Synthetic	Completed	Synthetic		Completed		Synthetic	Completed
Intercept	8.080	7.070	7.929	8.230	6.885	7.254	0.089	0.108
highschool_only	0.163	0.322	-0.031	0.357	0.231	0.413	0.090	0.053
somecollege	0.375	0.551	0.204	0.546	0.476	0.627	0.074	0.046
college_only	0.680	0.860	0.415	0.945	0.735	0.985	0.124	0.075
graduate	0.797	1.169	0.461	1.133	1.018	1.320	0.156	0.091
disab	-0.400	-0.631	-0.533	-0.267	-0.763	-0.499	0.062	0.075
foreign_born	0.082	0.046	-0.098	0.262	-0.106	0.197	0.084	0.084
hispanic	-0.030	0.156	-0.128	0.067	0.017	0.296	0.051	0.084
ser_totyrs_2000	0.173	0.388	0.154	0.191	0.336	0.440	0.011	0.030
ser_totyrs_2000_2	-0.067	-0.240	-0.078	-0.055	-0.284	-0.197	0.007	0.025
ser_totyrs_2000_3	0.013	0.067	0.009	0.018	0.053	0.080	0.003	0.008
ser_totyrs_2000_4	-0.001	-0.007	-0.002	-0.001	-0.008	-0.005	0.000	0.001

Log Annual Benefit Amount

Table 51: Log of MBA 2000 for retired individuals (TOB_2000=1)

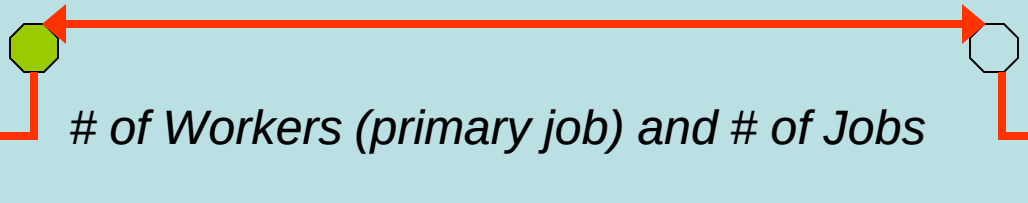
	Coefficient		Confidence Interval				Standard Error	
Explanatory Variables	Synthetic	Completed	Synthetic		Completed		Synthetic	Completed
Intercept	21.365	16.985	11.656	31.074	7.475	26.496	4.487	4.459
age_2000	0.009	0.009	0.005	0.012	0.007	0.011	0.002	0.001
blackfemale	-0.272	-0.250	-0.339	-0.206	-0.301	-0.199	0.032	0.026
blackmale	-0.095	-0.092	-0.122	-0.068	-0.127	-0.056	0.014	0.020
whitefemale	-0.211	-0.192	-0.284	-0.139	-0.257	-0.126	0.032	0.029
highschool_only	0.054	0.054	0.039	0.068	0.024	0.084	0.007	0.015
somecollege	0.098	0.078	0.075	0.121	0.052	0.104	0.012	0.014
college_only	0.136	0.128	0.103	0.169	0.084	0.172	0.016	0.022
graduate	0.157	0.150	0.132	0.182	0.125	0.175	0.013	0.015
disab	-0.019	0.001	-0.035	-0.003	-0.019	0.021	0.009	0.011
hispanic	-0.102	-0.059	-0.164	-0.040	-0.108	-0.009	0.028	0.025
divorced	0.110	0.126	0.071	0.148	0.075	0.177	0.021	0.027
married	0.107	0.081	0.073	0.142	0.051	0.110	0.019	0.017
widowed	0.232	0.217	0.187	0.278	0.171	0.264	0.024	0.026
famwelpart1999	-0.048	-0.022	-0.103	0.008	-0.058	0.013	0.026	0.019
hicovannual1999	0.023	-0.001	0.013	0.033	-0.012	0.010	0.006	0.007
log_totnetworth	0.012	0.040	0.004	0.021	0.035	0.046	0.004	0.003
ser_pct_yrs_wrked	0.948	1.001	0.473	1.423	0.593	1.409	0.202	0.174
year_initial_entitle	-0.008	-0.006	-0.013	-0.003	-0.011	-0.001	0.002	0.002

On The Map

- The Census Bureau's first public-use synthetic data application (publicly released on February 3, 2006)
- Developed by the U.S. Census Bureau's Longitudinal Employer-Household Dynamics Program (LEHD)
- Data on commute patterns between Census Blocks and area characteristics

Origin-Destination Database

Home Block
(U/I Wage Data)



# of Workers (primary job) and # of Jobs

Workplace Block
(ES-202 Data)

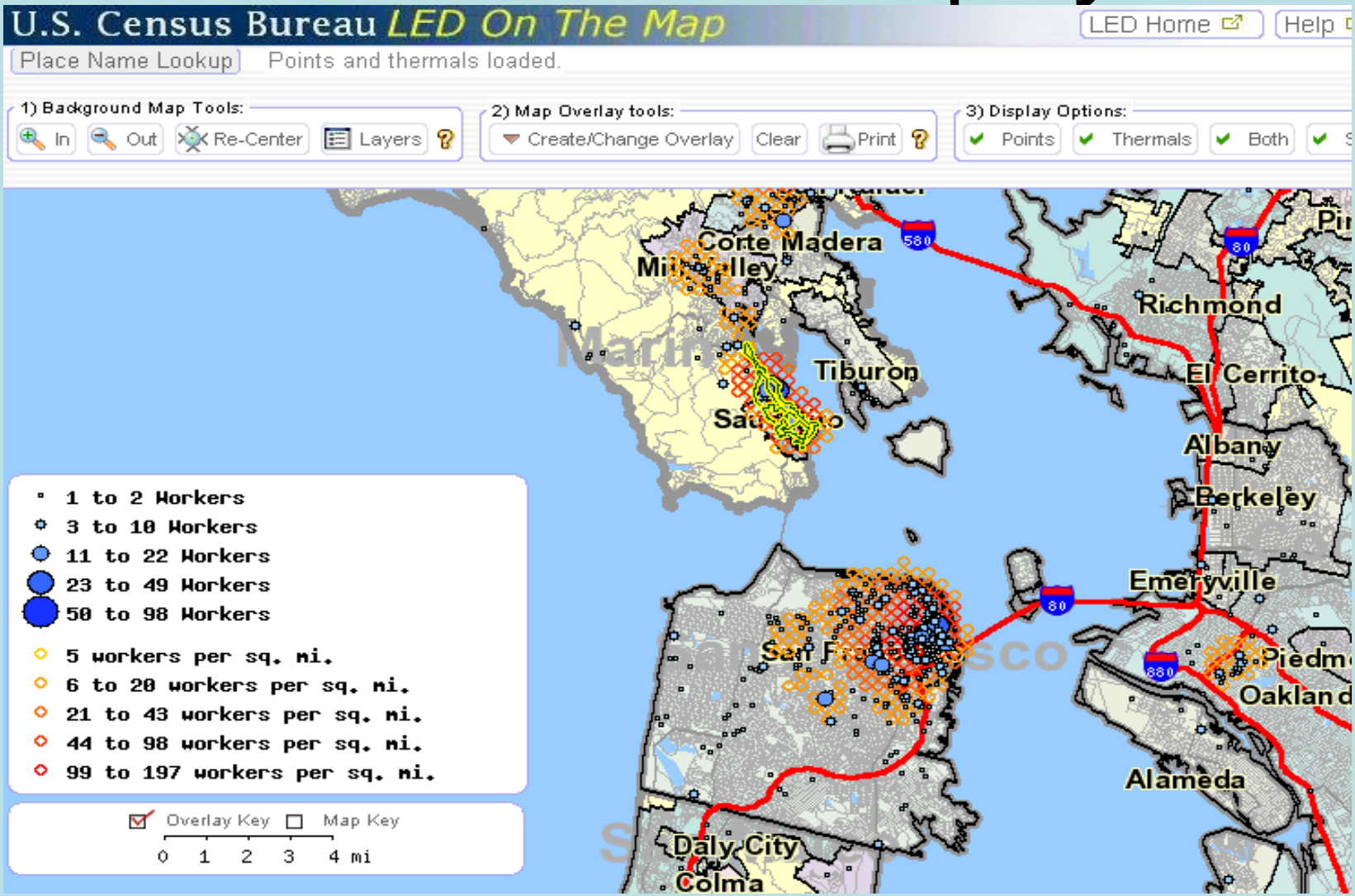
- Home Profile - block group
  - # workers
  - Worker distribution
    - by age range (-30; 31-54; 55+)
    - by monthly earnings range (-$1,199; $1,200-$3,399; $3,400+)
    - by industry (20 NAICS)

- Work Profile - block group
  - # establishments
  - # workers
  - Worker distribution
    - by age range (-30; 31-54; 55+)
    - by monthly earnings range (-$1,199; $1,200-$3,399; $3,400+)
    - by industry (20 NAICS)
  - Demand/growth indicators
    - Job creation/loss
    - Hires/separations
    - Earnings hires/separations

On The Map

- Origin block data are synthetic
  - Sampled from the posterior predictive distribution of origin blocks given destination block, worker characteristics
- All 10 implicates of the synthetic O/D data are available via the Virtual RDC
- Data available within a mapping application online: On The Map

Where Are Workers Residing in Sausalito, CA Employed?



Shed Report

	2003		2002	
	Count	Share	Count	Share
Number of Jobs Held by Residents				
* All Jobs	22,559	100.0%	21,129	100.0%
* All Jobs (Private Sector Only)	20,174	89.4%	21,129	100%
* All Primary Jobs (Worker's highest paying job)	21,418	94.9%	20,159	95.4%
* All Primary Jobs (Private Sector Only)	19,128	84.8%	20,159	95.4%

WHERE AREA WORKERS ARE EMPLOYED	2003		2002	
	Count	Share	Count	Share
Total Workers (Primary Jobs - Private Sector)	19,128	100.0%	20,159	100.0%
Cities/Towns Where Residents are Employed				
* San Ramon	1,894	9.9%	2,057	10.2%
* Unincorporated Areas	1,739	9.1%	1,703	8.4%
* San Francisco	1,515	7.9%	1,670	8.3%
* Pleasanton	1,097	5.7%	1,162	5.8%
* Danville	1,077	5.6%	1,017	5%
* All Other Locations	11,806	61.7%	12,550	62.3%
Counties Where Residents are Employed				
* Contra Costa	5,991	31.3%	6,228	30.9%
* Alameda	4,799	25.1%	5,092	25.3%
* Santa Clara	1,773	9.3%	1,855	9.2%
* San Francisco	1,515	7.9%	1,670	8.3%
* Los Angeles	1,095	5.7%	1,133	5.6%
* All Other Locations	3,955	20.7%	4,177	20.7%

Area Profile Report

	2003		2002	
	Count	Share	Count	Share
Number of Jobs Held by Resident Workers				
* All Jobs	22,559	100.0%	21,129	100.0%
* All Jobs (Private Sector Only)	20,174	89.4%	21,129	100%
* All Primary Jobs (Worker's highest paying job)	21,418	94.9%	20,159	95.4%
* All Primary Jobs (Private Sector Only)	19,128	84.8%	20,159	95.4%

	2003		2002	
	Count	Share	Count	Share
Total Workers (Primary Jobs - Private Sector)	19,128	100.0%	20,159	100.0%
Workers by Age				
* Age 30 or younger	3,476	18.2%	3,764	18.7%
* Age 31 to 54	12,352	64.6%	12,959	64.3%
* Age 55 or older	3,300	17.3%	3,436	17%
Workers by Earnings Paid				
* $1,200 per month or less	2,929	15.3%	3,217	16%
* $1,201 to $3,400 per month	3,462	18.1%	3,716	18.4%
* More then $3,400 per month	12,737	66.6%	13,226	65.6%
Workers by Primary Industry (2-digit NAICS)				
* Agriculture, Forestry, Fishing and Hunting	90	0.5%	82	0.4%
* Mining	121	0.6%	129	0.6%
* Utilities	110	0.6%	140	0.7%
* Construction	1,043	5.5%	1,099	5.5%
* Manufacturing	2,073	10.8%	2,239	11.1%
* Wholesale Trade	1,444	7.5%	1,491	7.4%

Disclosure Protection System

- Goal: “to protect confidentiality while preserving analytical validity of data”
  - No cell suppression
  - Synthetic place of residence data conditional on data on workplace and other characteristics
  - Bayesian techniques to estimate the posterior predictive distribution
  - Workplace data protected by QWI confidentiality rules
  - “Noise” in data increases as population in work place cell decreases

Synthetic Data Model

$$p(y_{i|jk} \mid \theta_{i|jk}) \propto \prod_{i=1}^{I} \theta_{i|jk}^{y_{ijk}}$$

$$\theta \sim \text{Dirichlet}\left(\alpha_{1|jk} + y_{1jk}, \ldots, \alpha_{I|jk} + y_{Ijk}\right)$$

- y_{ijk} are the counts for residence block *i*, work place block *j* and characteristics *k*
- Characteristics are age groups, earnings group, industry (NAICS sector), ownership sector

Complications

- Informative prior “shape”
- Prior “sample size”
- Work place counts must be compatible with the protection system used by Quarterly Workforce Indicators (QWI)
  - Dynamically consistent noise infusion

Design of Prior

- Unique priors for each of the *J x K* cells in the contingency table
- Only consider priors that have support across at least 10 residence blocks
- Search algorithm
  1. Work place tract, age category, earnings category, industry and ownership sector, else
  2. Work place tract, age category, earnings category, else
  3. Work place county, age category, earnings category, else
  4. Work place county
- Shape parameters based on observed distribution
- Scale parameters are confidential
  - The relative weight of the prior when sampling from the posterior distribution is larger for smaller populations

Residence Block (i)	Work Block (j)				Total
	W1	**W2**	**….**	**WJ**	
R1	2	5	…	…	50
R2	3		…	…	400
R3			…	…	50
R4		90	…	…	200
R5			…	…	100
R6			…	…	20
R7			…	…	20
R8			…	…	20
R9			…	…	40
R10			…	…	100
Total	5	95	…	…	1000

Residence Block (i)	Prior distribution (Aggregated Work Block Distribution)	Likelihood (Original Work Block Distribution)	Posterior Expected Counts	Posterior Probabilities	Synthetic Data
R1	0.050	0.400	2.350	0.196	1
R2	0.400	0.600	5.800	0.483	2
R3	0.050		0.350	0.029	
R4	0.200		1.400	0.117	
R5	0.100		0.700	0.058	1
R6	0.020		0.140	0.012	
R7	0.020		0.140	0.012	
R8	0.020		0.140	0.012	1
R9	0.040		0.280	0.023	
R10	0.100		0.700	0.058	1
Total	1.000	1.000	12.000	1.000	6

Work block	5
Prior	7
QWI estimate	6

Analytic Validity

- Assess the bias
- Assess the incremental variation

Census Tract	(1) Workers	(2) Average commute distance in true data (in miles)	(3) Average commute distance in synthetic data (in miles)	(4) Difference in miles	(5) Standard deviation across 10 implicates over (1)
1	6,747	17.9	17.9	0.0	0.019
2	4,535	14.6	14.8	0.1	0.013
3	2,251	18.5	19.3	0.9	0.018
4	1,932	12.0	13.2	1.3	0.043
5	1,996	15.0	15.0	-0.1	0.028
6	2,135	14.3	15.7	1.3	0.036
7	1,809	12.8	13.9	1.1	0.036
8	2,004	8.5	8.5	0.0	0.039
9	1,515	11.8	12.1	0.3	0.021
10	1,365	21.1	23.2	2.0	0.040
11	1,233	16.3	17.4	1.1	0.031
12	879	15.1	16.8	1.8	0.067
13	811	11.3	11.3	0.0	0.072
14	634	10.4	10.4	-0.1	0.051
15	618	9.6	9.6	0.0	0.046
16	526	11.4	10.1	-1.3	0.088
17	531	17.1	18.4	1.3	0.045
18	541	14.4	14.5	0.2	0.063
19	378	15.0	14.4	-0.6	0.069
20	372	7.7	7.2	-0.5	0.069
21	138	7.8	8.1	0.3	0.064
Total	32,951				

	Size weighted average of absolute difference in commute distance in confidential and synthetic data				
Population in Work Block	Mean	P20	P40	P60	P80
1-5	9.33	4.72	5.72	8.98	13.95
6-10	5.89	1.88	3.13	4.69	8.71
11-20	3.82	1.76	2.42	2.68	4.24
21-50	3.34	1.19	1.76	2.27	3.58
51-100	2.21	0.69	1.55	1.44	2.36
101-250	1.38	0.40	0.65	1.92	2.12
250-500	0.96	0.16	0.38	0.72	1.64
501-high	0.27	0.05	0.12	0.15	0.13

Confidentiality Protection

$$RI = \sum_{i=1}^{I} \frac{\text{abs}\left(\tilde{y}_{ij} - \bar{y}_{ij}\right)}{2\bar{y}_{j}}$$

- The reclassification index (*RI*) is a measure of how many workers were geographically relocated by the synthetic data
- Interpretation: Proportion of workers that need to be reallocated across residence areas in synthetic data in order to replicate confidential data
  - If counts identical in synthetic and confidential data, *RI* = 0
  - If no overlap, *RI* = 1

In aggregate synthetic data mimic residence patterns in confidential data well

Definition of residence area	Reclassification index	Size-weighted coefficient of variation across10 implicates
County	0.85%	0.0085
Census Tract	2.80%	0.0495
Block	7.25%	0.1895

Level of protection increases as population in work block decreases

Mean proportion of workers that need to be reallocated across selected residence areas in the synthetic data to replicate confidential data

Population in Work Block	Counties	Census Tracts	Blocks
1-5	30%	36%	43%
6-10	23%	25%	29%
11-20	18%	23%	24%
21-50	12%	18%	19%
51-100	10%	15%	17%
101-250	6%	11%	13%
250-500	5%	9%	13%
501-high	3%	7%	11%