RECONSTRUCTING THE 1890 CENSUS

A PROJECT

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

PETER DAVIS CLARK

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Advisor

Prof. Jaideep Srivastava

December, 2011

Final Version

© Peter Clark, 2011

**Acknowledgments**

I wish to thank Dr. Jaideep Srivastava for encouraging my work on this project. His ideas and suggestions during the course of the project have been invaluable.

I also wish to thank Dr. Steven Ruggles, the director of the Minnesota Population Center, for guidance and support as I developed an appreciation for historical demography and for first introducing me to the story of the lost census of 1890 and suggesting that it would be a neat problem to explore.

I thank my parents for wishing for a child blessed with curiosity. I hope they got their wish.

Lastly, I wish to thank my wife, Molly, for her support and patience as I engaged in many late nights of development, debugging, analysis, and writing.

**Dedication**

This project is dedicated to my children. If Daddy can get his homework done, so can you.

**Abstract**

On January 10, 1921, the original United States 1890 census schedules—the household-by-household enumeration forms—were severely damaged in a fire. Most of the surviving records were destroyed in the mid-1930s. The loss represents irreplaceable information about the condition of the country at the end of the nineteenth century. However, census microdata for the surrounding enumerations do survive, as do tabulations from 1890. We propose a method for generating a synthetic microdata dataset with similar statistical properties to 1890 that, while of limited use, would nevertheless fill a hole in the historical census records. This method constructs a new dataset by sampling with replacement of entire households from the 1880 and 1900 census microdata records to match selected tabulations from 1890. We show that the dataset produced by this method has error properties not dissimilar to those found when comparing tabulations produced from the IPUMS 1880 and 1900 microdata samples to the actual tabulations from 1880 and 1900.

**Table of Contents**

Introduction and Historical Context 1

Motivation 5

Related Work 7

Method 9

Setup 9

Aggregate Data 9

Microdata 11

Fitting and Substitutions 12

Results 16

Analysis of Literacy by Sex 20

Occupational Class by Sex by Year 22

Conclusions and Future Work 29

Appendix I 30

Appendix II 36

Bibliography 37

**List of Tables**

Table 1—Aggregate statistics comparison for Connecticut 16

Table 2—Literacy by Sex, population of age > 10, 1880–1900 20

Table 3—Occupational Class by Sex, for employed persons age > 10, 1880–1900 23

Table 4—Questions asked of each household member in 1890 30

Table 5—Questions asked of the household as a whole in 1890 31

Table 6—Aggregate variables used for fitting 31

Table 7—Household-level variables included in extract 34

Table 8—Person-level variables included in extract 34

**List of Figures**

Figure 1—Percentage difference between microdata tabulation and census statistics, literacy by sex for population age > 10, 1880–1900 22

Figure 2—Percentage Difference for Occupational Class, employed males age > 10, 1880–1900 26

Figure 3—Percentage Difference for Occupational Class, employed females age > 10, 1880–1900 27

Figure 4—Percentage Difference for Occupational Class, employed persons age > 10, 1880–1900 28

# Introduction and Historical Context

Census information, both in the form of tabulations and individual-level microdata, represent a rich resource for studying populations at a point in time, and a microdata series enables demographers to study change in population characteristics over time. This is particularly important and useful when things are changing quickly. The late 1800s represented a time of great change in the United States—it was a time of rapid population growth, high immigration (nearly 9% of the total population of Norway emigrated to America), industrialization, labor turmoil, and the closing of the western frontier. The United States conducts censuses once per decade (a sampling frequency of roughly 3.171\*10-9 Hz), and accordingly, censuses were taken in 1870, 1880, 1890, and 1900, providing snapshots of the state of the country during the late 1800s. However, the mechanics of conducting the census presented increasing challenges.

The tenth census of the United States began on June 1, 1880. Census returns were tallied mechanically using machines developed for the 1870 census. Congress appropriated money for the enumeration and tabulation of the census, but these funds were exhausted in the first half of 1881, with tabulation work still far from completion. In fact, tabulation work continued into late 1888, with the final volumes comprising over 19,000 pages. The tabulation work was left incomplete—data about several subject areas, including churches, educational institutions, libraries, and insurance, were collected during the census but no reports were published based on this data. (Wright 1900, 68)

The time and level of effort required by the 1880 census set the stage for a crisis in preparing for the eleventh census in 1890. The country had continued to grow, and tabulating the 1890 census using the methods from the 1880 census would mean that the results wouldn’t be complete until after 1900. Not only would some of the tabulations be a decade out-of-date upon publication (as was the case for the 1880 census), but this would also mean that preparations for the 1900 census would need to be conducted while the 1890 census was still being processed. Clearly, changes in methods were needed.

In 1888, the US Census Bureau held a competition to find a more efficient method. Entrants had to capture and prepare a small subset of 1880 census data for tabulation and then tabulate it. Of three entrants, one was the clear winner—a machine designed by Herman Hollerith (Hollerith 1889). Hollerith’s punch-card based electric tabulator completed the tabulation challenge in 5.5 hours, well under the 44.5 hours of the second-fastest entry (The Hollerith Tabulator 2011).

Census day for the eleventh census was set for June 2, 1890, and the enumeration process was expected to be completed within two weeks. Each household was enumerated on a separate form and was asked the questions in Table 4 and Table 5 given in Appendix I.

Once completed, the census schedules were “transcribed” to punched cards and tabulated using several passes through Hollerith’s tabulation machine (Report on Population of the United States at the Eleventh Census: 1890—Part I 1895). Hollerith’s machine was instrumental in rapidly compiling tabulations. In contrast to the pace of the 1880 census, preliminary results for some tabulations were available by August 1890, and the final general results were available to the public in May 1895 (Wright 1900, 75), although work continued until July 1897. Variants of Hollerith’s machine were in use at Census until the 1950s, when the first UNIVAC-I was installed.

Many of the original paper enumeration forms were stored in the basement of the US Department of Commerce Building, which contained a fireproof and waterproof vault. As of 1921, the vault contained records from 1830, 1840, 1880, 1900, and 1910. However, the enumeration records for the 1890 census were not kept in the vault, but rather were stacked outside. When the building caught fire on January 10, 1921, the 1890 records were “first in the path of the firemen.” The morning after the fire, Census director Sam Rogers reported that 25% of the 1890 records were destroyed and 50% of the remainder had been damaged by smoke, water, and fire. The surviving records sat in storage between 1922 and 1933. Congress authorized the destruction of the remaining records in 1933, and the records were indeed destroyed in either 1934 or 1935. The only surviving originals are records on just over six thousand individuals from a variety of locations (Alabama, Georgia, Illinois, Minnesota, New Jersey, New York, North Carolina, Ohio, South Dakota, Texas, and Washington DC). (Blake 1996).

Although the 1890 microdata were lost, extensive tabulations at the national, state, and county levels were published and distributed, and these are readily available. In addition, microdata records for the censuses of 1880 and 1900 survive as well. These records provide the basis for a method of creating synthetic data with the same statistical properties held by the 1890 census.

# Motivation

The primary motivation for this work is to eventually fill the hole left by the loss of the 1890 data in the IPUMS USA microdata series. IPUMS USA offers microdata samples for every decennial US census from 1850 to 2000, with the exception of 1890. For researchers interested in changes in population characteristics over time, this gap is at best an annoyance, at worst a disaster. In many cases, linear interpolation between 1880 and 1900 works fine. However, there are some specific cases where it does not. One example of this is age at first marriage, which increased from 1880 to 1890 and then began a decline that would last until 1960 (Fitch and Haines 2006). As the 1890 data represent an inflection point in this trend, it would be useful to have some sort of data available.

Although there are numerous methods of generating synthetic data, most of them create records by drawing variable values from a statistical distribution or by interpolating from existing records. Each individual record is fictional. In addition, methods that create records by interpolating values from existing records (such as SMOTE (Chawla, et al. 2002)) can fail for nominal variables, and methods that create records by drawing from a distribution may miss correlations that exist in the source dataset but for which a joint distribution wasn’t provided. As a simple example, within a household, spousal ages are generally close to each other and children’s ages are generally at least 15 years less than the ages of their parents. Inferences about family structure made from a synthetic dataset that doesn’t preserve intra-household relationships like this will be of questionable value.

Our method avoids both of these issues. Each output household for the synthetic 1890 sample is a household configuration that actually existed in either the 1880 or 1900 censuses and contains individuals that were actually enumerated in 1880 or 1900. Correlations between variable values for an individual, as well as correlations that are present across household members, are preserved.

# Related Work

The generation of synthetic data has been a topic of interest for a variety of reasons, but one of the most significant has been for providing research microdata that is representative of a population of interest but preserves privacy. A variety of approaches have been used. Lee (Lee, 2007; Lee, 2009) generates a complete joint distribution model for a set of variables of interest (presumably not for all the variables present in the microdata) from marginal tables and then generates synthetic individual-level microdata from that joint distribution. Beckman (1996) uses a similar method to generate synthetic households for use in transportation simulations by sampling households from Census Bureau PUMS files to match a joint distribution. Beckman’s validation method is essentially the same as ours, comparing crosstabs of the synthetic data to the actual reported crosstabs for variables not used in the synthetic data generation.

Both Deming and Stephan (Deming and Stephan 1940); (Stephan 1942) describe methods for iteratively adjusting frequency tables where expected marginal totals are known. Their Iterative Proportional Fitting (IPF) algorithm underlies both Beckman’s and Lee’s approach for creating complete joint distributions.

Abowd and Lane (2004) address concerns about the validity of inferences drawn from synthetic data by proposing a model by which researchers could use synthetic data for study and research and then validate the conclusions on the actual data within the bounds of a restricted access data center.

In several places, Reiter (2003) (2004) discusses the generation of partially synthetic data and methods for making inferences from it using multiple imputation techniques to suppress outlier values or other values that would involve high risk of identity disclosure.

Lastly, Dong and Meeden (Forthcoming) describe a different approach to the same problem of generating synthetic 1890 data based on drawing from the 1880 and 1900 censuses using simulated annealing as an optimization technique. Their simulated annealing method randomly selects one household to remove and selects a separate random household as a replacement.

# Method

Construction of the synthetic data proceeds in two phases. First, we replicate complete households from the 1880 and 1900 input file and add them to the output, wrapping around to the beginning of the input file as necessary until the state-level total population count is matched. Second, once the state-level total population count is matched, we iterate through the input file and, for each household *h* in the input, evaluate whether replacing one of the existing households in the output with *h* would provide a better or worse match to all the other state-level statistics. If making the replacement results in a better fit, then we make the replacement and move on to check the next household in the input. This process repeats until a complete pass has been made through the input file with no replacements. The steps and actual algorithm are presented in more detail in the following sections.

## Setup

The complete process consists of three steps:

1. Loading and parsing of aggregate data variables and associated universe statements
2. Loading and parsing of input microdata records
3. Generation and fitting of output data

## Aggregate Data

The process uses the 1890 US Census aggregate files contained in the 1890 dataset from Historical, Demographic, Economic and Social Data: The United States, 1790–2002 (Dataset 0018), downloaded from the Inter-university Consortium for Political and Social Research (ICPSR); Historical, Demographic, Economic, and Social Data (Haines 2010).

Although aggregate census data are often reported as tables, in this case each table cell is treated as an individual variable. For example, rather than having a table or crosstab of “nativity by sex by age,” we utilize the data as a series of variables—“native white males aged 5–20,” “foreign white males aged 5–20,” and “colored females aged 5–20” are treated as separate entities. Each aggregate variable has a name/mnemonic: for the three variables given above, the mnemonics are “nbwm520,” “fbwm520,” and “colf520.”

Each aggregate variable has an associated universe statement that connects the aggregate variables to the microdata variables. A small domain-specific language was implemented to concisely capture how to tally the microdata records–this language was closely modeled after the syntax for the WHERE clause in SQL. For instance, the aggregate variable “nbwm520” (native white males aged 5–20), has a universe statement of “NATIVITY in [1,2,3,4] and RACE = 1 and SEX = 1 and AGE >= 5 and AGE <= 20.” NATIVITY, RACE, AGE, and SEX are mnemonics for microdata variables. When tallying records from the input microdata, a record having a NATIVITY value of 1, 2, 3, or 4; RACE of 1; SEX of 1; and AGE between 5 and 20 would result in an increment to the nbwm520 aggregate variable.

For 1890, the aggregate data files also contain variables that are not relevant to population characteristics, mainly agricultural data such as “Value of Livestock” and “Acres of Barley.” These variables are marked with universe statements of “Nontabulated,” indicating that they are to be ignored. The 1890 aggregate dataset from ICPSR contains a total of 198 aggregate variables, of which 71 are useful for tabulation and 127 are ignored. Of the tabulated variables, 37 deal with birthplace and 18 deal with nativity in one form or another.

The specific aggregate variables used are given in Appendix I as Table 6—Aggregate variables used for fitting.

## Microdata

The 1890 synthesis process also takes as input a hierarchically structured microdata file consisting of households from the 10% 1880 and 5% 1900 IPUMS samples (Ruggles, et al. 2010). To produce the results discussed below, the sample was restricted to households from Connecticut (Connecticut is the first state listed in the ICPSR coding sequence). The specific extract used as input consists of 44 person-level variables and 10 household-level variables and contained a total of 26,125 households. IPUMS provides a codebook consisting of the variable name and field widths; reading the IPUMS data file is simplified by parsing the codebook and using it to determine the structure of the data file itself.

The microdata file is hierarchically structured; the file consists of a sequence of households organized as a household record (typically containing information about geographic location, housing stock characteristics, and economic characteristics, such as mortgage and property taxes and agricultural info), followed by a series of zero or more individual records containing information about the residents of the household. The specific variables included in the extract are given in Appendix I as Table 7 and Table 8.

Reading the microdata file results in an array of household objects, each of which contains an array of person records. The microdata records are connected to the aggregate variables through the aggregate variable universe statements: each universe statement declares the microdata variable mnemonic and categories to which it applies, as discussed above. Since households are treated as indivisible units in this process but the aggregate variables tabulate individuals, not households, deltas for each tabulated aggregate variable are computed and cached when the household is instantiated.

## Fitting and Substitutions

As noted above, the microdata variables are connected to the aggregate variables through universe statements. The universe statement is a logical expression that evaluates to true or false for each individual microdata record. If the expression evaluates to true for the microdata record, then the tally of the given aggregate variable should be incremented. If it evaluates to false, then there’s no change to the tally.

Once the aggregate data and microdata are read, a 100% “sample” that approximates the 1890 census is constructed as follows:

Iteratively copy households from the input array to the output list, in order, until the total population aggregate statistic is matched. If adding a household would cause the output list to exceed the total population count, then skip that household and proceed to the next one.

Once the total population count is satisfied, iterate through the input array *H*. For each household *h* in the input array *H,* compute the change in distance between the aggregate statistics and the tallies for the output dataset that would result in using *h* to replace each equally-sized household *o* in the output array *O.* Pick the output household *o* that offers the best improvement when removed, and make the replacement. Repeat until a full pass through *H* has been made with no replacements.

The distance between the known aggregate statistics and the tallies for the output dataset is computed as the Cartesian distance. Given a target in the form of the set of *i* known aggregate variables *A*, and the equivalent set of tallied values *T* from the output dataset, the distance between the output dataset and the target is .

Additionally, the index of the best substitution from the output is identified as follows. Given the set of *i* known aggregate variables *A*, the equivalent set of tabulated values *T* from the output dataset, deltas to the tabulated values for a household *H* at index *k* in the input dataset, and the deltas to the tabulated values for household *O* at each index *x* in the output dataset; then the optimal substitution for replacing *Ox* in the output with *Hk* from the input is .

Pseudocode for the algorithm is given below.

Algorithm FitAggregates(A, T, H, O) {

Input: collection of aggregate variables A and collection of current tallies T for contents of the output array,  
an array of input households H,  
an array of output households O such that T(total population) = A(total population)

Output: returns T and O with modifications

foundSubstitution = false

do {

foundSubstitution = false

for each h in H do {

i = FindSubstitution(A, T, h, O)

if (i != -1) {

RecordSubstitution(T, h, O, i)

foundSubstitution = true

}

}

} while (foundSubstitution)

}

FindSubstitution(A, T, h, O) {

Input: collection of aggregate variables A and collection of current tallies T for contents of the output array,   
a specific household h from the input array,  
an array of output households O

Output: index of household in O which, when removed and replaced with h, provides the best improvement in fit, or -1 if there is no such household

initialDistance = Distance(A, T)

distances = O.map(o =>DistanceWithSubstitution(A, T, h, o) - initialDistance)

let i = the index of o in O having the minimum value of distances(o), where the number of people in h equals the number of people in o.

if (distances(i) >= 0) return -1

else return i

}

RecordSubstitution(T, h, O, i) {

Input: collection of current tallies T for contents of the output array,  
a specific household h from the input array,  
array of output households O,  
index of household in O to replace with h

Output: None, but T and O are updated in place as side effects.

detabulate O(i) in T

tabulate h in T

O(i) = h

}

Distance(A, T) {

Input: collection of aggregate variables A and collection of current tallies T for the contents of the output array.

Output: cartesian distance between aggregate variables A and current tallies T

result = 0

for each aggregate variable a in A do

result = result + (A(a).value - T(a).value)^2

return sqrt(result)

}

DistanceWithSubstitution(A, T, h, o) {

Input: collection of aggregate variables A and collection of current tallies T for the contents of the output array,  
an input household h to substitute for an output household o

Output: cartesian distance between aggregate variables A and current tallies T given the substitution

result = 0

for each aggregate variable a in A do

result = result + (A(a).value - (T(a).value + h(a).value - o(a).value))^2

return sqrt(result)

}

# Results

The algorithm given above was coded in the Scala programming language (Odersky, et al. 2006) and run on the IPUMS extract described above. The program was run on a SunBlade X6270 server with two quad-core Intel Xeon X5560 processors (2.8 GHz, two threads per core) and 24 GB of memory, running the Solaris 10 operating system, and running Scala 2.9.1 on Sun/Oracle JDK 1.7.0\_01-b08. The JVM was invoked with the following tuning flags:

-Dfile.encoding=UTF8 -d64 -server -Xmx8g -Xms2g -Xss1M -XX:+UseLargePages -XX:+UseMPSS -XX:+UseParallelOldGC -XX:+UseBiasedLocking -XX:+UseAdaptiveSizePolicy -XX:MaxPermSize=256m

The program took 962 seconds to run on the above hardware and made 45 passes over the source data to make substitutions. A dataset consisting of 176,102 households and 746,258 people was generated.

The results of the fitting process resulted in the following comparison on an aggregate-variable-by-aggregate-variable basis:

Table —Aggregate statistics comparison for Connecticut

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mnemonic | 1890 Actual | Program | 𝚫 | Description |
| chitot | 272 | 296 | -24 | Total Chinese population 1890 |
| colf520 | 2043 | 2040 | 3 | Colored females aged 5–20 |
| colftot | 6498 | 6469 | 29 | Colored females |
| colm1844 | 3095 | 3094 | 1 | Colored males aged 18–44 |
| colm21 | 3976 | 3977 | -1 | Colored males aged 21+ |
| colm520 | 1786 | 1786 | 0 | Colored males aged 5–20 |
| colmtot | 6322 | 6443 | -121 | Colored males |
| fbftot | 90383 | 90422 | -39 | Foreign-born females |
| fbmtot | 93218 | 93192 | 26 | Foreign-born males |
| fbwf520 | 14572 | 14570 | 2 | Foreign white females aged 5–20 |
| fbwftot | 90324 | 91821 | -1497 | Foreign-born white females |
| fbwm1844 | 52947 | 52948 | -1 | Foreign white males aged 18–44 |
| fbwm21 | 78062 | 77993 | 69 | Foreign white males aged 21+ |
| fbwm520 | 13801 | 13732 | 69 | Foreign white males aged 5–20 |
| fbwmtot | 92831 | 91821 | 1010 | Foreign-born white males |
| ftot | 376720 | 376771 | -51 | Total female population |
| indtot | 228 | 244 | -16 | Civilized Indian population 1890 |
| japtot | 18 | 37 | -19 | Total Japanese population 1890 |
| mtot | 369538 | 369487 | 51 | Total male population |
| nbftot | 286337 | 286349 | -12 | Native-born females |
| nbmtot | 276320 | 276295 | 25 | Native-born males |
| nbwf520 | 94161 | 94163 | -2 | Native white females aged 5–20 |
| nbwffp | 97648 | 97674 | -26 | Native white females/for parents |
| nbwfnp | 182250 | 182275 | -25 | Native white females/nat parents |
| nbwm1844 | 107823 | 107820 | 3 | Native white males aged 18–44 |
| nbwm21 | 142054 | 142053 | 1 | Native white males aged 21+ |
| nbwm520 | 94882 | 94882 | 0 | Native white males aged 5–20 |
| nbwmfp | 95400 | 95820 | -420 | Native white males/for parents |
| nbwmnp | 174985 | 175403 | -418 | Native white males/nat parents |
| negtot | 12302 | 12335 | -33 | Total Negro population 1890 |
| pbasia | 33 | 45 | -12 | Persons born in Asia n.s. |
| pbatlisl | 183 | 183 | 0 | Persons born in Atlantic Islands |
| pbaustra | 86 | 77 | 9 | Persons born in Australia |
| pbaustri | 1187 | 1184 | 3 | Persons born in Austria |
| pbbelg | 165 | 142 | 23 | Persons born in Belgium |
| pbcanada | 21231 | 21229 | 2 | Persons born in Canada/Newfoundland |
| pbcentam | 10 | 11 | -1 | Persons born in Central America |
| pbchina | 298 | 297 | 1 | Persons born in China |
| pbcubawi | 200 | 196 | 4 | Persons born in Cuba/West Indies |
| pbdenmar | 1474 | 721 | 753 | Persons born in Denmark |
| pbenglan | 20572 | 20570 | 2 | Persons born in England |
| pbeurns | 39 | 38 | 1 | Persons born in Europe n.s. |
| pbfrance | 2048 | 2047 | 1 | Persons born in France |
| pbgerman | 28176 | 28175 | 1 | Persons born in Germany |
| pbgreece | 5 | 0 | 5 | Persons born in Greece |
| pbhollan | 121 | 142 | -21 | Persons born in Holland |
| pbhungar | 1146 | 1145 | 1 | Persons born in Hungary |
| pbindia | 48 | 50 | -2 | Persons born in India |
| pbirelan | 77880 | 77879 | 1 | Persons born in Ireland |
| pbitaly | 5285 | 5284 | 1 | Persons born in Italy |
| pbjapan | 22 | 38 | -16 | Persons born in Japan |
| pbluxemb | 2 | 10 | -8 | Persons born in Luxemburg |
| pbmexico | 12 | 16 | -4 | Persons born in Mexico |
| pbnorden | 0 | 751 | -751 | Persons born in Norway/Denmark |
| pbnorway | 523 | 30 | 493 | Persons born in Norway |
| pbpacifi | 15 | 0 | 15 | Persons born in Pacific Islands |
| pbpoland | 1504 | 1504 | 0 | Persons born in Poland |
| pbportug | 230 | 229 | 1 | Persons born in Portugal |
| pbrussia | 3027 | 3024 | 3 | Persons born in Russia |
| pbscot | 5992 | 5991 | 1 | Persons born in Scotland |
| pbsea | 60 | 57 | 3 | Persons born at sea |
| pbsoamer | 65 | 68 | -3 | Persons born in South America |
| pbspain | 45 | 45 | 0 | Persons born in Spain |
| pbsweden | 10021 | 10020 | 1 | Persons born in Sweden |
| pbswitz | 998 | 997 | 1 | Persons born in Switzerland |
| pbturkey | 49 | 37 | 12 | Persons born in Turkey |
| pbwales | 629 | 629 | 0 | Persons born in Wales |
| totpop | 746258 | 746258 | 0 | Total population 1890 |
| wftot | 370222 | 370302 | -80 | Total white females |
| whtot | 733438 | 733346 | 92 | Total white population |
| wmtot | 363216 | 363044 | 172 | Total white males |

There are some discrepancies in Table 1, such as the count of 0 persons in the aggregate statistics for persons born in Norway/Denmark but 751 persons in the synthesized data. The universe statement for tallying the Norway/Denmark variable includes people reported as born in either Norway or Denmark, which may be incorrect. Another is the count of 15 people born in the Pacific Islands in the aggregate statistics, but there are 0 in the output data. Because the input data did not contain any persons born in the Pacific Islands, the pbpacifi aggregate statistic can’t be fitted.

Evaluating the quality of the synthesized dataset can be done in a variety of ways. However, the utility of the dataset lies in the potential to draw inferences that aren’t connected with the aggregate variables used to generate the data. If the dataset is generated by fitting to nativity and birthplace variables, one would expect that analyses related to birthplace and nativity would generally be accurate. Such analyses would also be of limited value, since tabulations of nativity and birthplace already exist and are derived from ground truth rather than from synthetic data. Therefore, the quality of the synthesized dataset must be measured in comparison to how well it produces tabulations of aggregate variables that weren’t used in its production.

In addition, since the dataset is being assembled from samples drawn from the 1880 and 1900 censuses, an exact match can’t be expected. However, that leaves the question of how loose of a fit to the aggregate statistics is acceptable? For this analysis, the fit between the IPUMS 1880 and 1900 samples themselves and the corresponding Census aggregate statistics are treated as the gold standard.

Crosstabs for two tabulations—Literacy by Sex and Occupation Category by Sex—were produced from the 1880 and 1900 samples and compared to the corresponding 1880 and 1900 aggregate statistics. These tabulations were selected because they involve variables that weren’t used in the synthesis of the 1890 dataset and because the Occupation Category by Sex tabulation is not part of the ICPSR machine-readable aggregate dataset—the values for this table were hand-entered from the printed census volumes. The same crosstabs were run against the 1890 synthetic sample and compared to the actual 1890 aggregate statistics. As reported below, % Difference is computed as the value tabulated from the microdata minus the reported census value, over the “population at risk” – defined as the total for the category. For example, the % Difference for Male Illiterate is the tabulated value for Male Illiterate minus the census value for Male Illiterate, divided by the census value for Total Illiterate.

## Analysis of Literacy by Sex

The Census Bureau reports tabulations of Literacy by Sex for the population aged 10 and older. The statistics for Connecticut in 1880, 1890, and 1900 were compared to the same crosstabs generated from the IPUMS 10% 1880 and 5% 1900 samples (with weights).

Table —Literacy by Sex, population of age > 10, 1880–1900

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  | 1880 | 1890 (Synthetic) | 1900 |
| Male | Illiterate | Census tabulation | 12039 | 15233 | 20929 |
|  |  | Microdata | 11973 | 14273 | 19743 |
|  |  | Difference | -66 | -960 | -1186 |
|  |  | % Difference | -0.23% | -2.98% | -2.76% |
|  | Total | Census tabulation | 242392 | 300675 | 365130 |
|  |  | Microdata | 242020 | 298823 | 364358 |
|  |  | Difference | -372 | -1852 | -772 |
|  |  | % Difference | -0.07% | -0.30% | -0.11% |
| Female | Illiterate | Census tabulation | 16385 | 16961 | 22044 |
|  |  | Microdata | 15912 | 17737 | 21054 |
|  |  | Difference | -473 | 776 | -990 |
|  |  | % Difference | -1.66% | 2.41% | -2.30% |
|  | Total | Census tabulation | 254911 | 309155 | 365324 |
|  |  | Microdata | 254128 | 303558 | 367632 |
|  |  | Difference | -783 | -5597 | 2308 |
|  |  | % Difference | -0.16% | -0.92% | 0.32% |
| Total | Illiterate | Census tabulation | 28424 | 32194 | 42973 |
|  |  | Microdata | 27885 | 32010 | 40797 |
|  |  | Difference | -539 | -184 | -2176 |
|  |  | % Difference | -1.90% | -0.57% | -5.06% |
|  | Total | Census tabulation | 497303 | 609830 | 730454 |
|  |  | Microdata | 496148 | 602381 | 731990 |
|  |  | Difference | -1155 | -7449 | 1536 |
|  |  | % Difference | -0.23% | -1.22% | 0.21% |

The % difference values are plotted in Figure 1 to show the deviation from the census statistics for each of the IPUMS 10% 1880 sample, the synthetic 1890 dataset, and the IPUMS 5% 1900 sample. The 1890 dataset is further from the census statistics in comparison to the 1880 and 1900 samples, for most of the statistics. The comparison for Total Illiterate is deceptively good for 1890, because the 1890 sample is over for males and under for females, and these cancel each other out.

Figure —Percentage difference between microdata tabulation and census statistics, literacy by sex for population age > 10, 1880–1900

## Occupational Class by Sex by Year

A more interesting comparison is by Occupational Class. Occupational Class groups the two hundred-some occupational categories (the exact number of occupations varies from decade to decade) into six broader classes—Agricultural, Professional Service, Domestic and Personal Service, Trade and Transportation, and Manufacturing and Mechanical. The way in which these classes are constructed varies somewhat during the period of interest—in 1880 and 1890 for example, fishermen, oystermen, miners, and quarrymen were considered Agricultural, but in 1900 they were reclassified as Manufacturing and Mechanical. For the purposes of this analysis, occupations were grouped the way that Census did it for each decade, as we’re interested in how well the microdata matches the tallies reported by Census.

Table —Occupational Class by Sex, for employed persons age > 10, 1880–1900

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  | 1880 | 1890 (Synthetic) | 1900 |
| Male | Ag, Fishing, Mining | Census tabulation | 46943 | 47909 | 43247 |
|  |  | Microdata | 49218 | 51070 | 43622 |
|  |  | Difference | 2275 | 3161 | 375 |
|  |  | % Difference | 1.18% | 1.29% | 0.13% |
|  | Professional Service | Census tabulation | 5492 | 7509 | 10327 |
|  |  | Microdata | 4871 | 6840 | 10268 |
|  |  | Difference | -621 | -669 | -59 |
|  |  | % Difference | -0.32% | -0.27% | -0.02% |
|  | Domestic and Personal Service | Census tabulation | 25762 | 35166 | 45022 |
|  |  | Microdata | 22512 | 29590 | 41075 |
|  |  | Difference | 3250 | 5576 | 3947 |
|  |  | % Difference | 1.69% | 2.27% | 1.33% |
|  | Trade and Transportation | Census tabulation | 29897 | 44457 | 61656 |
|  |  | Microdata | 27448 | 40942 | 60907 |
|  |  | Difference | -2449 | -3515 | -749 |
|  |  | % Difference | -1.27% | -1.43% | -0.25% |
|  | Manufacturing and Mechanical | Census tabulation | 84659 | 110593 | 136719 |
|  |  | Microdata | 81425 | 104900 | 133118 |
|  |  | Difference | -3234 | -5693 | -3601 |
|  |  | % Difference | -1.68% | -2.32% | -1.21% |
|  | Total | Census tabulation | 192753 | 245634 | 296971 |
|  |  | Microdata | 185474 | 233342 | 288990 |
|  |  | Difference | -7279 | -12292 | -7981 |
|  |  | % Difference | -3.78% | -5.00% | -2.69% |
| Female | Ag, Fishing, Mining | Census tabulation | 97 | 707 | 1549 |
|  |  | Microdata | 227 | 769 | 1784 |
|  |  | Difference | 130 | 62 | 235 |
|  |  | % Difference | 0.27% | 0.09% | 0.27% |
|  | Professional Service | Census tabulation | 3485 | 4976 | 6642 |
|  |  | Microdata | 3468 | 4750 | 6932 |
|  |  | Difference | -17 | -226 | 290 |
|  |  | % Difference | -0.03% | -0.32% | 0.33% |
|  | Domestic and Personal Service | Census tabulation | 17158 | 24907 | 30295 |
|  |  | Microdata | 19104 | 23760 | 28475 |
|  |  | Difference | 1946 | -1147 | -1820 |
|  |  | % Difference | 4.00% | -1.61% | -2.05% |
|  | Trade and Transportation | Census tabulation | 1065 | 4926 | 10058 |
|  |  | Microdata | 1153 | 3917 | 9659 |
|  |  | Difference | 88 | -1009 | -399 |
|  |  | % Difference | 0.18% | -1.41% | -0.45% |
|  | Manufacturing and Mechanical | Census tabulation | 26865 | 35804 | 40095 |
|  |  | Microdata | 27075 | 32125 | 41268 |
|  |  | Difference | 210 | -3679 | 1173 |
|  |  | % Difference | 0.43% | -5.16% | 1.32% |
|  | Total | Census tabulation | 48670 | 71320 | 88639 |
|  |  | Microdata | 51027 | 65321 | 88118 |
|  |  | Difference | 2357 | -5999 | -521 |
|  |  | % Difference | 4.84% | -8.41% | -0.59% |
| Total | Ag, Fishing, Mining | Census tabulation | 47040 | 48616 | 44796 |
|  |  | Microdata | 49445 | 51839 | 45406 |
|  |  | Difference | 2405 | 3223 | 610 |
|  |  | % Difference | 1.00% | 1.02% | 0.16% |
|  | Professional Service | Census tabulation | 8977 | 12485 | 16969 |
|  |  | Microdata | 8339 | 11590 | 17200 |
|  |  | Difference | -638 | -895 | 231 |
|  |  | % Difference | -0.26% | -0.28% | 0.06% |
|  | Domestic and Personal Service | Census tabulation | 42920 | 60073 | 75317 |
|  |  | Microdata | 41616 | 53350 | 69550 |
|  |  | Difference | -1304 | -6723 | -5767 |
|  |  | % Difference | -0.54% | -2.12% | -1.50% |
|  | Trade and Transportation | Census tabulation | 30962 | 49383 | 71714 |
|  |  | Microdata | 28601 | 44859 | 70566 |
|  |  | Difference | -2361 | -4524 | -1148 |
|  |  | % Difference | -0.98% | -1.43% | -0.30% |
|  | Manufacturing and Mechanical | Census tabulation | 111524 | 146397 | 176814 |
|  |  | Microdata | 108500 | 137025 | 174386 |
|  |  | Difference | -3024 | -9372 | -2428 |
|  |  | % Difference | -1.25% | -2.96% | -0.63% |
|  | Total | Census tabulation | 241423 | 316954 | 385610 |
|  |  | Microdata | 236501 | 298663 | 377108 |
|  |  | Difference | -4922 | -18291 | -8502 |
|  |  | % Difference | -2.04% | -5.77% | -2.20% |

The patterns are difficult to see in the table above, so the % difference rows are broken out into separate graphs below for male, female, and total. By this comparison, it is apparent that the 1890 synthetic microdata fits to the aggregate statistics less well than the 1880 and 1900 microdata do. The error is quite high in Female/Manufacturing and Mechanical, but for Female/Domestic and Personal Service, the 1890 synthetic microdata fits the corresponding aggregate statistic better than the IPUMS samples do.

Figure —Percentage Difference for Occupational Class, employed males age > 10, 1880–1900

For employed females in 1880 for Agriculture, Fishing, and Mining, total numbers involved are quite low—less than 100 people, so the apparent goodness of fit may be misleading.

Figure —Percentage Difference for Occupational Class, employed females age > 10, 1880–1900

Lastly, here are the results for both males and females together. As with the sex-specific breakouts, the synthetic 1890 sample fits less well than the non-synthetic 1880 and 1900 data.

Figure —Percentage Difference for Occupational Class, employed persons age > 10, 1880–1900

# Conclusions and Future Work

The results discussed above involve fitting to 1890 aggregate statistics with both state and county availability but only using them at the state level, and these statistics only involve microdata variables for race, nativity, sex, age, and birthplace. In general, the synthetic data matches the 1890 aggregate statistics quite well using the fit between IPUMS microdata samples from surrounding census years and the corresponding aggregate statistics as a basis for comparison.

A number of potential improvements to the process should result in a better dataset. These steps have not yet been implemented and represent opportunities for future work.

1. Reweight the input data so that 1880 and 1900 are equally represented. The current implementation doesn’t compensate for the fact that the 1880 and 1900 input samples are of different densities.
2. Adjust the weight variable on the input data, instead of replicating households in output and then substituting.
3. Fit at the county, state, and national levels simultaneously.
4. Incorporate all published aggregate variables, not just the ones available at state and county level. There are additional statistics available at the state and national levels that are not available at the county level. Adding these variables to the fitting process should result in a better match.

# Appendix I

Questions as transcribed from original 1890 enumeration schedule (1890 Questionnaire n.d.).

Table —Questions asked of each household member in 1890

|  |  |
| --- | --- |
| Questions asked of each household member | |
| 1 | Name. |
| 2 | Military Service during the Civil War (or widowhood of such a person). |
| 3 | Relationship to head of family. |
| 4 | Race (white, black, mulatto, quadroon, octoroon, Chinese, Japanese, or Indian). |
| 5 | Sex. |
| 6 | Age at nearest birthday. |
| 7 | Marital Status (single, married, widowed, or divorced). |
| 8 | Whether married during the census year (June 1, 1889 to May 31st, 1890). |
| 9 | Mother of how many children, and how many living. |
| 10 | Place of birth. |
| 11 | Place of birth of Father. |
| 12 | Place of birth of Mother. |
| 13 | Number of years in the United States. |
| 14 | Whether naturalized. |
| 15 | Whether naturalization papers have been taken out. |
| 16 | Profession, trade, or occupation. |
| 17 | Months unemployed during the census year (June 1, 1889 to May 31st, 1890). |
| 18 | Attendance at school (in months) during the census year (June 1, 1889 to May 31st, 1890). |
| 19 | Able to Read. |
| 20 | Able to Write. |
| 21 | Able to speak English. If not, the language or dialect spoken. |
| 22 | Whether suffering from acute or chronic disease, with name of disease and length of time afflicted. |
| 23 | Whether defective in mind, sight, hearing, or speech, or whether crippled, maimed, or deformed, with name of defect. |
| 24 | Whether a prisoner, convict, homeless child, or pauper. |
| 25 | Supplemental Schedule and Page if more detail is required on questions 22–24. |

Table —Questions asked of the household as a whole in 1890

|  |  |
| --- | --- |
| Questions asked of the household as a whole | |
| 26 | Is the home you live in hired, owned by head or by a member of the family? |
| 27 | If owned by head or member of family, is the home free from mortgage incumbrance? |
| 28 | If the head of family is a farmer, is the farm which he cultivates hired, or is it owned by him or by a member of the family? |
| 29 | If owned by head or member of family, is the farm free from mortgage incumbrance? |
| 30 | If the home or farm is owned by head of member of family, and mortgaged, give the post-office address of owner. |

Table —Aggregate variables used for fitting

|  |  |  |
| --- | --- | --- |
| Mnemonic | Description | Universe Statement |
| totpop | Total population 1890 | All |
| negtot | Total Negro population 1890 | RACE = 2 |
| chitot | Total Chinese population 1890 | RACE = 4 |
| japtot | Total Japanese population 1890 | RACE = 5 |
| indtot | Civilized Indian population 1890 | RACE = 3 |
| nbmtot | Native-born males | SEX = 1 and NATIVITY in [1, 2, 3, 4] |
| nbftot | Native-born females | SEX = 2 and NATIVITY in [1, 2, 3, 4] |
| fbmtot | Foreign-born males | SEX = 1 AND NATIVITY = 5 |
| fbftot | Foreign-born females | SEX = 2 AND NATIVITY = 5 |
| nbwmnp | Native white males/nat parents | NATIVITY = 1 and RACE = 1 and SEX = 1 |
| nbwfnp | Native white females/nat parents | NATIVITY = 1 and RACE = 1 and SEX = 2 |
| nbwmfp | Native white males/for parents | NATIVITY in [2, 3, 4] and RACE = 1 and SEX = 1 |
| nbwffp | Native white females/for parents | NATIVITY in [2, 3, 4] and RACE = 1 and SEX = 2 |
| fbwmtot | Foreign-born white males | NATIVITY = 5 and RACE = 1 and SEX = 1 |
| fbwftot | Foreign-born white females | NATIVITY = 5 and RACE = 1 and SEX = 1 |
| colmtot | Colored males | RACE > 1 and SEX = 1 |
| colftot | Colored females | RACE > 1 and SEX = 2 |
| mtot | Total male population | SEX = 1 |
| ftot | Total female population | SEX = 2 |
| nbwm520 | Native white males aged 5–20 | NATIVITY in [1, 2, 3, 4] and RACE = 1 and SEX = 1 and AGE >= 5 and AGE <= 20 |
| nbwf520 | Native white females aged 5–20 | NATIVITY in [1, 2, 3, 4] and RACE = 1 and SEX = 2 and AGE >= 5 and AGE <= 20 |
| fbwm520 | Foreign white males aged 5–20 | NATIVITY = 5 and RACE = 1 and SEX = 1 and AGE >= 5 and AGE <= 20 |
| fbwf520 | Foreign white females aged 5–20 | NATIVITY = 5 and RACE = 1 and SEX = 2 and AGE >= 5 and AGE <= 20 |
| colm520 | Colored males aged 5–20 | RACE = 2 and SEX = 1 and AGE >= 5 and AGE <= 20 |
| colf520 | Colored females aged 5–20 | RACE = 2 and SEX = 2 and AGE >= 5 and AGE <= 20 |
| nbwm1844 | Native white males aged 18–44 | NATIVITY in [1, 2, 3, 4] and RACE = 1 and SEX = 1 and AGE >= 18 and AGE <= 44 |
| fbwm1844 | Foreign white males aged 18–44 | NATIVITY = 5 and RACE = 1 and SEX = 1 and AGE >= 18 and AGE <= 44 |
| colm1844 | Colored males aged 18–44 | RACE = 2 and SEX = 1 and AGE >= 18 and AGE <= 44 |
| nbwm21 | Native white males aged 21+ | NATIVITY in [1, 2, 3, 4] and RACE = 1 and SEX = 1 and AGE >= 21 |
| fbwm21 | Foreign white males aged 21+ | NATIVITY = 5 and RACE = 1 and SEX = 1 and AGE >= 21 |
| colm21 | Colored males aged 21+ | RACE > 1 and SEX = 1 and AGE >= 21 |
| pbasia | Persons born in Asia n.s. | BPL >= 502 and BPL <= 599 and BPL != 521 |
| pbatlisl | Persons born in Atlantic Islands | BPL = 160 |
| pbaustra | Persons born in Australia | BPL = 700 |
| pbaustri | Persons born in Austria | BPL = 450 |
| pbbelg | Persons born in Belgium | BPL = 425 |
| pbcanada | Persons born in Canada/Newfoundland | BPL = 150 |
| pbcentam | Persons born in Central America | BPL = 210 |
| pbchina | Persons born in China | BPL = 500 |
| pbcubawi | Persons born in Cuba/West Indies | BPL = 250 or BPL = 260 |
| pbdenmar | Persons born in Denmark | BPL = 400 |
| pbenglan | Persons born in England | BPL = 410 |
| pbeurns | Persons born in Europe n.s. | BPL = 499 |
| pbfrance | Persons born in France | BPL = 421 |
| pbgerman | Persons born in Germany | BPL = 453 |
| pbgreece | Persons born in Greece | BPL = 433 |
| pbhollan | Persons born in Holland | BPL = 425 |
| pbhungar | Persons born in Hungary | BPL = 454 |
| pbindia | Persons born in India | BPL = 521 |
| pbirelan | Persons born in Ireland | BPL = 414 |
| pbitaly | Persons born in Italy | BPL = 434 |
| pbjapan | Persons born in Japan | BPL = 501 |
| pbluxemb | Persons born in Luxemburg | BPL = 423 |
| pbmexico | Persons born in Mexico | BPL = 200 |
| pbnorway | Persons born in Norway | BPL = 404 |
| pbnorden | Persons born in Norway/Denmark | BPL in [404, 400] |
| pbpacifi | Persons born in Pacific Islands | BPL = 710 |
| pbpoland | Persons born in Poland | BPL = 455 |
| pbportug | Persons born in Portugal | BPL = 436 |
| pbrussia | Persons born in Russia | BPL = 465 |
| pbscot | Persons born in Scotland | BPL = 411 |
| pbsoamer | Persons born in South America | BPL = 300 |
| pbspain | Persons born in Spain | BPL = 438 |
| pbsweden | Persons born in Sweden | BPL = 405 |
| pbswitz | Persons born in Switzerland | BPL = 426 |
| pbturkey | Persons born in Turkey | BPL = 542 |
| pbwales | Persons born in Wales | BPL = 412 |
| pbsea | Persons born at sea | BPL = 900 |
| wmtot | Total white males | SEX = 1 and RACE = 1 |
| wftot | Total white females | SEX = 2 and RACE = 1 |
| whtot | Total white population | RACE = 1 |

Table —Household-level variables included in extract

|  |  |
| --- | --- |
| Household Variable | Label |
| YEAR | Census Year |
| DATANUM | Data Set Number |
| SERIAL | Household Serial Number |
| HHWT | Household Weight |
| STATEICP | State (ICPSR Code) |
| STATEFIP | State (FIPS Code) |
| County | County |
| URBAN | Urban/Rural Status |
| GQ | Group quarters status |

Table —Person-level variables included in extract

|  |  |
| --- | --- |
| Person Variable | Label |
| PERNUM | Person number in sample unit |
| PERWT | Person weight |
| RELATE | Relationship to household head |
| AGE | Age |
| SEX | Sex |
| MARST | Marital status |
| AGEMONTH | Age in months |
| BIRTHMO | Month of birth |
| DURMARR | Duration of current marital status |
| CHBORN | Children ever born |
| CHSURV | Children surviving |
| RACE | Race |
| BPL | Birthplace |
| MBPL | Mother’s birthplace |
| FBPL | Father’s birthplace |
| NATIVITY | Foreign birthplace or parentage |
| CITIZEN | Citizenship status |
| YRIMMIG | Year of immigration |
| YRSUSA1 | Years in the United States |
| YRSUSA2 | Years in the United States, intervalled |
| SPEAKENG | Speaks English |
| HISPAN | Hispanic origin |
| SPANNAME | Spanish surname |
| HISPRULE | Hispanic origin rule |
| RACESIGN | Race: Single race identification |
| SCHOOL | School attendance |
| SCHLMNTH | Months in school |
| LIT | Literacy |
| LABFORCE | Labor force status |
| OCC | Occupation |
| QTRUNEMP | Quarters unemployed last year |
| MOUNEMP | Months unemployed last year |
| OCCSCORE | Occupational income score |
| SEI | Duncan Socioeconomic Index |

# Appendix II

SPSS syntax used for recoding and analysis of 1890 synthetic sample

Compute over10 = (AGE >= 10).

variable labels over10 'age >= 10 (Filter)'.

Value Labels over10 0 'Not Selected' 1 'Selected'.

Formats over10 (f1.0).

Filter by over10.

compute illiterate = (lit < 4).

value labels illiterate 0 'literate' 1 'illiterate'.

recode occ (001 thru 012, 177, 196, 207, 227, 262 = 1)

(013 thru 015, 017, 019, 021, 022, 024 thru 028, 030, 031, 038, 045, 047, 048, 049, 052, 054, 055, 058 = 2)

(018, 019, 020, 029, 032 thru 037, 039, 040, 044, 046, 050, 051, 053, 056 = 3)

(016, 023, 042, 043, 059 thru 129 = 4)

(057, 130 thru 195, 197 thru 206, 208 thru 226, 228 thru 261, 263 thru 266 = 5)

(301 thru 999 = 6)  
 (else = 7) into occ\_class.

Value Labels occ\_class 1 'Agricultural' 2 'Professional Services' 3 'Domestic and Personal Services' 4 'Trade and Transportation' 5 'Manufacturing and Mechanical' 6 'Non-Occ Response' 7 'Other'.

compute gainfully\_employed = (occ\_class < 6).

value labels gainfully\_employed 0 'not gainfully employed' 1 'gainfully employed'.

cross tables=illiterate by sex.

cross tables=sex by occ\_class

execute.

# Bibliography

*1890 Questionnaire.* n.d. http://www.census.gov/history/pdf/1890\_questionnaire.pdf (accessed November 6, 2011).

Abowd, John, and Julia Lane. "New approaches to confidentiality protection: Synthetic data, remote access and research data centers." *Privacy in Statistical Databases.* Barcelona: Springer-Verlag, 2004. 282-289.

Beckman, Richard J, Keith A Baggerly, and Michael D McKay. "Creating Synthetic Baseline Populations." *Transportation Research Part A: Policy and Practice* (Elsevier Science, Ltd) 30, no. 6 (November 1996): 415-429.

Blake, Kellee. "First in the Path of the Firemen." *Prologue Magazine*, Spring 1996.

Chawla, Nitesh V, Kevin W Bowyer, Lawrence O Hall, and W. Phillip Kegelmeyer. "SMOTE: Synthetic Minory Over-sampling TEchnique." *Journal of Artificial Intelligence Research* 16 (2002): 341-378.

Deming, W. Edwards, and Frederick F Stephan. "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known ." *The Annals of Mathematical Statistics* (Institute of Mathematical Statistics) 11, no. 4 (Dec 1940): 427-444.

Dong, Hua, and Glen Meeden. "Using Simulated Annealing to Construct Synthetic Samples." Forthcoming.

Fitch, Catherine, and Michael R Haines. "Median age at first marriage, by sex and race: 1850-1999." Table Ae481-488." In *Historical Statistics of the United States, Earliest Times to the Present: Millennial Edition*, Table Ae481-488. Cambridge University Press, 2006.

Haines, Michael R., and Inter-university Consortium for Political and Social Research. Historical, Demographic, Economic, and Social Data. "The United States, 1790-2002 [Computer file]." *ICPSR02896-v3.* Ann Arbor, MI, May 21, 2010.

Hollerith, Herman. "An Electric Tabulating System." *The Quarterly, Columbia University School of Mines*, 1889: 238-255.

Lee, Alan. "Generating Synthetic Microdata from Published Marginal Tables and Confidentialised Files." Manuscript, Department of Statistics, University of Auckland, Auckland, NZ, 2009.

Lee, Alan. *Generating Synthetic Unit-Record Data from Published Marginal Tables.* Manuscript, Department of Statistics, University of Auckland, Auckland, NZ: University of Auckland, 2007.

Odersky, Martin, et al. *An Overview of the Scala Programming Language, 2nd Edition.* Technical Report LAMP-REPORT-2006-001, Lausanne: École Polytechnique Fédérale de Lausanne, 2006.

Reiter, Jerome. "Inference for partially synthetic, public use microdata sets." *Survey Methodology* (Statistics Canada) 29, no. 2 (December 2003): 181-188.

Reiter, Jerome. "Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation ." *Survey Methodology* (Statistics Canada) 30, no. 2 (December 2004): 235-242.

*Report on Population of the United States at the Eleventh Census: 1890—Part I.* US Department of the Interior, 1895.

Ruggles, Steven, J Trent Alexander, Katie Genadek, Ronald Goeken, Matthew B Schroeder, and Matthew Sobek. *Integrated Public Use Microdata Series: Version 5.0 [Machine-readable database].* University of Minnesota. Minneapolis, 2010.

Stephan, Frederick F. "An Iterative Method of Adjusting Sample Frequency Tables When Expected Marginal Totals are Known ." *The Annals of Mathematical Statistics* (Institute of Mathematical Statistics) 13, no. 2 (Jun 1942): 166-178.

Staff, Census Bureau History, ed. *The Hollerith Tabulator.* October 2011. http://www.census.gov/history/www/innovations/technology/the\_hollerith\_tabulator.html (accessed November 6, 2011).

Wright, Carroll. *History and Growth of the United States Census.* Washington DC: U.S. Senate. 56th Congress, 1st Session. (Senate Document 194. Serial Set 3856), 1900.