DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 1, 2016, pp. 113–127, http://dx.doi.org/10.1515/JOS-2016-0005

Constructing Synthetic Samples

Hua Dong[1] *and Glen Meeden*[2]

We consider the problem of constructing a synthetic sample from a population of interest which cannot be sampled from but for which the population means of some of its variables are known. In addition, we assume that we have in hand samples from two similar populations. Using the known population means, we will select subsamples from the samples of the other two populations which we will then combine to construct the synthetic sample. The synthetic sample is obtained by solving an optimization problem, where the known population means, are used as constraints. The optimization is achieved through an adaptive random search algorithm. Simulation studies are presented to demonstrate the effectiveness of our approach. We observe that on average, such synthetic samples behave very much like actual samples from the population of interest. As an application we consider constructing a one-percent synthetic sample for the missing 1890 decennial sample of the United States.

Key words: Sample survey; missing data; synthetic samples.

1. Introduction

The Minnesota Population Center (MPC) is an interdepartmental demography research group at the University of Minnesota. One major goal of the MPC is to create databases that can be utilized in the study of economic and social behavior. The Center has developed the Integrated Public Use Microdata Series (IPUMS-USA), which is available online and which consists, in part, of high-precision one-percent samples of the American population drawn from fifteen decennial federal censuses. A sample is composed of microdata consisting of a record for each person. These records are in turn organized into households, making it possible to study the characteristics of people in the context of their families or other coresidents. Unfortunately the complete records for the 1890 census were destroyed and now only certain summary statistics are available. For example, the family incomes for each particular family are missing but the average 1890 family income is known for many small regions of the country. Hence the Center now does not have a one-percent sample based on the complete 1890 census. In this article we will present a method that will allow a synthetic sample to be created for 1890 using the partial information from 1890 and the samples from 1880 and 1900.

Since overall the 1890 US population should not be that different from the 1880 and 1900 populations, it should be possible to construct a synthetic one-percent sample for 1890 using the one-percent samples from the 1880 and 1900 populations. The records in the synthetic sample should be chosen in such a way that their summary statistics closely

[1] Gilead Sciences, Inc. Foster City, CA 94404, U.S.A. Email: hdong@gilead.com
[2] School of Statistics, University of Minnesota, Minneapolis, MN 55455, U.S.A. Email: glen@stat.umn.edu

match the partial information for 1890. To accomplish this, we define a function that measures just how closely a possible synthetic sample matches the known population means. Since there will be many possible synthetic samples that nearly achieve the minimum of this function, our goal will not be to find an optimal synthetic sample. Instead we will be looking for one which is nearly optimal. Before considering this problem, however, we will first consider some simpler problems and present simulation results that demonstrate that our approach works well in these cases. More information about the MPC can be found at https://www.pop.umn.edu/.

From one point of view, the lack of a one-percent sample from the 1890 US population can be thought of as a massive missing-data problem where the entire sample is missing. Creating a synthetic sample is impossible unless there is some additional information that can be used, and we believe that this is indeed the case here. In the following, we will consider simpler versions of this problem and present simulation results which show that our approach can work. These simulations might suggest that our approach could be helpful in more standard missing-data problems where just some of the sample is missing. Here, however, our focus will be on the problem of creating a synthetic one-percent sample for the 1890 census.

In Section 2 we introduce a simple version of our problem. In Section 3 we propose an adaptive random search algorithm that will find a nearly optimal synthetic sample. Given an objective function defined over a large space, this technique is used to locate a point in the space whose value given by the objective function is very close to the global optimum of the objective function. In Section 4 we present simulations which show that our method works well for some simple versions of our problem. In Section 5 we use our algorithm on census data from 1900 and 1920 and partial information from 1910 to produce synthetic samples for 1910. If our approach produces good synthetic samples for this situation, then we believe it should produce a good synthetic sample for 1890 when using the 1880 and 1900 census data. Section 6 contains some final remarks.

2. A Simple Problem

Assume that there are three populations, Population 1, Population 2, and Population 3, and we believe that in some sense Population 2 is the "average" of the other two. (For our problem, the three populations can be thought as the records for 1880, 1890, and 1900 respectively.) Attached to each unit in the populations there is a pair of variables, say, X and Z. We suppose that in the three populations X and Z are related, but we make no model assumptions about this relationship. We do assume however that the mean of Z is known for the second population, that we have independent random samples from the first and third populations, and that for each unit in the samples the values of both X and Z are observed. A simple version of our problem is to use this limited information about the second population and the samples from the other two populations to construct a synthetic sample that is formed by taking elements from the other two samples and that will behave like an actual sample from the second population.

More formally, for $i = 1$ and 3 let $z_i = (z_{i,1}, \ldots, z_{i,n})$ be the observed values of Z in the random sample from population i where $n = 2m$. These will be considered fixed in what follows. If $s_1 = (i_1, \ldots, i_m)$ and $s_3 = (j_1, \ldots, j_m)$ where $1 \leq i_1 < i_2 < \cdots < i_m \leq n$ and

$1 \leq j_1 < j_2 < \cdots < j_m \leq n$ we denote the two possible subsamples of size m by z_{s_1} and z_{s_3} and denote the synthetic sample of size n formed by their union as

$$z_{s_1,s_3} = (z_{s_1}, z_{s_3}) = (z_{1,i_1}, \ldots, z_{1,i_m}, z_{3,j_1}, \ldots, z_{3,j_m}) = (z_{s_1,s_3,1}, \ldots, z_{s_1,s_3,n})$$

Finally, let μ_2 be the known mean of Z for the second population.

We need a function to measure how good the synthetic sample based on s_1 and s_3 actually is. For example, suppose that the sample mean of z_{s_1,s_3} is equal to μ_2; then we consider this to be an optimal solution for our problem. Although in theory there can be more than one such optimal solution, in practice there will almost never be even one synthetic sample that is optimal in this sense.

Let $p = (p_1, \ldots, p_n)$ be a probability vector belonging to Γ, the $n - 1$ dimensional simplex, and let

$$\Gamma_{\mu_2}(z_{s_1,s_3}) = \left\{ p \,:\, p \in \Gamma \text{ and } \sum_{i=1}^{n} p_i z_{s_1,s_3,i} = \mu_2 \right\}$$

This is the set of all probability vectors on z_{s_1,s_3} whose mean is equal to μ_2.

Let

$$h(p) = \sum_{i=1}^{n} (p_i - 1/n)^2 \quad \text{and} \tag{1}$$

$$p_{s_1,s_3} = \arg\min \{h(p) : p \in \Gamma_{\mu_2}(z_{s_1,s_3})\} \tag{2}$$

Then $h(p_{s_1,s_3})$ is our measure of how good z_{s_1,s_3} is as a synthetic sample for the second population. Given two possible synthetic samples, we will prefer the one that yields the smaller value of this function. So an optimal solution for our problem is any choice of s_1 and s_3 that gives the minimum value of $h(p_{s_1,s_3})$ over all possible synthetic samples. Our approach involves two steps. First, for a given s_1 and s_3, we need to find p_{s_1,s_3}. The second step involves searching for an s_1 and s_3 that minimize $h(p_{s_1,s_3})$.

Now for fixed s_1 and s_3, finding the value $h(p_{s_1,s_3})$ is just a standard quadratic programming problem and many software packages will have a function that will find a solution. That said, we do not know how to find explicitly the choices of s_1^* and s_3^*, which minimize $h(p_{s_1,s_3})$ over all possible synthetic samples. Instead we will conduct a random search over this space to find an approximate solution for our problem. There are $\binom{2m}{m}^2$ possible choices for s_1 and s_3, so one possibility would be to just randomly select a large number of choices for s_1 and s_3 and use the one that gives the best answer. But as m increases, the space we are searching over can become quite large and there are better search algorithms than random sampling. In the next section, we will explain our adaptive random search algorithm that seems to give sensible answers to our problem.

Finally, we note that we can include constraints on more than one variable. In particular, we could have more than one constraint involving the same variable. For example, if the mean and variance of Z were known, we could add a second constraint using its second moment.

3. The Algorithm

As we noted in the previous section, we cannot find explicitly s_1^* and s_3^*, a solution for our desired problem. On the other hand, even though an optimal solution must exist, there will usually be many other solutions that are almost as good. Our goal is not to find an optimal solution but to find just one of the possibly many synthetic samples that are nearly optimal.

To do this we will carry out a random search in the space of all possible synthetic samples. As we just noted above, one possibility would be to select at random a large number of values for (s_1,s_3) and keep the one which gives the smallest value of $h(p_{s_1,s_3})$. This is not very efficient, however, and better methods are available. One approach is to pick a starting point at random and then select at random a second point that is close to it. If the value of the function $h(p_{s_1,s_3})$ is smaller than its value at the first point, then we should move to this new point. If it is not, then we can pick another point at random from the neighborhood of the first point and repeat the process. If our function has a global minimum and no local minimums, we will eventually arrive in the neighborhood of the minimum. If there are local minimums, however, then we could get stuck at one of those points and never reach the neighborhood of the global minimum. A way to avoid this is to sometimes allow a move to a point with a larger $h(p_{s_1,s_3})$ value with positive probability. This probability should depend on both the relative sizes of the two values of the $h(p_{s_1,s_3})$s and the point we are at in the search process.

More formally, suppose we are in step l of our search, where (s_1^l, s_3^l) is our current state and we are considering moving to a new state or point in the space of synthetic samples, say (s_1^{l+1}, s_3^{l+1}). The first thing to note is that in the long run, rather than picking the new point at random, it is more efficient to pick one that is close by the current state. In our case, we will pick either s_1^l or s_3^l at random and then pick one of its entries at random and replace it by a new member, selected at random, from the appropriate full sample. Once we have determined (s_1^{l+1}, s_3^{l+1}), we can check if

$$h(p_{s_1^{l+1},s_3^{l+1}}) < h(p_{s_1^l,s_3^l}) \tag{3}$$

If this is the case then we should move to the new state. If the converse is true, then sometimes we will still want to move to the new state. This will allow us to escape from a point in the space which is a local minimum. For example, if the above equation is false then at step l one could move to the new synthetic sample with probability θ where

$$\theta = \frac{h(p_{s_1^l,s_3^l})}{h(p_{s_1^{l+1},s_3^{l+1}})}\frac{t}{a+l} \tag{4}$$

where $0 < t \leq a$ are specified constants. Note that this makes it less likely that we will move to a worse synthetic sample after lots of steps than earlier in the process. This makes sense, since we are more likely to be close to the optimal solution after many steps than when we were near the beginning of the process. We continue this process for a fixed, large number of steps and then stop. It is important to note that the "best" synthetic sample in the entire sequence need not be the state we were in when we stopped. It could have occurred much earlier and we just moved away from it. In fact, this is what usually happens in our problem.

The form of the function for θ in Equation 4 is just one of many that can be used in practice but it seemed to work well in our problem. This algorithm is just a special case of what is known as an adaptive random search. These methods have been used in a variety of problems for more than 50 years.

4. Some Simulation Examples

We conducted some simulation studies to see how our approach could work in practice.

4.1. First Example

We began by constructing three similar populations where we expect our approach to work well. Attached to each unit there are the values of two continuous variables and of two binary variables. We will denote these variables by U, V, X, and Y. The variable U will be a random sample from a gamma distribution with shape parameter γ and scale parameter one. The variable V will be a random sample from a gamma distribution with shape parameter λ and scale parameter one. The variable Y will be a random sample from a Bernoulli distribution where θ is the probability of observing a one. These variables will be independent. The final variable, X, will be constructed using logistic regression with the variable V. For a unit for which $V = v$ let $p(v)$ be the probability that its X variable has the value one. Then for our model

$$log(p(v)/(1 - p(v))) = \beta v$$

Using this model we generated three populations, each with 4,000 units. The parameter values for the three populations are given in Table 1. Note that the parameter values for the second population are the average of the other two in all cases. In addition, for each variable their distributions across the three populations are quite similar. In the second population the correlation between V and X was 0.18.

The first four rows of Table 2 give the results of 1,200 random samples, each of size 40, taken from the second population, where the population mean of each variable was estimated. For each variable the table gives the average value of the sample mean, its average absolute error, the average lower bound and average length of the usual 95% confidence interval and its frequency of coverage. The next four rows give the results when synthetic samples were constructed assuming that the true mean of V in the second population was known. These synthetic samples were also of size 40 and used 20 observations each from samples of size 40 taken from the other two populations. Note that the two results are very similar except that the confidence intervals for the synthetic

Table 1. Parameter values used to generate the three populations with four variables for the first example in Section 4

Population	γ	λ	β	θ
1	6	7	0.10	0.4
2	5	8	0.15	0.5
3	4	9	0.20	0.6

Table 2. Comparing the results from 1,200 samples of size 40 from Population 2 to 1,200 synthetic samples formed by combining samples from Populations 1 and 3 constraining on knowing the population mean of V for the first example in Section 4

	Mean	absErr	lowBd	Length	Coverage Rate
Variable	When sampling from the actual population				
U	5.02	0.29	4.34	1.37	0.932
V	7.99	0.35	7.12	1.74	0.950
X	0.75	0.056	0.62	0.27	0.943
Y	0.51	0.063	0.35	0.31	0.947
	Using synthetic samples				
U	5.08	0.26	4.38	1.40	0.962
V	7.98	0.0	7.05	1.86	1
X	0.74	0.052	0.60	0.27	0.968
Y	0.49	0.061	0.34	0.31	0.955

samples always contain the true mean of *V*. This must be the case because of the way they were formed. The constraint guarantees this.

One might wonder how the synthetic samples do when estimating population quantiles. In Table 3 the true values of the five quantiles of *V* in the second population are given. The next two rows give the average values of their standard estimates along with their average absolute error for 1,200 samples of size 40. The next two sets of two rows give the same information for synthetic samples formed by constraining on the true means of *U* and *X* in

Table 3. Comparing the results for estimating five quantiles of variable V for the first example in Section 4 when sampling from the population and when using three different constraining variables to construct synthetic samples. The results are based on 1,200 samples of size 40

	0.10 quantile	0.25 quantile	0.50 quantile	0.75 quantile	0.90 quantile
True	4.61	5.92	7.65	9.64	11.77
	When sampling from the actual population				
Mean of est	4.78	6.02	7.68	9.62	11.58
absErr	0.44	0.41	0.41	0.52	0.84
	Using synthetic samples formed by constraining on the mean of *U*				
Mean of est	4.66	5.97	7.66	9.70	11.81
absErr	0.44	0.41	0.43	0.51	0.81
	Using synthetic samples formed by constraining on the mean of *X*				
Mean of est	4.62	5.91	7.64	9.73	11.84
absErr	0.42	0.37	0.43	0.54	0.79
	Using synthetic samples formed by constraining on the mean of *V*				
Mean of est	4.58	5.87	7.61	9.67	11.79
absErr	0.37	0.30	0.28	0.33	0.60

the second population. We see that these results are very similar to those found using actual samples from the second population. Finally, the last two rows of the table give the results for synthetic samples formed by constraining on the mean of V from the second population. We see that these results are significantly better than using actual samples from the second population. What is the explanation for this perhaps surprising result?

This happens because knowing the mean of V in the second population is a very important piece of information. This fact, along with samples from two very similar populations, allows us to construct synthetic samples that on average are better than random samples drawn from the actual population. This is not a common situation, but we believe that something like this could hold for the 1890 census. Next we will consider an example where our approach does not work as well.

4.2. Second Example

Perhaps it is not so surprising that we can find good synthetic samples when the three populations are very similar. Here we will consider another example where they are less similar and in particular where a mean of the middle population is not approximately equal to the average of the means of the other two. In this example we assume that each population has two continuous variables, say U and V, which are independent and of course independent across the three populations. Suppose the mean of U in the ith population is $\mu_{u,i}$ while the mean in the ith population of V is $\mu_{v,i}$. In our simulation the values of the $\mu_{u,i}$s were equal to 8, 10, and 12, for $i = 1$, 2, and 3, while the corresponding values of the $\mu_{v,i}$s were 8, 9, and 12 respectively. All the distributions were normal with a common standard deviation equal to two. Each population contained 4,000 units and we constructed synthetic samples of size 60 for the second population using random samples of size 80 from the other two. Each synthetic sample contained 30 units from each of the other samples. We considered estimating the mean and the population quantiles of the variable U in the second population using synthetic samples based on various constraints.

The results for estimating the means are in Table 4. When constraining on the $E(V)$ our point estimate for $E(U)$ behaves just like the one based on samples from the actual population because $\mu_{u,2} = (\mu_{u,1} + \mu_{u,3})/2$. On the other hand, our point estimate for $E(V)$ when constraining on $E(U)$ performs poorly because $\mu_{v,2}$ is not the average of the other two means for V.

In addition, note that when constraining on $E(V)$ the confidence intervals for $E(U)$ are too long. In other words, even though our synthetic samples are centered properly they are too spread out. This happens despite the fact that the populations all have the same variance. So even though the average of the means for U for the first and third populations is equal to the mean of U for the second population, they are just too far apart to get good synthetic samples using just this one constraint. We can overcome this problem if we have more information about U for the second population. Suppose we know both its mean and variance; then we can constrain on both the first and second moments of U when selecting a synthetic sample. We did this in another simulation where we constrained on both $E(U)$ and $E(U^2)$ and we see from Table 4 that the length of the intervals, on average, are nearly the same as those based on random samples from U.

Table 4. Comparing the results for estimating E(U) and E(V) for the second example in Section 4 when sampling from the population and when constraining on moments of U and V. The results are based on 1,000 samples of size 60

	Mean	absErr	lowBd	Length	Coverage Rate
Variable	When sampling from the actual population				
U	9.97	0.20	9.47	1.00	0.949
V	9.01	0.20	8.52	0.99	0.950
	When constraining on $E(V)$				
U	10.02	0.19	9.31	1.42	0.998
V	9.01	0.0	8.29	1.45	1
	When constraining on $E(U)$				
U	9.98	0.0	9.27	1.42	1
V	10.07	1.06	9.34	1.45	0.11
	When constraining on $E(U)$ and $E(U^2)$				
U	9.98	0	9.48	1.00	1
V	10.11	1.10	9.35	1.51	0.068
	When constraining on $E(U)$, $E(U^2)$, $E(V)$, and $E(V^2)$				
U	9.99	0.01	9.49	1.01	1
V	9.03	0.02	8.52	1.01	1

To explore this further, we next considered the results for estimating the quantiles of U for the second population. We see from Table 5 that when constraining on either $E(U)$ or $E(V)$ our synthetic samples tend to underestimate the 0.10 quantile and overestimate the 0.90 quantile. That is, our synthetic samples are too spread out. However, when our constraints include both $E(U)$ and $E(U^2)$, our estimates based on synthetic samples perform better than random samples from the actual population. Although we have not included the simulation results, the story is the same for estimating the quantiles of V.

4.3. Third Example

We produced another example where each population consists of samples from three independent normal random variables, say U, V, and W. In all cases their standard deviations were 1.5. The means of U across the three populations were 8, 10, and 12 respectively, while for V they were 8, 9, and 12 and for W 8, 11, and 12. Each population contained 4,000 units. Then we took 1,000 samples of size 120 from the first and third populations to construct a synthetic sample for the middle population of size 40 by using 20 units each from the two samples. We constructed synthetic samples where for each variable their first two sample moments agreed with the first two population moments for the middle population. For each sample we estimated the 0.10, 0.25, 0.50, 0.75, and 0.90 population quantiles by their corresponding sample quantiles. Both the real samples and synthetic samples were approximately unbiased. Averaged over all samples and all

Table 5. Comparing the results for estimating five quantiles of variable U for the second example in Section 4 when sampling from the population and when constraining on moments of U and V. The results are based on 1,000 samples of size 60

	0.10 quantile	0.25 quantile	0.50 quantile	0.75 quantile	0.90 quantile
True	7.43	8.66	10.00	11.31	12.53
	When sampling from the actual population				
Mean of est	7.49	8.68	10.00	11.28	12.45
absErr	0.35	0.28	0.24	0.29	0.35
	When constraining on $E(V)$				
Mean of est	6.36	7.83	10.05	12.08	13.61
absErr	1.07	0.85	0.24	0.77	1.08
	When constraining on $E(U)$				
Mean of est	6.40	7.87	9.98	12.02	13.61
absErr	1.04	0.79	0.23	0.71	1.08
	When constraining on $E(U)$ and $E(U^2)$				
Mean of est	7.49	8.68	10.00	11.28	12.33
absErr	0.23	0.20	0.14	0.11	0.28
	When constraining on $E(U)$, $E(U^2)$, $E(V)$, and $E(V^2)$				
Mean of est	7.55	8.36	9.85	11.47	12.70
absErr	0.18	0.31	0.17	0.17	0.22

quantiles, the average absolute error for the real samples was 0.28 and for the synthetic samples 0.17. We repeated this example but now using a standard deviation of 2 instead of 1.5. In this case, averaged over all samples and all quantiles the average absolute error for the real samples was 0.28 and for the synthetic samples it was 0.20. So in both cases the synthetic samples seem to give a good picture of the unsampled populations. Once again, this shows that one can construct good synthetic samples from samples of similar populations for a population for which some true population parameters are known. How good the synthetic samples will actually be depends on how similar the populations are and how much is known about the middle population.

4.4. Fourth Example

So far we have seen that our method seems to work well when the three populations are quite similar and we are estimating means and quantiles. It is natural to wonder how our method will work if we are interested in estimating more complicated population parameters, say a regression coefficient.

Consider three populations, each of which consists of two variables X and Y. Let μ_i denote the mean of X in the ith population. In population i, X is normally distributed with mean μ_i and standard deviation 5. The distribution of Y given $X = x$ is normal with mean $50 + \beta_i x$ and standard deviation 15. All three populations will contain 4,000 units.

In the first example we let $\mu_i = 200$, 205, and 210 for $i = 1$, 2, and 3 and set β_i to be equal to 2 for all three populations. For the middle population the true value of the regression parameter was 1.99 and the correlation between X and Y was 0.58. We then took 500 random samples of size 120 from the first and third populations and for each pair of random samples found a synthetic sample of size 60 by selecting 30 units each from the two random samples where we assumed the population means of X and Y were known for the middle population. For these 500 synthetic samples, we found that the average value of their estimates for β_2 was 1.95 with average absolute error of 0.23. The average length of their 95% confidence interval was 0.53 with a frequency of containing the true value equal to 0.928. The corresponding values for 500 random samples from the middle population were 2.01, 0.31, 0.78, and 0.930. So in this example the synthetic samples perform very well.

In a second example we set the three μ_is equal to 200 but let the three β_is be equal to 2.00, 2.15, and 2.30 for the three populations. For this case with 500 synthetic samples formed as in the previous paragraph, we found that the average value of our estimates was 2.21 with an average absolute error of 0.70. The average length of the 95% confidence intervals was 1.71 with a frequency of containing the true value equal to 0.924. The corresponding values for 500 random samples from the middle population were 2.18, 0.33, 0.778, and 0.932. So here our synthetic samples are not doing so well. We believe this happens because in this second example the three populations are not quite as similar as those in the previous example. We find it interesting, however, that the confidence intervals based on the synthetic samples have approximately the correct coverage probability in both examples. In any case, it is clear from all our simulations that how well synthetic samples work depends not only on how similar the three populations are but also on what population parameters are being estimated.

4.5. Behavior of the Algorithm

Recall that our goal is not to find an optimal synthetic sample but just one among the large group of those who are nearly optimal. For the example in Subsection 4.3 where the standard deviation was 2, we ran our adaptive random search algorithm for 20,000 steps for each sample. We kept track of how many times it moved to a new state, the time it moved to the best state, our solution, and the time of the last move. For this example, on average, our chain moved to 96 new states, the last move occurred at step number 8,500 and our solution occurred at step number 8,055. The average of the minimum of the p_is in our solution was 0.024. Note that if our solution satisfied the constraints exactly, all the p_is would equal $1/40 = 0.025$. For the case where the standard deviation was 1.5, we ran our algorithm for 40,000 steps because there is more separation among the three populations. For this case, on average, our chain visited 174 states with the last move happening at step 19,537 and our solution occurring at step 15,533. The average of the minimum of the p_is in our solutions was 0.023.

Readers might have been questioning the need for using an adaptive random search algorithm and whether using random sampling for the searching could work just as well. For the above problem we took 100 random samples and for each sample we selected 20,000 possible synthetic samples at random. For each sample we found the value of the

vector p_{s_1,s_3}, which is the solution to the problem given in Equation 2. We then found the synthetic sample, which minimized the function h in Equation 1 over all the random samples. Averaged over these 100 samples the average minimum value of p was 0.0076. We repeated this but now included 100,000 random samples in our search. In this case, the average of the minimum values of p was 0.0084. Finally, we used 400,000 random samples for our search and found that this average was 0.0102. So even taking 20 times the number of synthetic samples that we do using our method, random sampling cannot find any synthetic samples that are as nearly balanced as ours.

Clearly our solutions are not optimal, but they are good enough for the synthetic samples to be good representations of real samples because we are constructing synthetic samples from samples of two populations that are similar to the population of interest.

5. A Simulation Using Census Data

To look at the potential performance of the proposed method for the missing 1890 population problem, we tested the proposed method on some actual census data from nearby decades. We used data supplied from the MPC for one geographical area out of a total of 56 possible geographical areas. We had approximately 2.3% samples from 1900, 1910, and 1920, which we treated as the entire populations. Associated with each individual was a vector of possible values indicating gender, age, marital status and race. We then selected 100 random samples of size about 100 from the 1900, 1910, and 1920 populations. We assumed that the sample from 1910 is missing and only the population means of five constrained variables were known. The five constrained variables were "married males", "single males", "married females", "single females", and "Negroes". We used the population means of these five variables from our 1910 population as our mean constraints and samples from the 1900 and 1920 "populations" to construct synthetic samples which contained about 50 individuals each from 1900 and 1920.

Because individuals are members of households, when a person was selected to be in our sample we included everyone in their household as well. Our samples always included at least 100 individuals. Our synthetic samples also always included at least 100 individuals. At each step of the search it was possible that we would need to remove more than one household to reduce the size of the current synthetic sample to be less than 100. By the same token, we might also need to add more than one household to ensure the number of individuals in the next synthetic sample would be at least 100. So a possible synthetic sample need not contain exactly 50 observations from 1900 and 1920 respectively.

To see what happens in this case, we constructed 100 synthetic samples using samples from 1900 and 1920 and the true 1910 population means as constraints. The results are given in Table 6.

To gain a better understanding of how the synthetic samples work we did another simulation where instead of constraining on the population means of the five variables we used sample information. That is, each time we took a sample from the 1910 population as well and used the sample means of our five constraining variables as the constraints when constructing a synthetic sample for 1910 from the samples from 1900 and 1920.

Table 6. The results for the synthetic samples for the 1910 population when the true population means are used as constraints

Variable	Mean	absErr	SD	Margin of error	Coverage rate
Married males	0.178	0.006	0.384	0.074	1.000
Divorced males	0.002	0.002	0.016	0.003	0.160
Widowed males	0.013	0.009	0.097	0.019	0.780
Single males	0.308	0.006	0.464	0.089	1.000
Negroes	0.352	0.004	0.480	0.092	1.000
Mulattoes	0.035	0.048	0.133	0.025	0.370
Married females	0.181	0.006	0.387	0.074	1.000
Divorced females	0.002	0.004	0.018	0.004	0.180
Widowed females	0.037	0.009	0.187	0.036	0.950
Single females	0.280	0.007	0.451	0.087	1.000
Foreign born	0.007	0.010	0.047	0.009	0.340
Age	23.093	1.451	18.628	3.571	0.950

Table 7 contains the results and is based on 100 samples. For comparison, we also calculated the estimates using the actual 100 samples from the 1910 population.

Note that the point estimates and the length of the confidence intervals based on the synthetic samples are very similar in the two tables. The intervals in Table 6 have better coverage rates, however. The better results occur because we are using better information, true population means, as our constraints.

For our purposes the more important fact is that the results for the synthetic samples are very similar to the results for the real samples in Table 6. This happens because the difference between the 1910 means and the average of the 1900 and 1920 means is quite small for most variables we considered. Because of the small size of our samples, it is not surprising that, especially for the rarer categories, the coverage rates of the confidence intervals can fall short of 95%. Moreover, we would expect the synthetic sample to perform poorly for a category whose 1910 mean is different from the average of its means from 1900 and 1920. For example, the coverage rate of the confidence intervals for "mulattoes" from the true 1910 sample is 0.53, which is much higher than 0.31, the coverage rate for the synthetic samples from 1900 and 1920. We believe that this stems from fact that the population proportion of "mulattoes" in 1910 is about 0.068, which is much higher than 0.028 which is the average of 1900 and 1920 population proportions. Note also that the margin of errors for the actual and synthetic samples are very similar. Because of the similarity of the three populations and the fact that the majority of the variables are binary, we see that just constraining on first moments is enough to obtain intervals with about the right length.

In our simulations, using the adaptive random search method based on Equation 4, we stopped the iterations after 5,000 steps. When trying to find one particularly good synthetic sample, there is no reason to stop after a particular number of steps. We did it here to make the running of a set of simulations easier. Since for the 1890 problem we are only interested in creating one sample, running the algorithm a long time is not a problem. However, it could take some experimentation to come up with a good choice for the values of t and a in Equation 4, as the number of variables used as constraints varies.

Table 7. A comparison of actual and synthetic samples for the census data when constraints based on sample information is used

Sample	Variable	Mean	absErr	SD	Margin of error	Coverage rate
1910	Married males	0.156	0.030	0.363	0.070	0.920
1910	Divorced males	0.001	0.002	0.008	0.002	0.080
1910	Widowed males	0.010	0.010	0.078	0.015	0.640
1910	Single males	0.333	0.042	0.471	0.091	0.950
1910	Negroes	0.325	0.101	0.450	0.087	0.510
1910	Mulattoes	0.077	0.052	0.224	0.043	0.530
1910	Married females	0.157	0.030	0.364	0.070	0.910
1910	Divorced females	0.003	0.004	0.029	0.006	0.280
1910	Widowed females	0.026	0.016	0.149	0.029	0.740
1910	Single females	0.314	0.050	0.463	0.089	0.870
1910	Foreign born	0.010	0.012	0.057	0.011	0.350
1910	Age	21.352	2.643	17.416	3.364	0.650
synthetic	Married males	0.157	0.028	0.364	0.070	0.940
synthetic	Divorced males	0.001	0.002	0.008	0.002	0.080
synthetic	Widowed males	0.009	0.009	0.074	0.014	0.620
synthetic	Single males	0.331	0.037	0.471	0.090	0.980
synthetic	Negroes	0.326	0.099	0.452	0.087	0.530
synthetic	Mulattoes	0.026	0.050	0.109	0.021	0.310
synthetic	Married females	0.159	0.027	0.366	0.070	0.960
synthetic	Divorced females	0.001	0.003	0.014	0.003	0.140
synthetic	Widowed females	0.027	0.016	0.151	0.029	0.760
synthetic	Single females	0.315	0.048	0.464	0.089	0.900
synthetic	Foreign born	0.010	0.011	0.063	0.012	0.410
synthetic	Age	21.614	2.170	17.418	3.336	0.810

Another approach to the 1890 census problem could be to try to use the information from the 1880 and 1900 censuses to create a model for the 1890 population that would then be used to generate a sensible one-percent census for 1890. Although such an approach could work, building a model for the entire US population would be a big problem. We believe, however, that the approach used here is simpler, and it effectively uses the information available in the 1880 and 1900 censuses in a simple and straightforward manner that bypasses the difficult problem of trying to construct a sensible model.

On the other hand, when constructing a one-percent sample for 1890 for a particular geographic area, historical information should be used when selecting the variables to constrain upon. These variables could depend on which area of the country you are considering. For rarer groups, you could make sure that each synthetic sample contains about the right proportion of individuals of that type. For example, if a family with a foreign-born individual is removed then it must be replaced by another family containing a foreign-born individual. If a proposed synthetic sample does not have approximately the correct mean for some variable not included in the constraining set, then one can always add this variable to the constraint set and find a new synthetic sample. Since a synthetic sample for the whole country will be made up of a collection of synthetic samples for a

large number of many small geographic areas, the approach given here should be able to construct a good synthetic sample for the 1890 population.

6. Final Remarks

Here we have considered the problem of constructing a synthetic sample from a population for which we have limited information. We proposed a novel approach that assumes the existence of two known populations which taken together are a good approximation to the missing population. We have seen in some cases that a synthetic sample can be constructed and used as a substitute for a missing sample and inferences based on it are as good as those based on the actual sample. In particular, we saw that to get synthetic samples that do a good job of estimating the quantiles of a variable one can constrain on the first two moments of the variable. To obtain the synthetic sample, we used an adaptive random search algorithm to solve an optimization problem which incorporates the available limited information about the population of interest. Simulations demonstrated the good performance of our approach for some small sample sizes.

As we have pointed out, creating a synthetic one-percent sample for the 1890 census is an extreme missing-data problem, and as far as we know this problem has never been considered in the literature. Although this is perhaps stating the obvious, we were not interested in combining or merging two data sets, a problem which has often been discussed in the literature (Kadane 2001). On the other hand, synthetic data has been considered in several contexts. It has been recommended to replace missing or censored observations with imputed or synthetic observations. In some such cases auxiliary information is used to model the missing observations. In the survey-sampling context, after a sample has been selected Hidiroglou and Laniel (2001) considered constructing synthetic variables at the estimation stage. In situations where confidentiality is an issue, Fienberg et al. (1998) considered constructing synthetic samples as part of a disclosure-avoidance methodology, but they were modifying existing samples rather than constructing new ones. Reiter (2002), Reiter (2005), and Drechsler and Reiter (2012) recommended constructing many synthetic samples and then using multiple imputation to make inferences. It was argued that valid inferences could still be made using such synthetic data. Multiple imputation is not an option for the MPC since the goal is to create a one-percent sample for the 1890 census. In a situation closer to our problem, Kohnen and Reiter (2009) considered combining information from two populations, but again they use multiple imputation to construct many synthetic samples. Meeden (2000) gives an approach to the standard missing-data problem involving constraints that is closer in spirit to what we are doing here. There, after one set of values are imputed for the missing observations, the observed and imputed values are then adjusted so that confidence intervals based on this adjusted sample will have the correct frequentist coverage probability under repeated sampling.

Another possible application of our methods is to create a synthetic sample for a population using samples from similar populations and constraints based on partial information from a sample taken from the population of interest. In one case here, we saw that such synthetic samples worked well. We have carried out other simulation studies, not included here, and observed that if the three populations are not too different such

synthetic samples behave very much like actual samples from the population. Although real data are always preferred, it seems clear to us that in some cases inferences based on synthetic data can perform almost as well as inferences based on actual data.

7. References

Drechsler, J. and J. Reiter. 2012. "Combining Synthetic Data with Subsampling to Create Public Use Microdata Files for Large Scale Surveys." *Survey Methodology* 38: 73–79.

Fienberg, S., U. Makov and R. Steele. 1998. "Disclosure Limitation Using Perturbation and Related Methods for Categorical Data." *Journal of Official Statistics* 14: 485–502.

Hidiroglou, M. and N. Laniel. 2001. "Sampling and Estimation Issues for Annual and Subannual Canadian Business Surveys." *International Statistical Review* 69: 487–504. Doi: http://dx.doi.org/10.1111/j.1751-5823.2001.tb00471.x.

Kadane, J. 2001. "Some Statistical Problems in Merging Datasets." *Journal of Official Statistics* 17: 423–433.

Kohnen, C. and J. Reiter. 2009. "Multiple Imputation for Combining Confidential Data Owned by Two Agencies." *Journal of the Royal Statistical Society, Series A* 172: 511–528. Doi: http://dx.doi.org/10.1111/j.1467-985X.2008.00574.x.

Meeden, G. 2000. "A Decision Theoretic Approach to Imputation in Finite Population Sampling." *Journal of the American Statistical Association* 95: 586–595. Doi: http://dx.doi.org/10.1080/01621459.2000.10474234.

Reiter, J. 2002. "Satisfying Disclosure Restrictions with Synthetic Data Sets." *Journal of Official Statistics* 18: 531–543.

Reiter, J. 2005. "Releasing Multiply Imputed, Synthetic Public Use Micro-Data: An Illustration and Empirical Study." *Journal of the Royal Statistical Society, Series A* 168: 185–205. Doi: http://dx.doi.org/10.1111/j.1467-985X.2004.00343.x.

Received January 2013
Revised February 2015
Accepted November 2015