ACS dataset Synthesis and Evaluation

Derek Burk, Kara Fisher, Dan Ehrlich

1/28/2022

Generate initial synthetic dataset

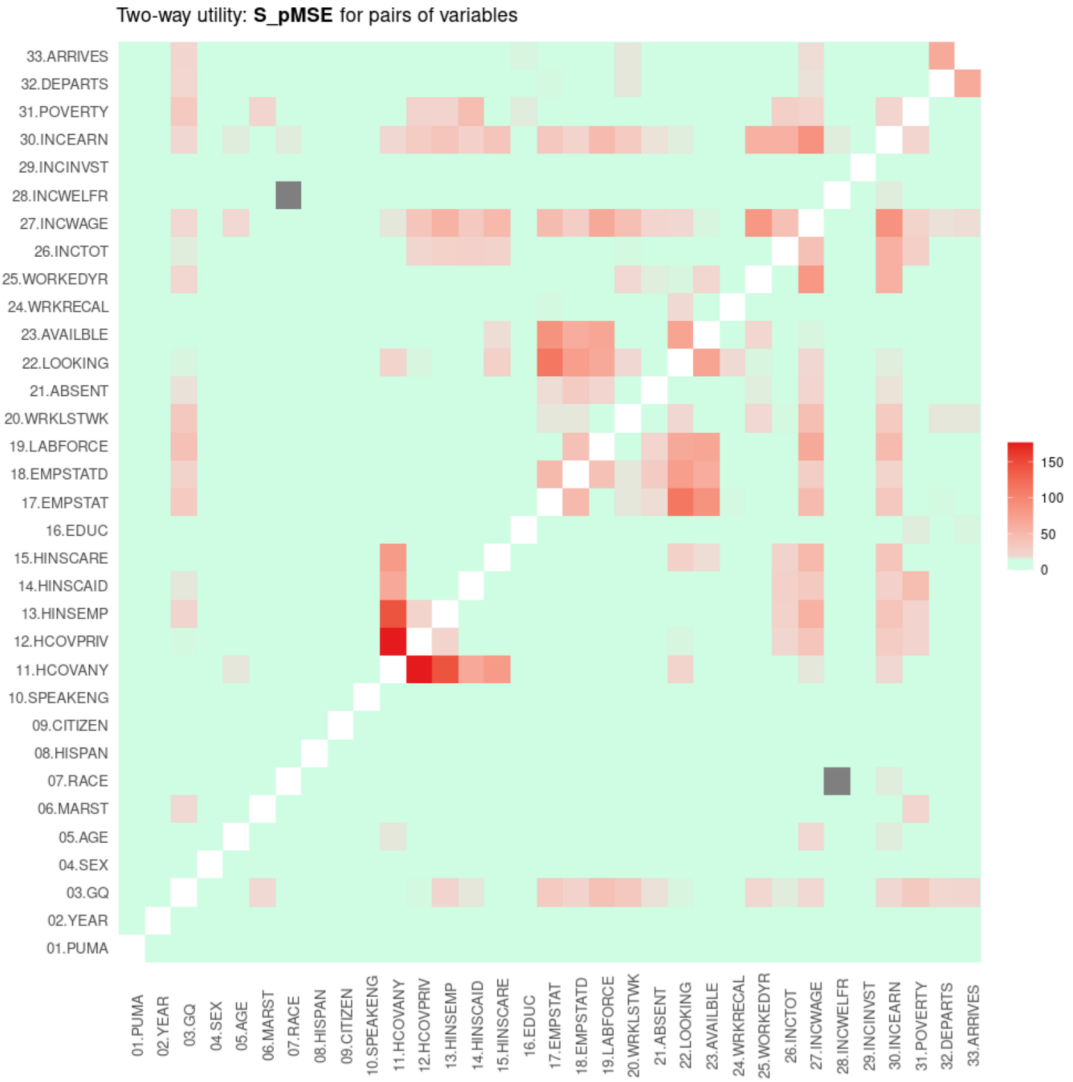
We begin by using Synthpop to generate an initial, full naive synthetic dataset. To facilitate analyses, we conduct this on a 1% sample (~10k records) of the original ACS data. We also drop the person and household level weights (PERWT, HHWT) of the original dataset, and a simulated variable from a different challenge (sim_individual_id).

We chose to stratify the data by geographic identifier, PUMA, and adjusted minimum stratumsize and factor levels, setting each to 10.

```
library(synthpop)

library(dplyr)

seed <- 13

acs_dat <- read.csv("data/IL_OH_10Y_PUMS.csv")

acs_dat <- acs_dat %>% slice_sample(prop = .01)

acs_dat <- acs_dat %>% select(-c(X, PERWT, HHWT,sim_individual_id))


synth_0 <- syn.strata(acs_dat,seed = seed, strata = "PUMA",minstratumsize =
10, minnumlevels = 10)
```

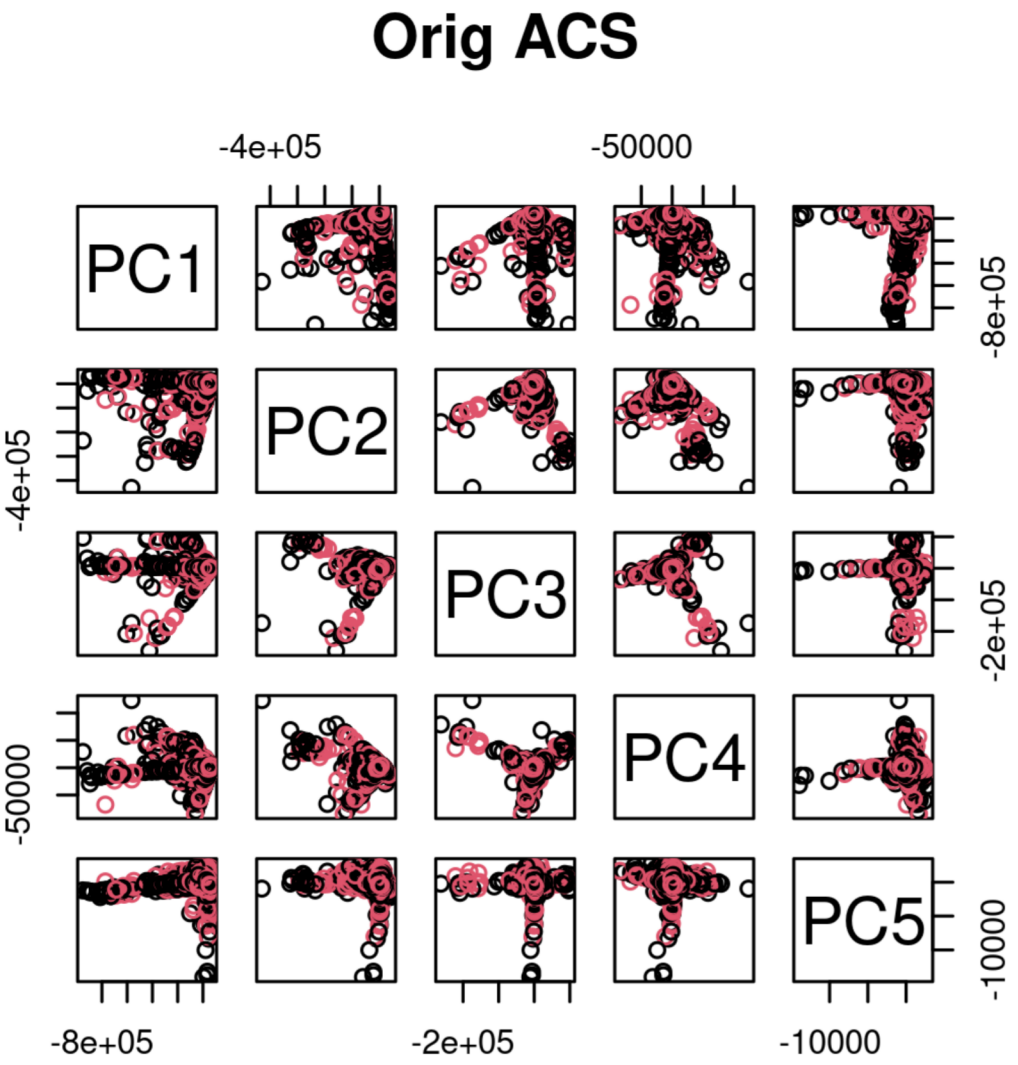
Evaluate initial synthetic dataset

We conduct an initial evaluation using the *utility.tables()* function of Synthpop, and find relatively low S_pMSE values for most two-way comparisons. S_pMSE values of 10 or less shown in green.



Evaluate utility

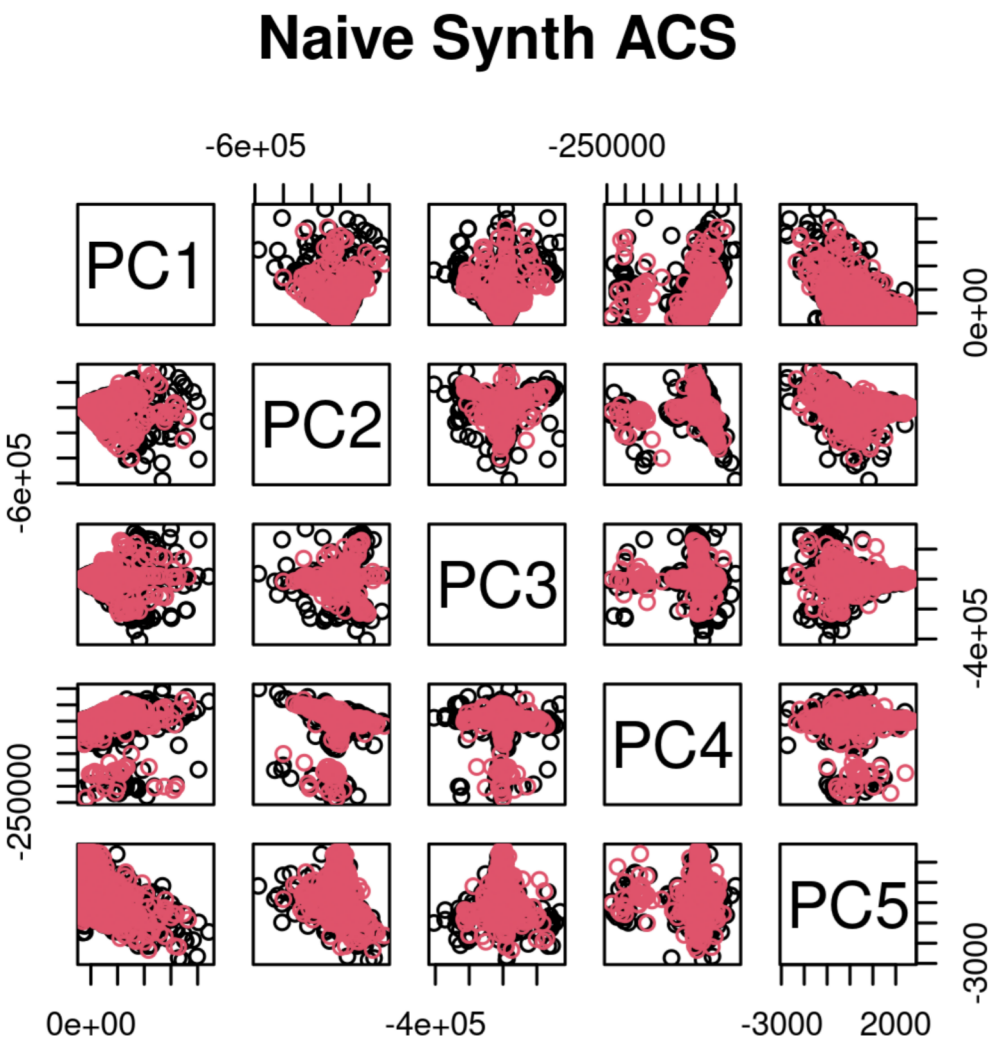
In order to evaluate the utility of such a large dataset, we apply Principal Component Analysis to better understand the multivariate structure of the data. PCA scatterplots (colored by higher-level geography, IL - BLACK, OH - RED) of the Original ACS data show a highly patterned dispersal. Individual observations radiate outward along one or more distinct trajectories from a relatively small central cluster of individuals.



In contrast, the Naive Synthetic data produces a more dispersed distribution across all PC axes. While this rounder dispersal is closer to a multivariate distribution, it is strikingly off base from the Original Data. Knowing what to look for in the original data, we can observe

the synthetic data moving in similar directions yet with significant spread not seen in the original data. This spread indicates that individuals in the synthetic data poses combinations of values that are explicitly not seen in the original data.

Failing to capture the precise data structure, and instead producing a more generalized pattern may be expected for a naive synthesis. However, this process seems to have **introduced new structure to the data**. PC 4 distinguishes data into two distinct, sizable clusters, which are **not present** in the original data. Additionally, data are colored by higher-order geography (IL, BLACK; OH, RED), yet compared to the original data, we seem to be over-representing individuals from OH, relative to IL.



Retooling to Improve Utility

Unfortunately, we did not have time to complete retooling for this dataset. The large number of variables, and complex relationships between them, require a large amount of time and effort to understand. This understanding is essential in order to specify less-naive parameters in order to achieve more realistic synthetic data.

To begin retooling, we begin by subdividing the variables by type. We identify 4 major variable types (demographic, employment, income, and healthcare). Breaking these down allows a more manageable number of variables to be considered at one time, using a partial synthesis.

Using a partial synthesis for each of these groups, we can look more closely at specific variables and use parameters to define logical rules, associated values, and define functions for passive inference. For example, LABFORCE variable can be defined by the EMPSTAT variable. An EMPSTAT value of 1 (employed) defines LABFORCE as true (2). All other values of EMPSTAT should result in a LABFORCE of 1(false). However, in reality there are two employment status variables, EMPSTAT, EMPSTATD. EMPSTATD is similar to EMPSTAT but uses 2 digit codes to add some specificity within certain categories. In reality, this means that EMPSTAT is simply a summarized version of the more detailed EMPSTATD. We suspect the *syn.nested()* function would be appropriate to first synthesize the nested values of EMPSTATD/EMPSTAT before using just one of these to define LABFORCE.

This small example is indicative of the types of complex relationships present within the ACS dataset. While we are confident that retooling will improve upon the synthetic quality of the synthetic data, there are several complications. Being very intentional about order of variables, and synthesis method employed should help, however there are actually numerous overlapping strata within this dataset, including the possibility of repeated measures.

The ACS dataset seems to include repeated observations of the same individuals across an 8 year timescale. In reality, individuals are likely present for some, but not all, of the years. There did not seem to be any immediately apparent ways to treat the temporal aspect of this data. As such, we chose to ignore it for the challenge, though we believe this presents a serious challenge to overcome in the creation and utility of synthetic data.

Use Cases

Our present synthetic data is not suitable for full release at this time. Each individual variable does match the source data reasonably well. As such, the synthetic data **may** be

suitable for certain targeted univariate analysis. Indeed, the overall low S_pMSE values seen in *utility.tables()* supports their use for simple two-way comparisons as well. These data may also serve reasonably well as a teaching resource, or for programmatic testing. However, the synthetic data fails to capture the full relationship between all variables. As such results from any targeted analysis, classroom analysis, or software testing should **not be considered representative or informative**.

While retooling should help improve upon these uses, the temporal component with repeated observations poses a substantial challenge to synthesizing accurately. A substantial amount of time would need to be devoted to retooling in order to achieve utility in a full data release.