Data from the American Community Survey for 2018

Working with a larger data set to optimise synthesis

As you might expect run times tend to be larger for a bigger data set

These data were selected from a data set made available for the Temporal Map component of the 2020 NIST Synthetic Data Challenge. This data was originally drawn from the **IPUMS harmonized archive of ACS data,** They include survey data, including demographic and financial features, representing a subset of IPUMS American Community Survey data for Ohio and Illinois. The data provided for the challenge included a large feature set of quantitative survey variables along with simulated individuals (with a sequence of records across years), time segments (years), and map segments (PUMA). It had already undergone some anonymisation procedures before it was made available. The full data set covered several years of data.

We have selected only the data from 2018, giving over 140 K records for 22 variables. Note that for large data sets the utility measures , that are relative to the stochastic errors in the data, tend to give larger values when the tables don't look too bad.

The data is an R data frame **`acs2018`** with the variables:listed on the next page saved as **acs2018.Rdata**

All categorical variables have been made into factors with levels assigned. Some details on next page

Practical 2 b)

The practical consists of using different methods to synthesise this data set. You can do it any way you want but here are some suggestions, code following them is in **synthesise_ACS2018.R.**

Suggested tasks

- Take a look at the data and try to understand the variables a little. Use codebook.syn to check the data.
- Are there any variables with many levels that you might want to put at the end of the visit.sequence or drop all together? Try it on your machine. I had to remove the variable to get it to run in just a few minutes.
- Synthesise the whole data set (or the slightly reduced one) with cart and with ctree and evaluate their utilities.
- Which variables by each method seem to give poor univariate utility, and which pairs give poor bivariate utility by each method? Are they the same or different?
- Now take one of these two syntheses and attempt to improve it by
    - Investigating plots or tables for difficult variables or pairs with **`multi.compare`**
    - Changing the synthesis order
    - Change the complexity parameters of the cart model
    - Stratify by one or more variables
    - Put a group of awkward variables together at the start and synthesise by "**`catall`**"

Details of the variables in acs2018

- PUMA (str) — Identifies the Public Use Microdata Area (PUMA) where the housing unit was located.
- SEX (uint8) — Reports whether the person was male or female.
- AGE (uint8) — Reports the person's age in years as of the last birthday.
- MARST (uint8) — Gives each person's current marital status.
- RACE (uint8) — Reports what race the person considers himself/herself to be.
- HISPAN (uint8) — Identifies persons of Hispanic/Spanish/Latino origin and classifies them according to their country of origin when possible.
- CITIZEN (uint8) — Reports the citizenship status of respondents, distinguishing between naturalized citizens and non-citizens.
- SPEAKENG (uint8) — Indicates whether the respondent speaks only English at home, and also reports how well the respondent, who speaks a language other than English at home, speaks English.
- HCOVANY, HCOVPRIV, HCOVEMP, HINSCAID, HINSCARE – Reports what type of health insurance held.
- EDUC (uint8) — Indicates respondents' educational attainment, as measured by the highest year of school or degree completed.
- EMPSTAT, LABFORCE, WRKLSTWK, ABSENT, LOOKING, AVAILBLE, WORKEDYR (uint8) — Indicates whether the respondent was a part of the labor force (working or seeking work), whether the person was currently unemployed, their work-related status in the last week, whether they were informed they would be returning to work (if not working in the last week), and whether they worked during the previous year.
- INCTOT (int32) — Reports each respondent's total pre-tax personal income or losses from all sources for the previous year, as well as the income from wages, welfare, investment, and wages or a person's own business or farm, respectively.

This data was originally drawn from the **IPUMS harmonized archive of ACS data,** and so additional details of the code values for all variables can be found on their website here:

- **Employment Variables:** https://usa.ipums.org/usa-action/variables/group/work
- **Health Insurance**: https://usa.ipums.org/usa-action/variables/group?id=insurance
- **Demographic Variables**: https://usa.ipums.org/usa-action/variables/group?id=demog
- **Education Variables**: https://usa.ipums.org/usa-action/variables/group?id=educ
- **Income Variables**: https://usa.ipums.org/usa-action/variables/group?id=income
- **Commute Variables**: https://usa.ipums.org/usa-action/variables/group?id=place_work