ACS - Fully Conditional Specification (FCS)

Evaluation Synthetic Data Creation

Steffen Moritz, Hariolf Merkle, Felix Geyer, Michel Reiffert, Reinhard Tent (DESTATIS)

January 31, 2022

Executive Summary

We applied **Fully Conditional Specification** (FCS) on the **ACS** dataset via the **synthpop** package. Our final FCS model was with **CART**. FCS seemed to us like a very interesting option for certain use cases. Out of all different methods we tested (FCS, IPSO, GAN, Simulation, Minutemen) FCS produced the **best usability** results for **ACS**. Of course by providing good usability there is usually a **trade-off** with the privacy measures. Only **IPSO** was behind **FCS** in our main privacy metrics. However, overall the privacy measures were still acceptable for some use cases.

We found **Fully Conditional Specification** (FCS) algorithm is not only very useful for the generation of synthetic "SAT"-data, it is also very suitable to generate synthetic data from the ACS dataset. Basically all marginal distributions are aligning. With one extreme exception, the S_pMSE shows only values below 10. The Pearson correlation coefficients for binary and (semi-)continuous variables are also practically identical to those of the original dataset. Also the absolute difference in densities and the Bhattacharyya distance support the overall impression. Only Mlodak's information loss criterion indicates this synthetic dataset as mediocre useful (based on 100k sample).

USE CASE RECOMMENDATIONS

Releasing_to_Public	Testing_Analysis	Education	Testing_Technology
NO	YES	NO	MAYBE

Since the usability results were clearly the best, FCS is interesting for every use case that requires high usability. Because of the trade-off with privacy we probably would only supply the FCS synthetic data to trusted partners. So **Testing Analysis**, where trusted researchers can develop and test their models before clearance for the actual microdata seems like a very good fit. **Releasing to Public** and

Education mostly wouldn't fit because of privacy issues. Internal **Technology Testing** could be a possible use case, but for most of these testing cases there are probably easier options requiring less computational power to provide synthetic data.

Dataset Considerations

When deciding, if data is released to the public it is of utmost importance to define, **which variables** are the most relevant in terms of **privacy and utility**. This process is very **domain and country** specific, since different areas of the world have different privacy legislation and feature specific overall circumstances. This step would require input and discussions with actual domain experts. Since we are foreign to US privacy law, the assumptions made for the Synthetic Data Challenge are basically an **educated guess** from our side. From a utility perspective it is important to know which variables and correlations are **most interesting** for actual users of the created synthetic dataset. Different use cases might require focus on different variables and correlations. We could not single out a most important variable, thus in our utility analysis we decided to focus on the overall utility and not to prioritize a specific variable. We decided to remove the first column of the **ACS** dataset, since it only contains column numbers and hence does not need to be altered by any means. From a privacy perspective it has to be decided, which variables are **confidential** and which are **identifying**. As already mentioned, specifying this depends on multiple factors e.g. regulations or also other public information, that could be used for **de-anonymization**. For our analysis, we made the following assumptions: Of course any information about **income** has to be considered as **confidential**, otherwise publishing income statistics would be a way easier task for NSOs than it actually is. So `INCTOT`, `INCWAGE`, `INCWELFR`, `INCINVST`, `INCEARN` and `POVERTY` are treated as confidential variables. Additionally the times a person is not at home also is an information that encroaches in personal right and might be to the respondents detriment e.g. by burglars. The features HHWT and PERWT are weights that only present information about the way the dataset was created and hence are neither confidential nor identifying. All the other information (like Sex, Age, Race...) contain observable information and hence, in our opinion, are **identifying variables**.

Method Considerations

We decided to use the **FCS** method for multiple reasons. For one the use of the FCS method is **fairly simple** and straightforward, since no prior knowledge of the relation between the data is necessary for fitting a first model. Secondly, the R package **synthpop** already comes with a good implementation of the method. Thirdly, and maybe most importantly, the method can be used for nearly all types of datasets and yield meaningful results.

For our first approach we chose to use the default settings of the method, i.e. the order of synthesisation of the variables in ascending order, and using the Classification and Regression Tree (CART) machine learning model for each variable. Since computing time increases sharply, when applied to larger datasets, applying FCS to the ACS dataset was rather challenging. Our first idea was to find out which of the features correlate, in order to decide which of these features should be fed to the algorithm simultaneously. Unfortunately we found a rather complex network of connections between many of the variables and hence had the problem, that it was not clear how the dataset could be split up in order to reduce it's complexity without losing correlations between the data. So we decided to first use a subsample of the dataset, by randomly drawing 100 000 data points (approx.

10%) of the original dataset and to apply the FCS method on all features. Again we chose to use the default settings of the method, i.e. the order of synthesisation of the variables in ascending order, and using the Classification and Regression Tree (CART) machine learning model for each variable. The computation time for this subsample only took a little more than one hour. Unfortunately, we couldn't finish a run on the complete dataset, thus have to rely on the sample.

Privacy and Risk Evaluation

Disclosure Risk (R-Package: synthpop with own Improvements)

Our starting point was the **matching of unique records**, as described in the disclosure risk measures chapter of the starter guide. The synthpop package provides us with an easy-to-use implementation of this method: `replicated.uniques`. However, one downside of just using `replicated.uniques` is that it does **not consider almost exact matches in numeric variables**. Imagine a data set with information about the respondents' income. If there is a matching data point in the synthetic data set for a unique person in the original data set, that only differs by a slight margin, the original function would not identify this as a match. **Our solution** is to borrow the notion of the **p% rule** from **cell suppression methods**, which identifies a data point as critical, if one can guess the original values with **some error of at most p%**. Thus, **our improved risk measure** is able to evaluate disclosure risk in numeric data. Our Uniqueness-Measure for **"almost exact"** matches provides us with the following outputs:

- **Replication Uniques** | Number of unique records in the synthetic data set that replicates unique records in the original data set w.r.t. their quasi-identifying variables. In brackets, the proportion of replicated uniques in the synthetical data set relative to the original data set size is stated.
- **Count Disclosure** | Number of replicated unique records in the synthetical data set that have a real disclosure risk in at least one confidential variable, i.e. there is at least one confidential variable where the record in the synthetical data set is "too close" to the matching unique record in the original data set. We identify two records as "too close" in a variable, if they differ in this variable by at most p%.
- **Percentage Disclosure** | Proportion of the number of replicated unique records in the synthetical data set that have a real disclosure risk in at least one confidential variable relating to the original data set size. For our selected best parametrized solution in this method-category, we got the following results:

Replication.Uniques	Number.Replications	Percentage.Replications
292	154	0.154

Perceived Disclosure Risk (R-Package: synthpop)

Unique records in the synthetic dataset may be **mistaken for unique records** based on the fact that **only the identifying variables match**. This can lead to problems, even if the associated confidential variables significantly differ from the original record. E.g. people might assume a certain income for a

person, because they believe to have identified her from the identifying variables. Even if her real income **is not leaked** (as the confidential variables are different), this assumed (but wrong) information about him **might lead to disadvantages**. The **perceived risk** is measured by matching the unique records among the quasi-identifying variables (compare with non-confidential variables in Section “Dataset Considerations”). We applied the method `replicated.uniques` of the synthpop package. There is no fixed threshold that must not be exceeded in this measure, however, a smaller percentage of unique matches (referred to as Number Replications) is preferred to minimize the perceived disclosure risk. These are the results variables for perceived disclosure risk:

- **Number Uniques** | Number of unique individuals in the original data set.
- **Number Replications** | The number of matching records in the synthetic data set (based only on identifying variables). This is the number of individuals, which might perceived as disclosed (real disclosures would also count into this metric).
- **Percentage Replications** | The calculated percentage of duplicates in the synthetic data. For our selected best parametrized solution in this method-category, we got the following results:

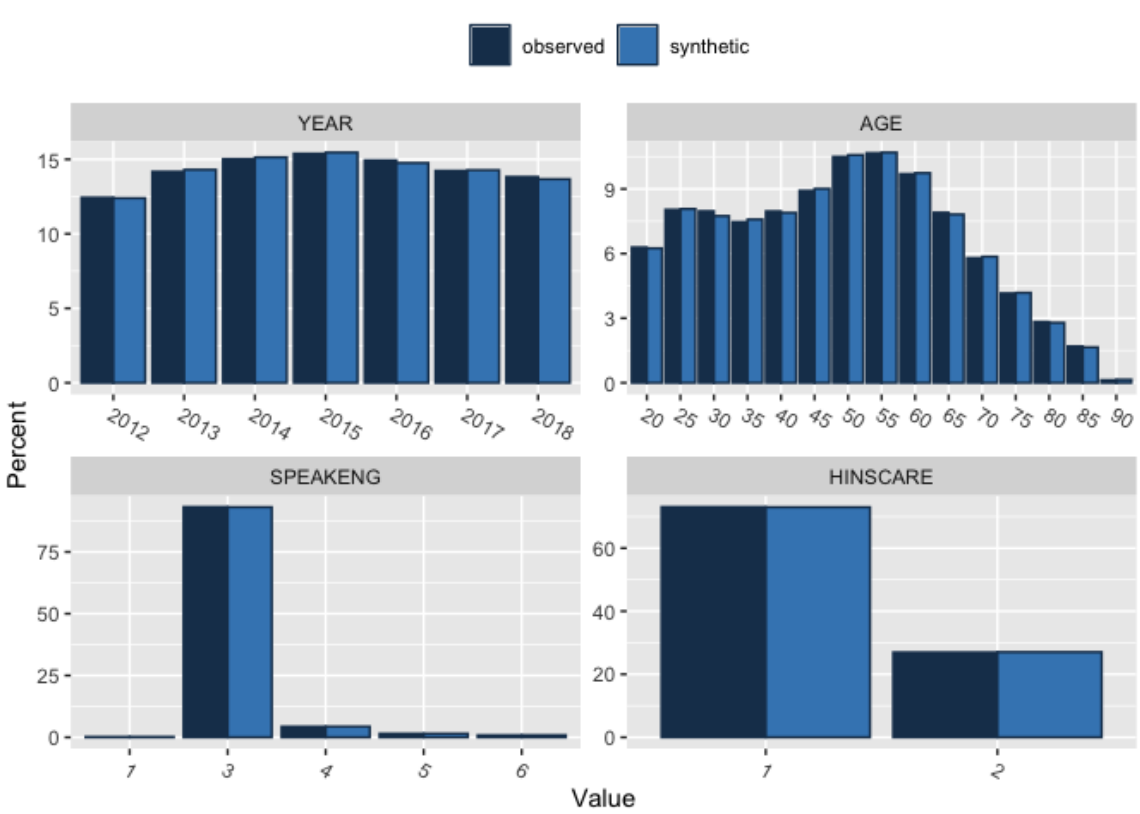
Metric	Number.Uniques	Number.Replications	Percentage.Replications
Perceived Risk	1e+05	12	0.012

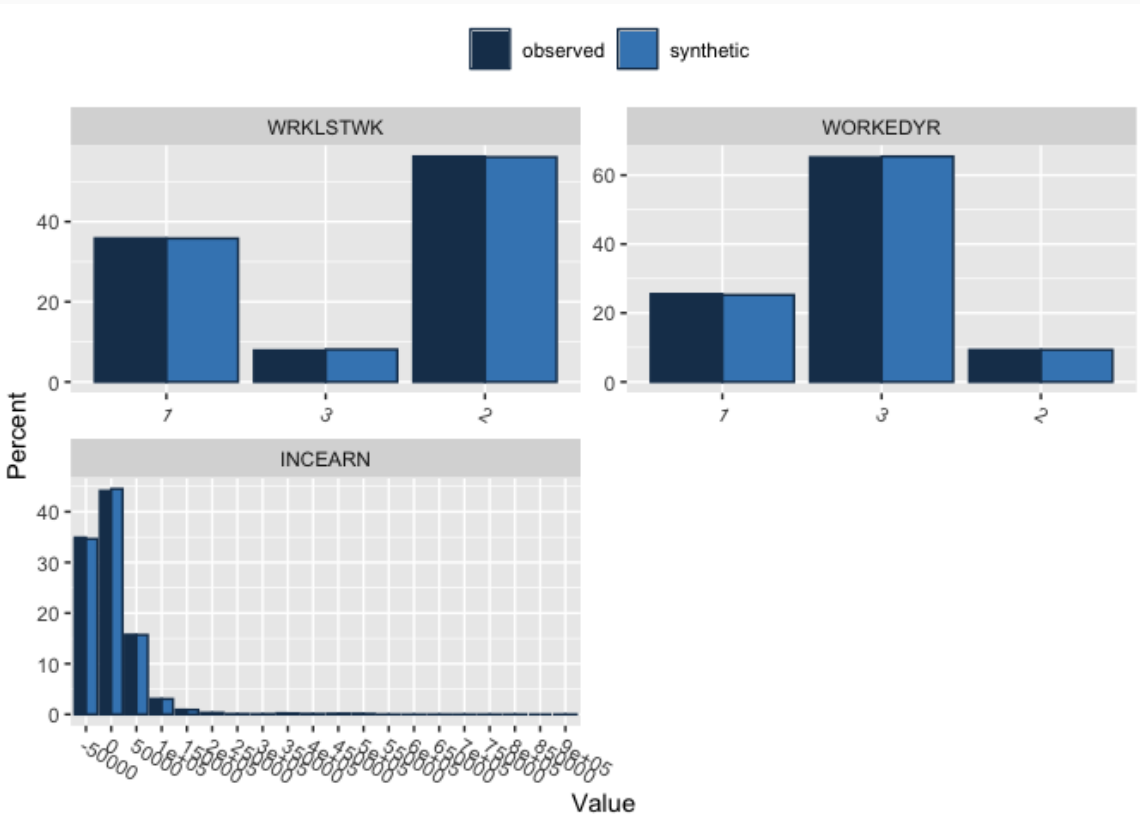
Utility Evaluation

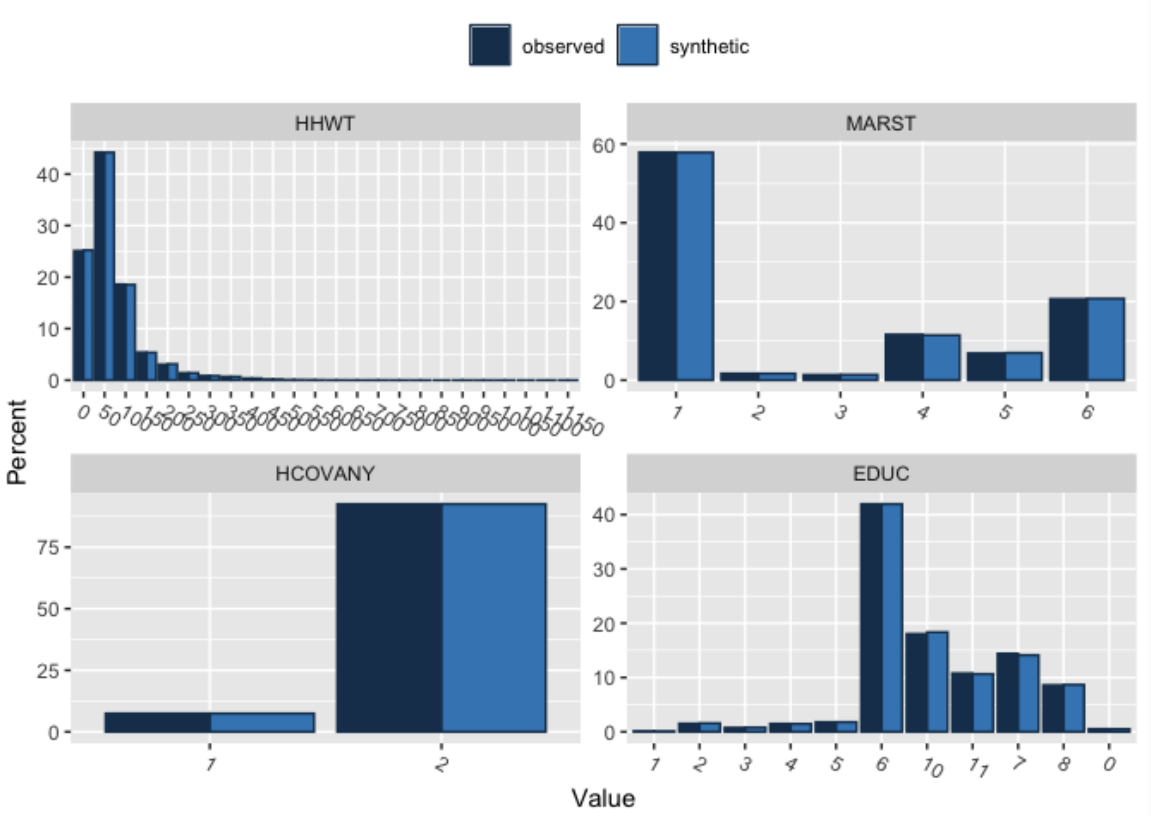
Different utility measures are applied in this section. These utility measures are the basis of utility evaluation for the generated synthetic dataset. The R packages synthpop, sdcMicro and corrplot were used to compute the following metrics. We do not use tests incorporating significance here. Confidence intervals in large surveys often tend to be extremely small so many slight differences appear to be significant. We do not consider the variable PUMA for our utility evaluation. During the ACS reports, some minor changes in availability regarding plots might occur. This is caused by the application of standardised scripts on different synthetic datasets.

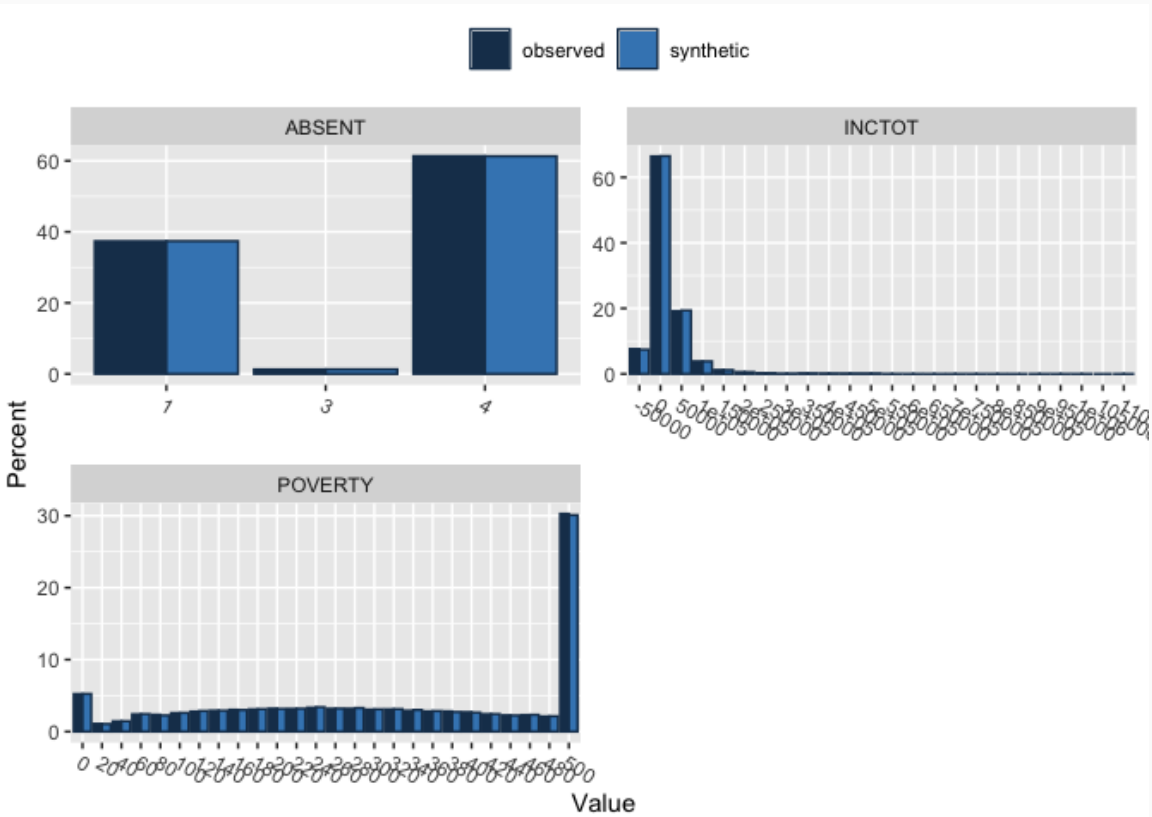
Graphical Comparison for Margins (R-Package: synthpop)

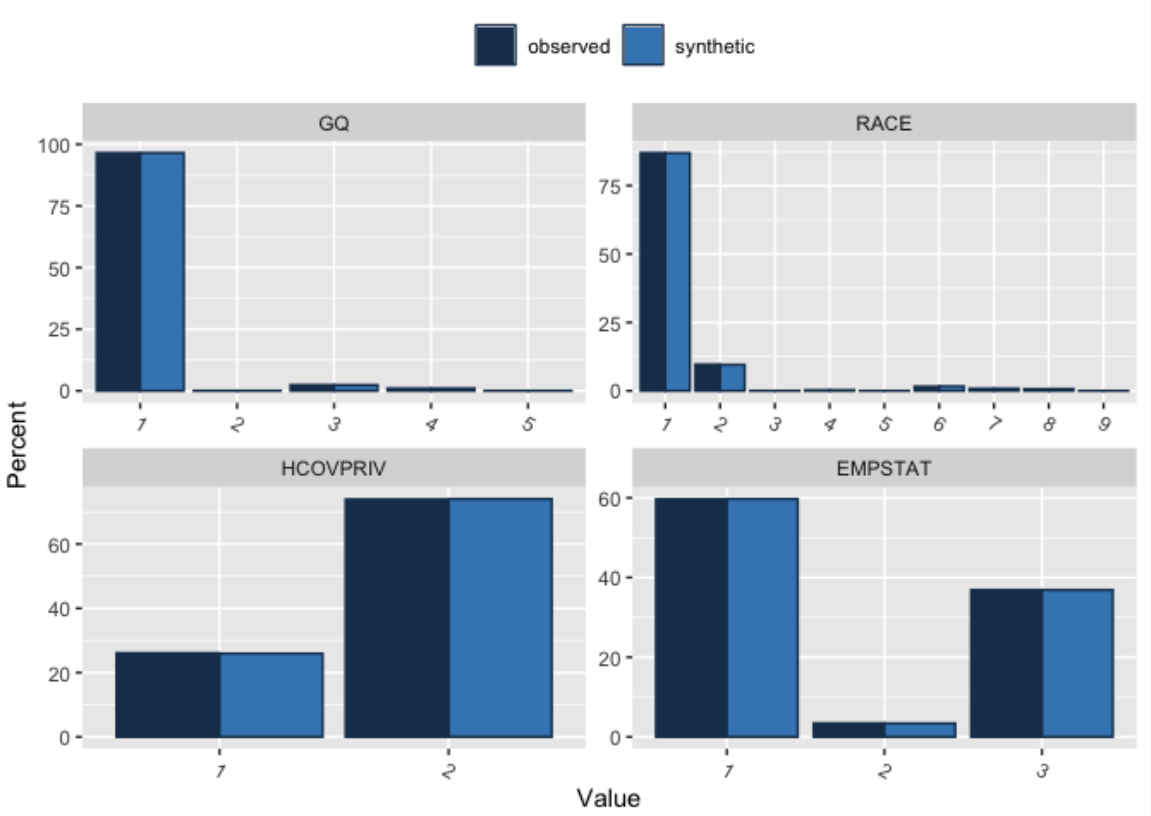
The following histograms provide an ad-hoc overview on the marginal distributions of the original and synthetic dataset. Matching or close distributions are related to a high data utility.

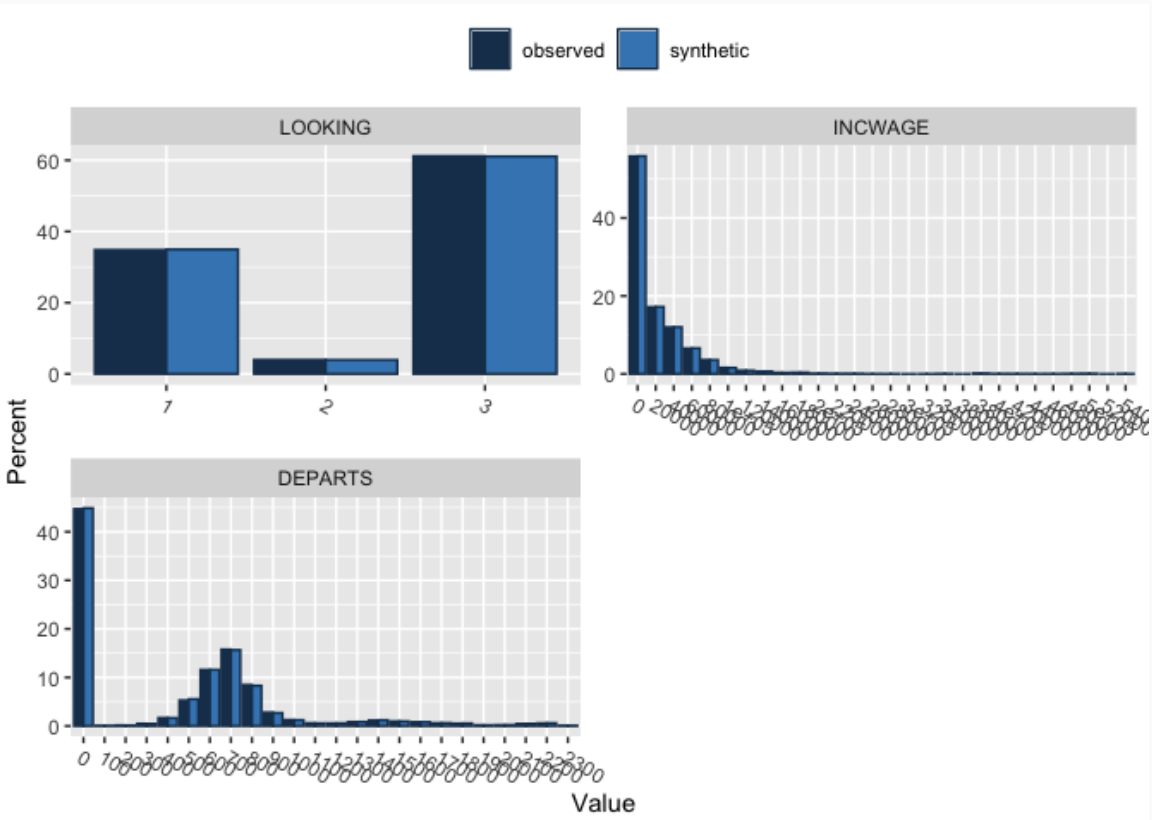

observed
synthetic
YEAR
AGE
SPEAKENG
HINSCARE
Percent
Value

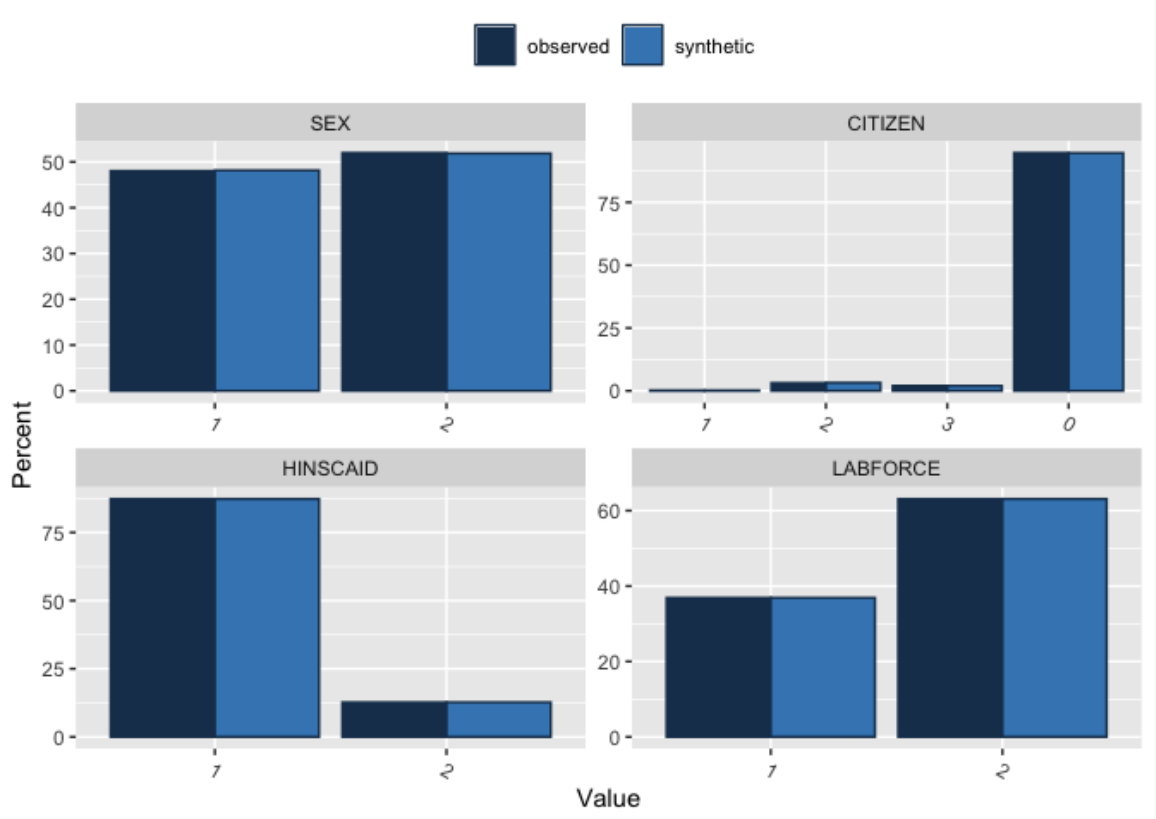

observed
synthetic
WRKLSTWK
WORKEDYR
INCEARN
Percent
Value

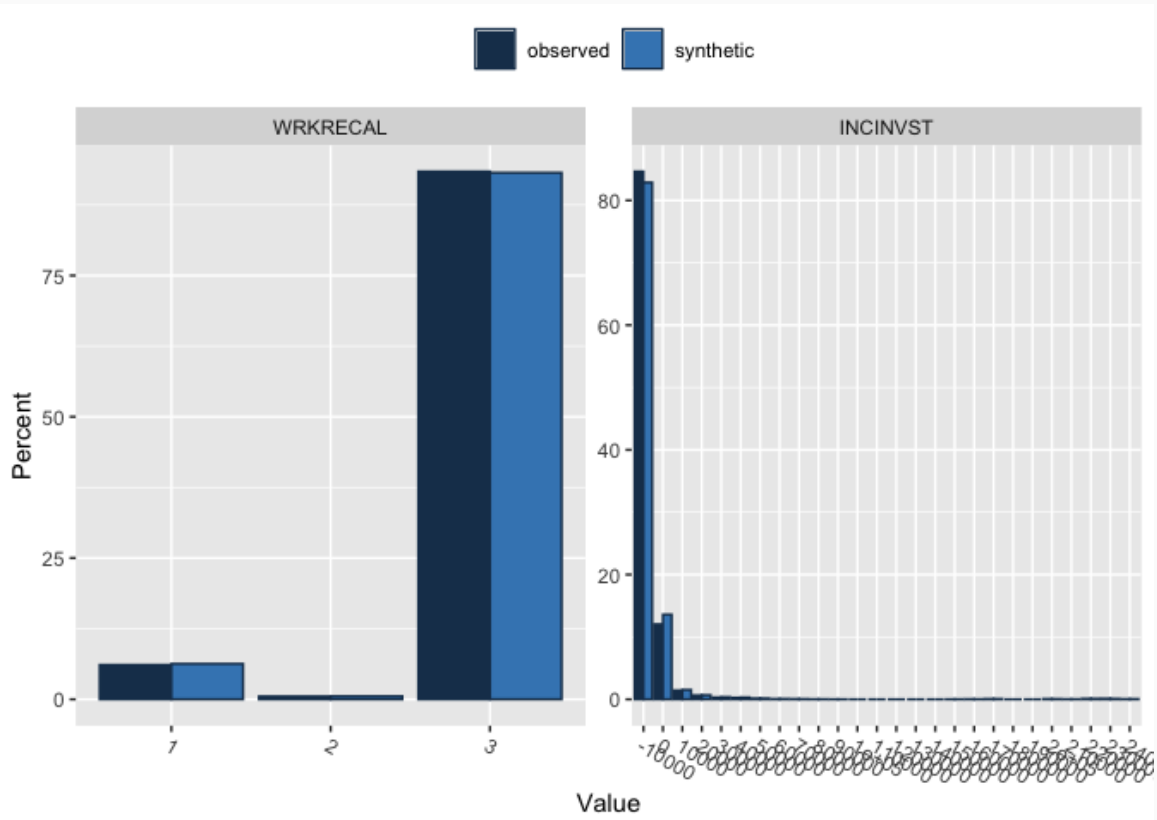

observed
synthetic
HHWT
MARST
HCOVANY
EDUC
Percent
Value


observed
synthetic
ABSENT
INCTOT
POVERTY
Percent
Value


observed
synthetic
GQ
RACE
HCOVPRIV
EMPSTAT
Percent
Value

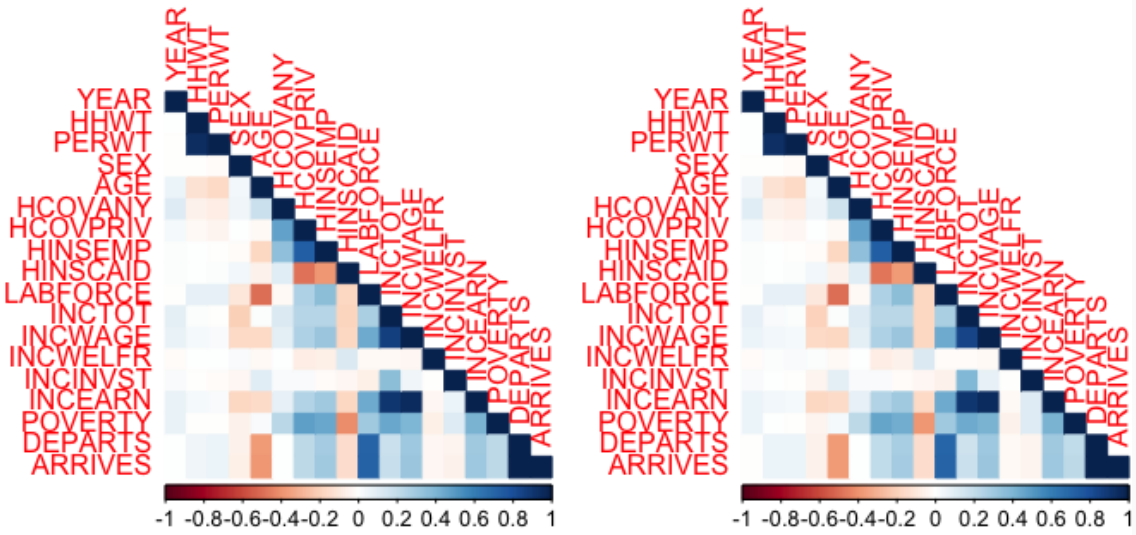

observed
synthetic
LOOKING
INCWAGE
DEPARTS
Percent
Value





Correlation Plots for Graphical Comparison of Pearson Correlation

Synthetic Datasets should represent the dependencies of the original datasets. The following correlation plots provide an ad-hoc overview on the Pearson correlations of the original and synthetic dataset. The left plot shows the original correlation whereas the right plot provides the correlation based on the synthetic dataset.



Distributional Comparison of Synthesised Data (R-Package: synthpop) by (S_)pMSE

Propensity scores are calculated on a combined dataset (original and synthetic). A model (here: CART) tries to identify the synthetic units in the dataset. Since both datasets should be identically structured, the pMSE should equal zero. The S_pMSE (standardised pMSE) should not exceed 10 and for a good fit below 3 according to Raab (2021, https://unece.org/sites/default/files/2021-12/SDC2021_Day2_Raab_AD.pdf)

	pMSE	S_pMSE	df
YEAR	4.2e-06	1.1165079	6
AGE	1.4e-06	0.5589858	4
SPEAKENG	1.5e-06	0.5824481	4
HINSCARE	2.0e-07	0.2559169	1
WRKLSTWK	3.8e-06	3.0699998	2
WORKEDYR	2.5e-06	2.0114386	2
INCEARN	1.6e-06	0.6573846	4

pMSE	S_pMSE
0.0018163	2.222492

	pMSE	S_pMSE	df
HHWT	4.80e-06	1.9233444	4

	pMSE	S_pMSE	df
MARST	3.60e-06	1.1611271	5
HCOVANY	0.00e+00	0.0468290	1
EDUC	1.17e-05	1.8754607	10
ABSENT	4.00e-07	0.3200769	2
INCTOT	3.70e-06	1.4913204	4
POVERTY	1.70e-06	0.9098795	3

pMSE	S_pMSE
0.0031569	2.554604

	pMSE	S_pMSE	df
GQ	1.10e-06	0.4215628	4
RACE	1.11e-05	2.2149785	8
HCOVPRIV	5.00e-07	0.8244981	1
EMPSTAT	5.00e-07	0.3745661	2
LOOKING	6.00e-07	0.4872763	2
INCWAGE	1.80e-06	0.9639562	3
DEPARTS	1.80e-06	1.4569282	2

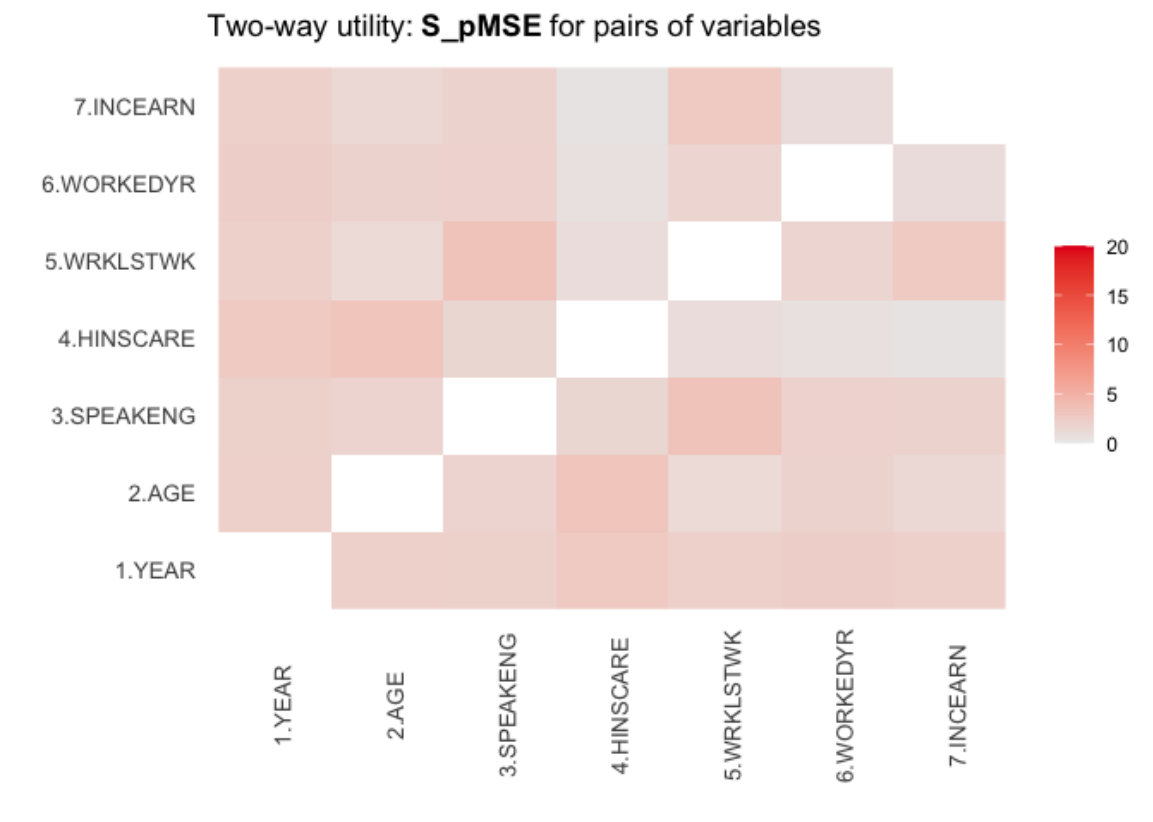
pMSE	S_pMSE
0.0008792	1.526391

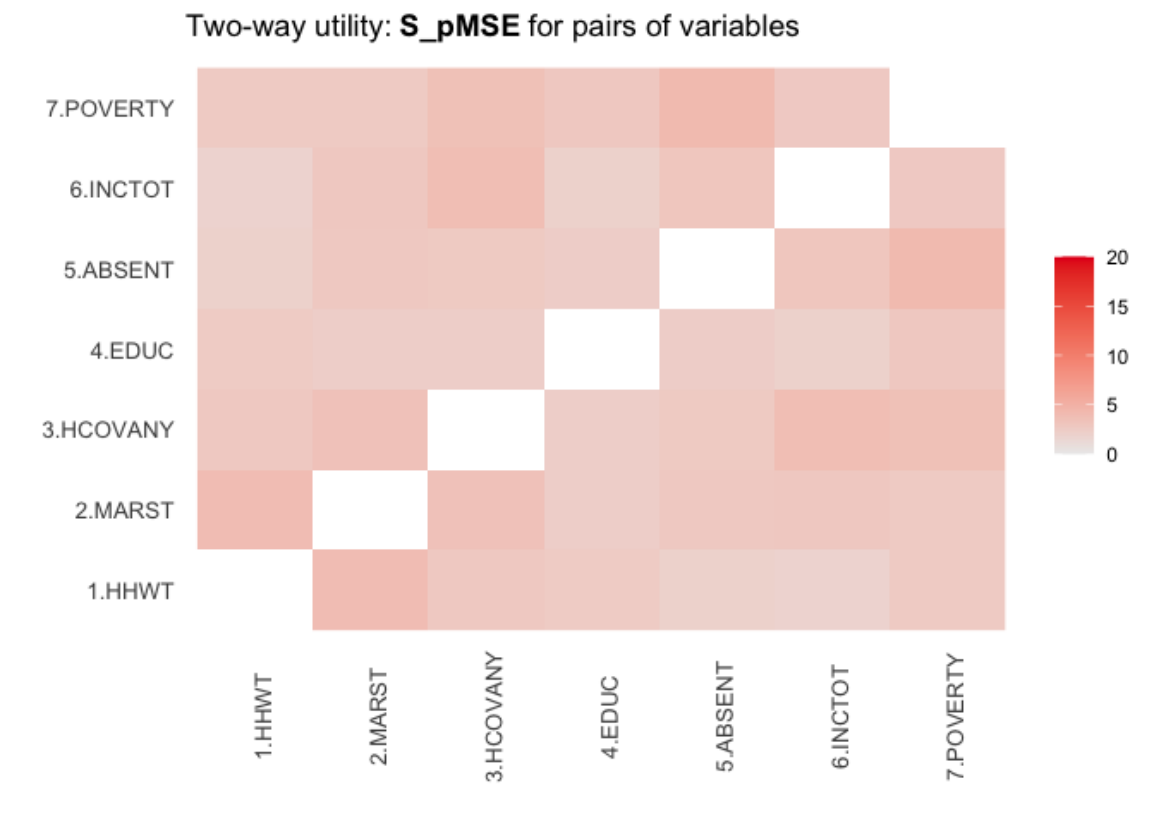
	pMSE	S_pMSE	df
SEX	0.0000004	0.5768435	1
CITIZEN	0.0000027	1.4147189	3
HINSCAID	0.0000000	0.0008118	1
LABFORCE	0.0000000	0.0003866	1
WRKRECAL	0.0000047	3.7820577	2
INCINVST	0.0001569	251.0007098	1

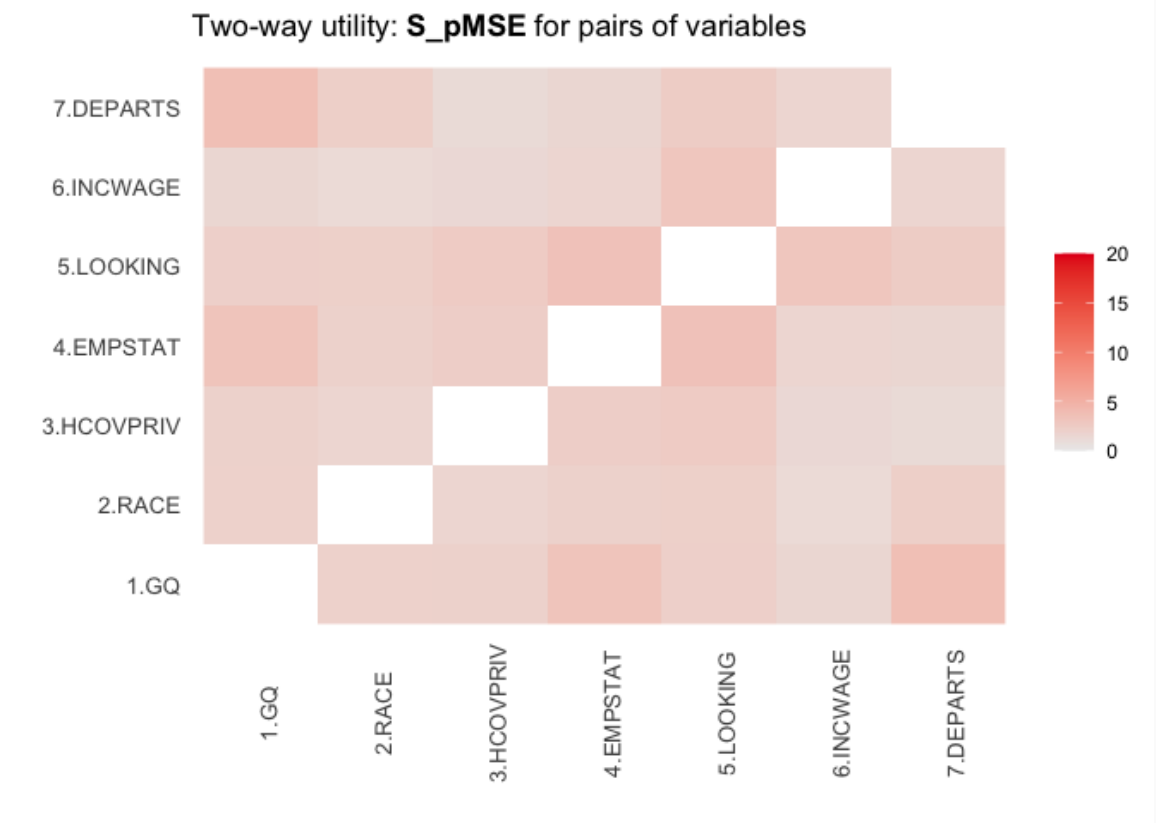
pMSE	S_pMSE
0.0004111	2.074587

Two-way Tables Comparison of Synthesised Data (R-Package: synthpop) by (S_)pMSE

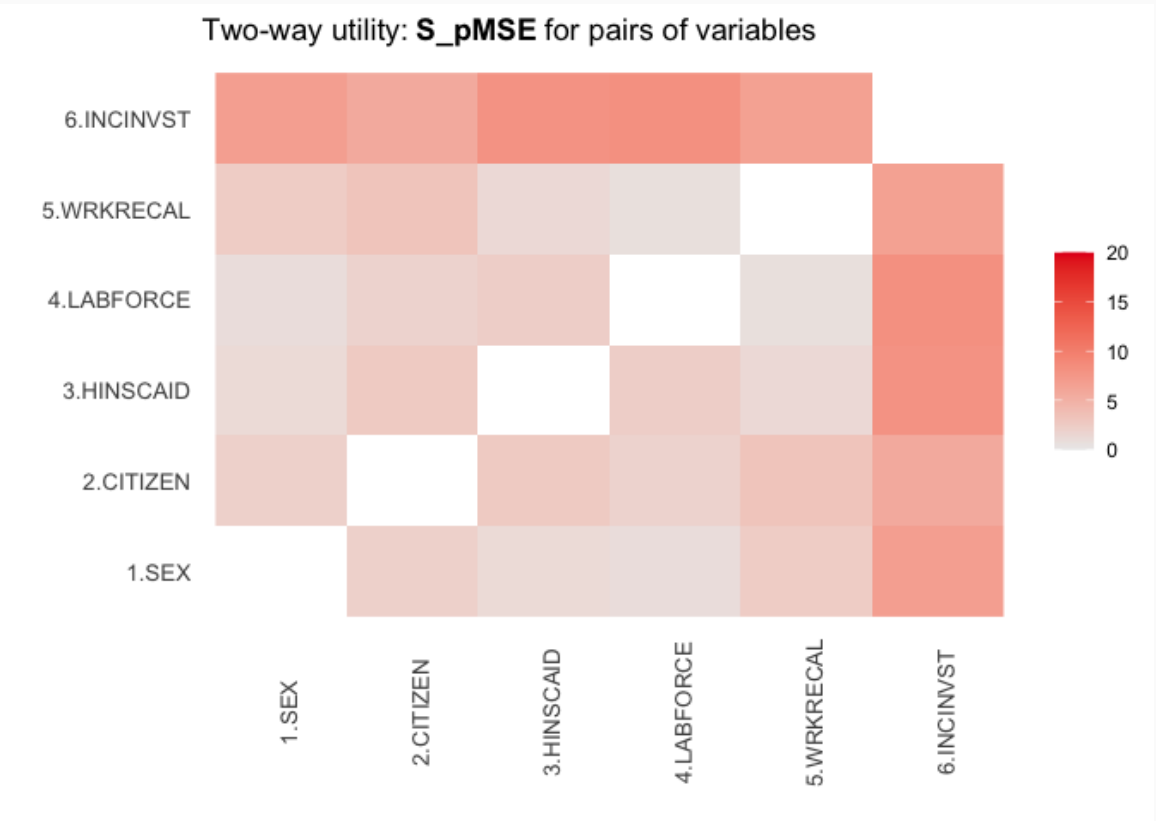
Two-way tables are evaluated based on the original and the synthetic dataset based on S_pMSE (see above). We also present the results for the mean absolute difference in densities (MabsDD) and the Bhattacharyya distance (dBhatt).

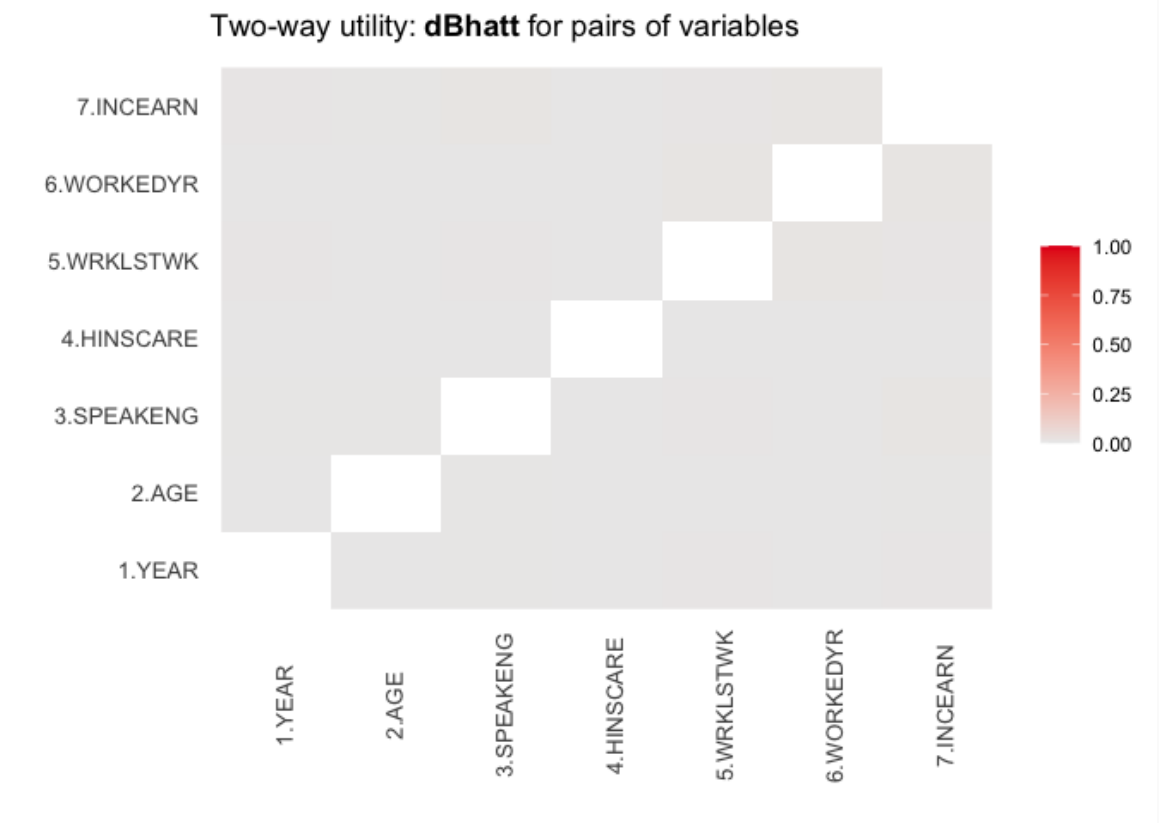


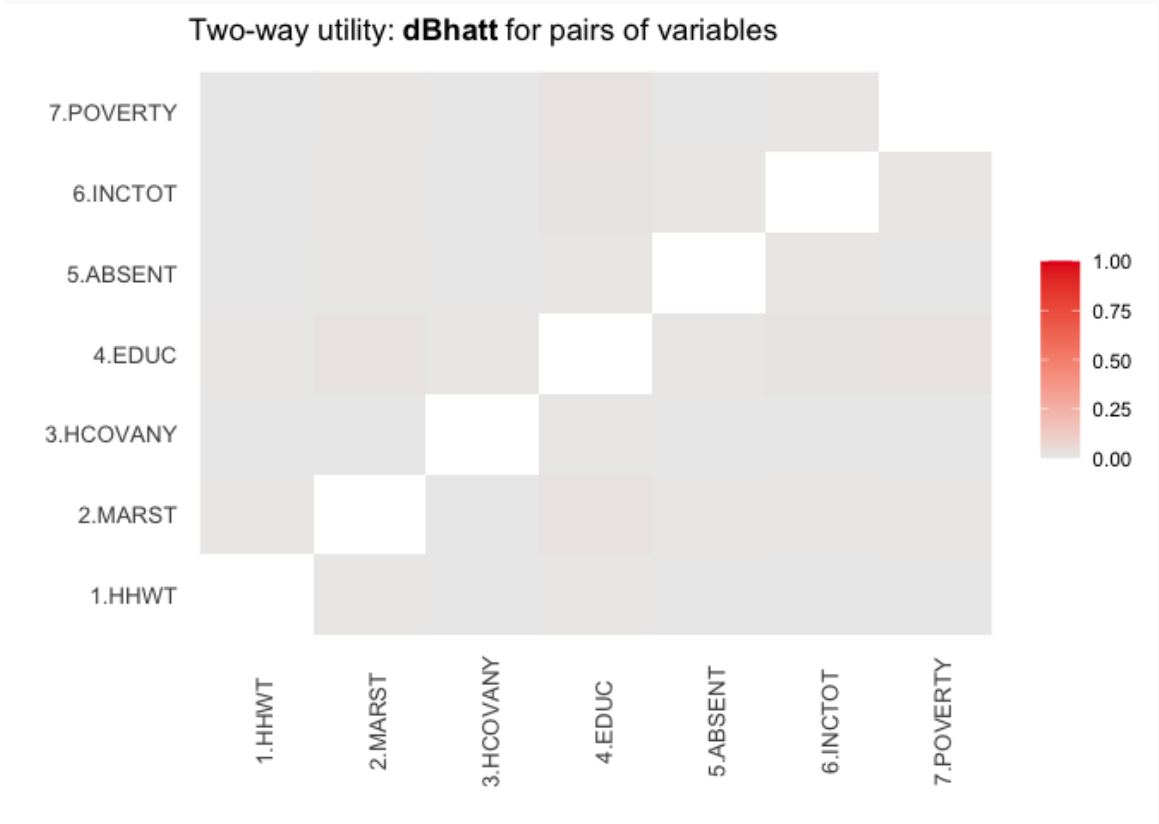


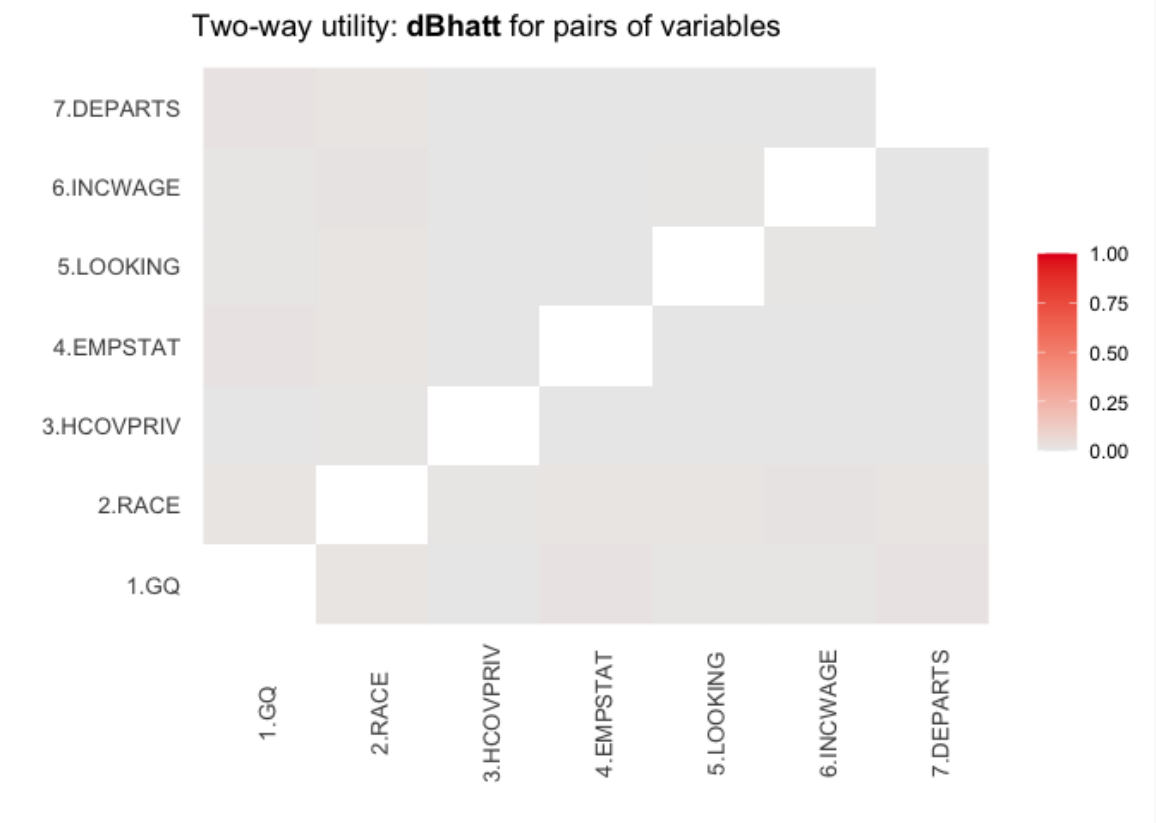

Two-way utility: S_pMSE for pairs of variables
7.DEPARTS
6.INCWAGE
5.LOOKING
4.EMPSTAT
3.HCOVPRIV
2.RACE
1.GQ
1.GQ
2.RACE
3.HCOVPRIV
4.EMPSTAT
5.LOOKING
6.INCWAGE
7.DEPARTS
20
15
10
5
0

```
## NULL
```

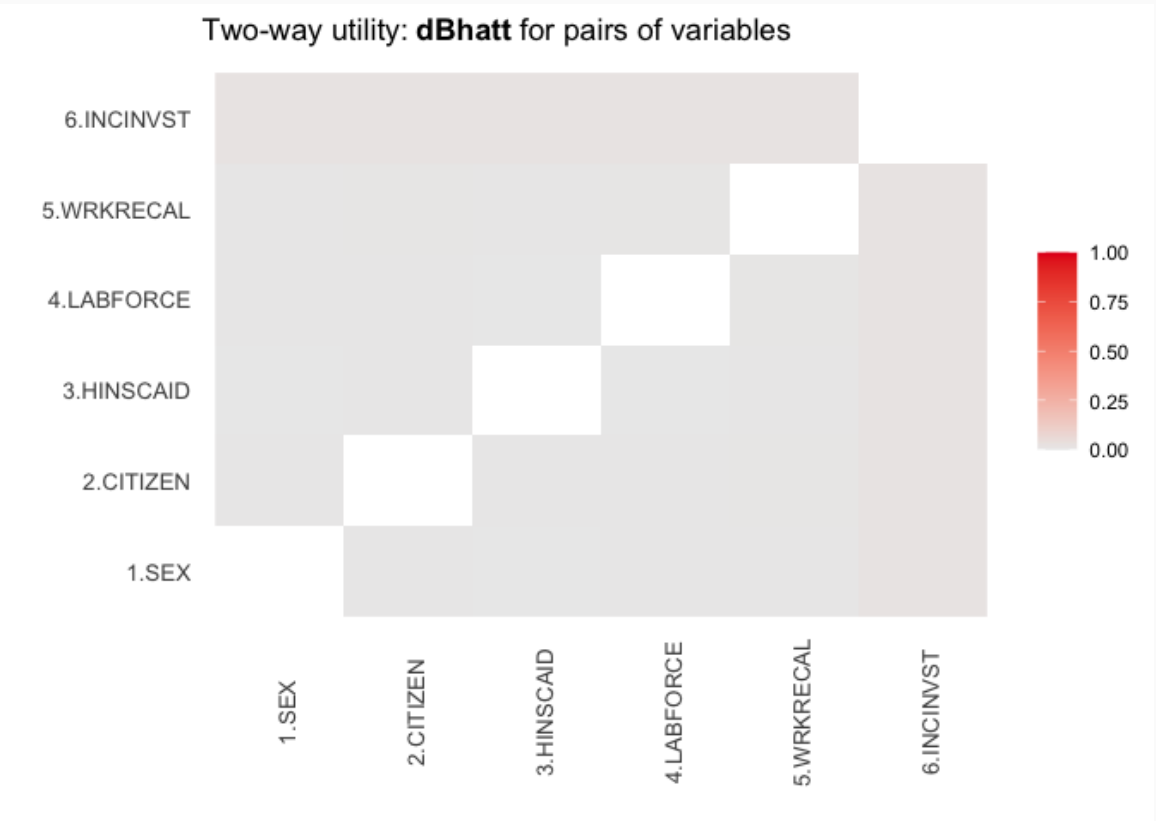

Two-way utility: S_pMSE for pairs of variables
6.INCINVST
5.WRKRECAL
4.LABFORCE
3.HINSCAID
2.CITIZEN
1.SEX
1.SEX
2.CITIZEN
3.HINSCAID
4.LABFORCE
5.WRKRECAL
6.INCINVST
20
15
10
5
0

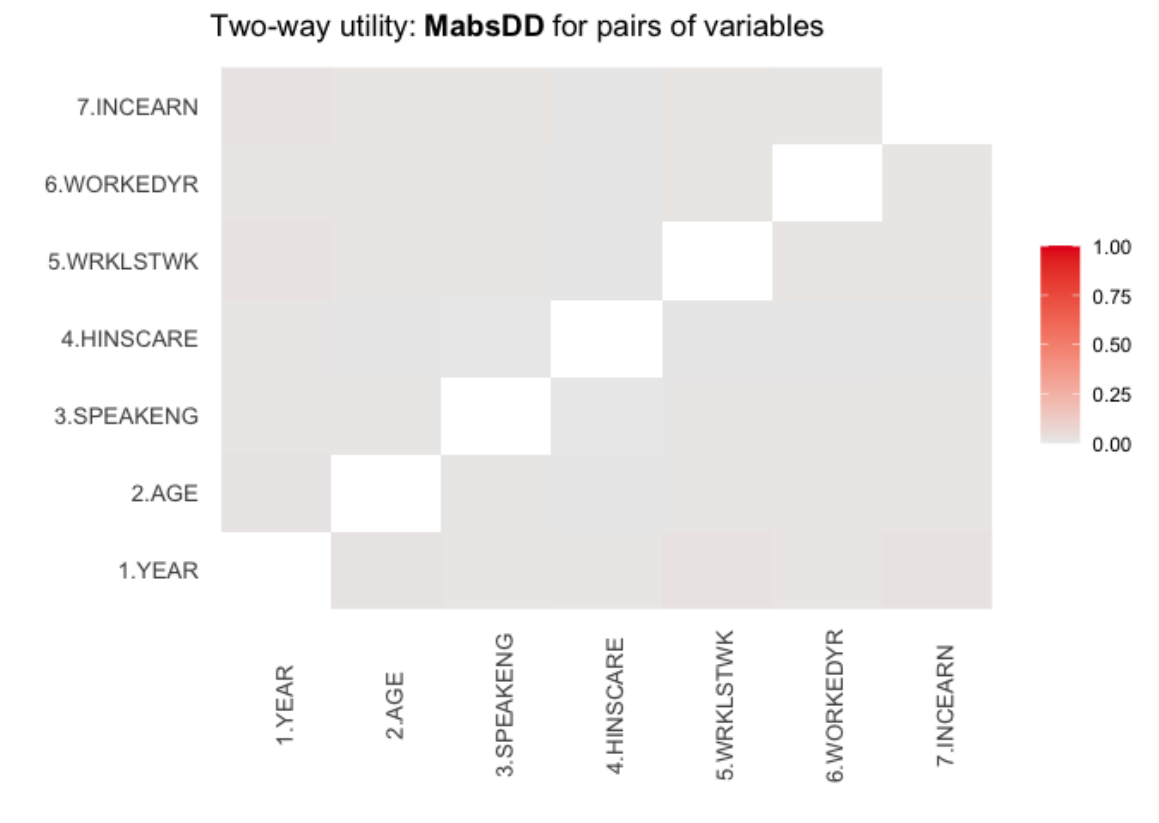

Two-way utility: dBhatt for pairs of variables
7.INCEARN
6.WORKEDYR
5.WRKLSTWK
4.HINSCARE
3.SPEAKENG
2.AGE
1.YEAR
1.00
0.75
0.50
0.25
0.00
1.YEAR
2.AGE
3.SPEAKENG
4.HINSCARE
5.WRKLSTWK
6.WORKEDYR
7.INCEARN

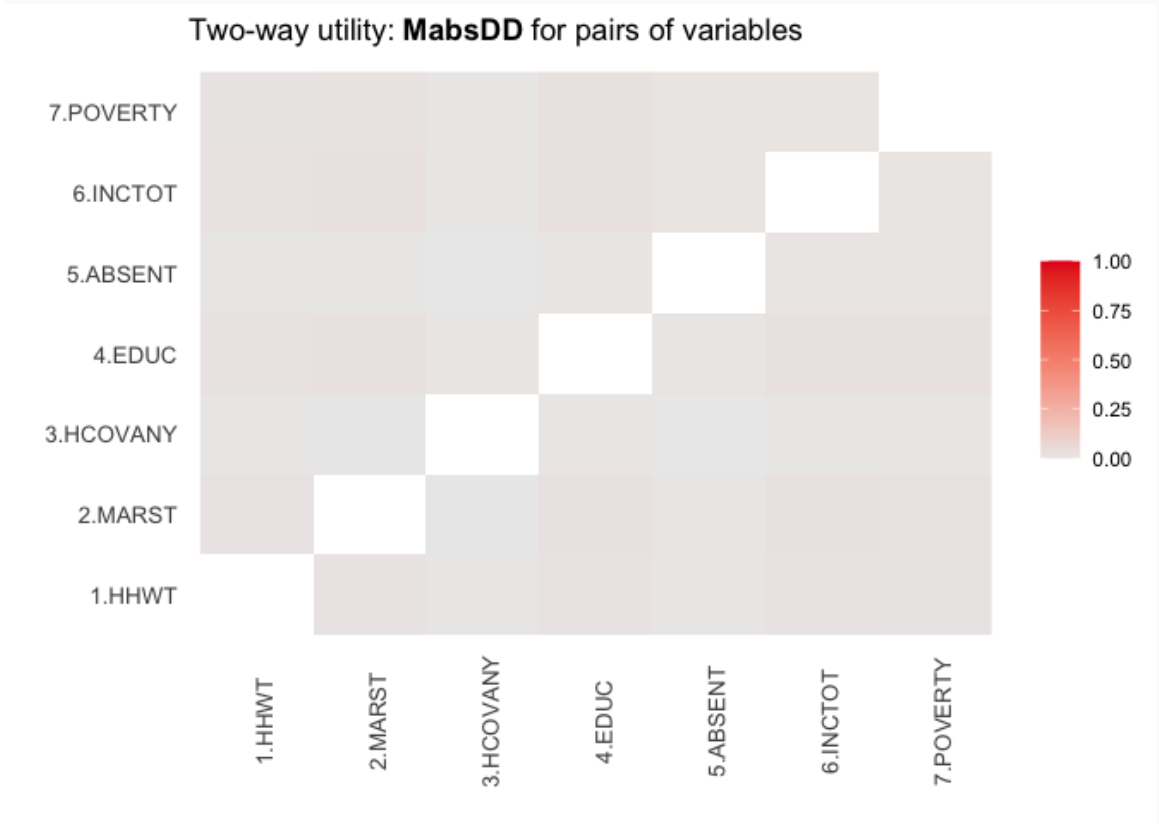

Two-way utility: dBhatt for pairs of variables
7.POVERTY
6.INCTOT
5.ABSENT
4.EDUC
3.HCOVANY
2.MARST
1.HHWT
1.00
0.75
0.50
0.25
0.00
1.HHWT
2.MARST
3.HCOVANY
4.EDUC
5.ABSENT
6.INCTOT
7.POVERTY

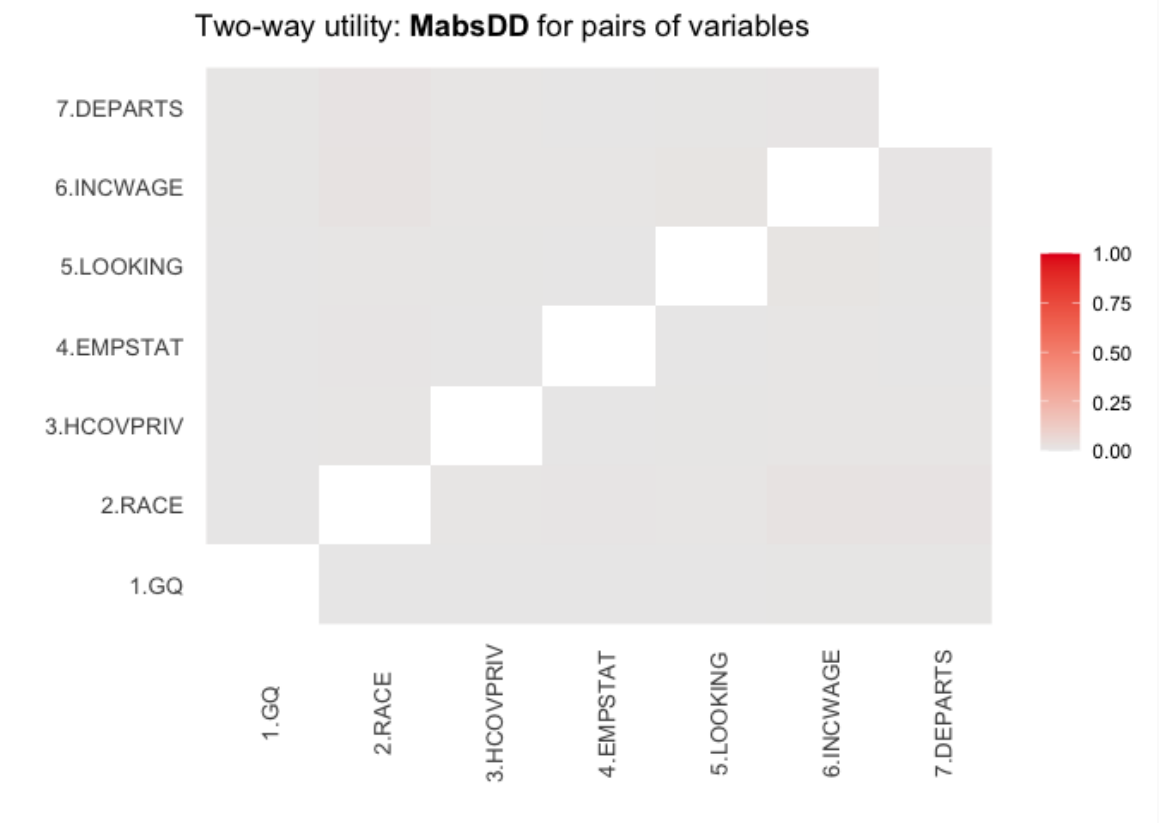

Two-way utility: dBhatt for pairs of variables
7.DEPARTS
6.INCWAGE
5.LOOKING
4.EMPSTAT
3.HCOVPRIV
2.RACE
1.GQ
1.00
0.75
0.50
0.25
0.00
1.GQ
2.RACE
3.HCOVPRIV
4.EMPSTAT
5.LOOKING
6.INCWAGE
7.DEPARTS

```
## NULL
```

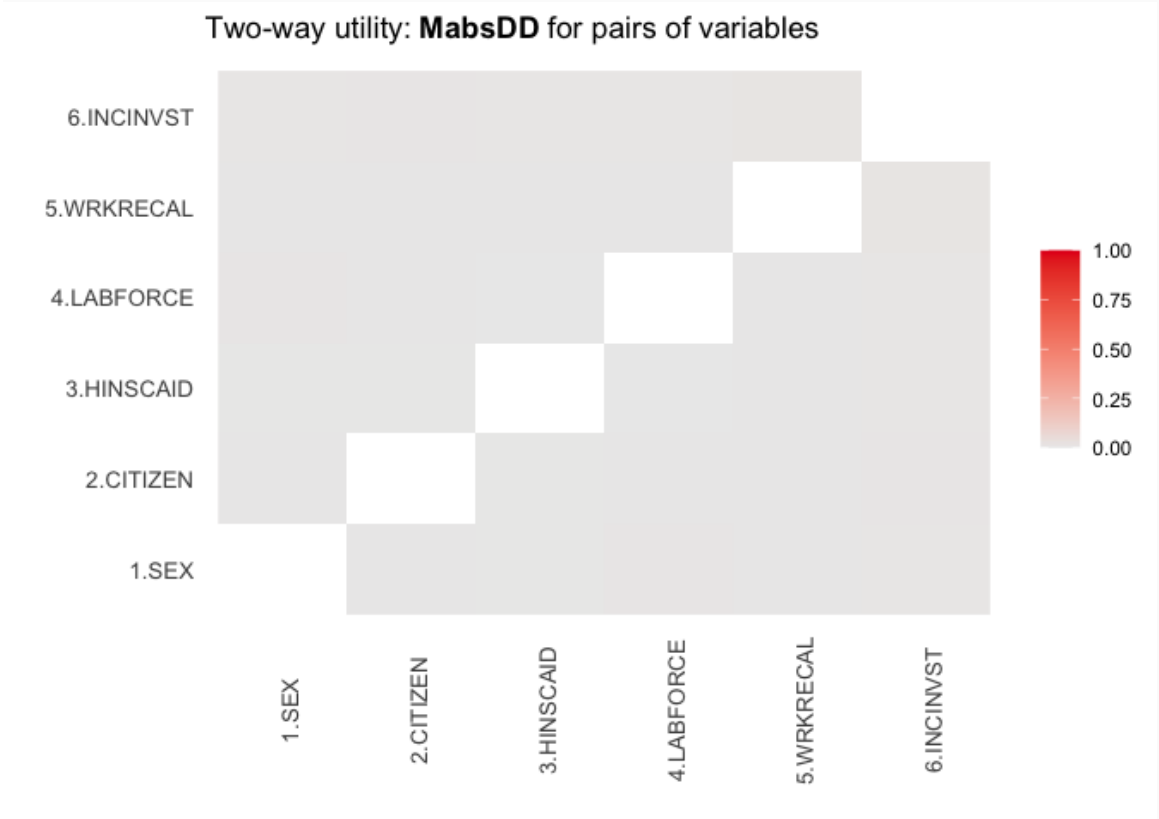

Two-way utility: dBhatt for pairs of variables
6.INCINVST
5.WRKRECAL
4.LABFORCE
3.HINSCAID
2.CITIZEN
1.SEX
1.00
0.75
0.50
0.25
0.00
1.SEX
2.CITIZEN
3.HINSCAID
4.LABFORCE
5.WRKRECAL
6.INCINVST


Two-way utility: MabsDD for pairs of variables
7.INCEARN
6.WORKEDYR
5.WRKLSTWK
4.HINSCARE
3.SPEAKENG
2.AGE
1.YEAR
1.00
0.75
0.50
0.25
0.00
1.YEAR
2.AGE
3.SPEAKENG
4.HINSCARE
5.WRKLSTWK
6.WORKEDYR
7.INCEARN


Two-way utility: MabsDD for pairs of variables
7.POVERTY
6.INCTOT
5.ABSENT
4.EDUC
3.HCOVANY
2.MARST
1.HHWT
1.00
0.75
0.50
0.25
0.00
1.HHWT
2.MARST
3.HCOVANY
4.EDUC
5.ABSENT
6.INCTOT
7.POVERTY




```
## NULL
```



Information Loss Measure Proposed by Andrzej Mlodak (R-Package: sdcMicro)

The value of this information loss criterion is between 0 (no information loss) and 1. It is calculated overall and for each variable.

Information.Loss
0.5033936

Individual Distances for Information Loss:

```
##       YEAR       HHWT         GQ      PERWT        SEX        AGE      MARST
## 0.85697000 0.95752370 0.06751000 0.95912056 0.50262000 0.91090934 0.60210000
##       RACE     HISPAN    CITIZEN   SPEAKENG    HCOVANY   HCOVPRIV    HINSEMP
## 0.23510000 0.05828000 0.10303000 0.13019000 0.13808000 0.38412000 0.47622000
##   HINSCAID   HINSCARE       EDUC    EMPSTAT   EMPSTATD   LABFORCE   WRKLSTWK
## 0.22071000 0.39305000 0.75087000 0.50792000 0.52176000 0.46653000 0.55036000
##     ABSENT    LOOKING   AVAILBLE   WRKRECAL   WORKEDYR     INCTOT     INCWAGE
## 0.48528000 0.50203000 0.17949000 0.12559000 0.50191000 0.99030473 0.85213910
##   INCWELFR   INCINVST    INCEARN    POVERTY    DEPARTS    ARRIVES
## 0.03123161 0.30548819 0.87593524 0.89627195 0.78453832 0.79219817
```