ACS - Simulation Models

Evaluation Synthetic Data Creation

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

January 28, 2022

Executive Summary

We fit two **multivariate normal distributions** for each gender and create synthetic data by drawing thereof. Out of all different methods we tested (`FCS`, `IPSO`, `GAN`, `Simulation`, `Minutemen`) this method actually scored best in our privacy measures. Of course, utility is not as good as with other methods. This is a relatively **easy and fast approach** which should make it interesting for **testing technology** and **education**. According to our utility measures, simulated data using a multivariate normal distribution for (semi-)continuous variables and expanding it using FCS (CART) for the categorical variables is not a useful strategy to generate suitable synthetic data from the ACS dataset in general. The first impression of barely aligning marginal distributions is underpinned by further metrics. Only the Pearson correlation coefficients for binary and (semi-)continuous variables are close to those of the original dataset ("lower right corner"). The S_pMSE for tables and for distributions shows extreme values. Also the absolute difference in densities and the Bhattacharyya distance support the overall impression. Mlodak's information loss criterion indicates this synthetic dataset as not useful apart from testing technology. There is a distinct limited usability according to Mlodak's information loss criterion.

USE CASE RECOMMENDATIONS

Releasing_to_Public	Testing_Analysis	Education	Testing_Technology
NO	NO	YES	YES

Since it is a rather simple and fast approach with very good privacy measures, **testing technology** and **education** is the prime use case for these simulations. **Releasing to the public** and **testing analysis** wouldn't be a good fit, since in our opinion the dataset doesn't have the required utility.

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 fit a multivariate normal distribution on the (semi-)continuous variables, e.g. income related variables, and expand the dataset using FCS (Cart) by further categorical variables. The fit of the multivariate normal distribution is crucial for the overall quality. One can assume a poor overall usability if the (few) starting variables do not mimic the original variables adequately. Some variables were censored at zero in cases where the respective draw delivered negative values if the original variable did not contain negative values. The variable for the total income was calculated by the sum of the other income components to assess consistency.

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
0	0	0

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:

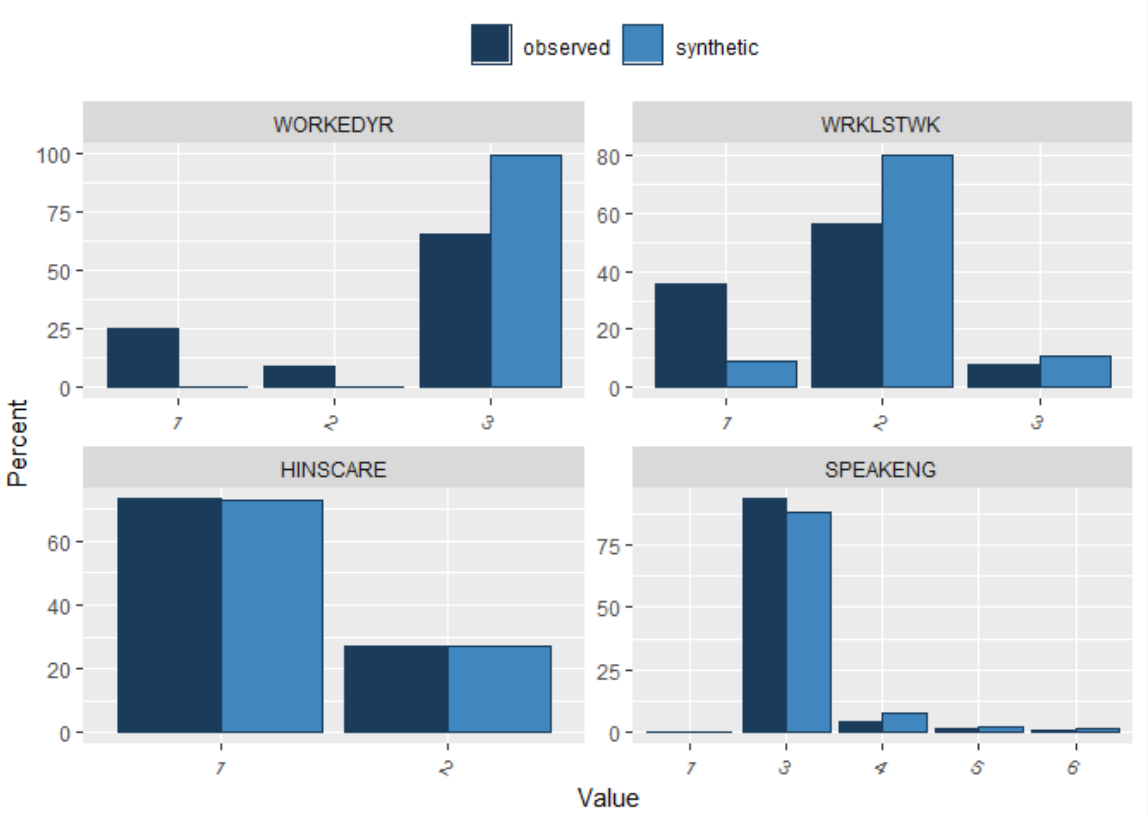
Metric	Number.Uniques	Number.Replications	Percentage.Replications
Perceived Risk	1033709	0	0

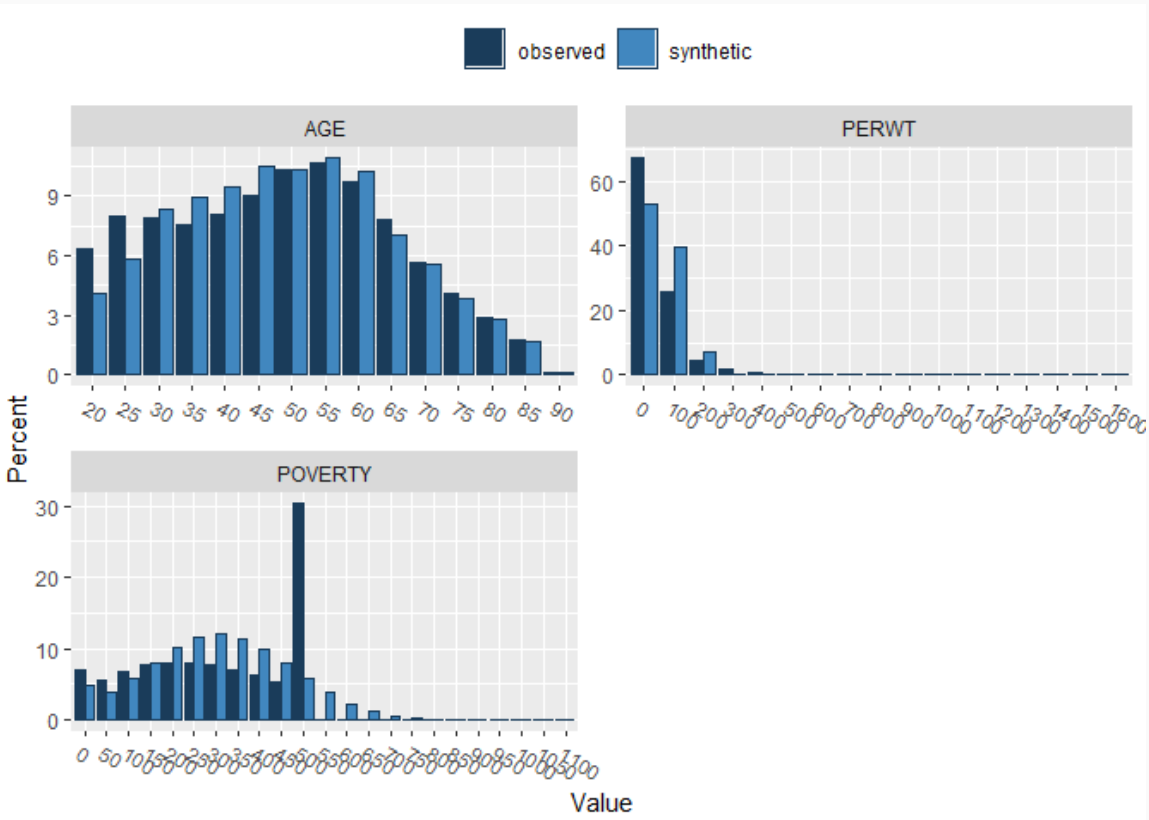
Utility Evaluation

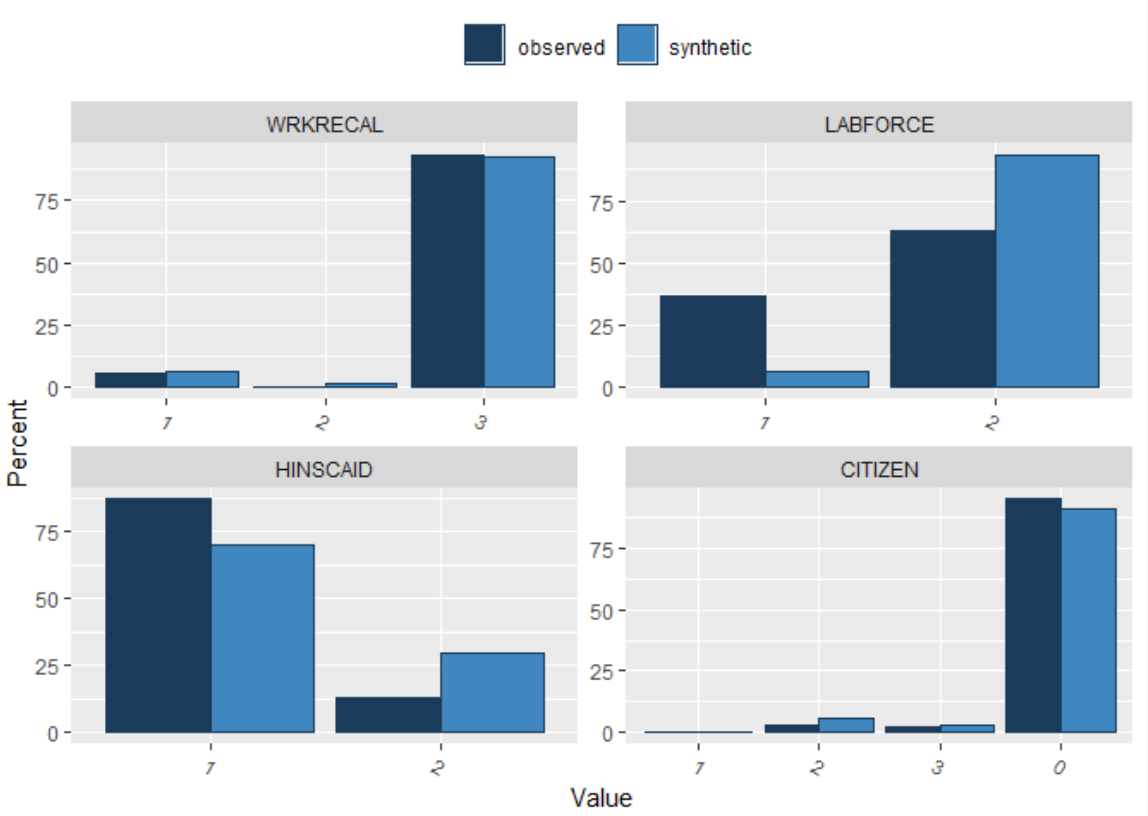
Different utility measures are applied in this section. These utility measures are the basis of utility evaluation for the generated synthetic dataset. The R packages synthpop, sdcMicro and corrplot were used to compute the following metrics. We do not use tests incorporating significance here. Confidence 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.

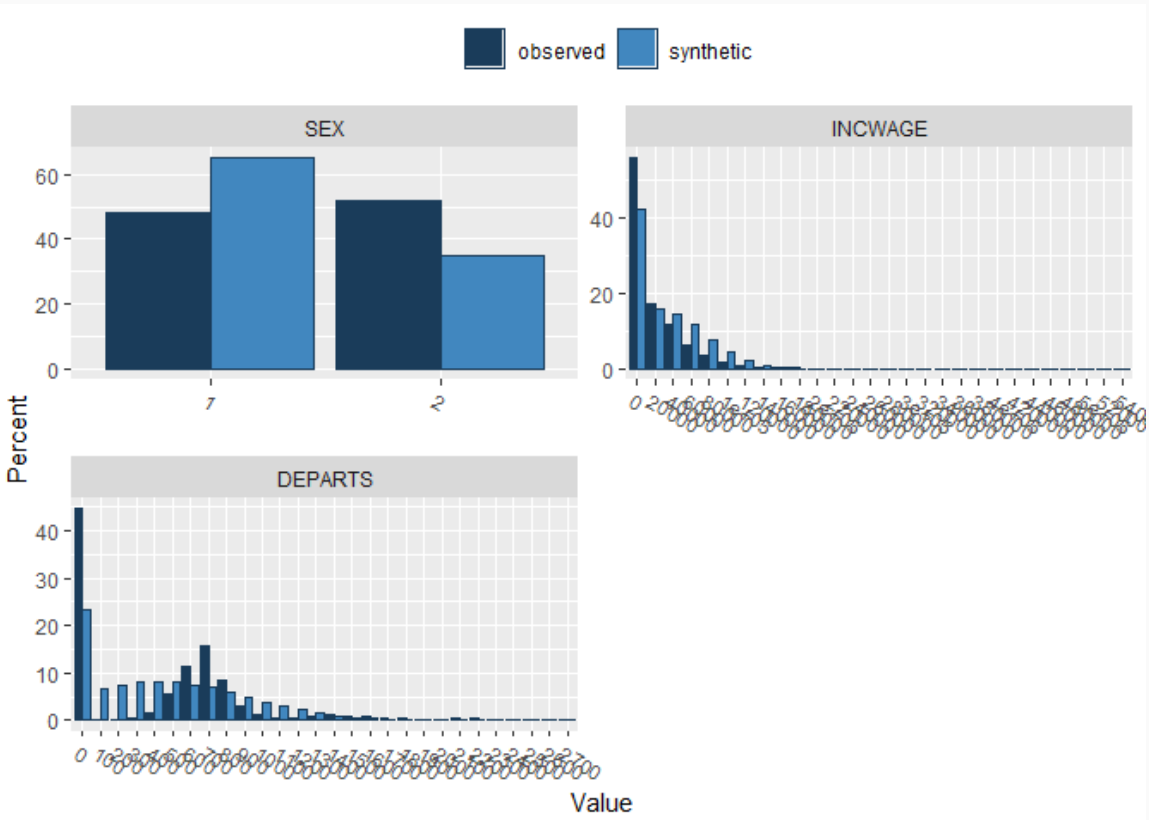
Graphical Comparison for Margins (R-Package: synthpop)

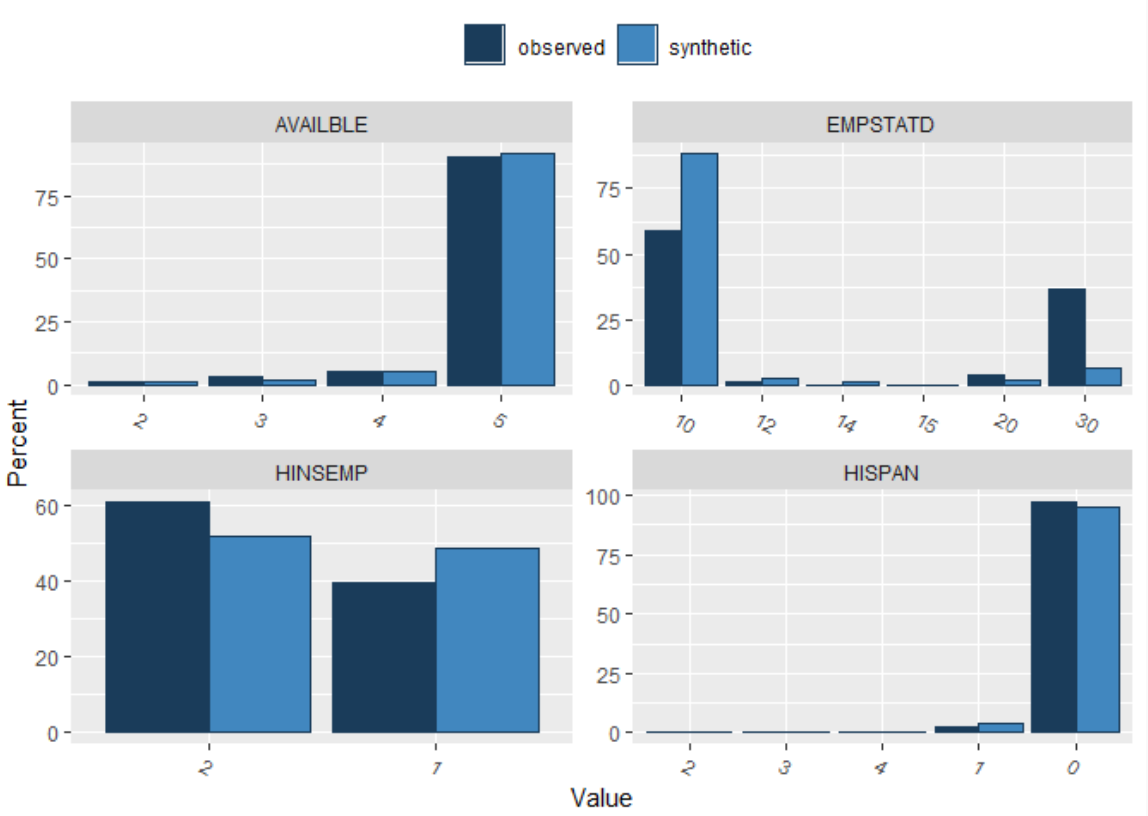
The following histograms provide an ad-hoc overview on the marginal distributions of the original and synthetic dataset. Matching or close distributions are related to a high data utility.

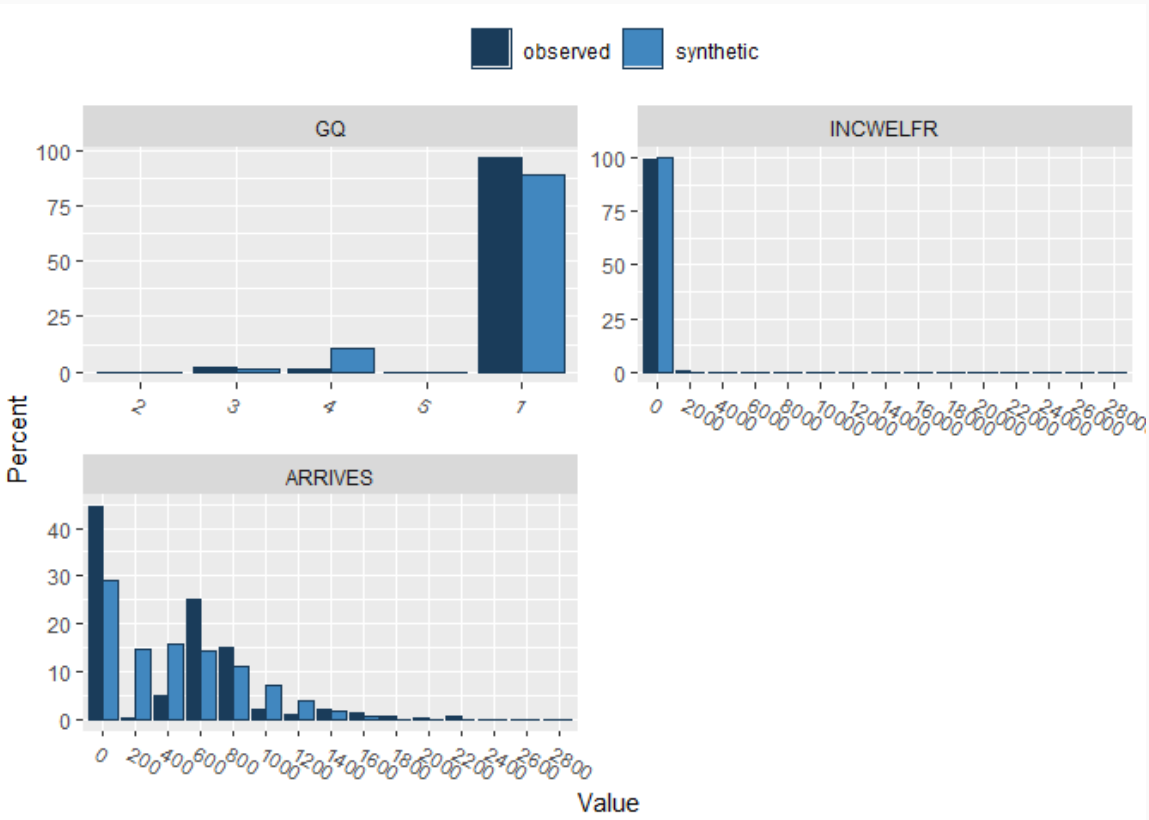

observed
synthetic
WORKEDYR
WRKLSTWK
HINSCARE
SPEAKENG
Percent
Value

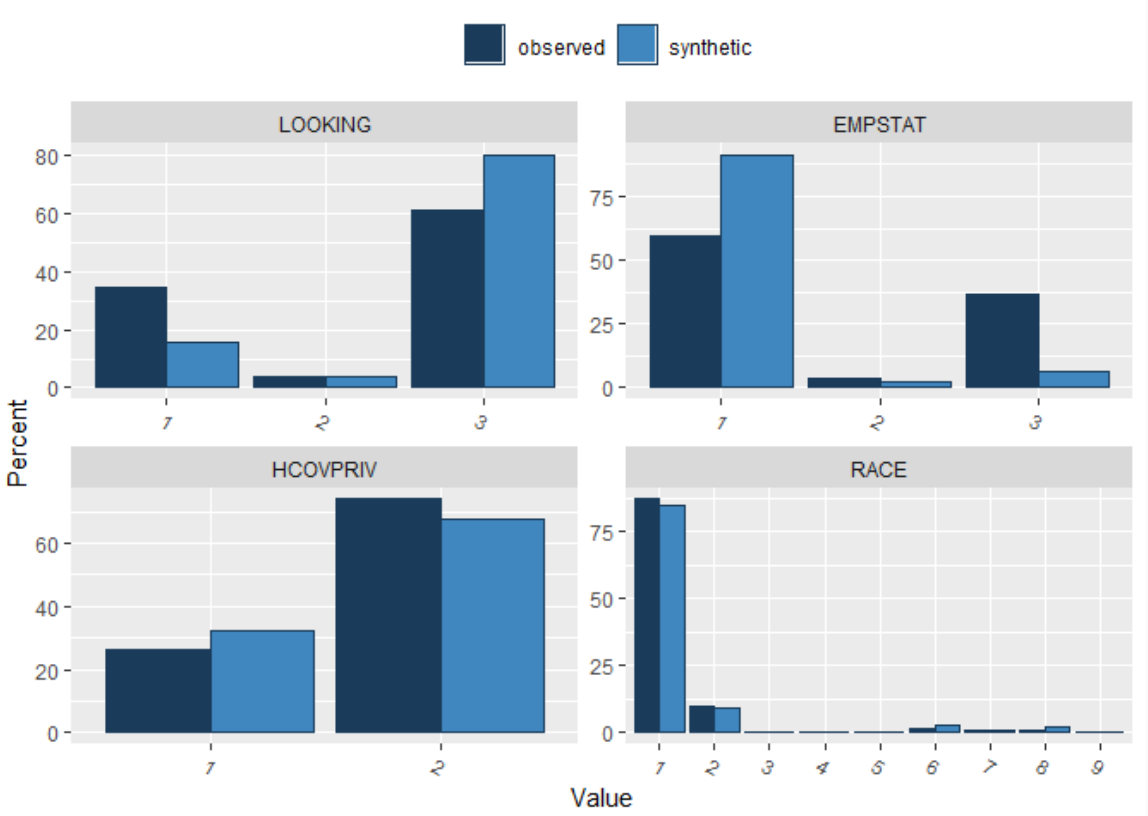

observed
synthetic
AGE
PERWT
POVERTY
Percent
Value

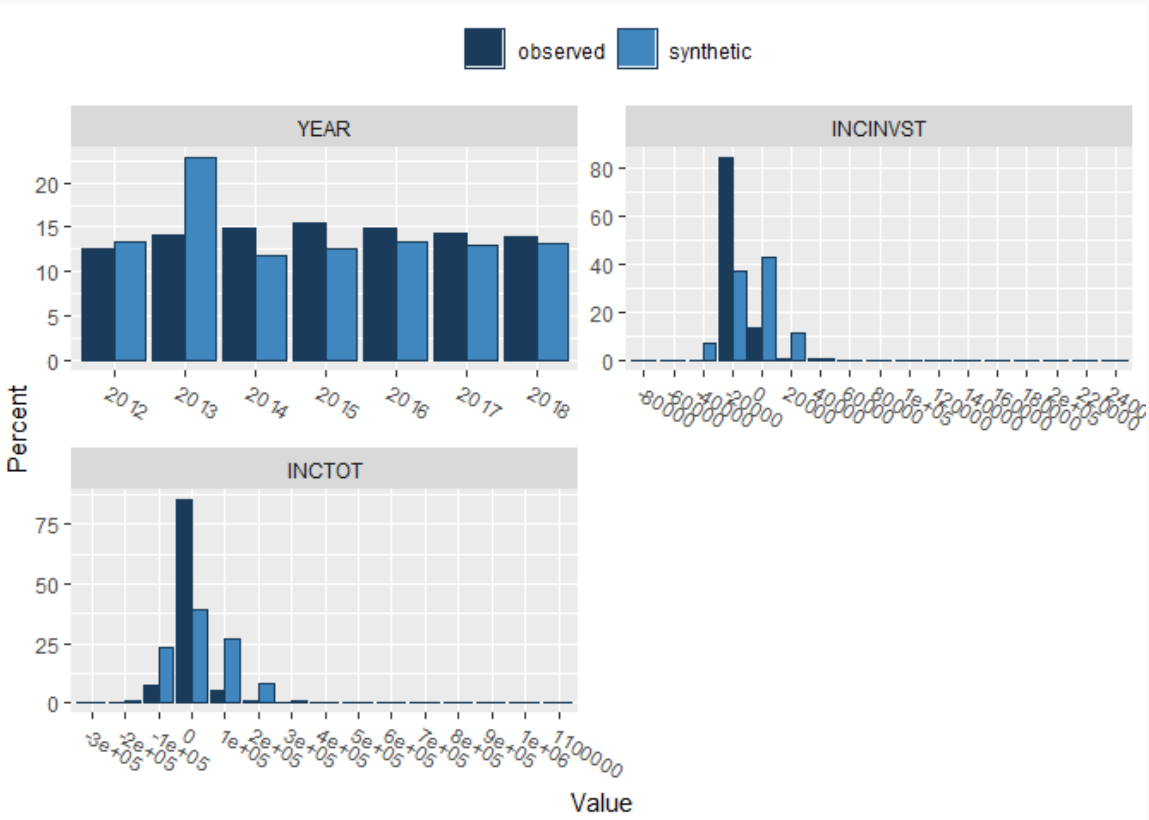

observed
synthetic
WRKRECAL
LABFORCE
HINSCAID
CITIZEN
Percent
Value

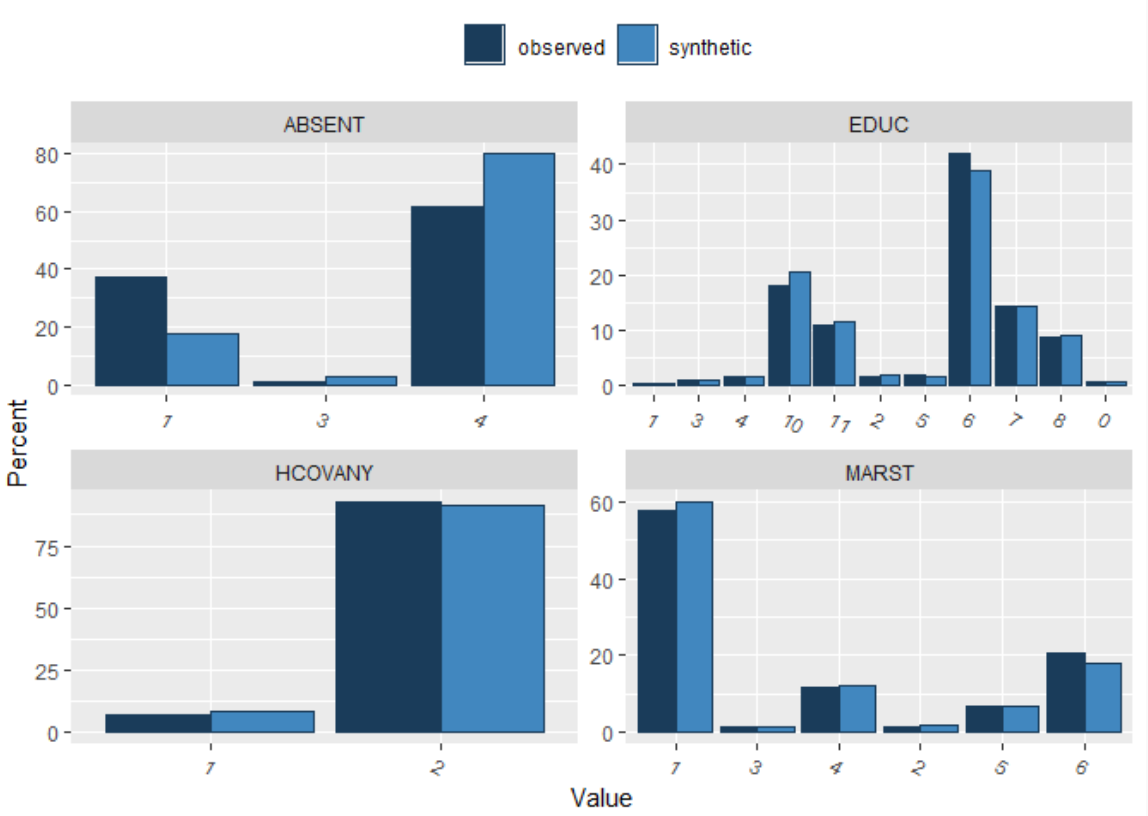

observed
synthetic
SEX
INCWAGE
DEPARTS
Percent
Value

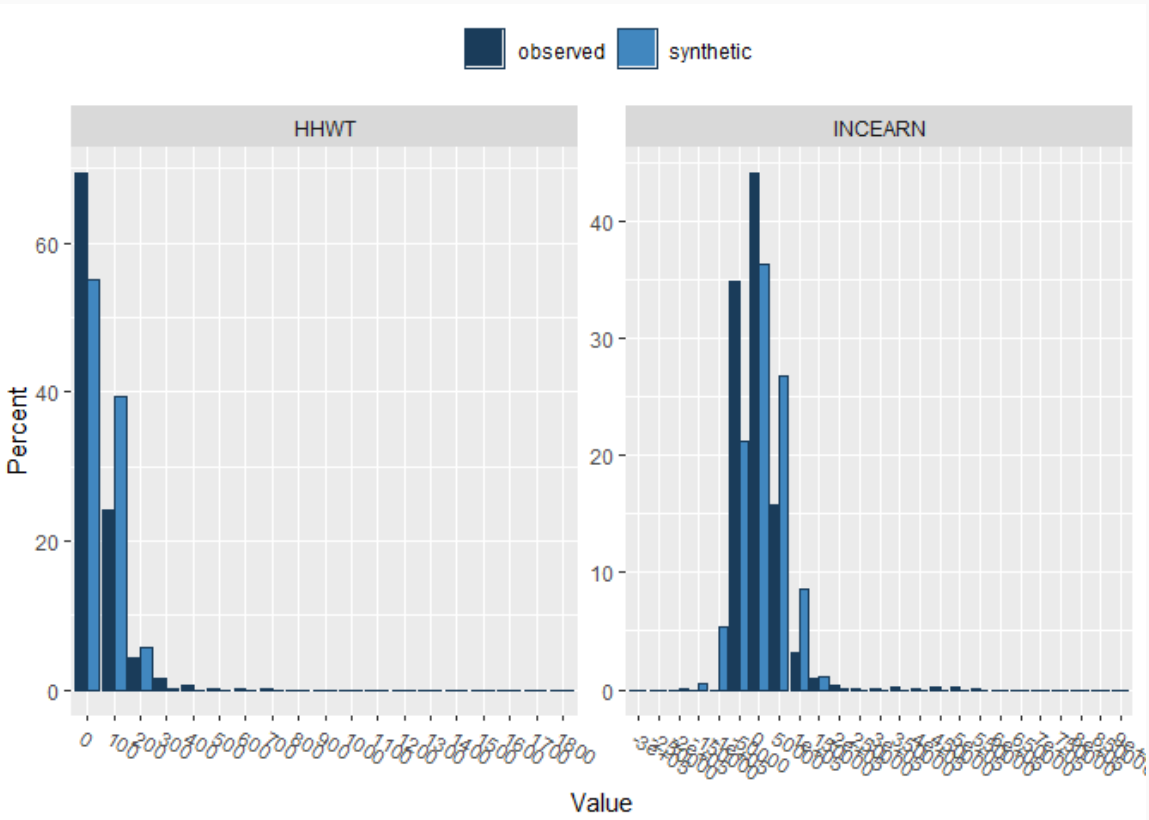

observed
synthetic
AVAILBLE
EMPSTATD
HINSEMP
HISPAN
Percent
Value


observed
synthetic
GQ
INCWELFR
ARRIVES
Percent
Value


observed
synthetic
LOOKING
EMPSTAT
HCOVPRIV
RACE
Percent
Value

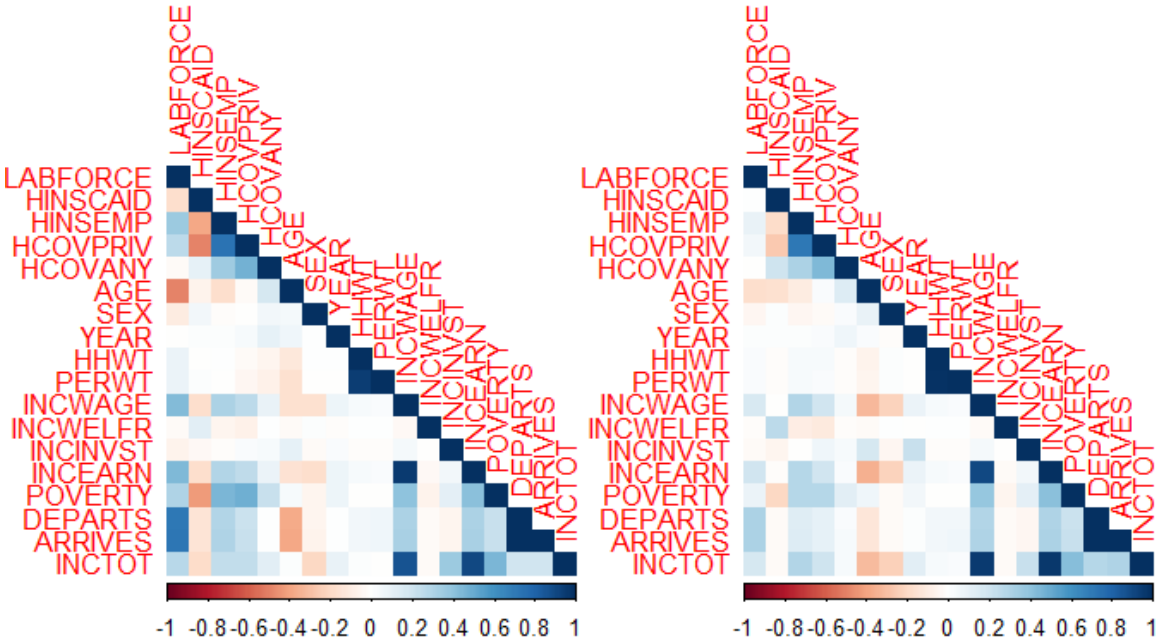

observed
synthetic
YEAR
INCINVST
INCTOT
Percent
Value





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
WORKEDYR	0.0498032	412450.63301	2
WRKLSTWK	0.0253896	210266.66023	2
HINSCARE	0.0000014	23.79855	1
SPEAKENG	0.0021720	8993.97073	4
AGE	0.0010205	4225.89233	4
PERWT	0.0108233	44816.96794	4
POVERTY	0.0130157	53895.64953	4

pMSE	S_pMSE
0.1310628	102.9972

	pMSE	S_pMSE	df
WRKRECAL	0.0004926	4079.799	2

	pMSE	S_pMSE	df
LABFORCE	0.0339964	563089.062	1
HINSCAID	0.0106940	177126.791	1
CITIZEN	0.0017075	9427.000	3
SEX	0.0074704	123734.293	1
INCWAGE	0.0064828	35792.172	3
DEPARTS	0.0139431	76980.894	3

pMSE	S_pMSE
0.1597504	228.9538

	pMSE	S_pMSE	df
AVAILBLE	0.0003161	1745.231	3
EMPSTATD	0.0359792	119186.102	5
HINSEMP	0.0021060	34881.594	1
HISPAN	0.0006612	2737.927	4
GQ	0.0109480	45333.358	4
INCWELFR	0.0544996	902688.824	1
ARRIVES	0.0175164	96709.067	3

pMSE	S_pMSE
0.1655773	317.258

	pMSE	S_pMSE	df
LOOKING	0.0119657	99095.067	2
EMPSTAT	0.0353746	292958.861	2
HCOVPRIV	0.0012108	20054.501	1
RACE	0.0013987	2895.781	8
YEAR	0.0036775	10152.001	6
INCINVST	0.1482503	818500.409	3
INCTOT	0.0586351	242796.252	4

pMSE	S_pMSE
0.2176001	268.8362

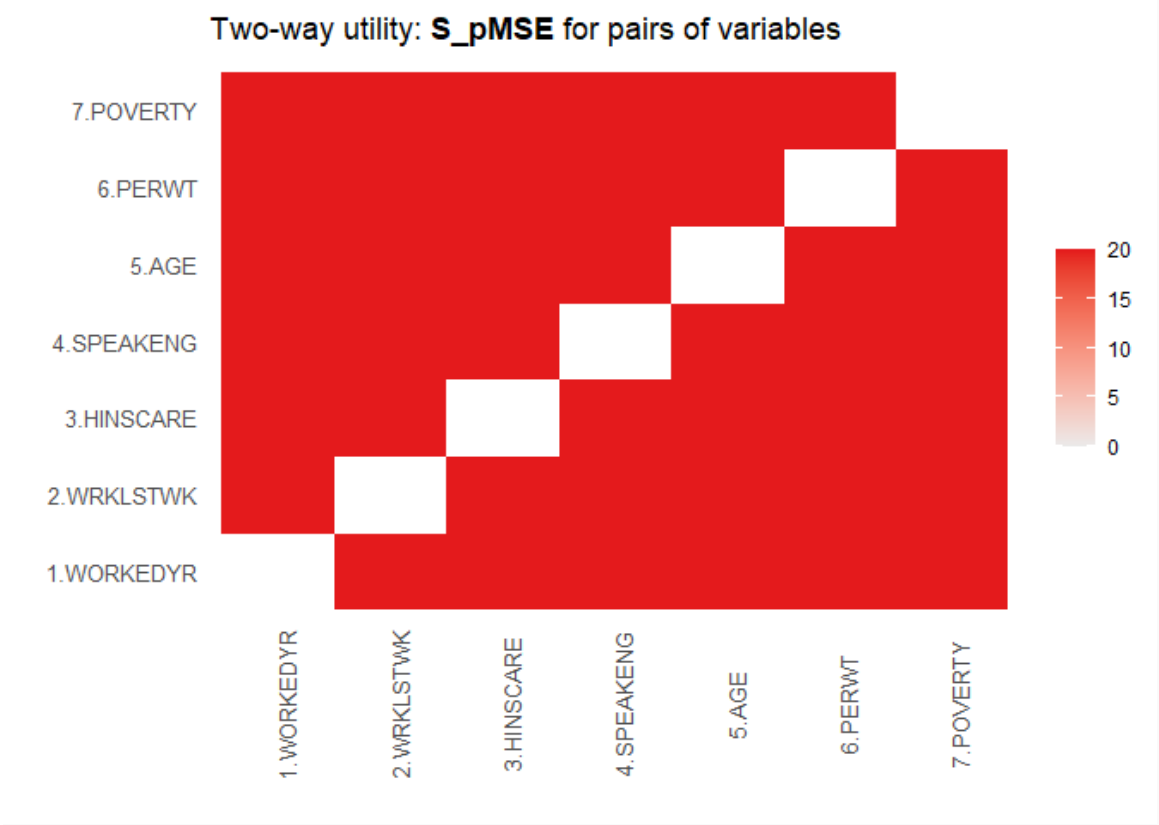
	pMSE	S_pMSE	df

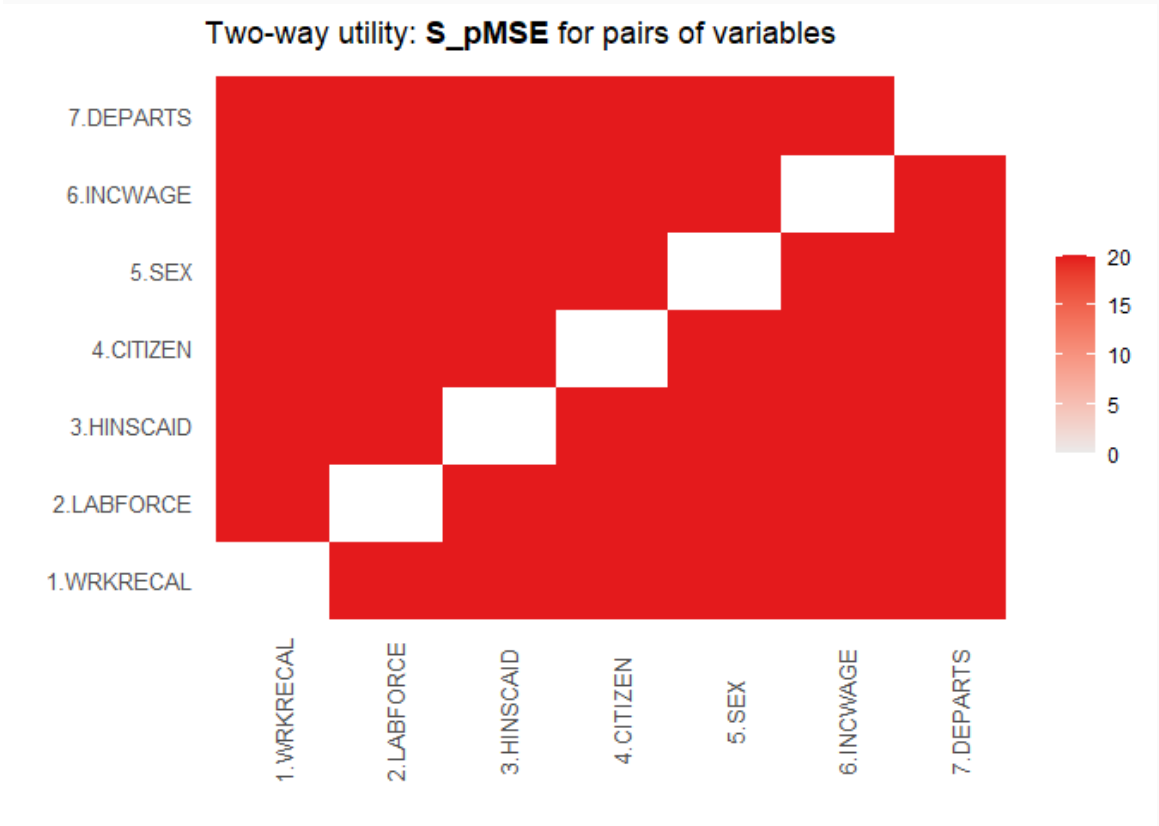
	pMSE	S_pMSE	df
ABSENT	0.0124855	103400.0692	2
EDUC	0.0004760	788.3904	10
HCOVANY	0.0001541	2552.1508	1
MARST	0.0002868	950.0797	5
HHWT	0.0104755	43376.9275	4
INCEARN	0.0729040	301881.2943	4

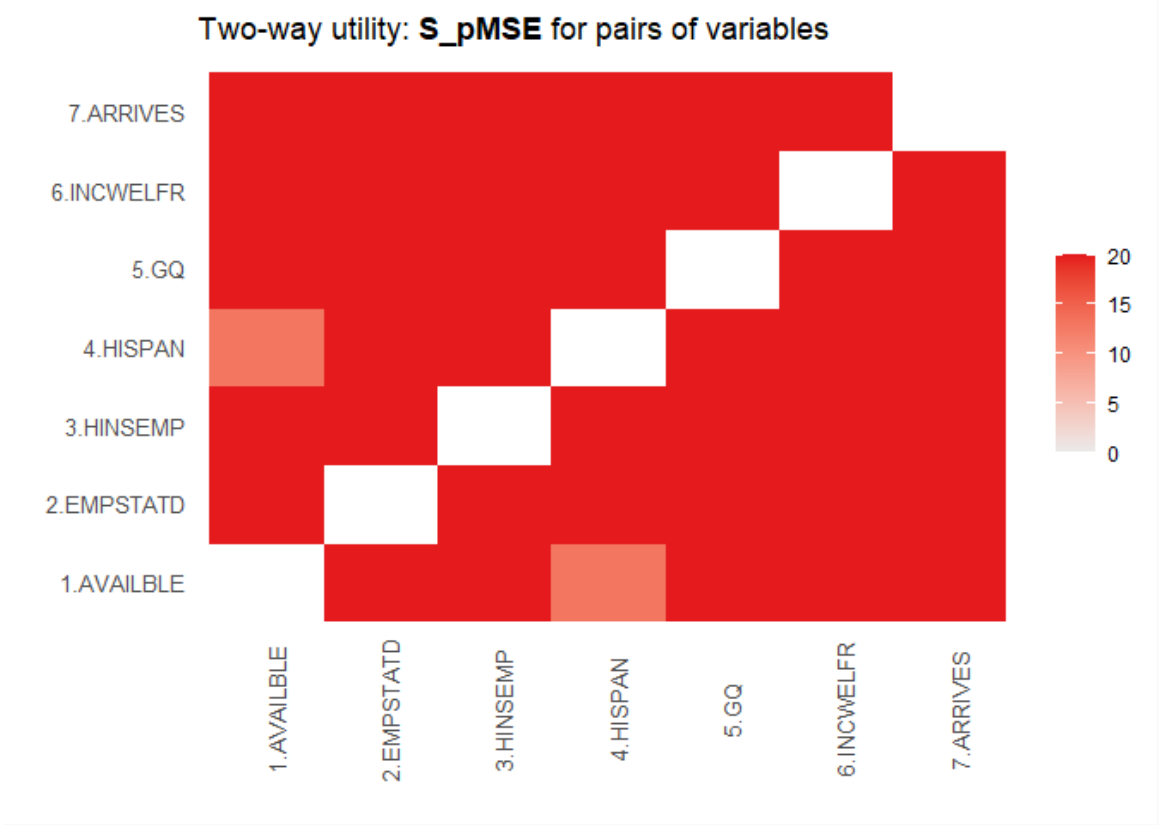
pMSE	S_pMSE
0.1159474	107.7212

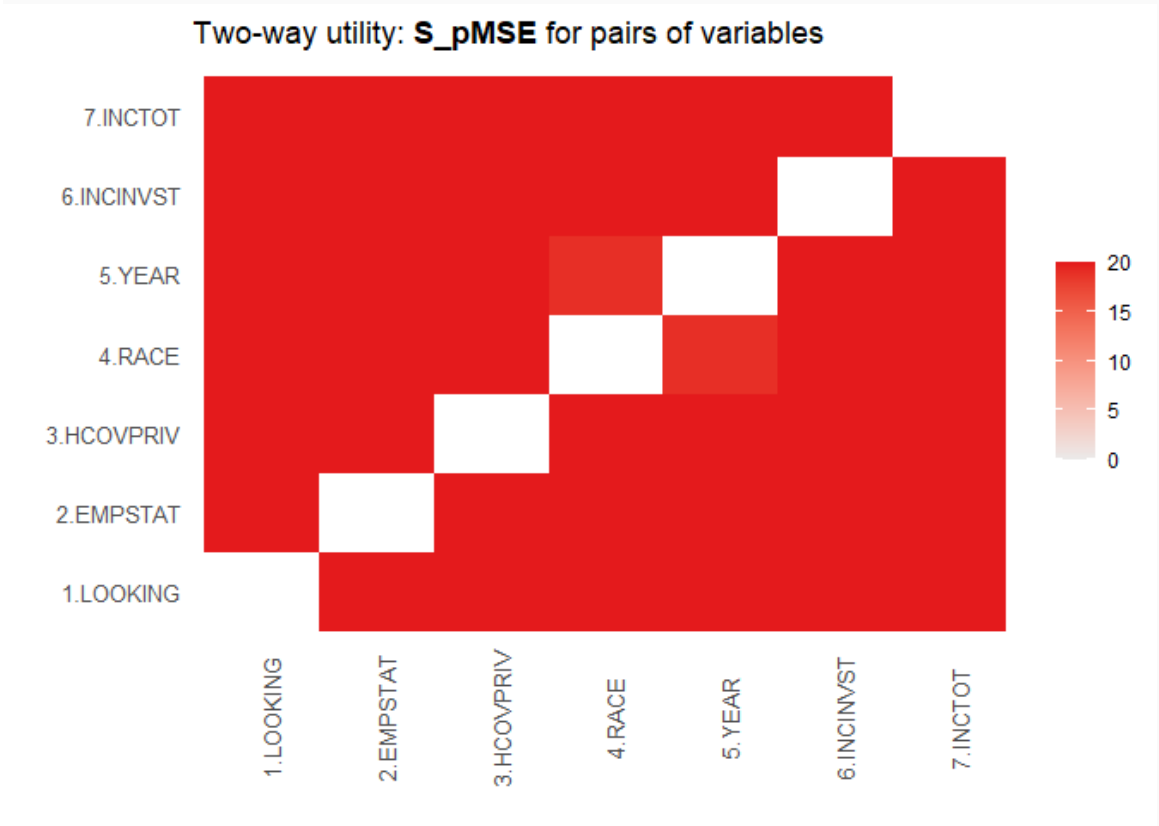
Two-way Tables Comparison of Synthesised Data (R-Package: synthpop) by (S_)pMSE

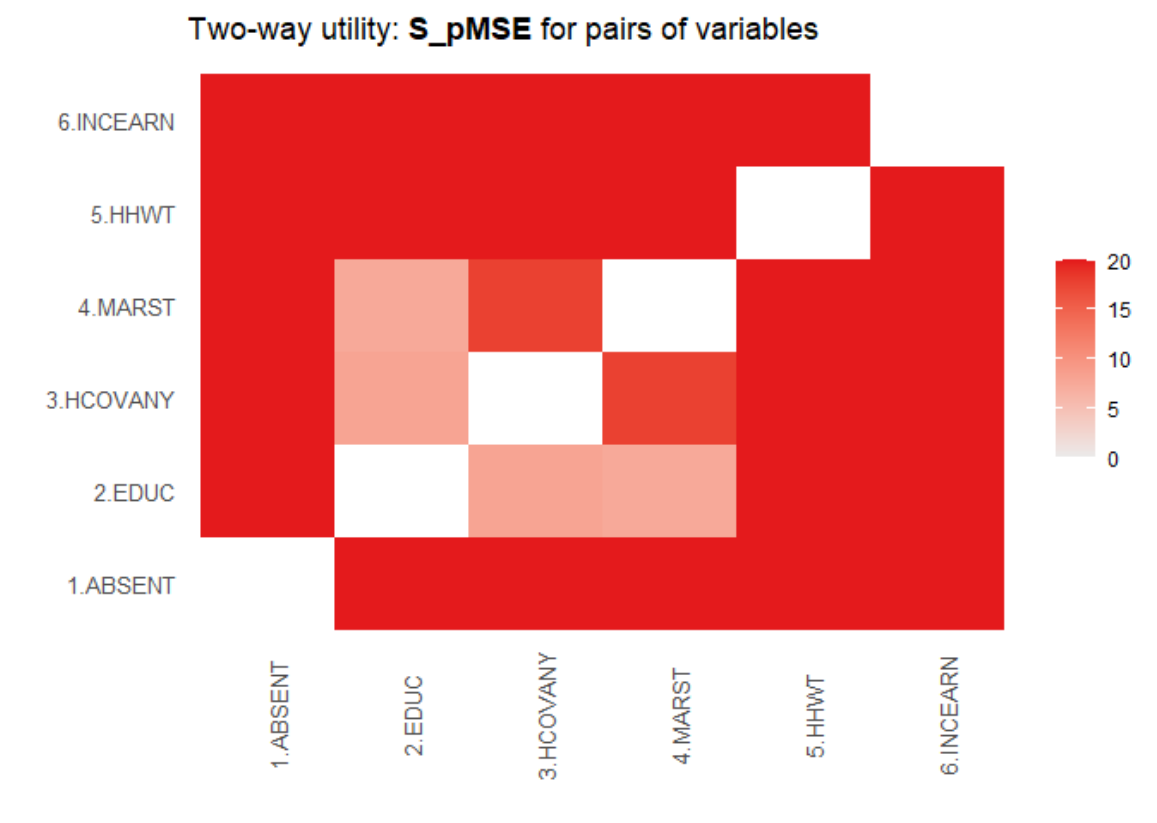
Two-way tables are evaluated based on the original and the synthetic dataset based on S_pMSE (see above). We also present the results for the mean absolute difference in densities (MabsDD) and the Bhattacharyya distance (dBhatt).

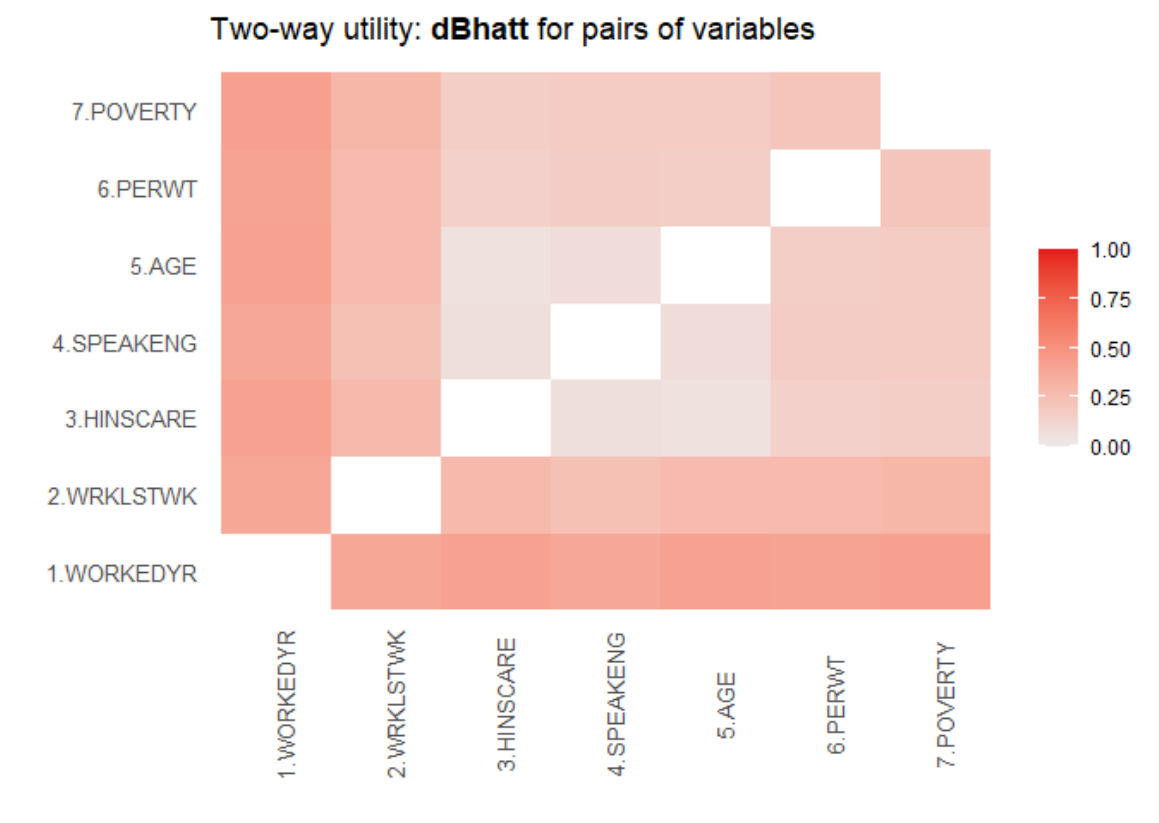

Two-way utility: S_pMSE for pairs of variables
7.POVERTY
6.PERWT
5.AGE
4.SPEAKENG
3.HINSCARE
2.WRKLSTWK
1.WORKEDYR
1.WORKEDYR
2.WRKLSTWK
3.HINSCARE
4.SPEAKENG
5.AGE
6.PERWT
7.POVERTY
20
15
10
5
0

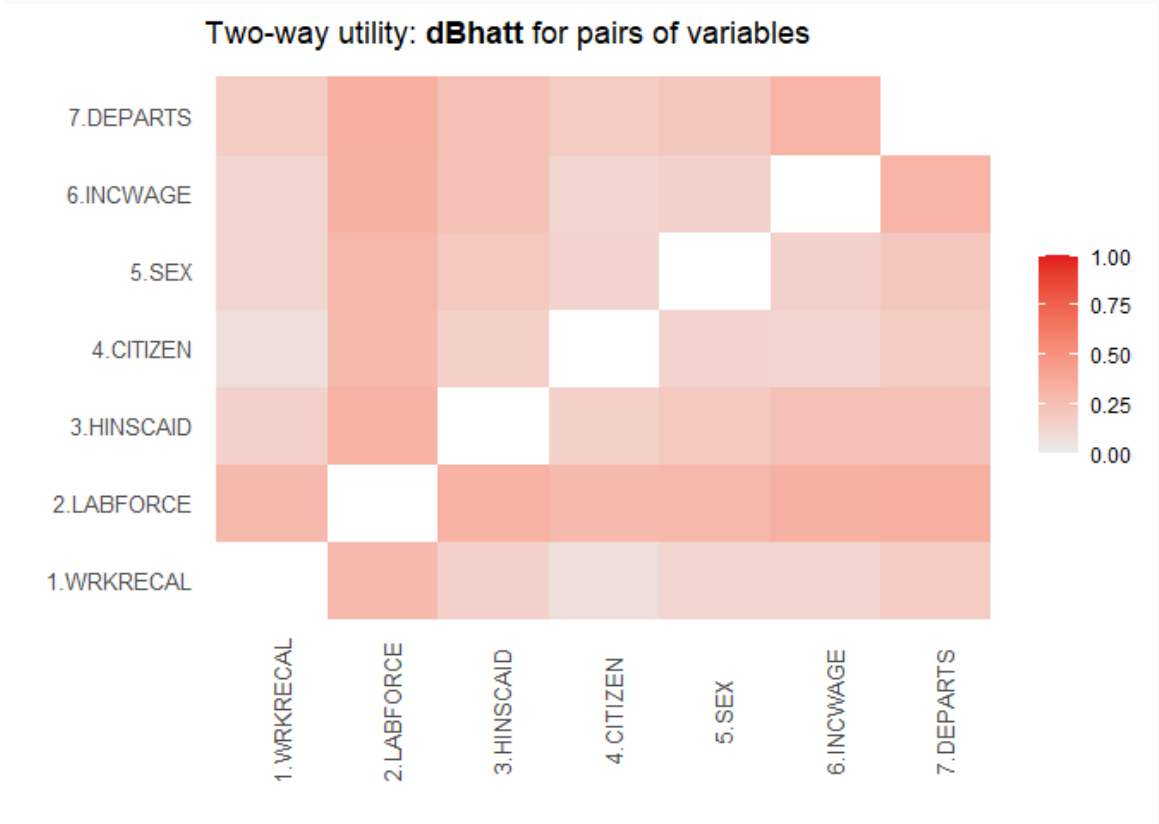

Two-way utility: S_pMSE for pairs of variables
7.DEPARTS
6.INCWAGE
5.SEX
4.CITIZEN
3.HINSCAID
2.LABFORCE
1.WRKRECAL
1.WRKRECAL
2.LABFORCE
3.HINSCAID
4.CITIZEN
5.SEX
6.INCWAGE
7.DEPARTS
20
15
10
5
0

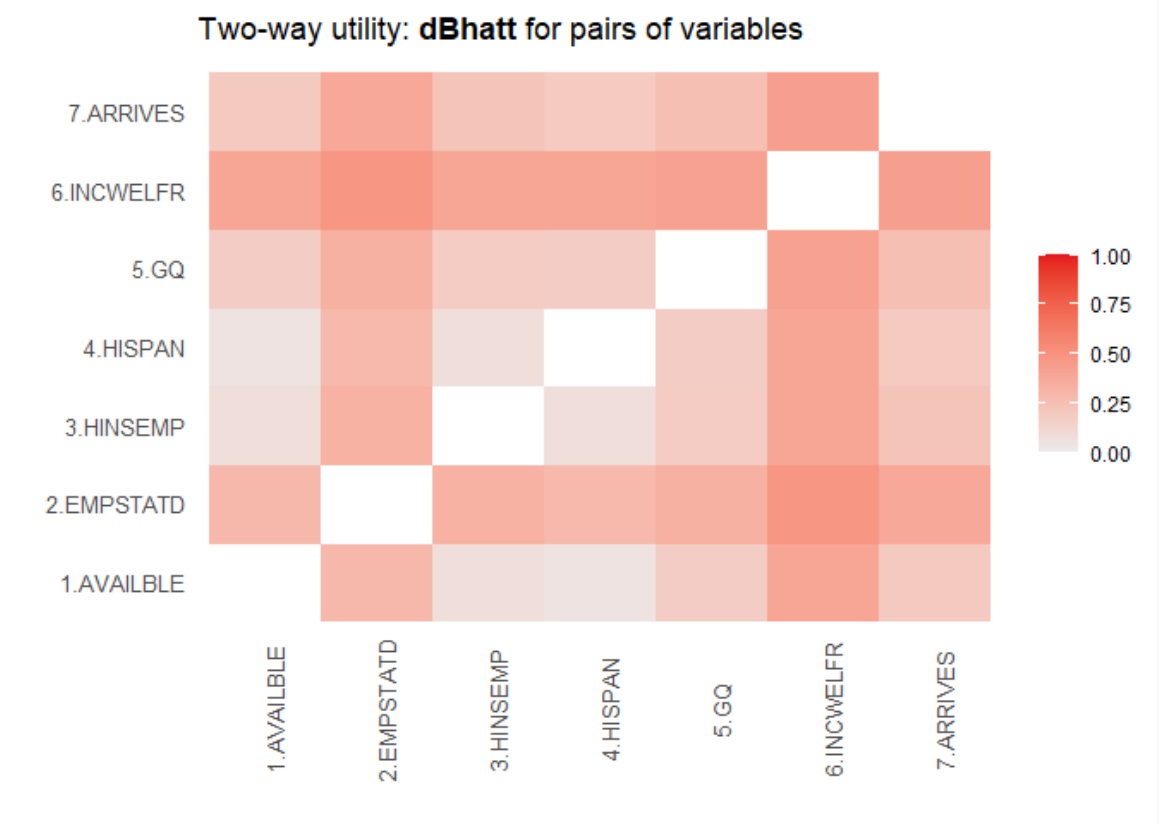

Two-way utility: S_pMSE for pairs of variables
7.ARRIVES
6.INCWELFR
5.GQ
4.HISPAN
3.HINSEMP
2.EMPSTATD
1.AVAILBLE
1.AVAILBLE
2.EMPSTATD
3.HINSEMP
4.HISPAN
5.GQ
6.INCWELFR
7.ARRIVES
20
15
10
5
0

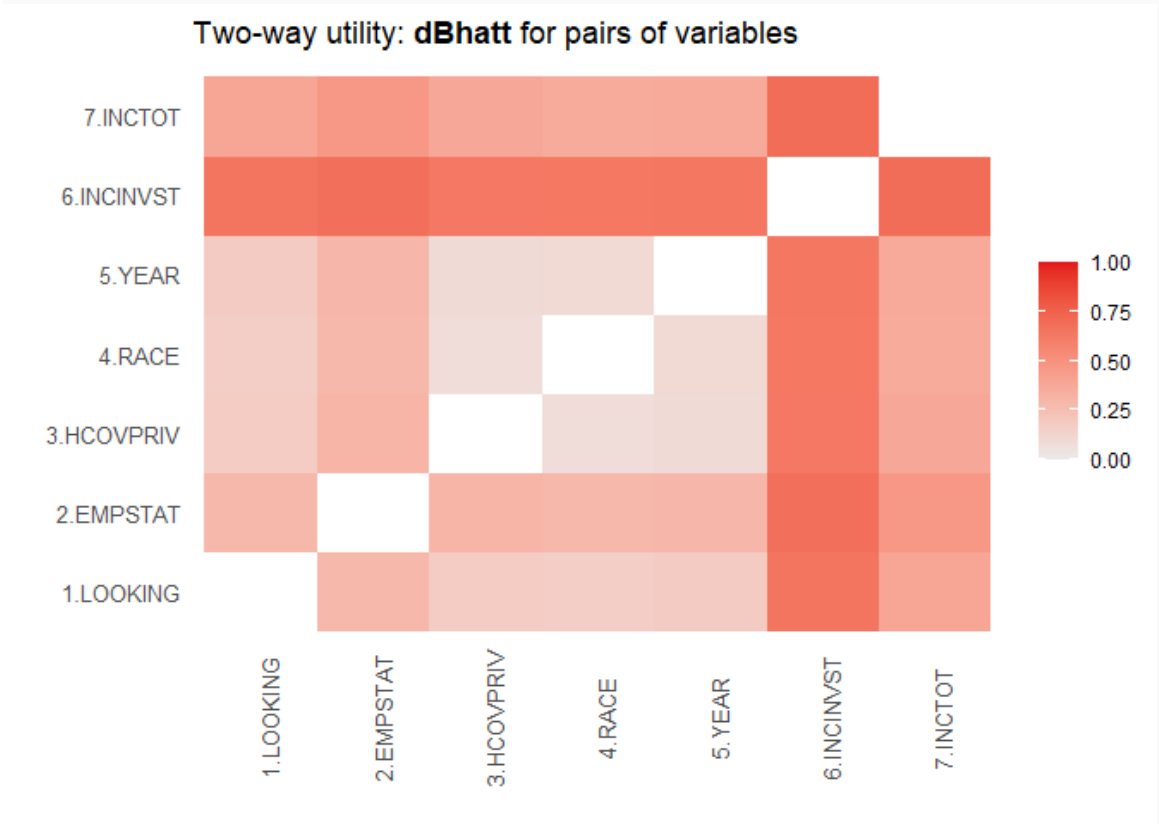

Two-way utility: S_pMSE for pairs of variables
7.INCTOT
6.INCINVST
5.YEAR
4.RACE
3.HCOVPRIV
2.EMPSTAT
1.LOOKING
1.LOOKING
2.EMPSTAT
3.HCOVPRIV
4.RACE
5.YEAR
6.INCINVST
7.INCTOT
20
15
10
5
0

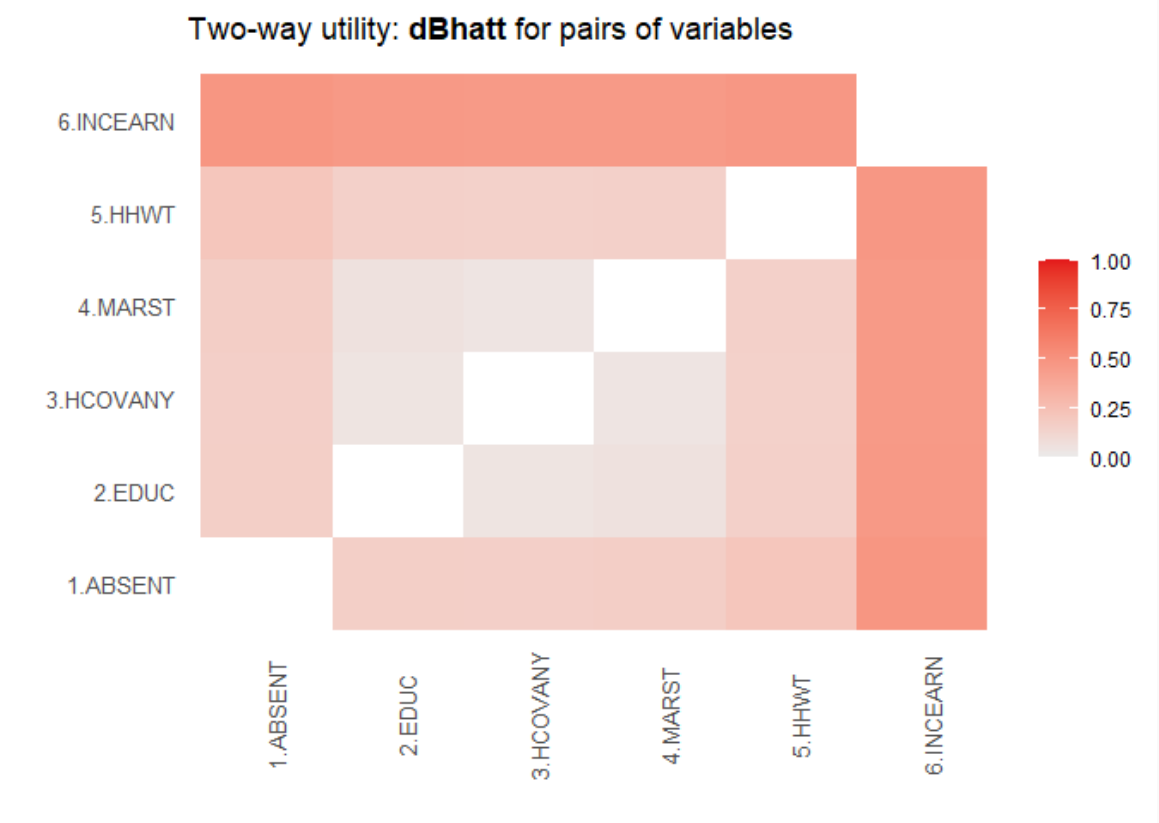

Two-way utility: S_pMSE for pairs of variables
6.INCEARN
5.HHWT
4.MARST
3.HCOVANY
2.EDUC
1.ABSENT
1.ABSENT
2.EDUC
3.HCOVANY
4.MARST
5.HHWT
6.INCEARN
20
15
10
5
0

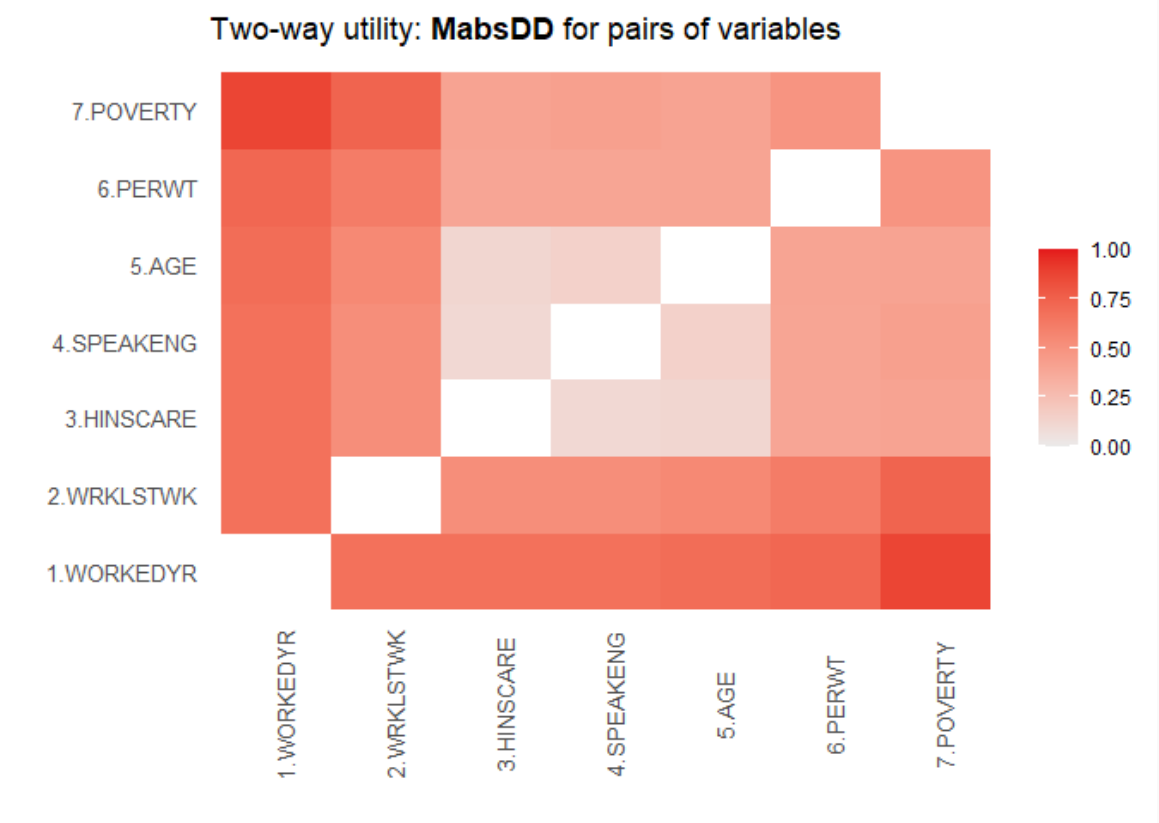

Two-way utility: dBhatt for pairs of variables
7.POVERTY
6.PERWT
5.AGE
4.SPEAKENG
3.HINSCARE
2.WRKLSTWK
1.WORKEDYR
1.00
0.75
0.50
0.25
0.00
1.WORKEDYR
2.WRKLSTWK
3.HINSCARE
4.SPEAKENG
5.AGE
6.PERWT
7.POVERTY

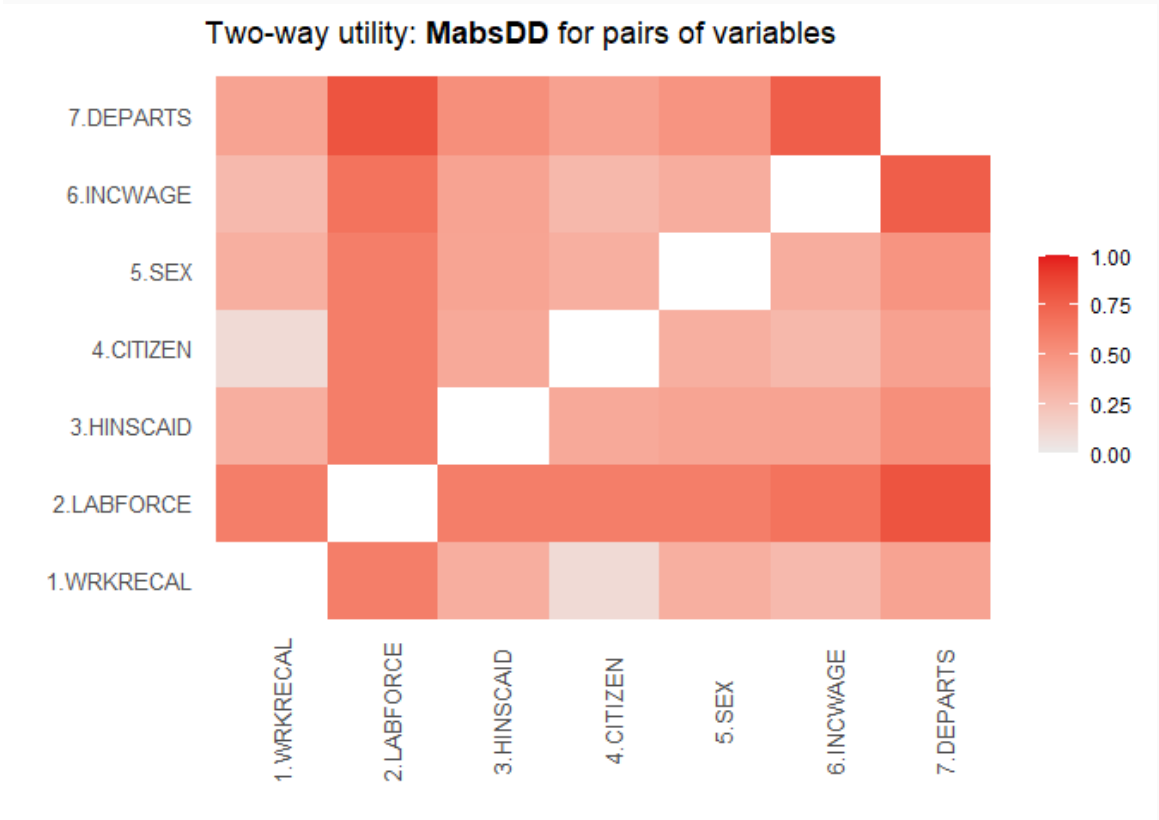

Two-way utility: dBhatt for pairs of variables
7.DEPARTS
6.INCWAGE
5.SEX
4.CITIZEN
3.HINSCAID
2.LABFORCE
1.WRKRECAL
1.00
0.75
0.50
0.25
0.00
1.WRKRECAL
2.LABFORCE
3.HINSCAID
4.CITIZEN
5.SEX
6.INCWAGE
7.DEPARTS

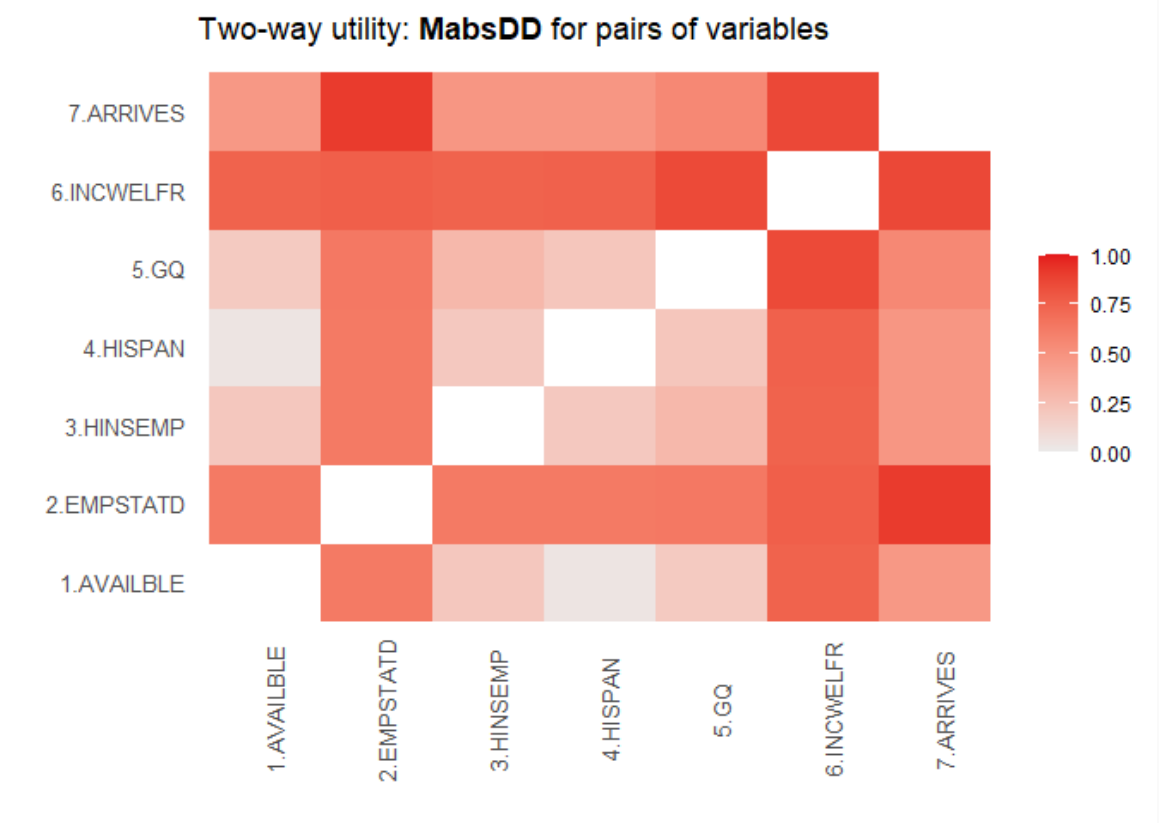

Two-way utility: dBhatt for pairs of variables
7.ARRIVES
6.INCWELFR
5.GQ
4.HISPAN
3.HINSEMP
2.EMPSTATD
1.AVAILBLE
1.AVAILBLE
2.EMPSTATD
3.HINSEMP
4.HISPAN
5.GQ
6.INCWELFR
7.ARRIVES
1.00
0.75
0.50
0.25
0.00

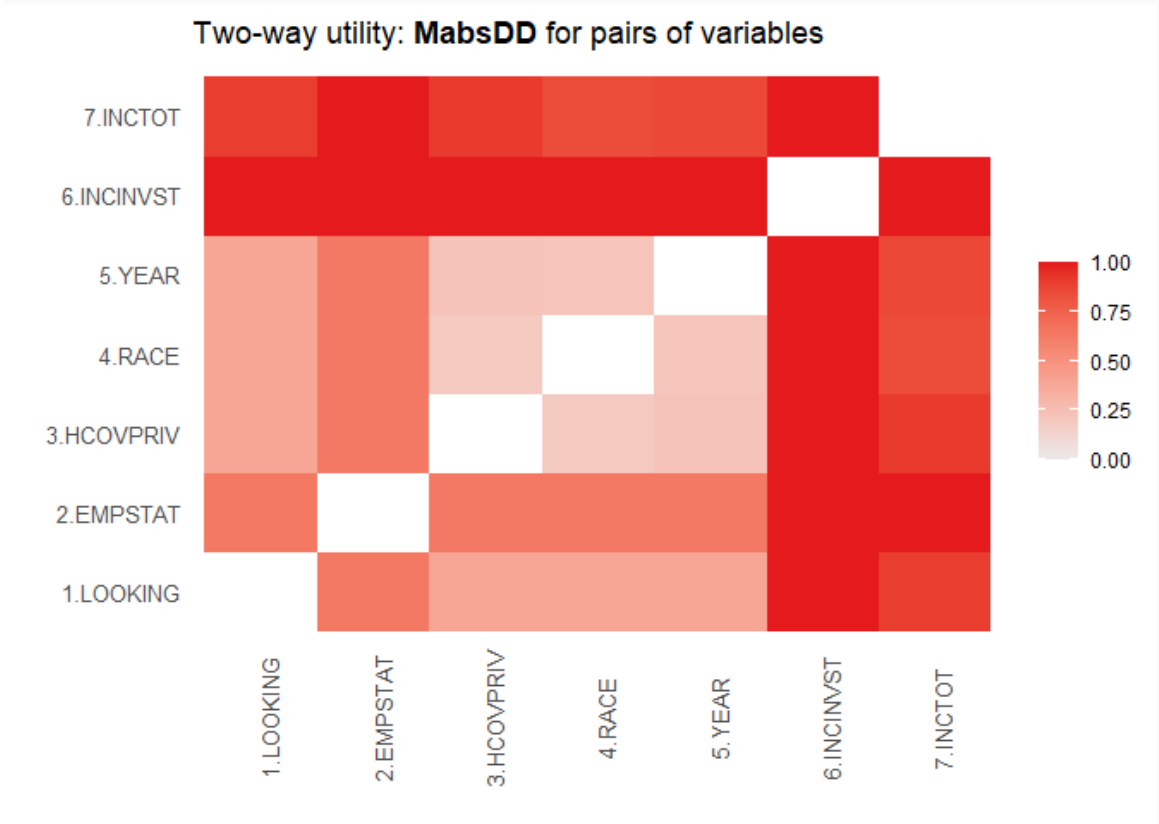

Two-way utility: dBhatt for pairs of variables
7.INCTOT
6.INCINVST
5.YEAR
4.RACE
3.HCOVPRIV
2.EMPSTAT
1.LOOKING
1.LOOKING
2.EMPSTAT
3.HCOVPRIV
4.RACE
5.YEAR
6.INCINVST
7.INCTOT
1.00
0.75
0.50
0.25
0.00

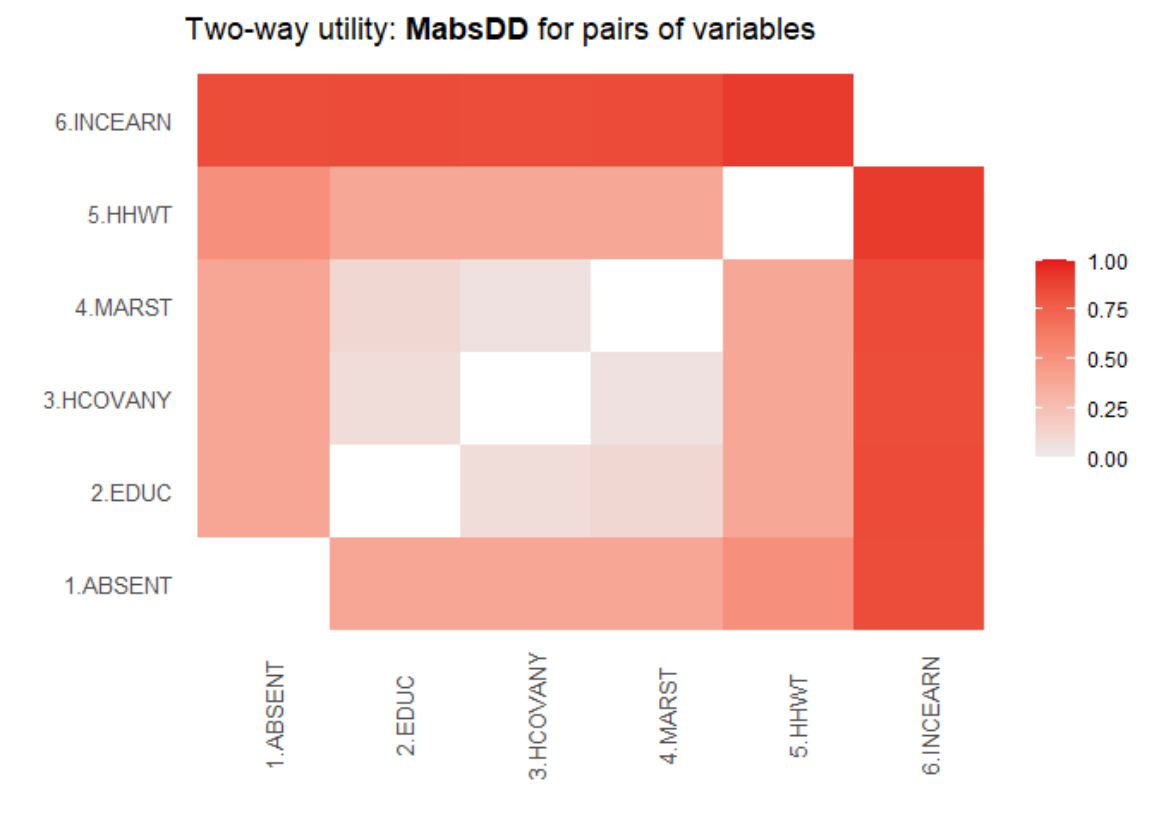

Two-way utility: dBhatt for pairs of variables
6.INCEARN
5.HHWT
4.MARST
3.HCOVANY
2.EDUC
1.ABSENT
1.00
0.75
0.50
0.25
0.00
1.ABSENT
2.EDUC
3.HCOVANY
4.MARST
5.HHWT
6.INCEARN


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


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


Two-way utility: MabsDD for pairs of variables
7.ARRIVES
6.INCWELFR
5.GQ
4.HISPAN
3.HINSEMP
2.EMPSTATD
1.AVAILBLE
1.AVAILBLE
2.EMPSTATD
3.HINSEMP
4.HISPAN
5.GQ
6.INCWELFR
7.ARRIVES
1.00
0.75
0.50
0.25
0.00


Two-way utility: MabsDD for pairs of variables
7.INCTOT
6.INCINVST
5.YEAR
4.RACE
3.HCOVPRIV
2.EMPSTAT
1.LOOKING
1.LOOKING
2.EMPSTAT
3.HCOVPRIV
4.RACE
5.YEAR
6.INCINVST
7.INCTOT
1.00
0.75
0.50
0.25
0.00



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
0.5511464

Individual Distances for Information Loss:

```
##   WORKEDYR   WRKRECAL   AVAILBLE    LOOKING     ABSENT   WRKLSTWK   LABFORCE
## 0.35020542 0.13346780 0.17000370 0.44859694 0.44324339 0.50753815 0.38504310
##   EMPSTATD    EMPSTAT       EDUC   HINSCARE   HINSCAID    HINSEMP   HCOVPRIV
## 0.45931466 0.42927316 0.75877052 0.39350619 0.35090770 0.49670450 0.41670555
##    HCOVANY   SPEAKENG    CITIZEN     HISPAN       RACE      MARST        AGE
## 0.14876241 0.17623051 0.13806980 0.07408127 0.25387920 0.59861225 0.90945726
##        SEX         GQ       YEAR       HHWT      PERWT    INCWAGE   INCWELFR
## 0.50577907 0.14450430 0.85777351 0.96739287 0.96902225 0.89616969 0.53305494
##   INCINVST    INCEARN    POVERTY    DEPARTS    ARRIVES     INCTOT
## 0.99965790 0.99989706 0.98568138 0.91795015 0.91978442 0.99993769
```

Tuning and Optimizations

Addtionally to fitting multivariate normal distributions, we tested an approach with non-normal multivariate distributions. We were not able to fit a more flexibel multivariate distributions, supposedly caused among other characteristics by extreme skewness in some variables.