ACS - Information Preserving Statistical Obfuscation (IPSO)

Evaluation Synthetic Data Creation

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

January 31, 2022

- Executive Summary
- Dataset Considerations
- Method Considerations
- Privacy and Risk Evaluation
- Utility Evaluation
- Tuning and Optimizations

Executive Summary

We created two versions of the IPSO-generated synthetic ACS dataset. The difference is the variable HHWT, which is considered as confidential and is hence synthetically generated in the second version and presented firstly. The results do not differ relevantly from the first version (see section **tuning and optimization**). Since IPSO did not score too well on our privacy metrics, we would only release the synthetic data to trusted partners. Thus, **education** and **releasing to the public** does not seem like a good option for us. We also think there are better options for **technology testing**,. We could imagine **testing analysis** could be a good fit, if the choice of **confidetial** and **non-confidential** is a good fit for the analysis planned by the researchers.

We found **IPSO** algorithm is not suitable to generate synthetic data from the ACS dataset. Basically all marginal distributions are aligning. Hence, a high utility for the original variables should not be surprising. The utility measures for the synthetic part are not supporting the usage. The S_pMSE for tables is usually high for the synthetic part. Also the absolute difference in densities and the Bhattacharyya are not supporting a high utility accordingly. Mlodak's information loss criterion underpins the overall impression.

USE CASE RECOMMENDATIONS

Releasing_to_Public	Testing_Analysis	Education	Testing_Technology
NO	YES	NO	MAYBE

Because of the trade-off with privacy we probably would only supply the IPSO synthetic data to highly 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. **Realeasing 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

Similar to the FCS-method, IPSO is easy to understand and to explain, since it is based on classic linear regression. For applying IPSO, we chose the R package RegSDC that provides a framework containing several versions of the method. We applied the classical version of IPSO provided by the function RegSDCipso.

IPSO requires to split up the variables in non-confidential ones and confidential ones. It assumes statistical independence among the non-confidential variables and multivariate normally distributed confidential variables. The assumption is strong and holds in general not for the ACS dataset, therefore poor quality of the synthetical data is inevitable. We classified the sat-related variables and sex as non-confidential and the gpa-related variables as confidential.

The computation time for IPSO was superb. Since the basic assumptions of IPSO are not fulfilled in the ACS data set, we have dispensed with parameter tuning.

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:

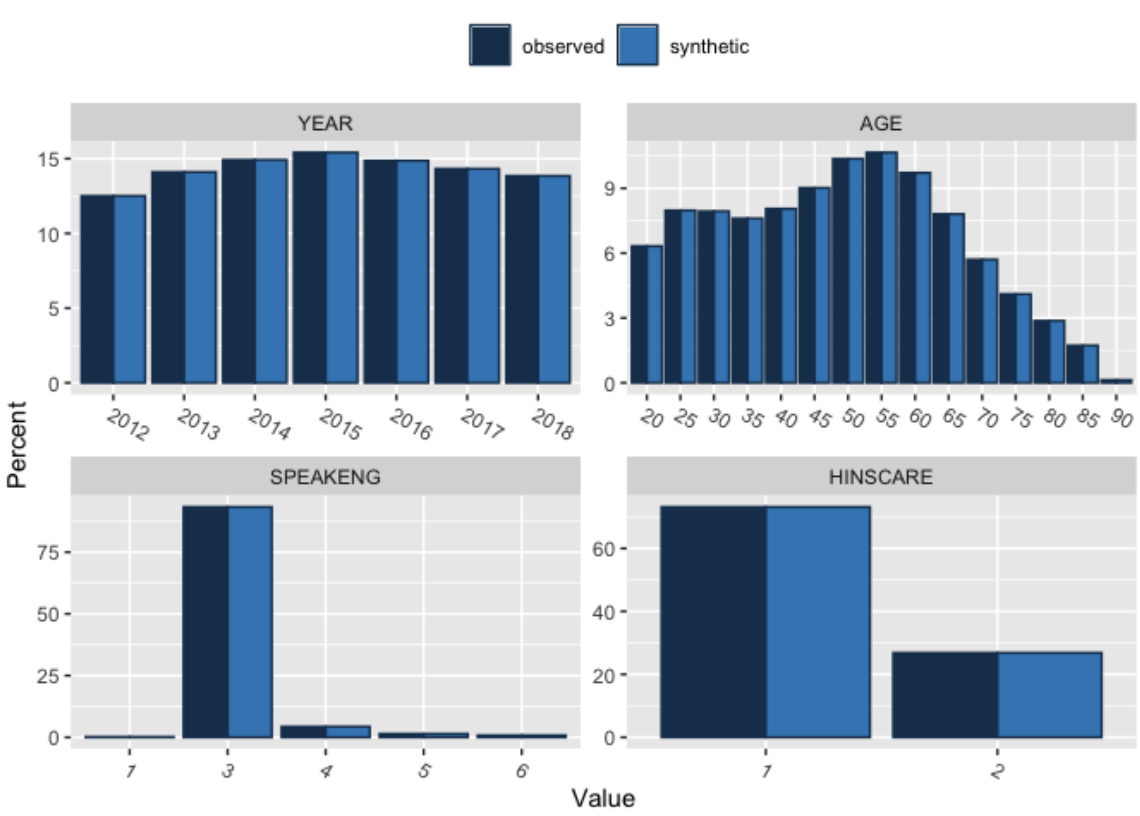
Metric	Number.Uniques	Number.Replications	Percentage.Replications
Perceived Risk	1001862	0	0

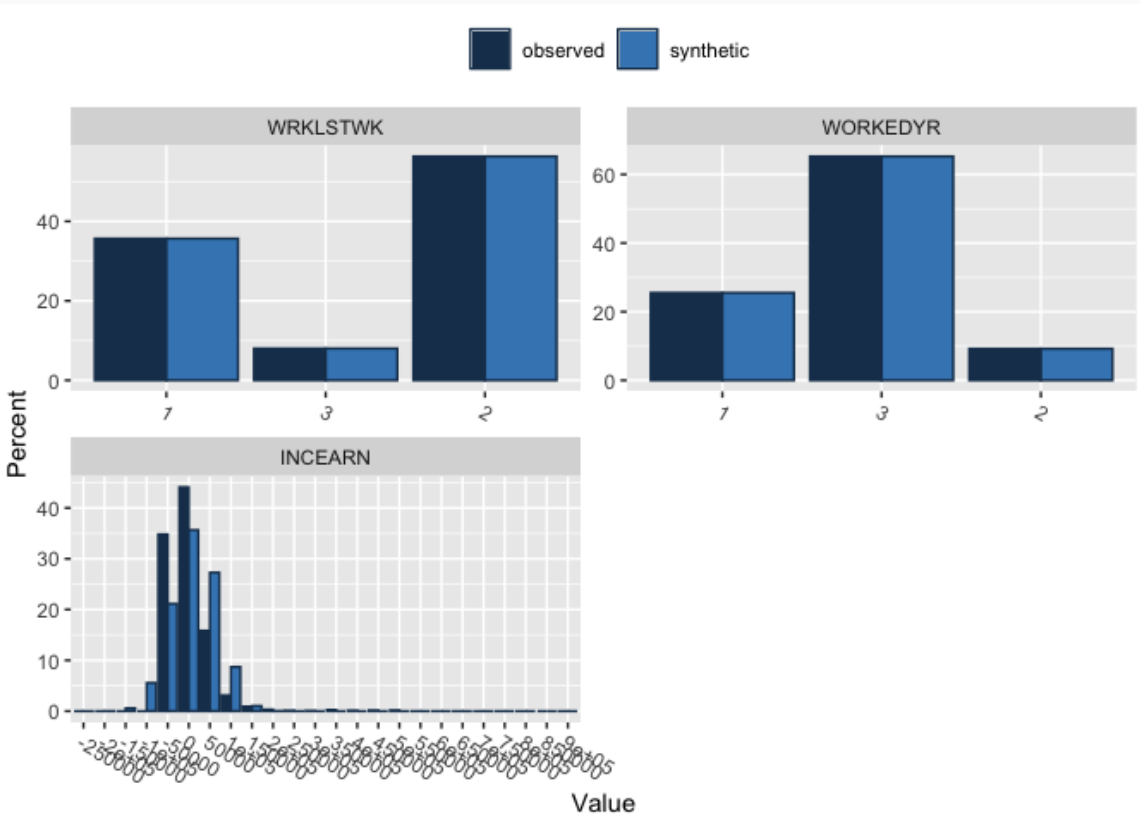
Utility Evaluation

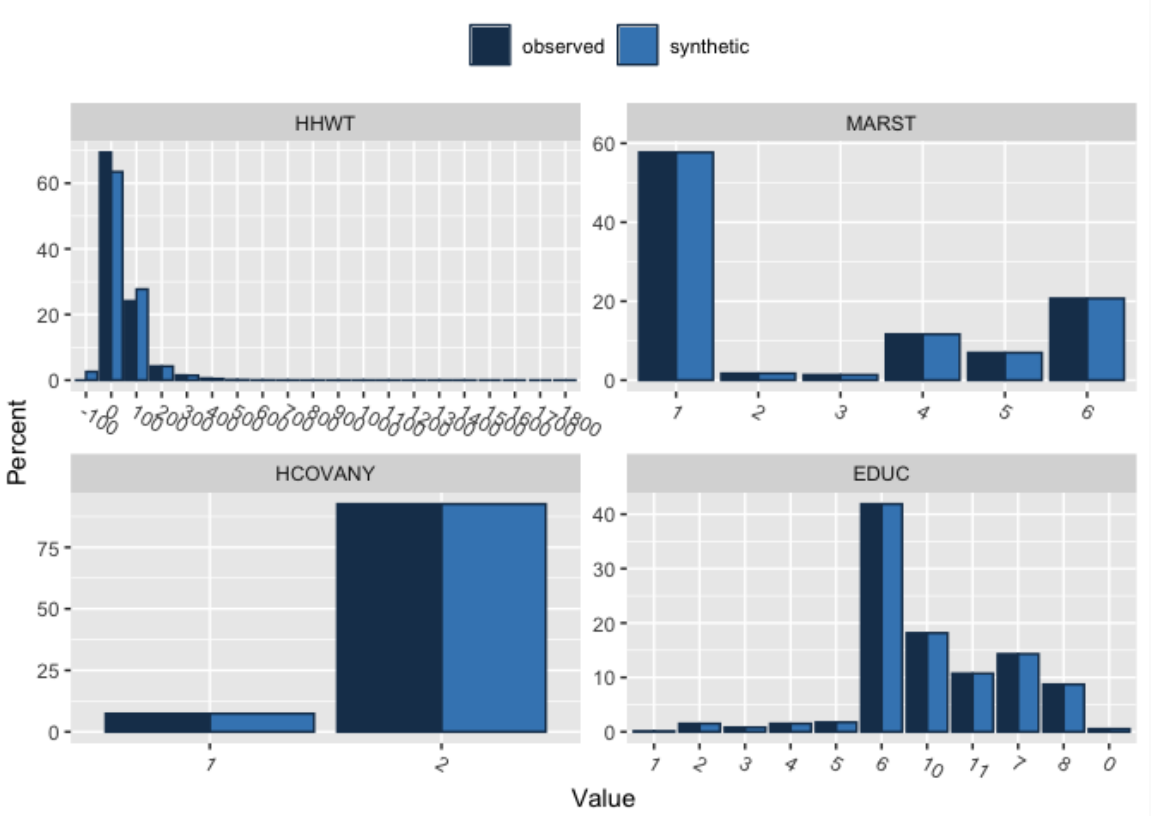
Different utility measures are applied in this section. These utility measures are the basis of utility evaluation for the generated synthetic dataset. The R packages synthpop, sdcMicro and corrplot were used to compute the following metrics. We do not use tests incorporating significance here. Confidence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

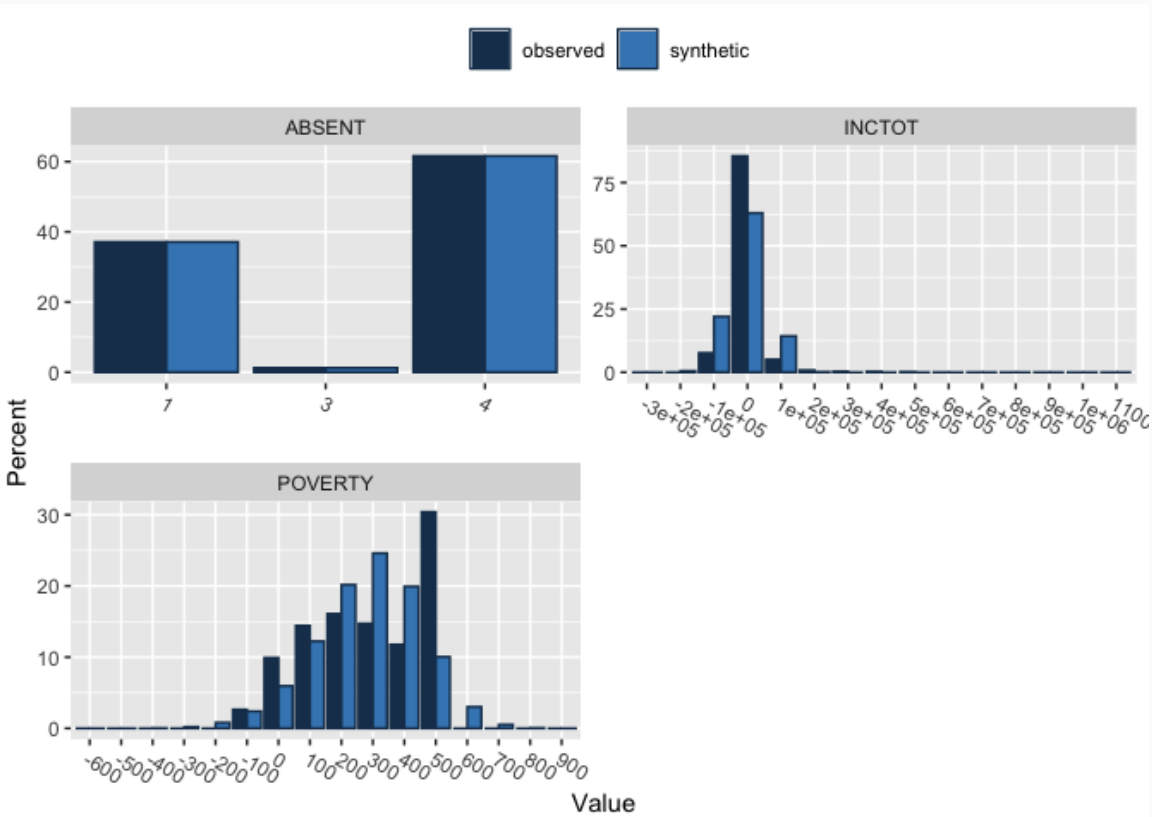
Graphical Comparison for Margins (R-Package: synthpop)

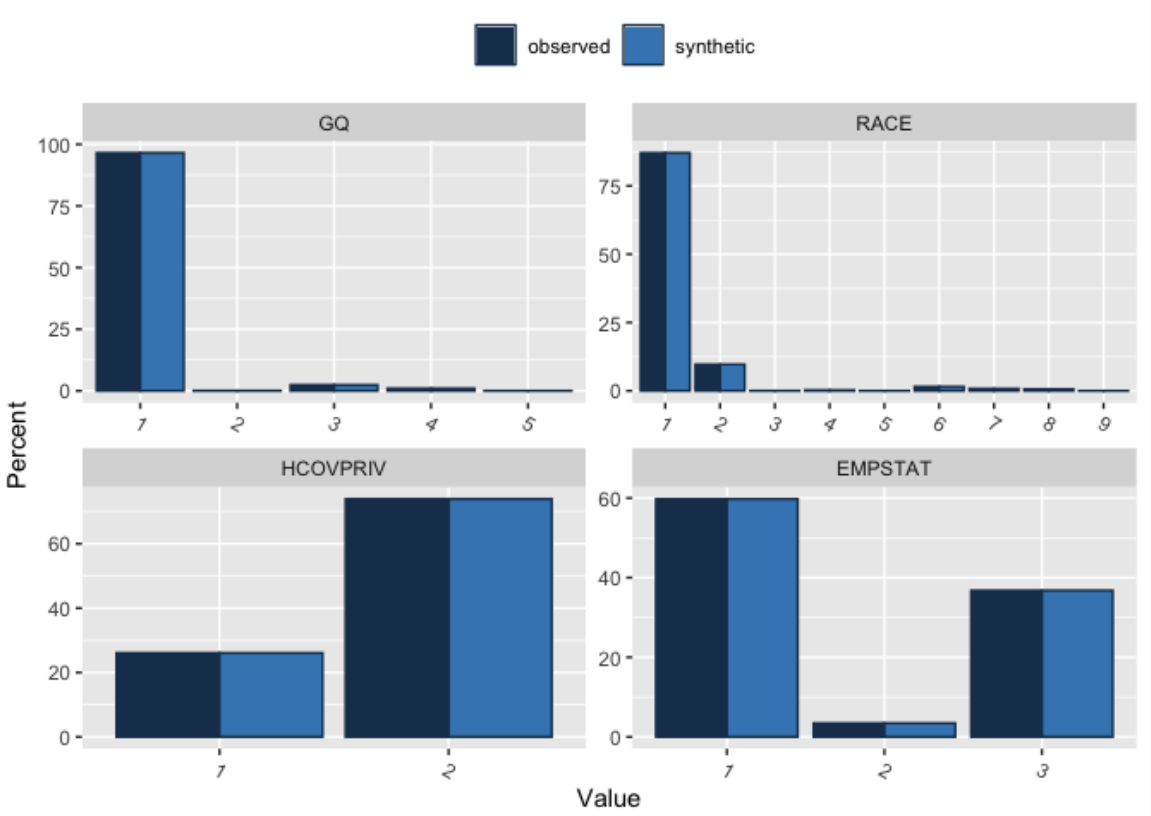
The following histograms provide an ad-hoc overview on the marginal distributions of the original and synthetic dataset. Matching or close distributions are related to a high data utility.

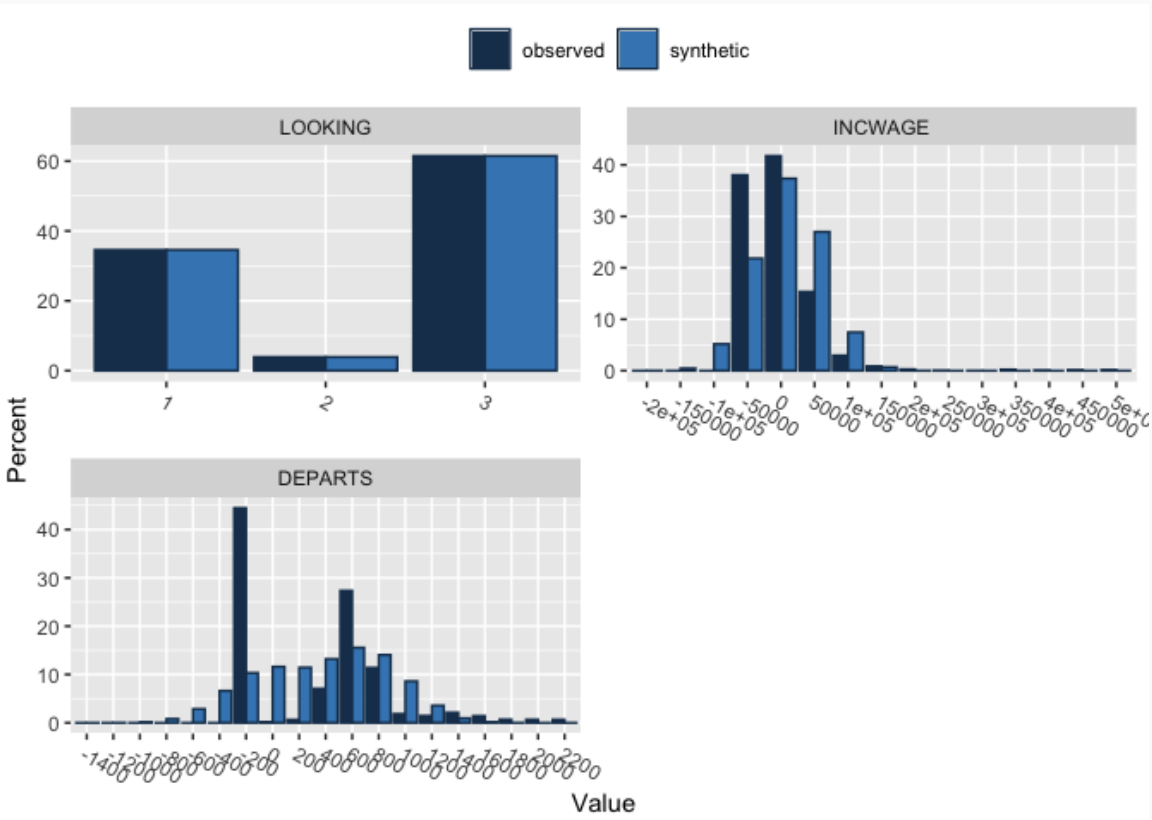

observed
synthetic
YEAR
AGE
SPEAKENG
HINSCARE
Percent
Value

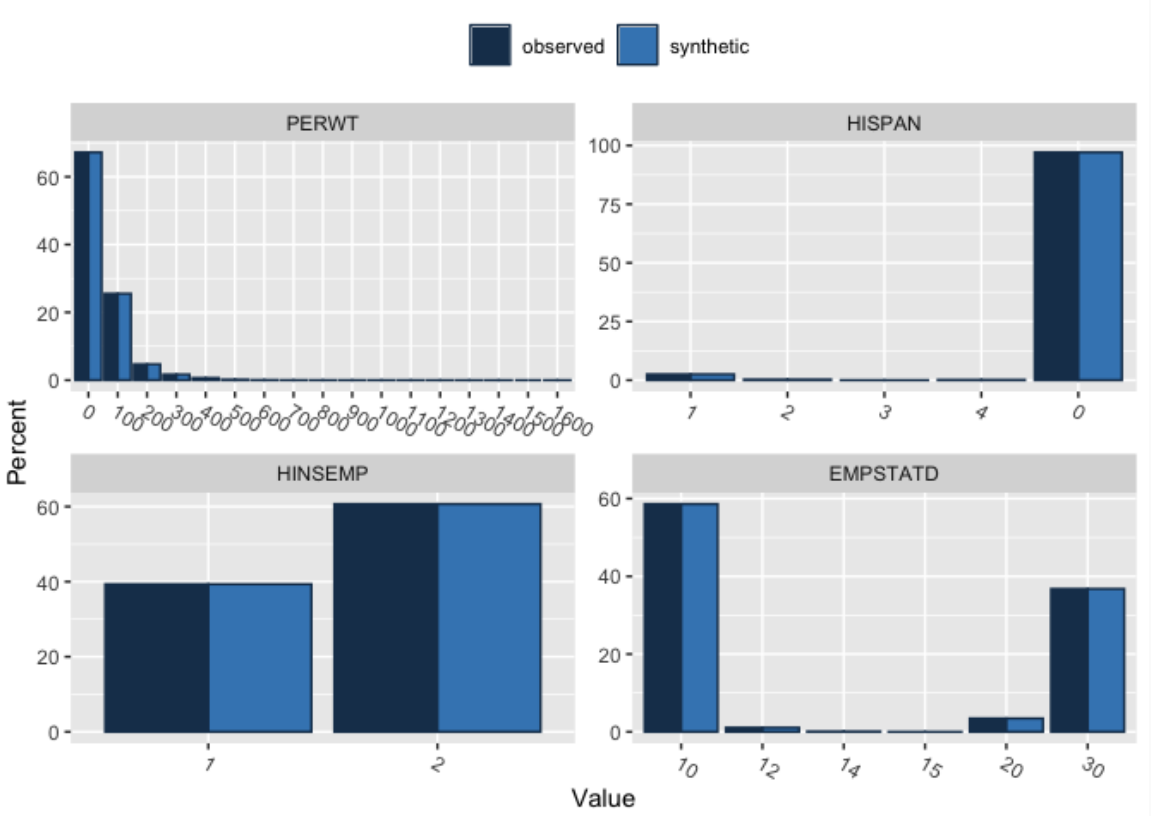

observed
synthetic
WRKLSTWK
WORKEDYR
INCEARN
Percent
Value

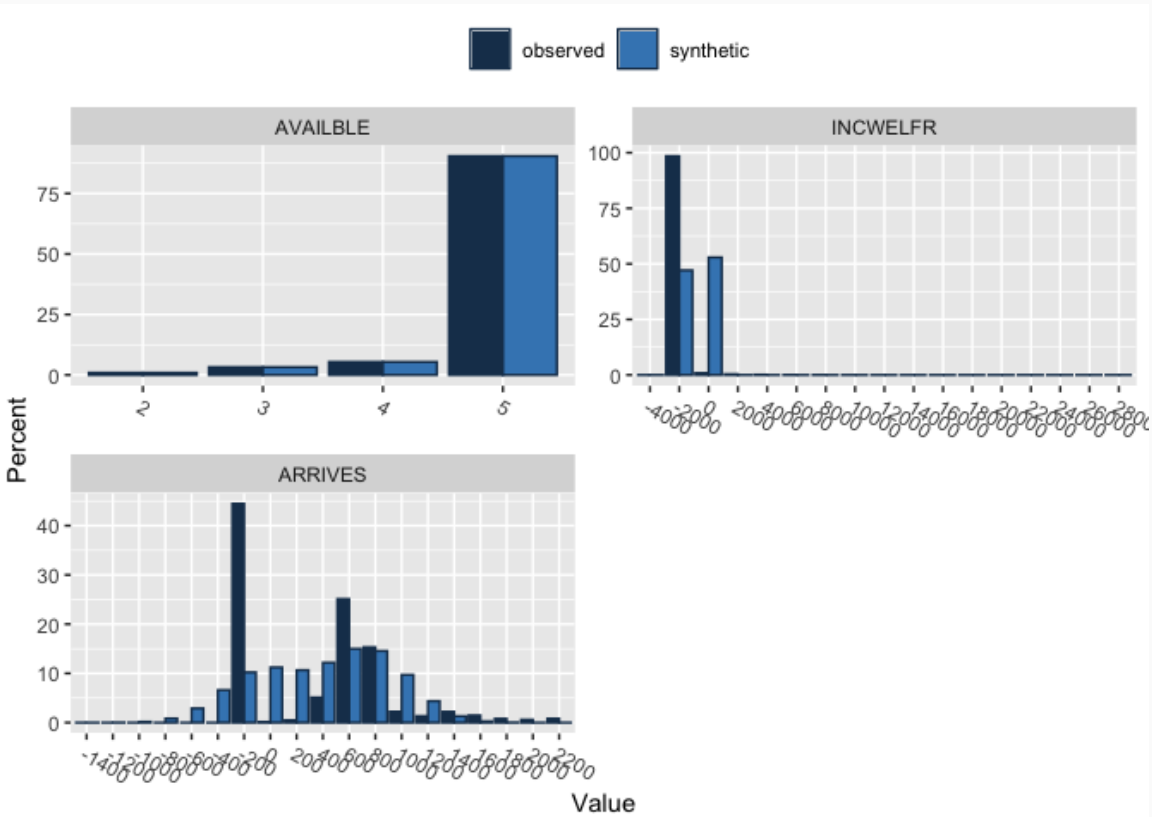

observed
synthetic
HHWT
MARST
HCOVANY
EDUC
Percent
Value

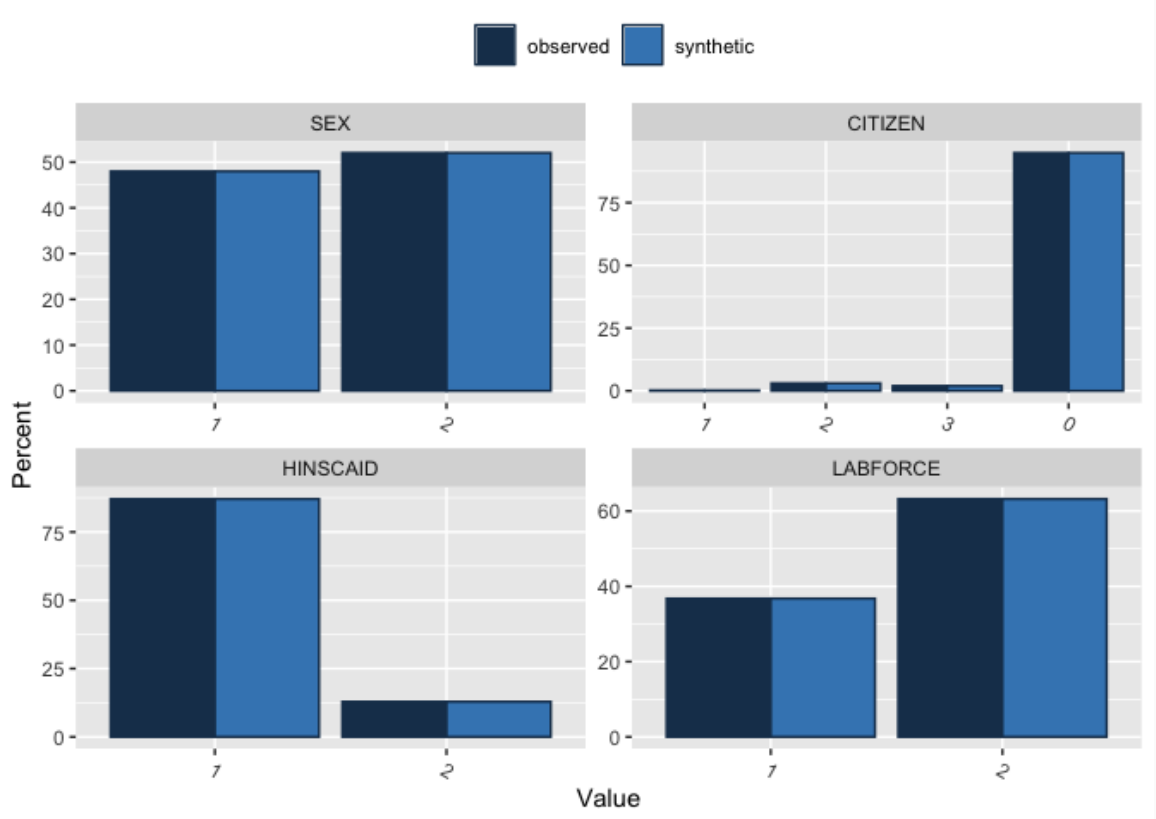

observed
synthetic
ABSENT
INCTOT
POVERTY
Percent
Value

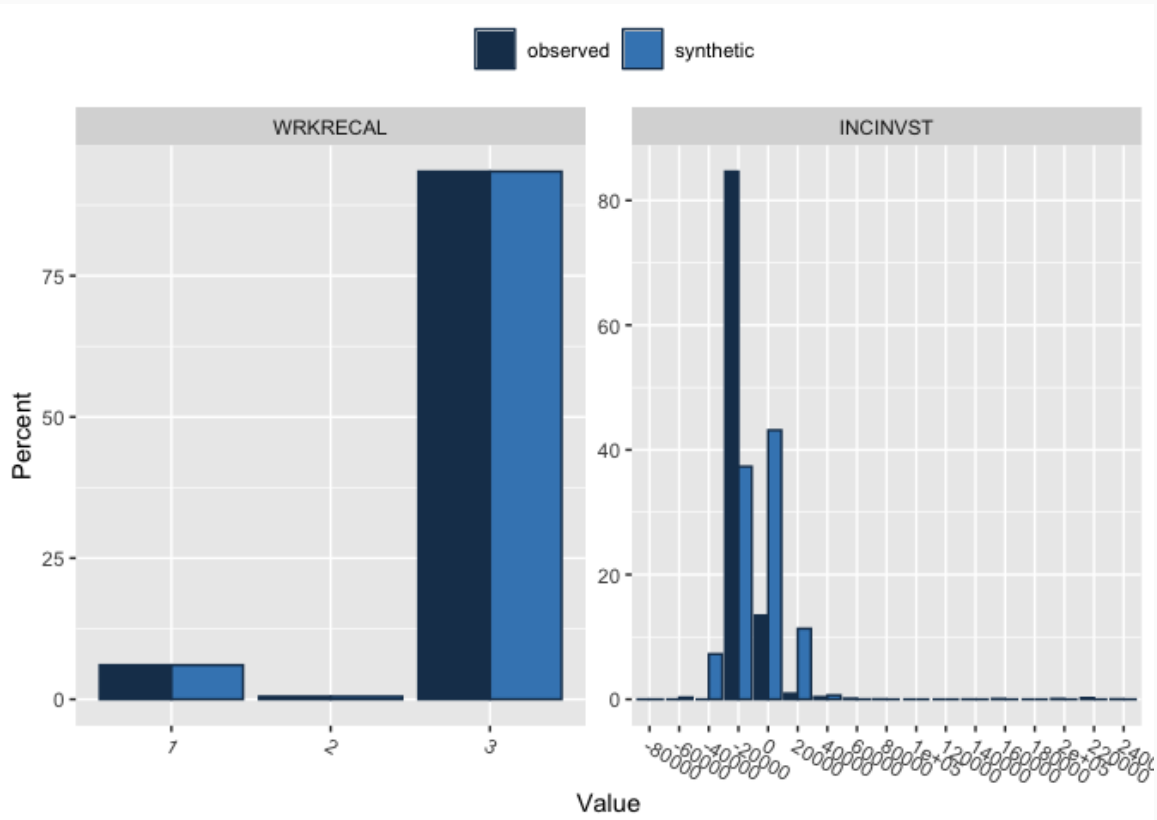

observed
synthetic
GQ
RACE
HCOVPRIV
EMPSTAT
Percent
Value


observed
synthetic
LOOKING
INCWAGE
DEPARTS
Percent
Value


observed
synthetic
PERWT
HISPAN
HINSEMP
EMPSTATD
Percent
Value

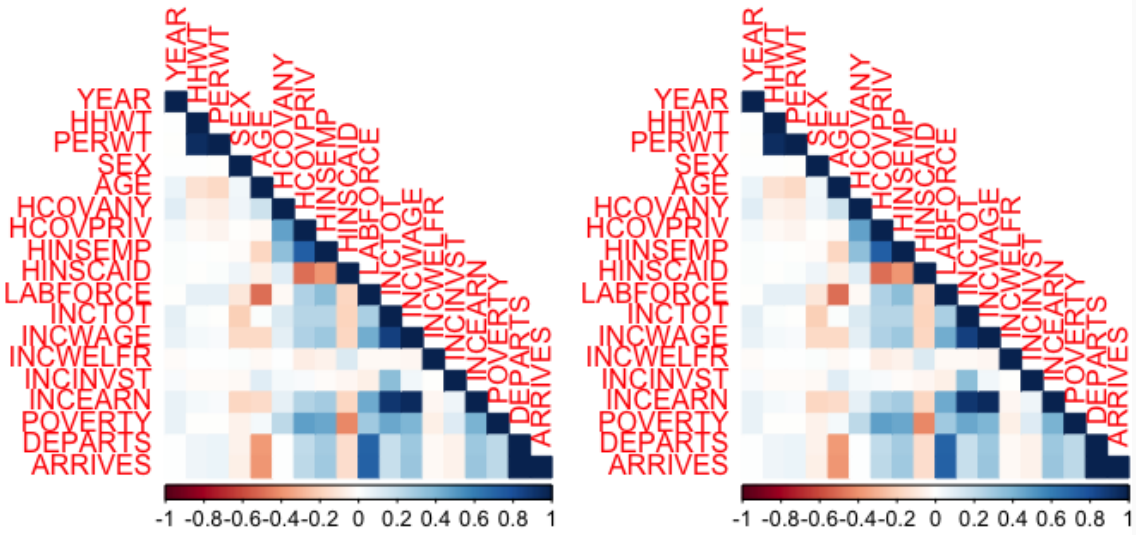

observed
synthetic
AVAILBLE
INCWELFR
ARRIVES
Percent
Value





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	0.0000000	0.0	6
AGE	0.0000000	0.0	4
SPEAKENG	0.0000000	0.0	4
HINSCARE	0.0000000	0.0	1
WRKLSTWK	0.0000000	0.0	2
WORKEDYR	0.0000000	0.0	2
INCEARN	0.0736501	304970.6	4

pMSE	S_pMSE
0.1642919	194.7155

	pMSE	S_pMSE	df
HHWT	0.0013717	5679.863	4

	pMSE	S_pMSE	df
MARST	0.0000000	0.000	5
HCOVANY	0.0000000	0.000	1
EDUC	0.0000000	0.000	10
ABSENT	0.0000000	0.000	2
INCTOT	0.0236127	97775.741	4
POVERTY	0.0151811	62861.860	4

pMSE	S_pMSE
0.1452538	102.303

	pMSE	S_pMSE	df
GQ	0.0000000	0.0	4
RACE	0.0000000	0.0	8
HCOVPRIV	0.0000000	0.0	1
EMPSTAT	0.0000000	0.0	2
LOOKING	0.0000000	0.0	2
INCWAGE	0.0758063	313899.0	4
DEPARTS	0.0579339	239892.8	4

pMSE	S_pMSE
0.2058324	317.3236

	pMSE	S_pMSE	df
PERWT	0.0000000	0.0	4
HISPAN	0.0000000	0.0	4
HINSEMP	0.0000000	0.0	1
EMPSTATD	0.0000000	0.0	5
AVAILBLE	0.0000000	0.0	3
INCWELFR	0.1813874	1001453.1	3
ARRIVES	0.0567102	234825.9	4

pMSE	S_pMSE
0.2489073	298.5004

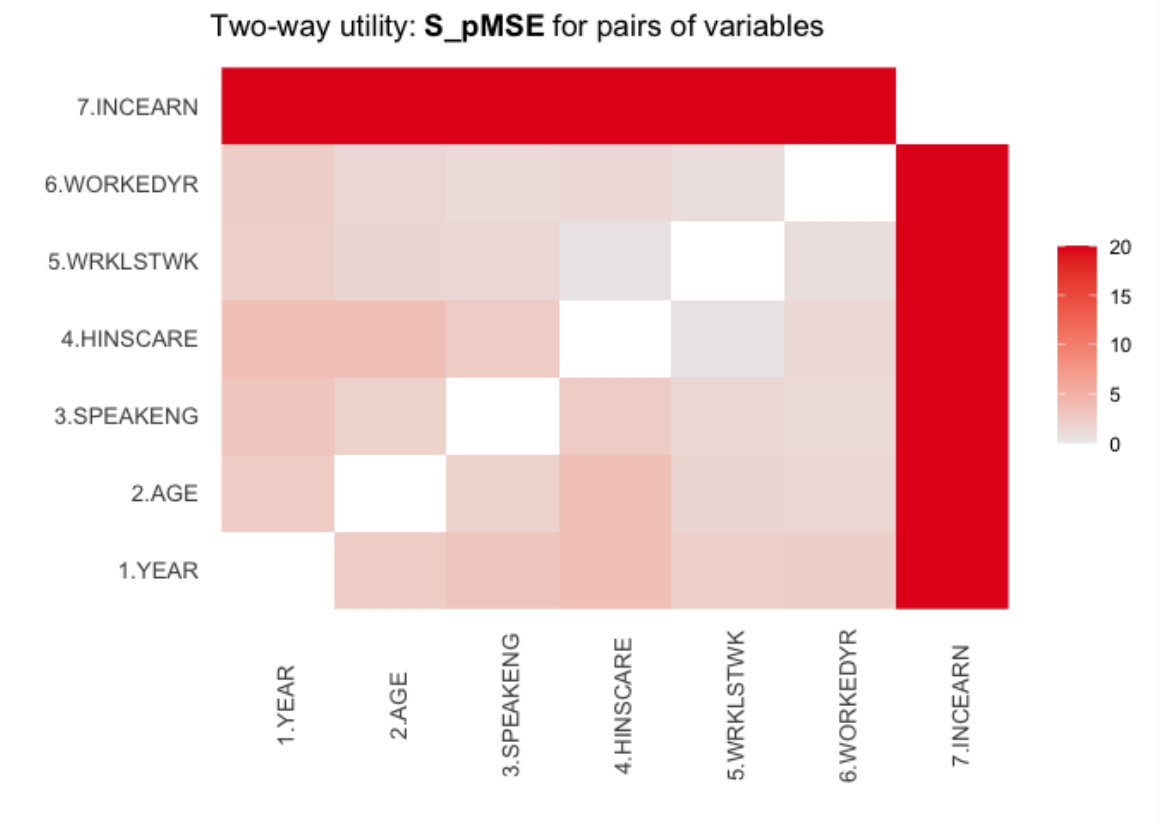
	pMSE	S_pMSE	df

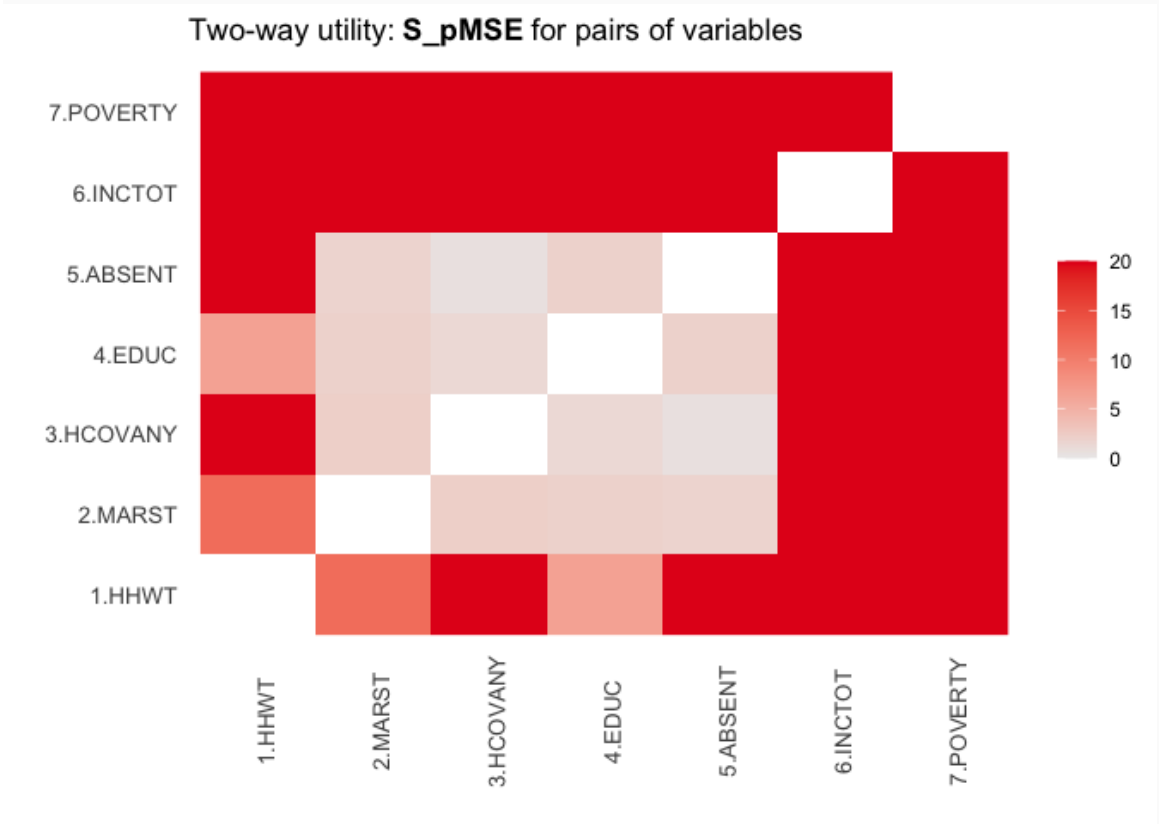
	pMSE	S_pMSE	df
SEX	0.0000000	0.0	1
CITIZEN	0.0000000	0.0	3
HINSCAID	0.0000000	0.0	1
LABFORCE	0.0000000	0.0	1
WRKRECAL	0.0000000	0.0	2
INCINVST	0.1484874	819809.7	3

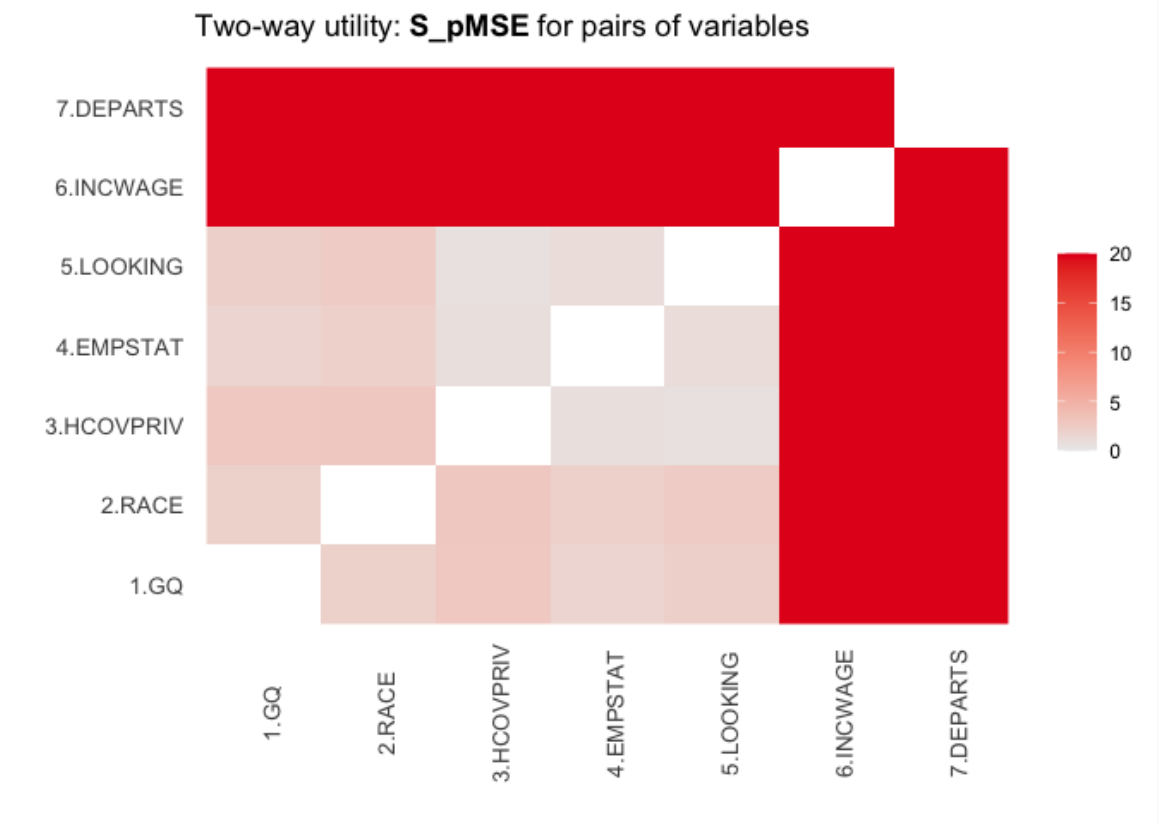
pMSE	S_pMSE
0.2103239	489.4489

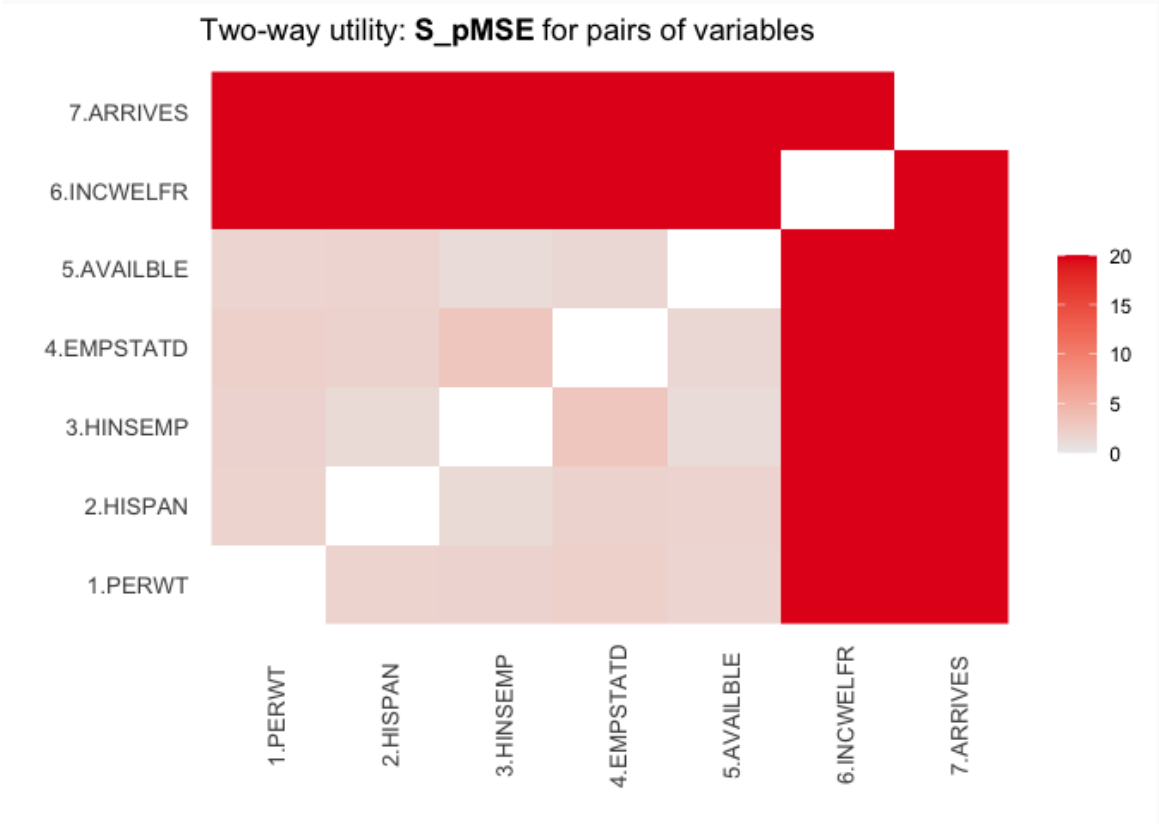
Two-way Tables Comparison of Synthesised Data (R-Package: synthpop) by (S_)pMSE

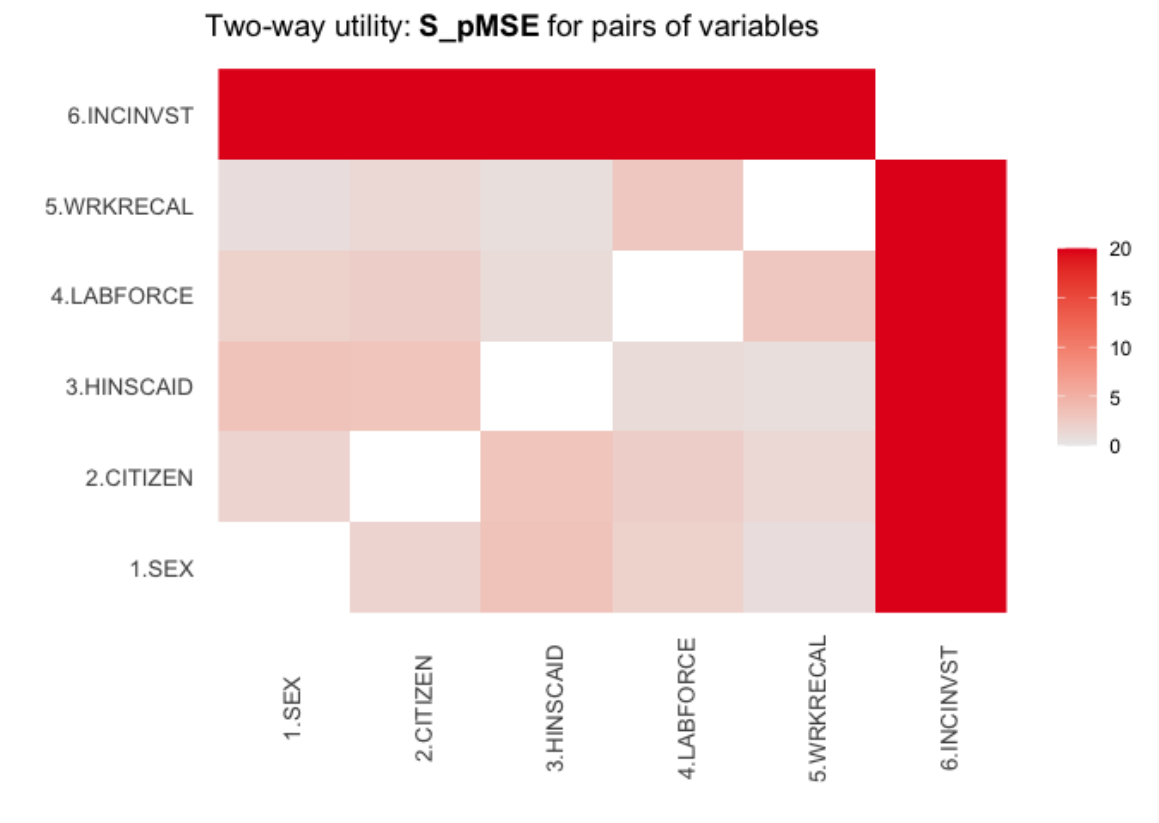
Two-way tables are evaluated based on the original and the synthetic dataset based on S_pMSE (see above). We also present the results for the mean absolute difference in densities (MabsDD) and the Bhattacharyya distance (dBhatt).

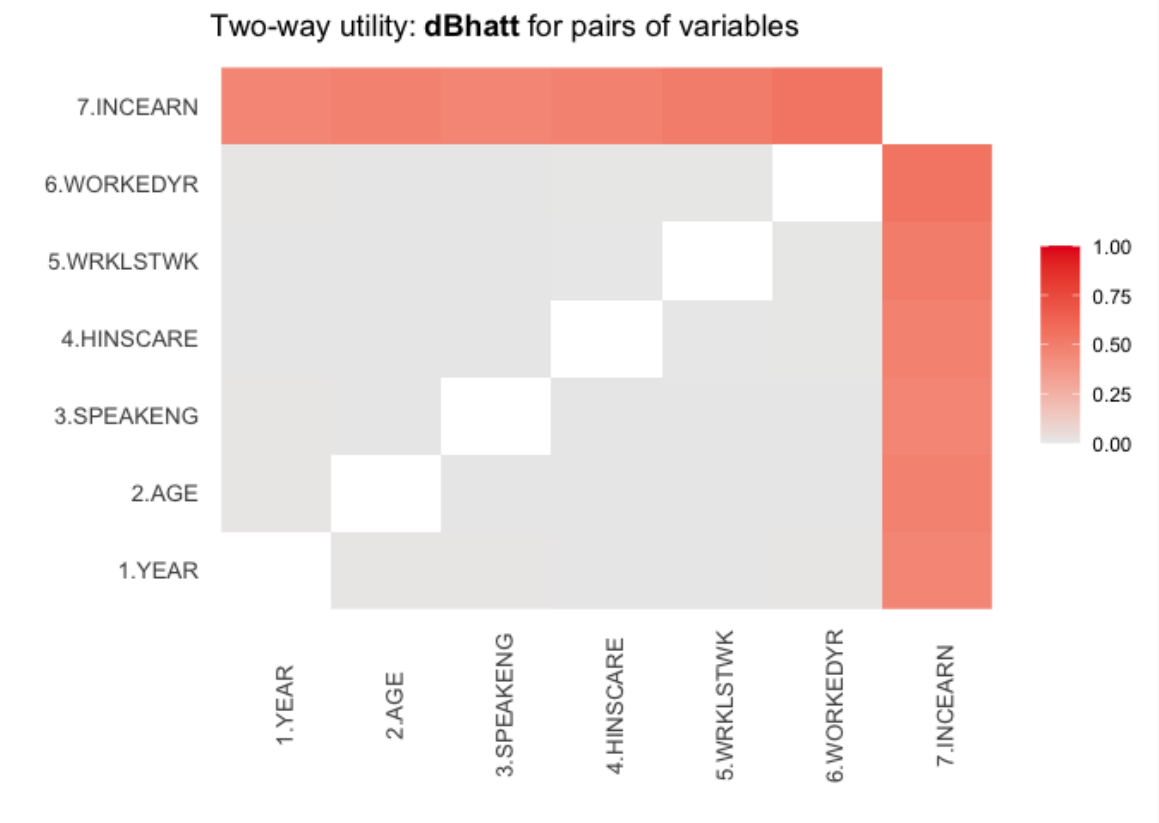

Two-way utility: S_pMSE for pairs of variables
7.INCEARN
6.WORKEDYR
5.WRKLSTWK
4.HINSCARE
3.SPEAKENG
2.AGE
1.YEAR
1.YEAR
2.AGE
3.SPEAKENG
4.HINSCARE
5.WRKLSTWK
6.WORKEDYR
7.INCEARN
20
15
10
5
0

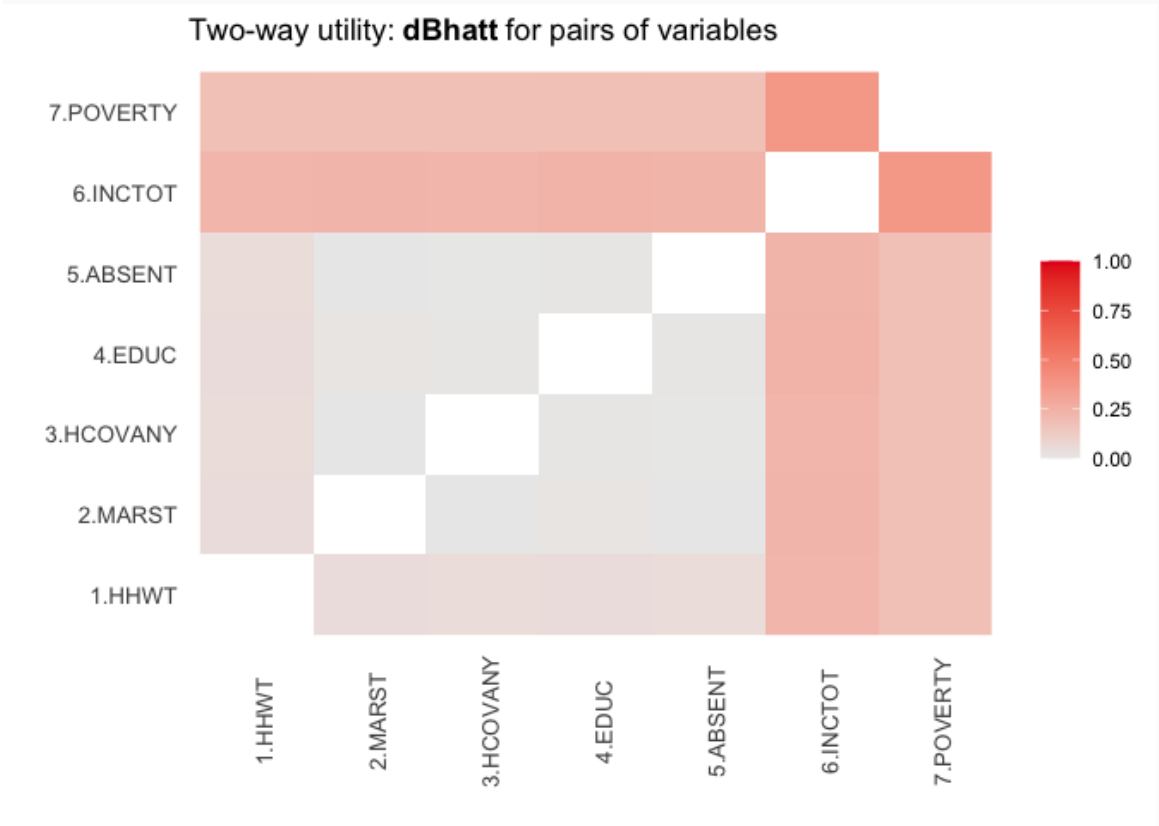

Two-way utility: S_pMSE for pairs of variables
7.POVERTY
6.INCTOT
5.ABSENT
4.EDUC
3.HCOVANY
2.MARST
1.HHWT
1.HHWT
2.MARST
3.HCOVANY
4.EDUC
5.ABSENT
6.INCTOT
7.POVERTY
20
15
10
5
0

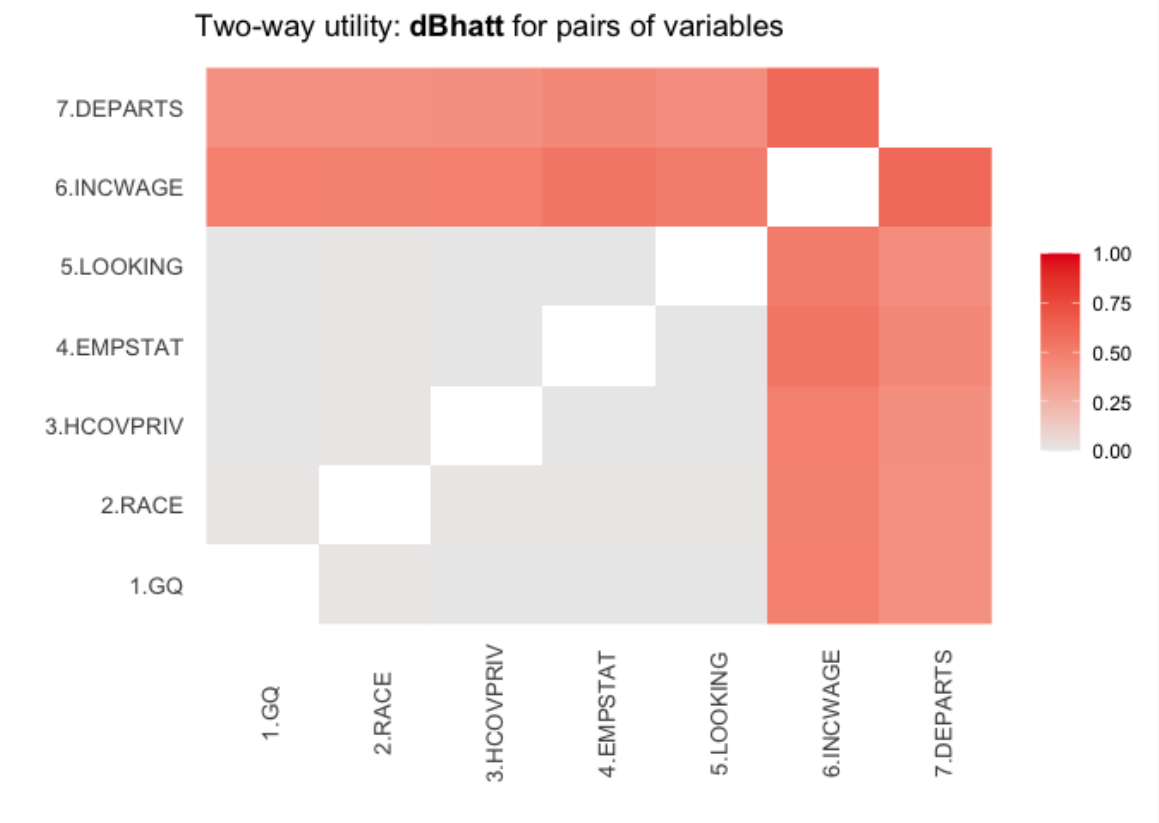

Two-way utility: S_pMSE for pairs of variables
7.DEPARTS
6.INCWAGE
5.LOOKING
4.EMPSTAT
3.HCOVPRIV
2.RACE
1.GQ
1.GQ
2.RACE
3.HCOVPRIV
4.EMPSTAT
5.LOOKING
6.INCWAGE
7.DEPARTS
20
15
10
5
0

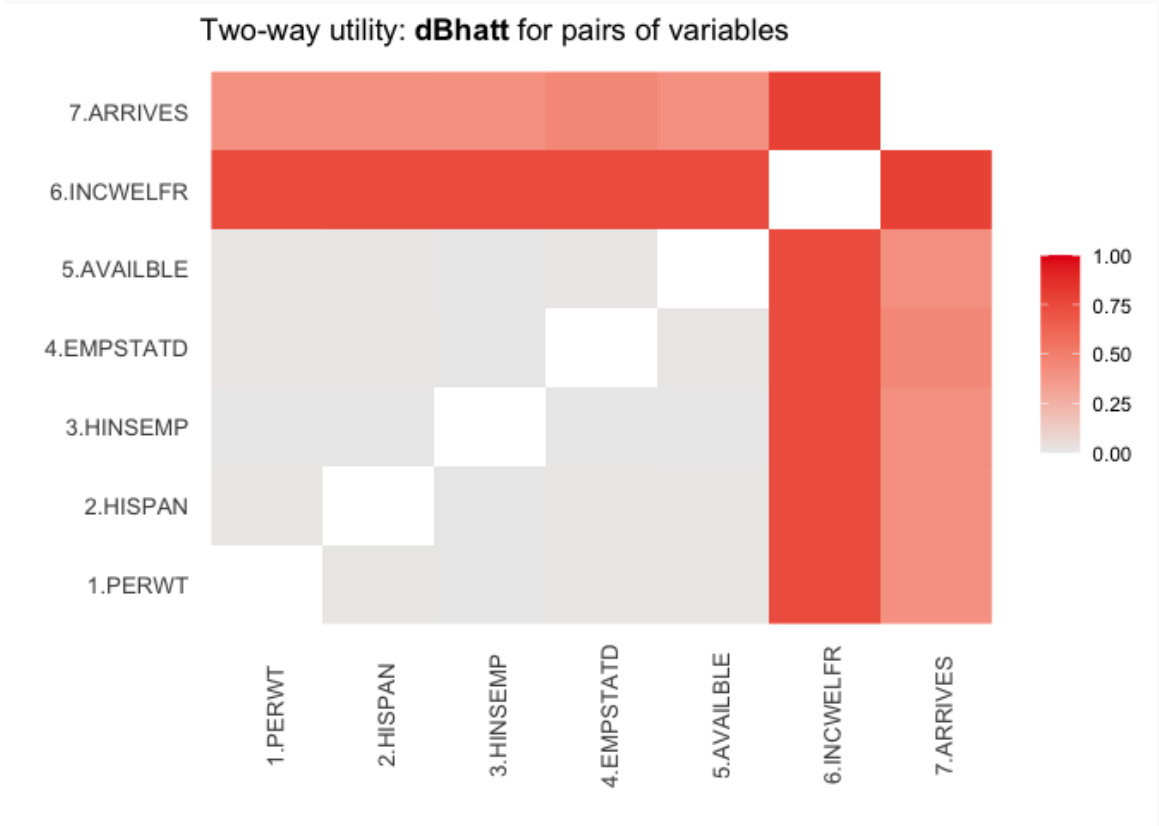

Two-way utility: S_pMSE for pairs of variables
7.ARRIVES
6.INCWELFR
5.AVAILBLE
4.EMPSTATD
3.HINSEMP
2.HISPAN
1.PERWT
1.PERWT
2.HISPAN
3.HINSEMP
4.EMPSTATD
5.AVAILBLE
6.INCWELFR
7.ARRIVES
20
15
10
5
0

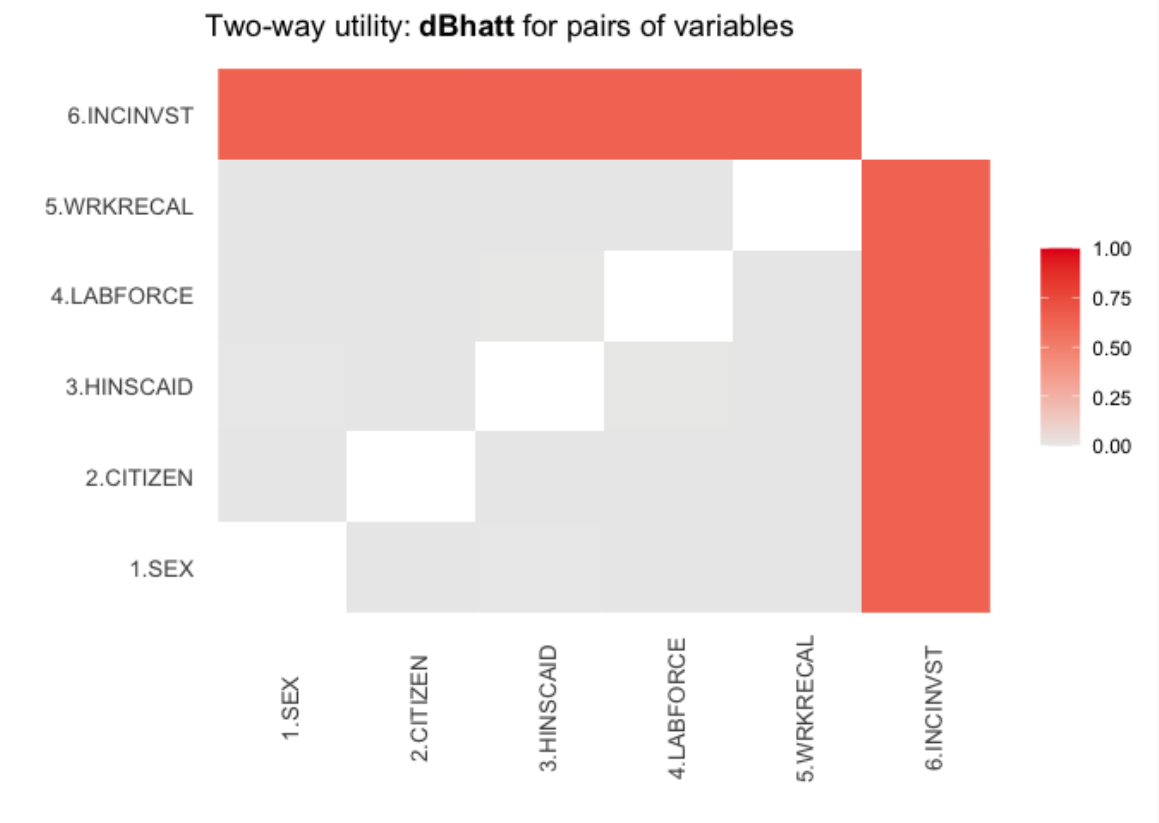

Two-way utility: S_pMSE for pairs of variables
6.INCINVST
5.WRKRECAL
4.LABFORCE
3.HINSCAID
2.CITIZEN
1.SEX
20
15
10
5
0
1.SEX
2.CITIZEN
3.HINSCAID
4.LABFORCE
5.WRKRECAL
6.INCINVST

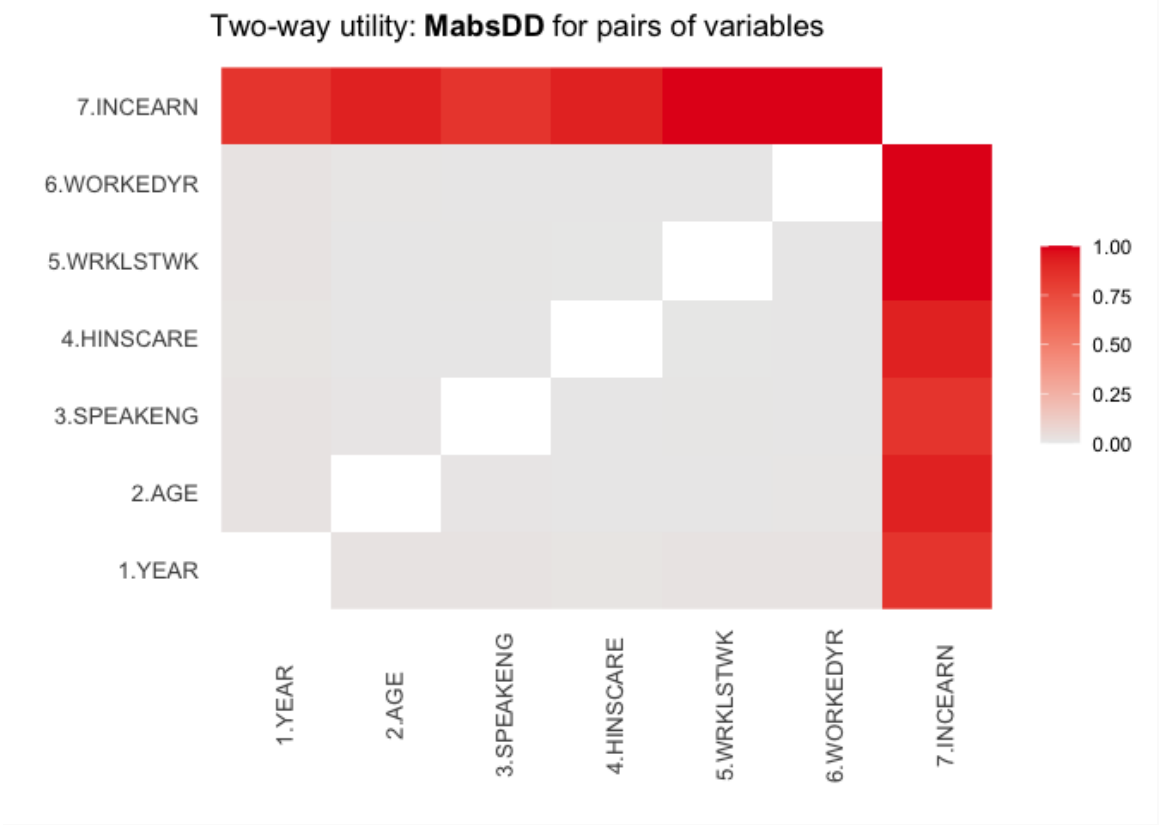

Two-way utility: dBhatt for pairs of variables
7.INCEARN
6.WORKEDYR
5.WRKLSTWK
4.HINSCARE
3.SPEAKENG
2.AGE
1.YEAR
1.YEAR
2.AGE
3.SPEAKENG
4.HINSCARE
5.WRKLSTWK
6.WORKEDYR
7.INCEARN
1.00
0.75
0.50
0.25
0.00

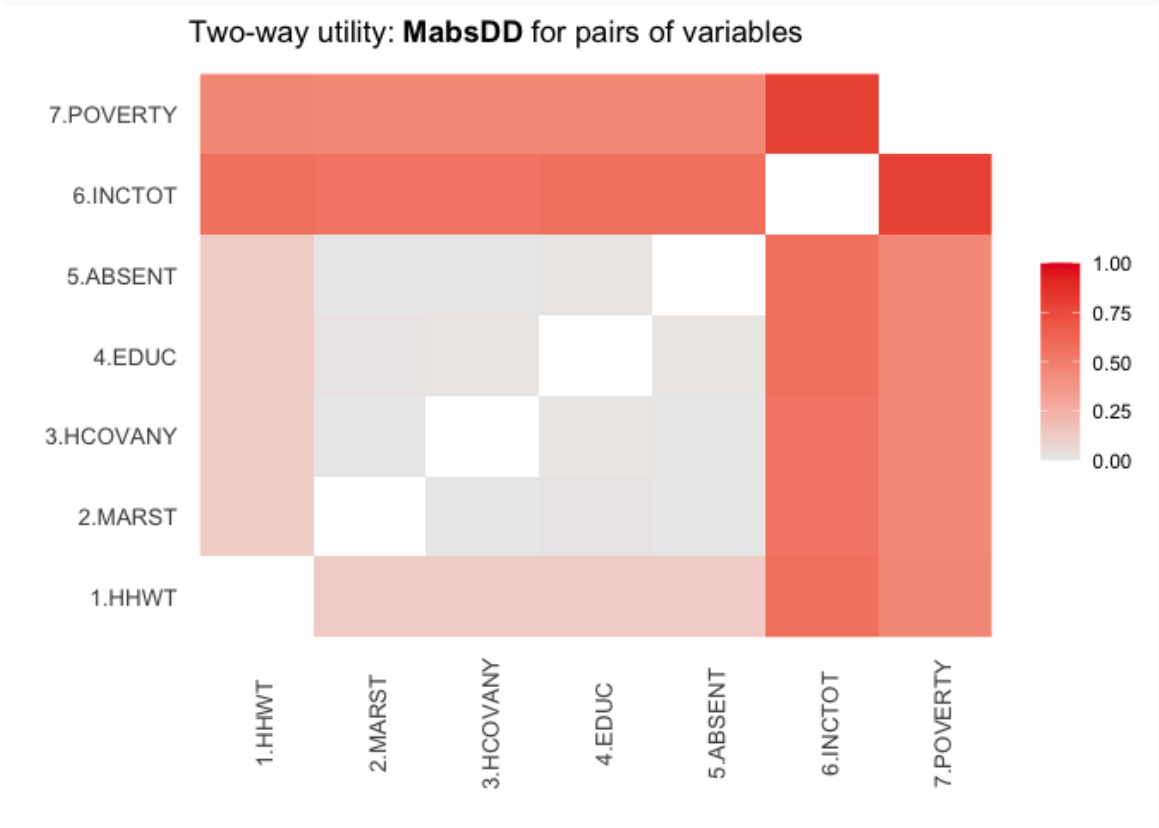

Two-way utility: dBhatt for pairs of variables
7.POVERTY
6.INCTOT
5.ABSENT
4.EDUC
3.HCOVANY
2.MARST
1.HHWT
1.HHWT
2.MARST
3.HCOVANY
4.EDUC
5.ABSENT
6.INCTOT
7.POVERTY
1.00
0.75
0.50
0.25
0.00

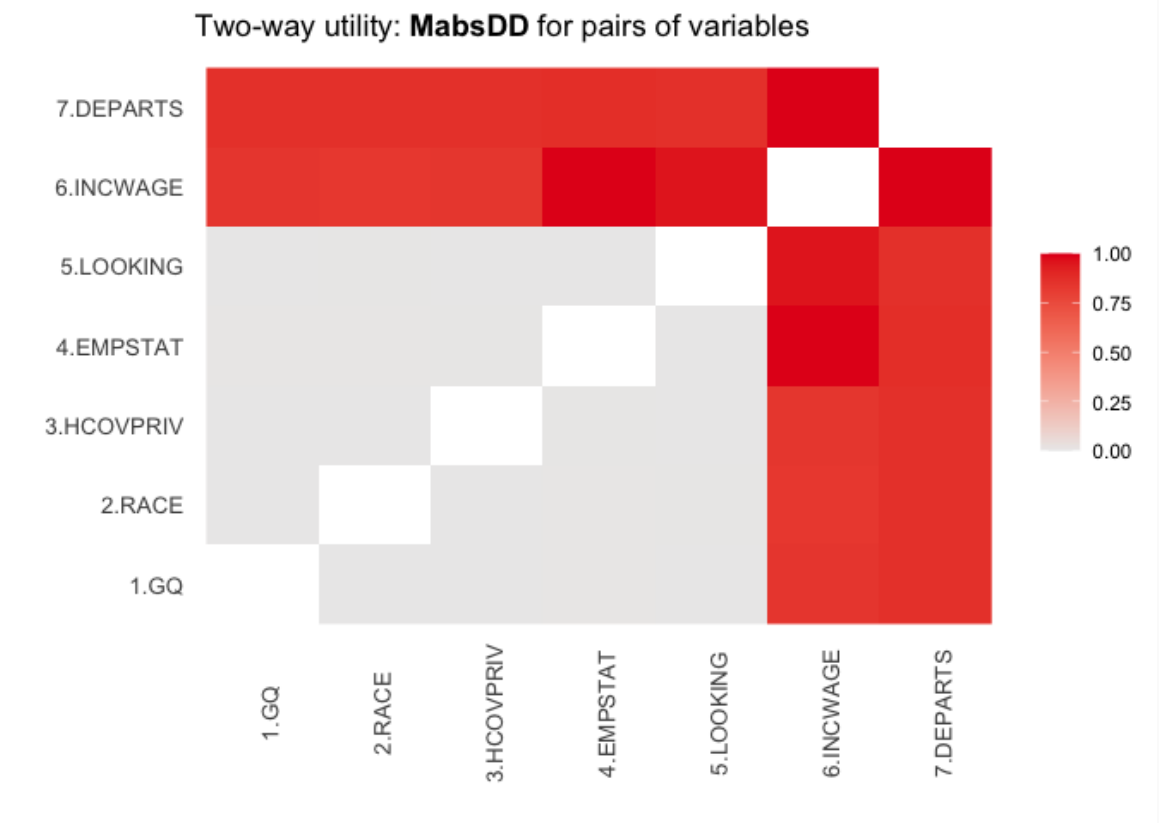

Two-way utility: dBhatt for pairs of variables
7.DEPARTS
6.INCWAGE
5.LOOKING
4.EMPSTAT
3.HCOVPRIV
2.RACE
1.GQ
1.GQ
2.RACE
3.HCOVPRIV
4.EMPSTAT
5.LOOKING
6.INCWAGE
7.DEPARTS
1.00
0.75
0.50
0.25
0.00

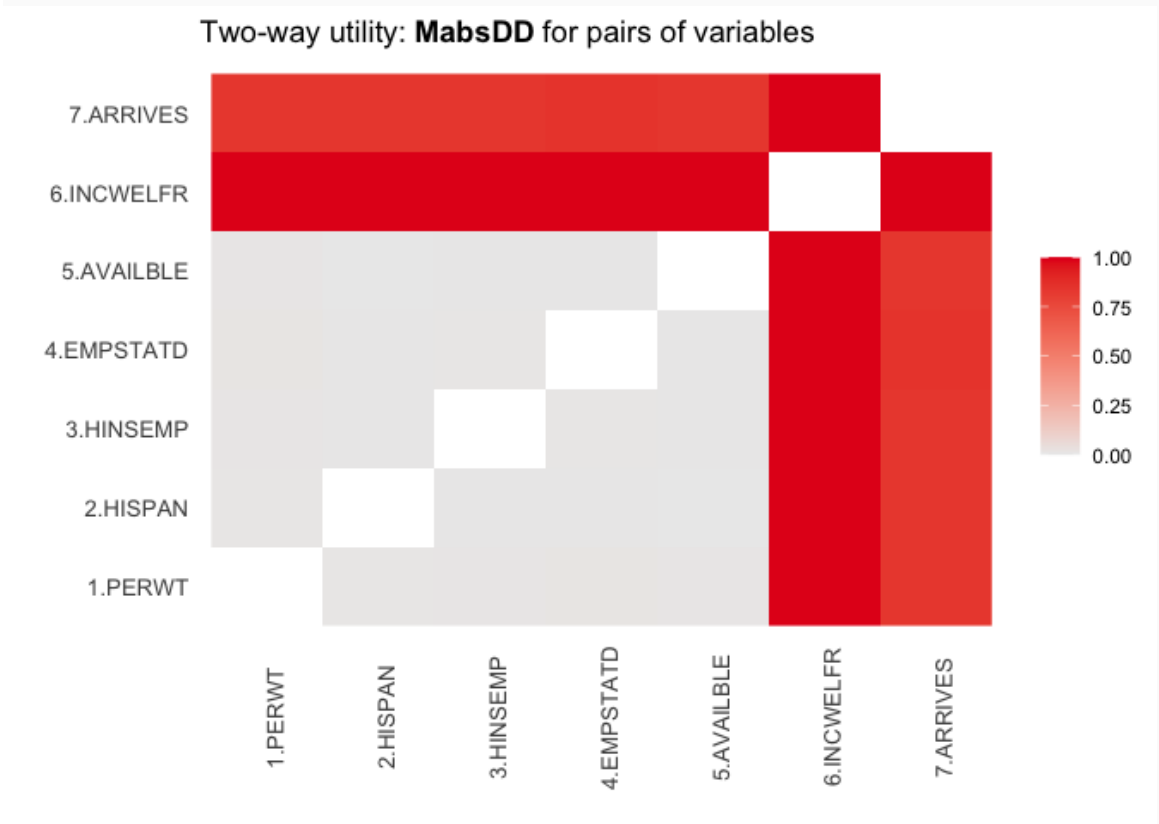

Two-way utility: dBhatt for pairs of variables
7.ARRIVES
6.INCWELFR
5.AVAILBLE
4.EMPSTATD
3.HINSEMP
2.HISPAN
1.PERWT
1.PERWT
2.HISPAN
3.HINSEMP
4.EMPSTATD
5.AVAILBLE
6.INCWELFR
7.ARRIVES
1.00
0.75
0.50
0.25
0.00

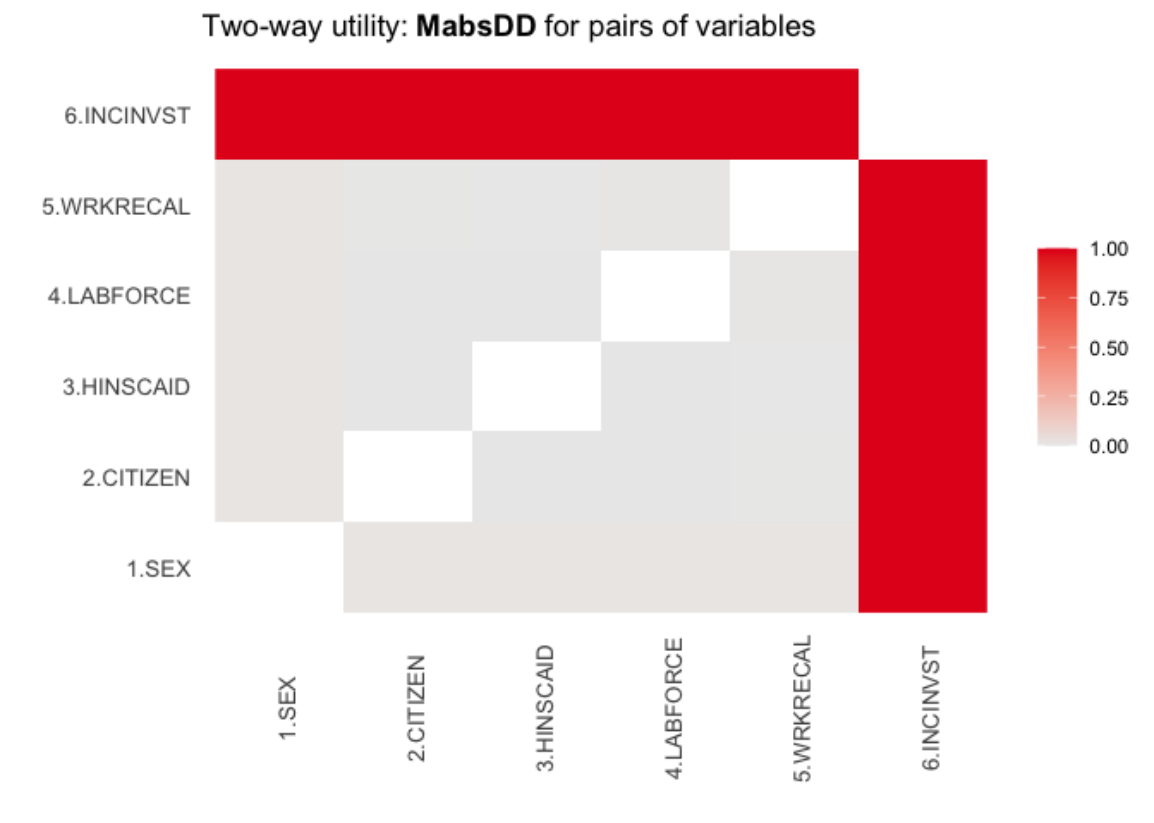

Two-way utility: dBhatt for pairs of variables
6.INCINVST
5.WRKRECAL
4.LABFORCE
3.HINSCAID
2.CITIZEN
1.SEX
1.00
0.75
0.50
0.25
0.00
1.SEX
2.CITIZEN
3.HINSCAID
4.LABFORCE
5.WRKRECAL
6.INCINVST


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


Two-way utility: MabsDD 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
1.00
0.75
0.50
0.25
0.00


Two-way utility: MabsDD for pairs of variables
7.ARRIVES
6.INCWELFR
5.AVAILBLE
4.EMPSTATD
3.HINSEMP
2.HISPAN
1.PERWT
1.PERWT
2.HISPAN
3.HINSEMP
4.EMPSTATD
5.AVAILBLE
6.INCWELFR
7.ARRIVES
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
0.2609665

Individual Distances for Information Loss:

```
##      YEAR      HHWT        GQ     PERWT       SEX       AGE     MARST      RACE
## 0.0000000 0.9180626 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
##    HISPAN   CITIZEN  SPEAKENG   HCOVANY  HCOVPRIV   HINSEMP  HINSCAID  HINSCARE
## 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
##      EDUC   EMPSTAT  EMPSTATD  LABFORCE  WRKLSTWK    ABSENT   LOOKING  AVAILBLE
## 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
##  WRKRECAL  WORKEDYR    INCTOT   INCWAGE  INCWELFR  INCINVST   INCEARN   POVERTY
## 0.0000000 0.0000000 0.9998829 0.9998724 0.9931795 0.9996511 0.9998798 0.9819239
##   DEPARTS   ARRIVES
## 0.9901850 0.9902226
```

Tuning and Optimizations

Results for IPSO when **HHWT** would not be confidential.

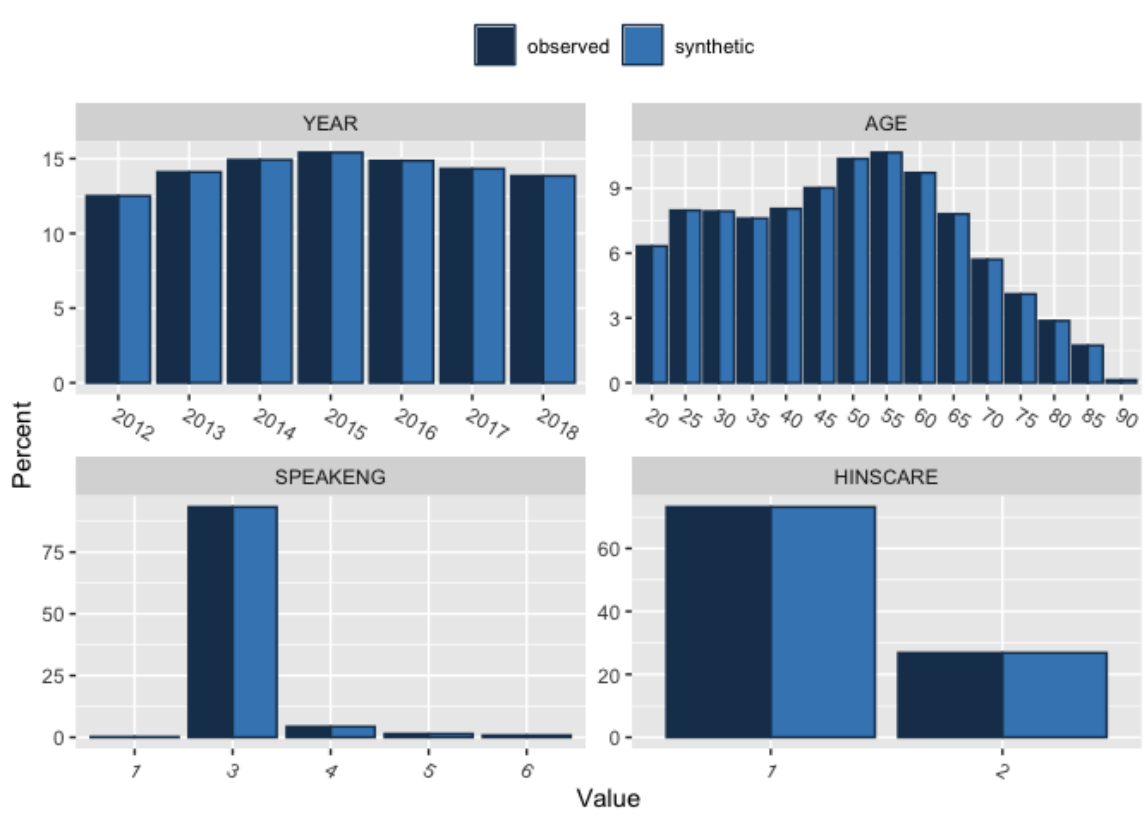
Replication.Uniques	Number.Replications	Percentage.Replications

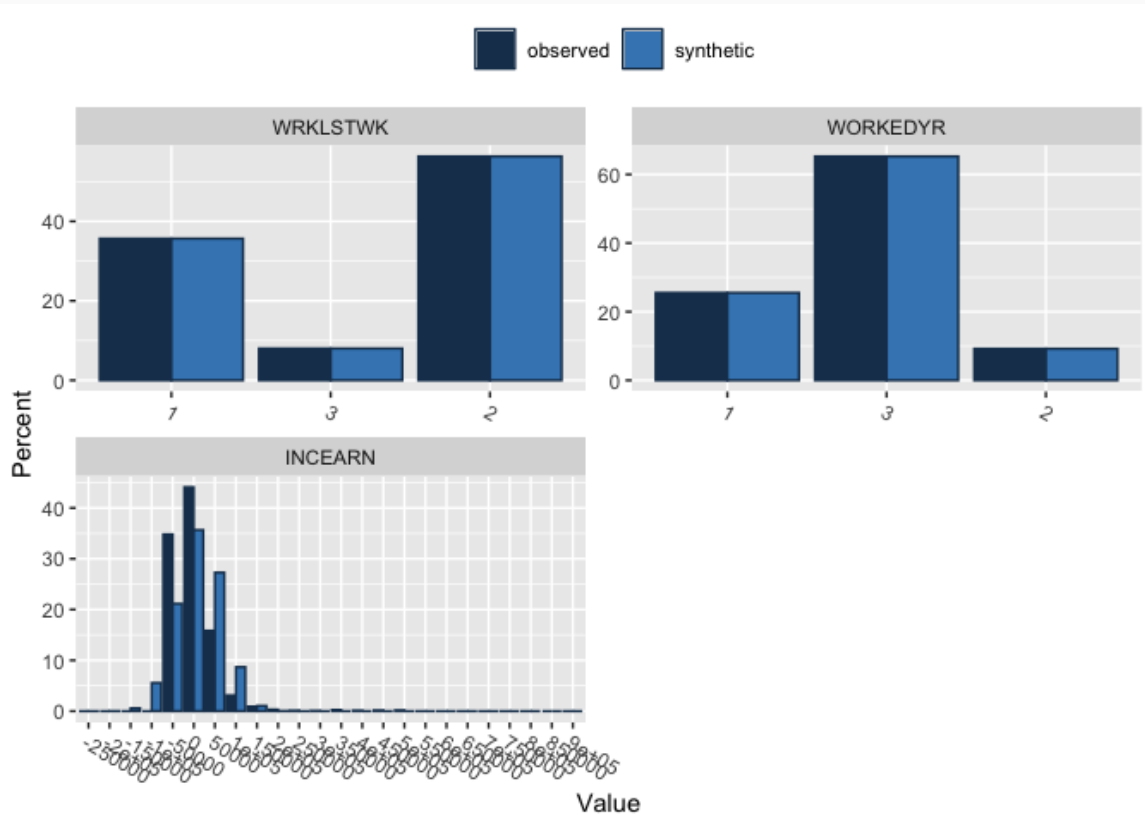
Replication.Uniques	Number.Replications	Percentage.Replications
1001862	177075	17.10537

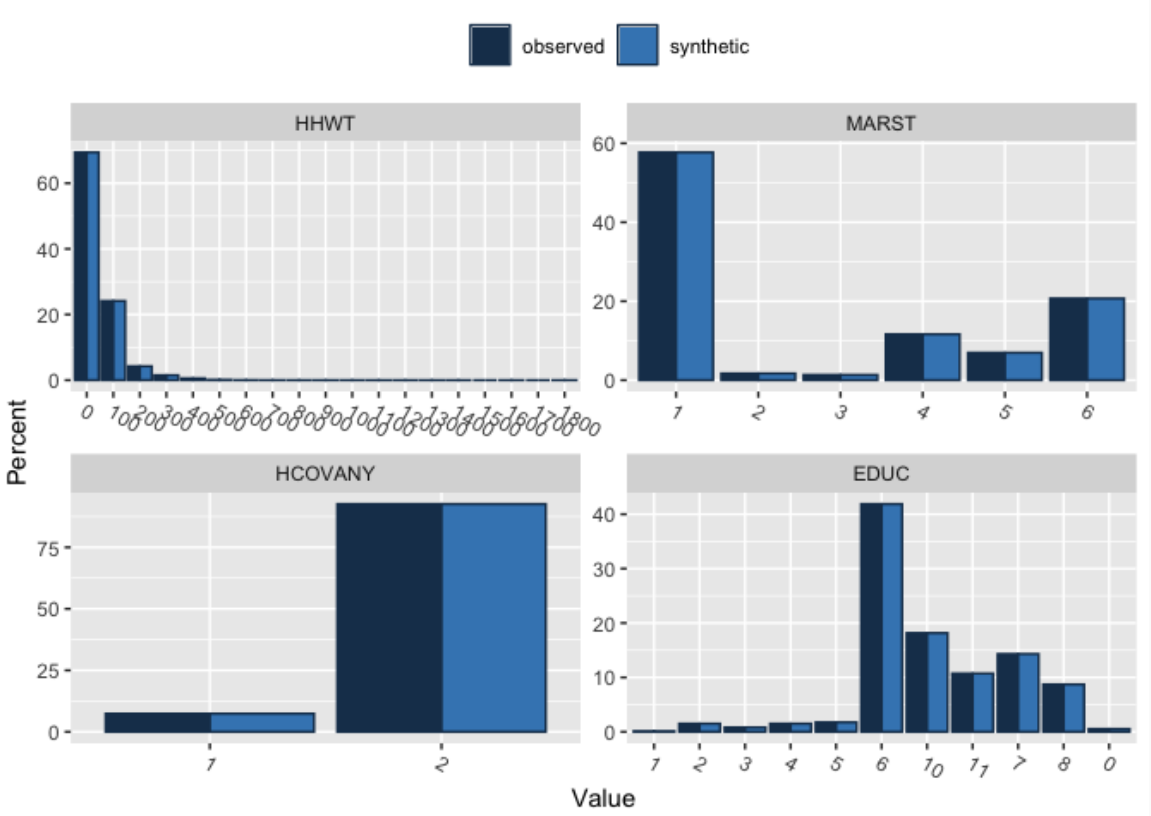
Perceived Disclosure Risk (R-Package: synthpop)

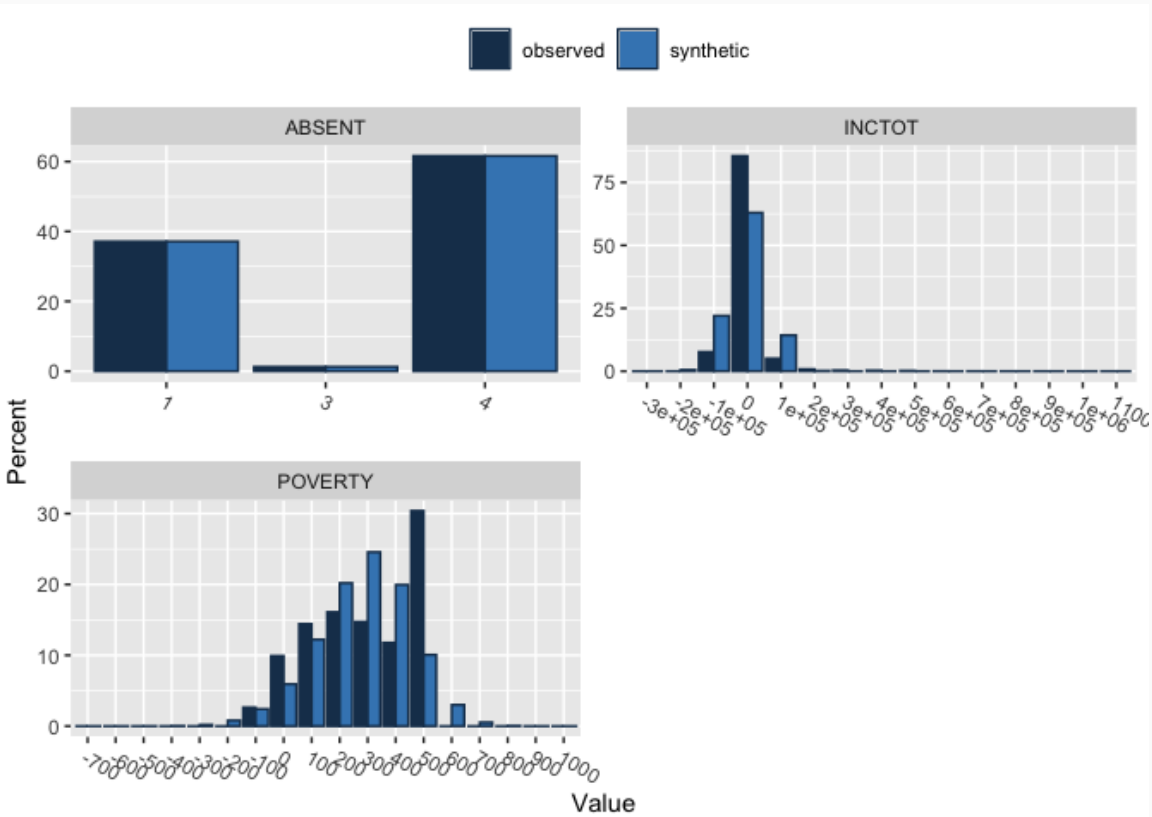
Metric	Number.Uniques	Number.Replications	Percentage.Replications
Perceived Risk	1001862	1001862	96.77947

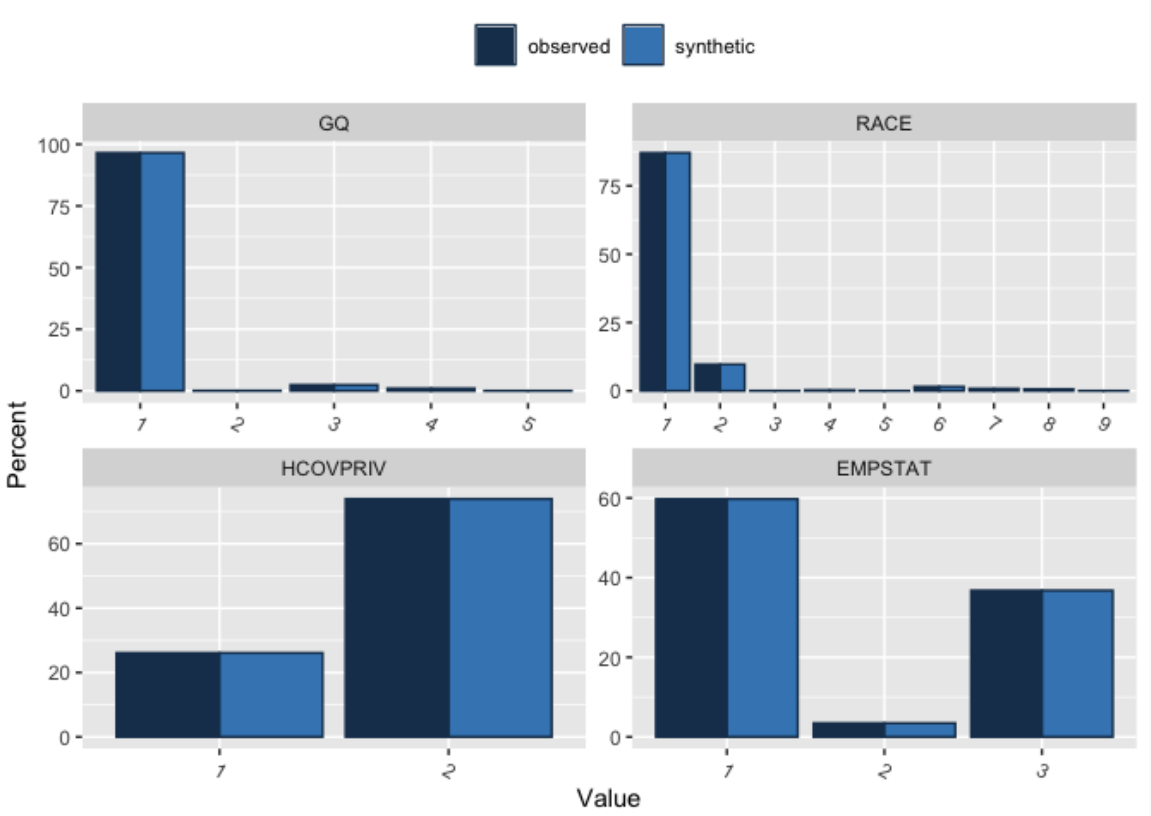
Graphical Comparison for Margins (R-Package: synthpop)

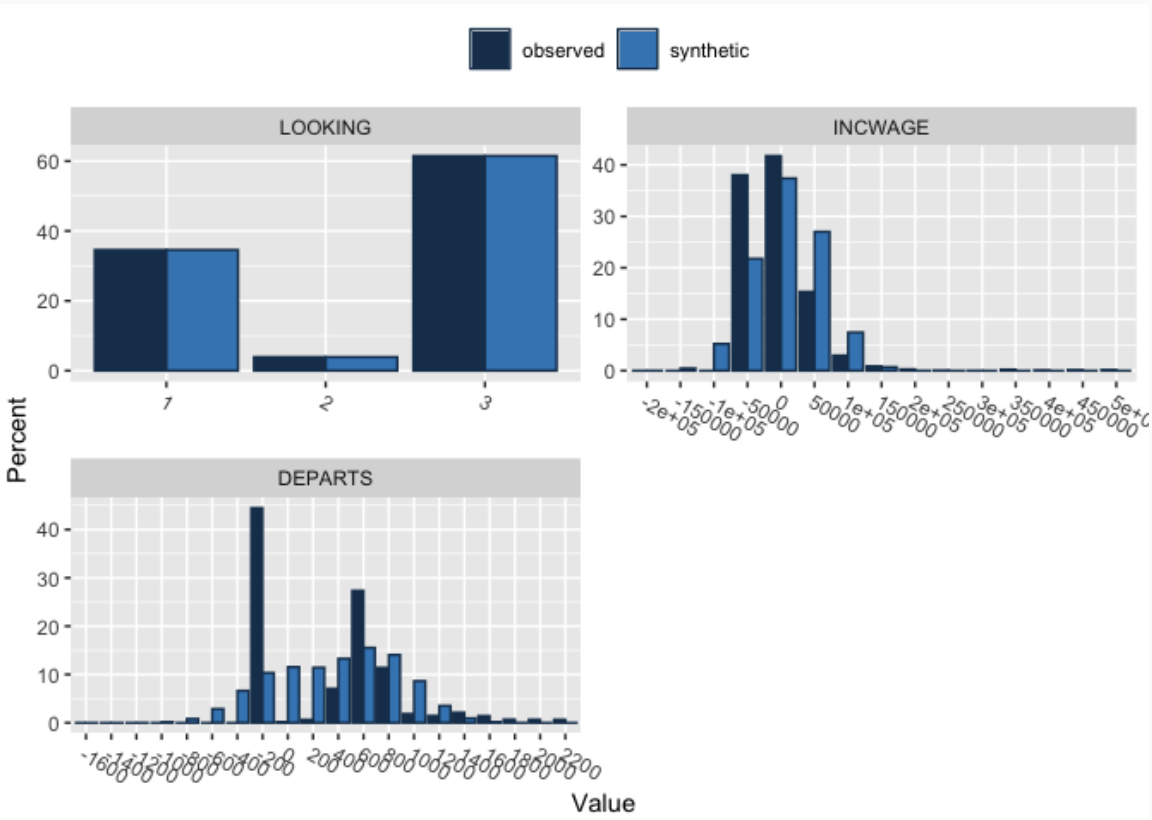

observed
synthetic
YEAR
AGE
SPEAKENG
HINSCARE
Percent
Value

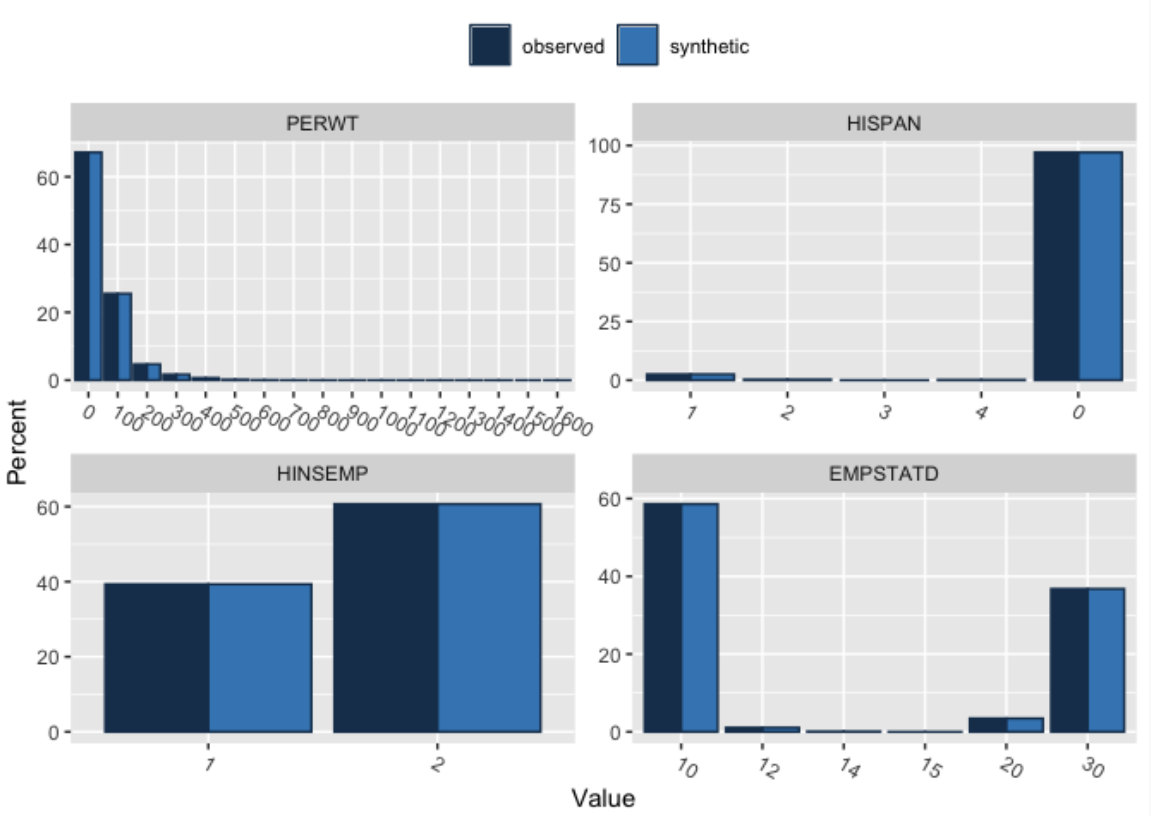

observed
synthetic
WRKLSTWK
WORKEDYR
INCEARN
Percent
Value

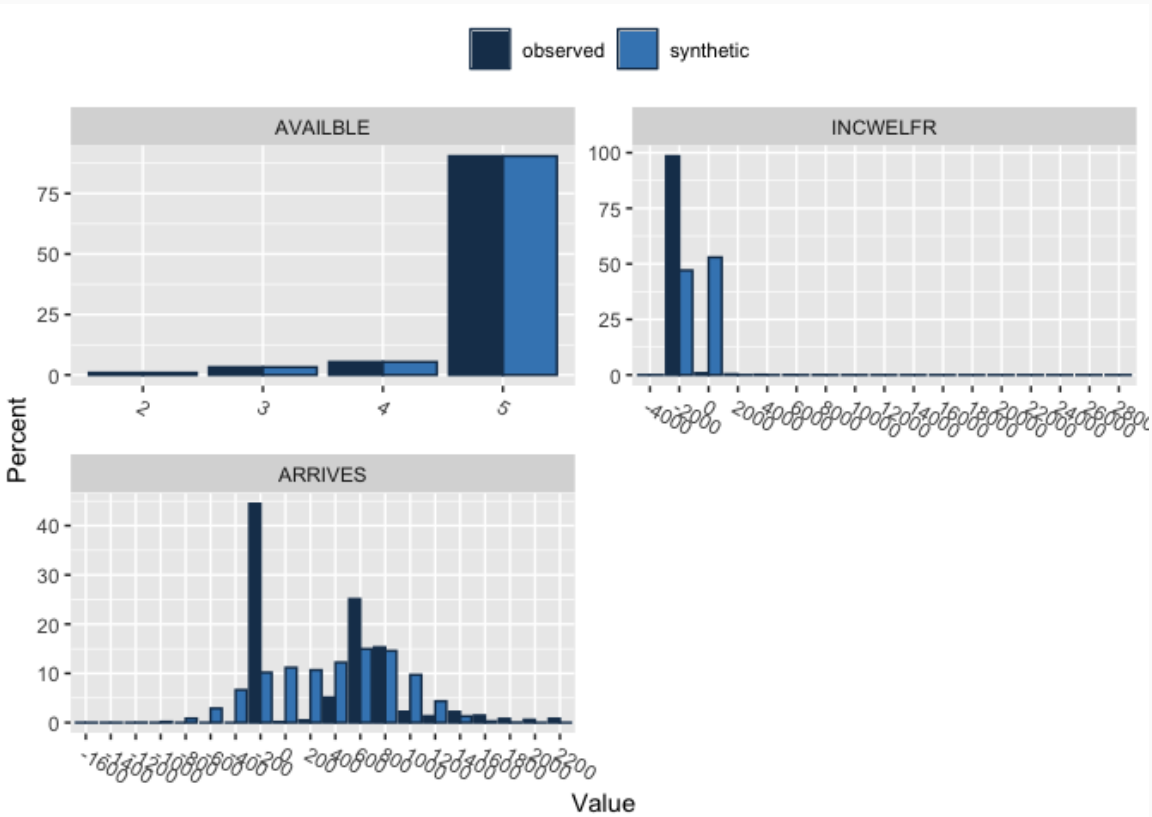

observed
synthetic
HHWT
MARST
HCOVANY
EDUC
Percent
Value

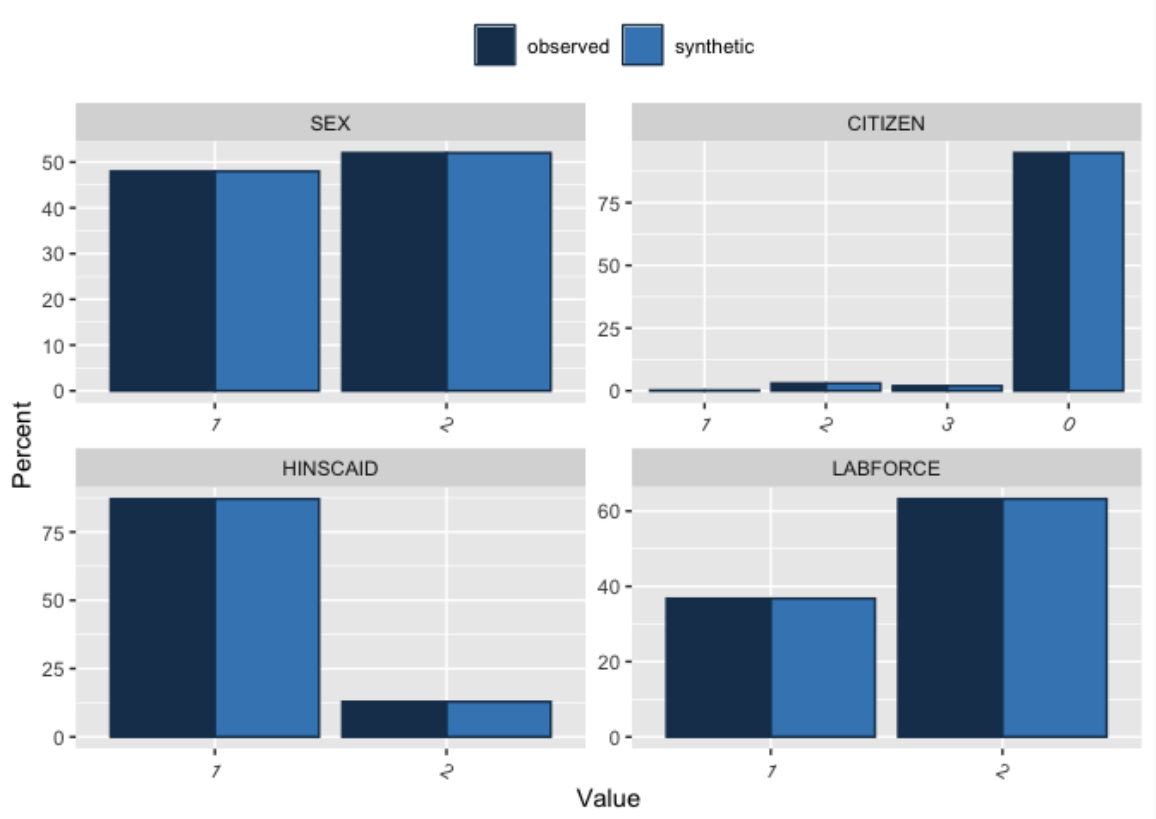

observed
synthetic
ABSENT
INCTOT
POVERTY
Percent
Value

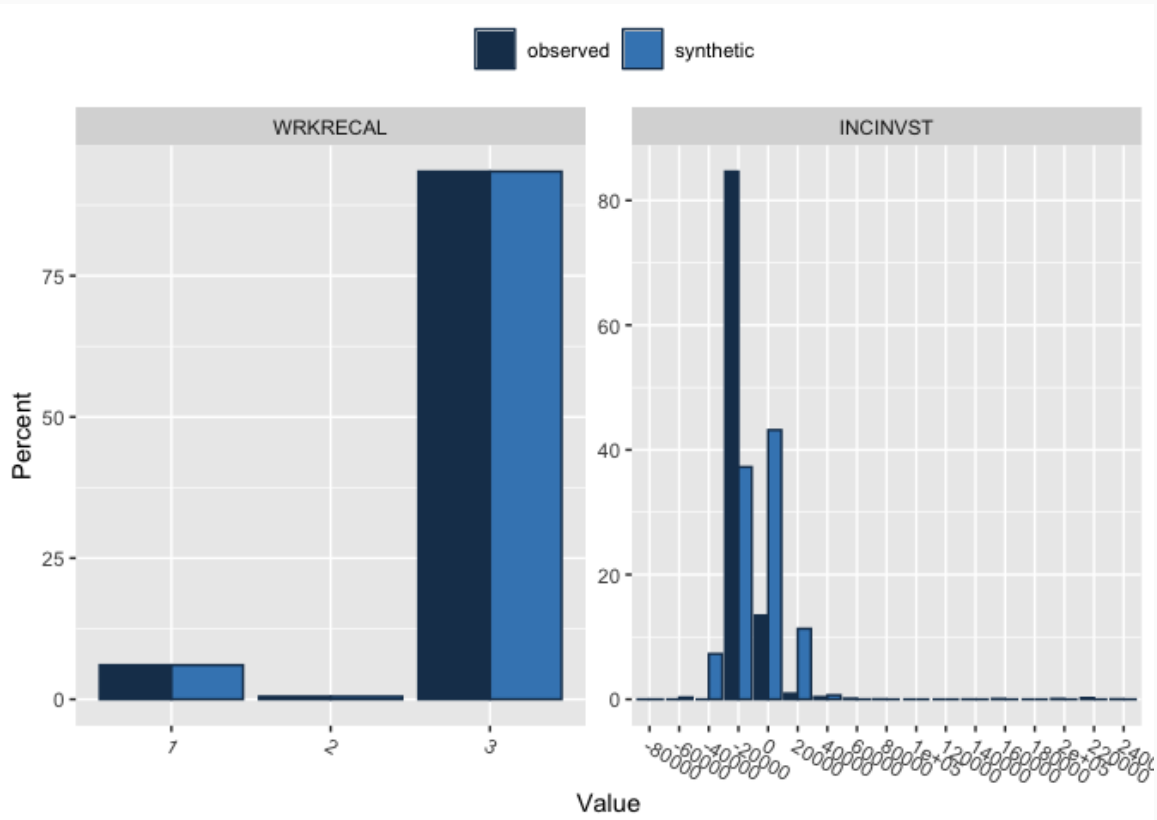

observed
synthetic
GQ
RACE
HCOVPRIV
EMPSTAT
Percent
Value


observed
synthetic
LOOKING
INCWAGE
DEPARTS
Percent
Value

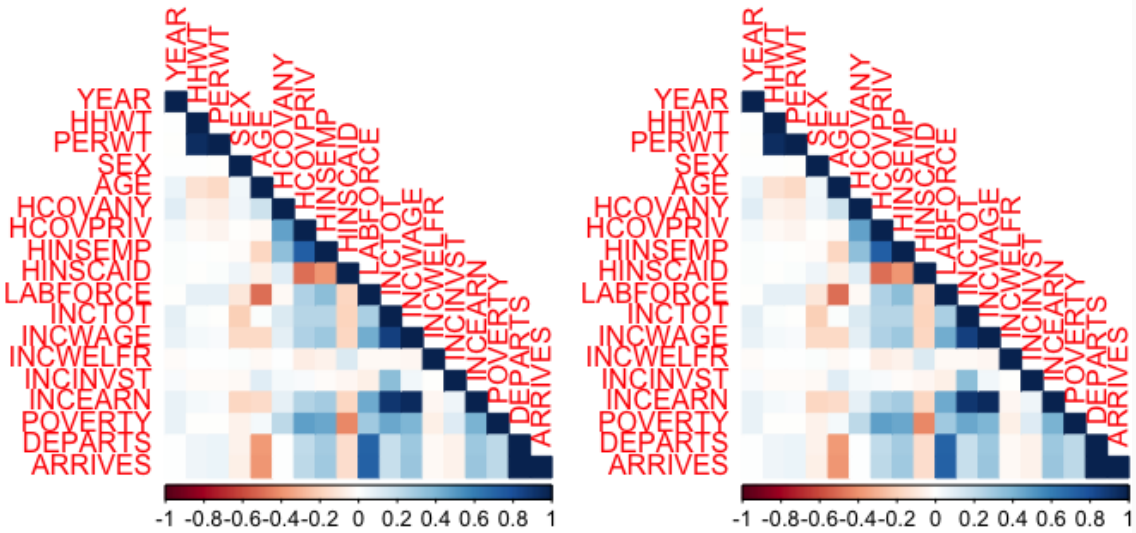

observed
synthetic
PERWT
HISPAN
HINSEMP
EMPSTATD
Percent
Value


observed
synthetic
AVAILBLE
INCWELFR
ARRIVES
Percent
Value





Correlation Plots for Graphical Comparison of Pearson Correlation



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

	pMSE	S_pMSE	df
YEAR	0.0000000	0.0	6
AGE	0.0000000	0.0	4
SPEAKENG	0.0000000	0.0	4
HINSCARE	0.0000000	0.0	1
WRKLSTWK	0.0000000	0.0	2
WORKEDYR	0.0000000	0.0	2
INCEARN	0.0736524	304980.3	4

	pMSE	S_pMSE
	0.1679304	188.8739

	pMSE	S_pMSE	df
HHWT	0.0000000	0.00	4
MARST	0.0000000	0.00	5
HCOVANY	0.0000000	0.00	1
EDUC	0.0000000	0.00	10
ABSENT	0.0000000	0.00	2

	pMSE	S_pMSE	df
INCTOT	0.0235894	97678.90	4
POVERTY	0.0151864	62884.01	4

pMSE	S_pMSE
0.1418994	102.9702

	pMSE	S_pMSE	df
GQ	0.0000000	0.0	4
RACE	0.0000000	0.0	8
HCOVPRIV	0.0000000	0.0	1
EMPSTAT	0.0000000	0.0	2
LOOKING	0.0000000	0.0	2
INCWAGE	0.0757311	313587.6	4
DEPARTS	0.0580187	240243.9	4

pMSE	S_pMSE
0.2056397	286.6373

	pMSE	S_pMSE	df
PERWT	0.0000000	0.0	4
HISPAN	0.0000000	0.0	4
HINSEMP	0.0000000	0.0	1
EMPSTATD	0.0000000	0.0	5
AVAILBLE	0.0000000	0.0	3
INCWELFR	0.1812949	1000942.1	3
ARRIVES	0.0566870	234729.6	4

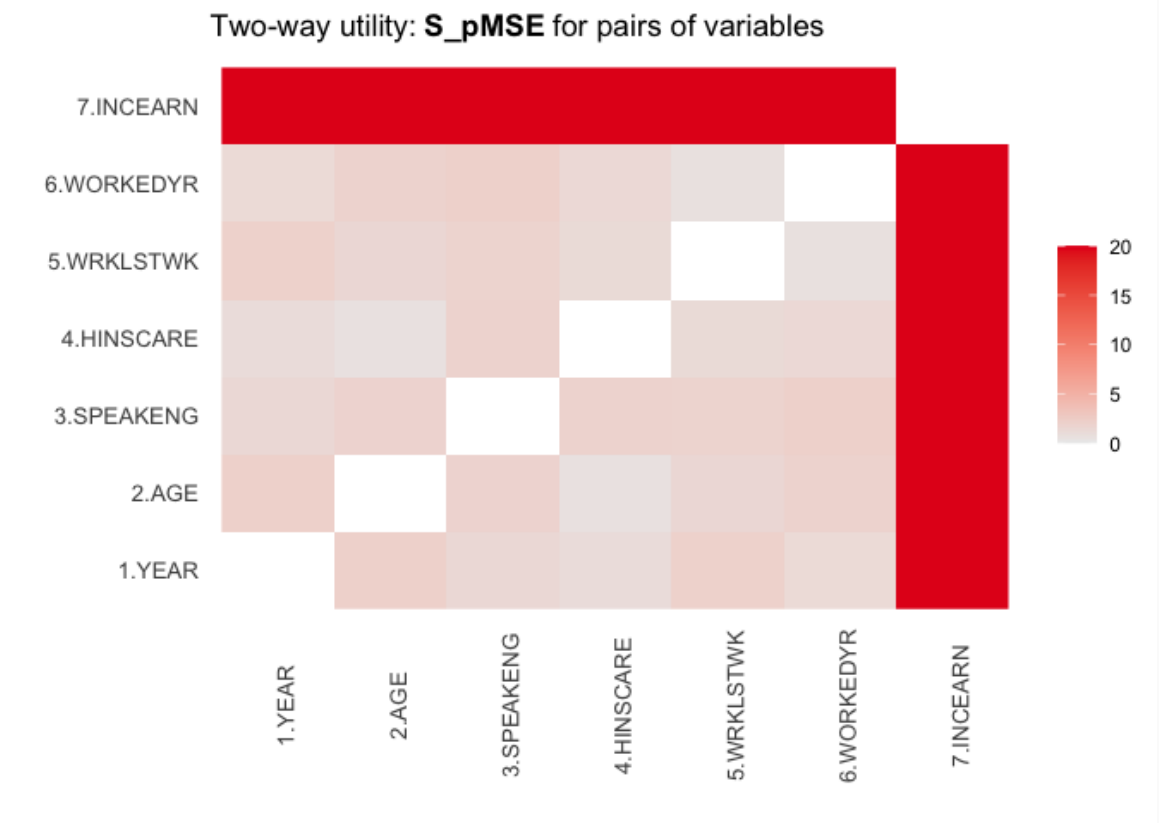
pMSE	S_pMSE
0.2485476	250.5728

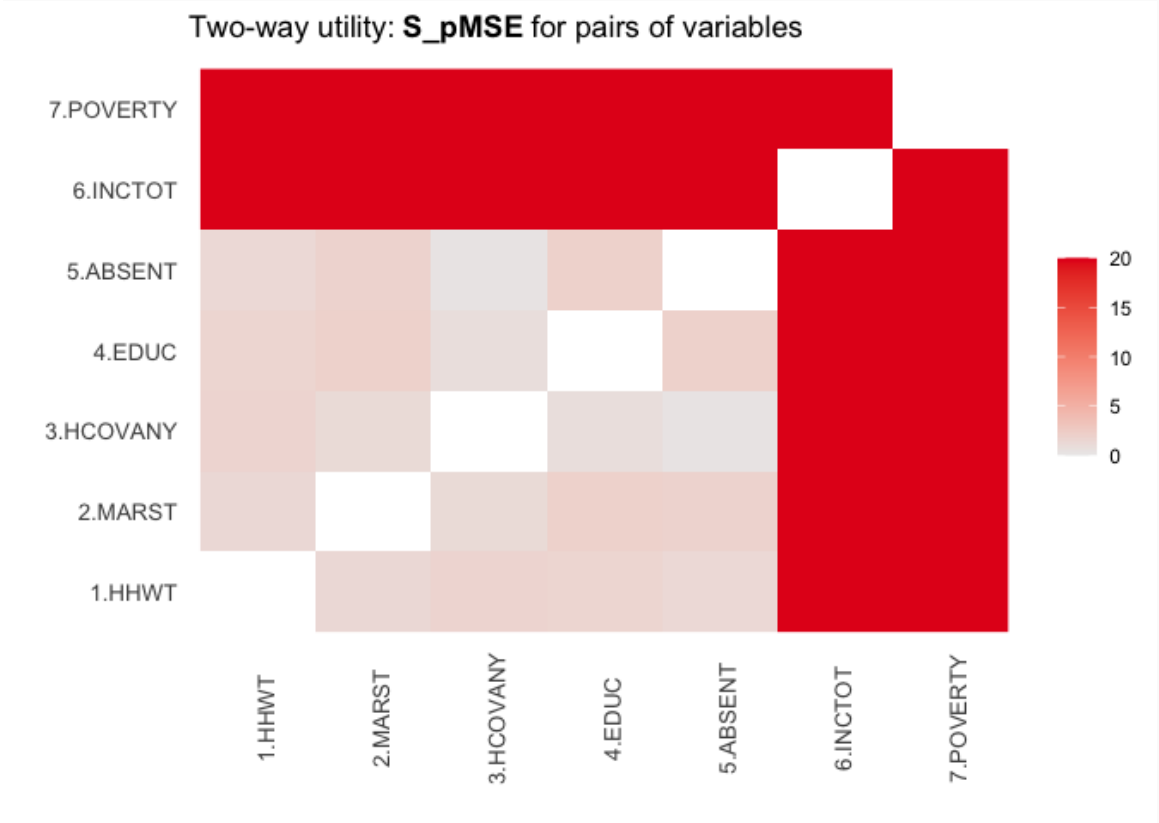
	pMSE	S_pMSE	df
SEX	0.0000000	0.0	1
CITIZEN	0.0000000	0.0	3
HINSCAID	0.0000000	0.0	1
LABFORCE	0.0000000	0.0	1

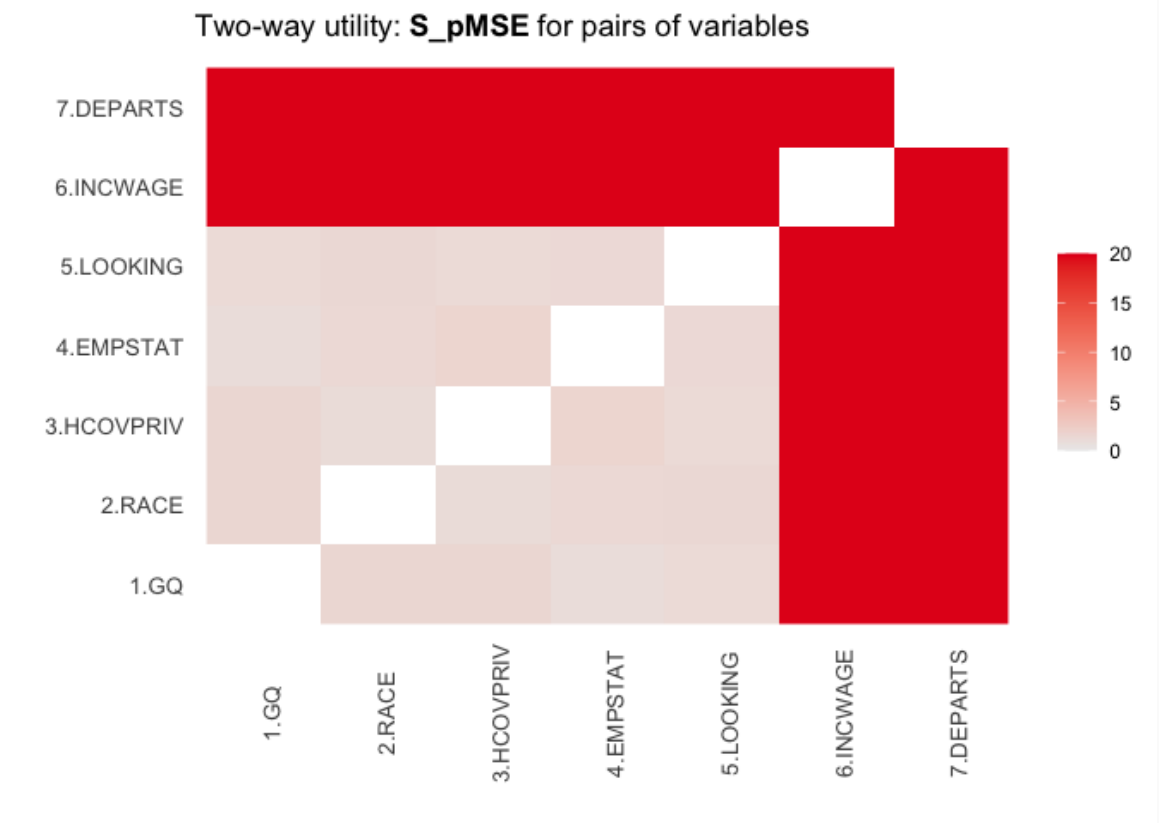
	pMSE	S_pMSE	df
WRKRECAL	0.0000000	0.0	2
INCINVST	0.1483913	819278.8	3

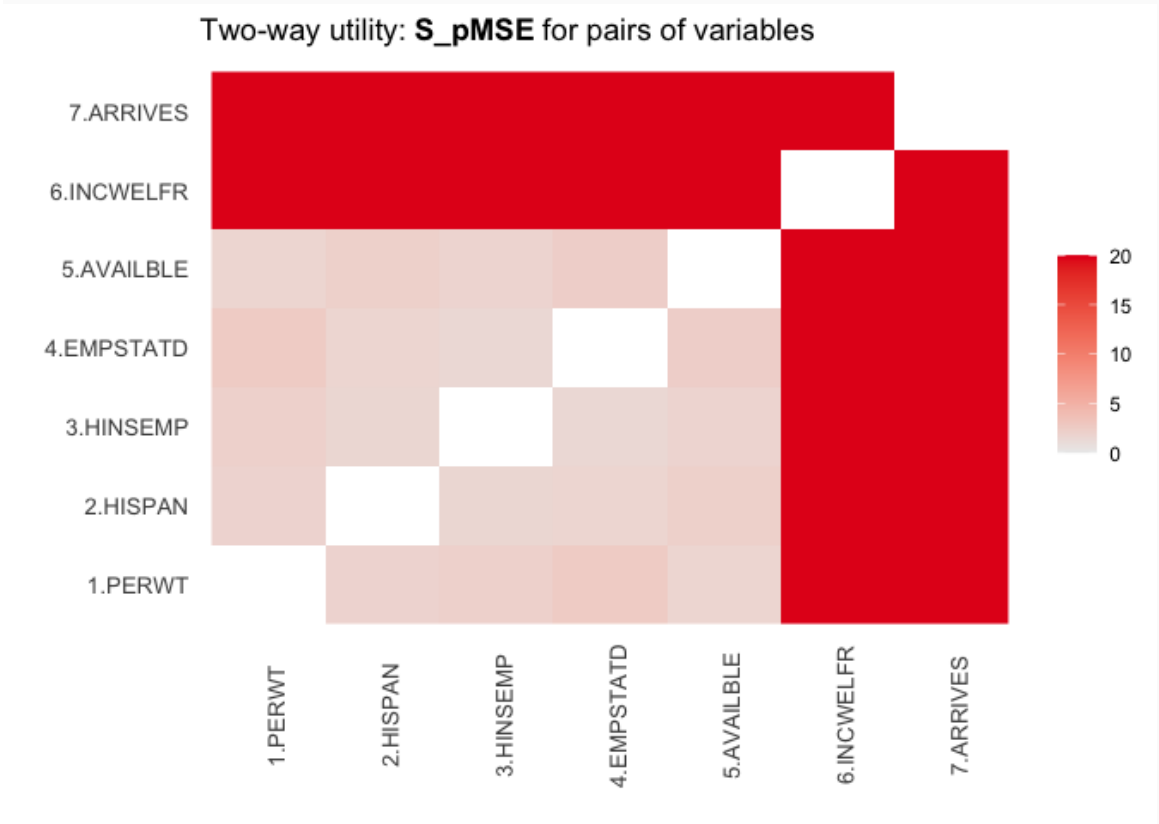
pMSE	S_pMSE
0.2109031	599.3116

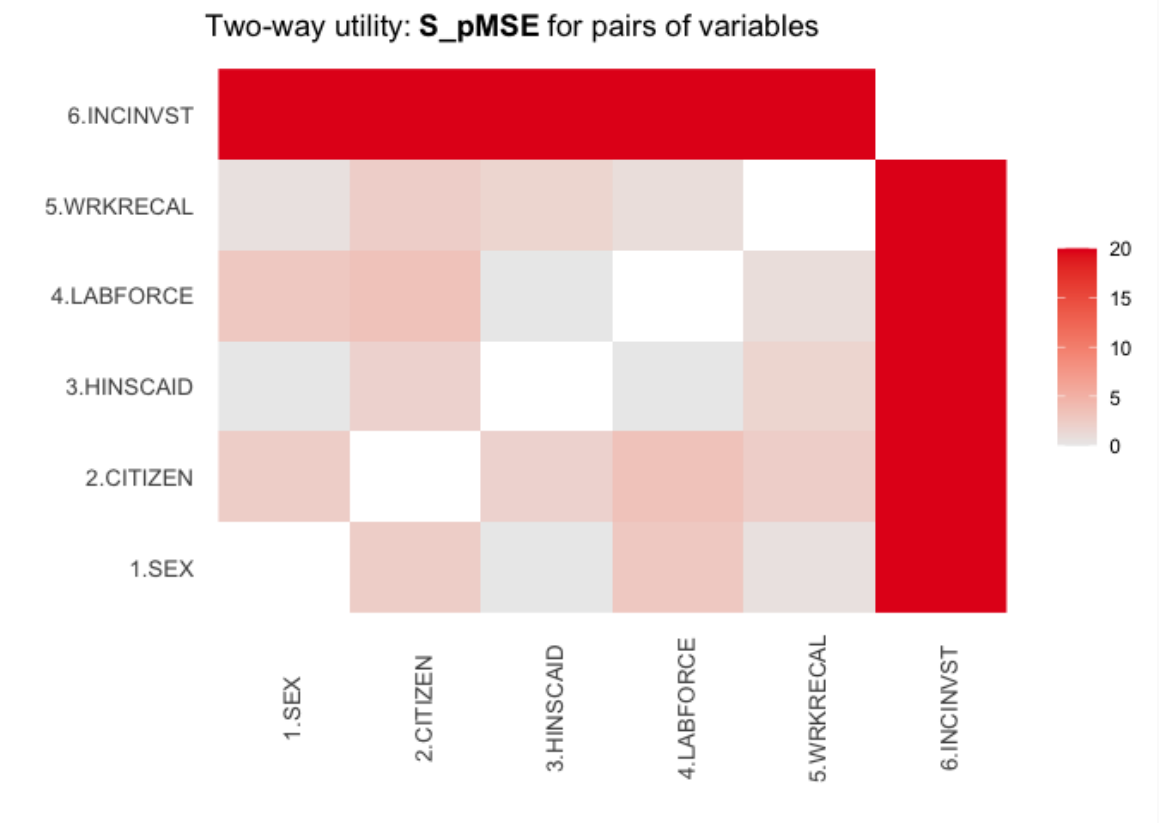
Two-way Tables Comparison of Synthesised Data (R-Package: synthpop) by (S_)pMSE

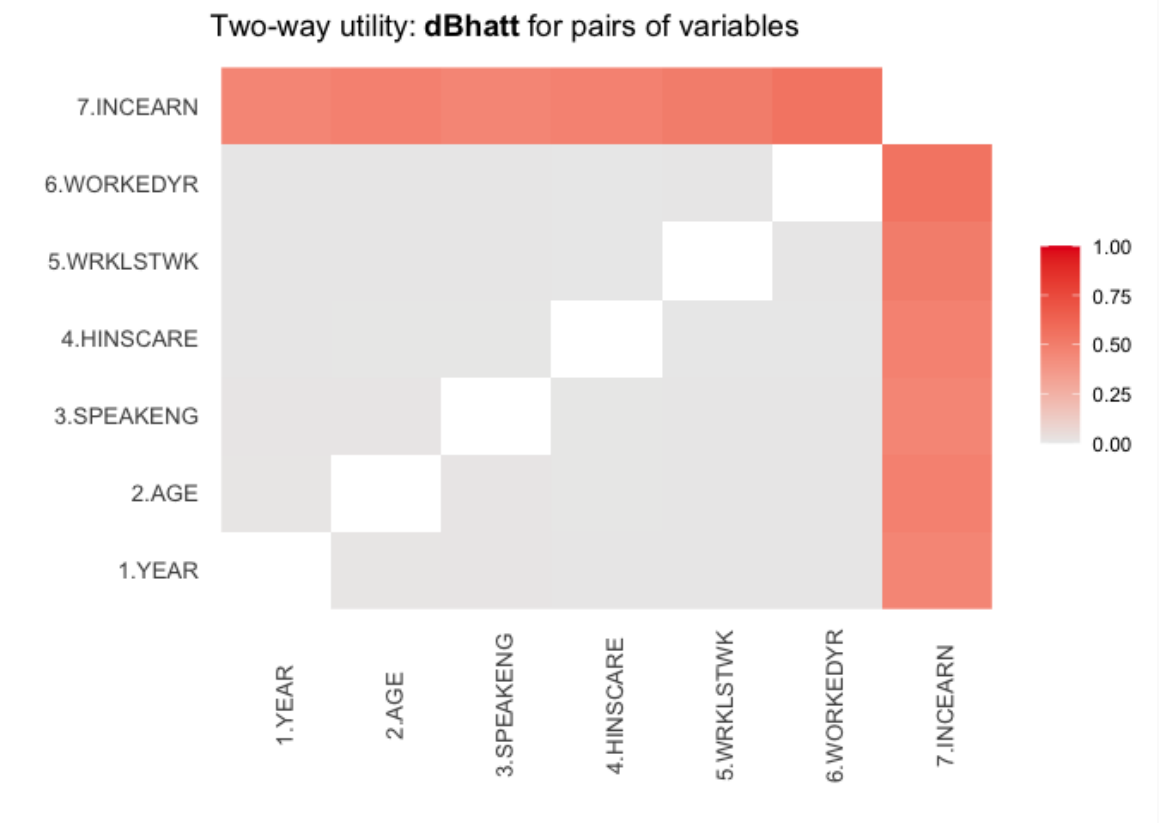

Two-way utility: S_pMSE for pairs of variables
7.INCEARN
6.WORKEDYR
5.WRKLSTWK
4.HINSCARE
3.SPEAKENG
2.AGE
1.YEAR
1.YEAR
2.AGE
3.SPEAKENG
4.HINSCARE
5.WRKLSTWK
6.WORKEDYR
7.INCEARN
20
15
10
5
0

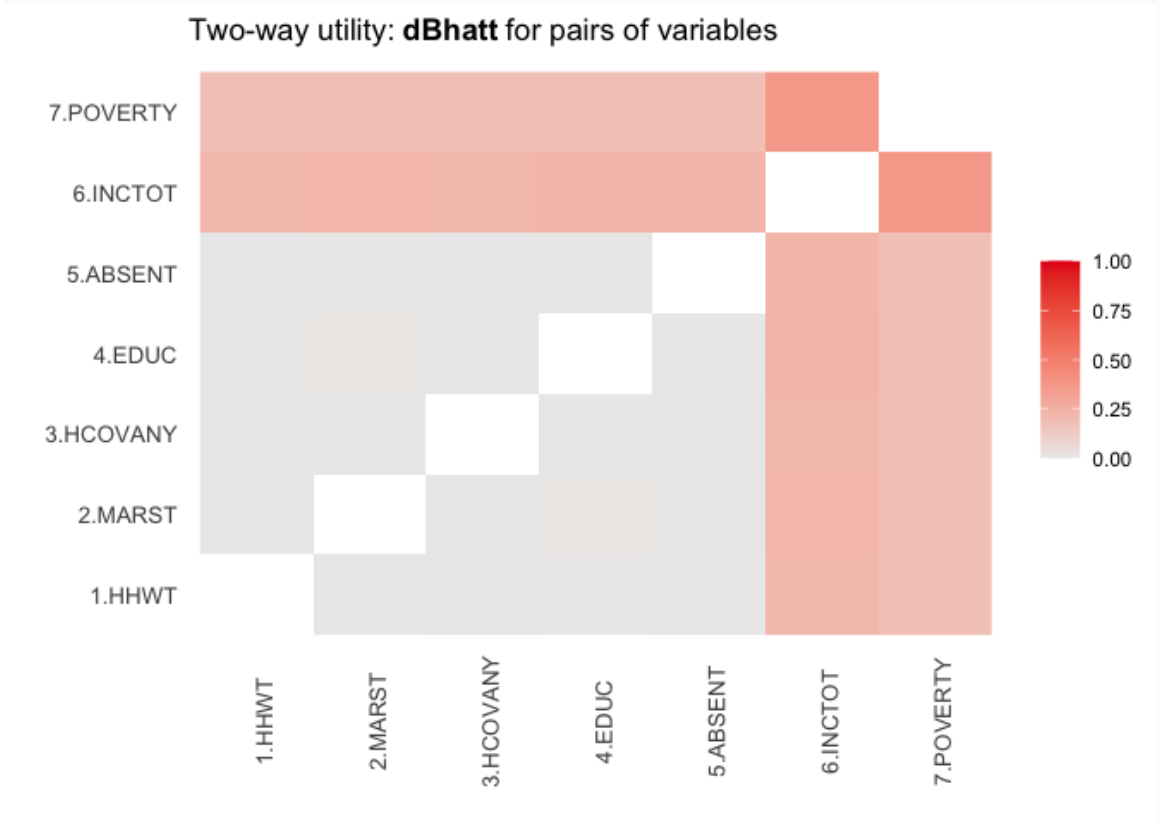

Two-way utility: S_pMSE for pairs of variables
7.POVERTY
6.INCTOT
5.ABSENT
4.EDUC
3.HCOVANY
2.MARST
1.HHWT
1.HHWT
2.MARST
3.HCOVANY
4.EDUC
5.ABSENT
6.INCTOT
7.POVERTY
20
15
10
5
0

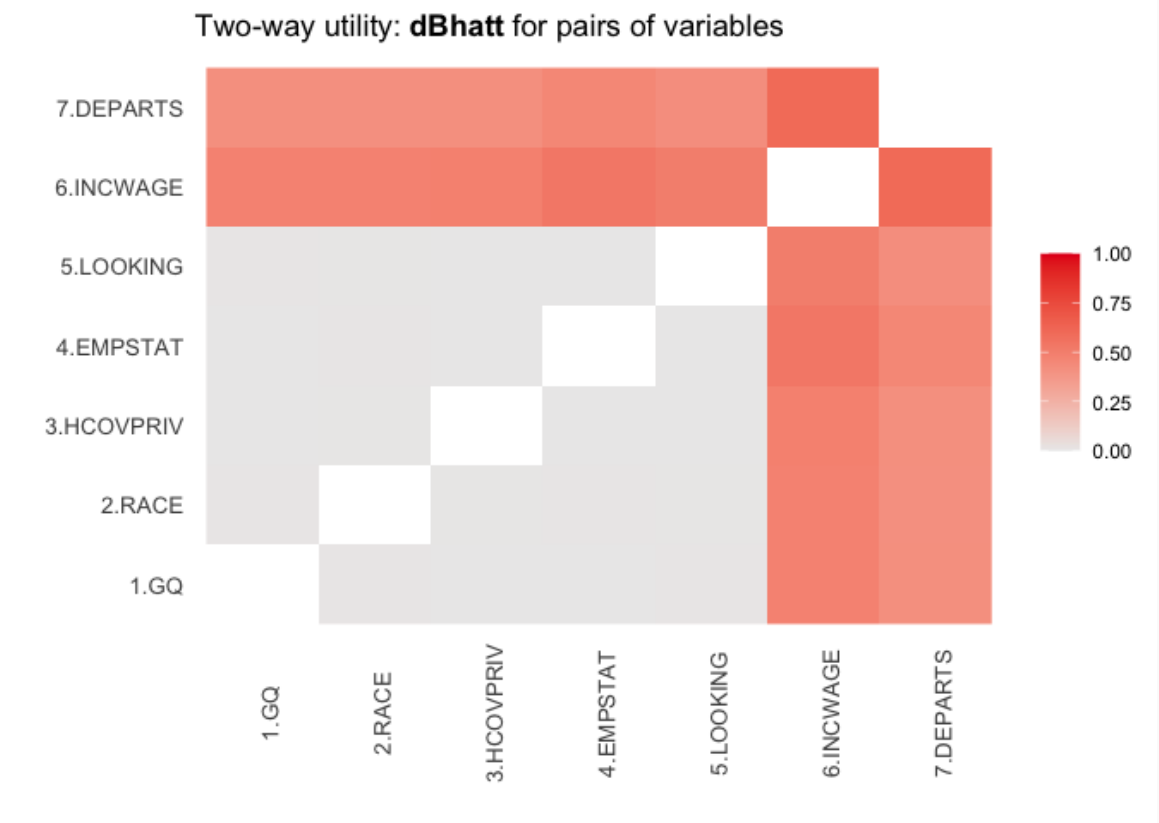

Two-way utility: S_pMSE for pairs of variables
7.DEPARTS
6.INCWAGE
5.LOOKING
4.EMPSTAT
3.HCOVPRIV
2.RACE
1.GQ
1.GQ
2.RACE
3.HCOVPRIV
4.EMPSTAT
5.LOOKING
6.INCWAGE
7.DEPARTS
20
15
10
5
0

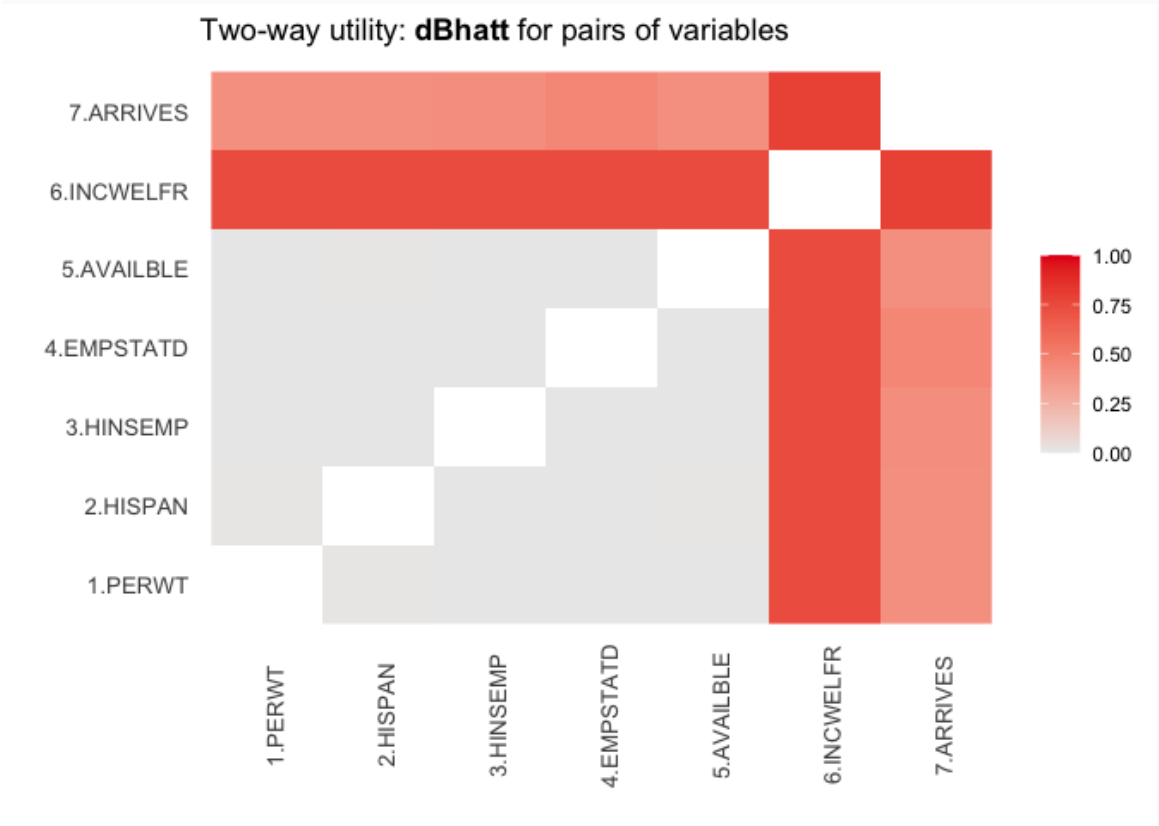

Two-way utility: S_pMSE for pairs of variables
7.ARRIVES
6.INCWELFR
5.AVAILBLE
4.EMPSTATD
3.HINSEMP
2.HISPAN
1.PERWT
1.PERWT
2.HISPAN
3.HINSEMP
4.EMPSTATD
5.AVAILBLE
6.INCWELFR
7.ARRIVES
20
15
10
5
0

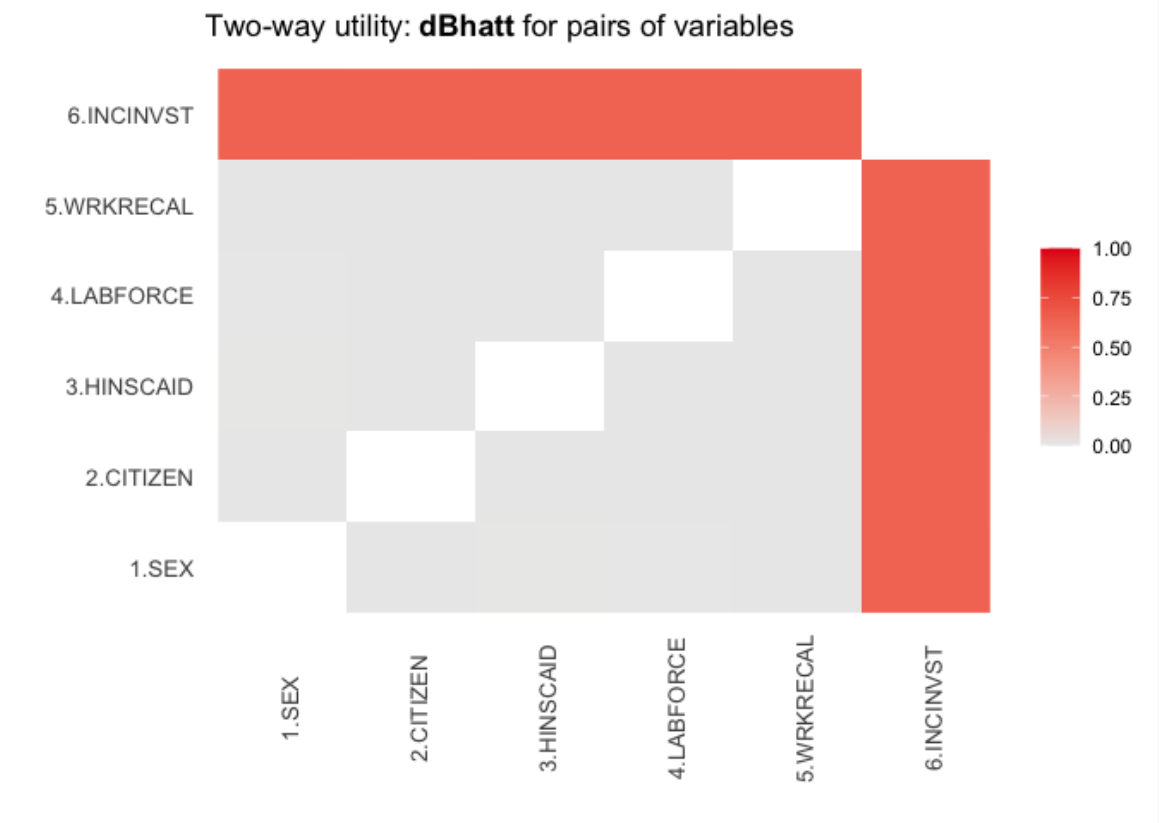

Two-way utility: S_pMSE for pairs of variables
6.INCINVST
5.WRKRECAL
4.LABFORCE
3.HINSCAID
2.CITIZEN
1.SEX
20
15
10
5
0
1.SEX
2.CITIZEN
3.HINSCAID
4.LABFORCE
5.WRKRECAL
6.INCINVST

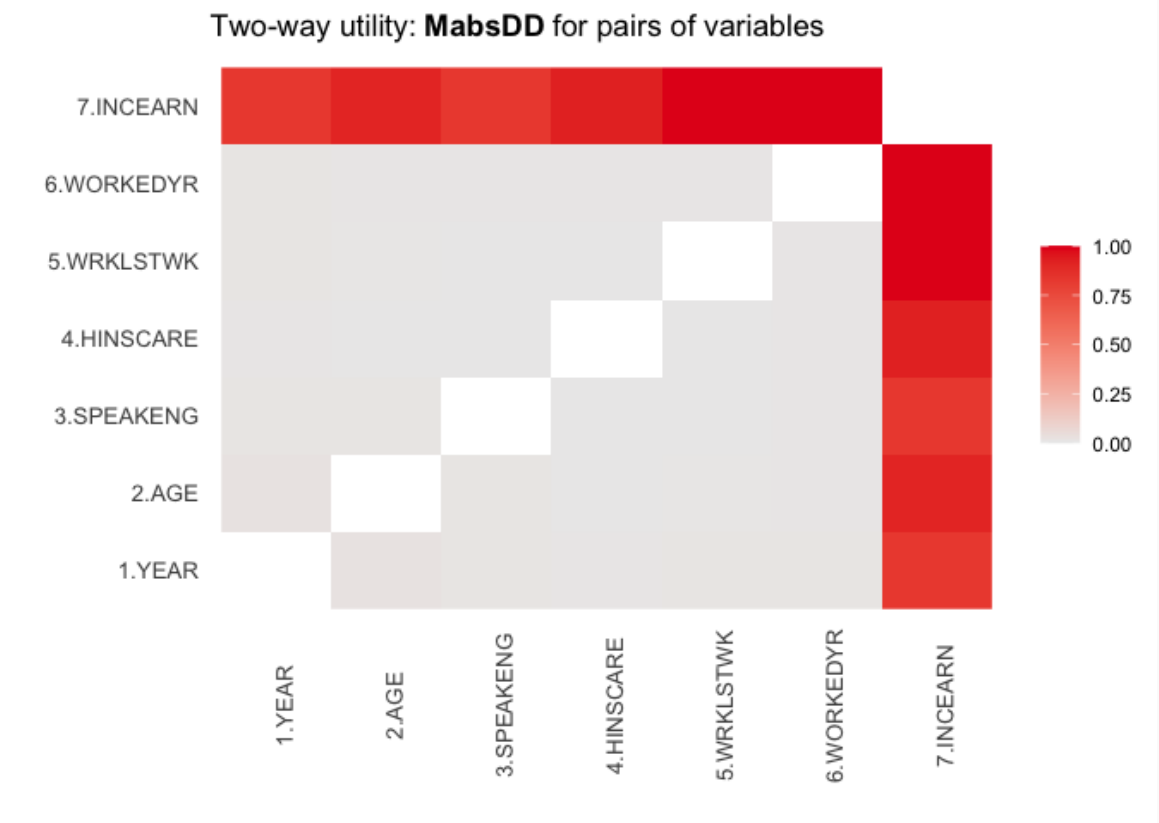

Two-way utility: dBhatt for pairs of variables
7.INCEARN
6.WORKEDYR
5.WRKLSTWK
4.HINSCARE
3.SPEAKENG
2.AGE
1.YEAR
1.00
0.75
0.50
0.25
0.00
1.YEAR
2.AGE
3.SPEAKENG
4.HINSCARE
5.WRKLSTWK
6.WORKEDYR
7.INCEARN

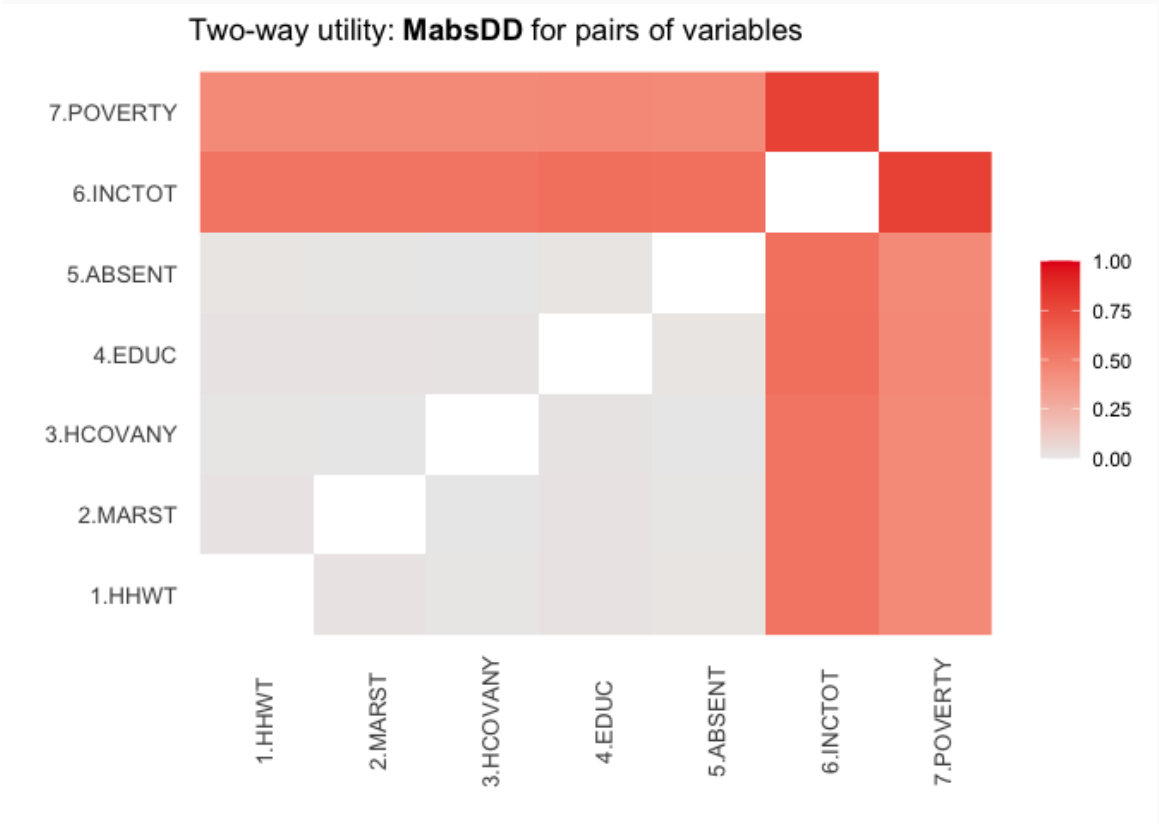

Two-way utility: dBhatt for pairs of variables
7.POVERTY
6.INCTOT
5.ABSENT
4.EDUC
3.HCOVANY
2.MARST
1.HHWT
1.00
0.75
0.50
0.25
0.00
1.HHWT
2.MARST
3.HCOVANY
4.EDUC
5.ABSENT
6.INCTOT
7.POVERTY

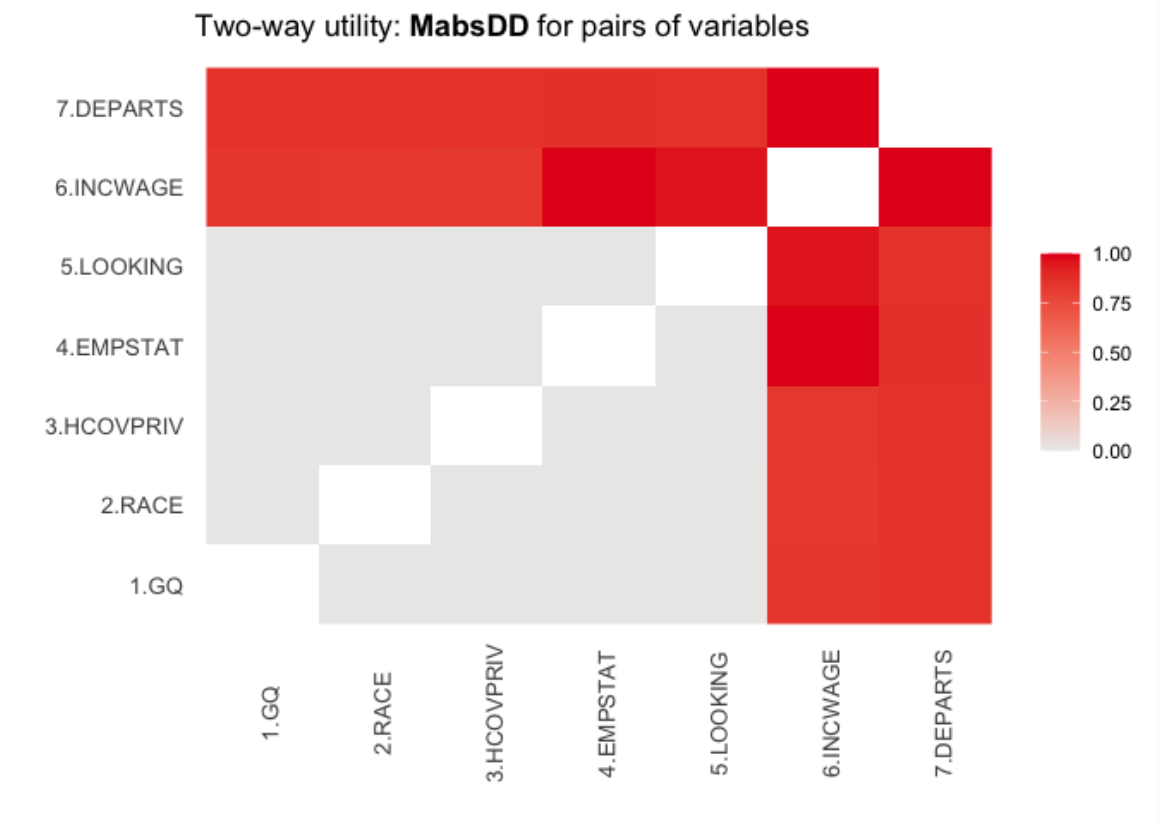

Two-way utility: dBhatt for pairs of variables
7.DEPARTS
6.INCWAGE
5.LOOKING
4.EMPSTAT
3.HCOVPRIV
2.RACE
1.GQ
1.GQ
2.RACE
3.HCOVPRIV
4.EMPSTAT
5.LOOKING
6.INCWAGE
7.DEPARTS
1.00
0.75
0.50
0.25
0.00

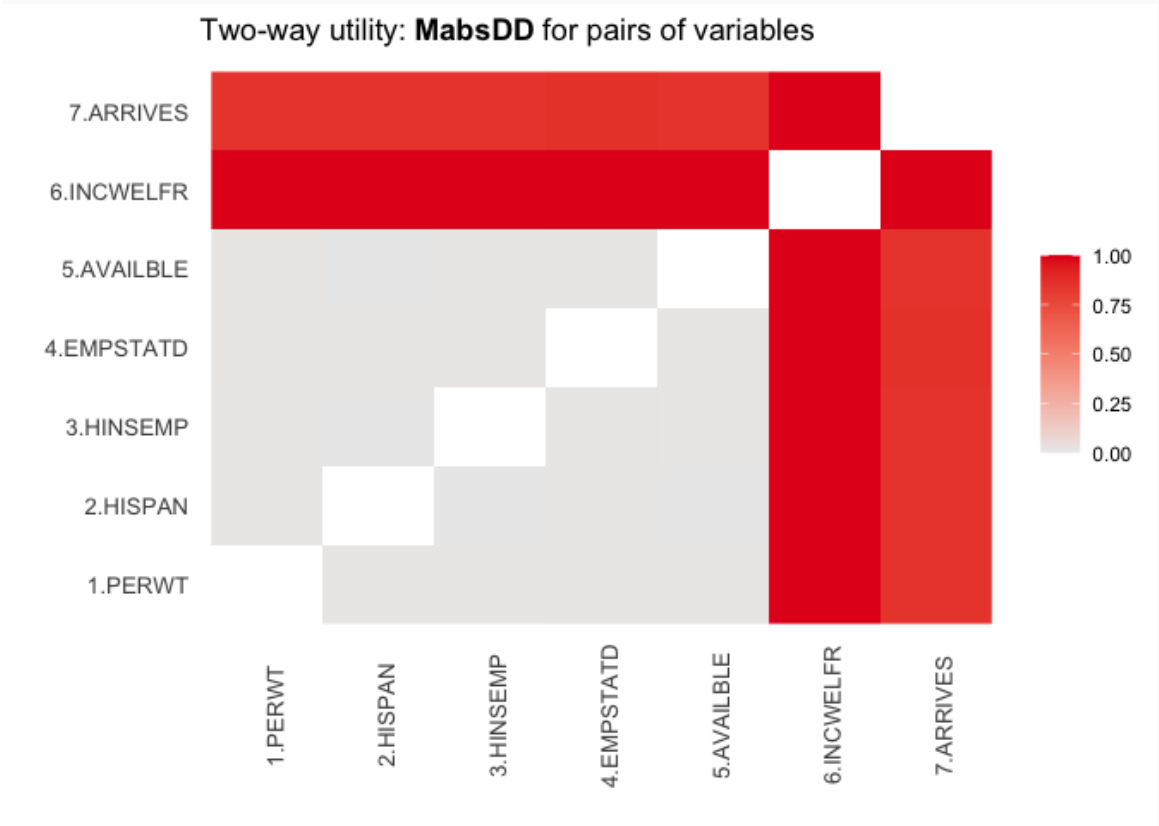

Two-way utility: dBhatt for pairs of variables
7.ARRIVES
6.INCWELFR
5.AVAILBLE
4.EMPSTATD
3.HINSEMP
2.HISPAN
1.PERWT
1.PERWT
2.HISPAN
3.HINSEMP
4.EMPSTATD
5.AVAILBLE
6.INCWELFR
7.ARRIVES
1.00
0.75
0.50
0.25
0.00

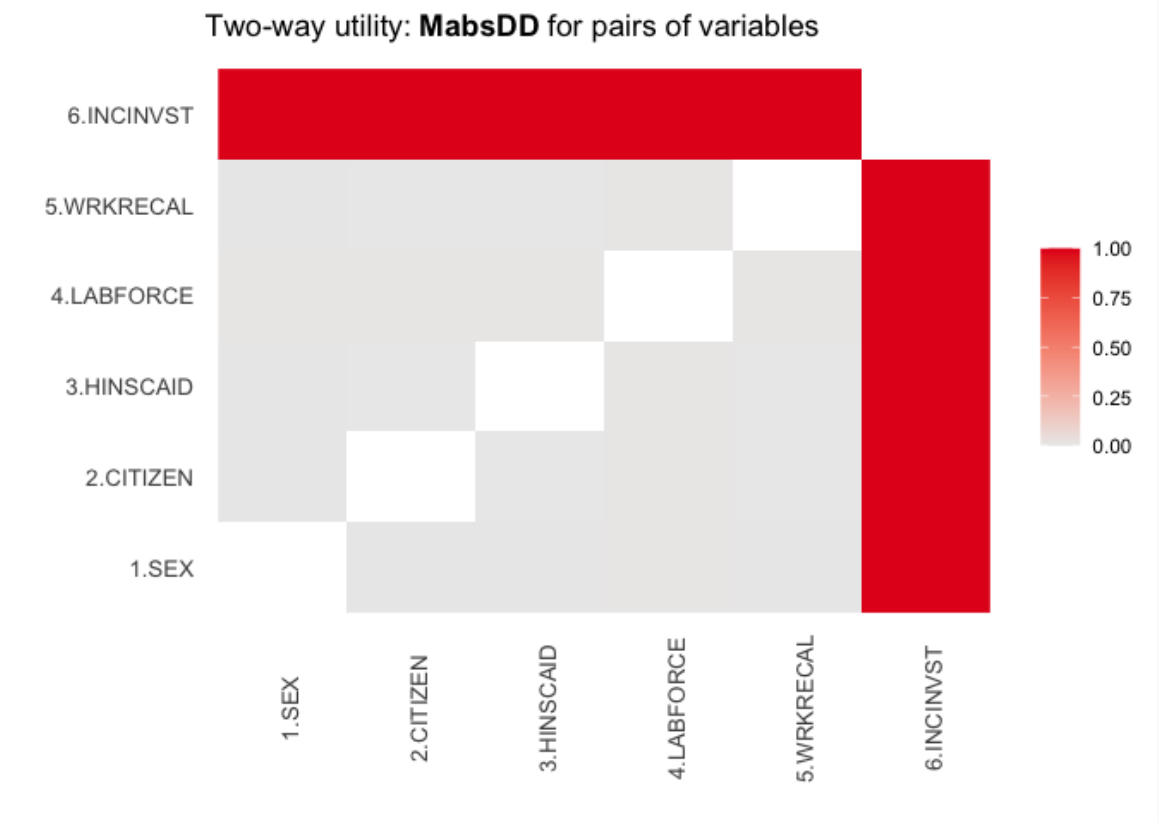

Two-way utility: dBhatt for pairs of variables
6.INCINVST
5.WRKRECAL
4.LABFORCE
3.HINSCAID
2.CITIZEN
1.SEX
1.00
0.75
0.50
0.25
0.00
1.SEX
2.CITIZEN
3.HINSCAID
4.LABFORCE
5.WRKRECAL
6.INCINVST


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


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


Two-way utility: MabsDD 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
1.00
0.75
0.50
0.25
0.00


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



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

Information.Loss
0.2339598

Individual Distances for Information Loss:

```
##      YEAR      HHWT        GQ     PERWT       SEX       AGE     MARST      RACE
## 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
##    HISPAN   CITIZEN  SPEAKENG   HCOVANY  HCOVPRIV   HINSEMP  HINSCAID  HINSCARE
## 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
##      EDUC   EMPSTAT  EMPSTATD  LABFORCE  WRKLSTWK    ABSENT   LOOKING  AVAILBLE
## 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
##  WRKRECAL  WORKEDYR    INCTOT   INCWAGE  INCWELFR  INCINVST   INCEARN   POVERTY
## 0.0000000 0.0000000 0.9998836 0.9998783 0.9931953 0.9996545 0.9998822 0.9817076
##   DEPARTS   ARRIVES
## 0.9902072 0.9902257
```