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:

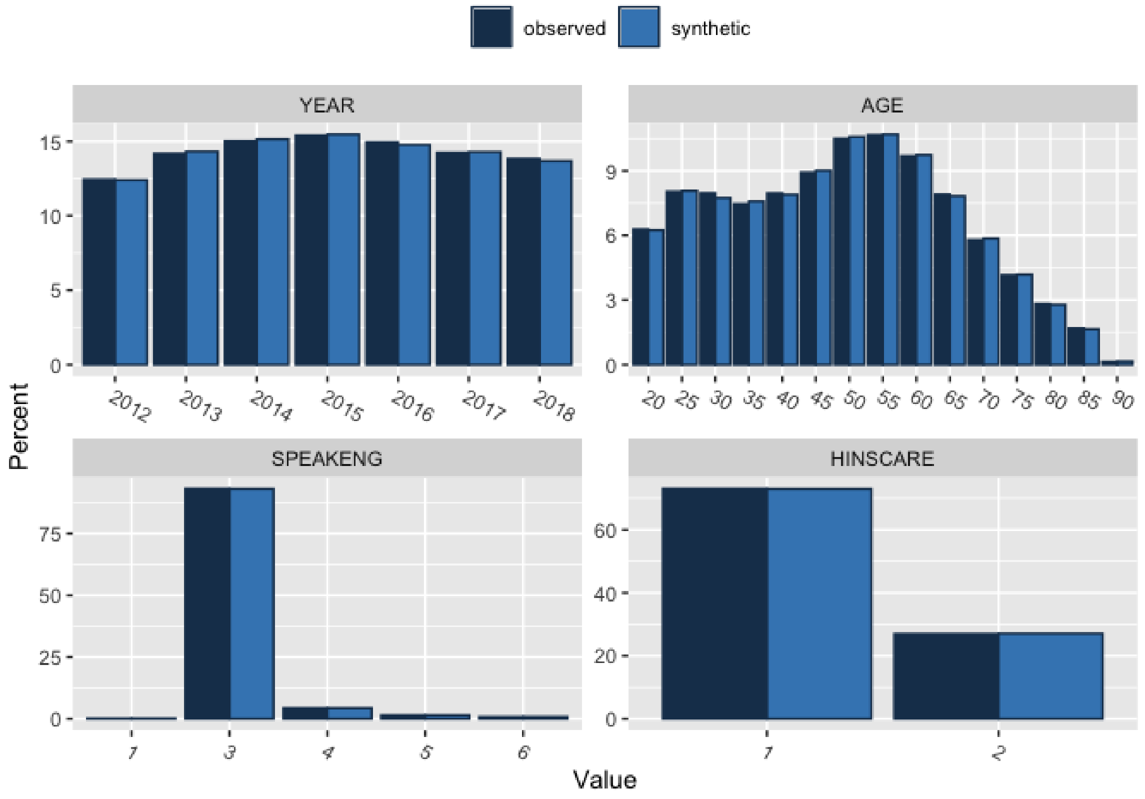
Metric	Number.Uniques	Number.Replications	Percentage.Replications
Perceived Risk	1e+05	12	0.012

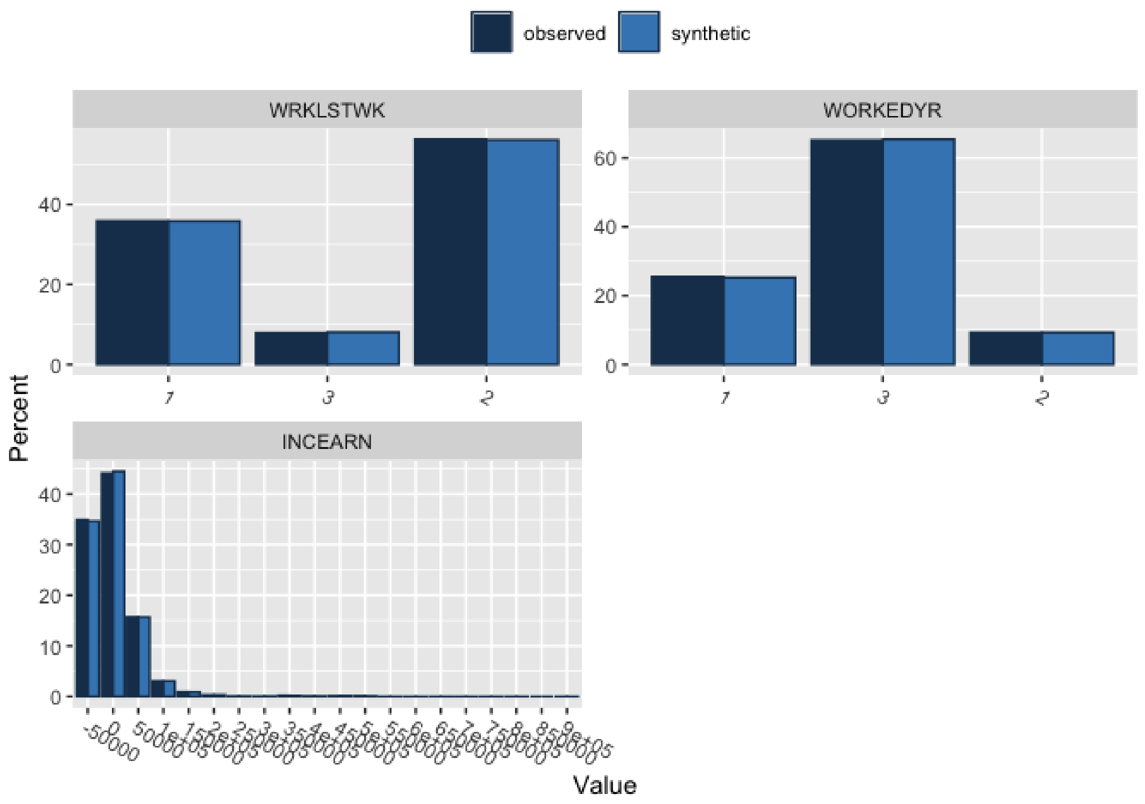
Utility Evaluation

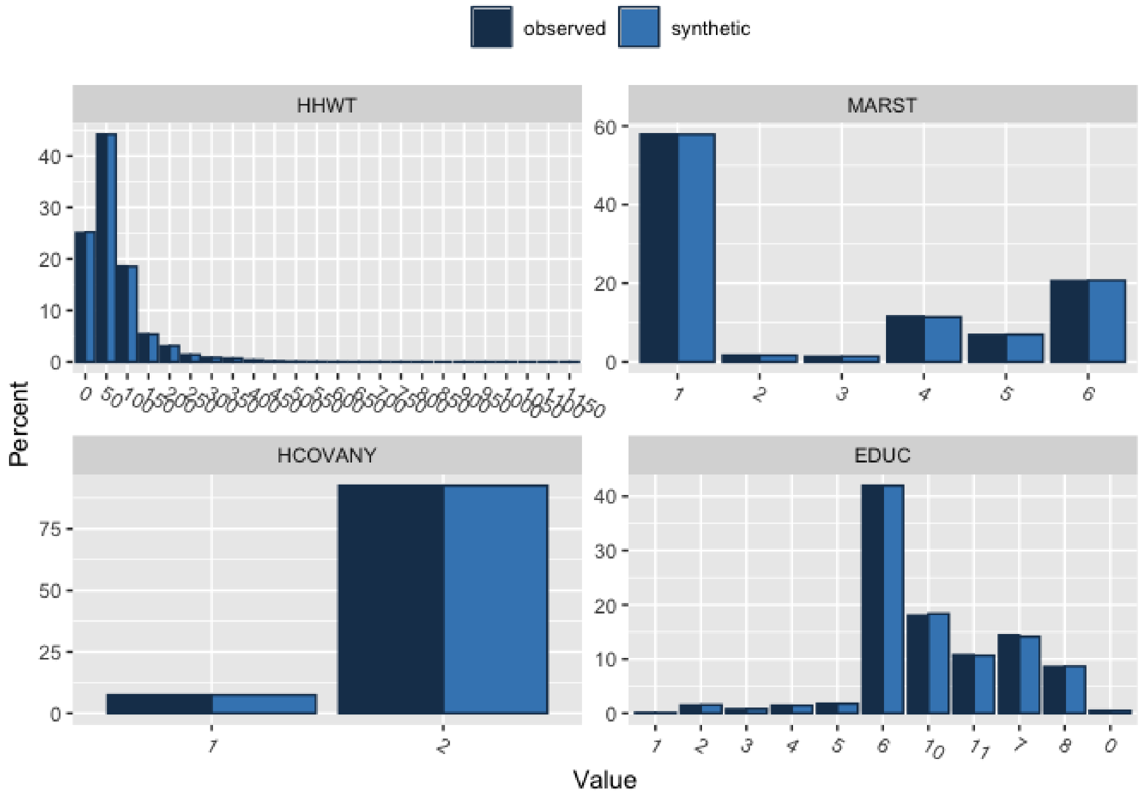
Different utility measures are applied in this section. These utility measures are the basis of utility evaluation for the generated synthetic dataset. The R packages synthpop, sdcMicro and corrplot were used to compute the following metrics. We do not use tests incorporating significance here. Confidence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

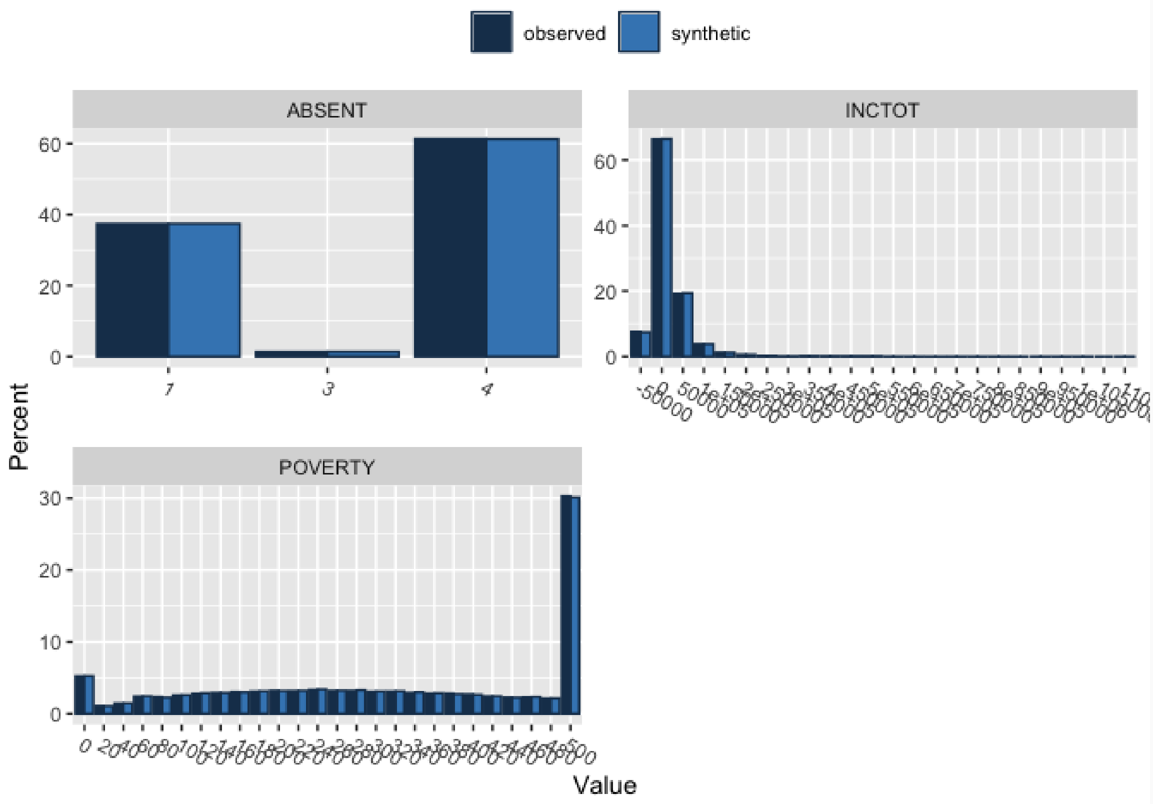
Graphical Comparison for Margins (R-Package: synthpop)

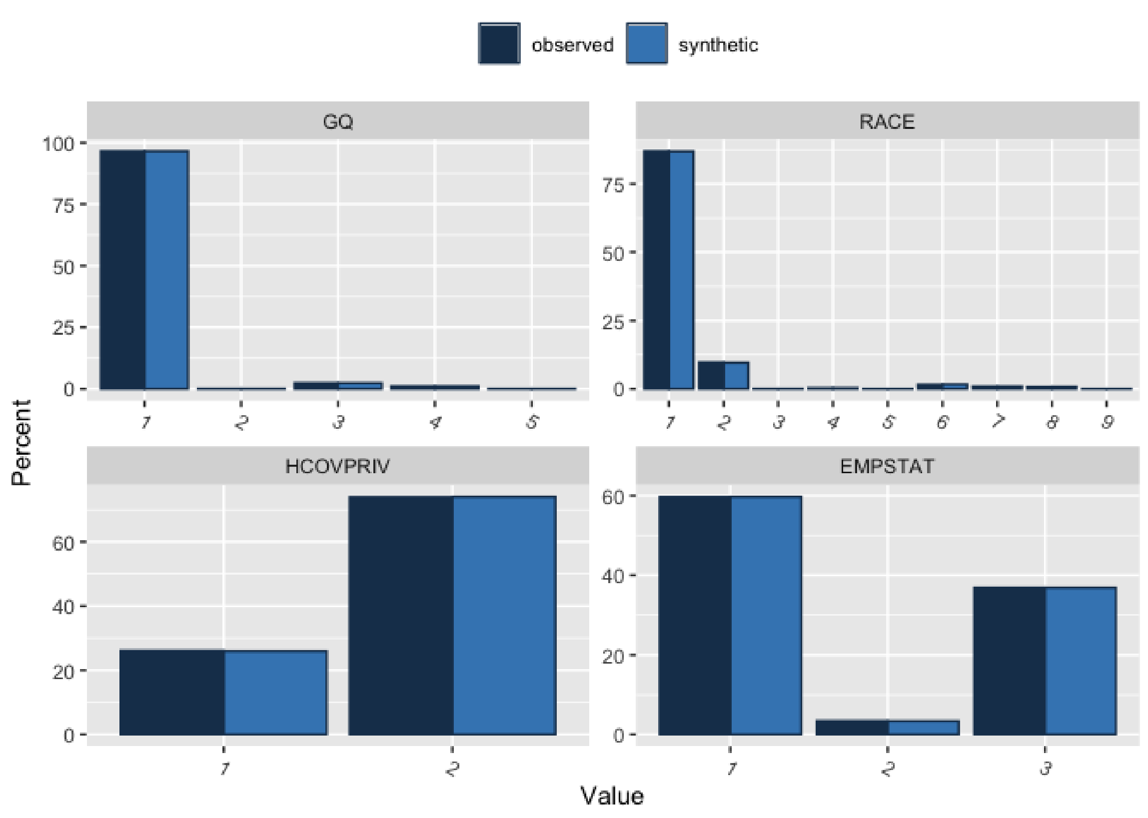
The following histograms provide an ad-hoc overview on the marginal distributions of the original and synthetic dataset. Matching or close distributions are related to a high data utility.

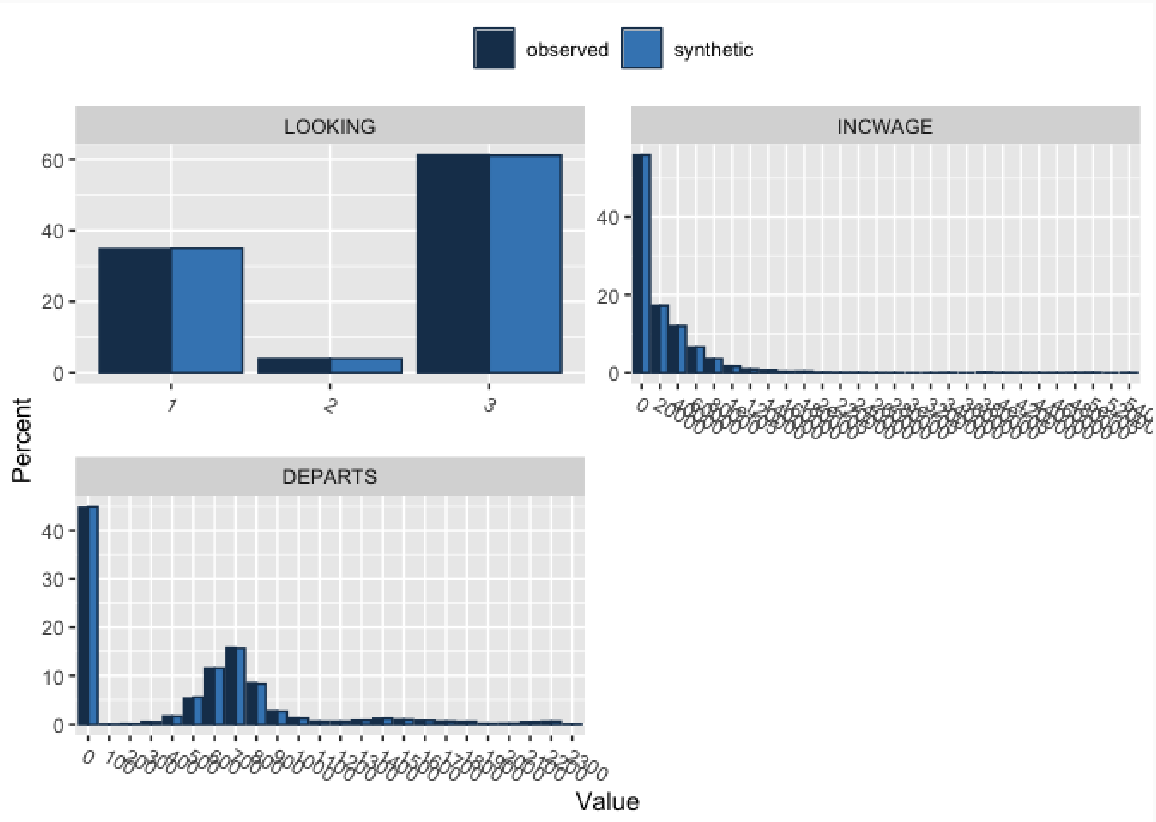

observed
synthetic
YEAR
AGE
SPEAKENG
HINSCARE
Percent
Value
2012
2013
2014
2015
2016
2017
2018
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
1
3
4
5
6
2

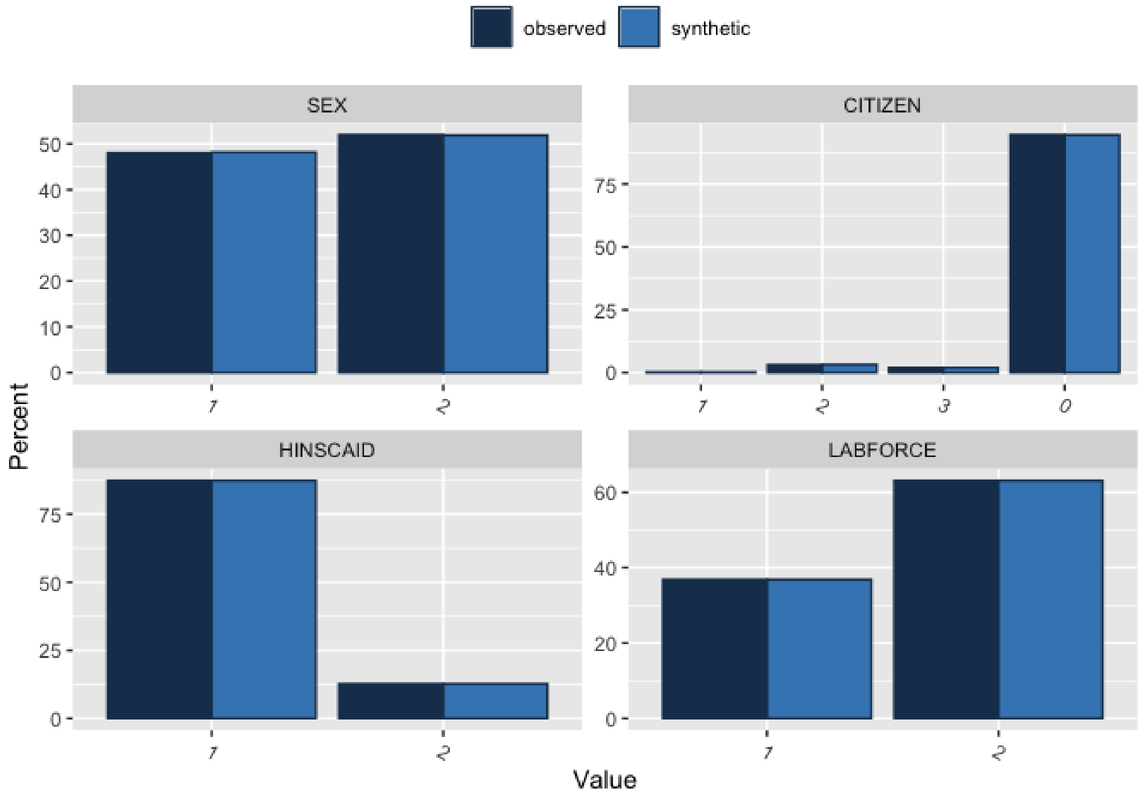

observed
synthetic
WRKLSTWK
WORKEDYR
INCEARN
Percent
Value
1
3
2

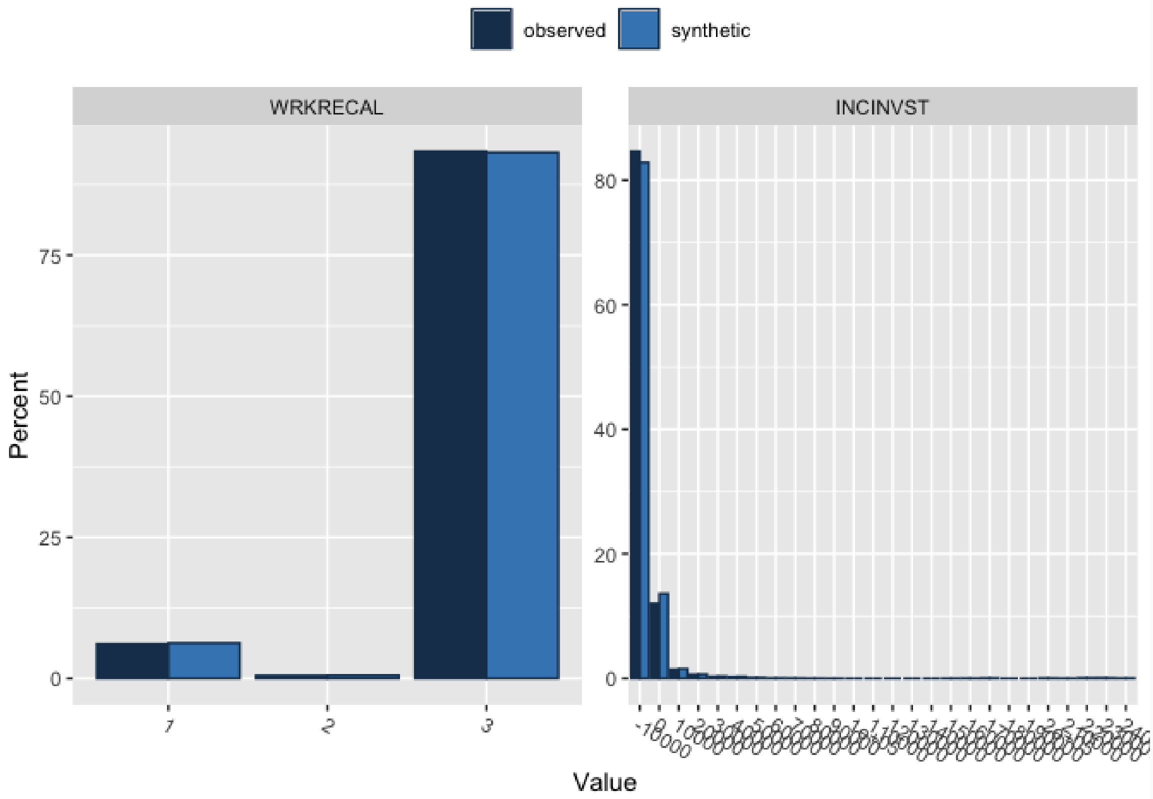

observed
synthetic
HHWT
MARST
HCOVANY
EDUC
Percent
Value


observed
synthetic
ABSENT
INCTOT
POVERTY
Percent
Value


observed
synthetic
GQ
RACE
HCOVPRIV
EMPSTAT
Percent
Value

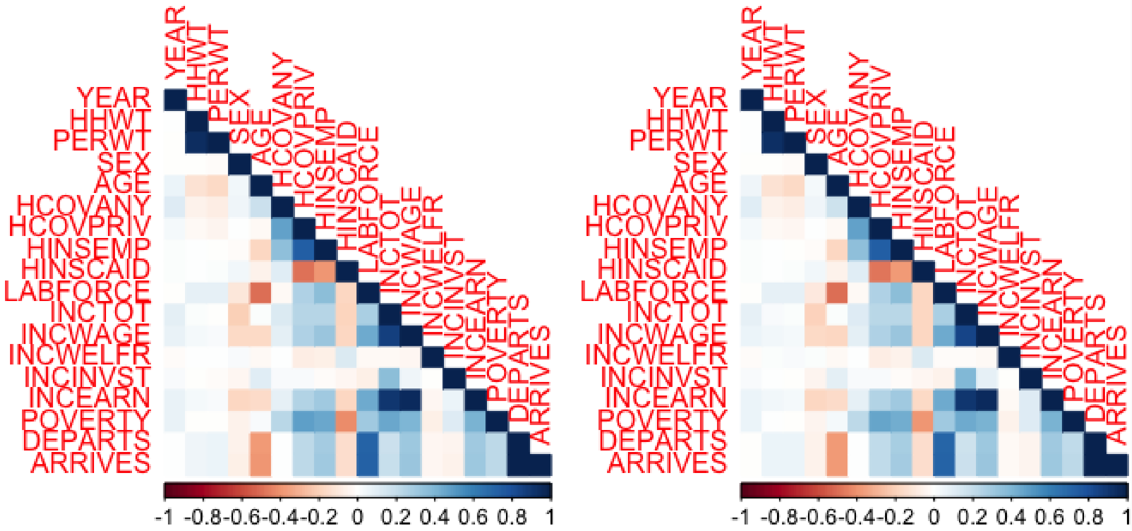

observed
synthetic
LOOKING
INCWAGE
DEPARTS
Percent
Value





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

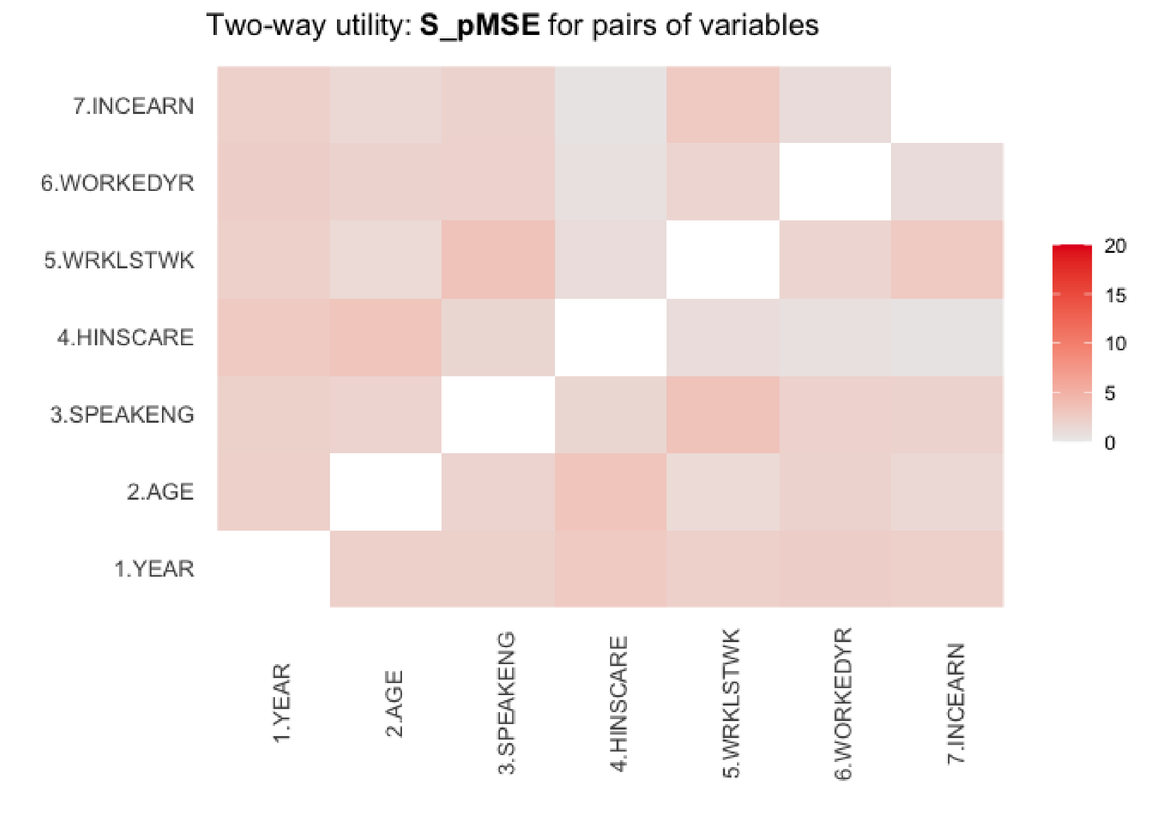
pMSE	S_pMSE
0.0008792	1.526391

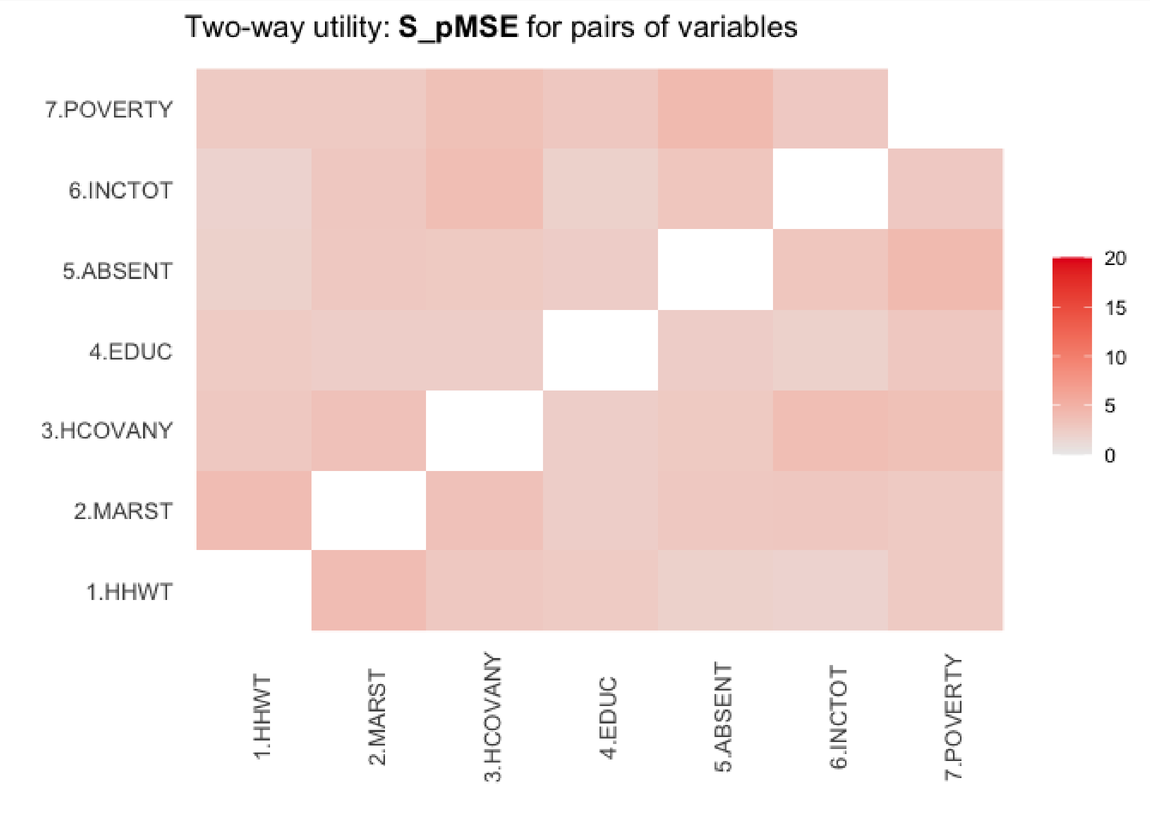
	pMSE	S_pMSE	df
SEX	0.0000004	0.5768435	1
CITIZEN	0.0000027	1.4147189	3
HINSCAID	0.0000000	0.0008118	1
LABFORCE	0.0000000	0.0003866	1
WRKRECAL	0.0000047	3.7820577	2
INCINVST	0.0001569	251.0007098	1

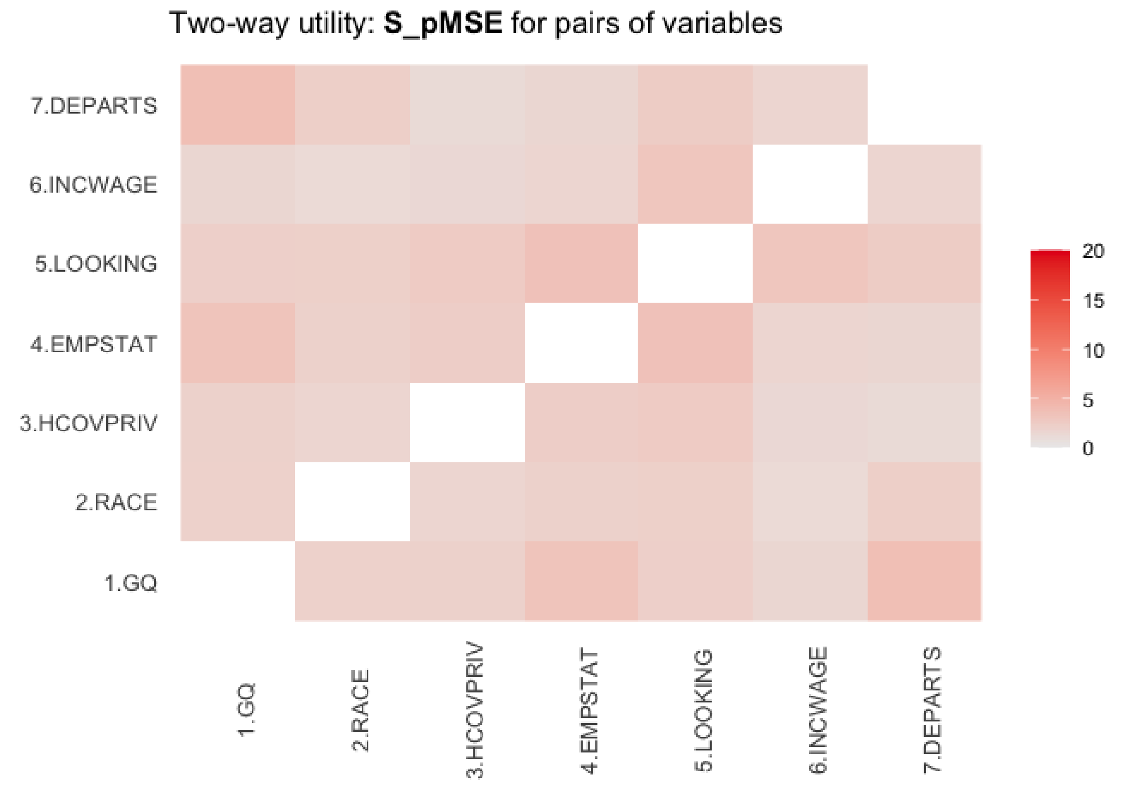
pMSE	S_pMSE
0.0004111	2.074587

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

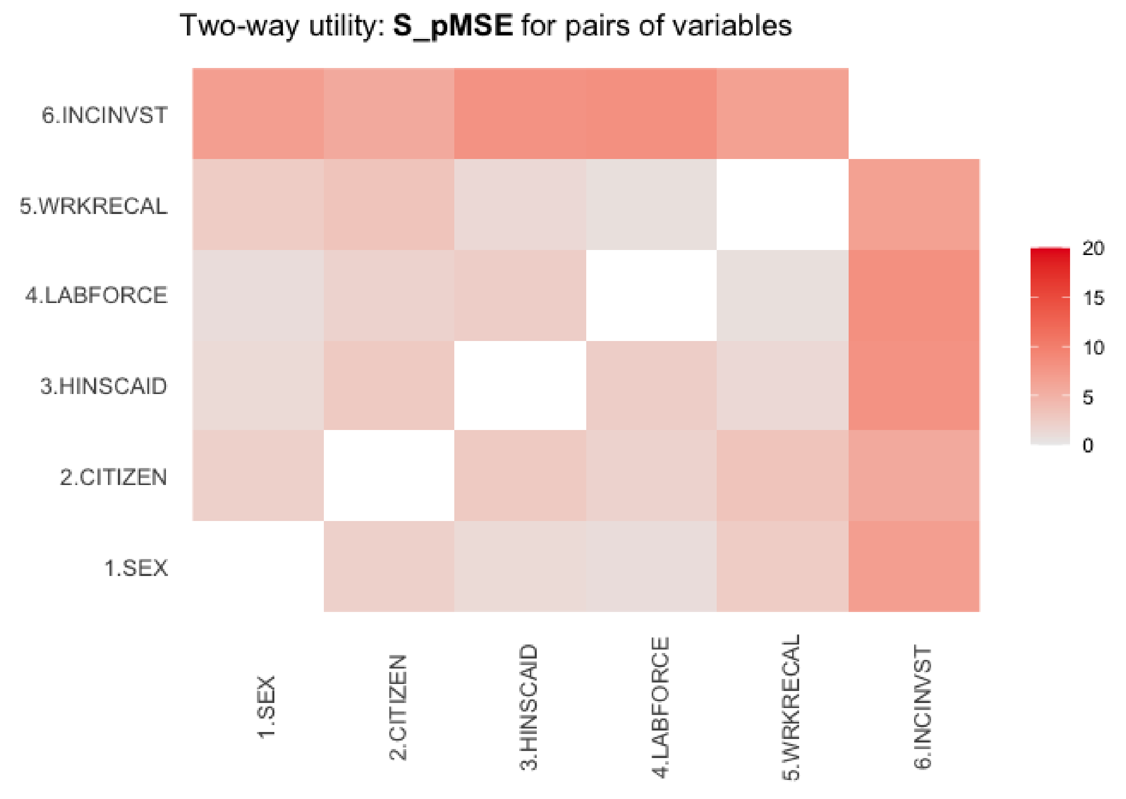
Two-way tables are evaluated based on the original and the synthetic dataset based on S_pMSE (see above). We also present the results for the mean absolute difference in densities (MabsDD) and the Bhattacharyya distance (dBhatt).

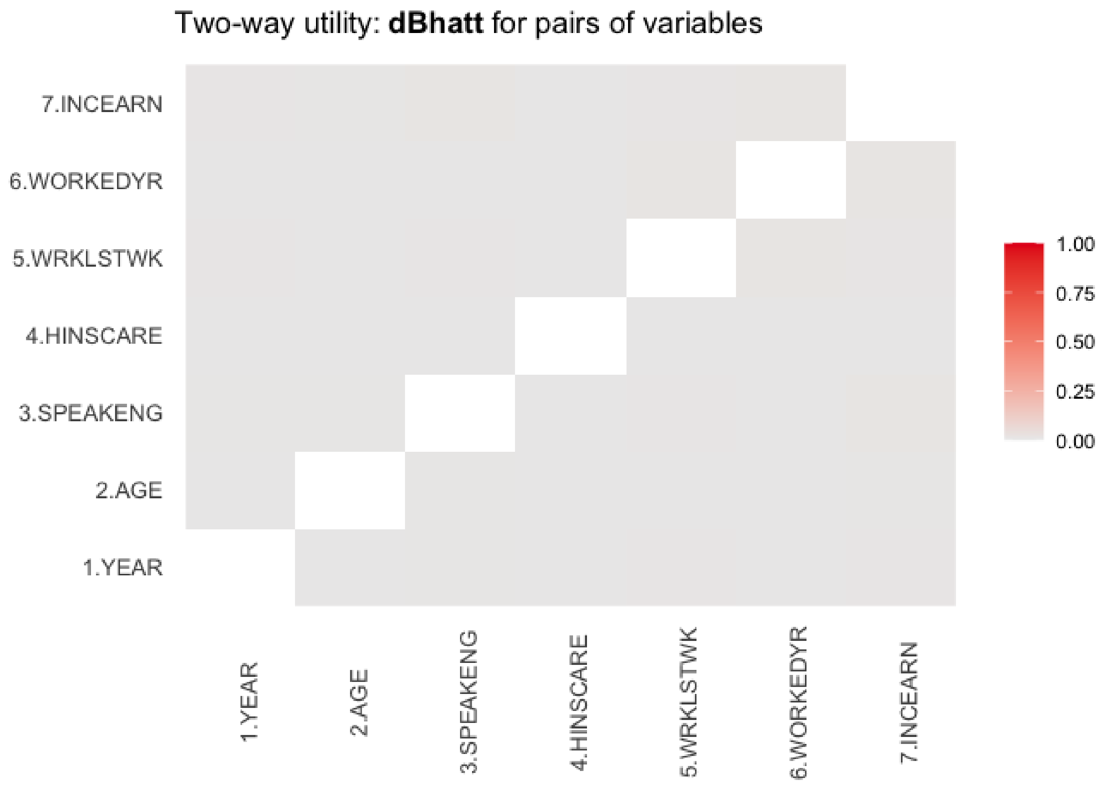




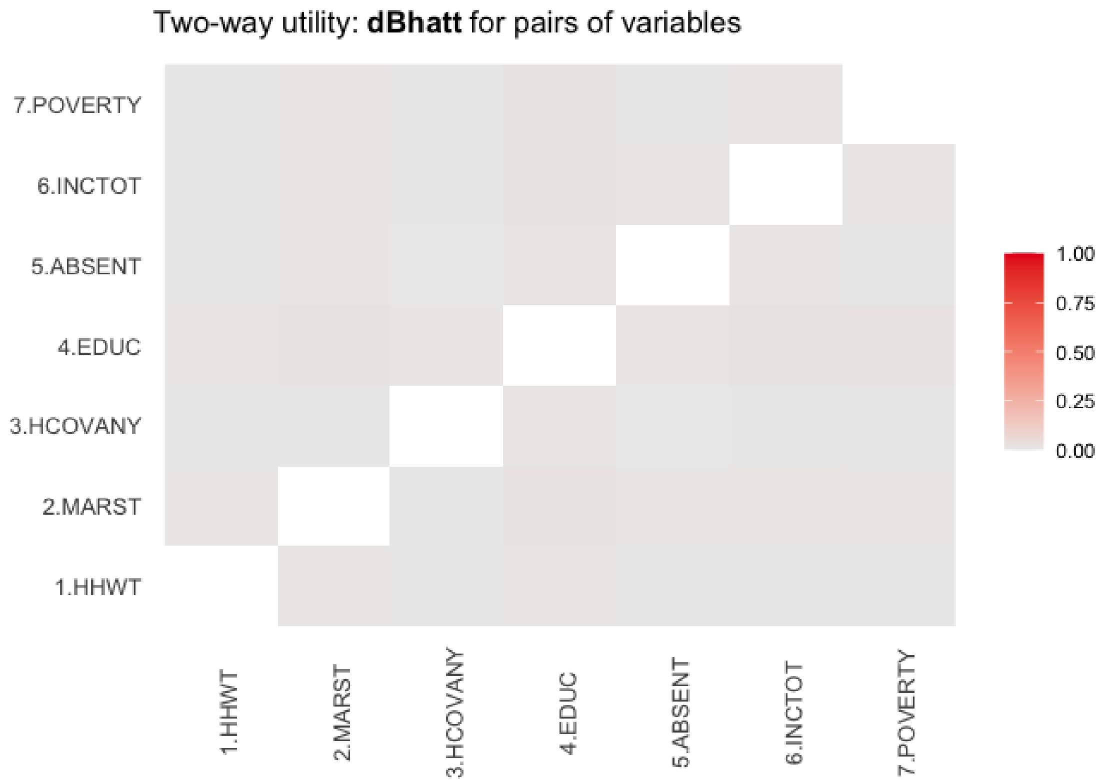

Two-way utility: S_pMSE for pairs of variables
7.DEPARTS
6.INCWAGE
5.LOOKING
4.EMPSTAT
3.HCOVPRIV
2.RACE
1.GQ
1.GQ
2.RACE
3.HCOVPRIV
4.EMPSTAT
5.LOOKING
6.INCWAGE
7.DEPARTS
20
15
10
5
0

```
## NULL
```

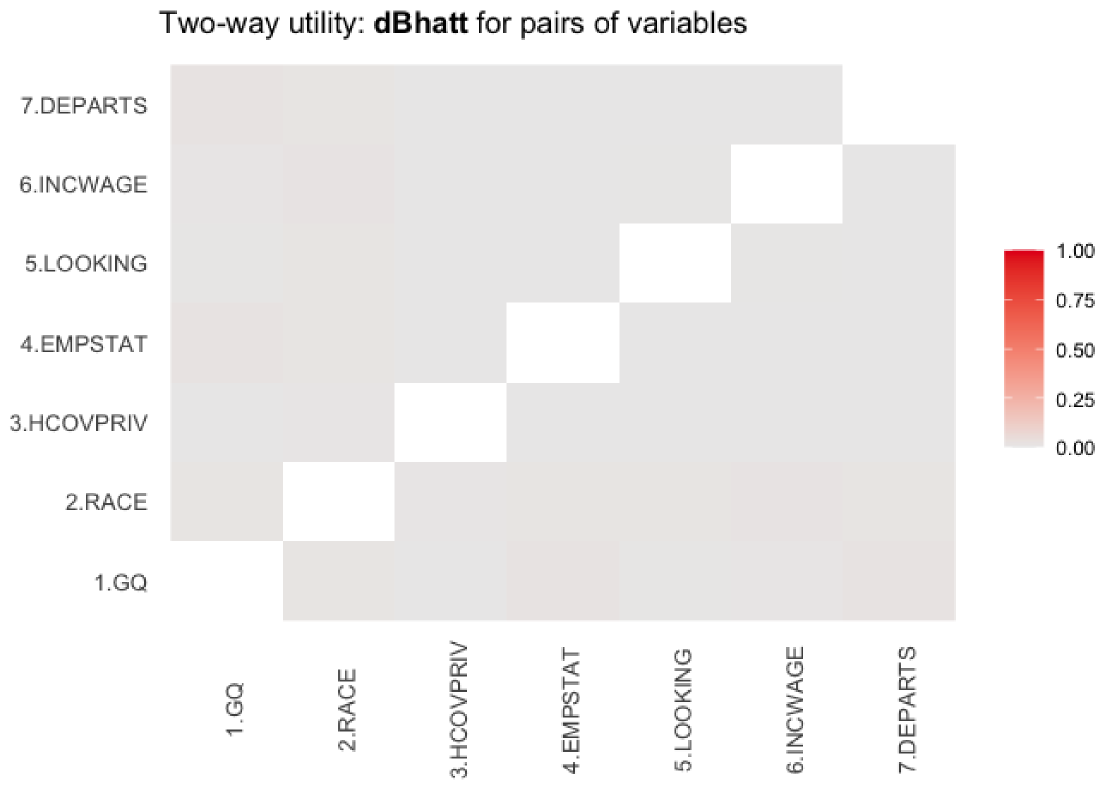

Two-way utility: S_pMSE for pairs of variables
6.INCINVST
5.WRKRECAL
4.LABFORCE
3.HINSCAID
2.CITIZEN
1.SEX
1.SEX
2.CITIZEN
3.HINSCAID
4.LABFORCE
5.WRKRECAL
6.INCINVST
20
15
10
5
0



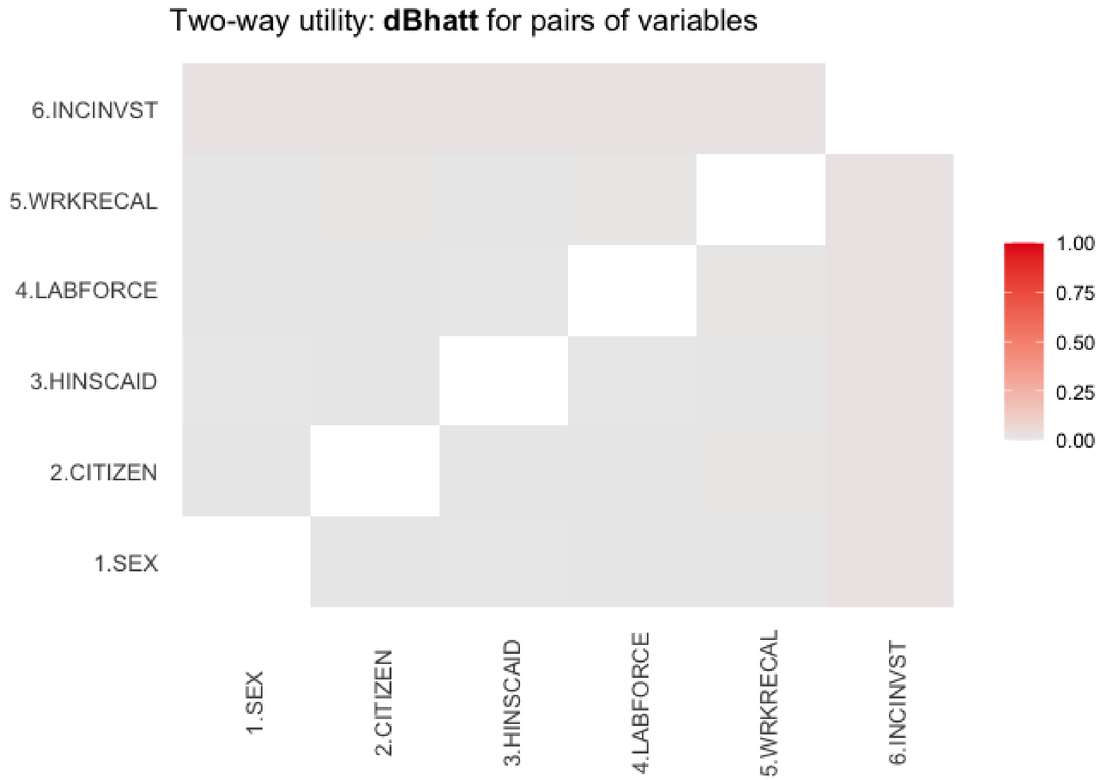
Two-way utility: dBhatt for pairs of variables
7.INCEARN
6.WORKEDYR
5.WRKLSTWK
4.HINSCARE
3.SPEAKENG
2.AGE
1.YEAR
1.YEAR
2.AGE
3.SPEAKENG
4.HINSCARE
5.WRKLSTWK
6.WORKEDYR
7.INCEARN
1.00
0.75
0.50
0.25
0.00

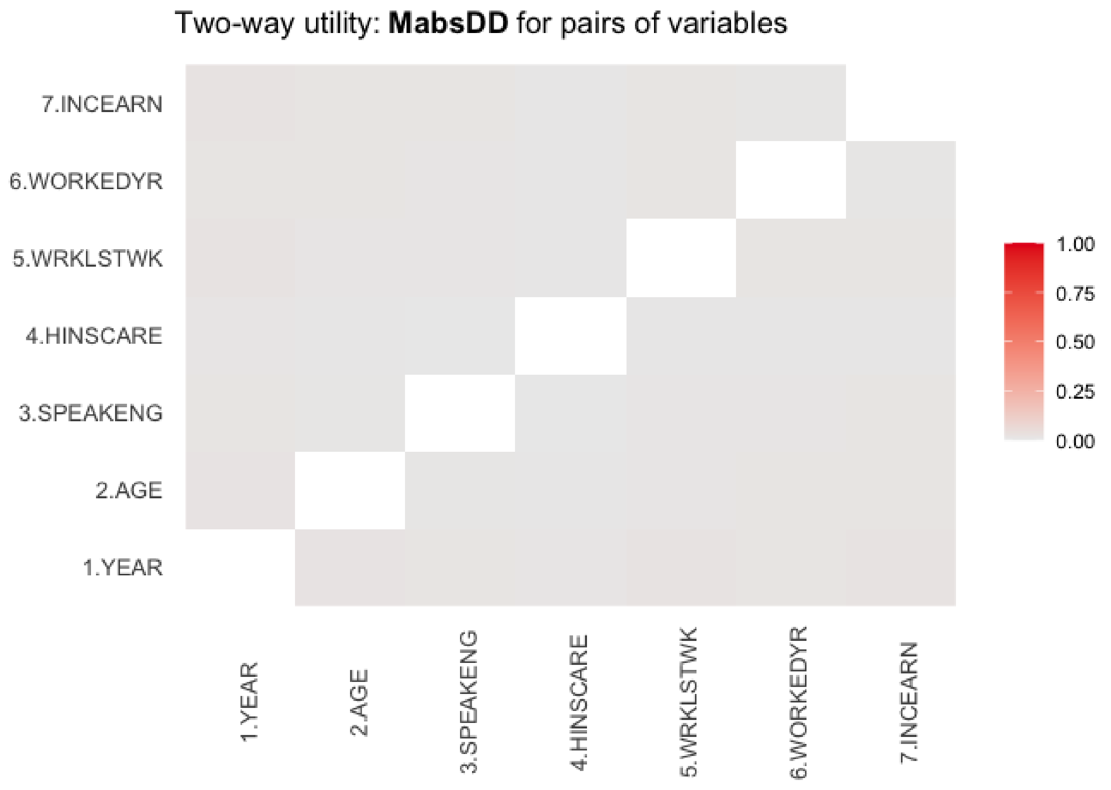


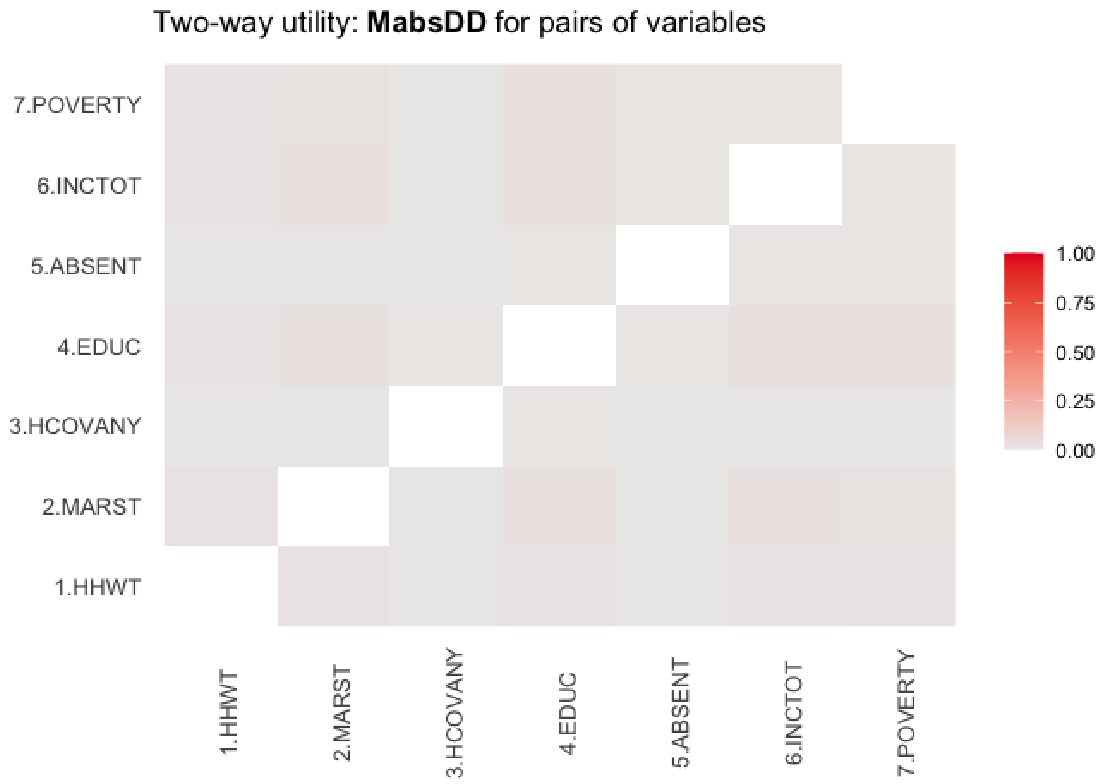
Two-way utility: dBhatt for pairs of variables
7.POVERTY
6.INCTOT
5.ABSENT
4.EDUC
3.HCOVANY
2.MARST
1.HHWT
1.HHWT
2.MARST
3.HCOVANY
4.EDUC
5.ABSENT
6.INCTOT
7.POVERTY
1.00
0.75
0.50
0.25
0.00

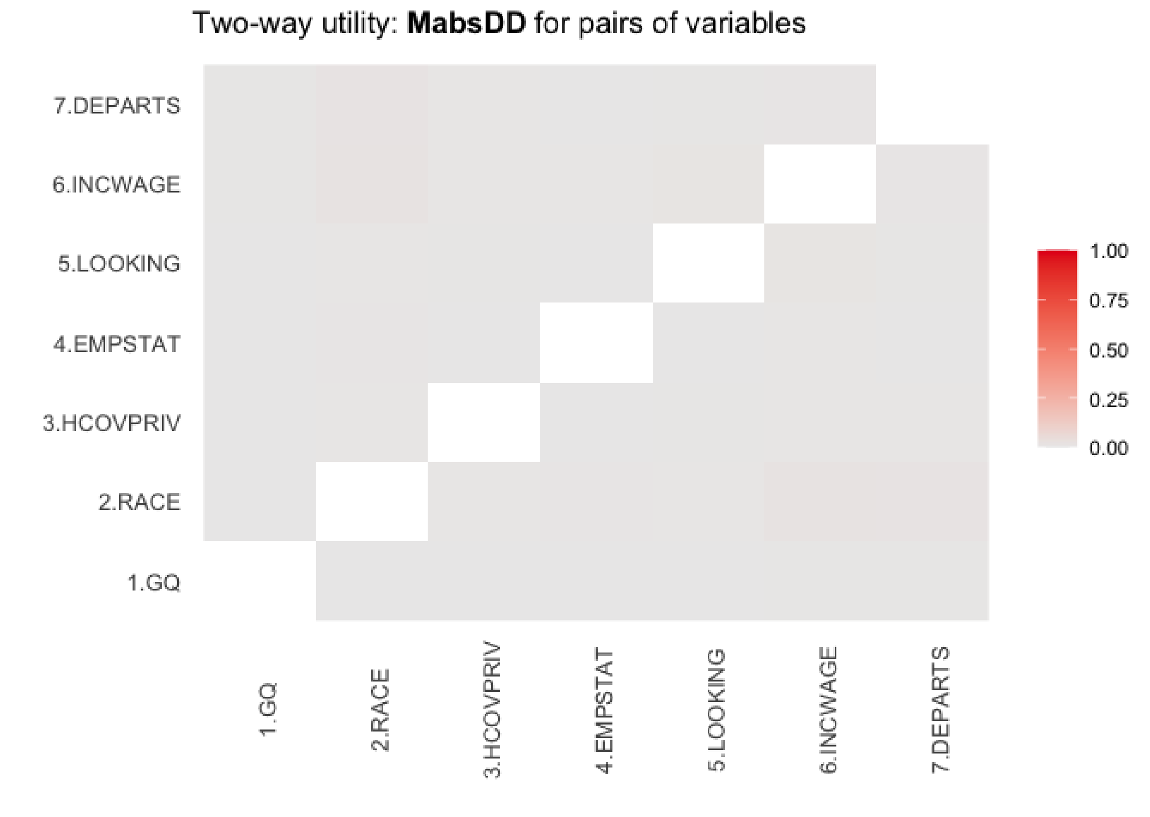

Two-way utility: dBhatt for pairs of variables
7.DEPARTS
6.INCWAGE
5.LOOKING
4.EMPSTAT
3.HCOVPRIV
2.RACE
1.GQ
1.GQ
2.RACE
3.HCOVPRIV
4.EMPSTAT
5.LOOKING
6.INCWAGE
7.DEPARTS
1.00
0.75
0.50
0.25
0.00

```
## NULL
```

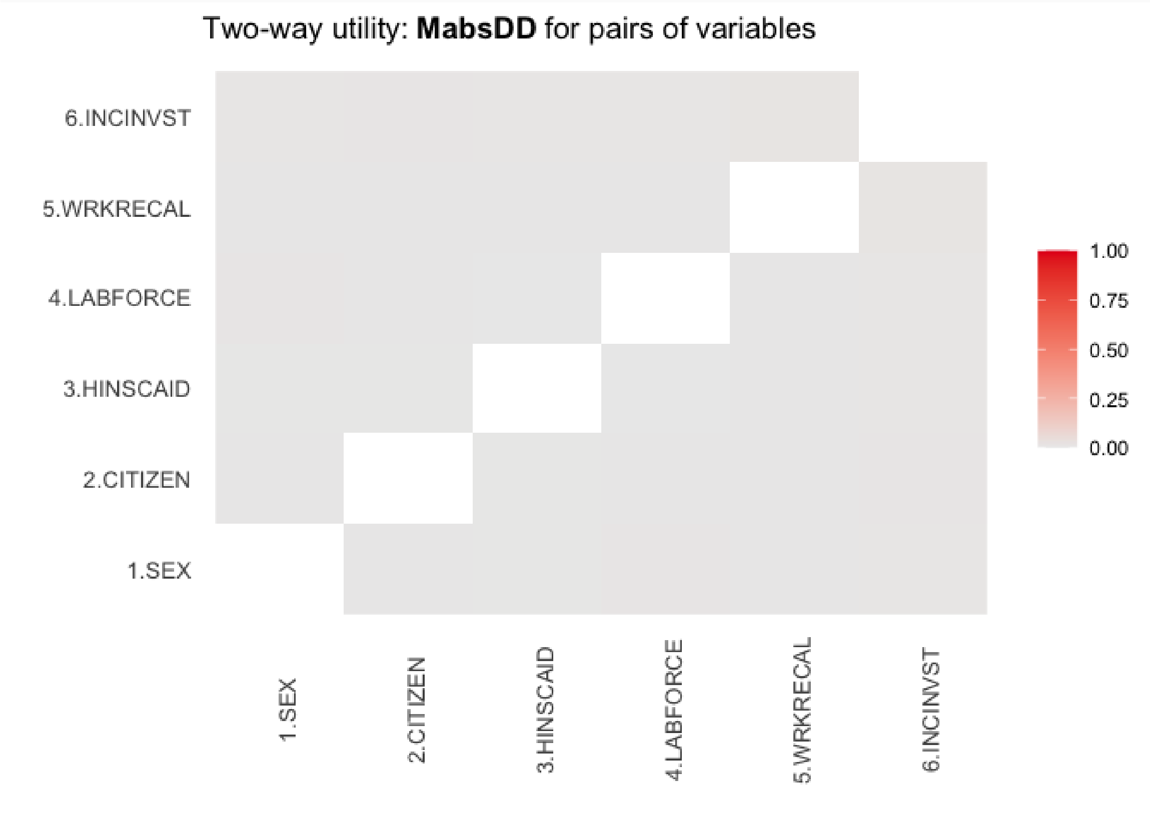

Two-way utility: dBhatt for pairs of variables
6.INCINVST
5.WRKRECAL
4.LABFORCE
3.HINSCAID
2.CITIZEN
1.SEX
1.SEX
2.CITIZEN
3.HINSCAID
4.LABFORCE
5.WRKRECAL
6.INCINVST
1.00
0.75
0.50
0.25
0.00


Two-way utility: MabsDD 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: MabsDD 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




```
## 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
```

ACS - Generative Adversarial Network (GAN)

Evaluation Synthetic Data Creation

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

January 28, 2022

Executive Summary

We used mainly the `sdv` python libraries to employ GANs and tested the R package `ganGenerativeData`. The GAN algorithms required quite a lot of computing power, which is a clear downside. From our main metrics the final GAN result seemed like a good trade-off between utility and privacy. Looking at utility measures weakens the first impression. The `S_pMSE` for tables and for distributions is extremely high. The Pearson correlation coefficients for binary and (semi-)continuous variables are also practically identical to those of the original dataset. The absolute difference in densities shows mediocre results whereas the Bhattacharyya distance gives a slight better impression. There is **no reasonable utility** according to Mlodak's information loss criterion. From a privacy perspective the GAN looks quite good (also when looking at more detailed metrics). From our perspective it seemed like the GAN algorithms tend to extrapolate more than other algorithms like FCS.

USE CASE RECOMMENDATIONS

Releasing_to_Public	Testing_Analysis	Education	Testing_Technology
NO	NO	YES	YES

The utility of **GAN** has some flaws, thus we **don't** think it is a good idea to **release this data to the public**. This could lead to false impressions. Also scientists may be led to false conclusions when using this data for **testing analysis**. We could imagine GAN generated data in **education** or in **technology testing**. On first sight, it seems like GAN data is somehow to computationally intensive to consider it for testing, but we also see an advantage in the fact, that they tend a little more to extrapolate, what could be beneficial for testing.

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

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:

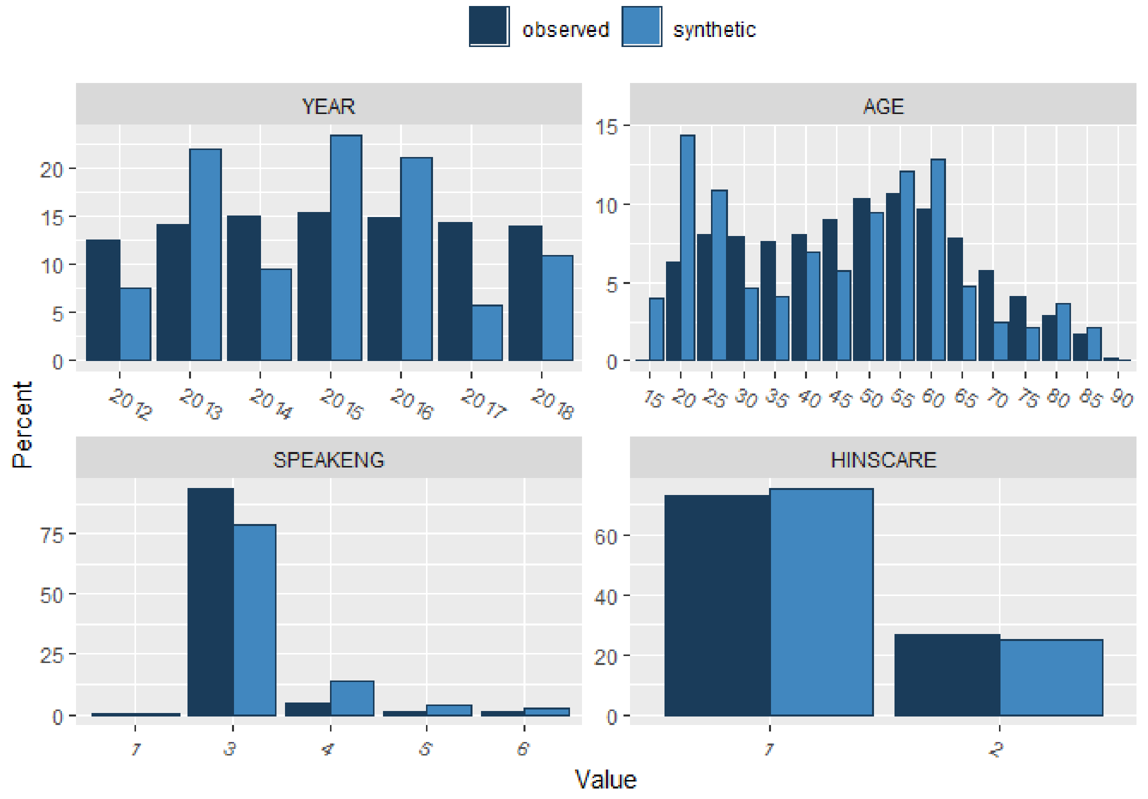
Metric	Number.Uniques	Number.Replications	Percentage.Replications
Perceived Risk	1035201	0	0

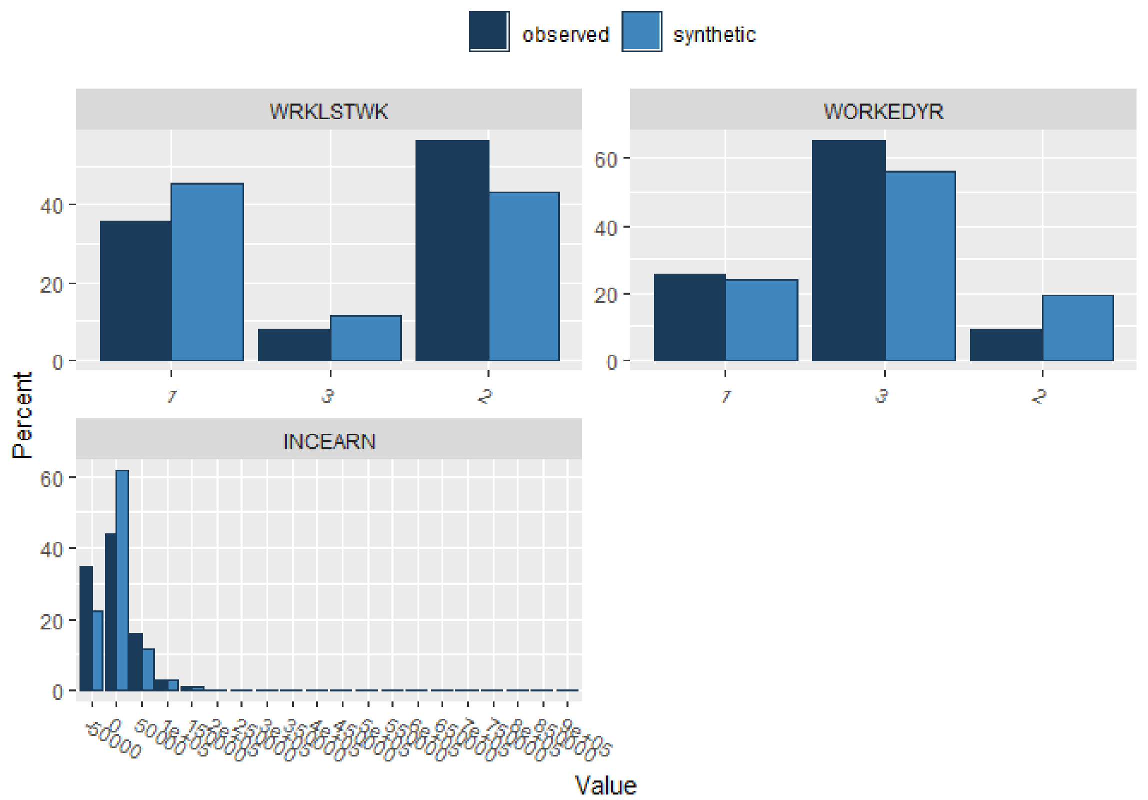
Utility Evaluation

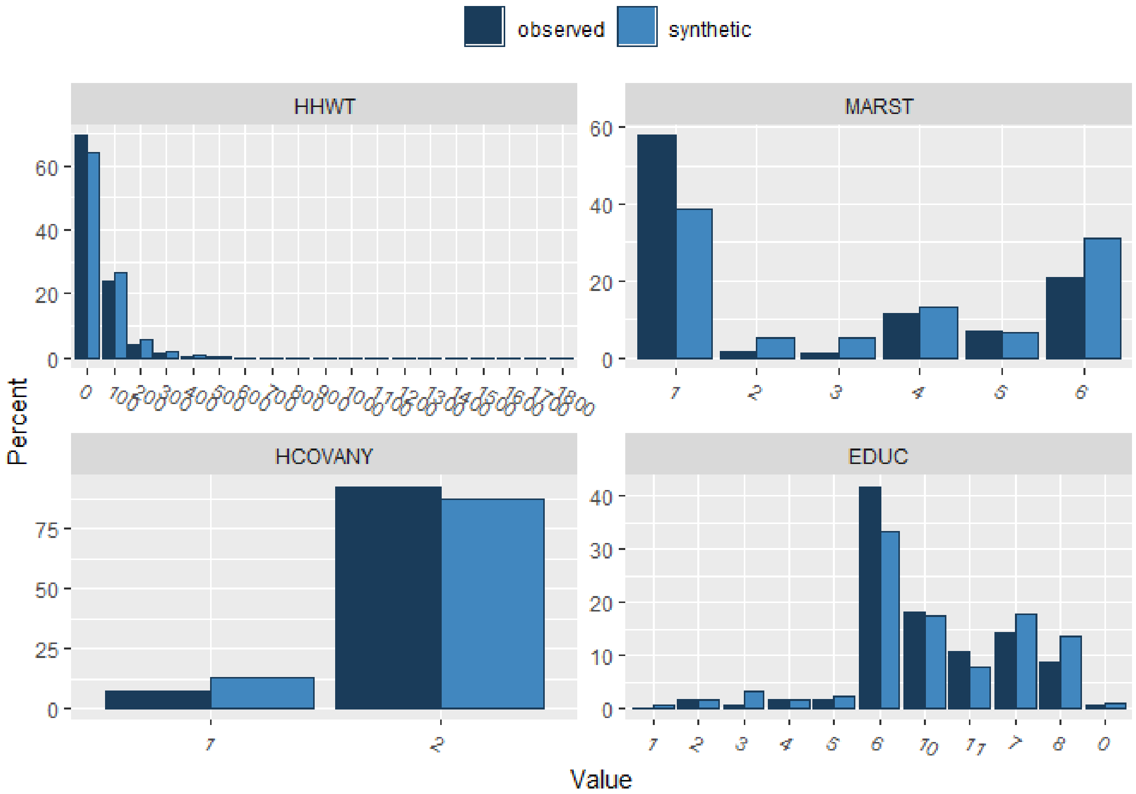
Different utility measures are applied in this section. These utility measures are the basis of utility evaluation for the generated synthetic dataset. The R packages synthpop, sdcMicro and corrplot were used to compute the following metrics. We do not use tests incorporating significance here. Confidence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

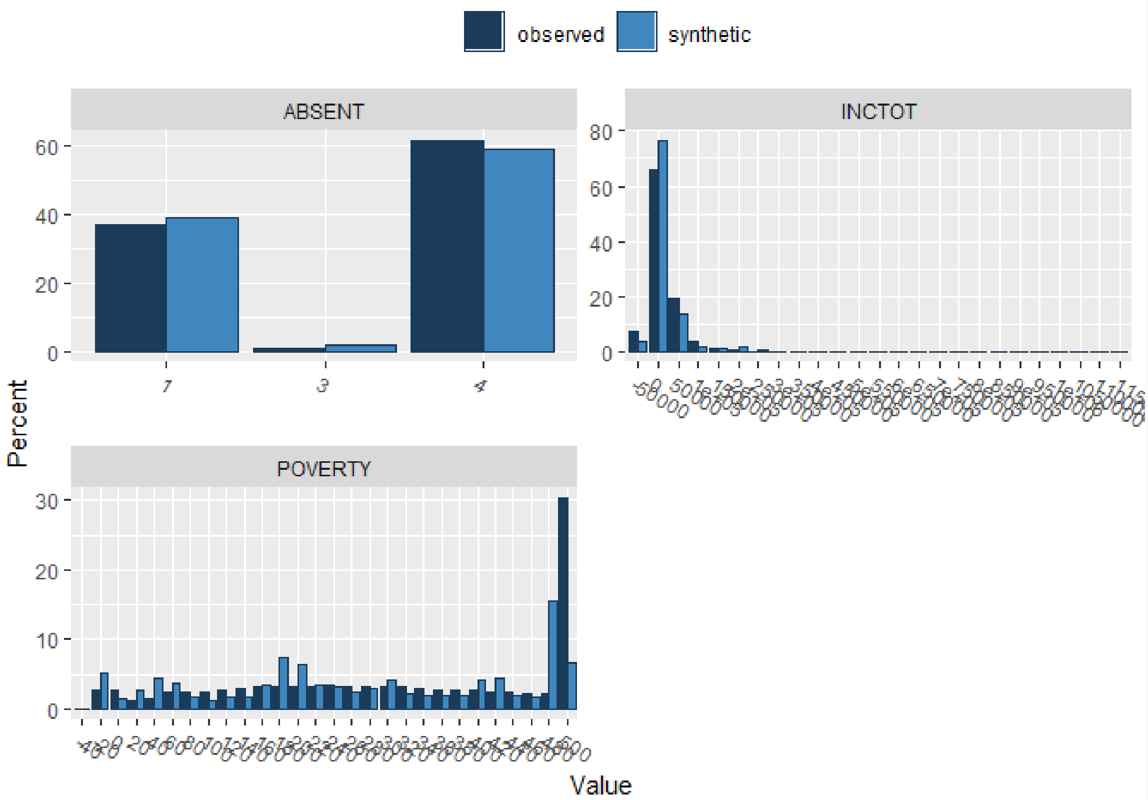
Graphical Comparison for Margins (R-Package: synthpop)

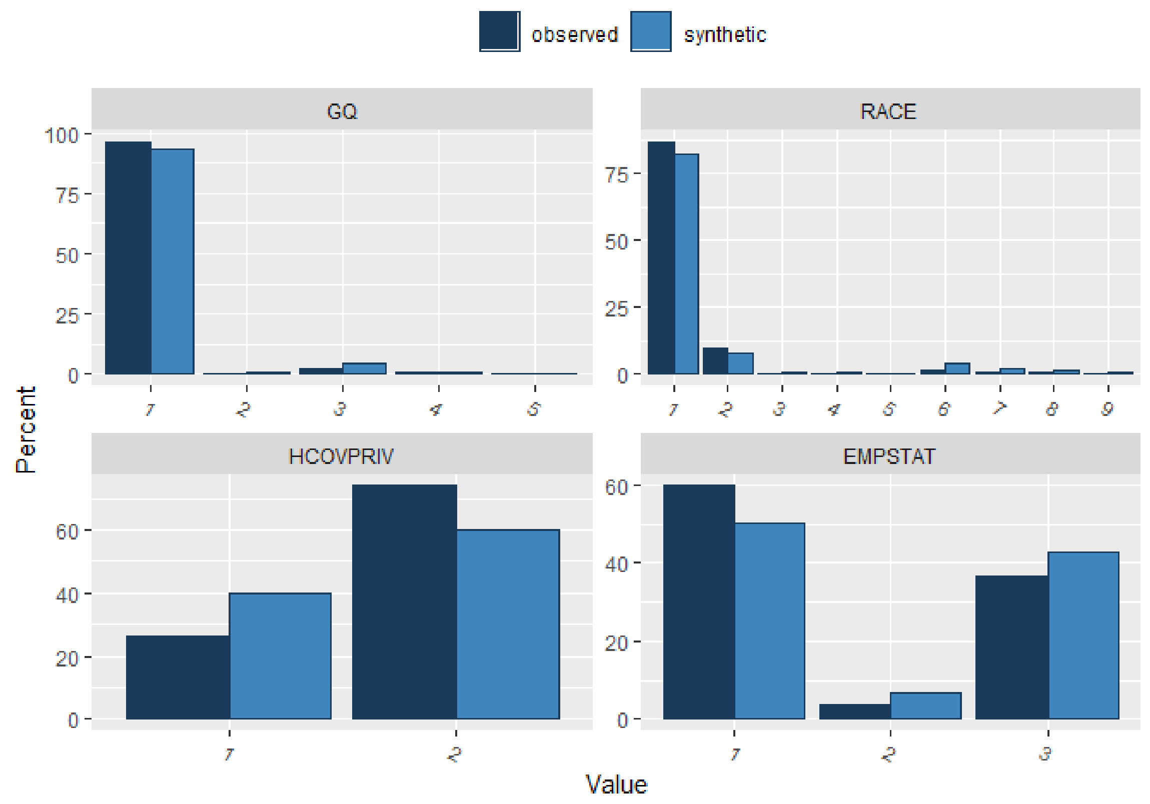
The following histograms provide an ad-hoc overview on the marginal distributions of the original and synthetic dataset. Matching or close distributions are related to a high data utility.

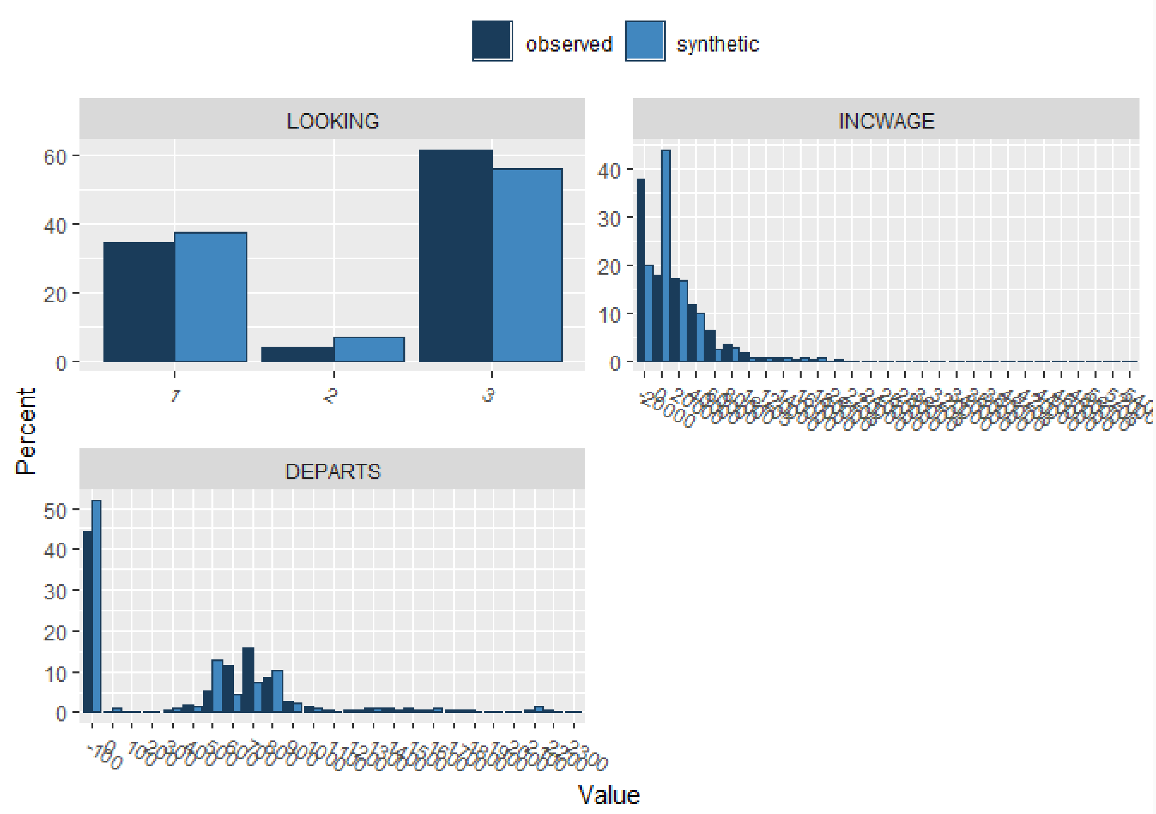

observed
synthetic
YEAR
AGE
SPEAKENG
HINSCARE
Percent
Value

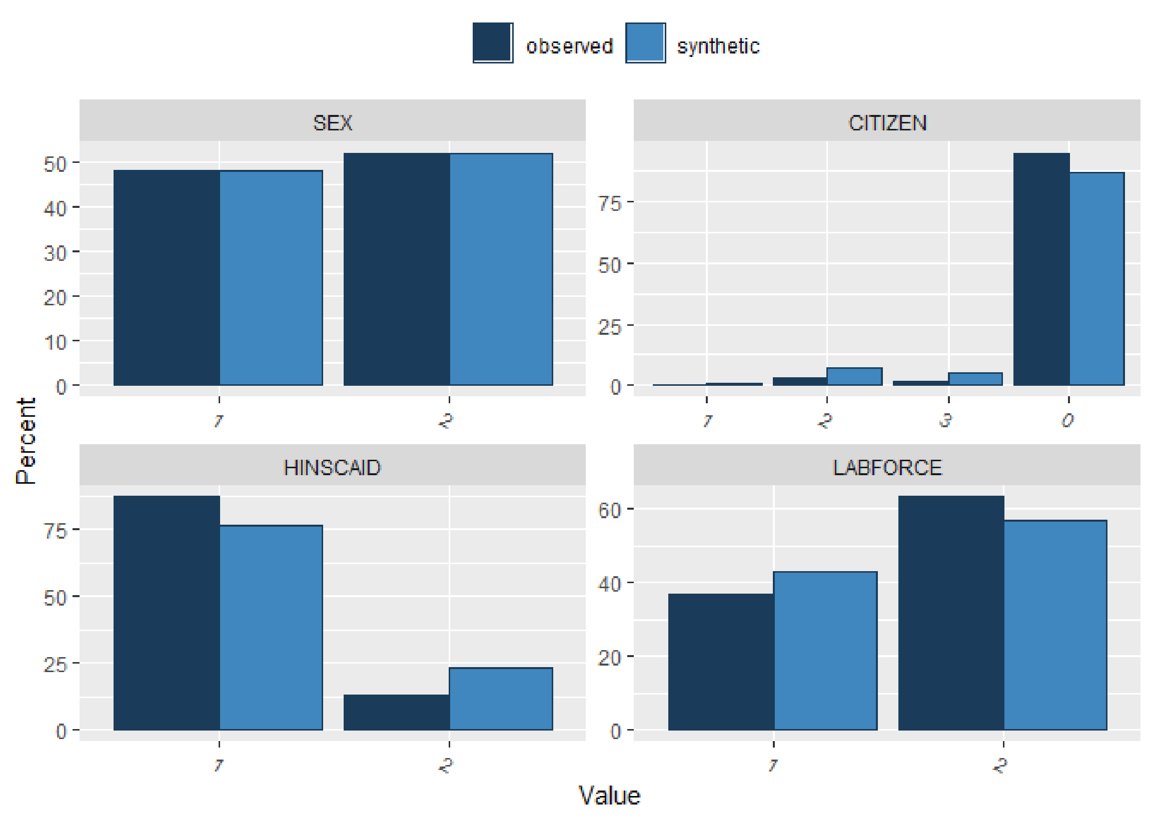

observed
synthetic
WRKLSTWK
WORKEDYR
INCEARN
Percent
Value

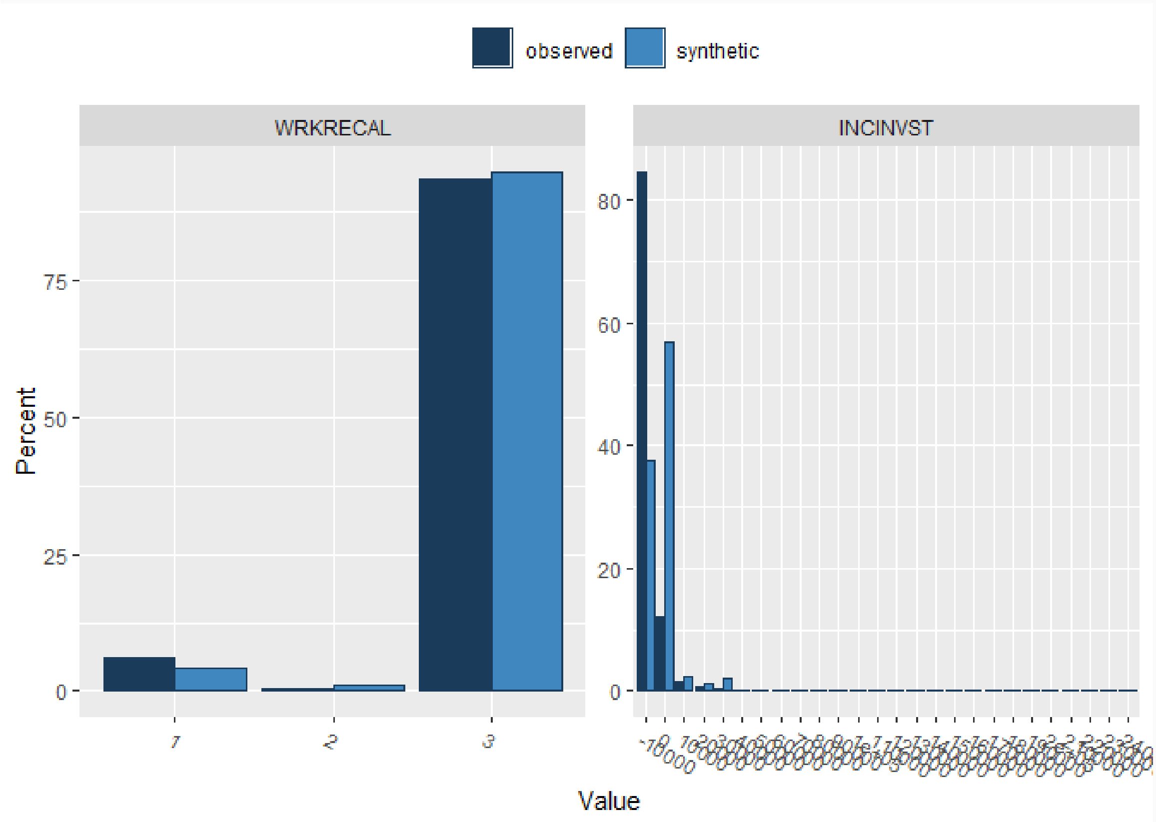

observed
synthetic
HHWT
MARST
HCOVANY
EDUC
Percent
Value


observed
synthetic
ABSENT
INCTOT
POVERTY
Percent
Value


observed
synthetic
GQ
RACE
HCOVPRIV
EMPSTAT
Percent
Value

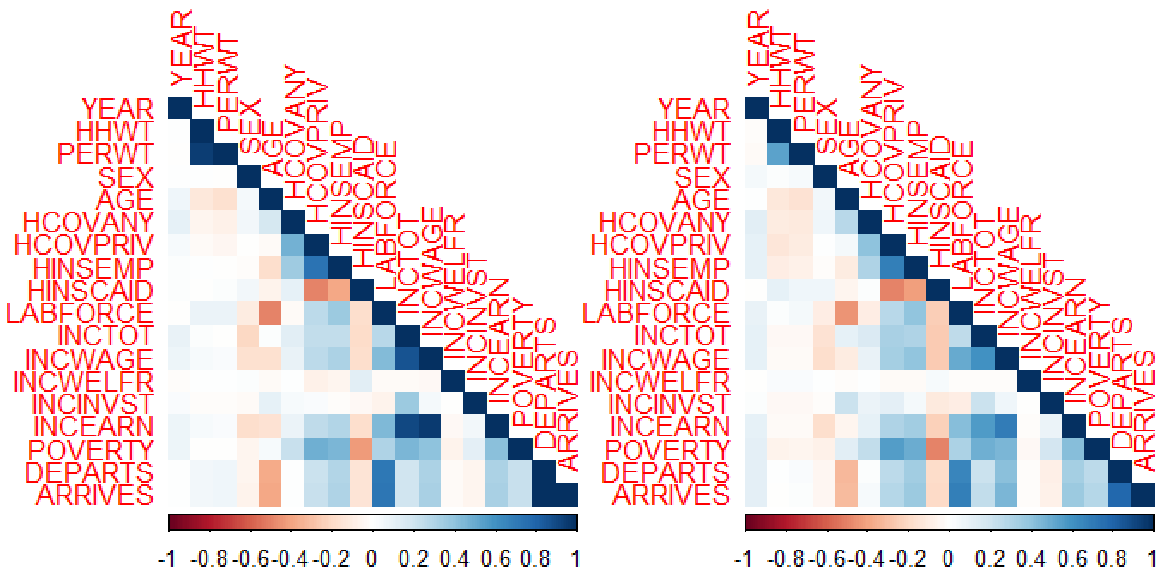

observed
synthetic
LOOKING
INCWAGE
DEPARTS
Percent
Value





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	0.0135632	37441.701	6
AGE	0.0116103	48076.115	4
SPEAKENG	0.0107432	44485.286	4
HINSCARE	0.0001238	2050.923	1
WRKLSTWK	0.0044304	36690.457	2
WORKEDYR	0.0055422	45898.084	2
INCEARN	0.0332500	137681.705	4

pMSE	S_pMSE
0.1371639	140.9911

	pMSE	S_pMSE	df
HHWT	0.0031583	13077.92	4

	pMSE	S_pMSE	df
MARST	0.0121554	40266.50	5
HCOVANY	0.0021357	35373.25	1
EDUC	0.0060960	10097.02	10
ABSENT	0.0003480	2882.38	2
INCTOT	0.0035312	14622.03	4
POVERTY	0.0043482	18005.02	4

pMSE	S_pMSE
0.0808016	57.79566

	pMSE	S_pMSE	df
GQ	0.0018455	7641.756	4
RACE	0.0050582	10472.443	8
HCOVPRIV	0.0052671	87240.296	1
EMPSTAT	0.0029431	24373.495	2
LOOKING	0.0013935	11540.811	2
INCWAGE	0.0335343	138858.787	4
DEPARTS	0.0097650	53913.483	3

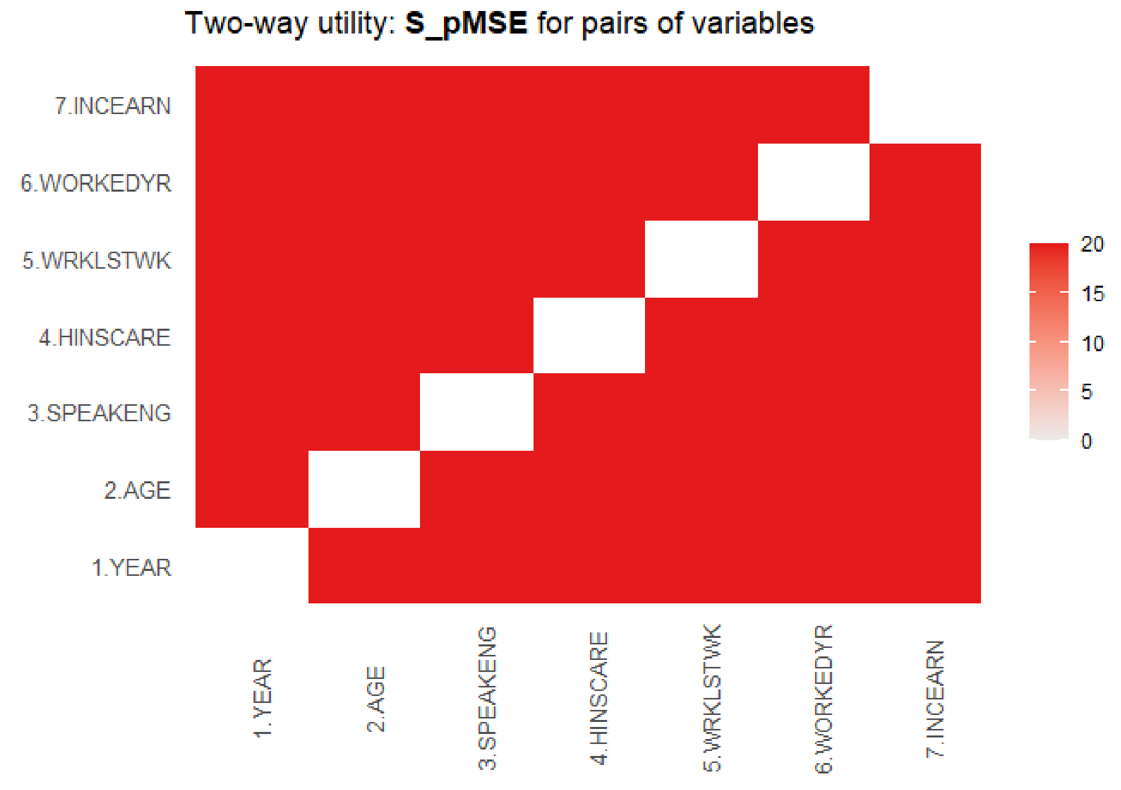
pMSE	S_pMSE
0.1917524	321.7915

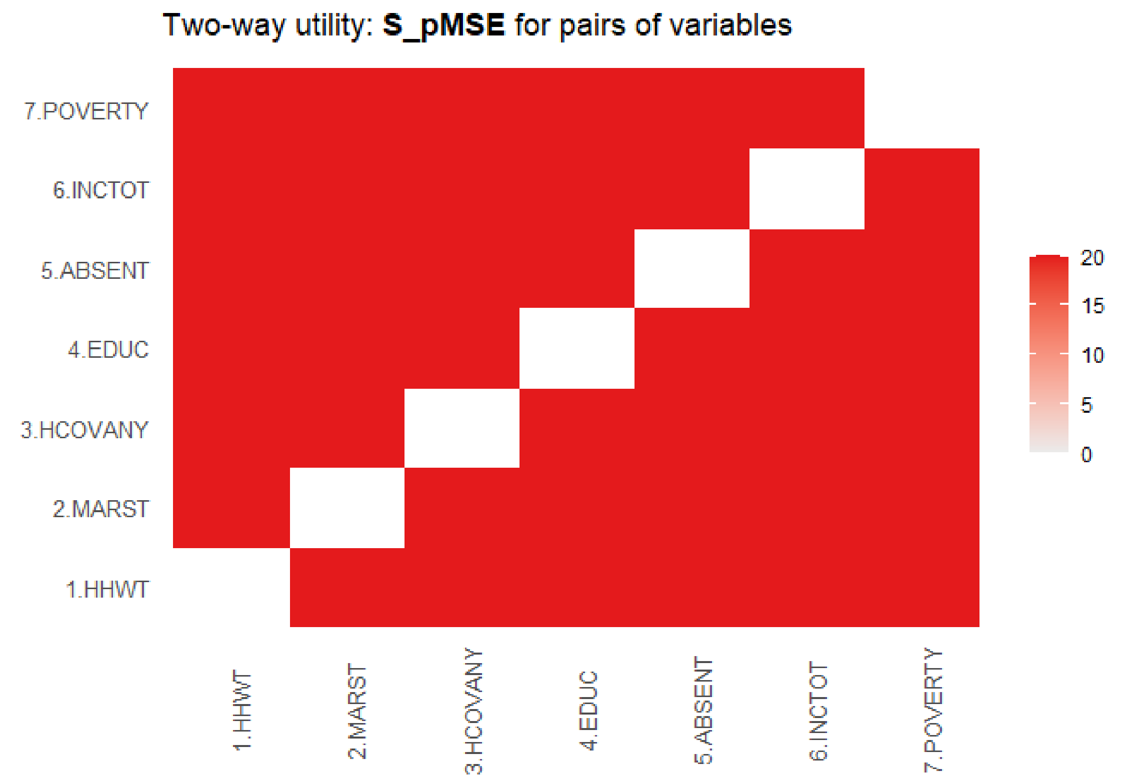
	pMSE	S_pMSE	df
SEX	0.0000001	9.521136e-01	1
CITIZEN	0.0046015	2.540531e+04	3
HINSCAID	0.0044901	7.437076e+04	1
LABFORCE	0.0009779	1.619639e+04	1
WRKRECAL	0.0007581	6.278456e+03	2
INCINVST	0.0553913	4.587289e+05	2

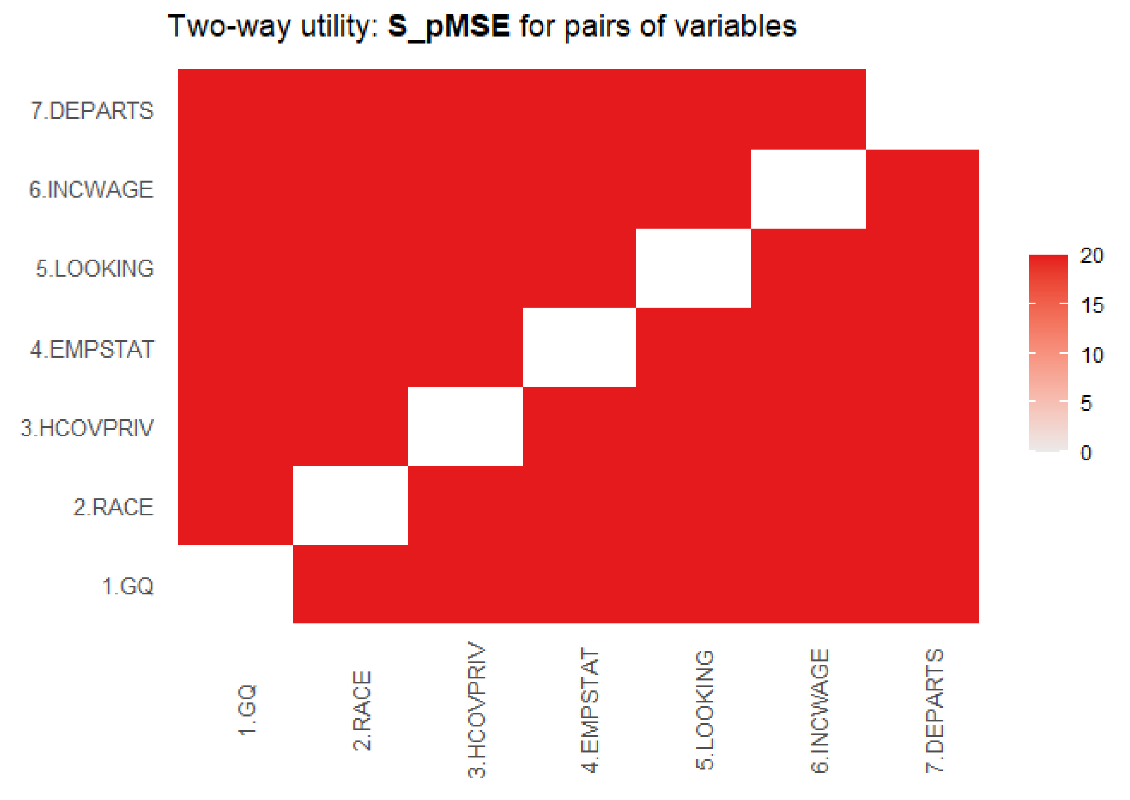
pMSE	S_pMSE
0.1884497	490.0153

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

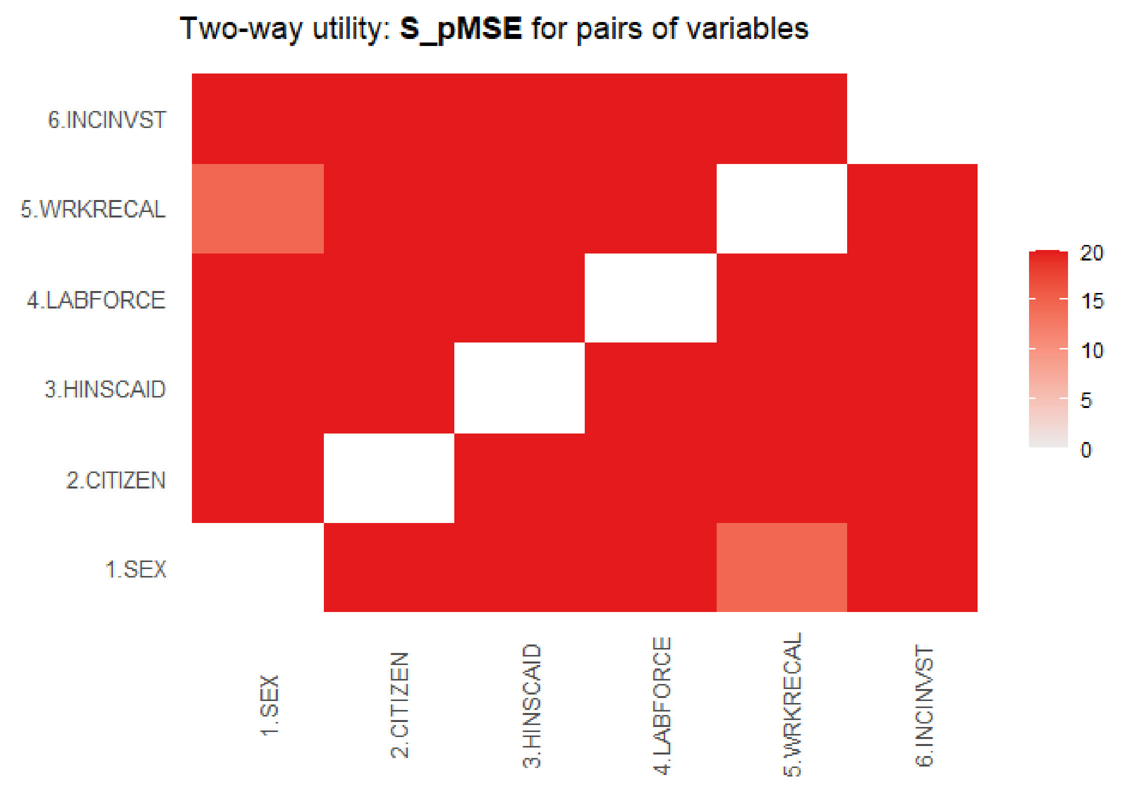
Two-way tables are evaluated based on the original and the synthetic dataset based on S_pMSE (see above). We also present the results for the mean absolute difference in densities (MabsDD) and the Bhattacharyya distance (dBhatt).

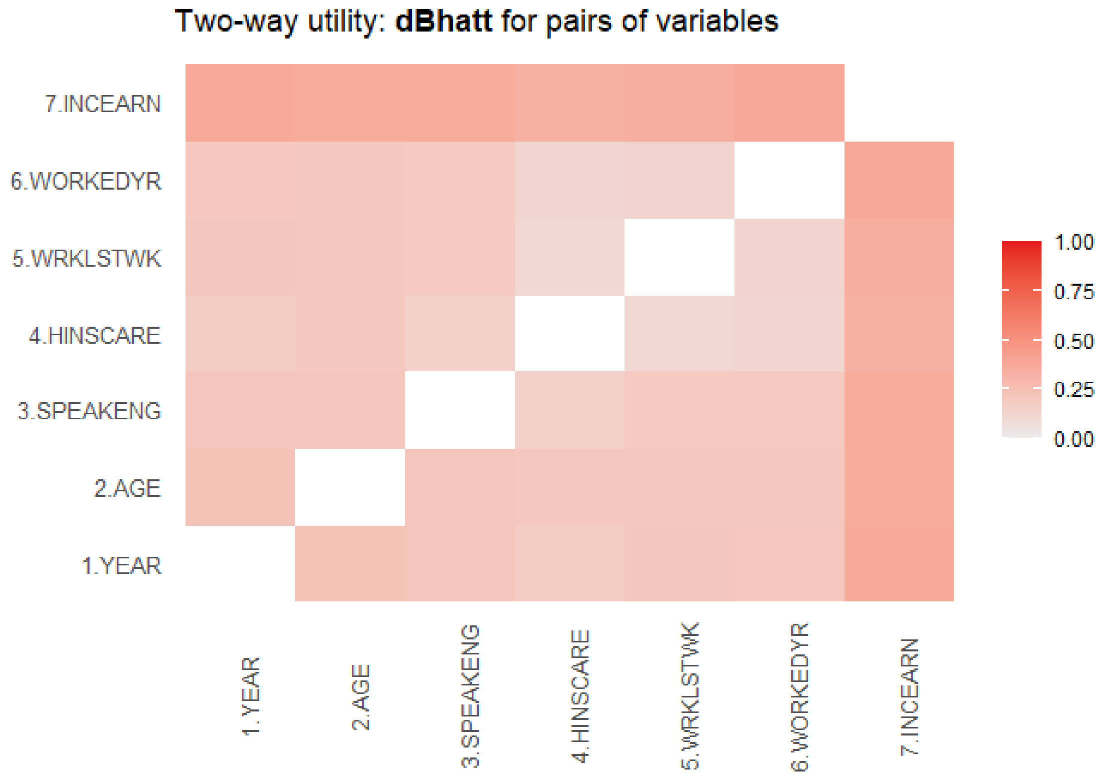


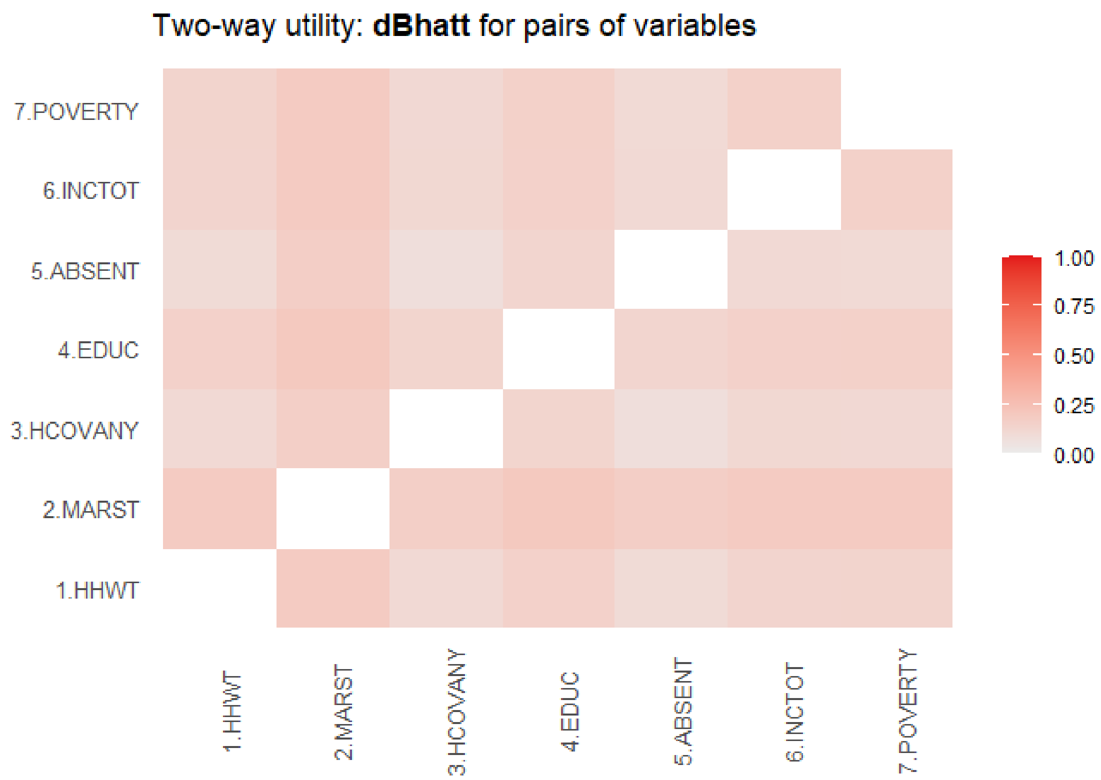


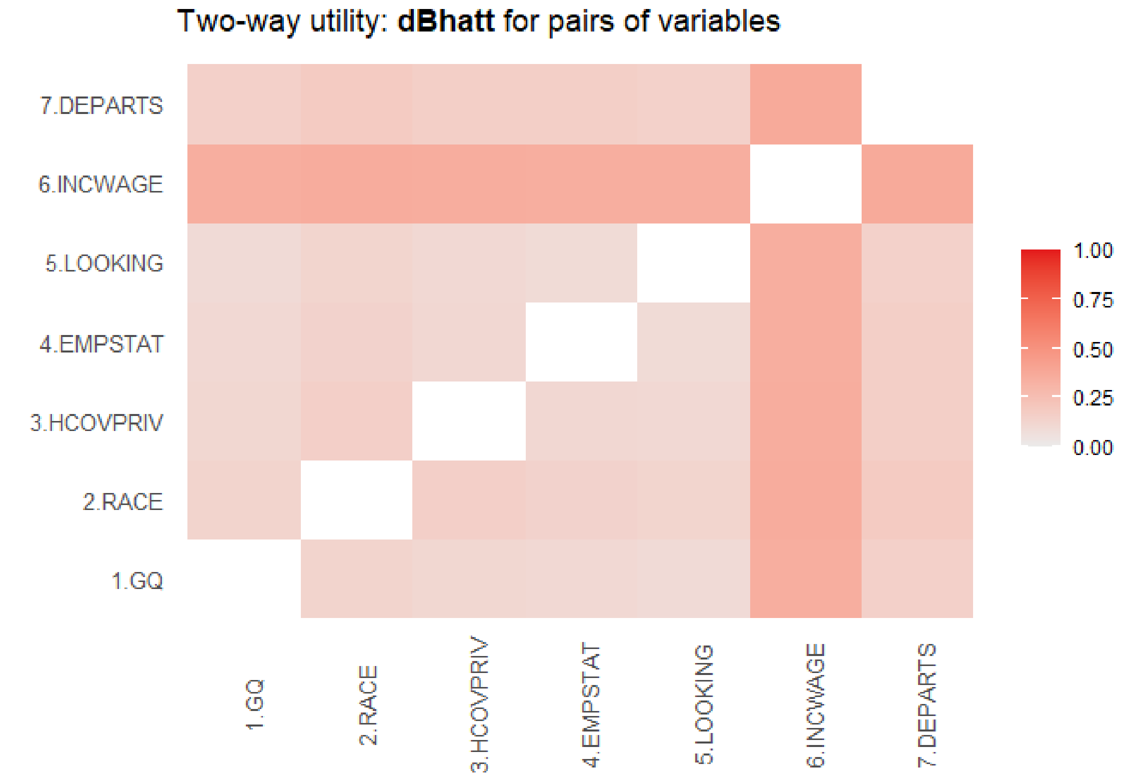

Two-way utility: S_pMSE for pairs of variables
7.DEPARTS
6.INCWAGE
5.LOOKING
4.EMPSTAT
3.HCOVPRIV
2.RACE
1.GQ
1.GQ
2.RACE
3.HCOVPRIV
4.EMPSTAT
5.LOOKING
6.INCWAGE
7.DEPARTS
20
15
10
5
0

```
## NULL
```

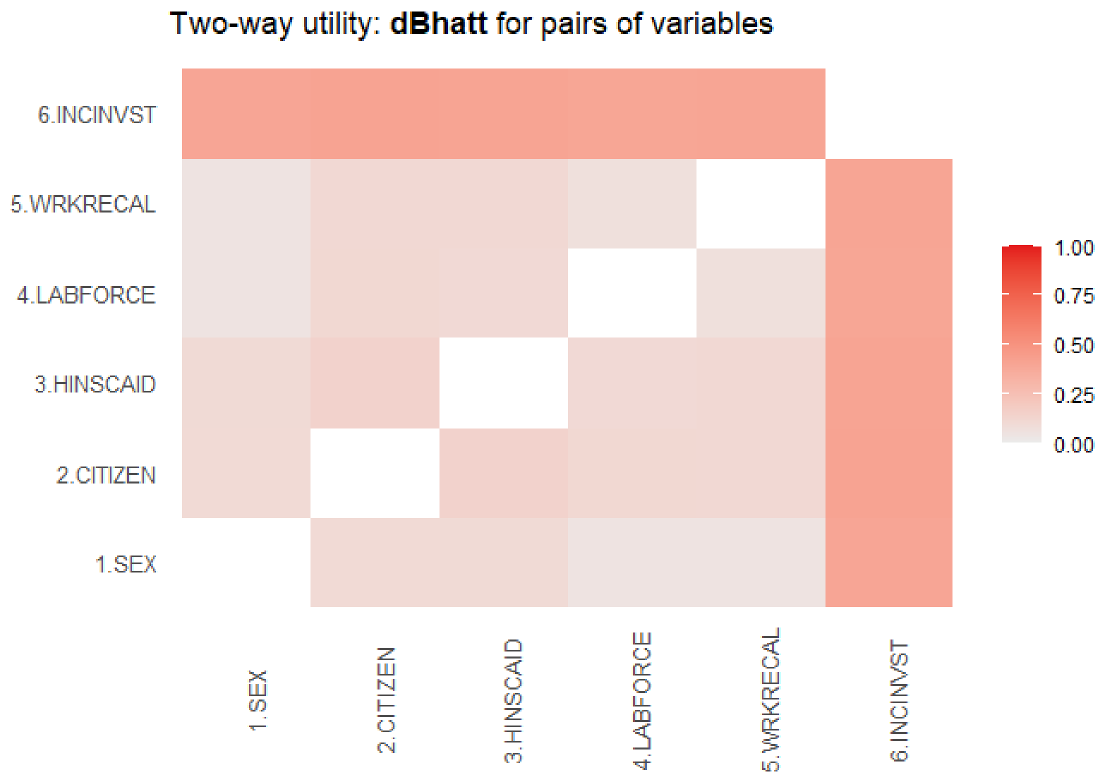

Two-way utility: S_pMSE for pairs of variables
6.INCINVST
5.WRKRECAL
4.LABFORCE
3.HINSCAID
2.CITIZEN
1.SEX
1.SEX
2.CITIZEN
3.HINSCAID
4.LABFORCE
5.WRKRECAL
6.INCINVST
20
15
10
5
0

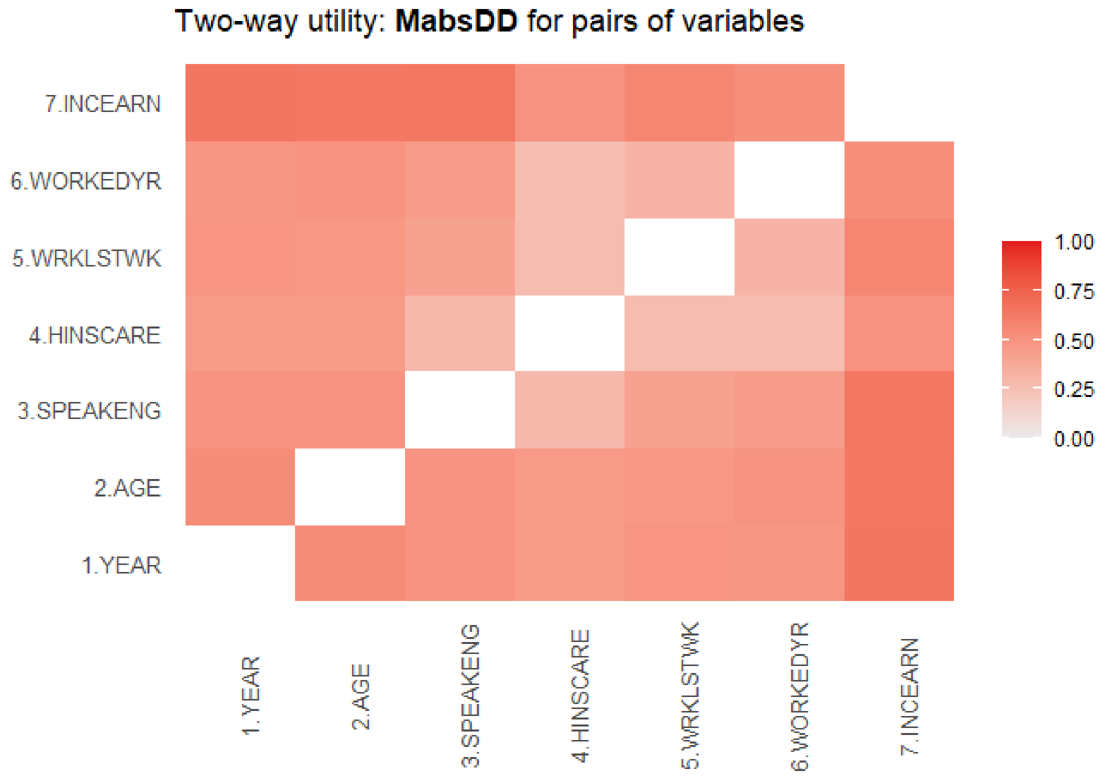

Two-way utility: dBhatt for pairs of variables
7.INCEARN
6.WORKEDYR
5.WRKLSTWK
4.HINSCARE
3.SPEAKENG
2.AGE
1.YEAR
1.YEAR
2.AGE
3.SPEAKENG
4.HINSCARE
5.WRKLSTWK
6.WORKEDYR
7.INCEARN
1.00
0.75
0.50
0.25
0.00

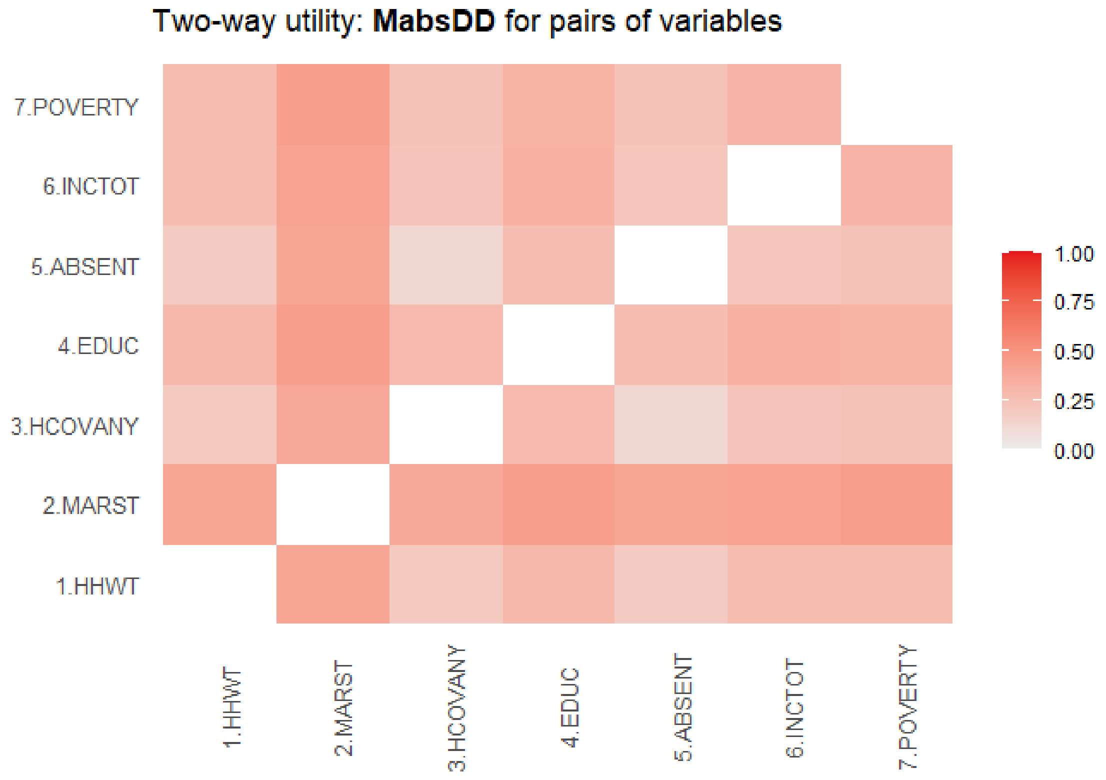

Two-way utility: dBhatt for pairs of variables
7.POVERTY
6.INCTOT
5.ABSENT
4.EDUC
3.HCOVANY
2.MARST
1.HHWT
1.HHWT
2.MARST
3.HCOVANY
4.EDUC
5.ABSENT
6.INCTOT
7.POVERTY
1.00
0.75
0.50
0.25
0.00

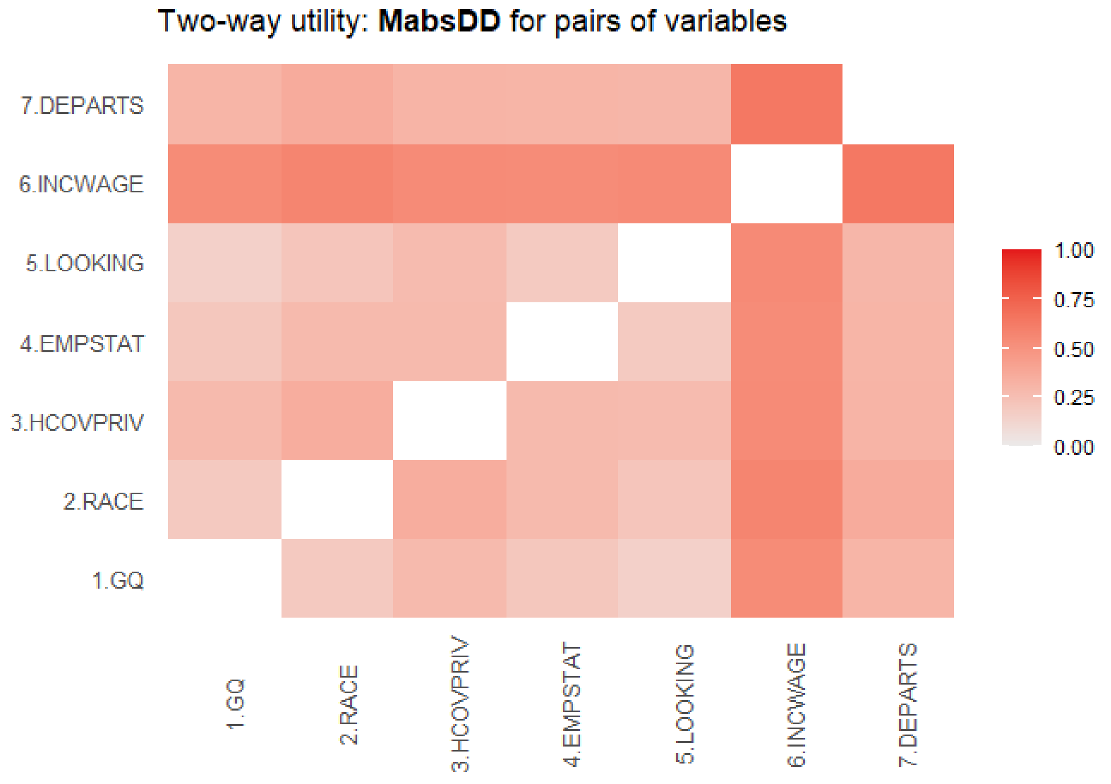

Two-way utility: dBhatt for pairs of variables
7.DEPARTS
6.INCWAGE
5.LOOKING
4.EMPSTAT
3.HCOVPRIV
2.RACE
1.GQ
1.00
0.75
0.50
0.25
0.00

```
## NULL
```

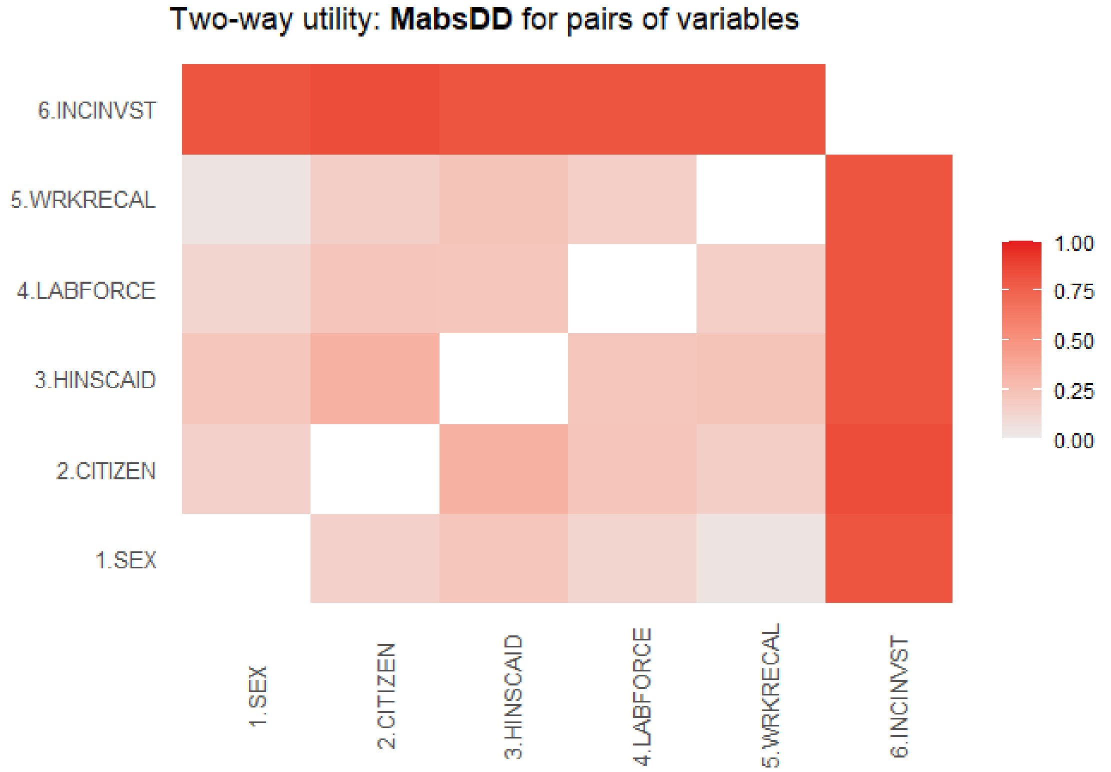

Two-way utility: dBhatt for pairs of variables
6.INCINVST
5.WRKRECAL
4.LABFORCE
3.HINSCAID
2.CITIZEN
1.SEX
1.00
0.75
0.50
0.25
0.00


Two-way utility: **MabsDD** 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: **MabsDD** 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




```
## 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
0.5616973

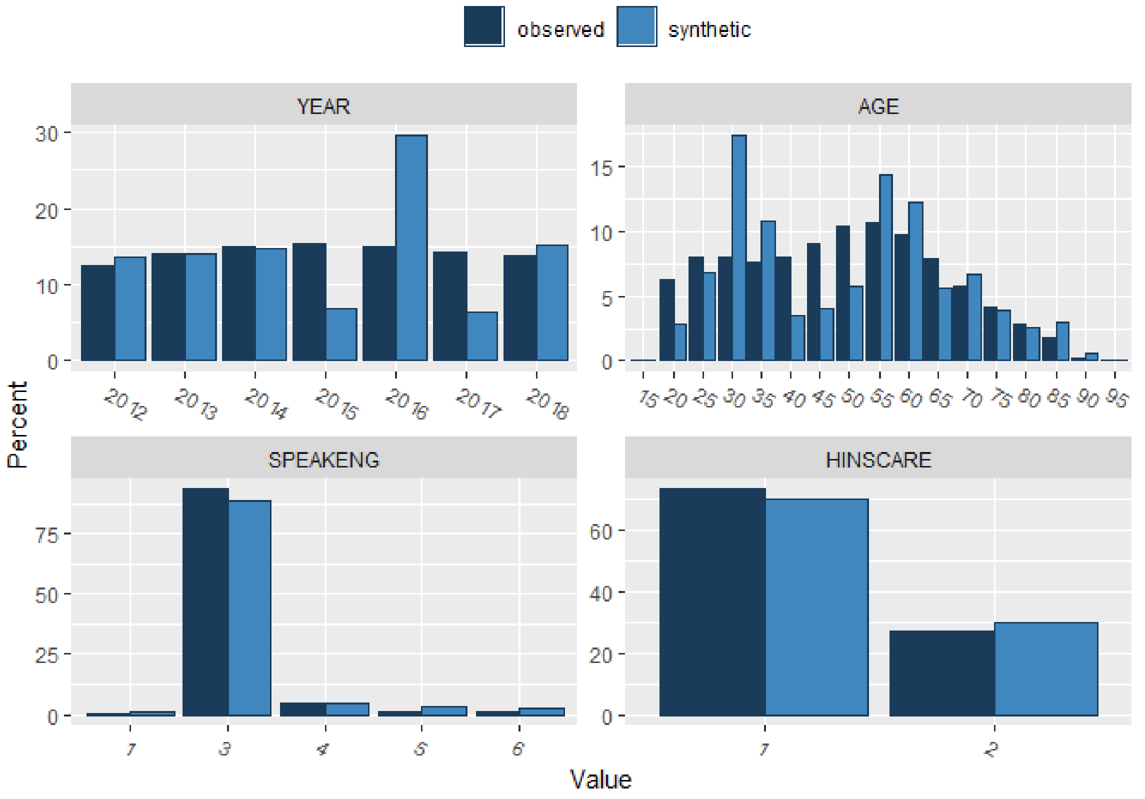
Individual Distances for Information Loss:

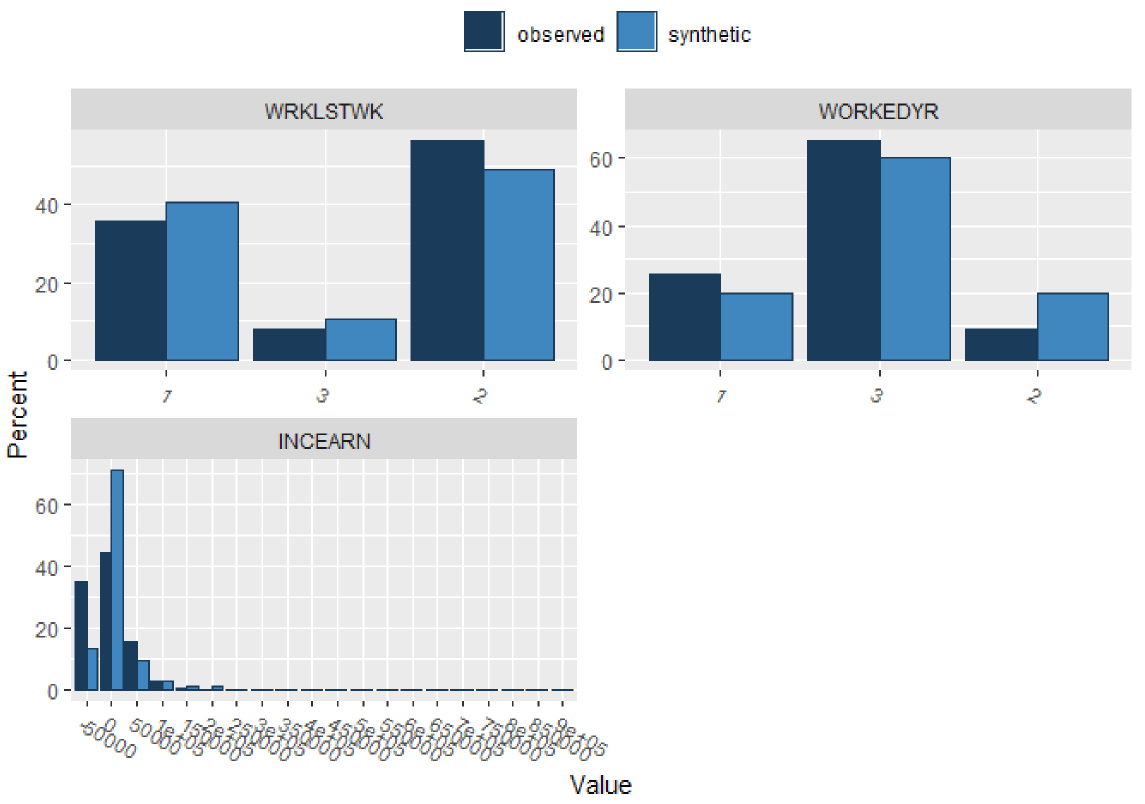
```
##        YEAR        HHWT          GQ       PERWT         SEX         AGE       MARST
## 0.85508418 0.95945851 0.09829975 0.95905994 0.49935230 0.91471230 0.69000996
##        RACE      HISPAN     CITIZEN    SPEAKENG     HCOVANY    HCOVPRIV     HINSEMP
## 0.27595704 0.11088185 0.17048380 0.25596575 0.18490612 0.45099937 0.50200492
##    HINSCAID    HINSCARE        EDUC     EMPSTAT    EMPSTATD    LABFORCE    WRKLSTWK
## 0.30127676 0.38379890 0.78324596 0.53886830 0.83006006 0.48095974 0.58594708
##      ABSENT     LOOKING    AVAILBLE    WRKRECAL    WORKEDYR      INCTOT     INCWAGE
## 0.49174122 0.52430687 0.26801945 0.11229703 0.55307327 0.99968486 0.99051450
##    INCWELFR    INCINVST     INCEARN     POVERTY     DEPARTS     ARRIVES
## 0.01535592 0.83495156 0.99219440 0.94552969 0.76413905 0.77456653
```

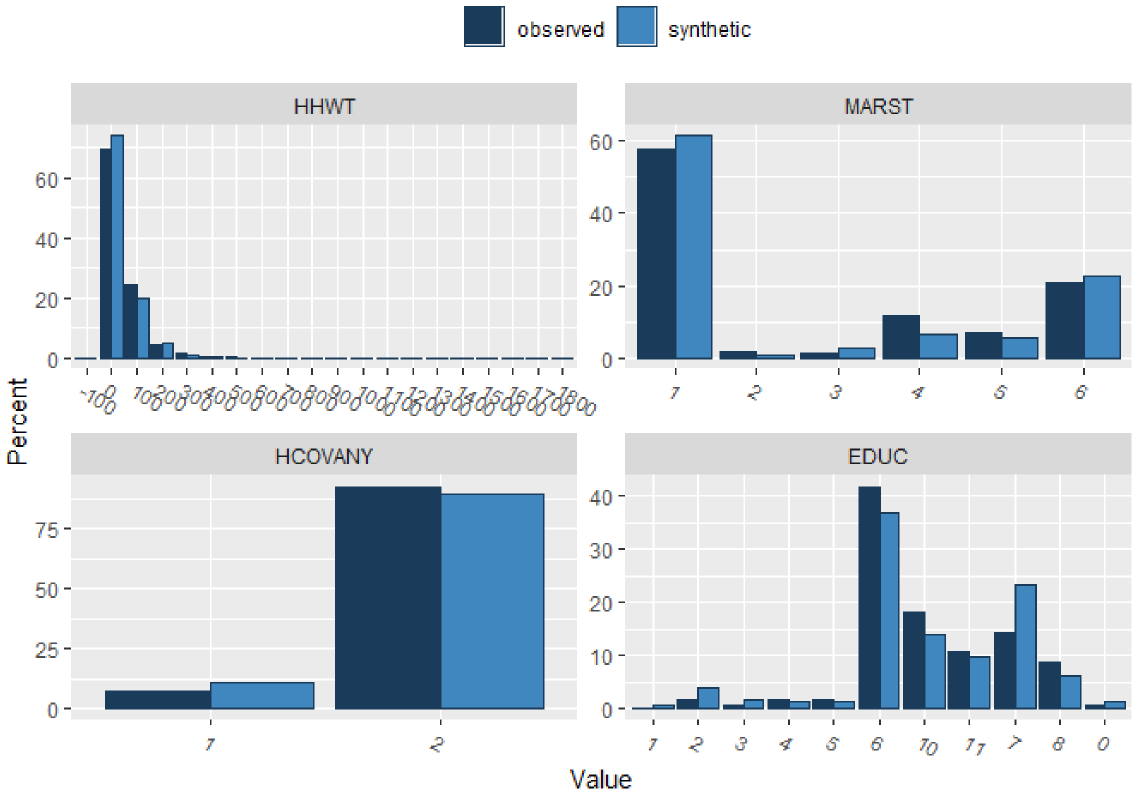
Tuning and Optimizations

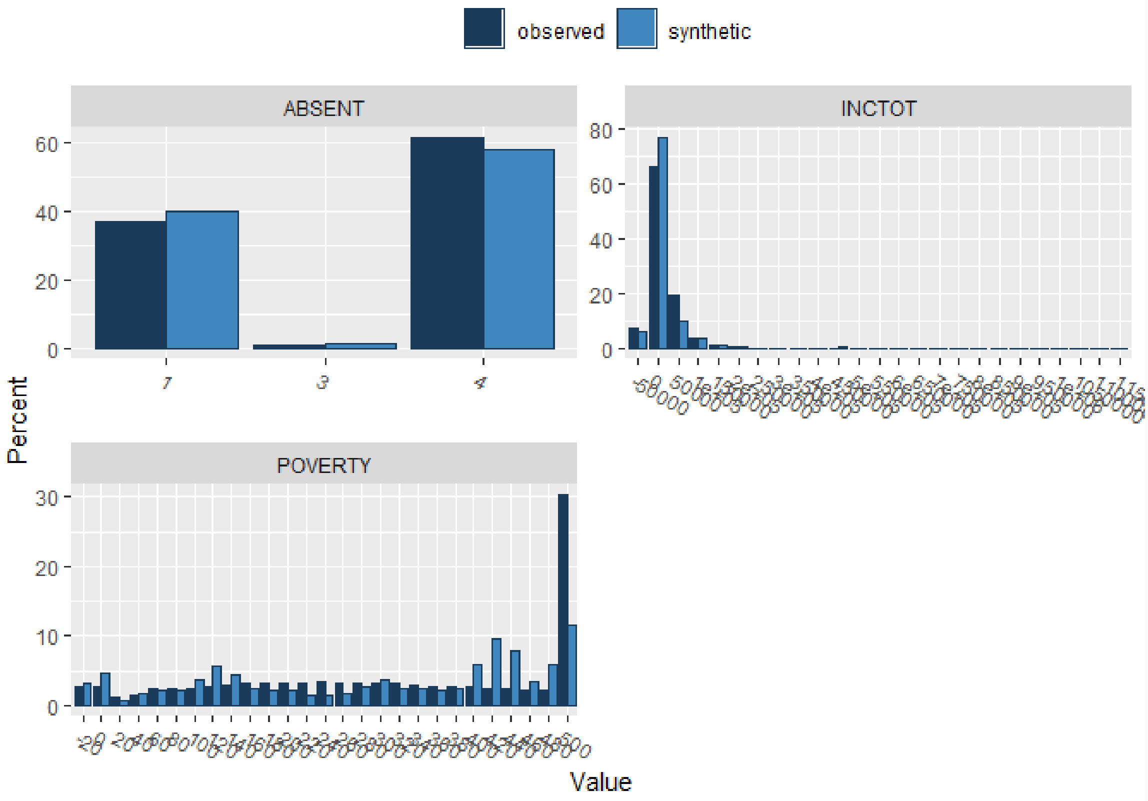
We also tried to optimize parameters and settings for the **GAN** methods on the ACS dataset. Our main problem here was our **limited computing time**. We tried using `CopulaGAN`, which we stopped (without result) after 8h computing time. Also for `ctgan` computing time was an issue. Our first try with `epochs = 10` only was of very limited utility. Increasing to `epochs = 30` for our final solution **increased usability** (still being on a rather low level). We assume we could have reached reasonable usability results with higher `epochs` values (we made good experiences with a value of 30 in the ACS dataset). Thus, with more time and computing resources parameters and results could probably be further improved. The privacy measures indicate a high level of privacy, which is not surprising considering their bad usability. So increasing `epochs` had no drawbacks on provacy measures.

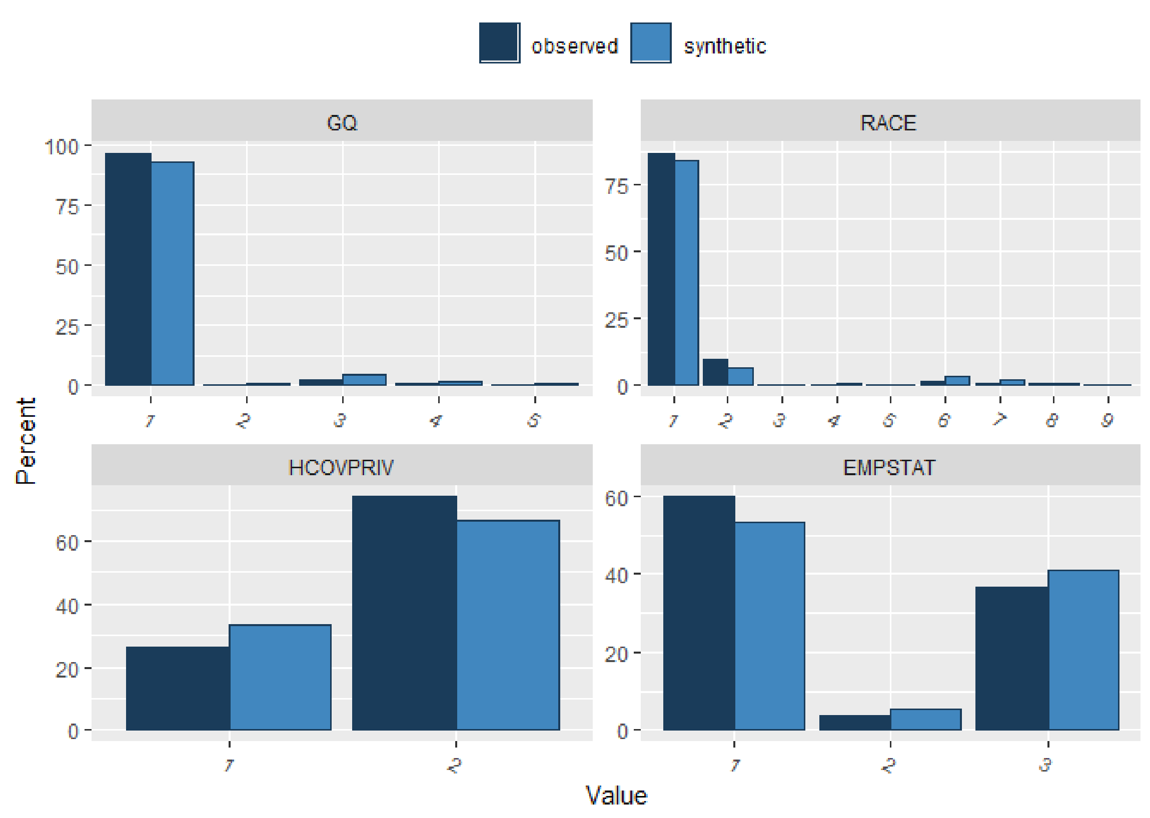
Here are some measures and plots for `epochs = 10`. As can be seen with lower usability results than our final model.

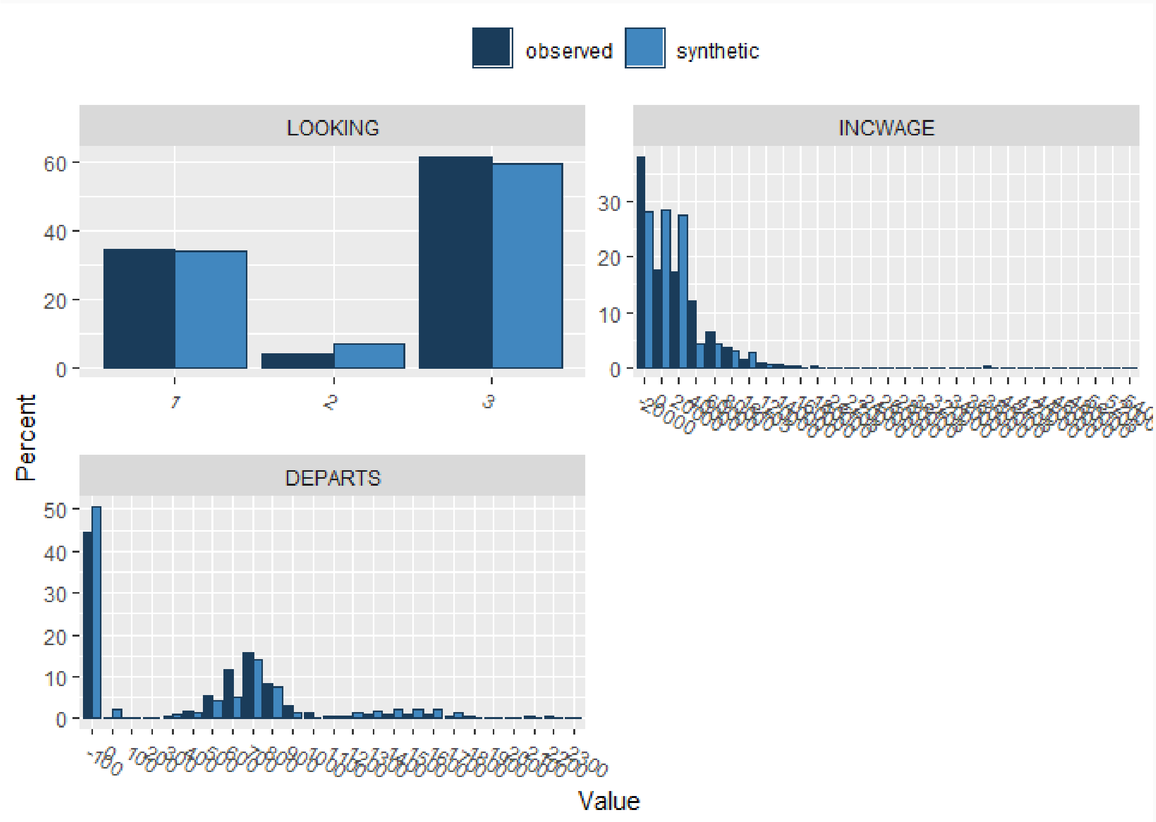

observed
synthetic
YEAR
AGE
SPEAKENG
HINSCARE
Percent
Value
2012
2013
2014
2015
2016
2017
2018
15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
1
3
4
5
6
1
2

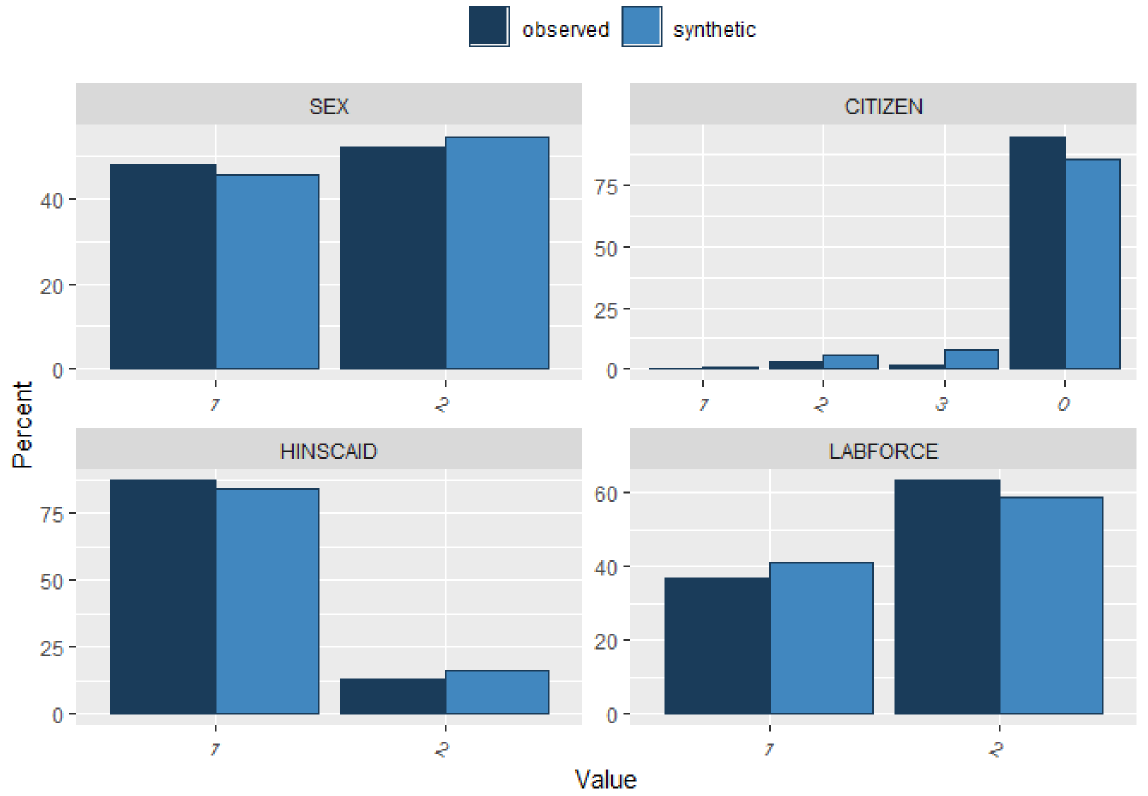

observed
synthetic
WRKLSTWK
WORKEDYR
INCEARN
Percent
Value
1
3
2
1
3
2

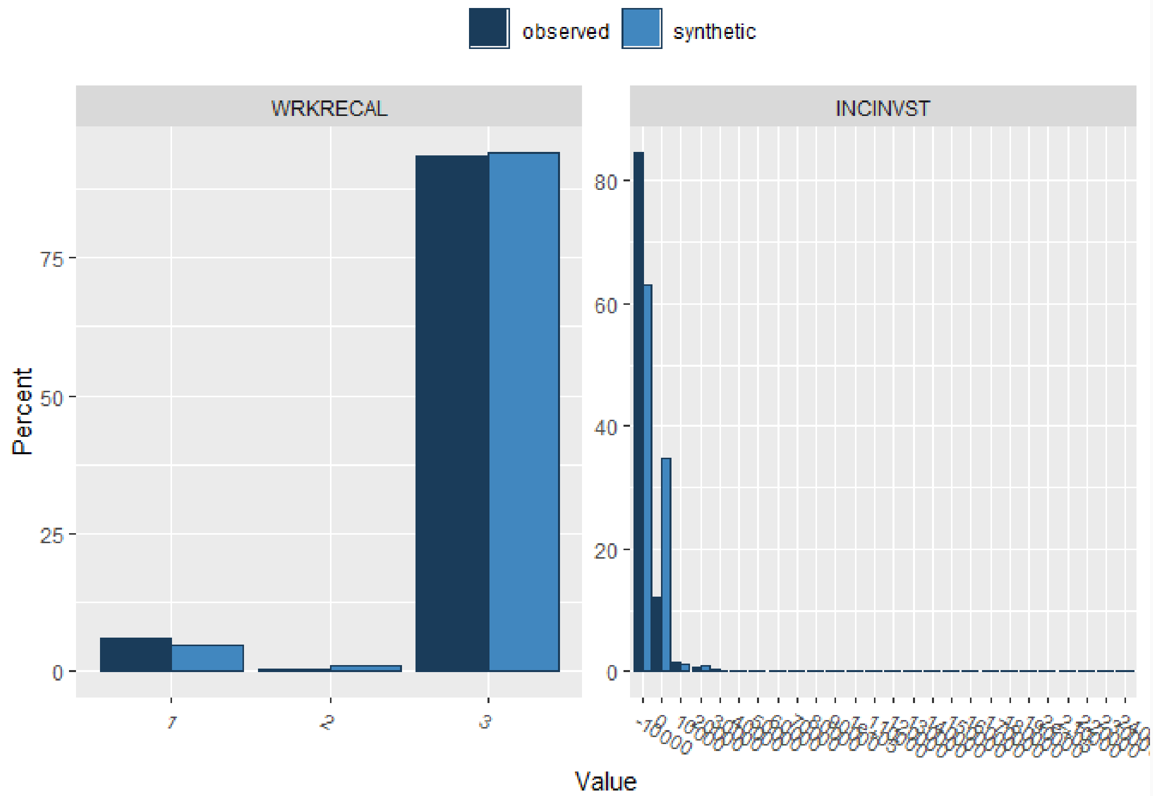

observed
synthetic
HHWT
MARST
HCOVANY
EDUC
Percent
Value


observed
synthetic
ABSENT
INCTOT
POVERTY
Percent
Value


observed
synthetic
GQ
RACE
HCOVPRIV
EMPSTAT
Percent
Value


observed
synthetic
LOOKING
INCWAGE
DEPARTS
Percent
Value





	pMSE	S_pMSE	df
YEAR	0.0142379	39304.26	6
AGE	0.0049334	20428.18	4
SPEAKENG	0.0027034	11194.42	4
HINSCARE	0.0002626	4349.33	1
WRKLSTWK	0.0015273	12648.37	2

	pMSE	S_pMSE	df
WORKEDYR	0.0059145	48981.88	2
INCEARN	0.0198820	82327.38	4

pMSE	S_pMSE
0.1378191	149.9058

	pMSE	S_pMSE	df
HHWT	0.0010559	4372.373	4
MARST	0.0029667	9827.479	5
HCOVANY	0.0007542	12491.534	1
EDUC	0.0068279	11309.130	10
ABSENT	0.0003845	3183.923	2
INCTOT	0.0095084	39372.292	4
POVERTY	0.0236237	97820.973	4

pMSE	S_pMSE
0.0771774	53.49455

	pMSE	S_pMSE	df
GQ	0.0024415	10109.949	4
RACE	0.0042393	8777.125	8
HCOVPRIV	0.0015846	26245.730	1
EMPSTAT	0.0013919	11526.775	2
LOOKING	0.0010216	8460.084	2
INCWAGE	0.0519297	215030.842	4
DEPARTS	0.0047654	26310.121	3

pMSE	S_pMSE
0.1537748	223.8978

	pMSE	S_pMSE	df
SEX	0.0001612	2670.187	1
CITIZEN	0.0066649	36797.356	3
HINSCAID	0.0005099	8445.848	1
LABFORCE	0.0005142	8516.440	1

	pMSE	S_pMSE	df
WRKRECAL	0.0004557	3773.945	2
INCINVST	0.1216055	671392.777	3

pMSE	S_pMSE
0.1835097	563.0436

ACS - Probabilistic Graphical Models (Minutemen DP-pgm)

Evaluation Synthetic Data Creation

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

January 31, 2022

Executive Summary

We used **minutemen** from the provided python scripts. As you could expect from the **second place submission** of 2021 NIST differential privacy, minutemen scored good in our main privacy metrics. Out of all different methods we tested (`FCS`, `IPSO`, `GAN`, `Simulation`, `Minutemen`) together with the simulation approach it was the **best method in terms of privacy**. Unfortunately, the minutemen method could not produce useful synthetic data based on ACS as well in our case. There is **barely any measure that would indicate a high utility**.

Looking at utility measures is not actually motivating. The `S_pMSE` for tables and for distributions is very high. The Pearson correlation coefficients for binary and (semi-)continuous are in many cases practically zero. The absolute difference in densities and the Bhattacharyya distance show extreme results. There is **no reasonable utility** according to Mlodak's information loss criterion. From a privacy perspective the minutemen looks quite good (also when looking at more detailed metrics).

USE CASE RECOMMENDATIONS

Releasing_to_Public	Testing_Analysis	Education	Testing_Technology
NO	NO	NO	NO

The utility of **minutemen** has some flaws, thus we **don't** think it is a good idea to **release this data to the public**. This could lead to false impressions. Also scientists may be led to false conclusions when using this data for **testing analysis**. We could not imagine our minutemen generated data to be used for **education** or in **technology testing** because of the lost variable dependencies.

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

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	1035201	0	0

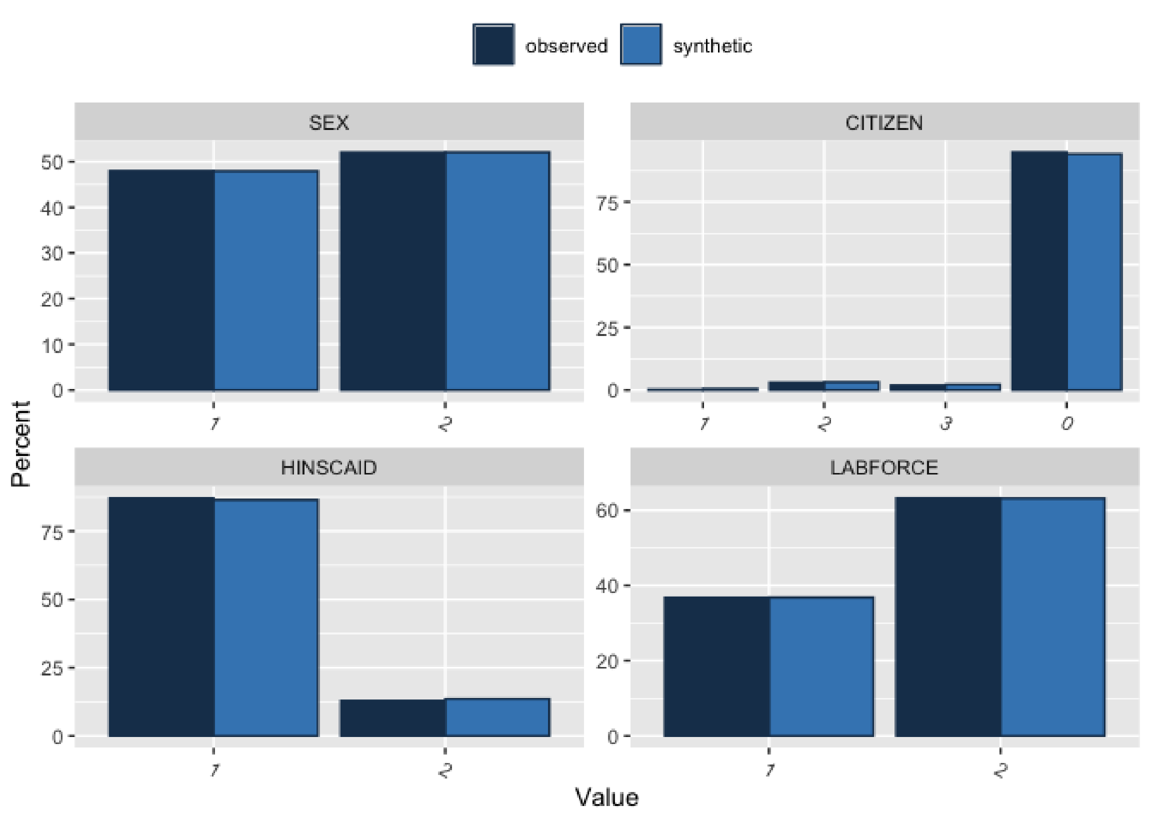
Utility Evaluation

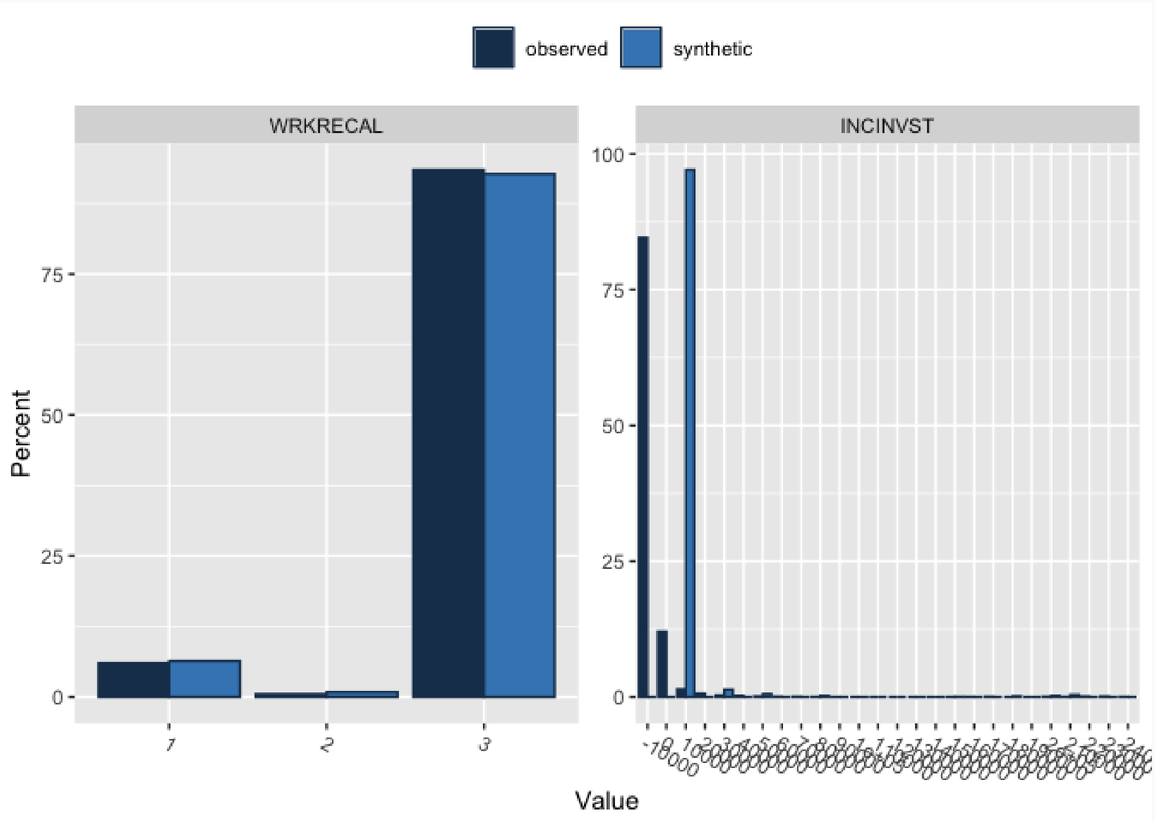
Different utility measures are applied in this section. These utility measures are the basis of utility evaluation for the generated synthetic dataset. The R packages synthpop, sdcMicro and corrplot were used to compute the following metrics. We do not use tests incorporating significance here.

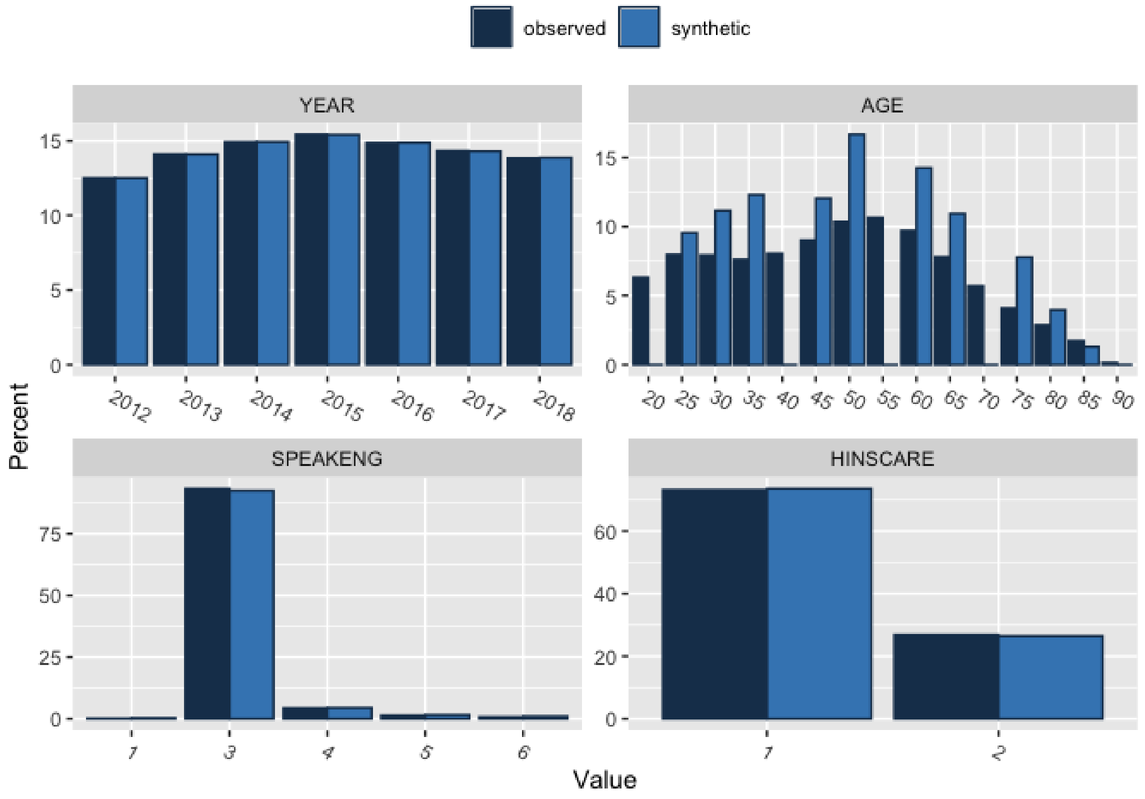
Confidence intervals in large surveys often tend to be extremely small so many slight differences appear to be significant. We do not consider the variable PUMA for our utility evaluation. During the ACS reports, some minor changes in availability regarding plots might occur. This is caused by the application of standardised scripts on different synthetic datasets.

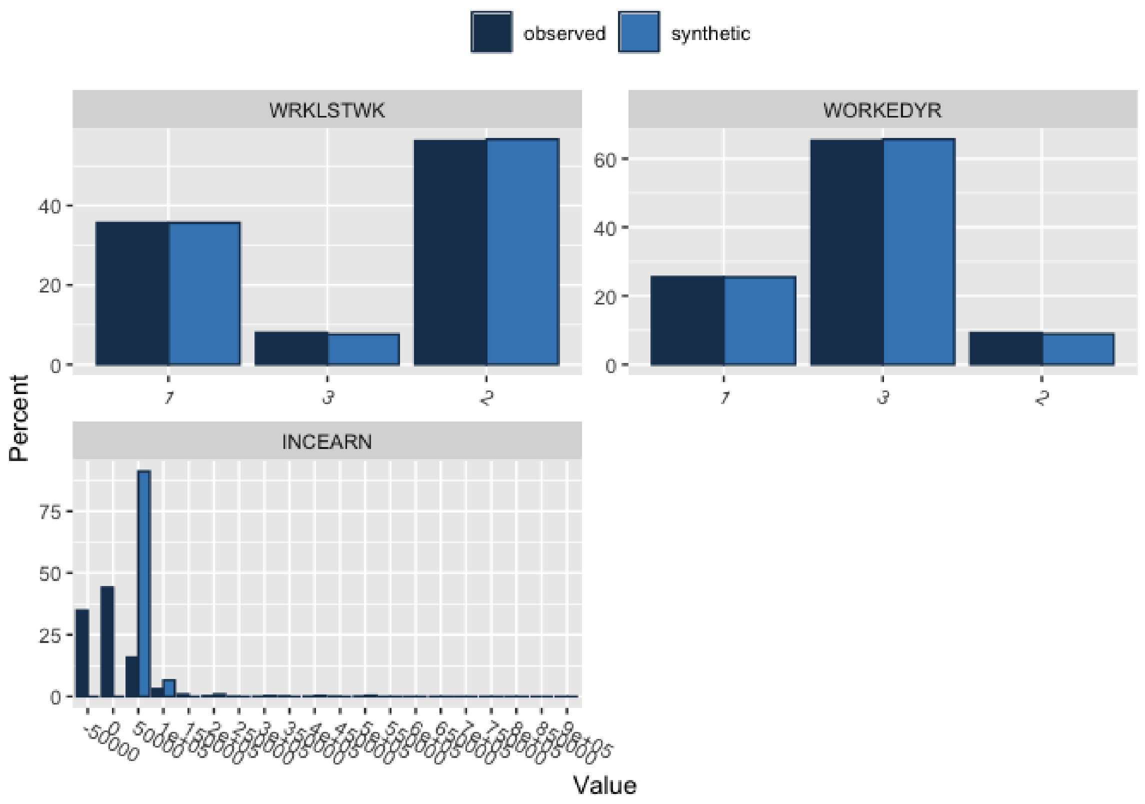
Graphical Comparison for Margins (R-Package: synthpop)

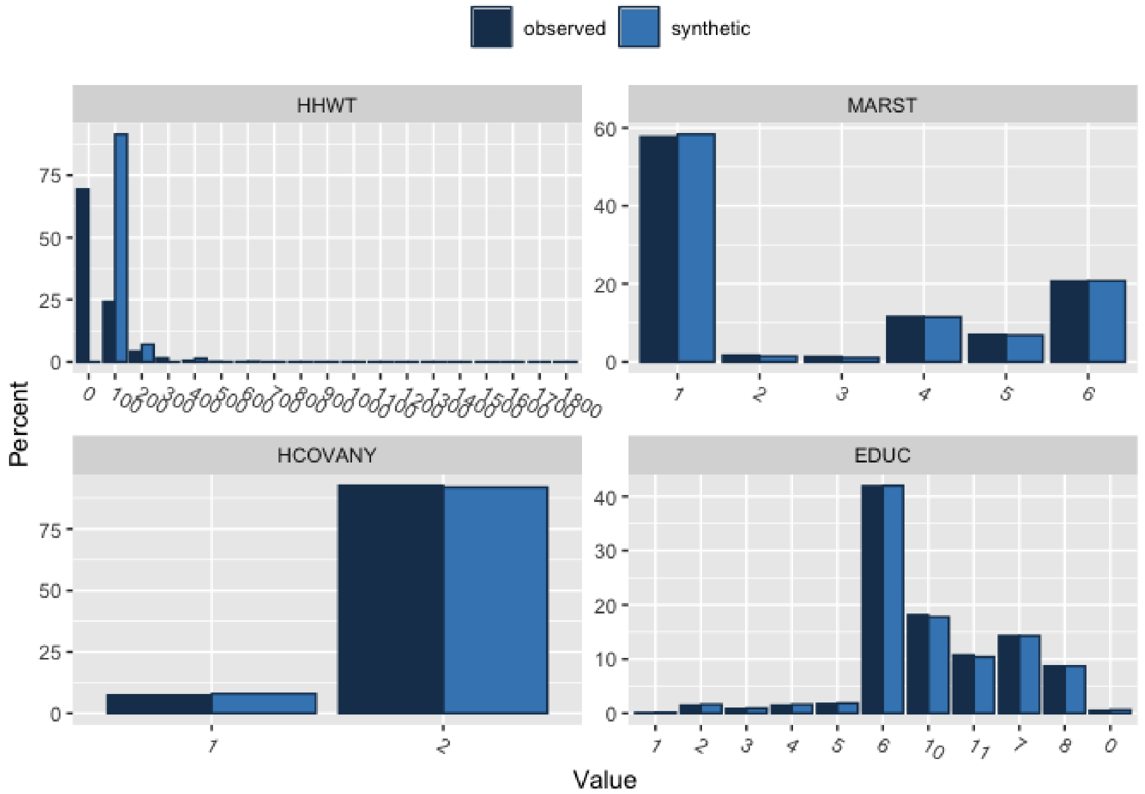
The following histograms provide an ad-hoc overview on the marginal distributions of the original and synthetic dataset. Matching or close distributions are related to a high data utility.

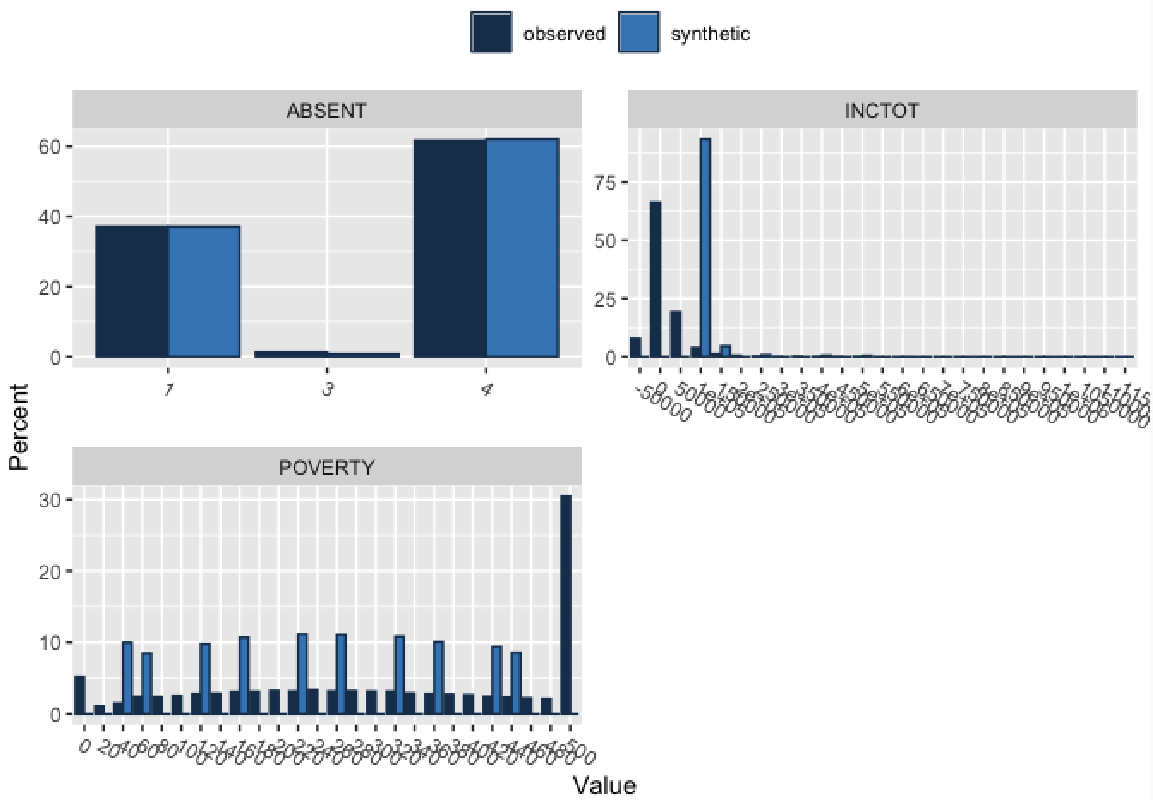


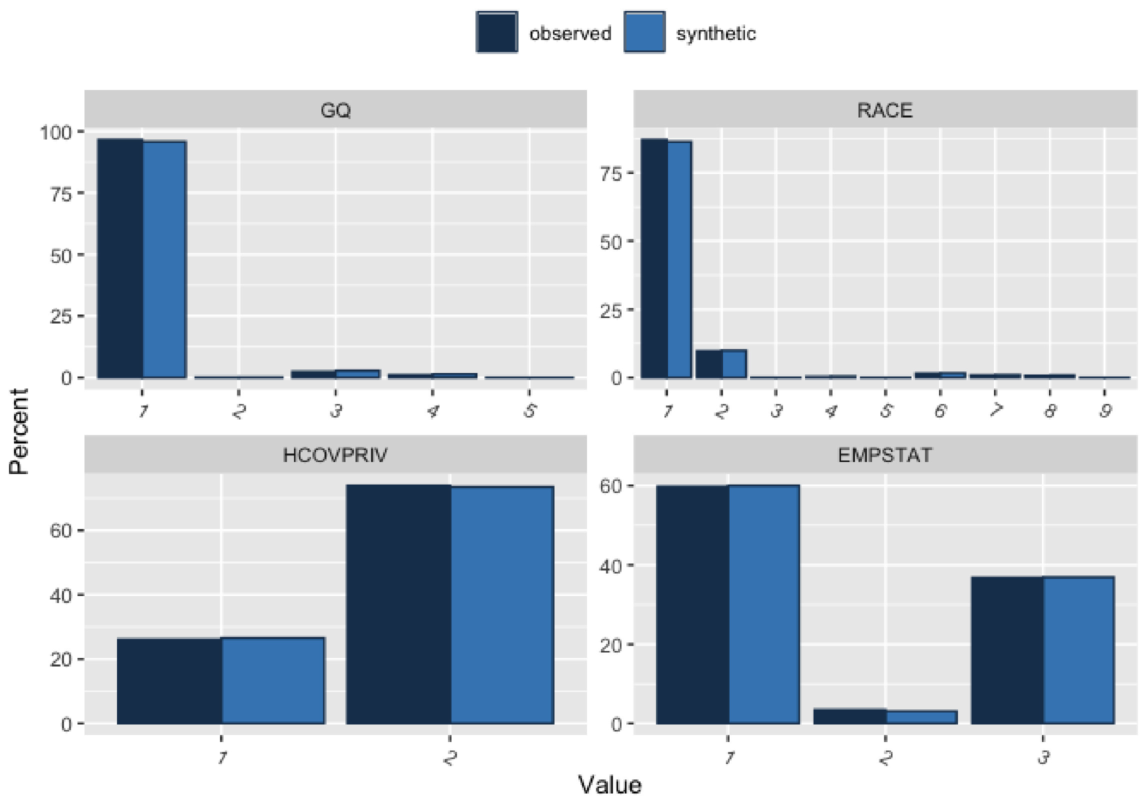


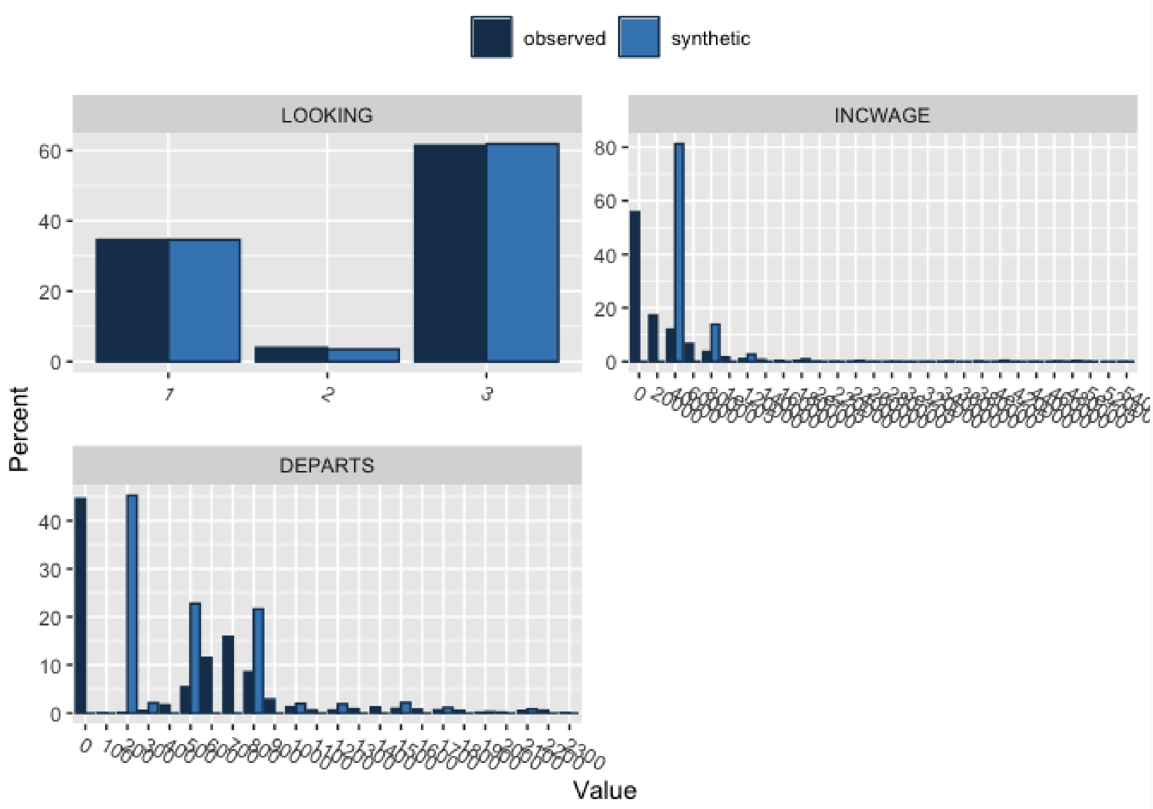

observed
synthetic
YEAR
AGE
SPEAKENG
HINSCARE
Percent
Value


observed
synthetic
WRKLSTWK
WORKEDYR
INCEARN
Percent
Value


observed
synthetic
HHWT
MARST
HCOVANY
EDUC
Percent
Value

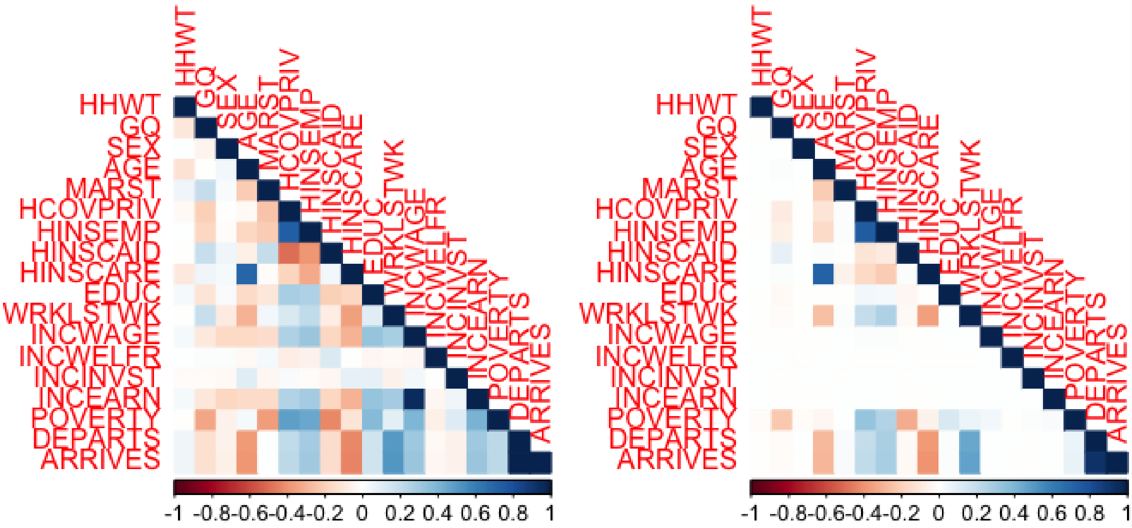

observed
synthetic
ABSENT
INCTOT
POVERTY
Percent
Value





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
SEX	0.0000000	5.102196e-01	1
CITIZEN	0.0001703	9.244680e+02	3
HINSCAID	0.0000276	4.492361e+02	1
LABFORCE	0.0000001	2.177926e+00	1
WRKRECAL	0.0001435	1.168322e+03	2
INCINVST	0.2395194	1.950079e+06	2

pMSE	S_pMSE
0.2484669	1626.453

	pMSE	S_pMSE	df
YEAR	0.0000001	3.018637e-01	6
AGE	0.0077622	3.159859e+04	4

	pMSE	S_pMSE	df
SPEAKENG	0.0001604	6.530851e+02	4
HINSCARE	0.0000045	7.382504e+01	1
WRKLSTWK	0.0000150	1.223837e+02	2
WORKEDYR	0.0000086	7.041004e+01	2
INCEARN	0.2106121	1.143151e+06	3

pMSE	S_pMSE
0.2485162	384.5451

	pMSE	S_pMSE	df
HHWT	0.2134683	1158653.3496	3
MARST	0.0000687	223.8566	5
HCOVANY	0.0000342	556.6046	1
EDUC	0.0000639	103.9779	10
ABSENT	0.0001302	1060.1785	2
INCTOT	0.2211937	1200585.2797	3
POVERTY	0.0288527	117453.8627	4

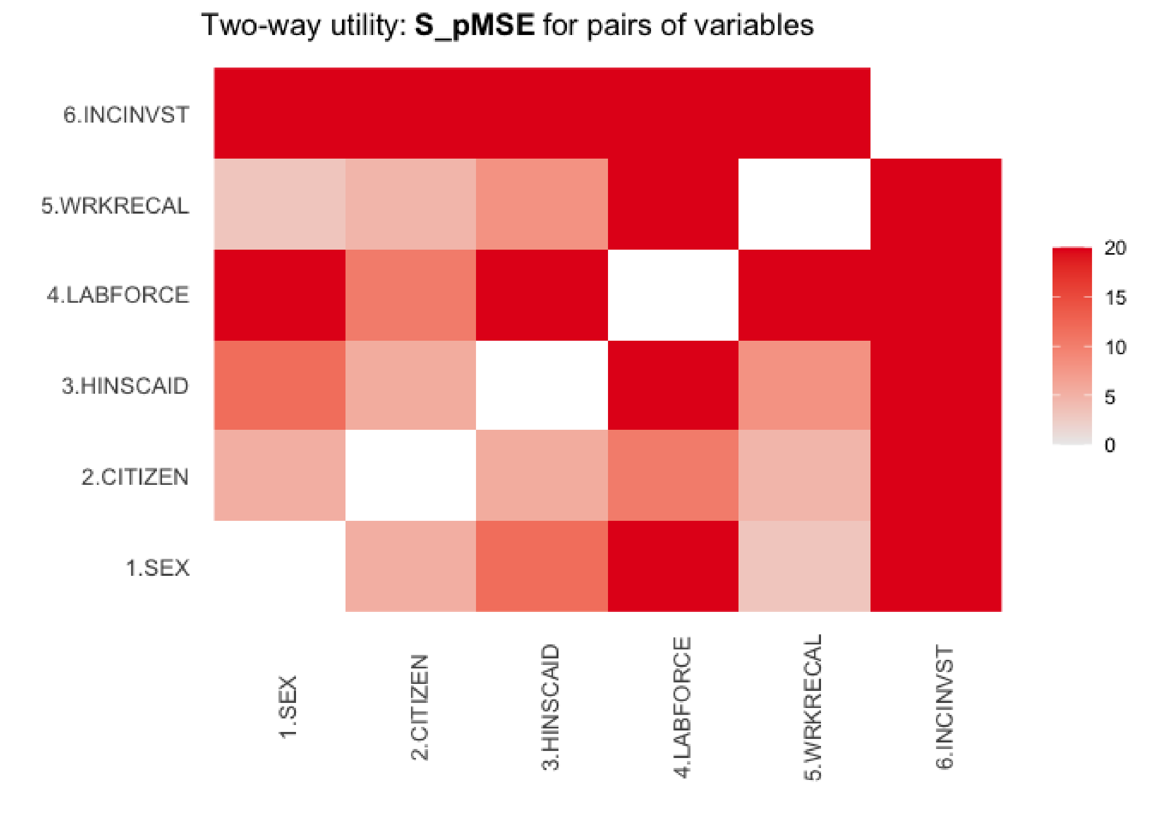
pMSE	S_pMSE
0.25	257.326

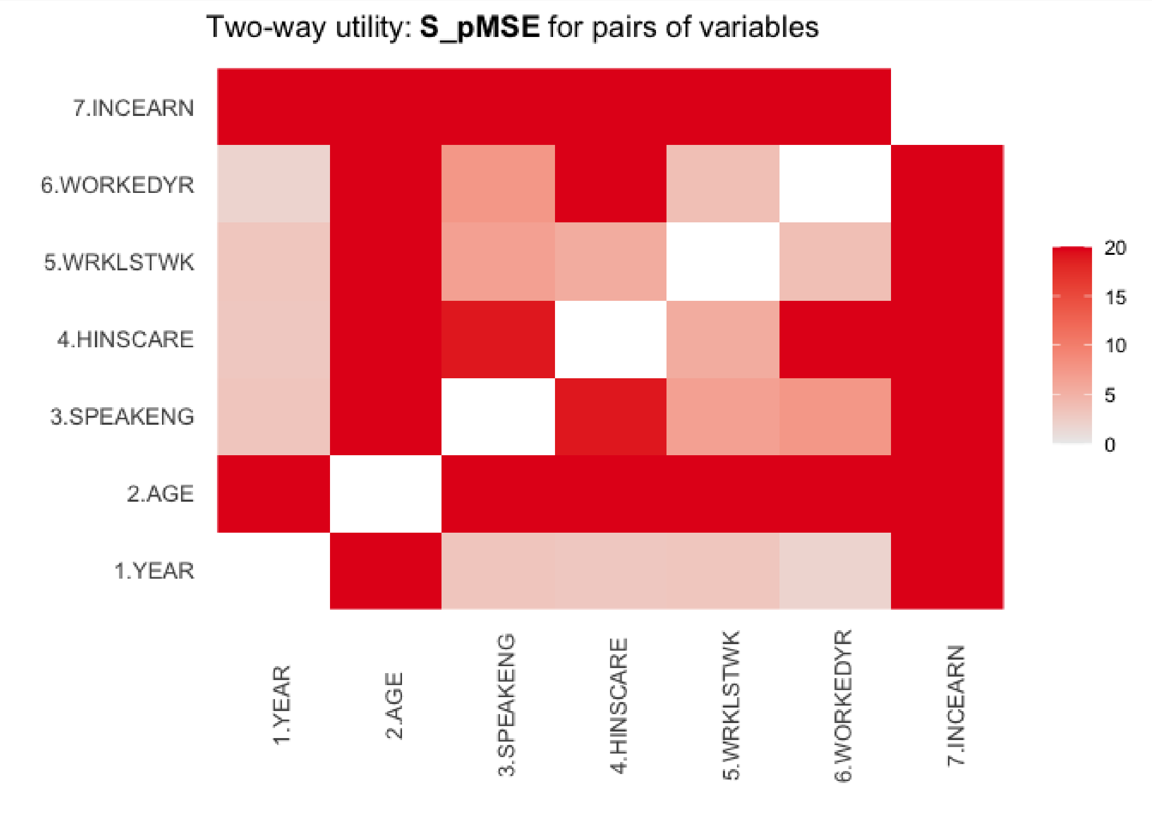
	pMSE	S_pMSE	df
GQ	0.0001484	604.0495	4
RACE	0.0001179	239.9770	8
HCOVPRIV	0.0000063	101.7801	1
EMPSTAT	0.0000283	230.0871	2
LOOKING	0.0000340	276.9484	2
INCWAGE	0.1710486	928409.9725	3
DEPARTS	0.1003127	544472.9242	3

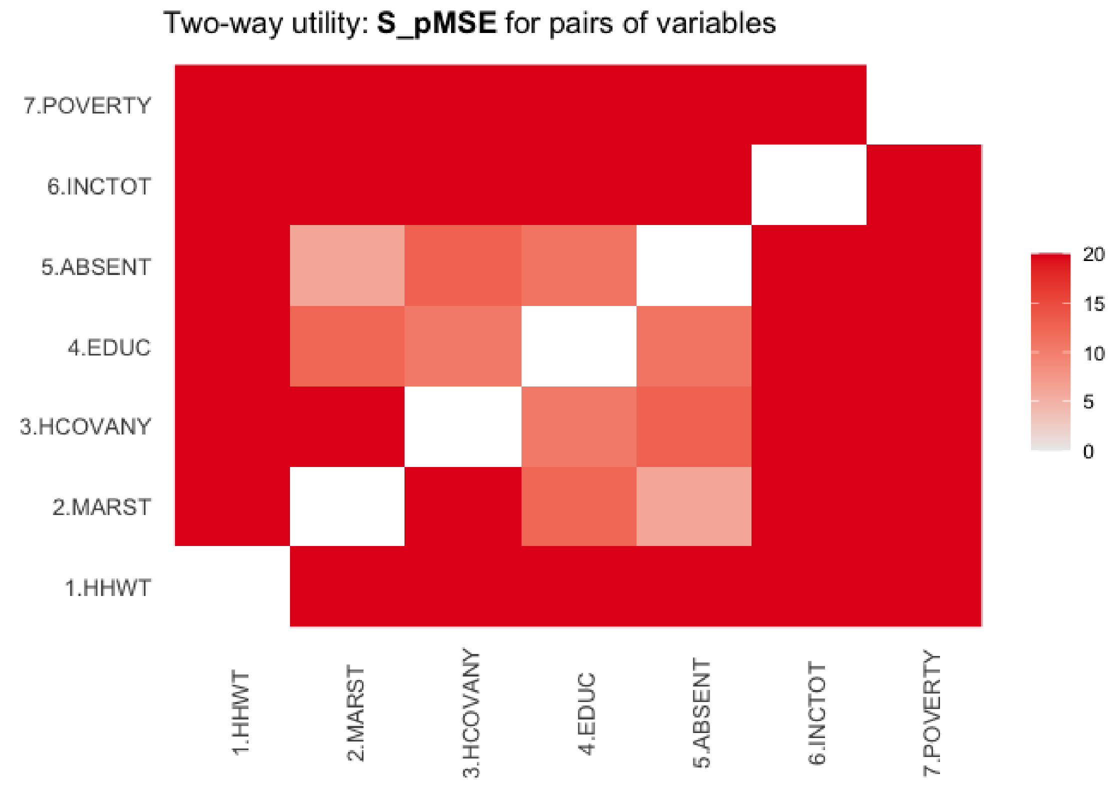
pMSE	S_pMSE
0.2487321	608.7562

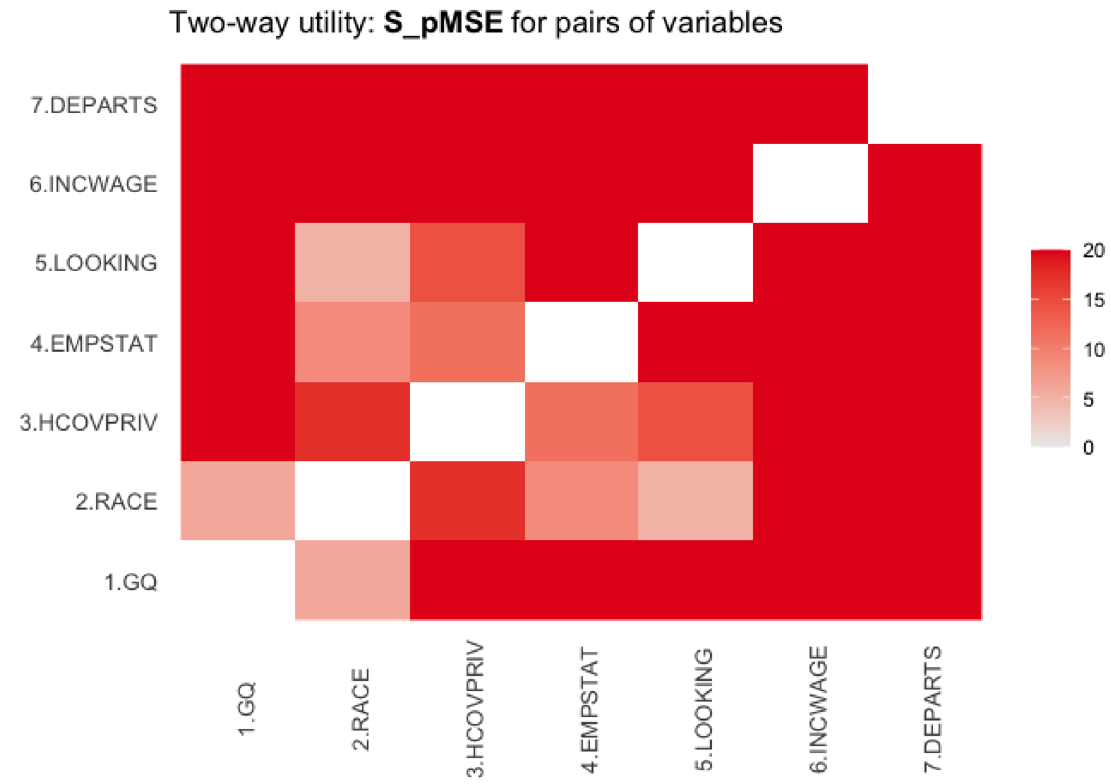
Two-way Tables Comparison of Synthesised Data (R-Package: synthpop) by (S_)pMSE

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).

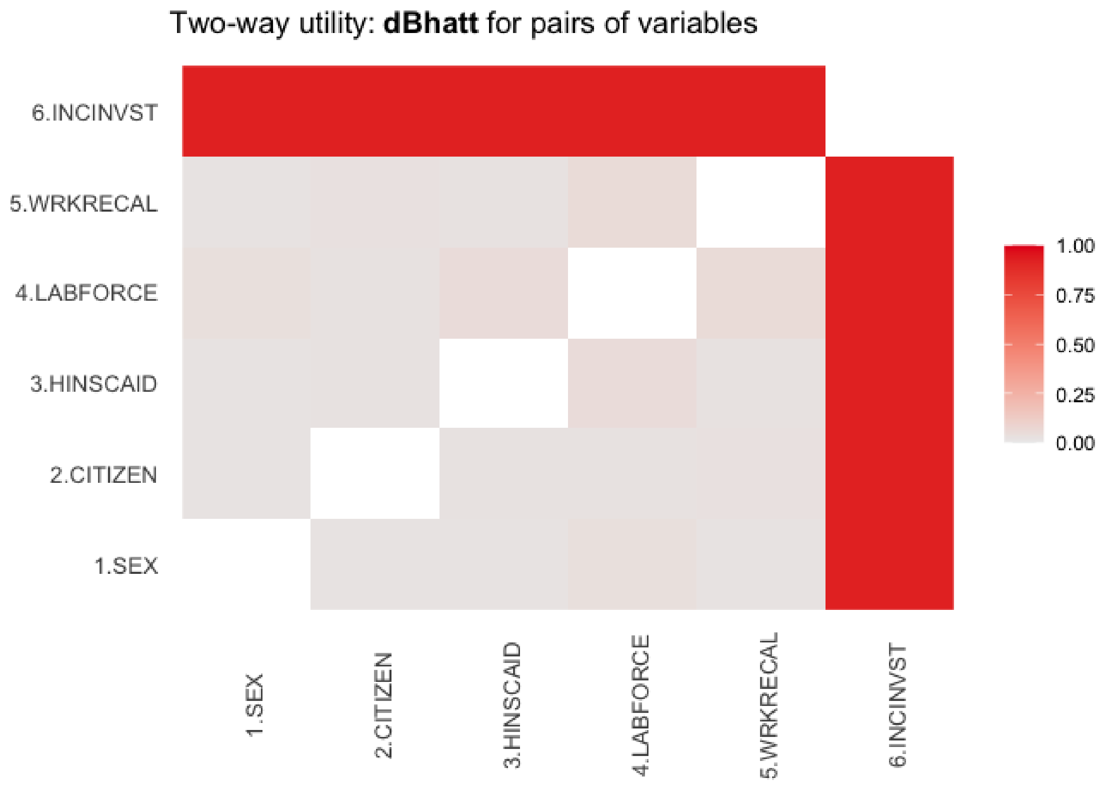


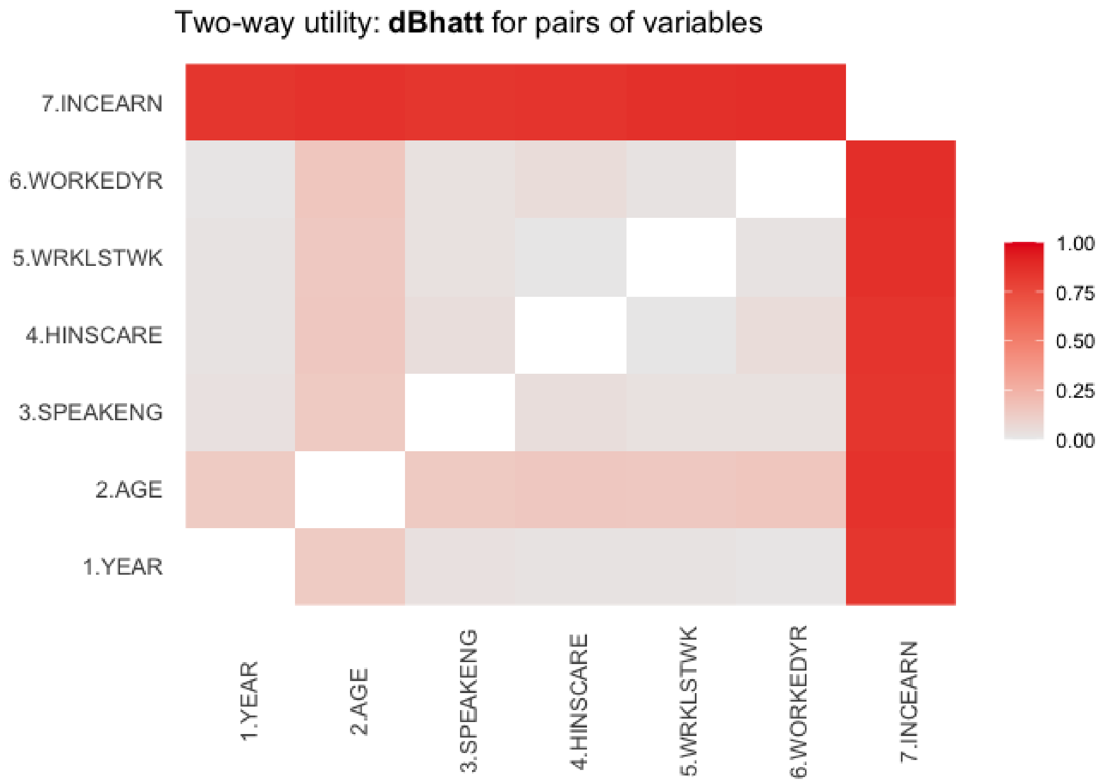


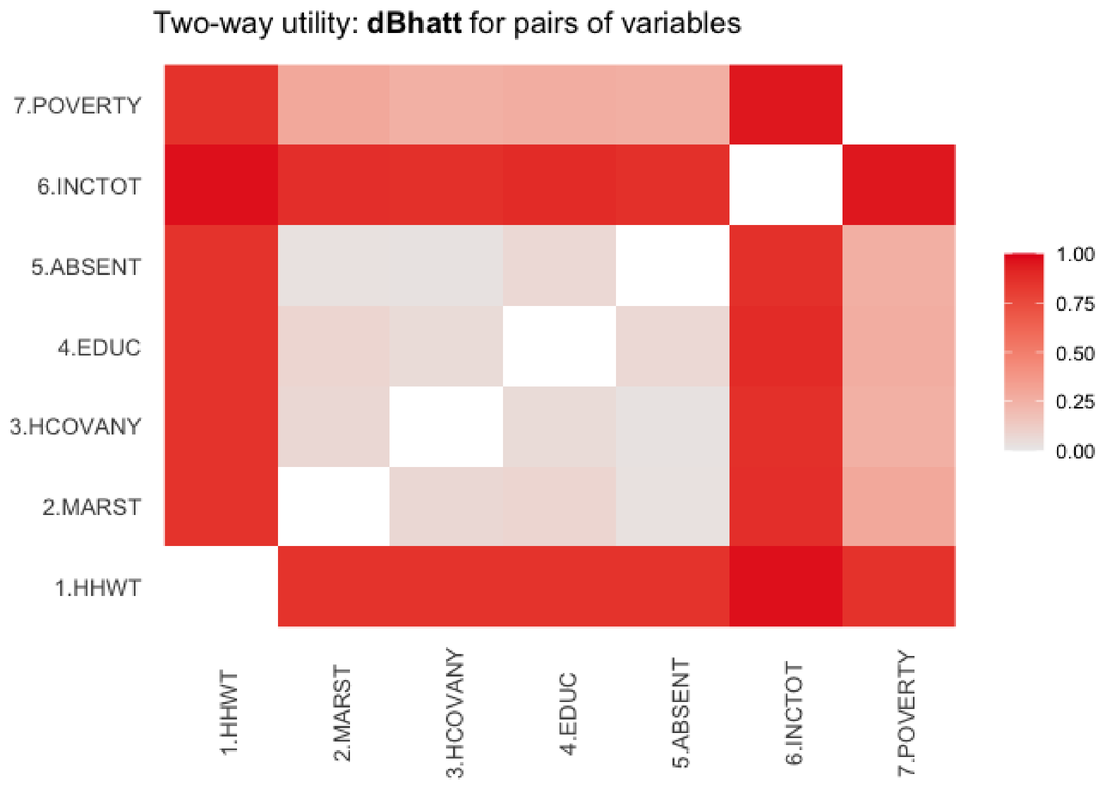

Two-way utility: **S_pMSE** for pairs of variables
7.POVERTY
6.INCTOT
5.ABSENT
4.EDUC
3.HCOVANY
2.MARST
1.HHWT
20
15
10
5
0

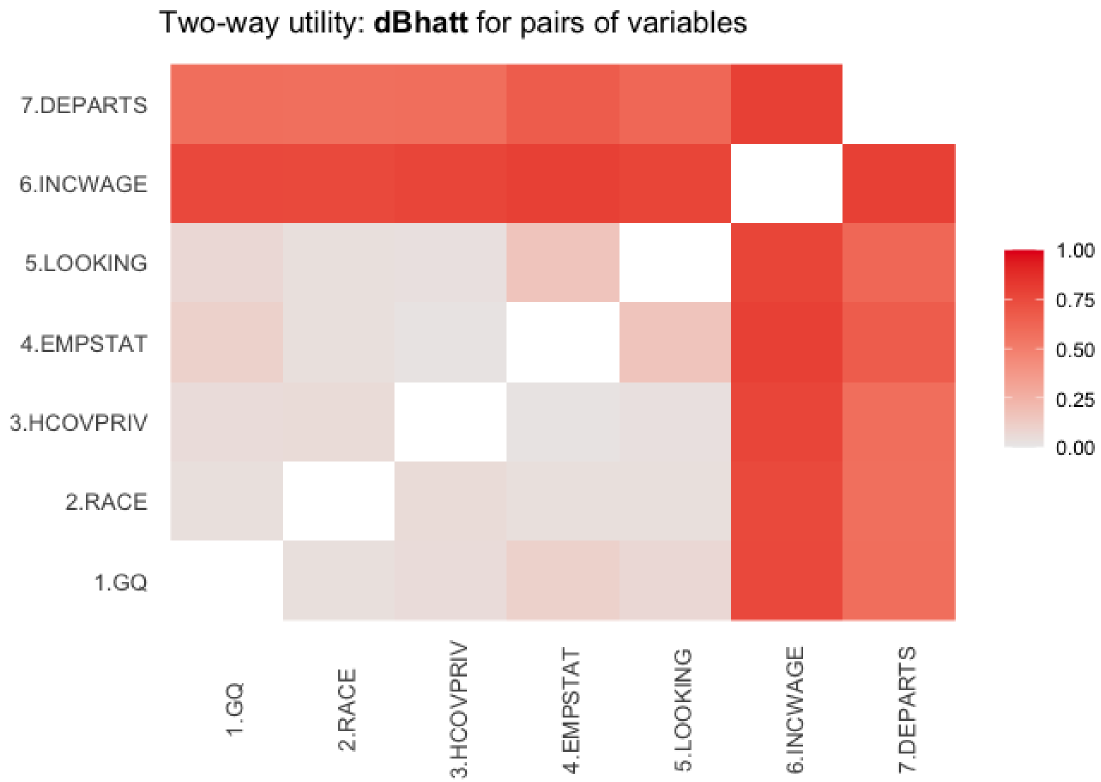

Two-way utility: **S_pMSE** for pairs of variables
7.DEPARTS
6.INCWAGE
5.LOOKING
4.EMPSTAT
3.HCOVPRIV
2.RACE
1.GQ
20
15
10
5
0

```
## NULL
```

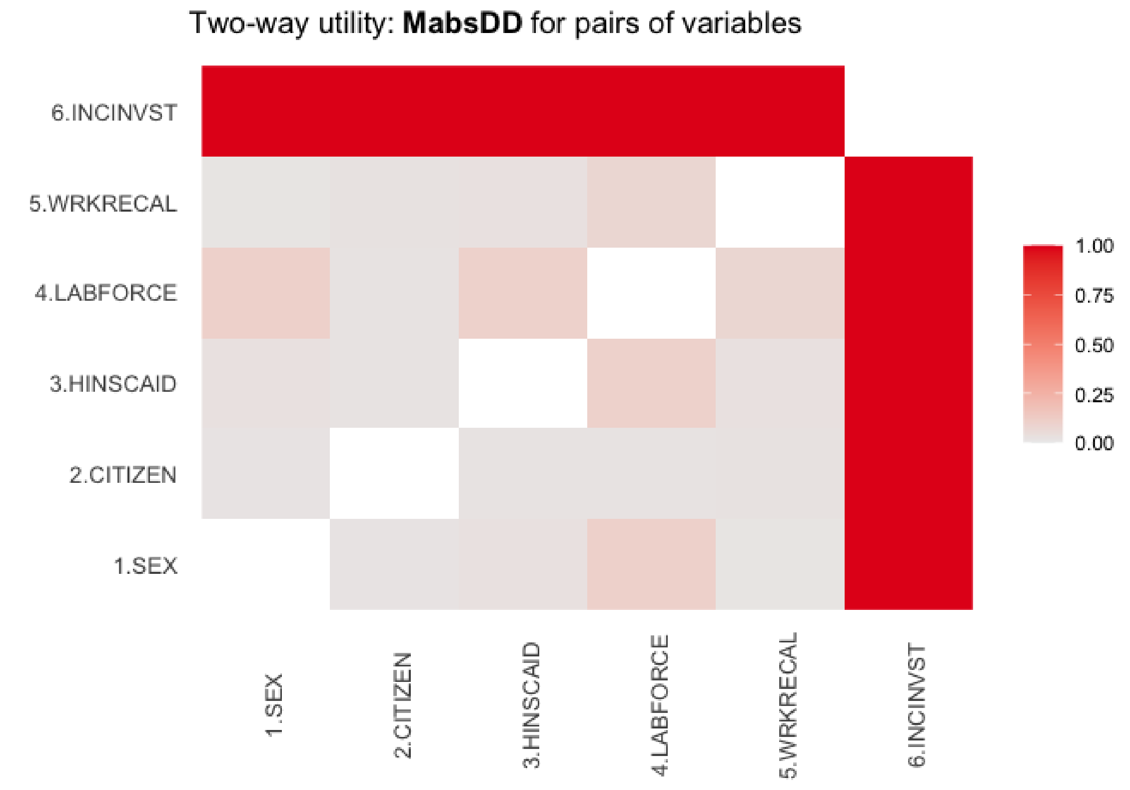

Two-way utility: **dBhatt** for pairs of variables
6.INCINVST
5.WRKRECAL
4.LABFORCE
3.HINSCAID
2.CITIZEN
1.SEX
1.SEX
2.CITIZEN
3.HINSCAID
4.LABFORCE
5.WRKRECAL
6.INCINVST
1.00
0.75
0.50
0.25
0.00

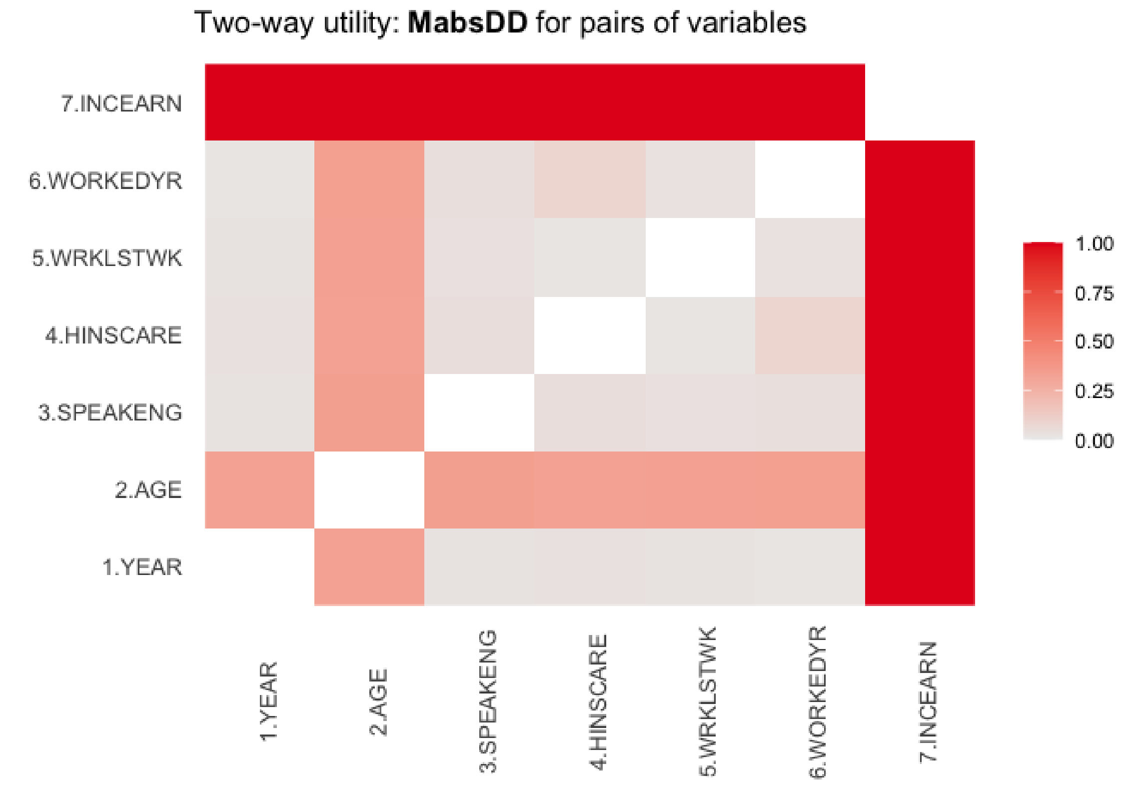

Two-way utility: **dBhatt** for pairs of variables
7.INCEARN
6.WORKEDYR
5.WRKLSTWK
4.HINSCARE
3.SPEAKENG
2.AGE
1.YEAR
1.YEAR
2.AGE
3.SPEAKENG
4.HINSCARE
5.WRKLSTWK
6.WORKEDYR
7.INCEARN
1.00
0.75
0.50
0.25
0.00

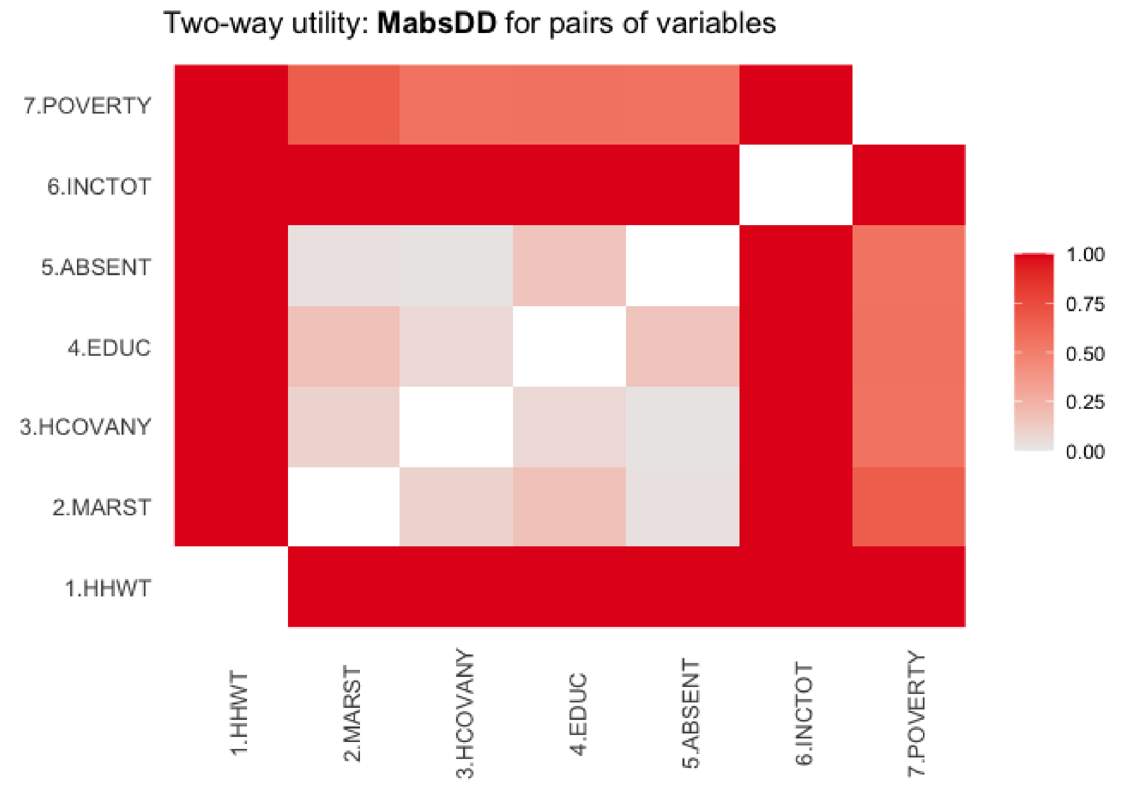

Two-way utility: dBhatt for pairs of variables
7.POVERTY
6.INCTOT
5.ABSENT
4.EDUC
3.HCOVANY
2.MARST
1.HHWT
1.HHWT
2.MARST
3.HCOVANY
4.EDUC
5.ABSENT
6.INCTOT
7.POVERTY
1.00
0.75
0.50
0.25
0.00

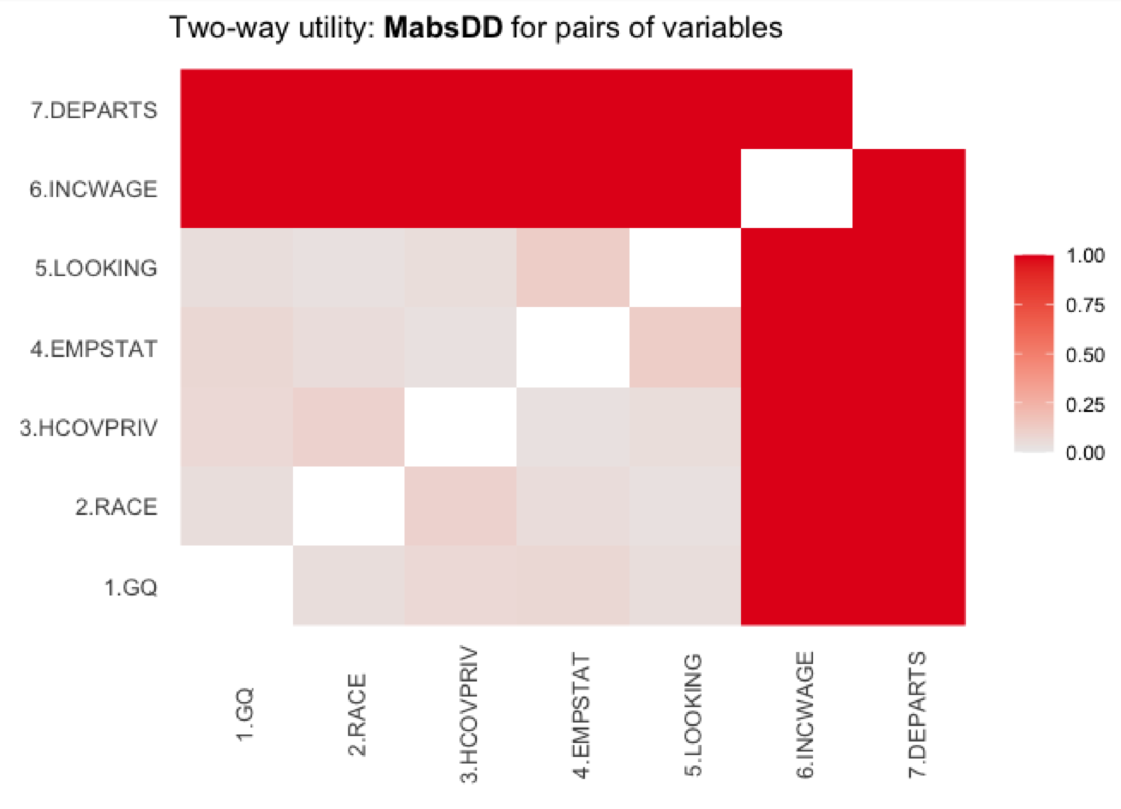

Two-way utility: dBhatt for pairs of variables
7.DEPARTS
6.INCWAGE
5.LOOKING
4.EMPSTAT
3.HCOVPRIV
2.RACE
1.GQ
1.GQ
2.RACE
3.HCOVPRIV
4.EMPSTAT
5.LOOKING
6.INCWAGE
7.DEPARTS
1.00
0.75
0.50
0.25
0.00

```
## NULL
```


Two-way utility: **MabsDD** 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
1.00
0.75
0.50
0.25
0.00


Two-way utility: **MabsDD** 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






```
## 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
0.5775517

Individual Distances for Information Loss:

```
##       YEAR       HHWT         GQ      PERWT        SEX        AGE      MARST
## 0.85665871 0.98917712 0.07539913 0.98600258 0.49894665 0.90964133 0.60246398
##       RACE     HISPAN    CITIZEN   SPEAKENG    HCOVANY   HCOVPRIV    HINSEMP
## 0.23887889 0.06500515 0.10625192 0.13442514 0.14261023 0.38777582 0.47623572
##   HINSCAID   HINSCARE       EDUC    EMPSTAT   EMPSTATD   LABFORCE   WRKLSTWK
## 0.23028798 0.39041018 0.75218236 0.50345287 0.51548387 0.46424501 0.54633273
##     ABSENT    LOOKING   AVAILBLE   WRKRECAL   WORKEDYR     INCTOT    INCWAGE
## 0.47846918 0.49724576 0.18486220 0.12909942 0.49744032 0.99998358 0.99995903
##   INCWELFR   INCINVST    INCEARN    POVERTY    DEPARTS    ARRIVES
## 0.99961495 0.99994939 0.99997459 0.98751727 0.99624002 0.99453489
```

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

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:

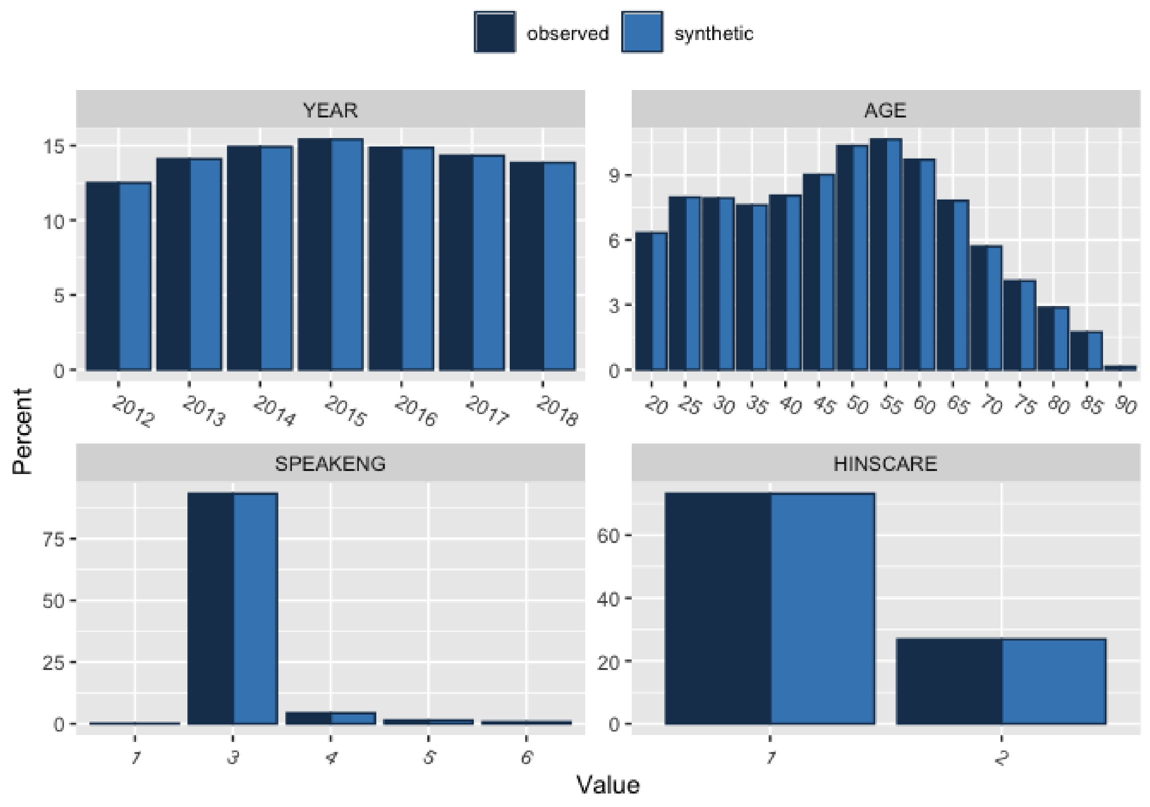
Metric	Number.Uniques	Number.Replications	Percentage.Replications
Perceived Risk	1001862	0	0

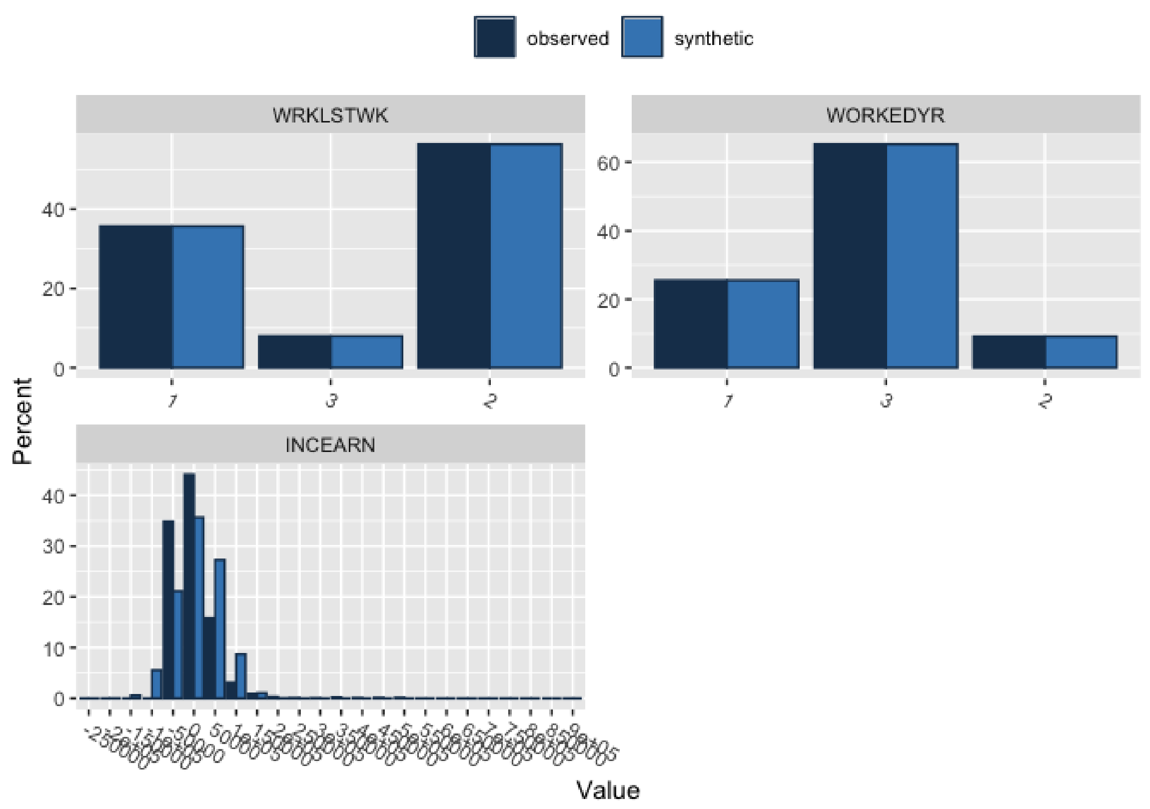
Utility Evaluation

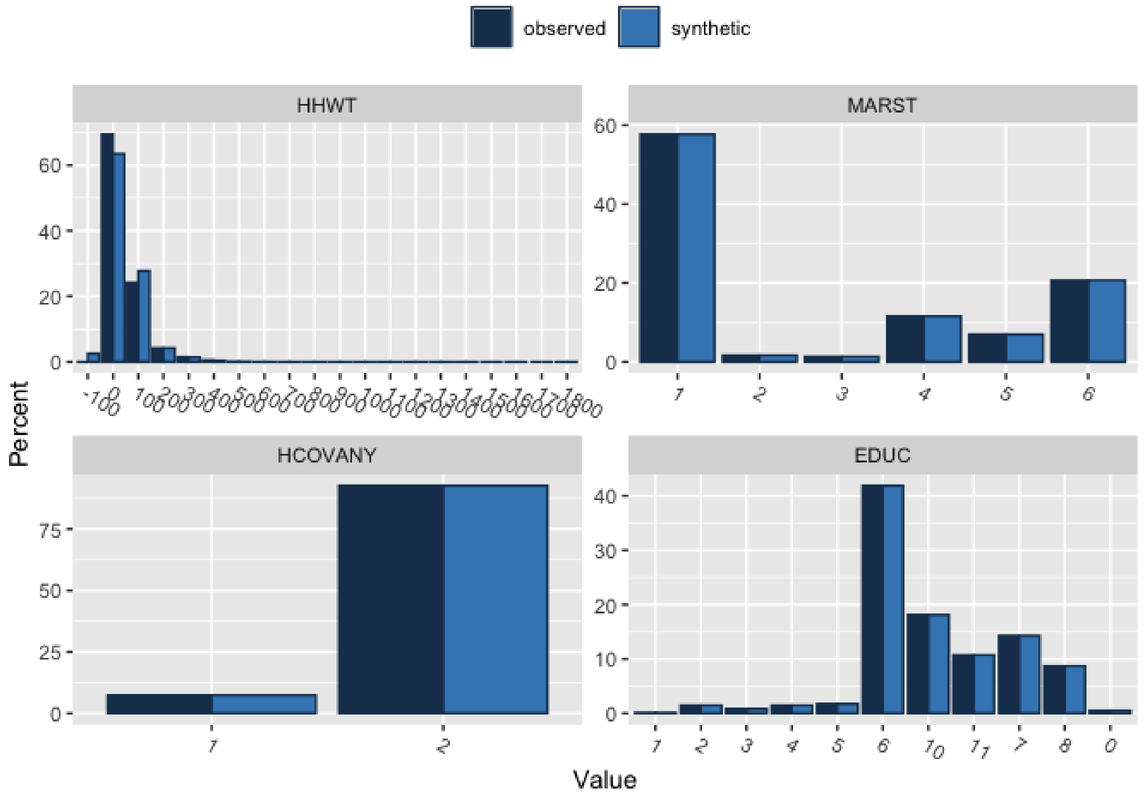
Different utility measures are applied in this section. These utility measures are the basis of utility evaluation for the generated synthetic dataset. The R packages synthpop, sdcMicro and corrplot were used to compute the following metrics. We do not use tests incorporating significance here. Confidence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

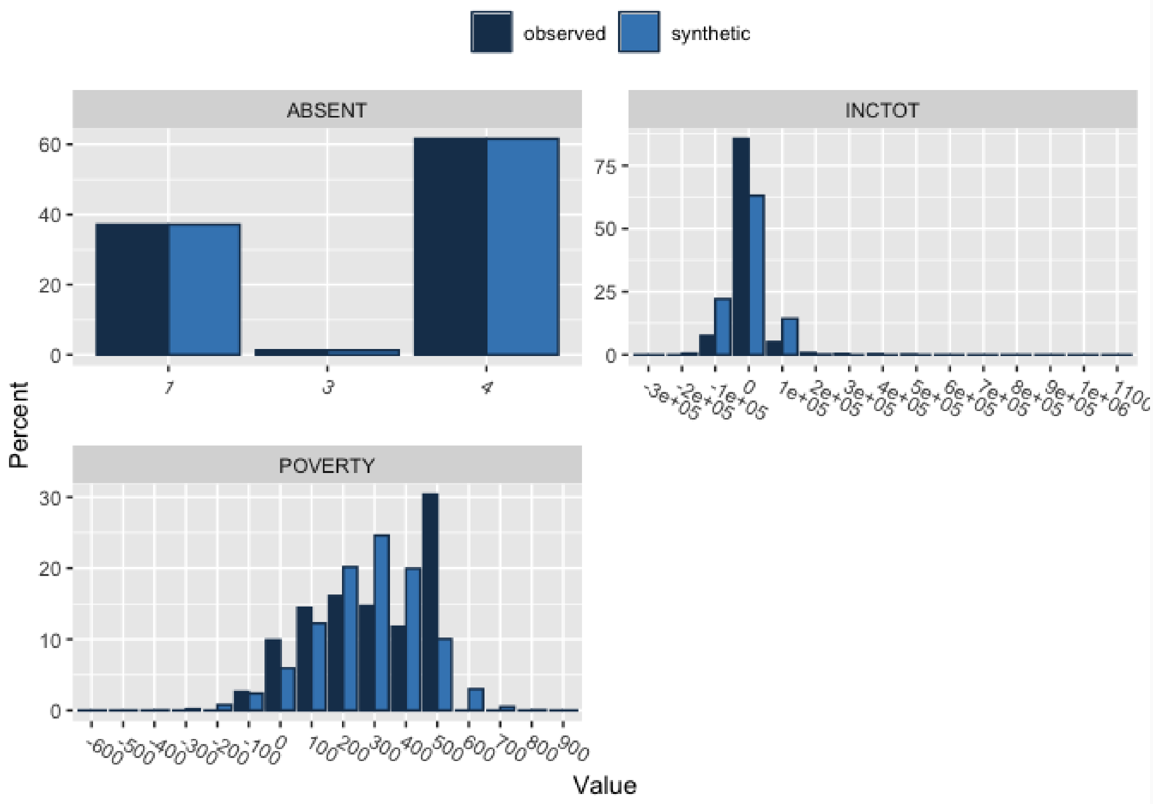
Graphical Comparison for Margins (R-Package: synthpop)

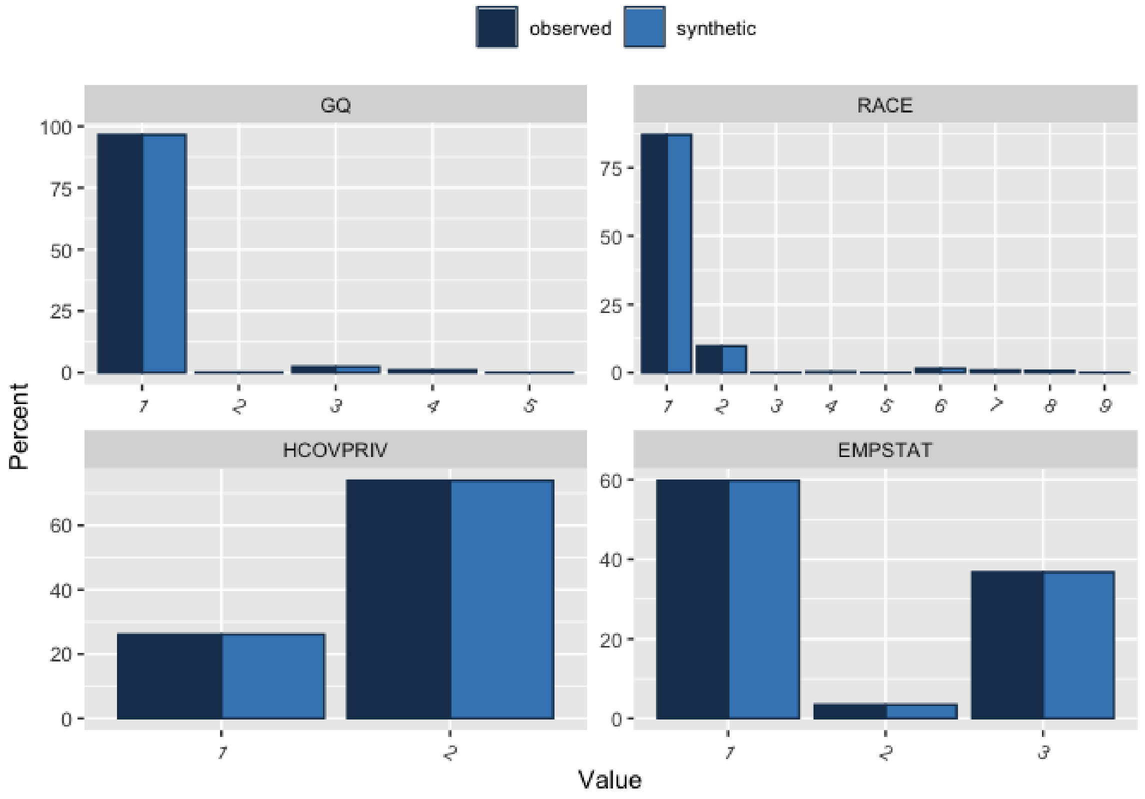
The following histograms provide an ad-hoc overview on the marginal distributions of the original and synthetic dataset. Matching or close distributions are related to a high data utility.

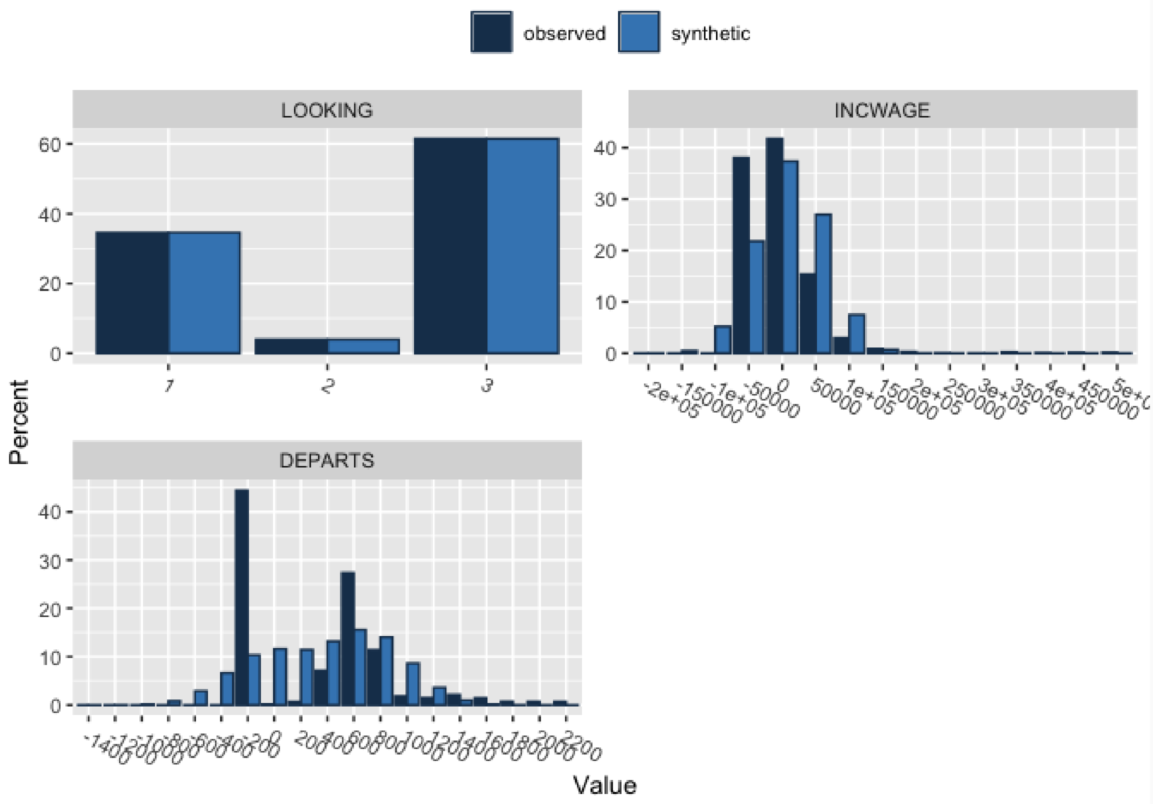

observed
synthetic
YEAR
AGE
SPEAKENG
HINSCARE
Percent
Value
2012 2013 2014 2015 2016 2017 2018
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90
1 3 4 5 6
1 2

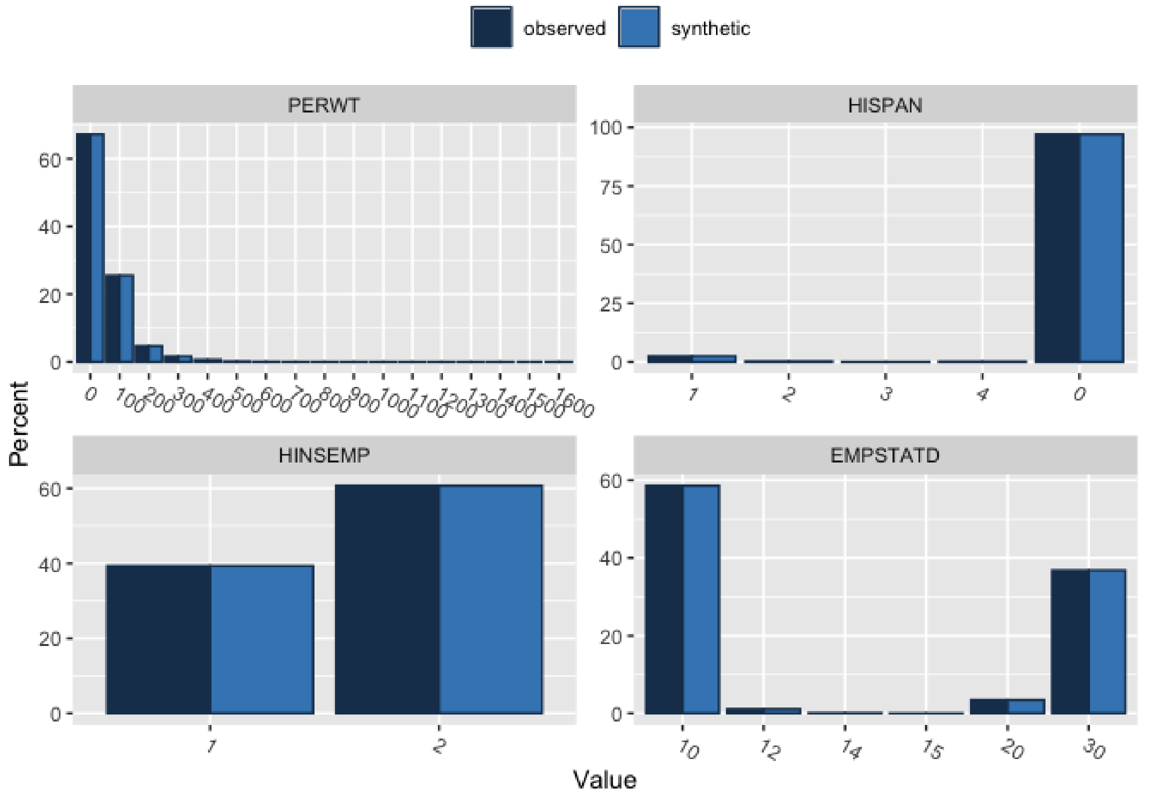

observed
synthetic
WRKLSTWK
WORKEDYR
INCEARN
Percent
Value
1 3 2
1 3 2

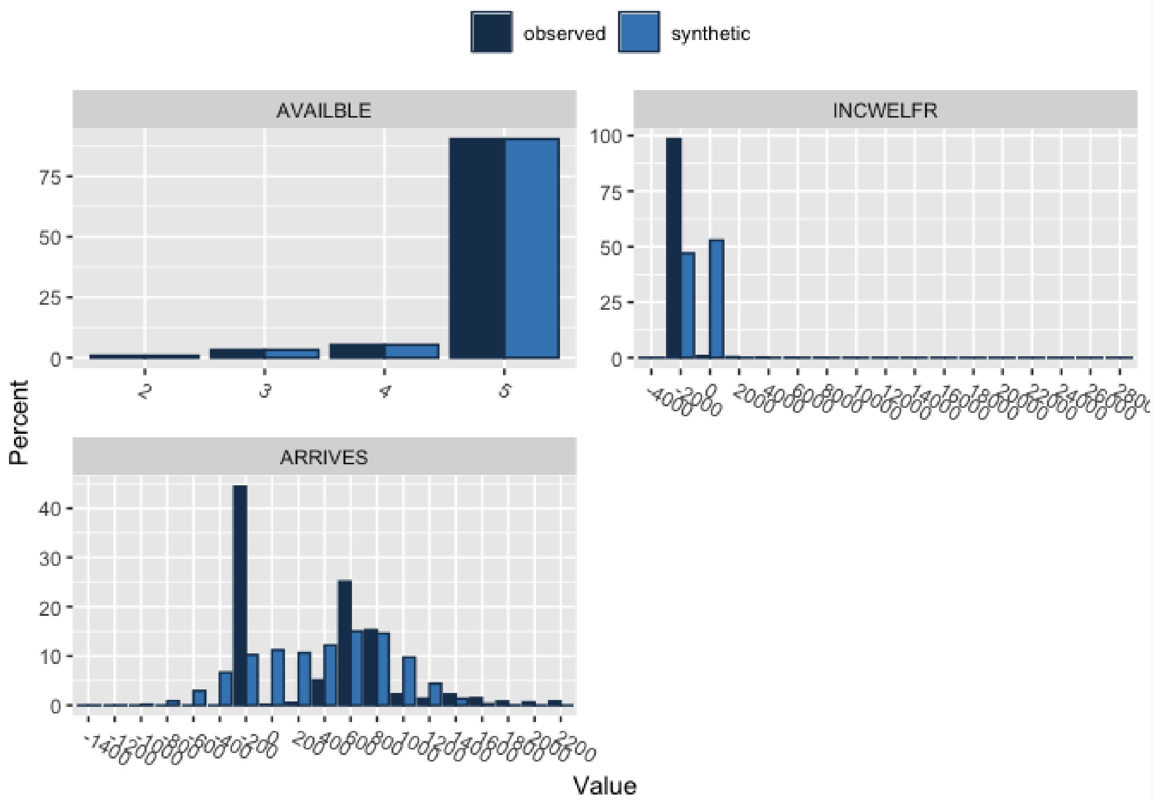

observed
synthetic
HHWT
MARST
HCOVANY
EDUC
Percent
Value

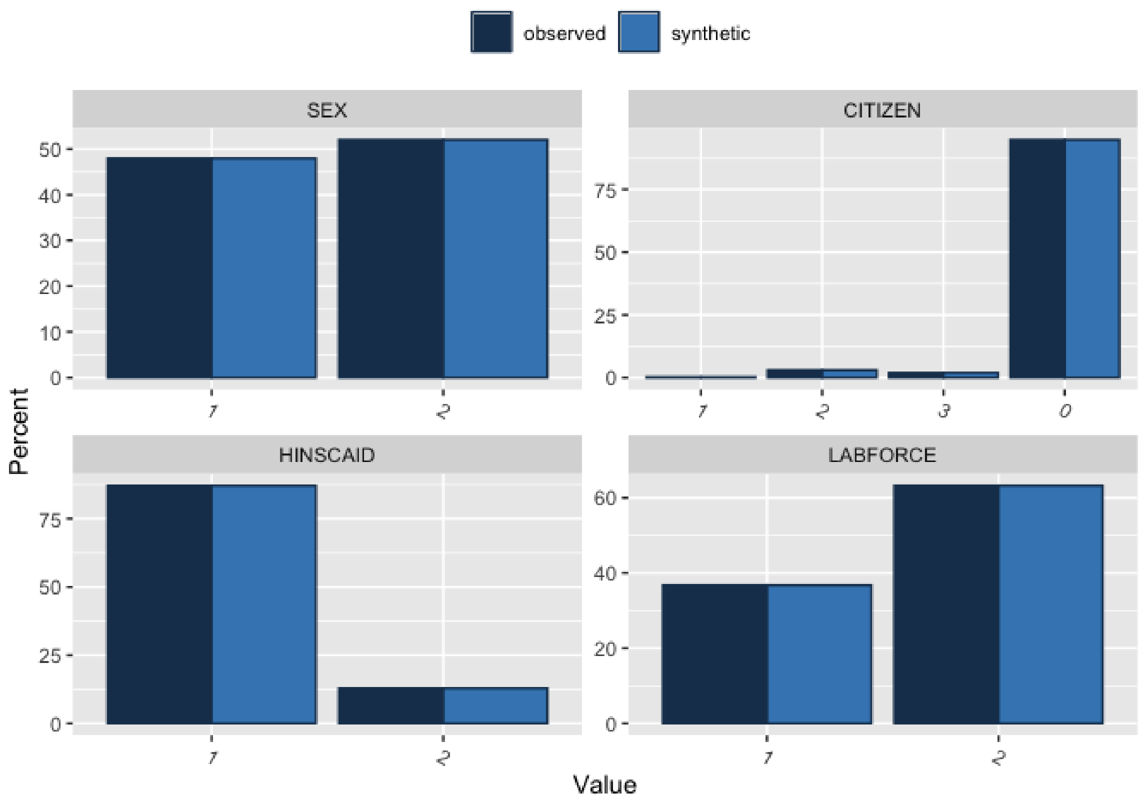

observed
synthetic
ABSENT
INCTOT
POVERTY
Percent
Value

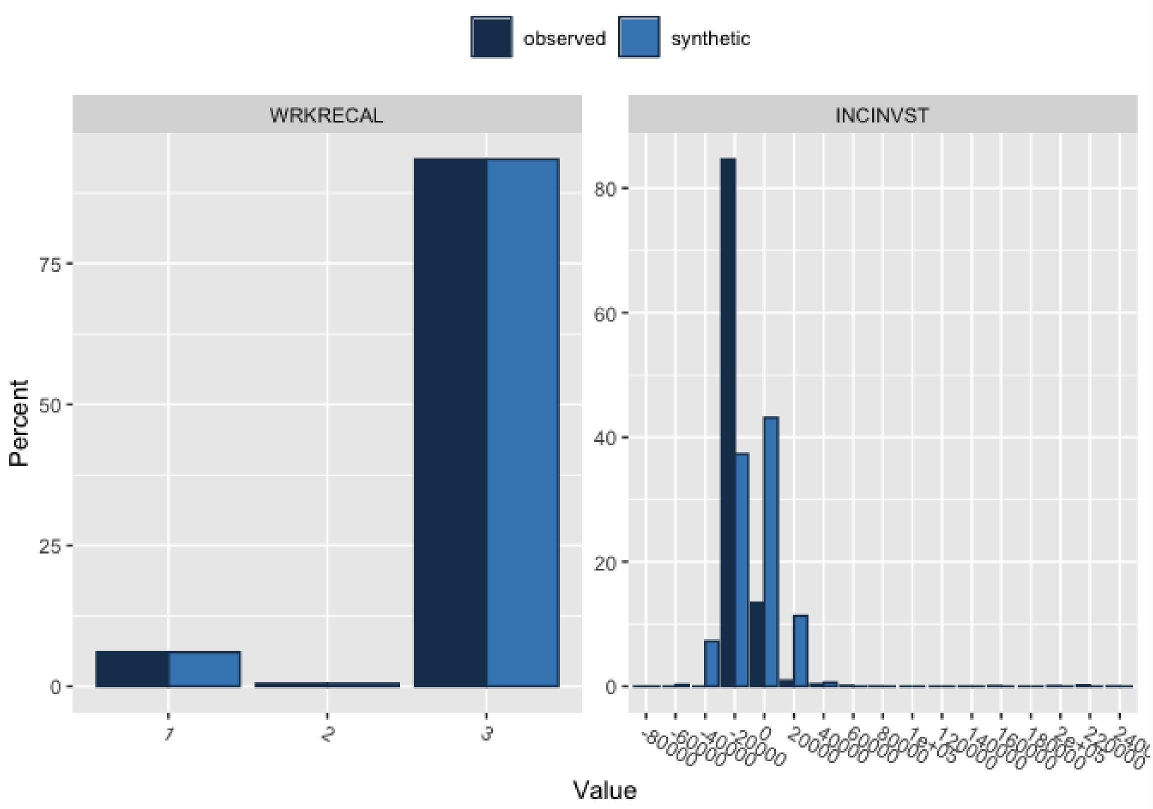

observed
synthetic
GQ
RACE
HCOVPRIV
EMPSTAT
Percent
Value


observed
synthetic
LOOKING
INCWAGE
DEPARTS
Percent
Value


observed
synthetic
PERWT
HISPAN
HINSEMP
EMPSTATD
Percent
Value

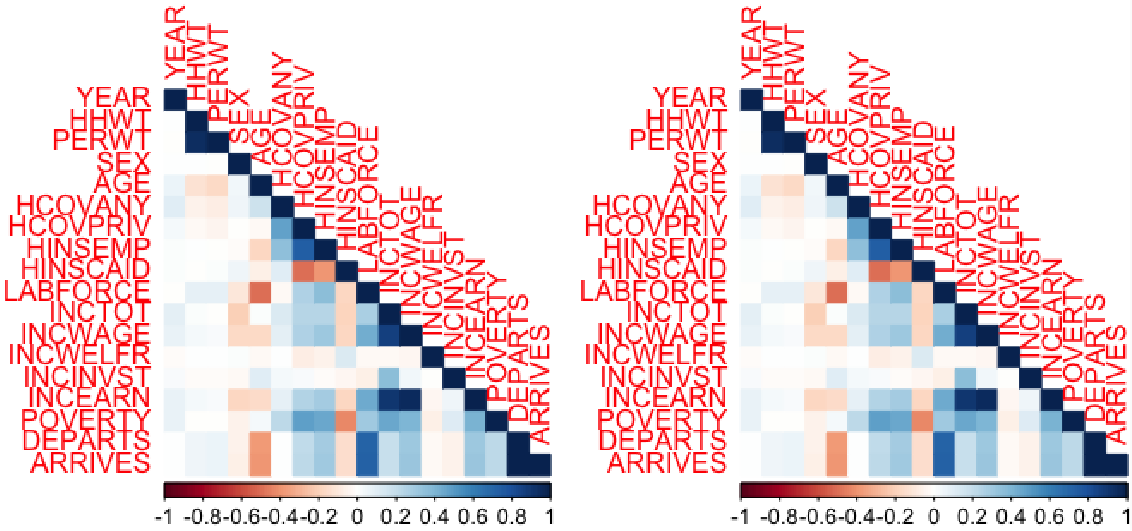

observed
synthetic
AVAILBLE
INCWELFR
ARRIVES
Percent
Value





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

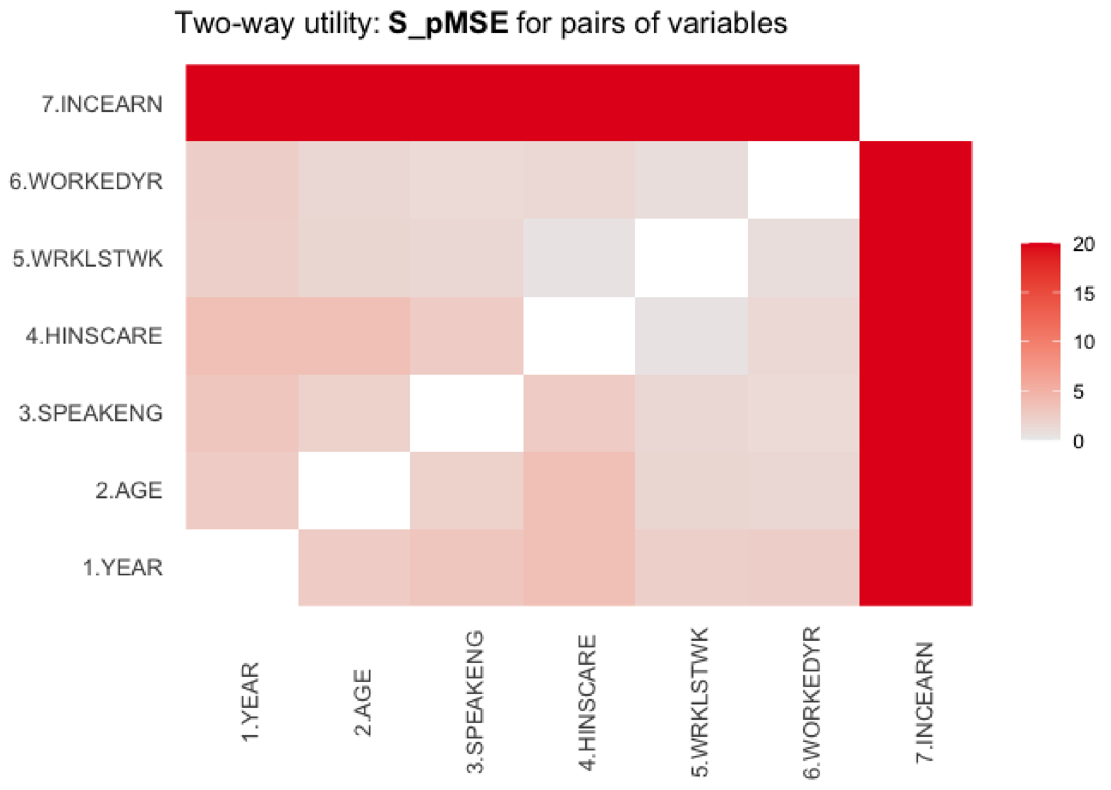
	pMSE	S_pMSE	df

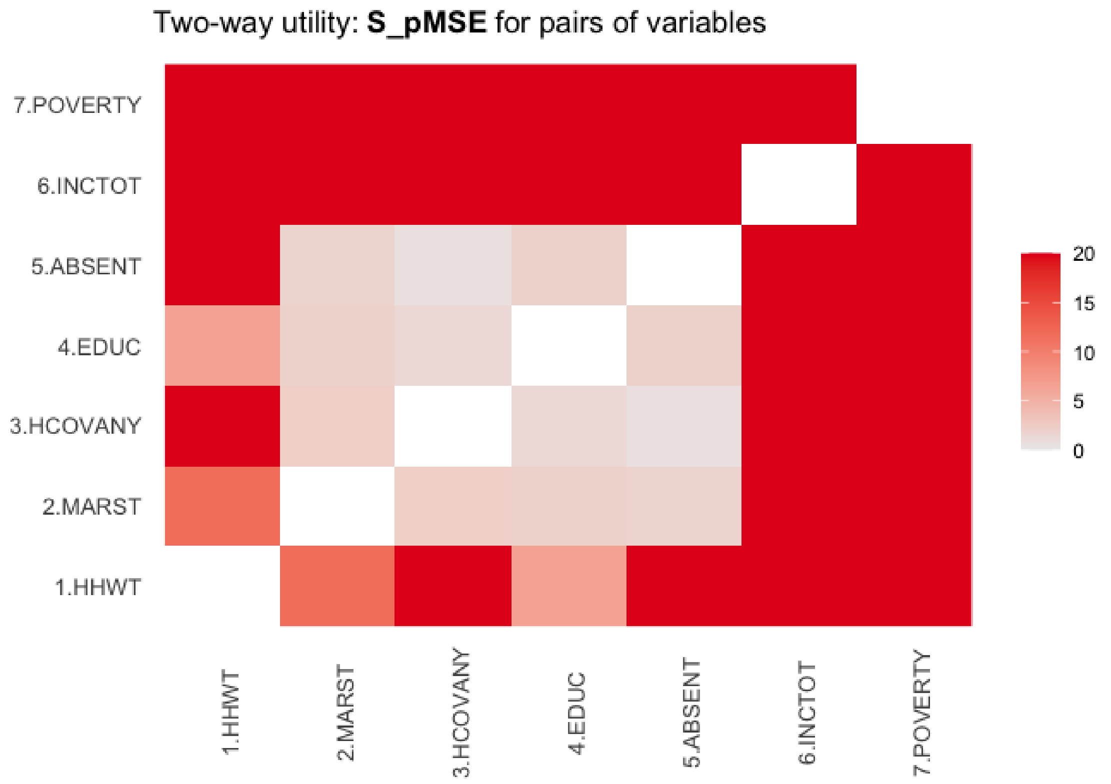
	pMSE	S_pMSE	df
SEX	0.0000000	0.0	1
CITIZEN	0.0000000	0.0	3
HINSCAID	0.0000000	0.0	1
LABFORCE	0.0000000	0.0	1
WRKRECAL	0.0000000	0.0	2
INCINVST	0.1484874	819809.7	3

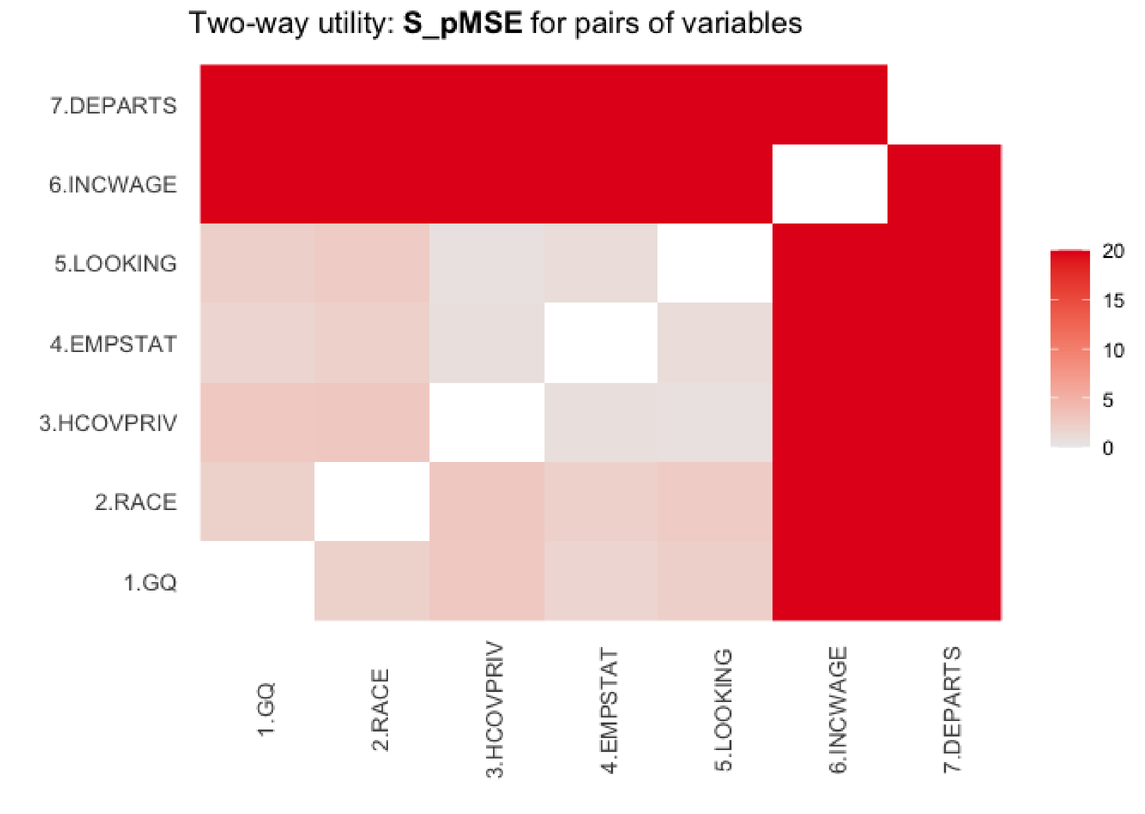
pMSE	S_pMSE
0.2103239	489.4489

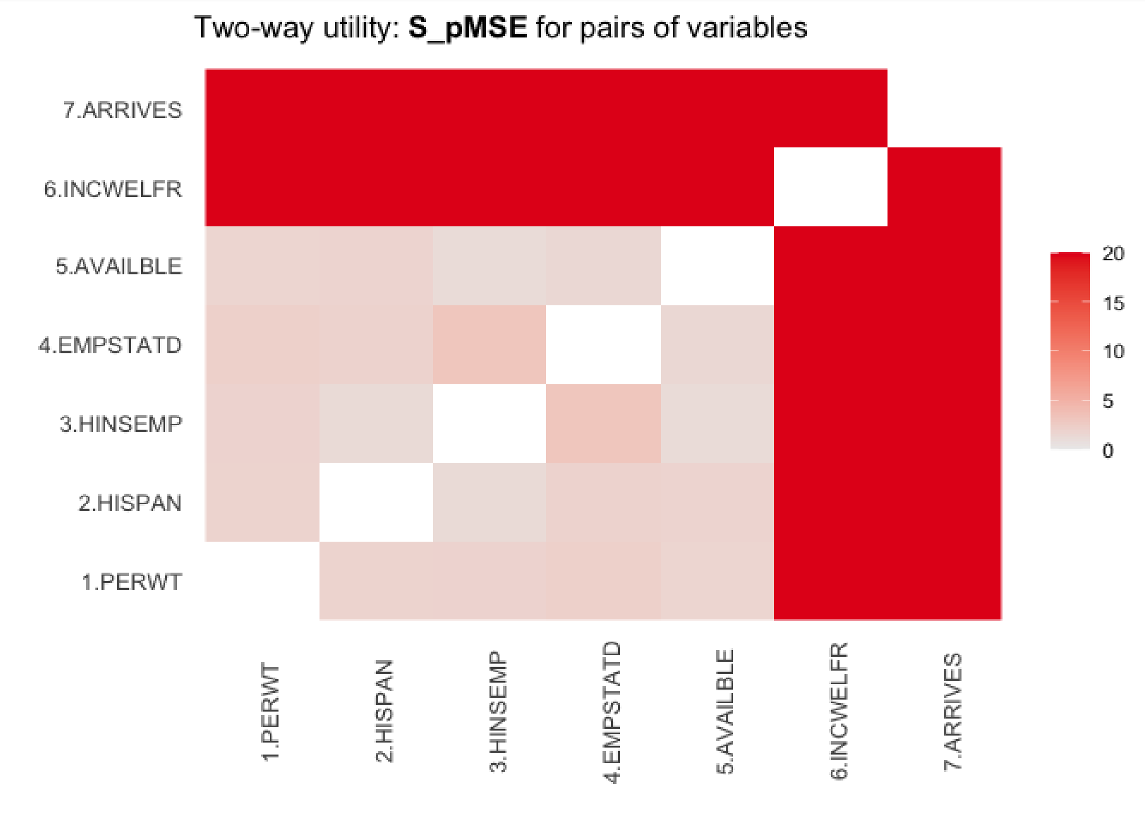
Two-way Tables Comparison of Synthesised Data (R-Package: synthpop) by (S_)pMSE

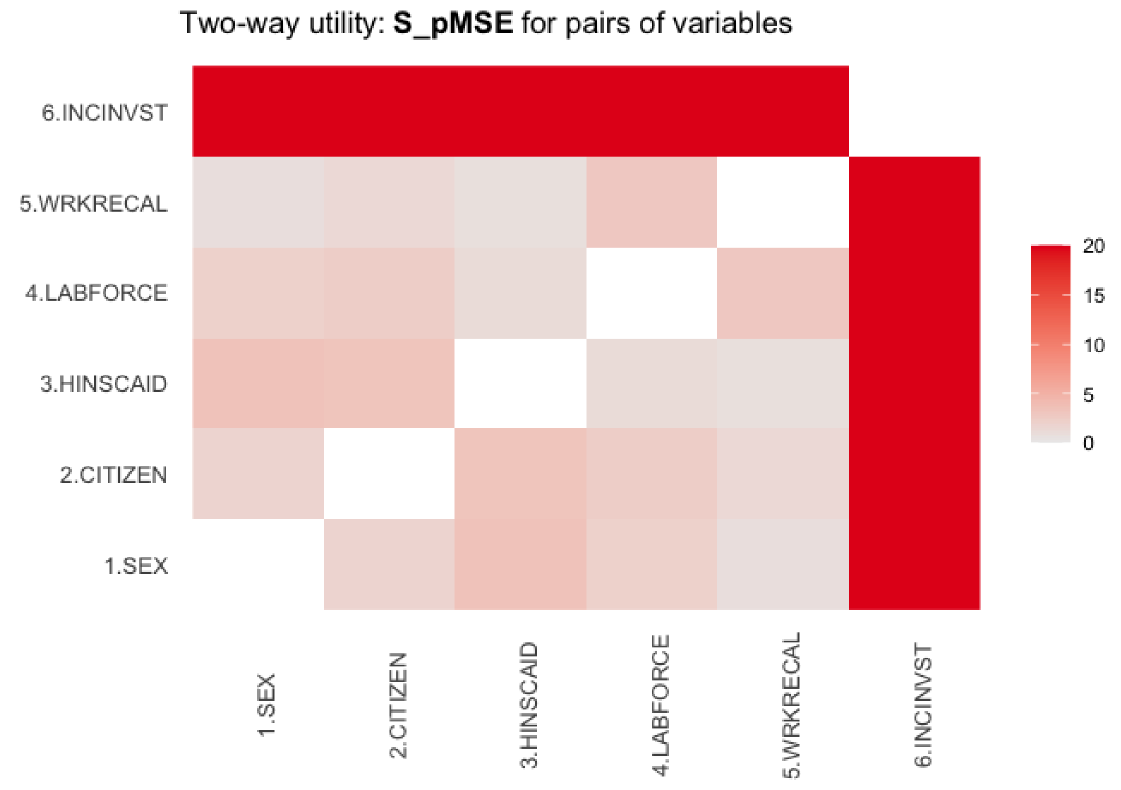
Two-way tables are evaluated based on the original and the synthetic dataset based on S_pMSE (see above). We also present the results for the mean absolute difference in densities (MabsDD) and the Bhattacharyya distance (dBhatt).

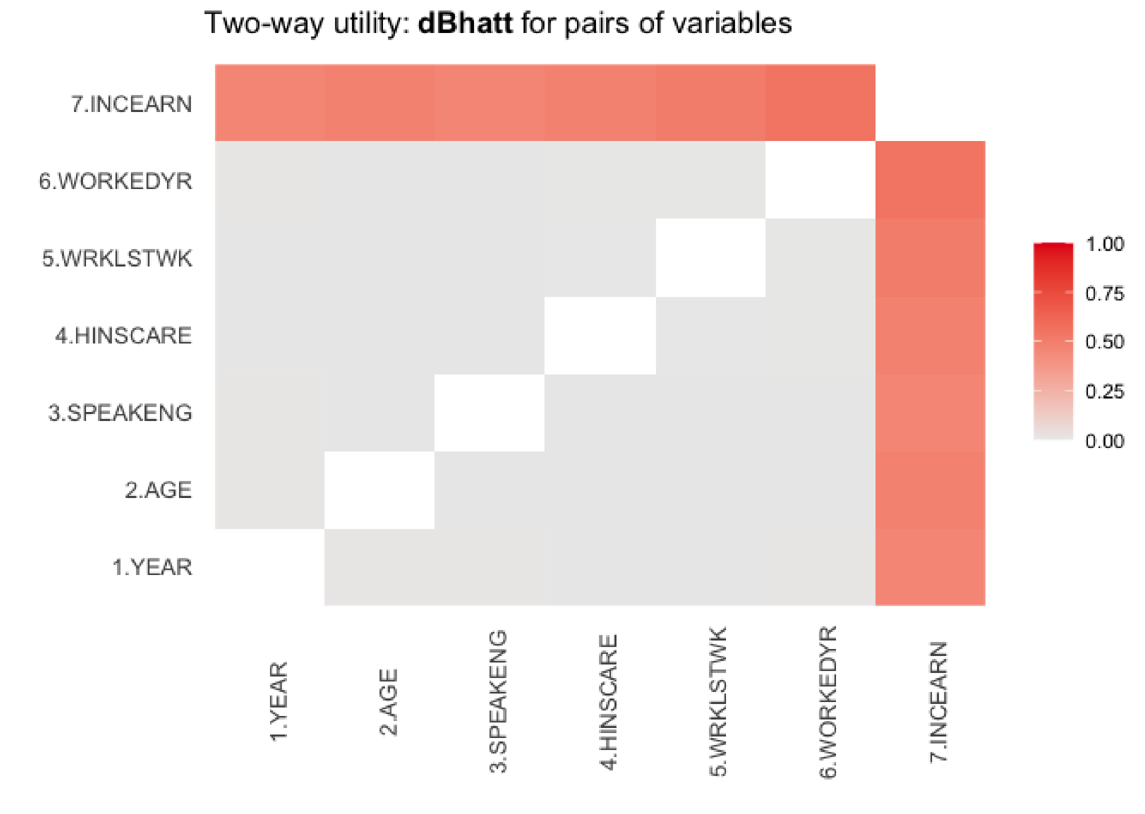

Two-way utility: S_pMSE for pairs of variables
7.INCEARN
6.WORKEDYR
5.WRKLSTWK
4.HINSCARE
3.SPEAKENG
2.AGE
1.YEAR
1.YEAR
2.AGE
3.SPEAKENG
4.HINSCARE
5.WRKLSTWK
6.WORKEDYR
7.INCEARN
20
15
10
5
0

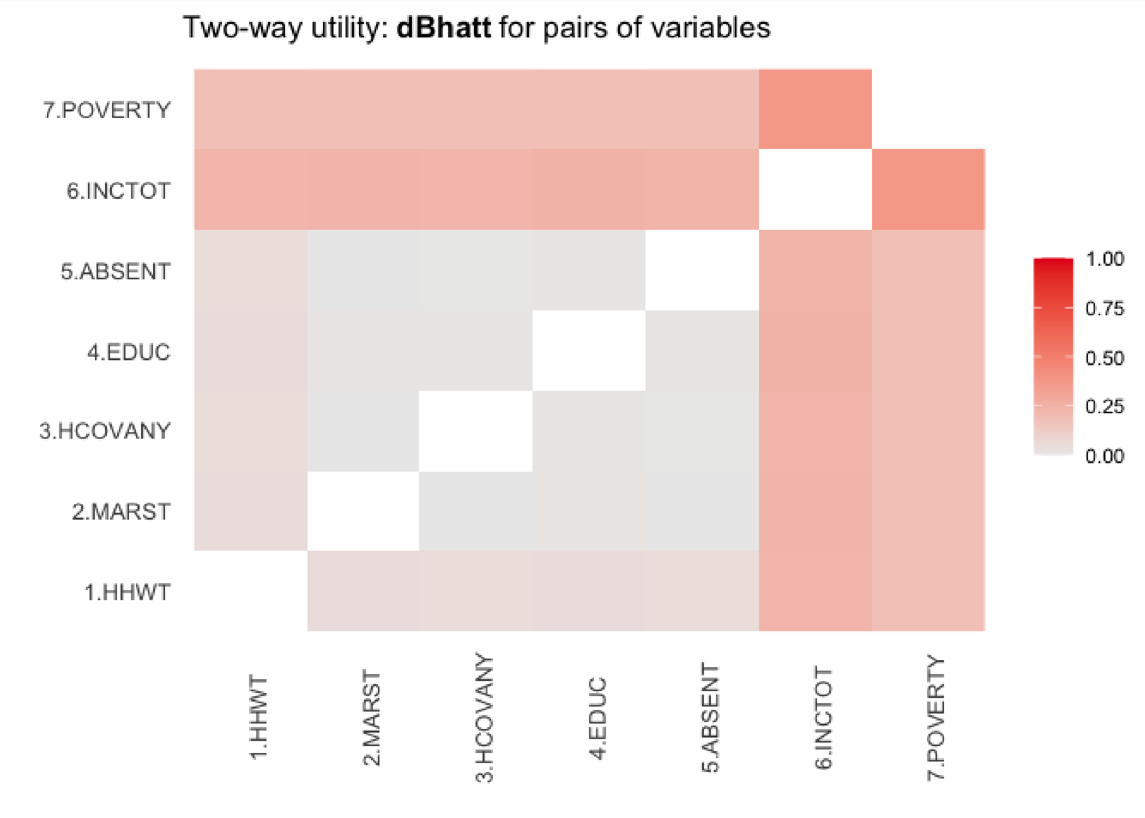

Two-way utility: S_pMSE for pairs of variables
7.POVERTY
6.INCTOT
5.ABSENT
4.EDUC
3.HCOVANY
2.MARST
1.HHWT
1.HHWT
2.MARST
3.HCOVANY
4.EDUC
5.ABSENT
6.INCTOT
7.POVERTY
20
15
10
5
0

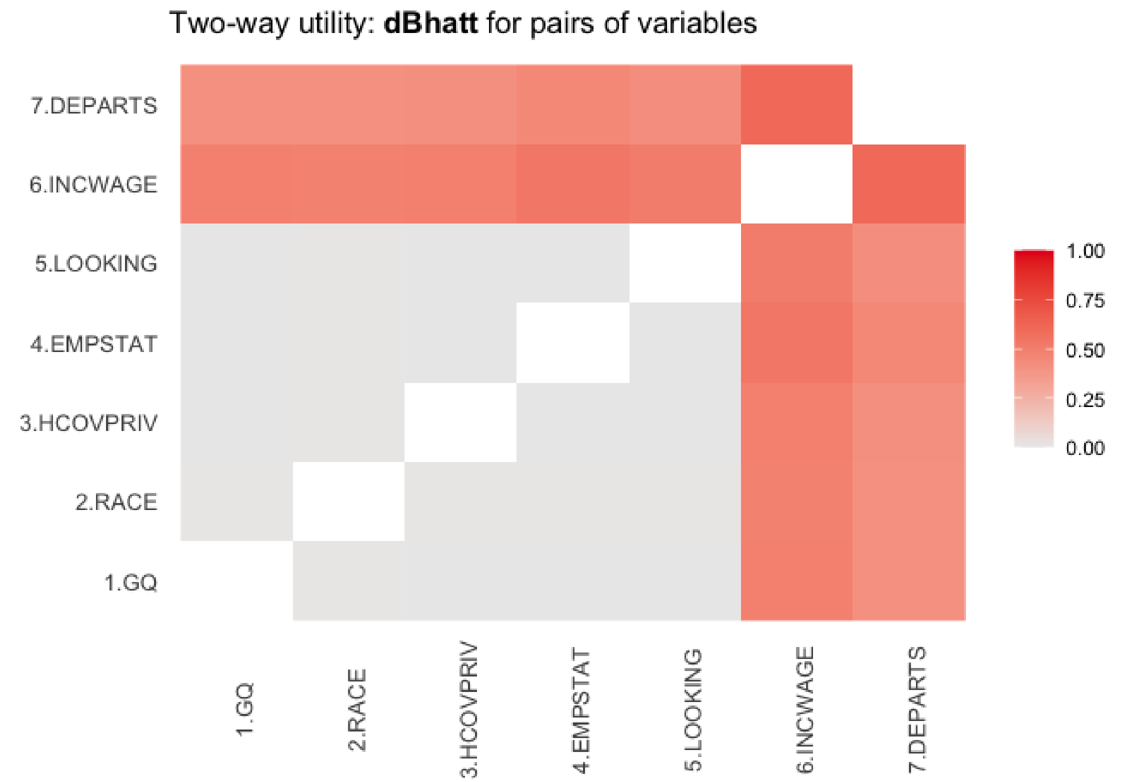

Two-way utility: S_pMSE for pairs of variables
7.DEPARTS
6.INCWAGE
5.LOOKING
4.EMPSTAT
3.HCOVPRIV
2.RACE
1.GQ
1.GQ
2.RACE
3.HCOVPRIV
4.EMPSTAT
5.LOOKING
6.INCWAGE
7.DEPARTS
20
15
10
5
0

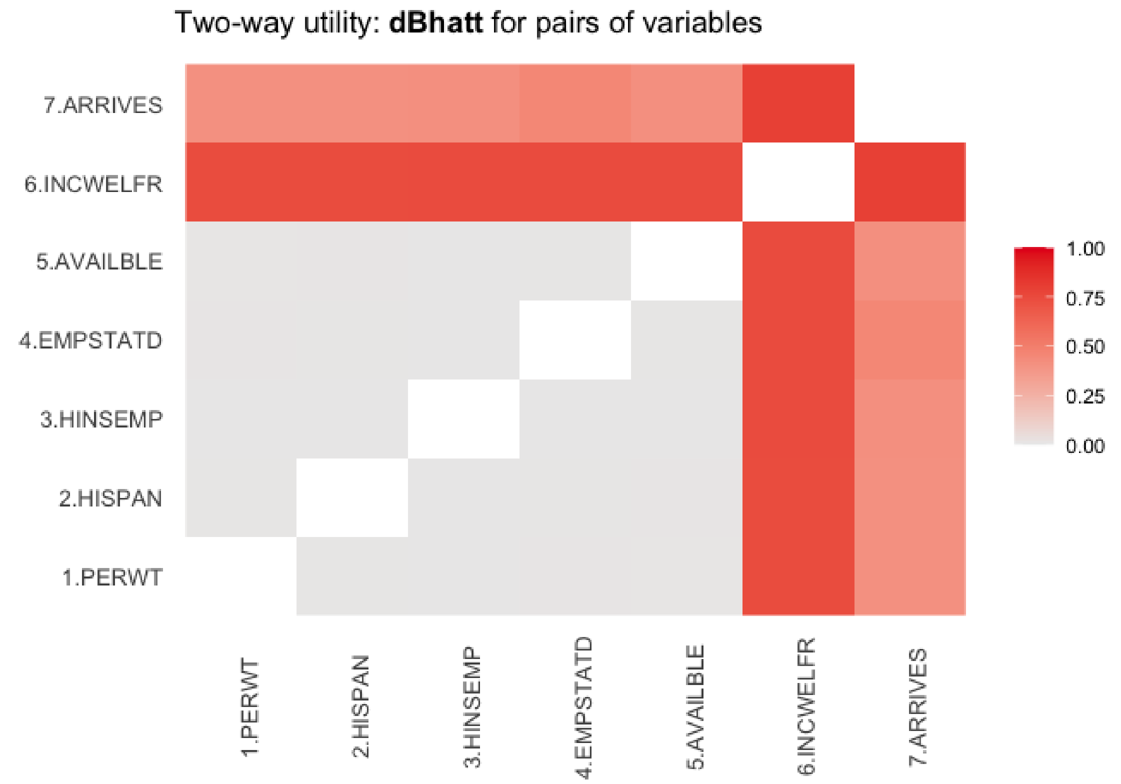

Two-way utility: S_pMSE for pairs of variables
7.ARRIVES
6.INCWELFR
5.AVAILBLE
4.EMPSTATD
3.HINSEMP
2.HISPAN
1.PERWT
1.PERWT
2.HISPAN
3.HINSEMP
4.EMPSTATD
5.AVAILBLE
6.INCWELFR
7.ARRIVES
20
15
10
5
0

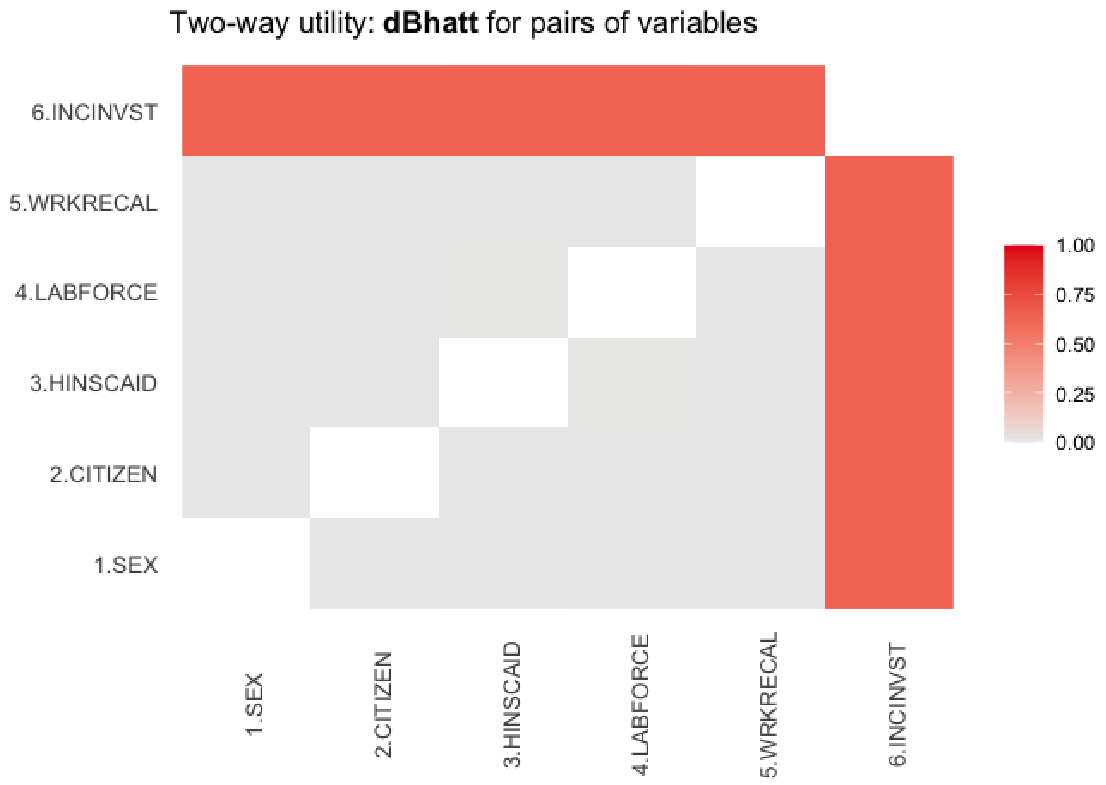

Two-way utility: **S_pMSE** for pairs of variables
6.INCINVST
5.WRKRECAL
4.LABFORCE
3.HINSCAID
2.CITIZEN
1.SEX
1.SEX
2.CITIZEN
3.HINSCAID
4.LABFORCE
5.WRKRECAL
6.INCINVST
20
15
10
5
0

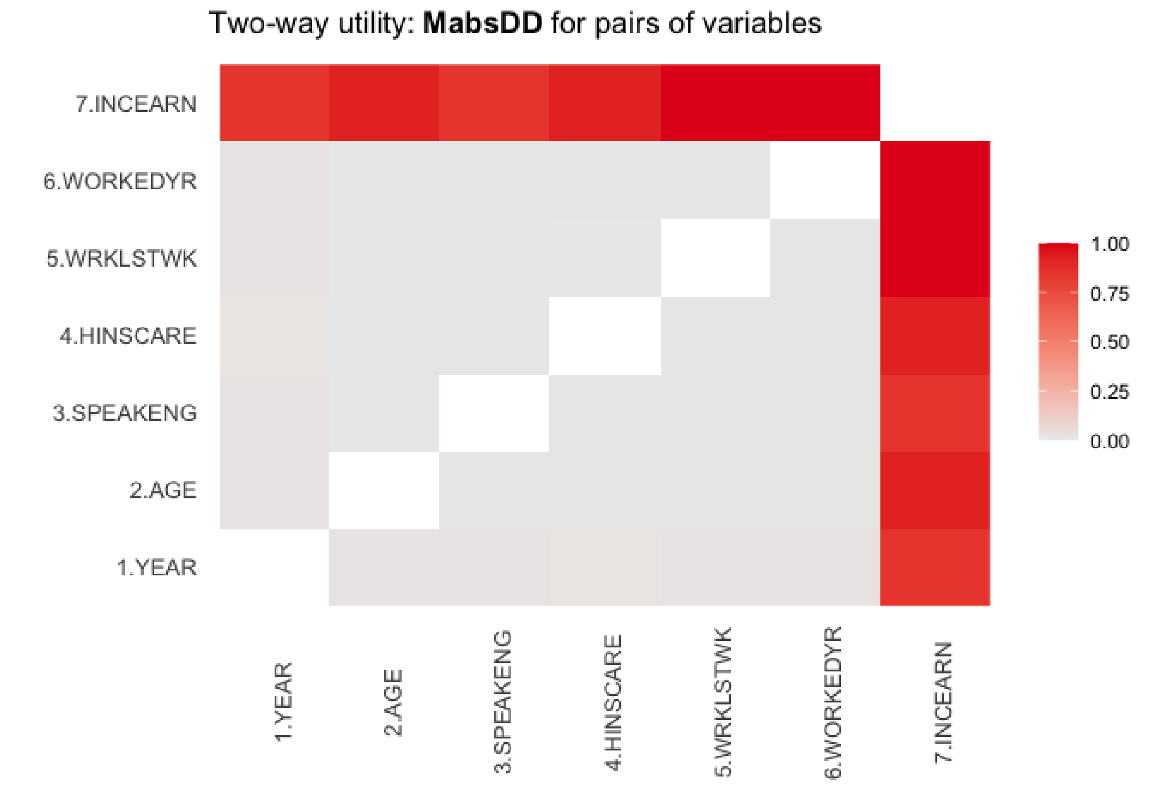

Two-way utility: **dBhatt** for pairs of variables
7.INCEARN
6.WORKEDYR
5.WRKLSTWK
4.HINSCARE
3.SPEAKENG
2.AGE
1.YEAR
1.YEAR
2.AGE
3.SPEAKENG
4.HINSCARE
5.WRKLSTWK
6.WORKEDYR
7.INCEARN
1.00
0.75
0.50
0.25
0.00

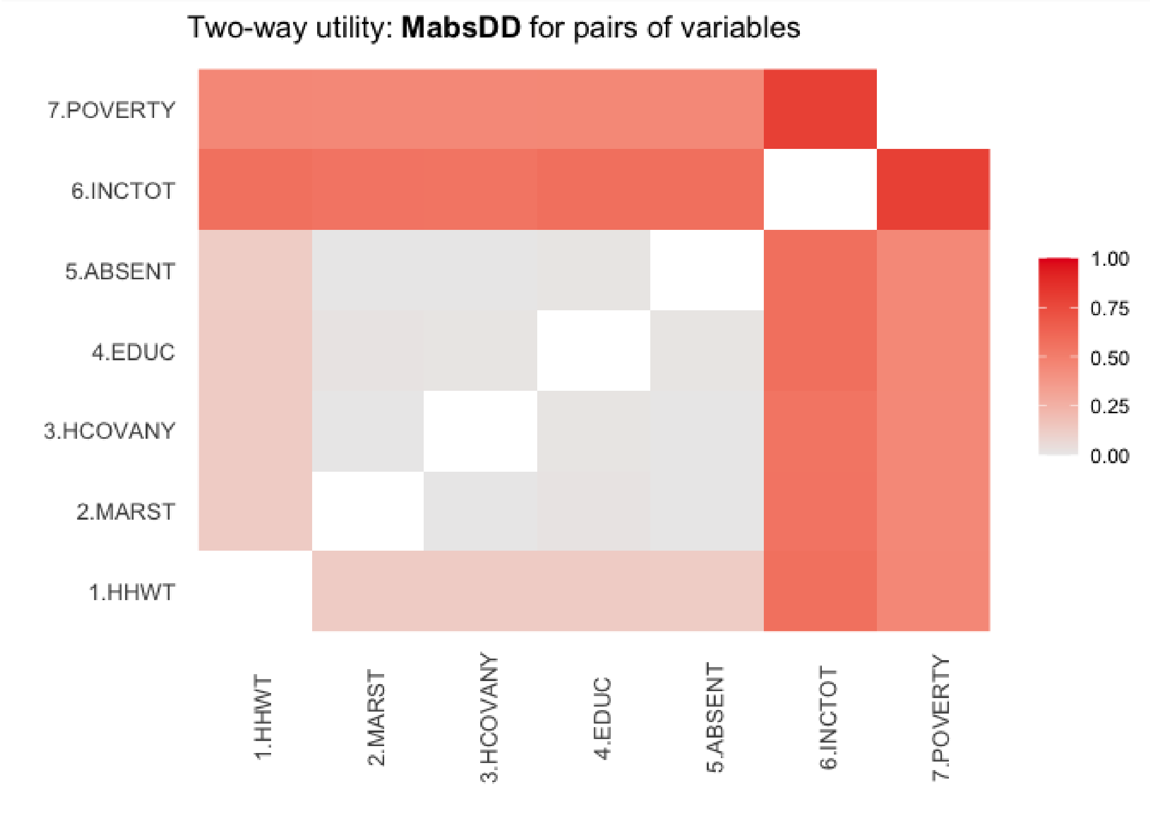

Two-way utility: **dBhatt** for pairs of variables
7.POVERTY
6.INCTOT
5.ABSENT
4.EDUC
3.HCOVANY
2.MARST
1.HHWT
1.HHWT
2.MARST
3.HCOVANY
4.EDUC
5.ABSENT
6.INCTOT
7.POVERTY
1.00
0.75
0.50
0.25
0.00

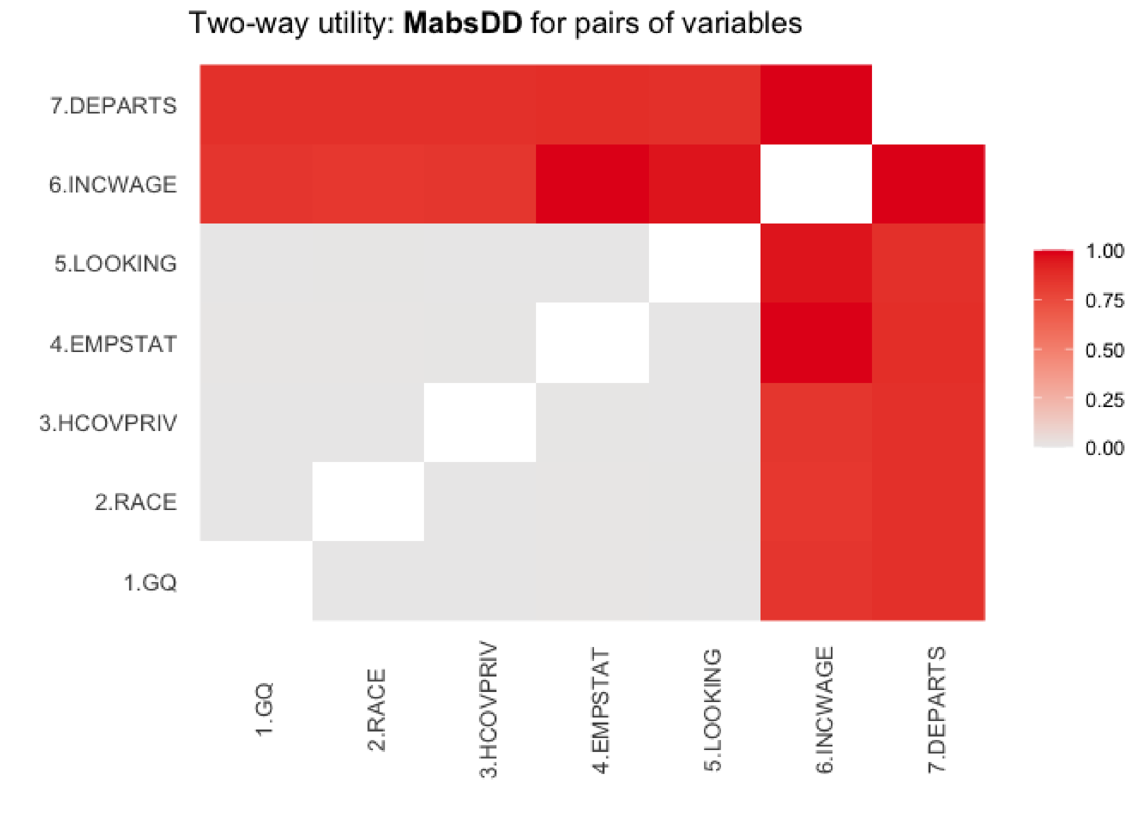

Two-way utility: dBhatt for pairs of variables
7.DEPARTS
6.INCWAGE
5.LOOKING
4.EMPSTAT
3.HCOVPRIV
2.RACE
1.GQ
1.00
0.75
0.50
0.25
0.00

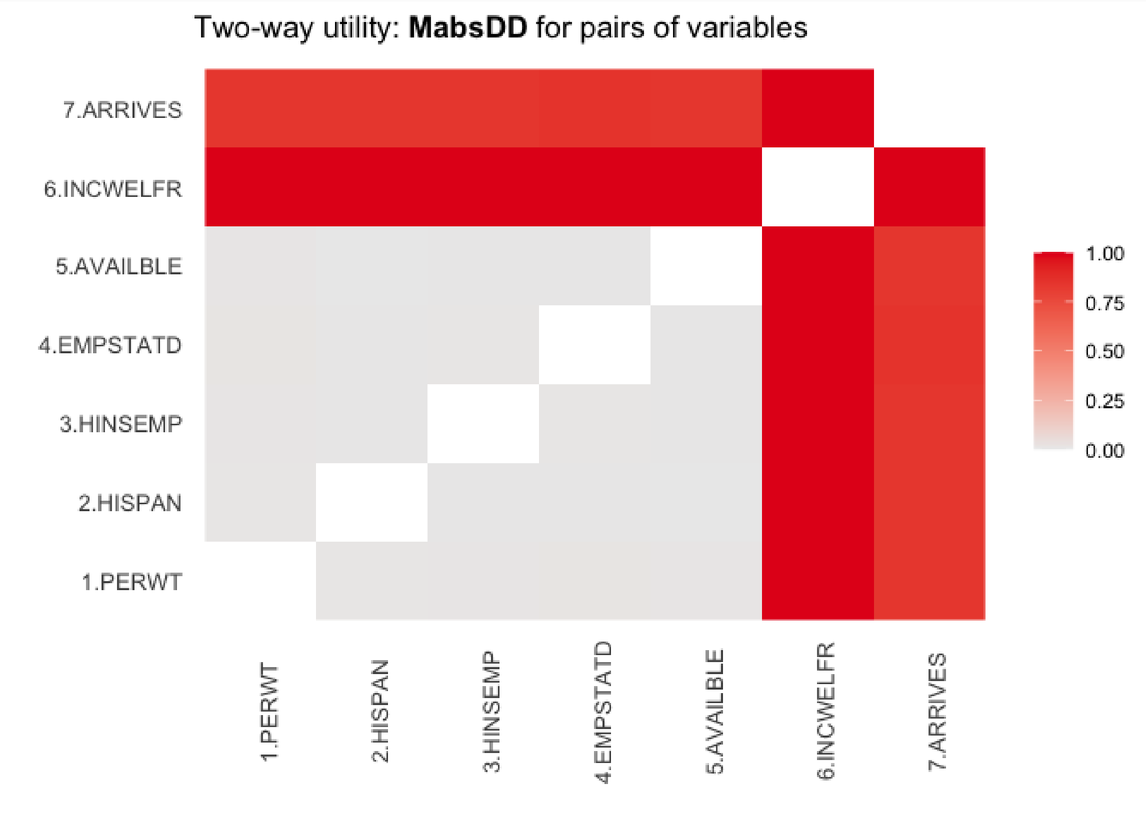

Two-way utility: dBhatt for pairs of variables
7.ARRIVES
6.INCWELFR
5.AVAILBLE
4.EMPSTATD
3.HINSEMP
2.HISPAN
1.PERWT
1.00
0.75
0.50
0.25
0.00

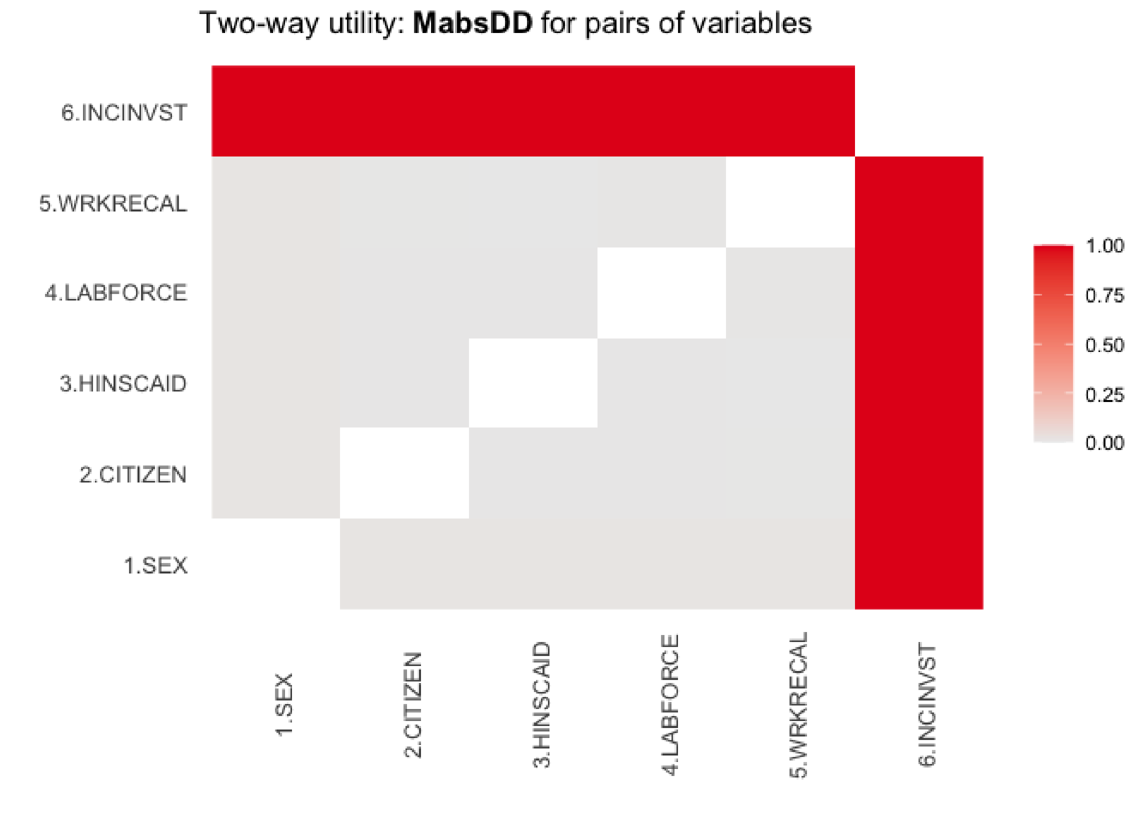

Two-way utility: dBhatt for pairs of variables
6.INCINVST
5.WRKRECAL
4.LABFORCE
3.HINSCAID
2.CITIZEN
1.SEX
1.SEX
2.CITIZEN
3.HINSCAID
4.LABFORCE
5.WRKRECAL
6.INCINVST
1.00
0.75
0.50
0.25
0.00


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)

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.

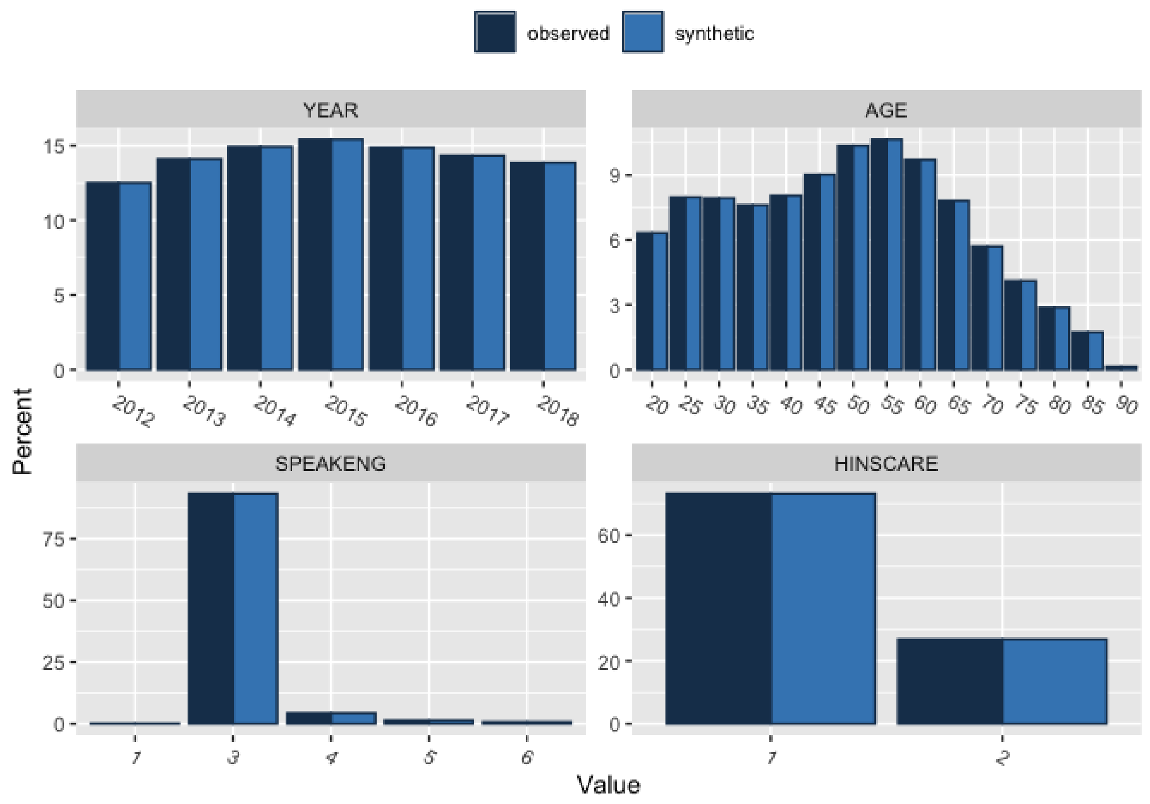
Replication.Uniques	Number.Replications	Percentage.Replications

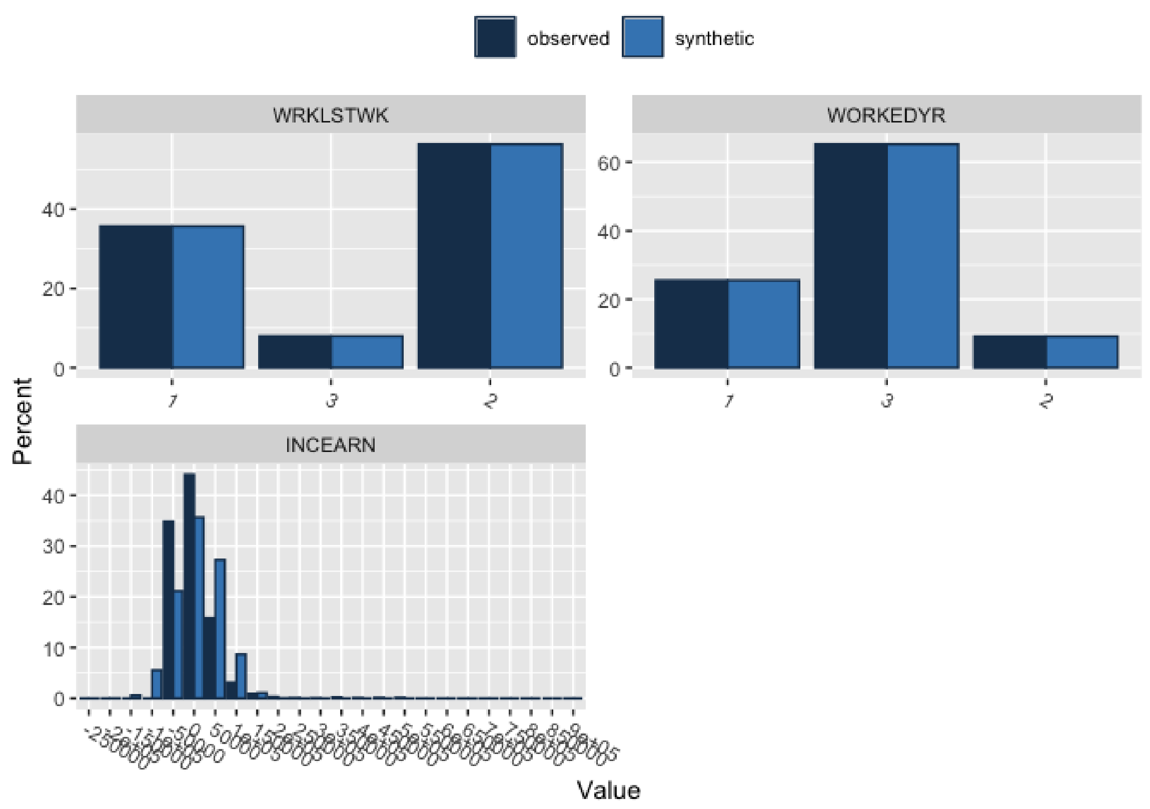
Replication.Uniques	Number.Replications	Percentage.Replications
1001862	177075	17.10537

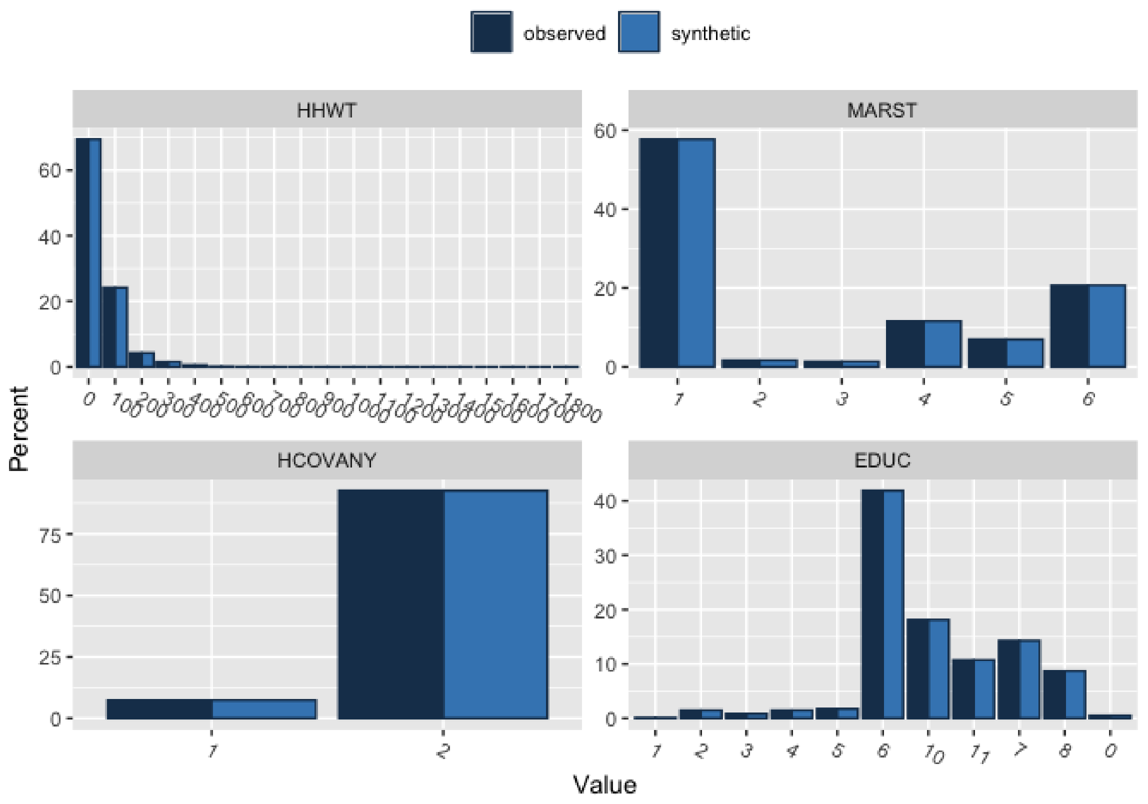
Perceived Disclosure Risk (R-Package: synthpop)

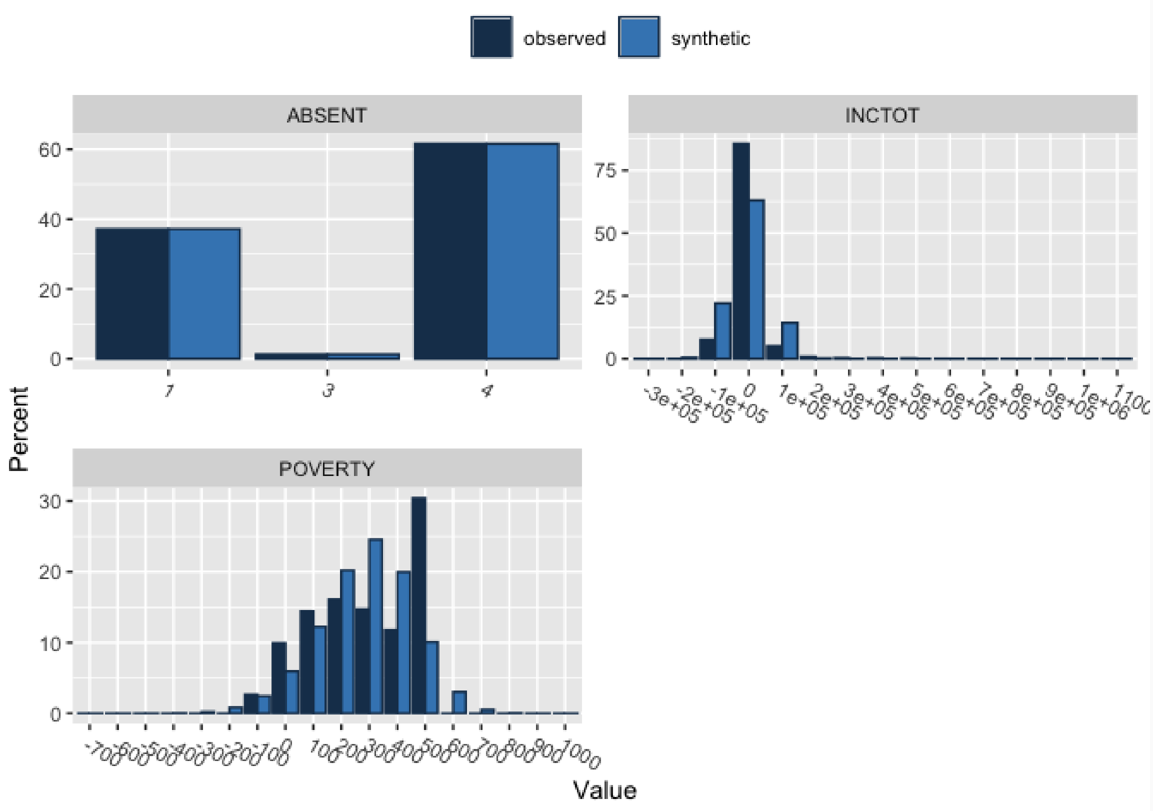
Metric	Number.Uniques	Number.Replications	Percentage.Replications
Perceived Risk	1001862	1001862	96.77947

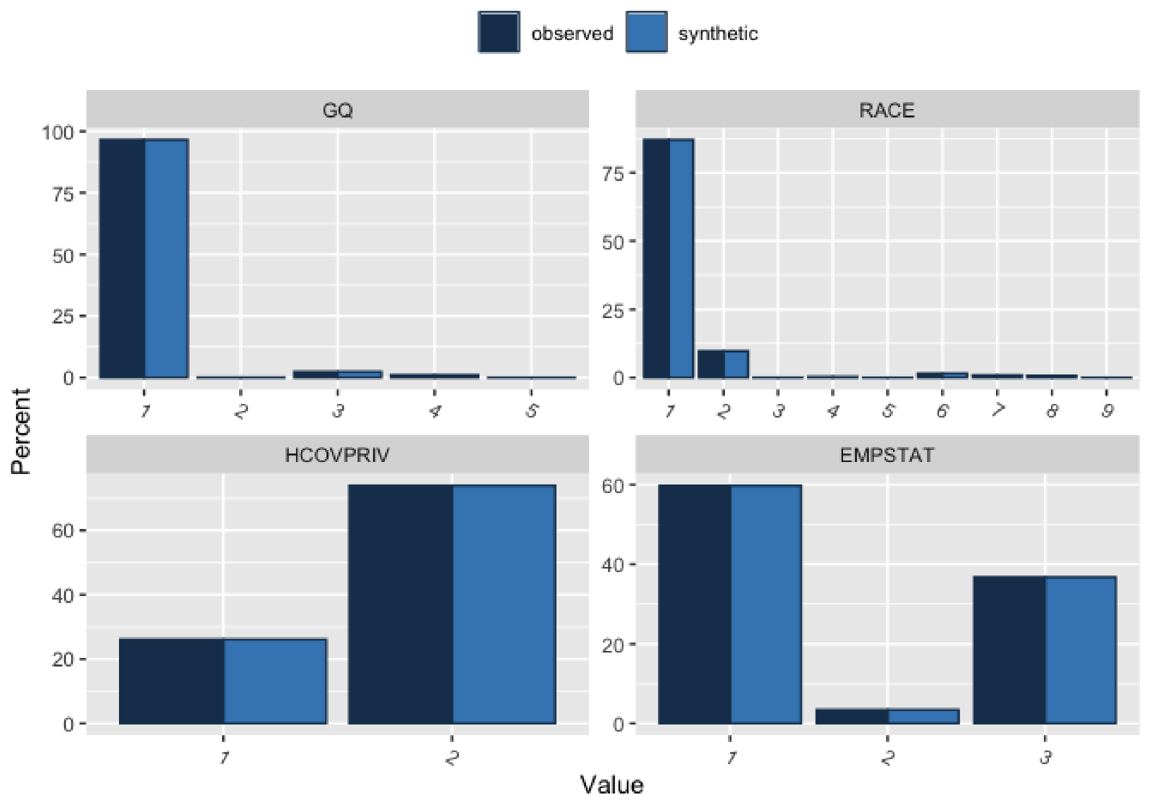
Graphical Comparison for Margins (R-Package: synthpop)

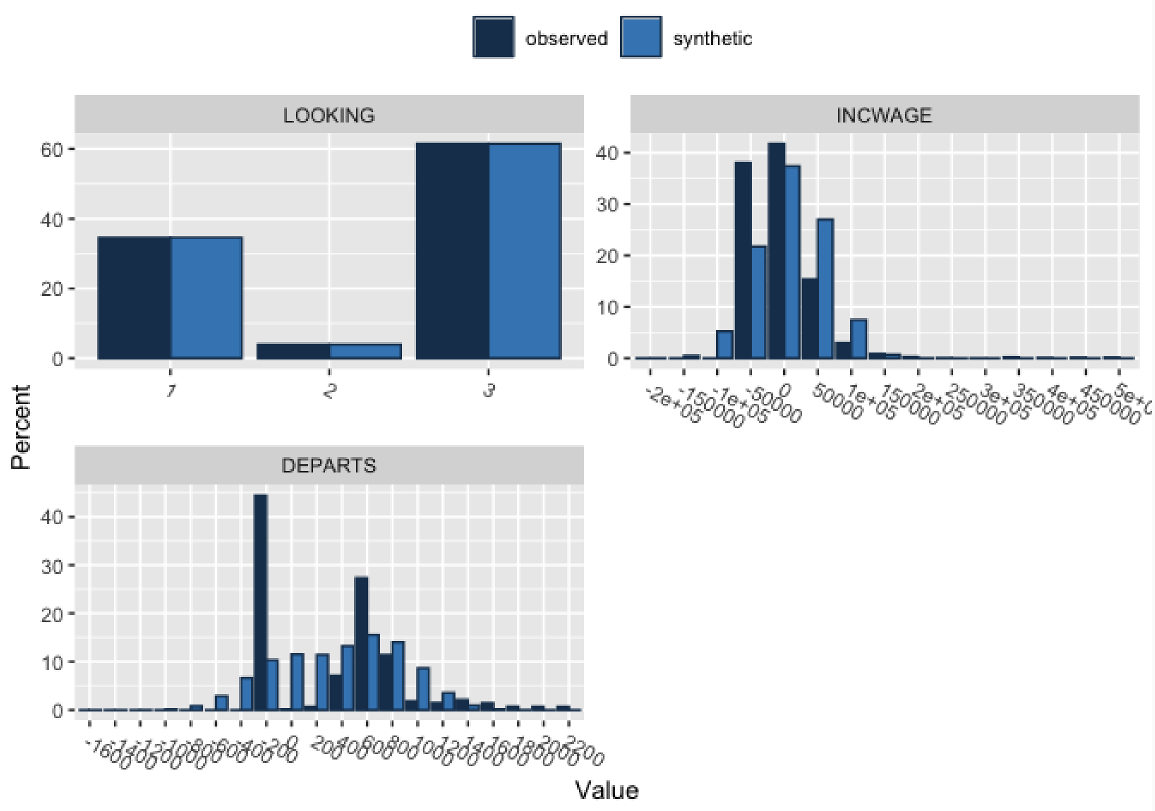

observed
synthetic
YEAR
AGE
SPEAKENG
HINSCARE
Percent
Value
2012
2013
2014
2015
2016
2017
2018
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
1
3
4
5
6
2

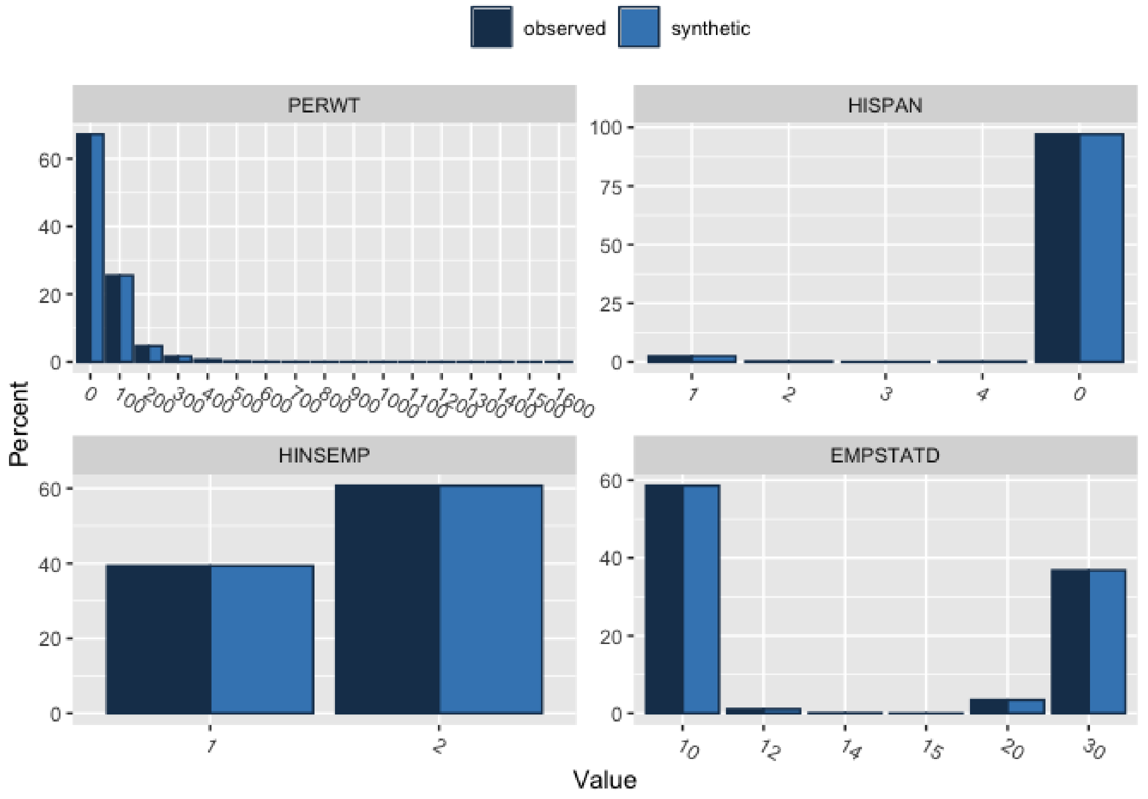

observed
synthetic
WRKLSTWK
WORKEDYR
INCEARN
Percent
Value
1
3
2


observed
synthetic
HHWT
MARST
HCOVANY
EDUC
Percent
Value

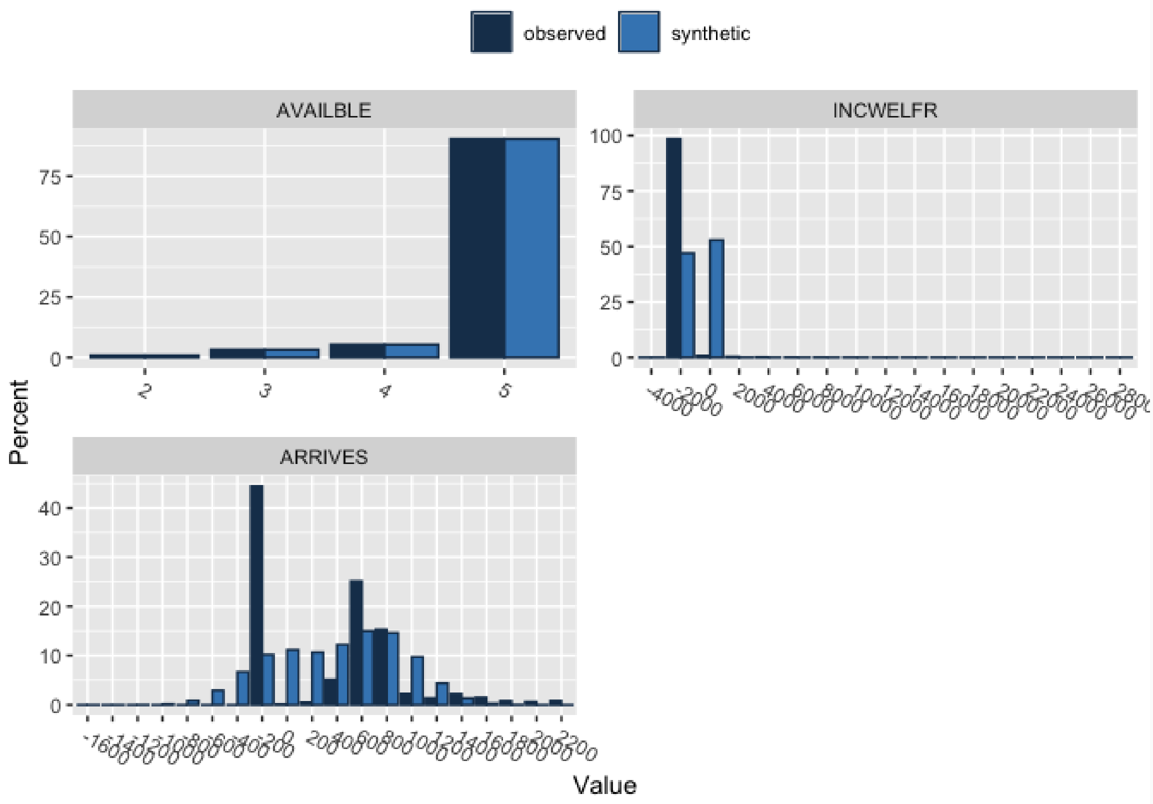

observed
synthetic
ABSENT
INCTOT
POVERTY
Percent
Value


observed
synthetic
GQ
RACE
HCOVPRIV
EMPSTAT
Percent
Value

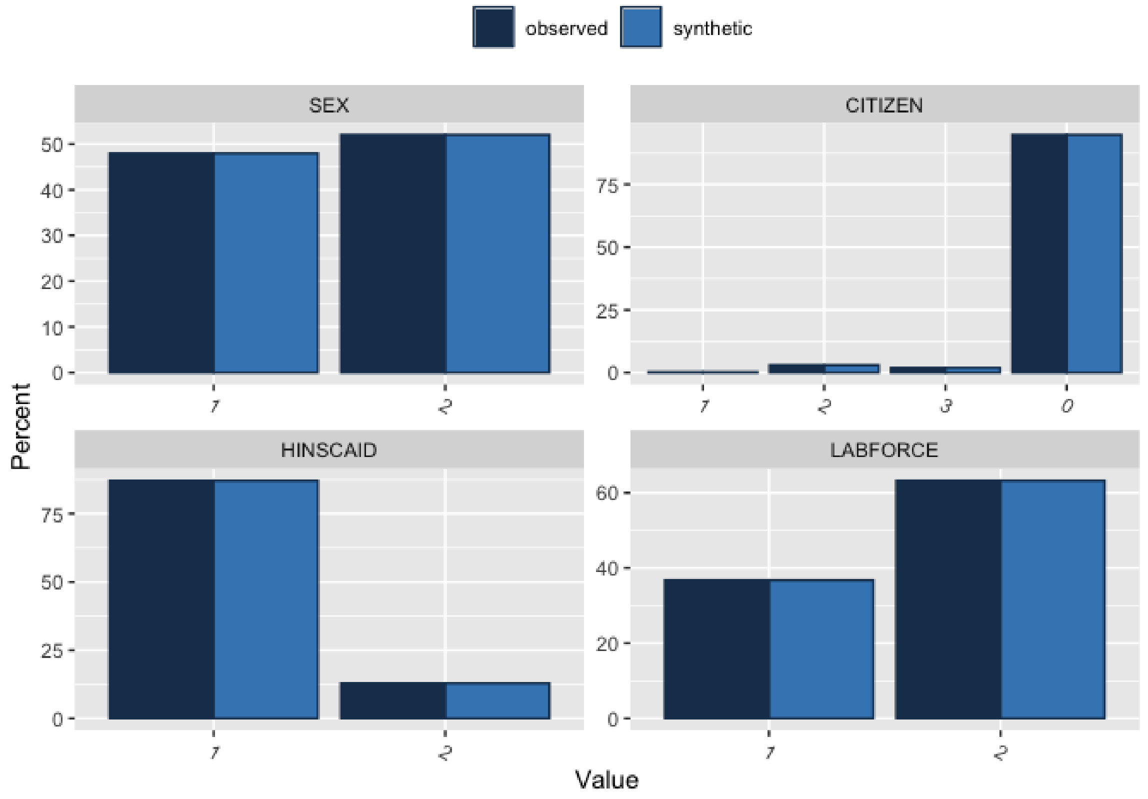

observed
synthetic
LOOKING
INCWAGE
DEPARTS
Percent
Value

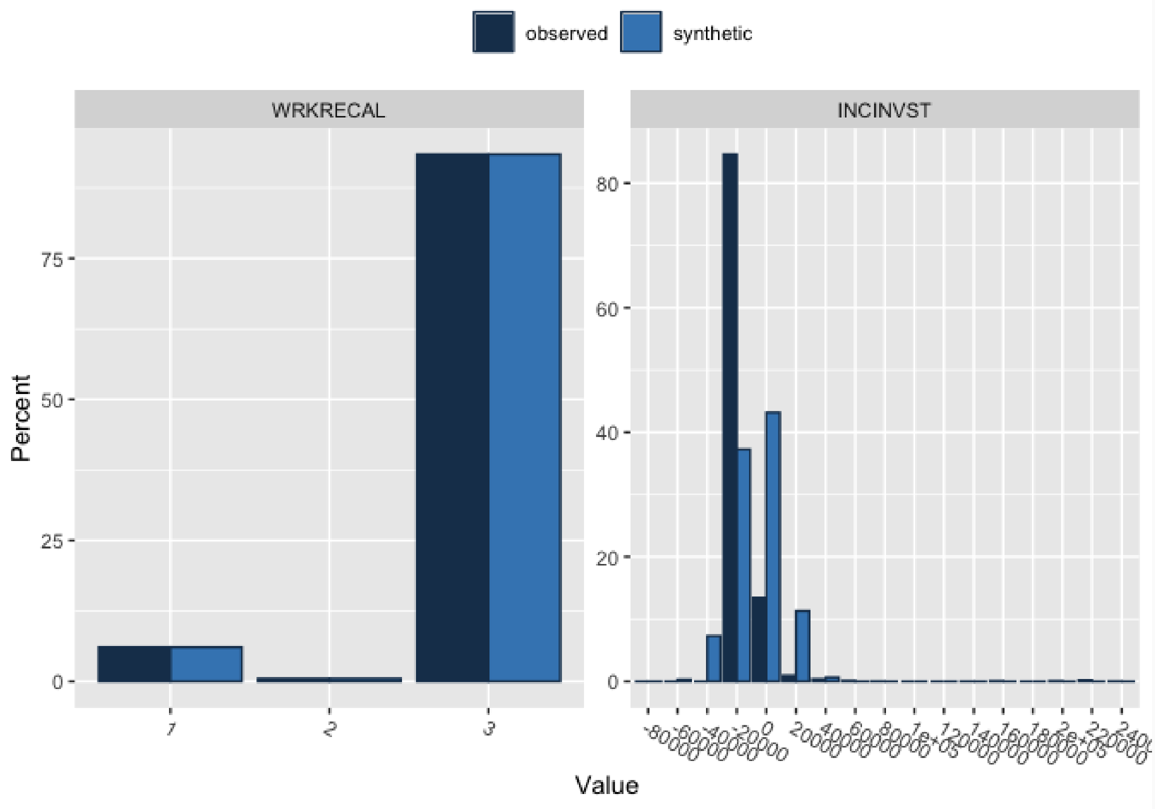


observed
synthetic
PERWT
HISPAN
HINSEMP
EMPSTATD
Percent
Value

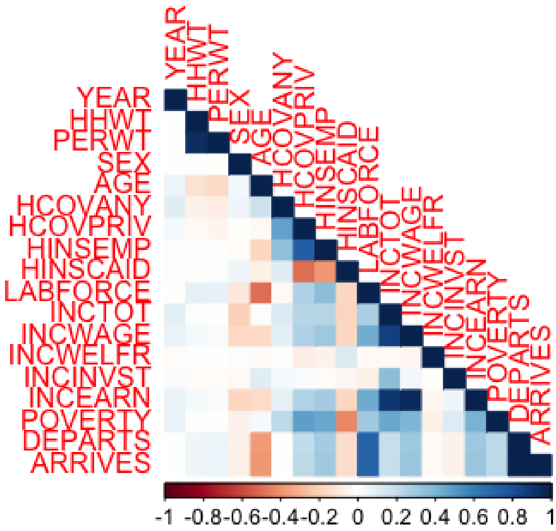


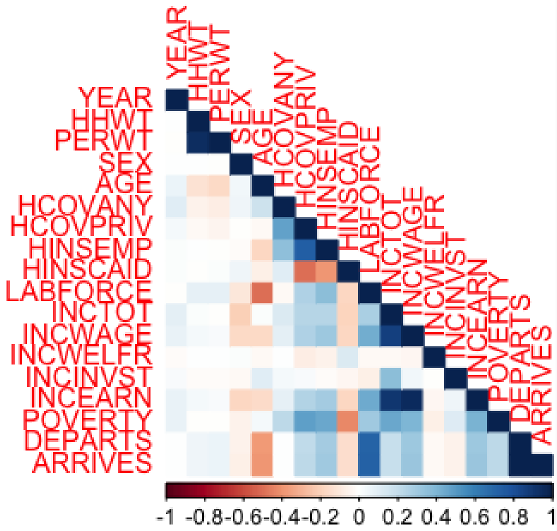
observed
synthetic
AVAILBLE
INCWELFR
ARRIVES
Percent
Value





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

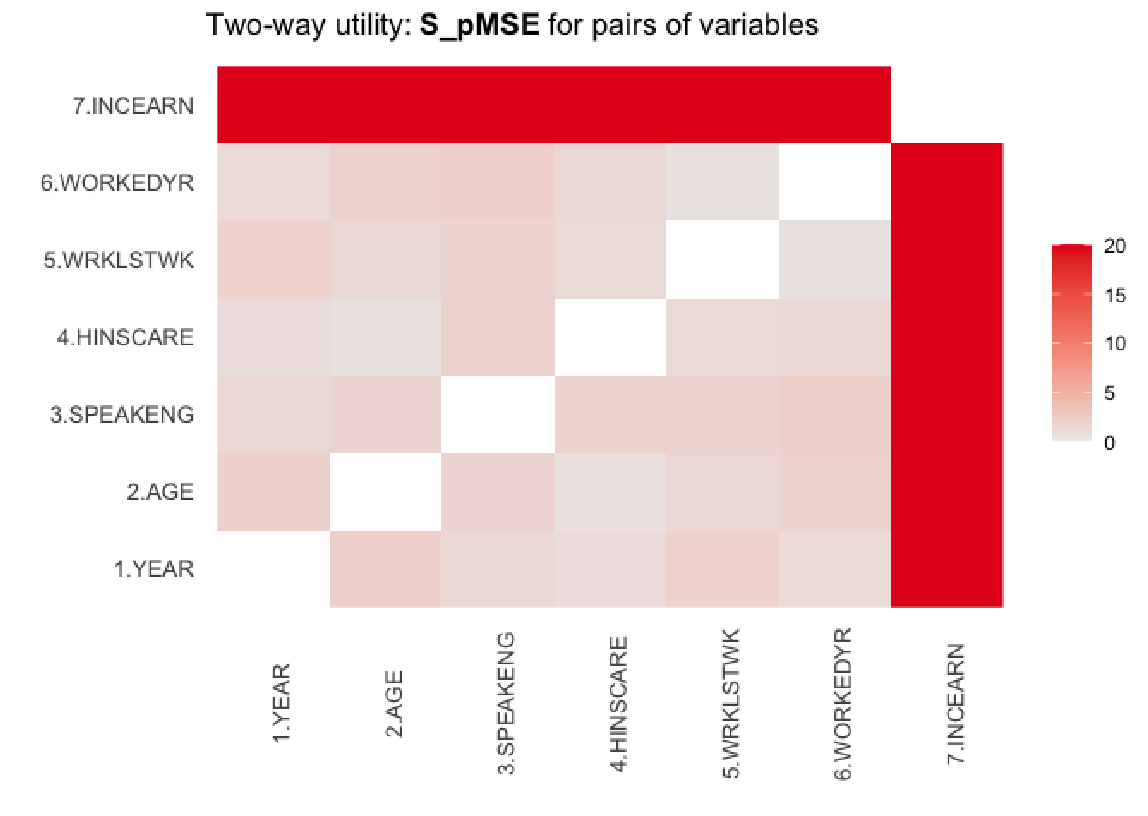
pMSE	S_pMSE
0.2485476	250.5728

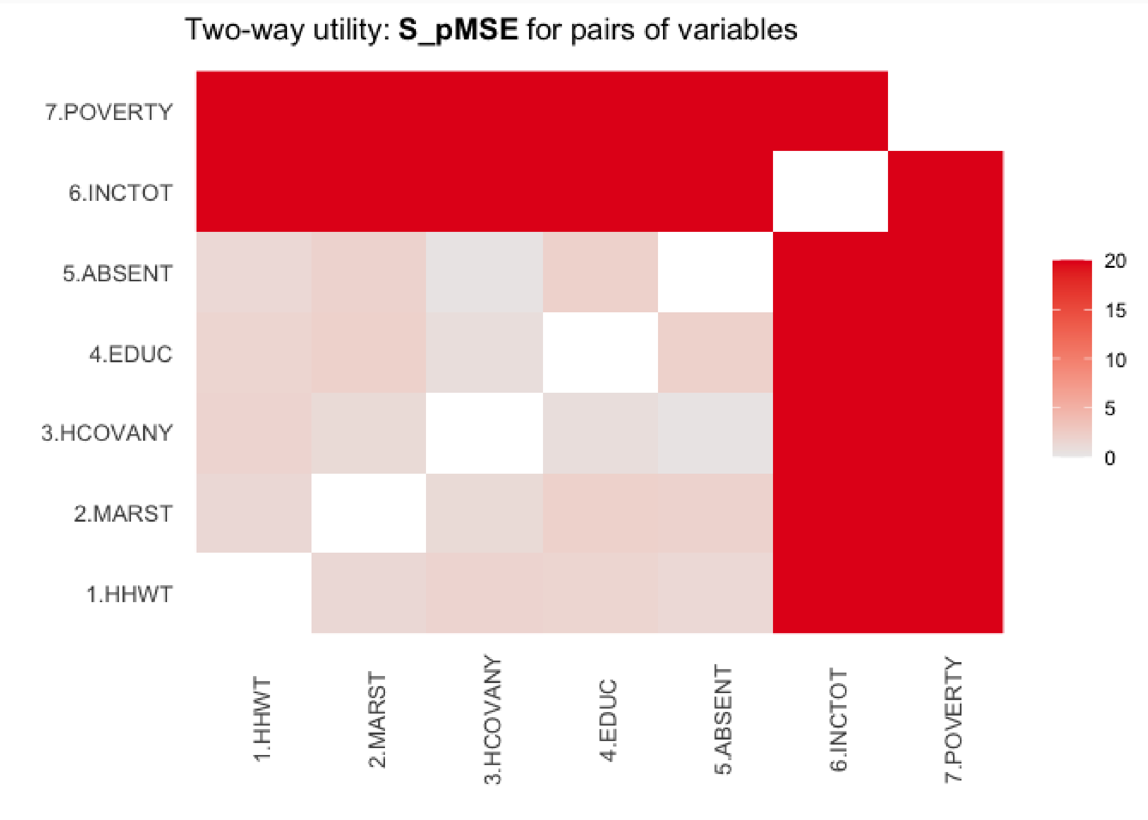
	pMSE	S_pMSE	df
SEX	0.0000000	0.0	1
CITIZEN	0.0000000	0.0	3
HINSCAID	0.0000000	0.0	1
LABFORCE	0.0000000	0.0	1

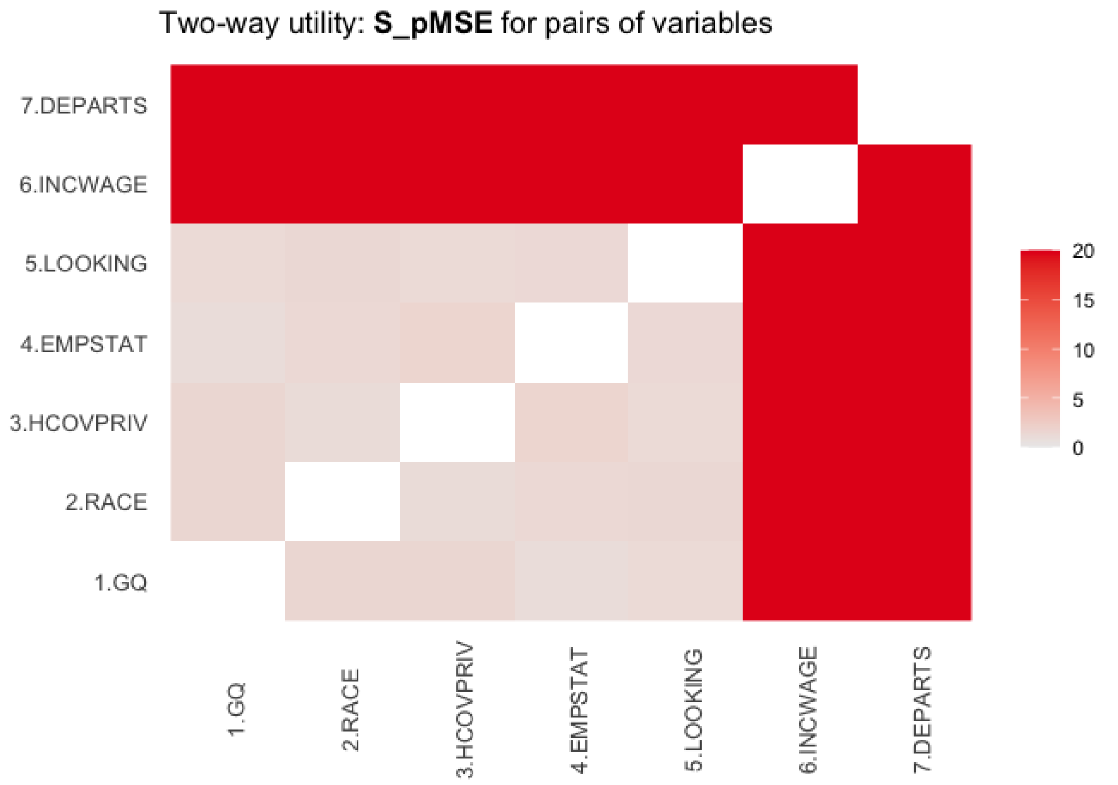
	pMSE	S_pMSE	df
WRKRECAL	0.0000000	0.0	2
INCINVST	0.1483913	819278.8	3

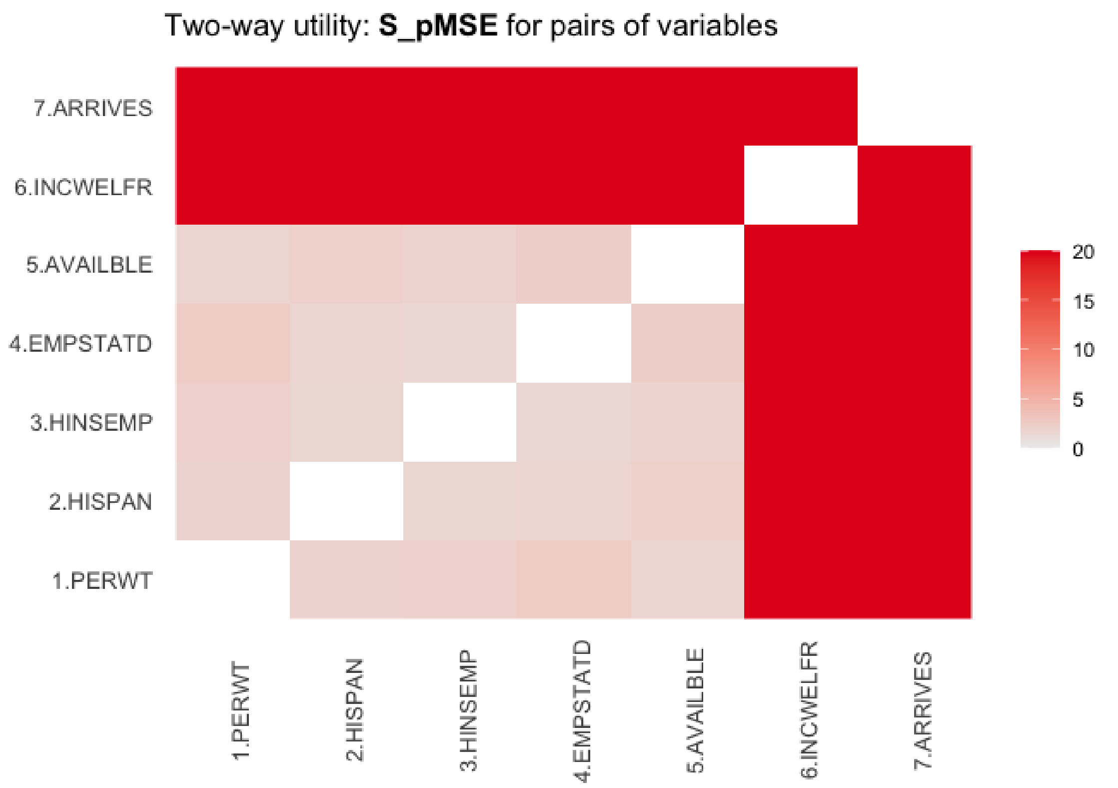
pMSE	S_pMSE
0.2109031	599.3116

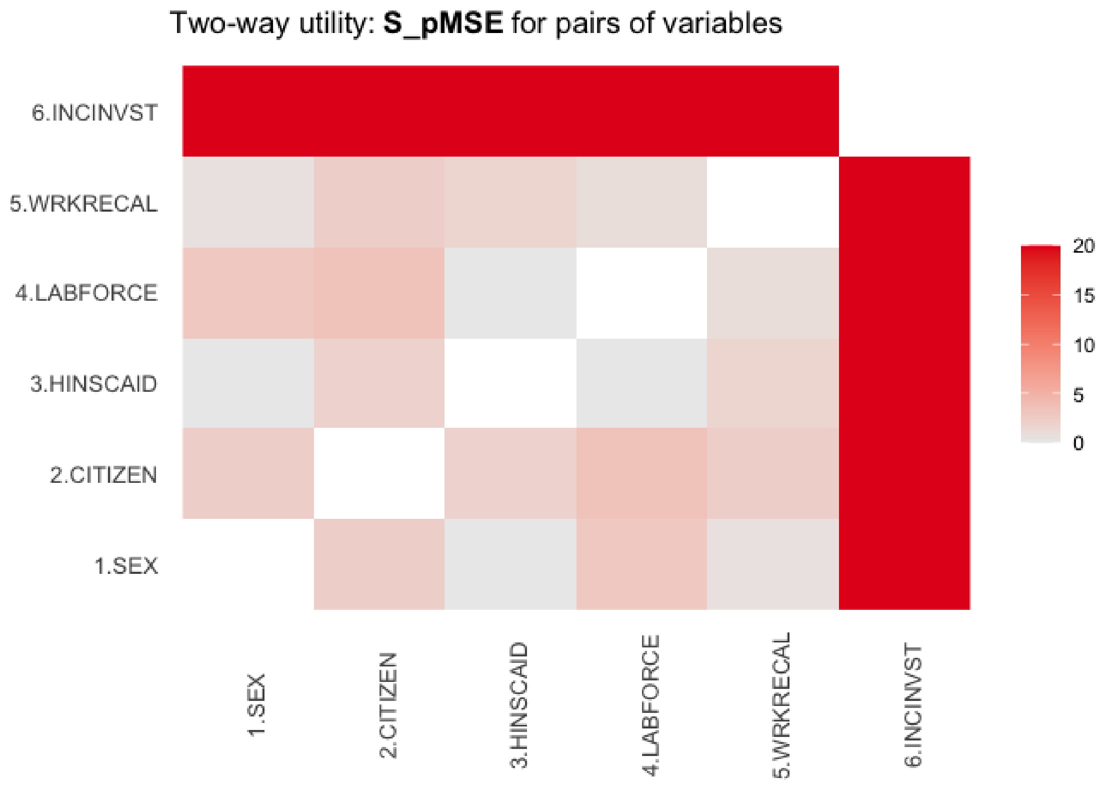
Two-way Tables Comparison of Synthesised Data (R-Package: synthpop) by (S_)pMSE

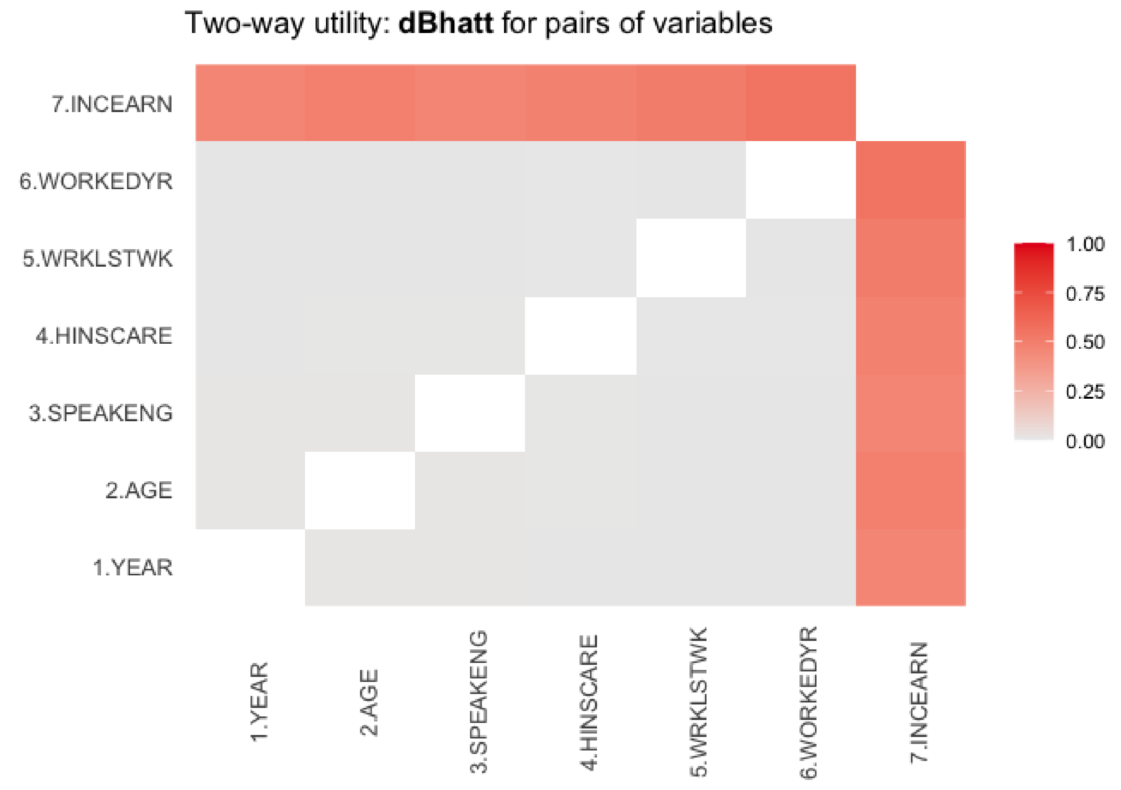

Two-way utility: S_pMSE for pairs of variables
7.INCEARN
6.WORKEDYR
5.WRKLSTWK
4.HINSCARE
3.SPEAKENG
2.AGE
1.YEAR
1.YEAR
2.AGE
3.SPEAKENG
4.HINSCARE
5.WRKLSTWK
6.WORKEDYR
7.INCEARN
20
15
10
5
0

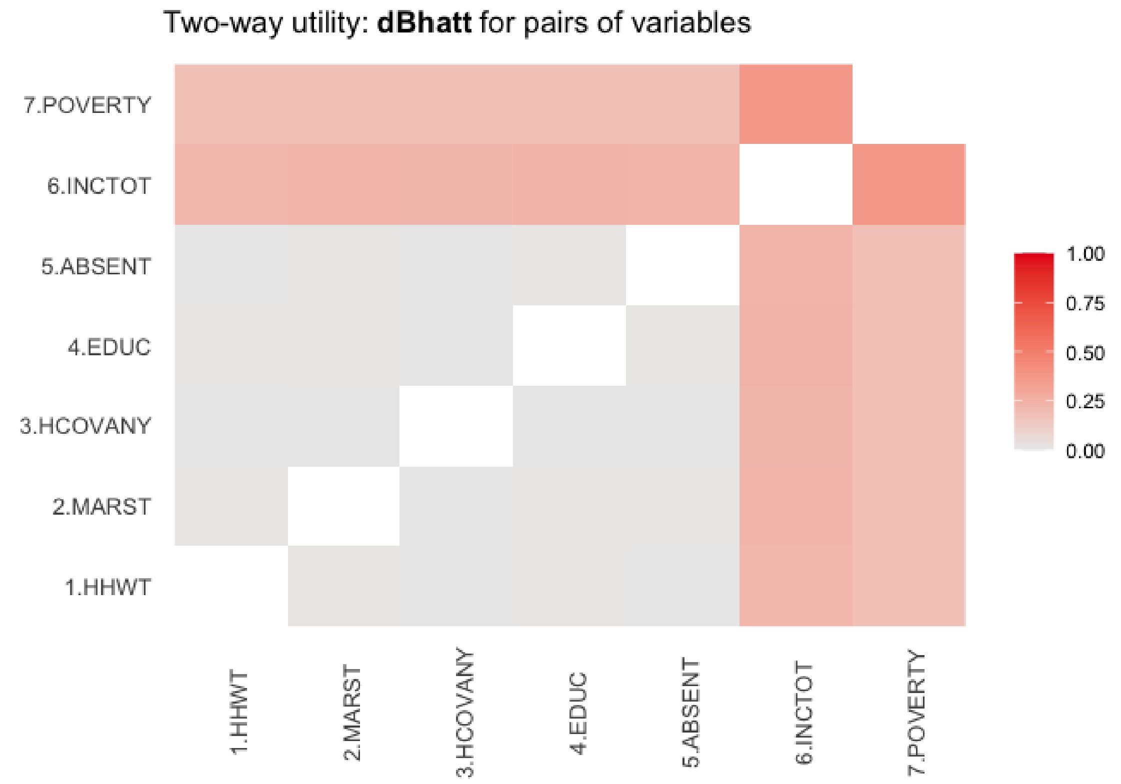

Two-way utility: S_pMSE for pairs of variables
7.POVERTY
6.INCTOT
5.ABSENT
4.EDUC
3.HCOVANY
2.MARST
1.HHWT
1.HHWT
2.MARST
3.HCOVANY
4.EDUC
5.ABSENT
6.INCTOT
7.POVERTY
20
15
10
5
0

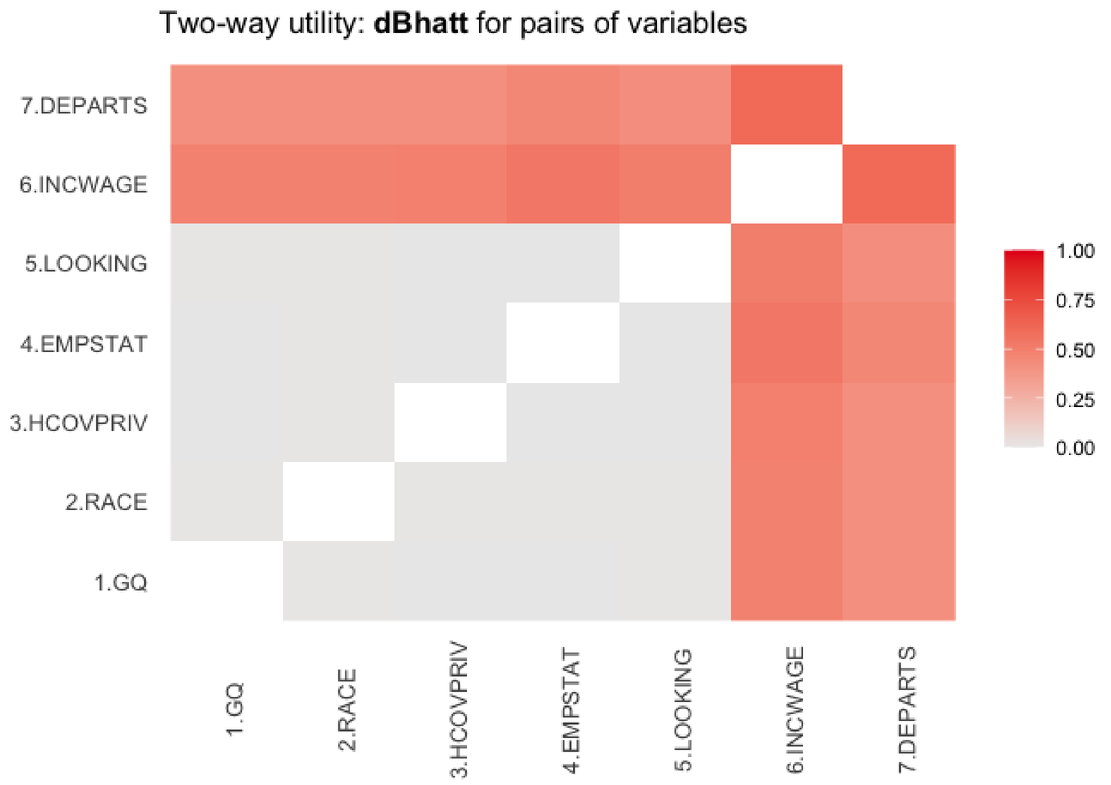

Two-way utility: **S_pMSE** for pairs of variables
7.DEPARTS
6.INCWAGE
5.LOOKING
4.EMPSTAT
3.HCOVPRIV
2.RACE
1.GQ
1.GQ
2.RACE
3.HCOVPRIV
4.EMPSTAT
5.LOOKING
6.INCWAGE
7.DEPARTS
20
15
10
5
0

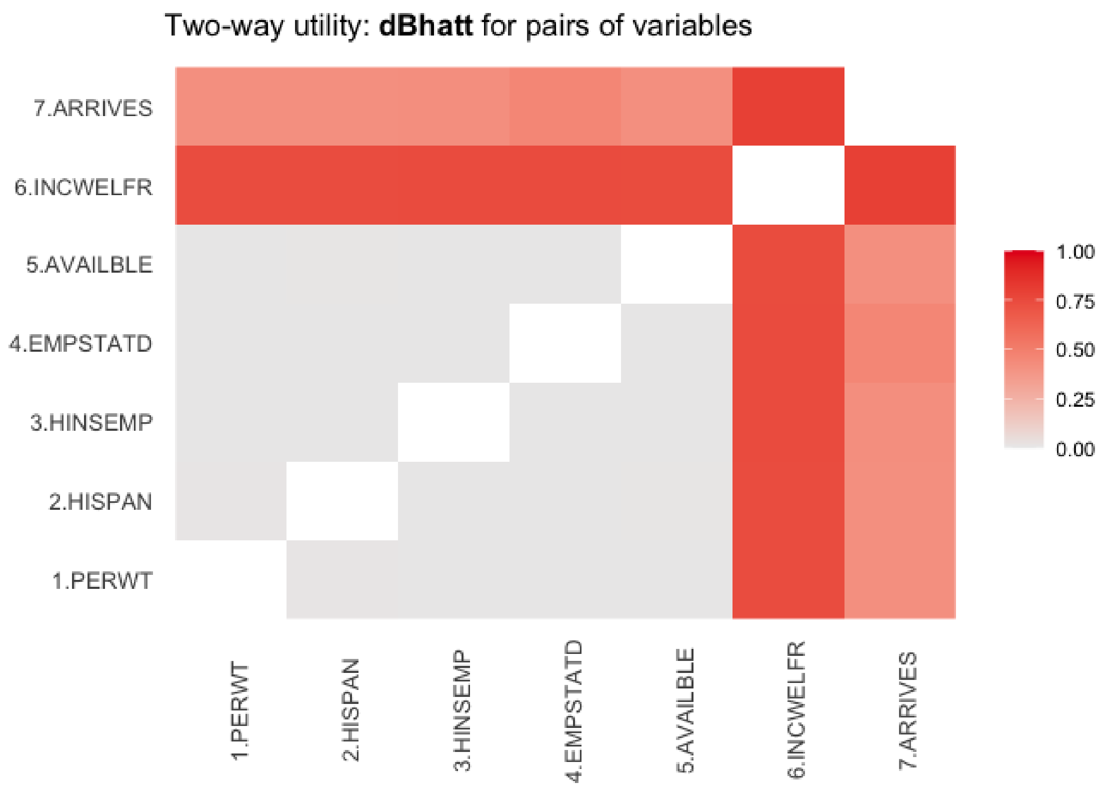

Two-way utility: **S_pMSE** for pairs of variables
7.ARRIVES
6.INCWELFR
5.AVAILBLE
4.EMPSTATD
3.HINSEMP
2.HISPAN
1.PERWT
1.PERWT
2.HISPAN
3.HINSEMP
4.EMPSTATD
5.AVAILBLE
6.INCWELFR
7.ARRIVES
20
15
10
5
0

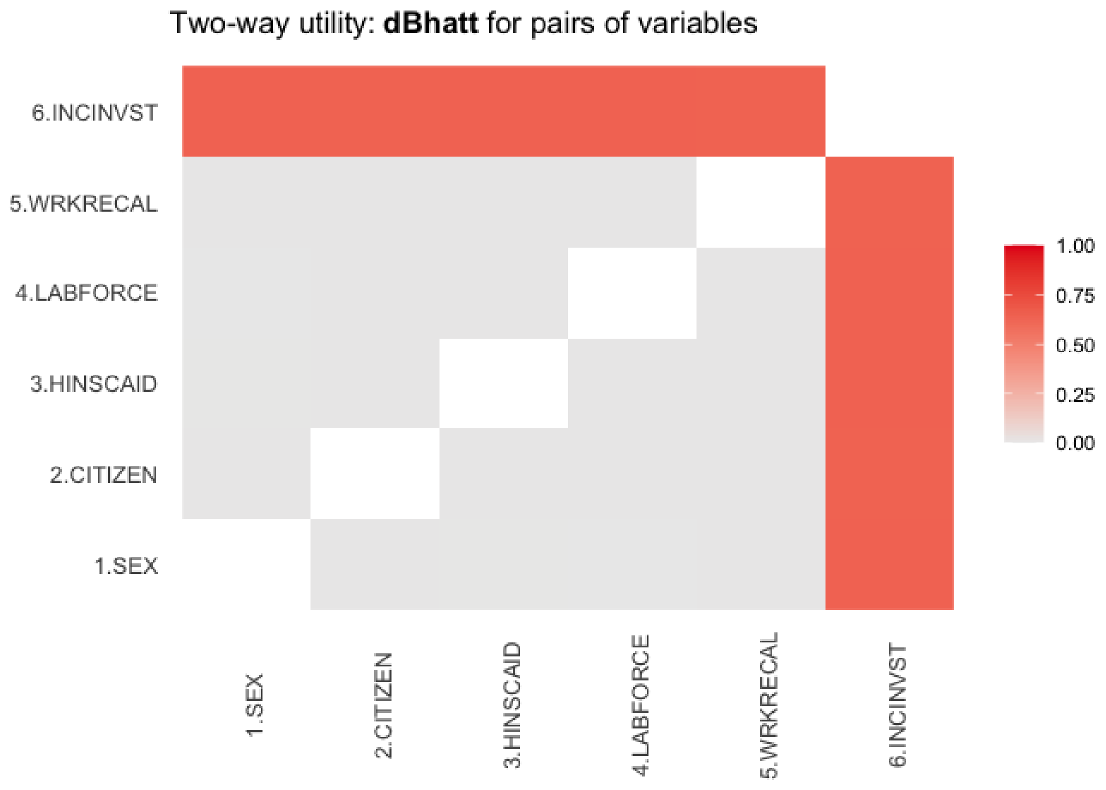

Two-way utility: S_pMSE for pairs of variables
6.INCINVST
5.WRKRECAL
4.LABFORCE
3.HINSCAID
2.CITIZEN
1.SEX
1.SEX
2.CITIZEN
3.HINSCAID
4.LABFORCE
5.WRKRECAL
6.INCINVST
20
15
10
5
0

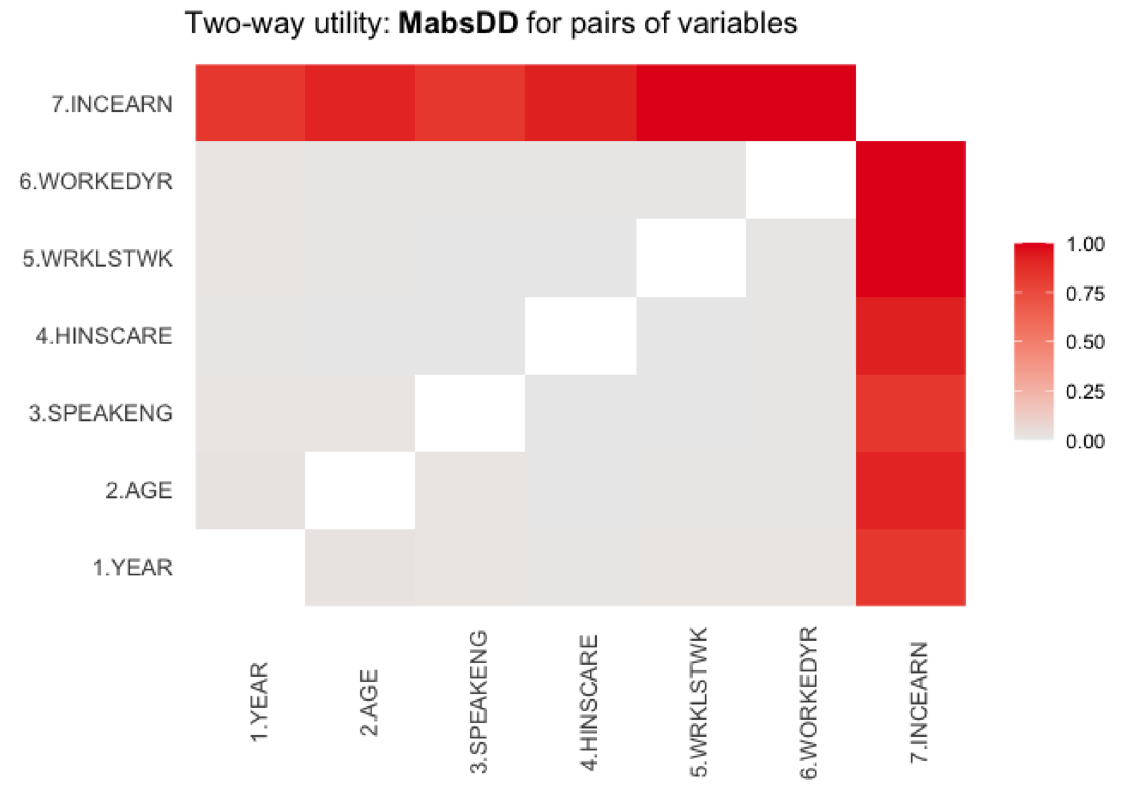

Two-way utility: dBhatt for pairs of variables
7.INCEARN
6.WORKEDYR
5.WRKLSTWK
4.HINSCARE
3.SPEAKENG
2.AGE
1.YEAR
1.00
0.75
0.50
0.25
0.00

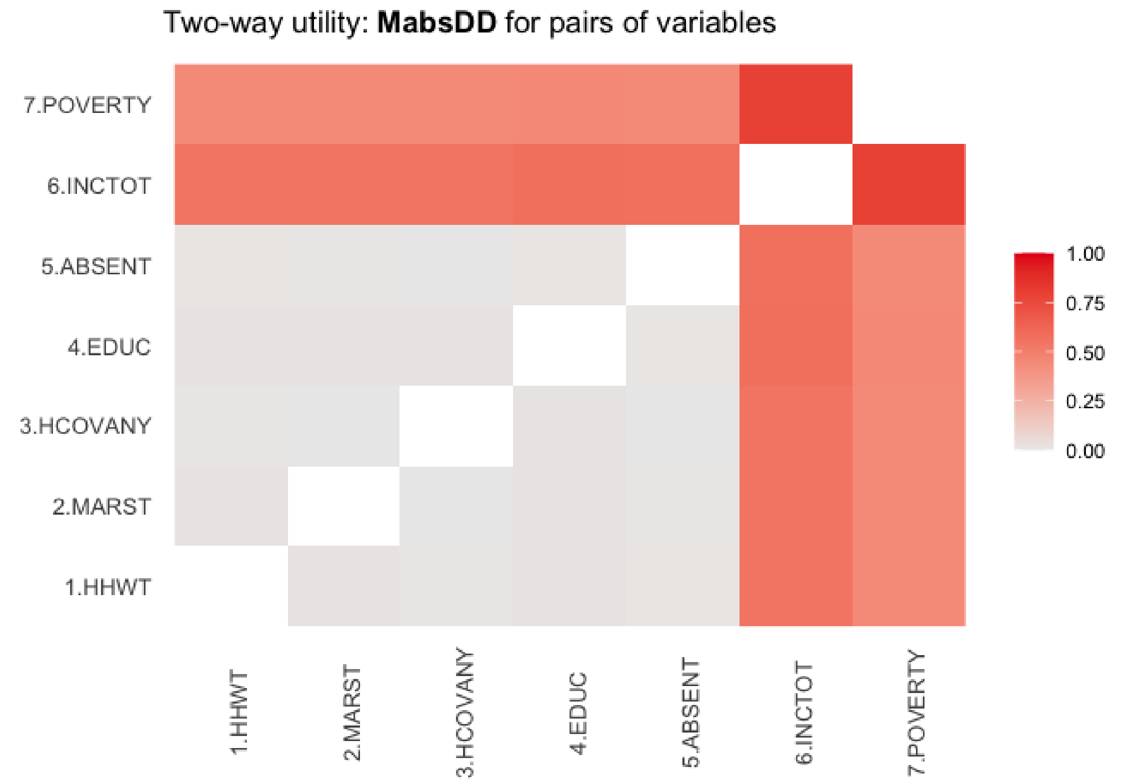

Two-way utility: dBhatt for pairs of variables
7.POVERTY
6.INCTOT
5.ABSENT
4.EDUC
3.HCOVANY
2.MARST
1.HHWT
1.00
0.75
0.50
0.25
0.00


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

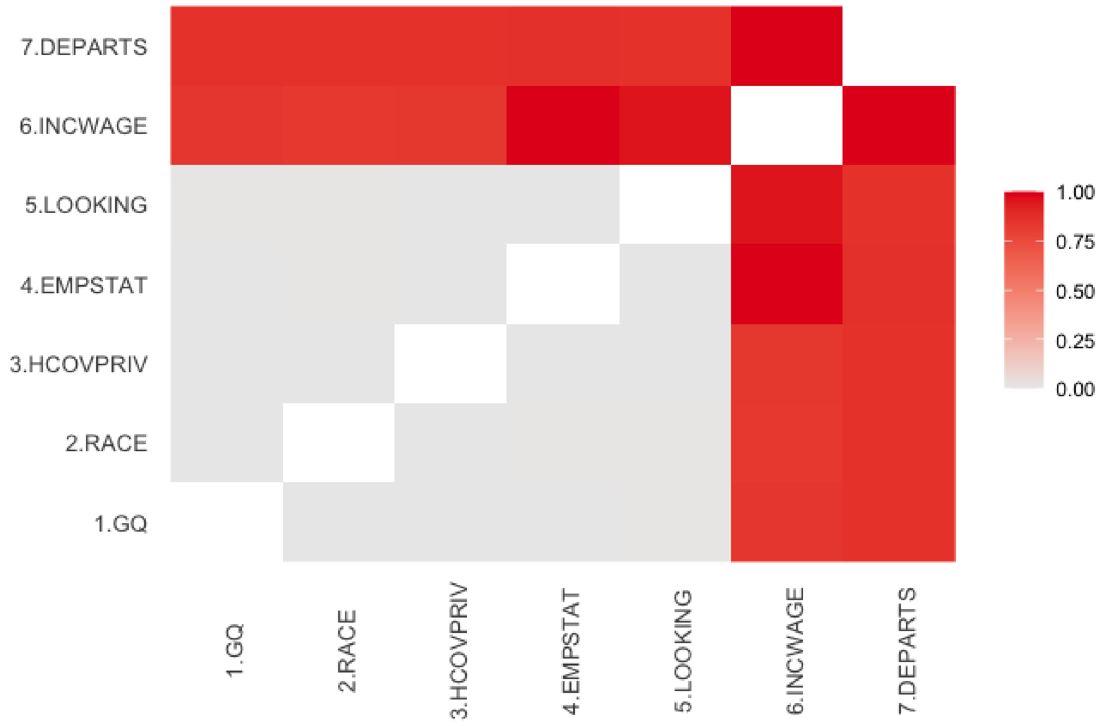

Two-way utility: dBhatt for pairs of variables
7.ARRIVES
6.INCWELFR
5.AVAILBLE
4.EMPSTATD
3.HINSEMP
2.HISPAN
1.PERWT
1.PERWT
2.HISPAN
3.HINSEMP
4.EMPSTATD
5.AVAILBLE
6.INCWELFR
7.ARRIVES
1.00
0.75
0.50
0.25
0.00


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

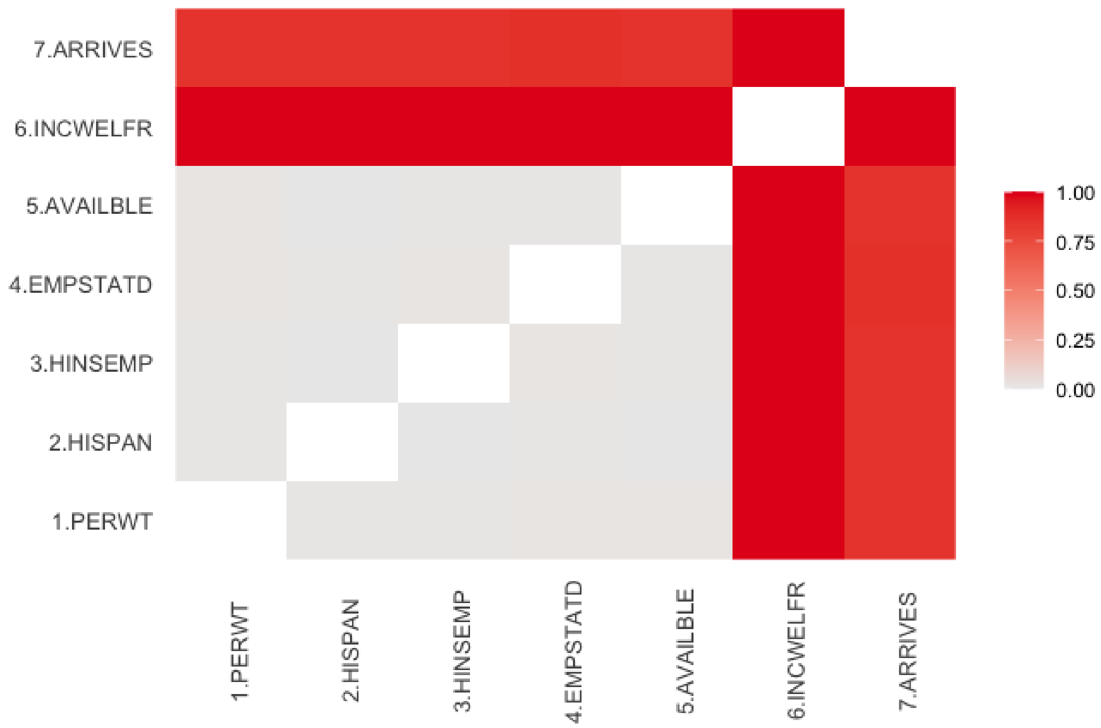

Two-way utility: MabsDD for pairs of variables
7.INCEARN
6.WORKEDYR
5.WRKLSTWK
4.HINSCARE
3.SPEAKENG
2.AGE
1.YEAR
1.YEAR
2.AGE
3.SPEAKENG
4.HINSCARE
5.WRKLSTWK
6.WORKEDYR
7.INCEARN
1.00
0.75
0.50
0.25
0.00

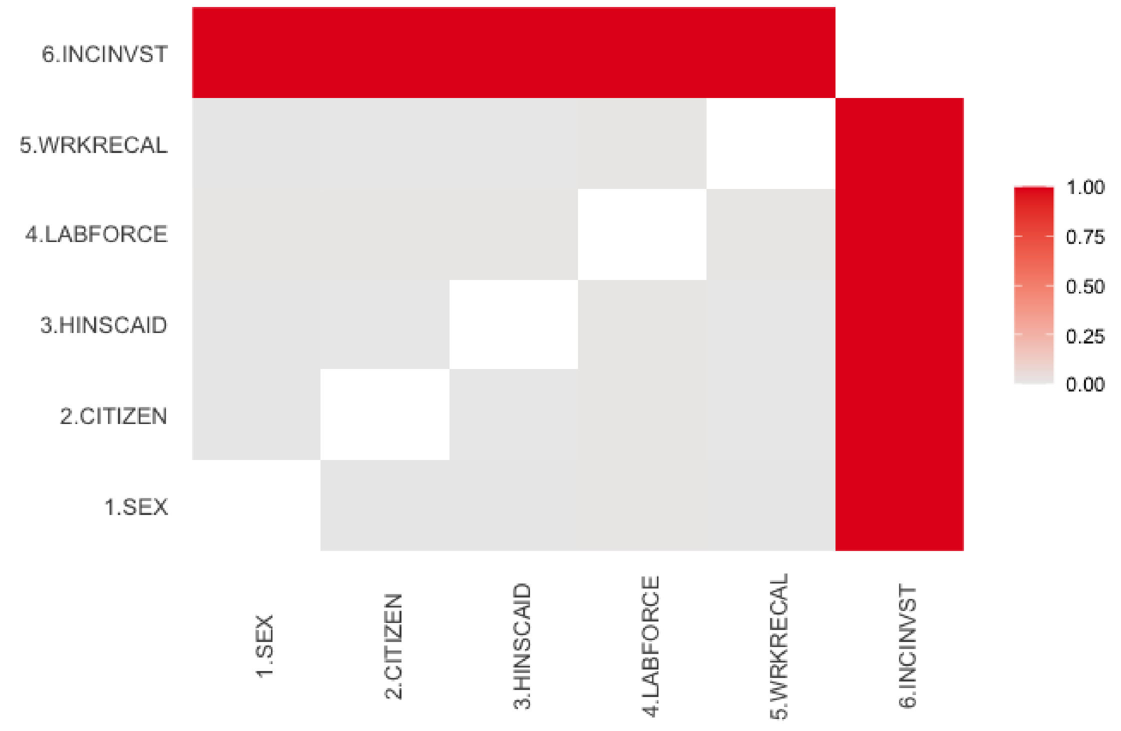

Two-way utility: MabsDD for pairs of variables
7.POVERTY
6.INCTOT
5.ABSENT
4.EDUC
3.HCOVANY
2.MARST
1.HHWT
1.HHWT
2.MARST
3.HCOVANY
4.EDUC
5.ABSENT
6.INCTOT
7.POVERTY
1.00
0.75
0.50
0.25
0.00

Two-way utility: **MabsDD** for pairs of variables



Two-way utility: **MabsDD** for pairs of variables





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
```

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:

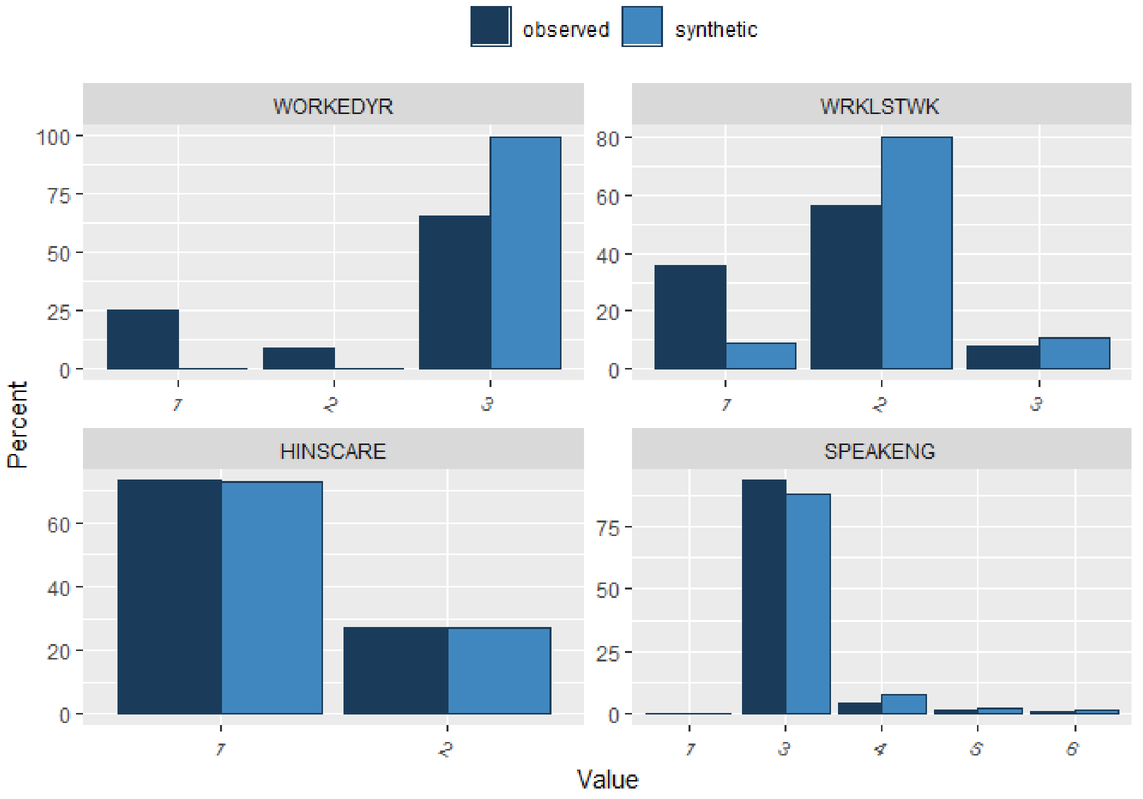
Metric	Number.Uniques	Number.Replications	Percentage.Replications
Perceived Risk	1033709	0	0

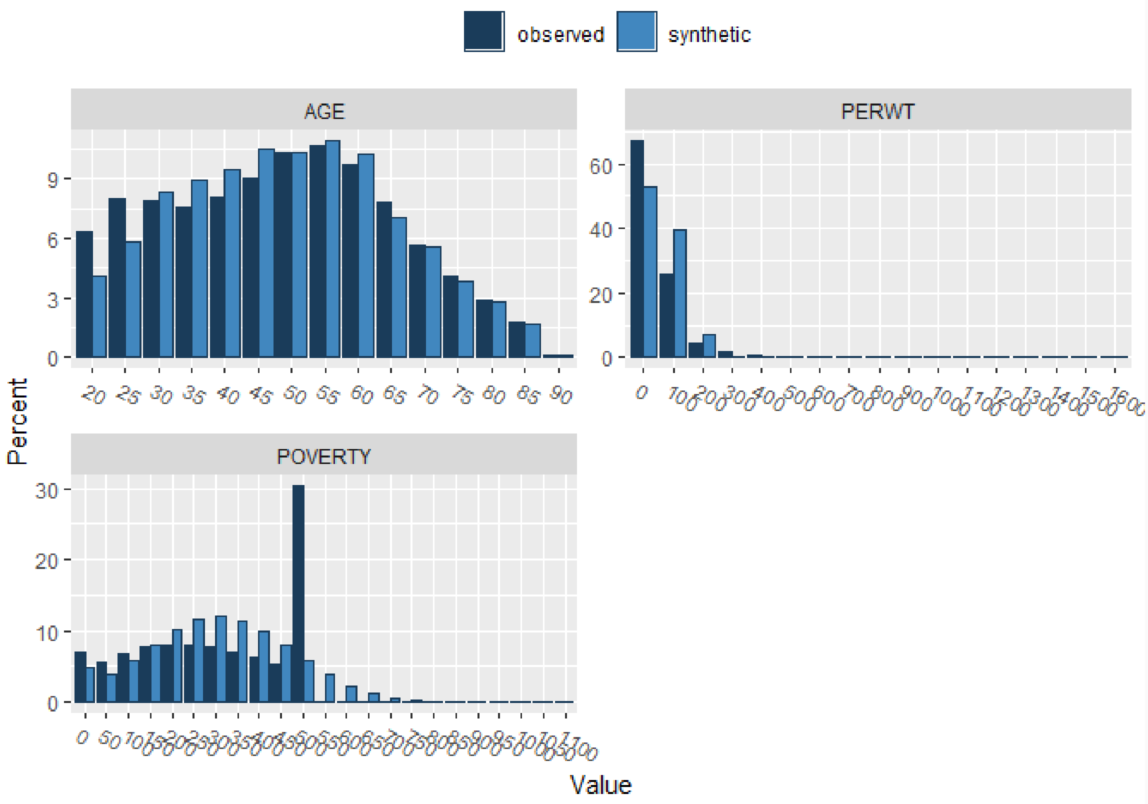
Utility Evaluation

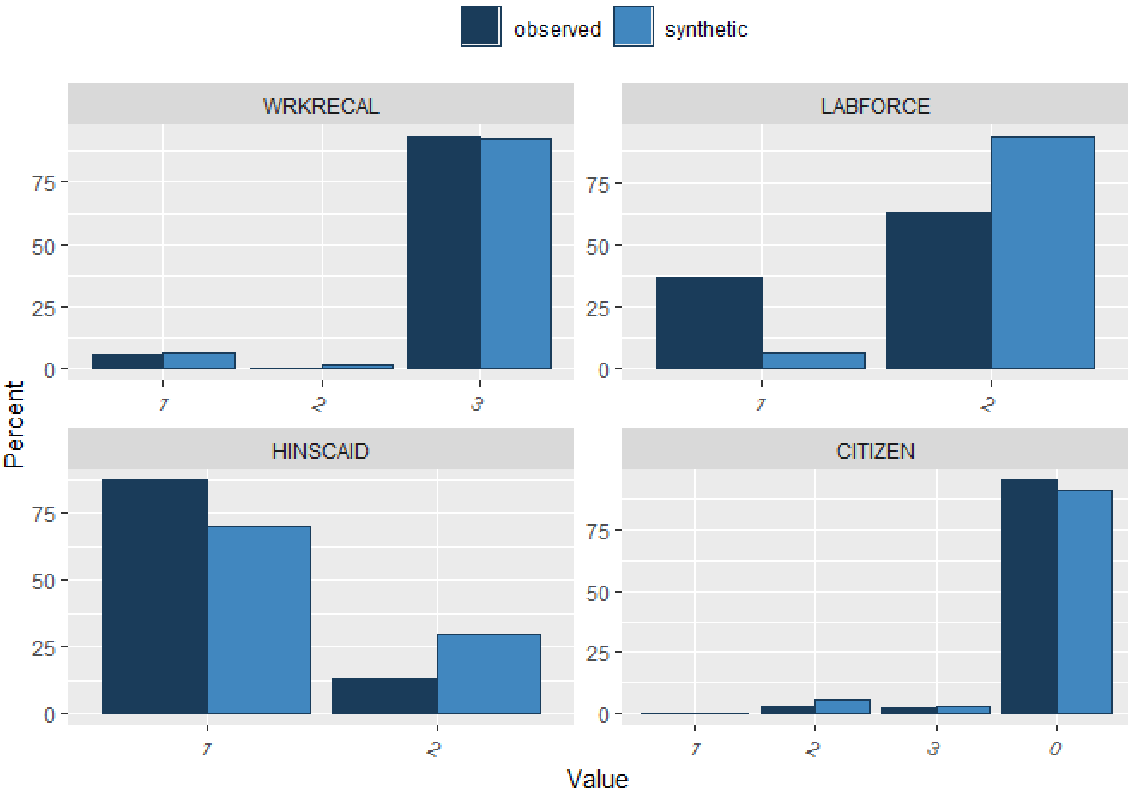
Different utility measures are applied in this section. These utility measures are the basis of utility evaluation for the generated synthetic dataset. The R packages synthpop, sdcMicro and corrplot were used to compute the following metrics. We do not use tests incorporating significance here. Confidence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

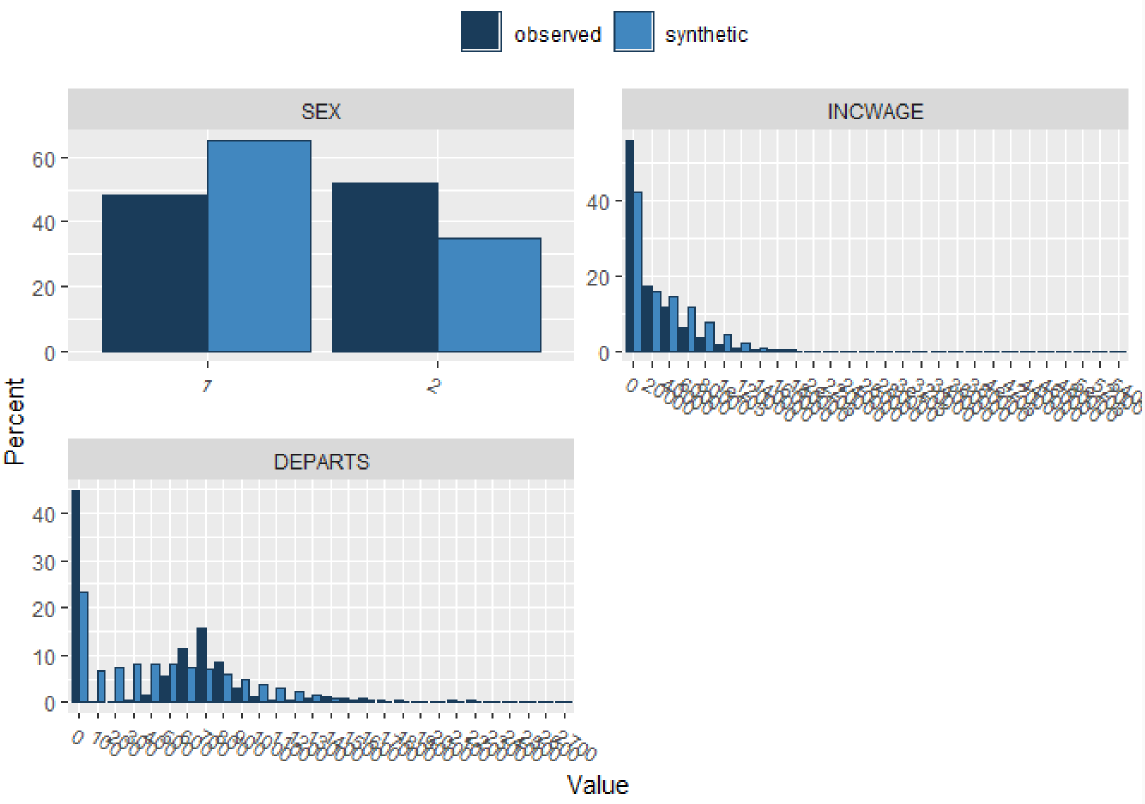
Graphical Comparison for Margins (R-Package: synthpop)

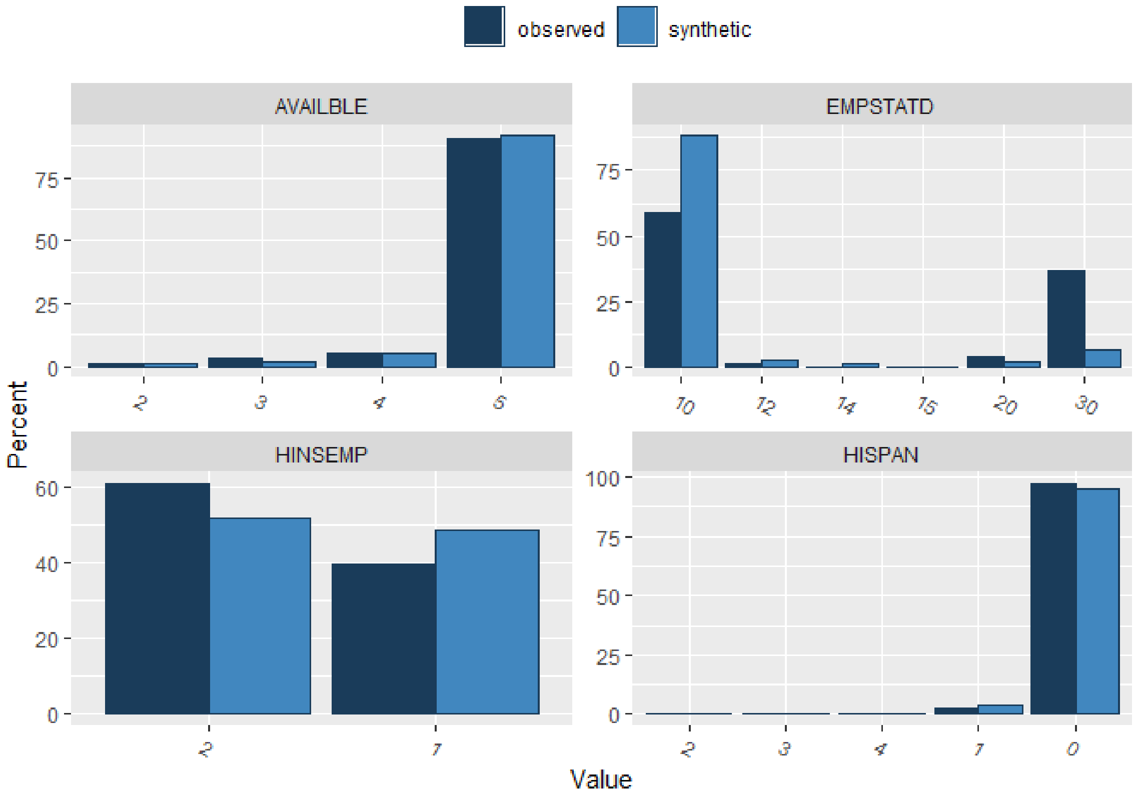
The following histograms provide an ad-hoc overview on the marginal distributions of the original and synthetic dataset. Matching or close distributions are related to a high data utility.

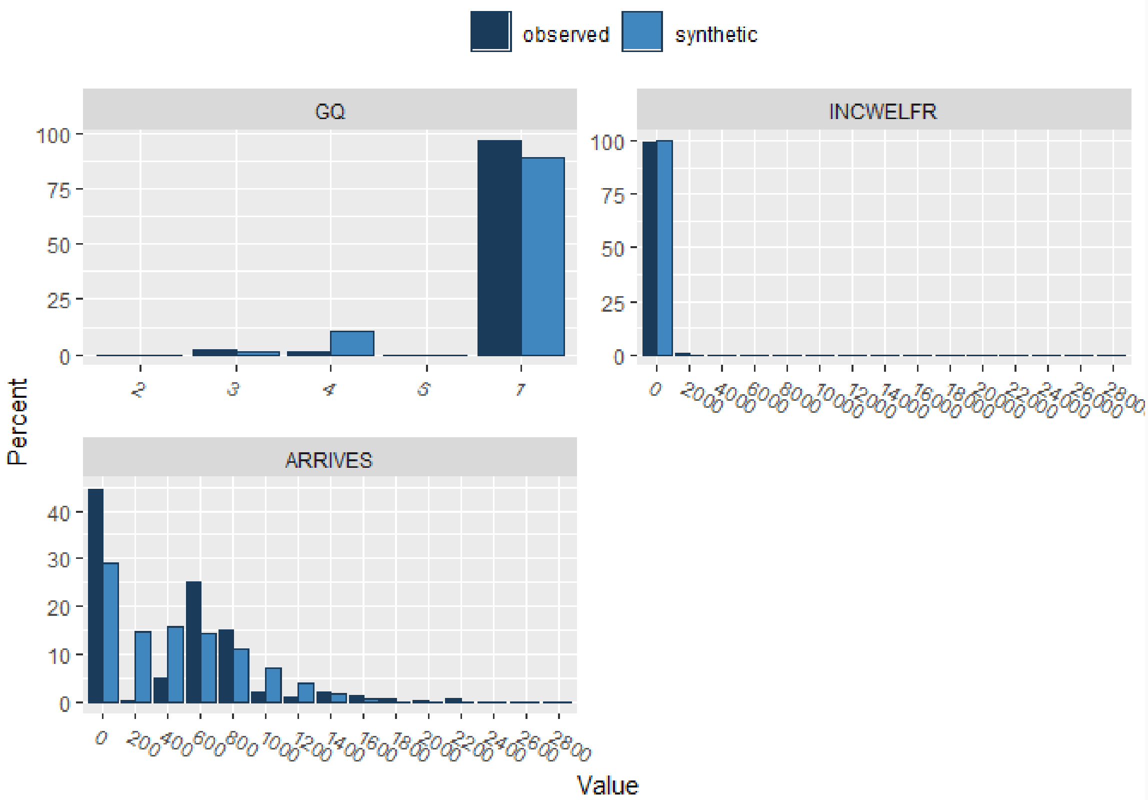

observed
synthetic
WORKEDYR
WRKLSTWK
HINSCARE
SPEAKENG
Percent
Value

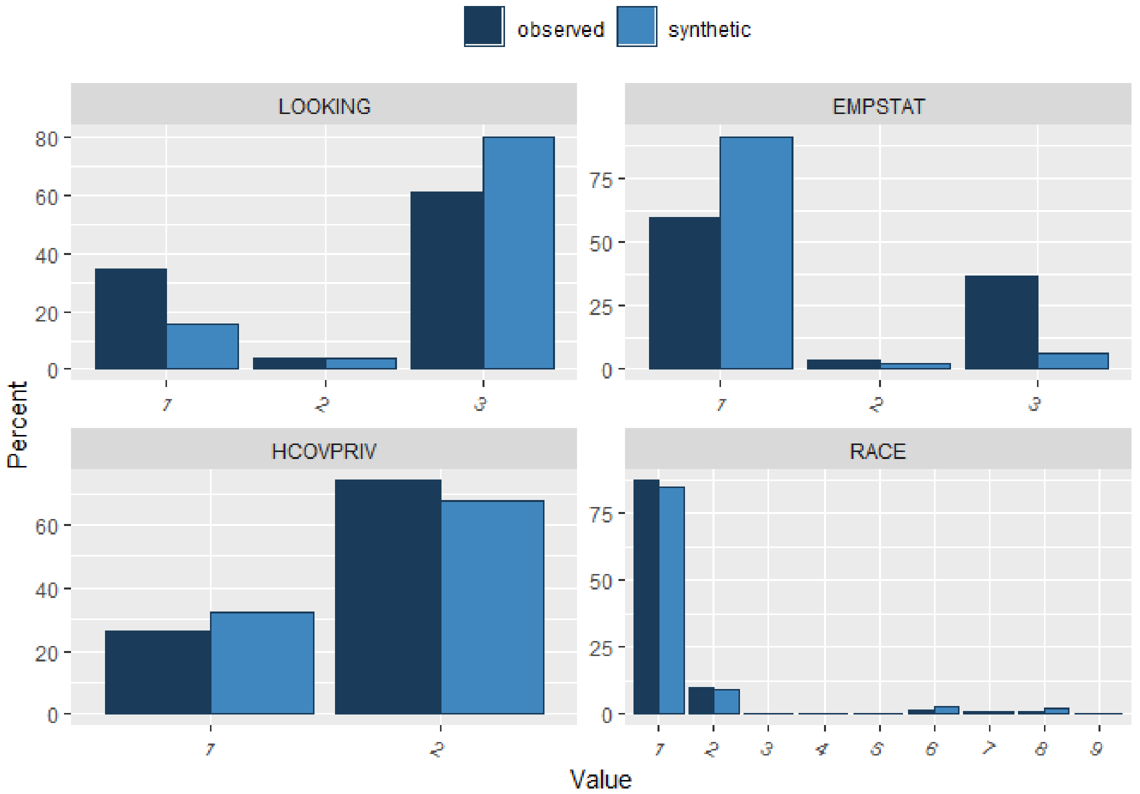

observed
synthetic
AGE
PERWT
POVERTY
Percent
Value

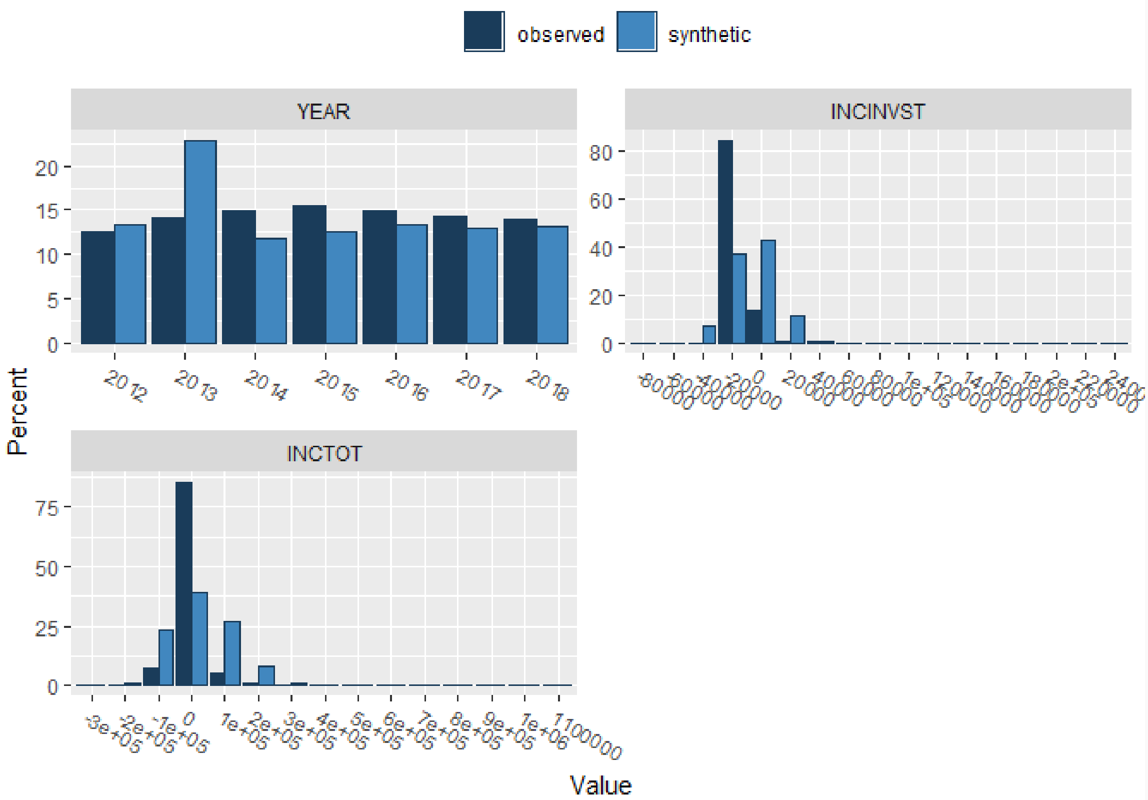

observed
synthetic
WRKRECAL
LABFORCE
HINSCAID
CITIZEN
Percent
Value
0
25
50
75
1
2
3
0

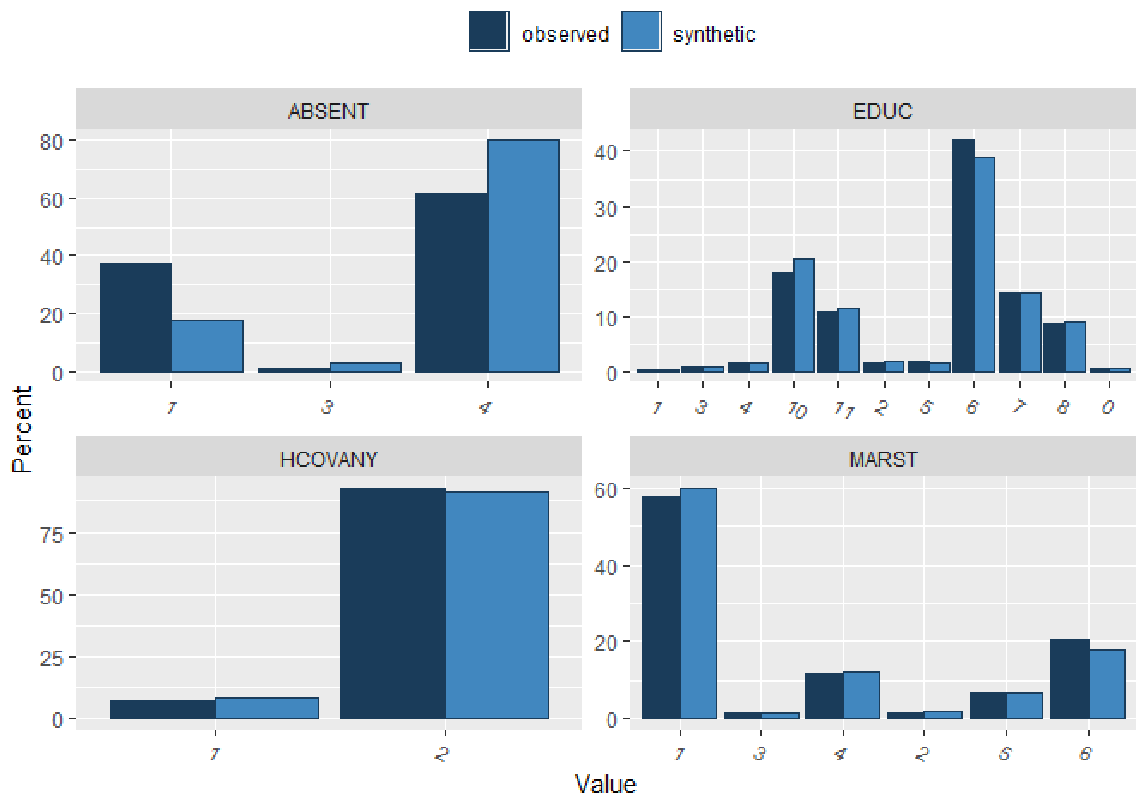

observed
synthetic
SEX
INCWAGE
DEPARTS
Percent
Value
0
10
20
30
40
60
1
2

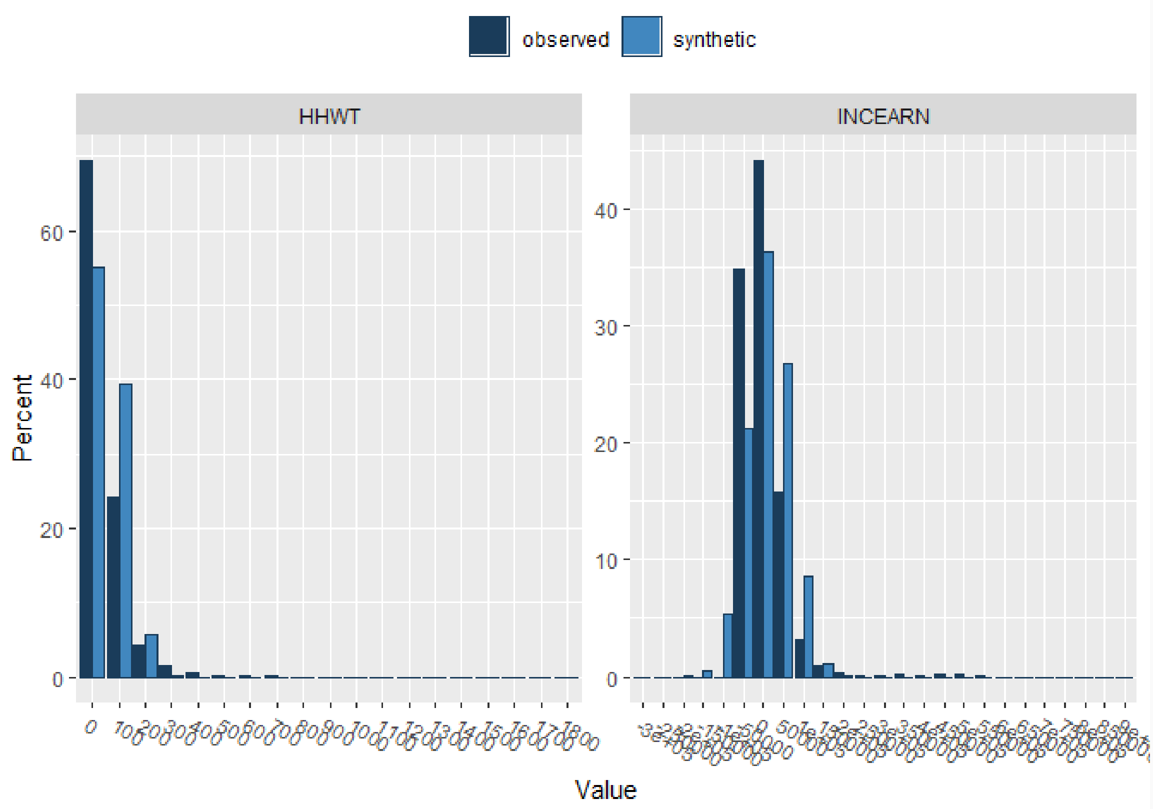

observed
synthetic
AVAILBLE
EMPSTATD
HINSEMP
HISPAN
Percent
Value


observed
synthetic
GQ
INCWELFR
ARRIVES
Percent
Value


observed
synthetic
LOOKING
EMPSTAT
HCOVPRIV
RACE
Percent
Value

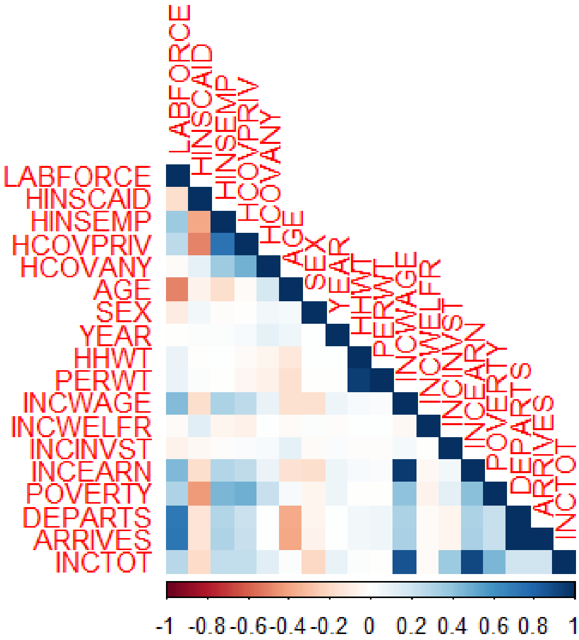

observed
synthetic
YEAR
INCINVST
INCTOT
Percent
Value

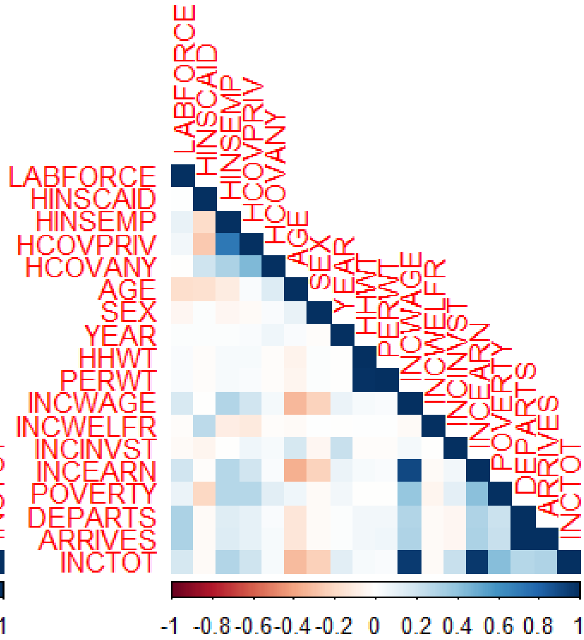




Correlation Plots for Graphical Comparison of Pearson Correlation

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

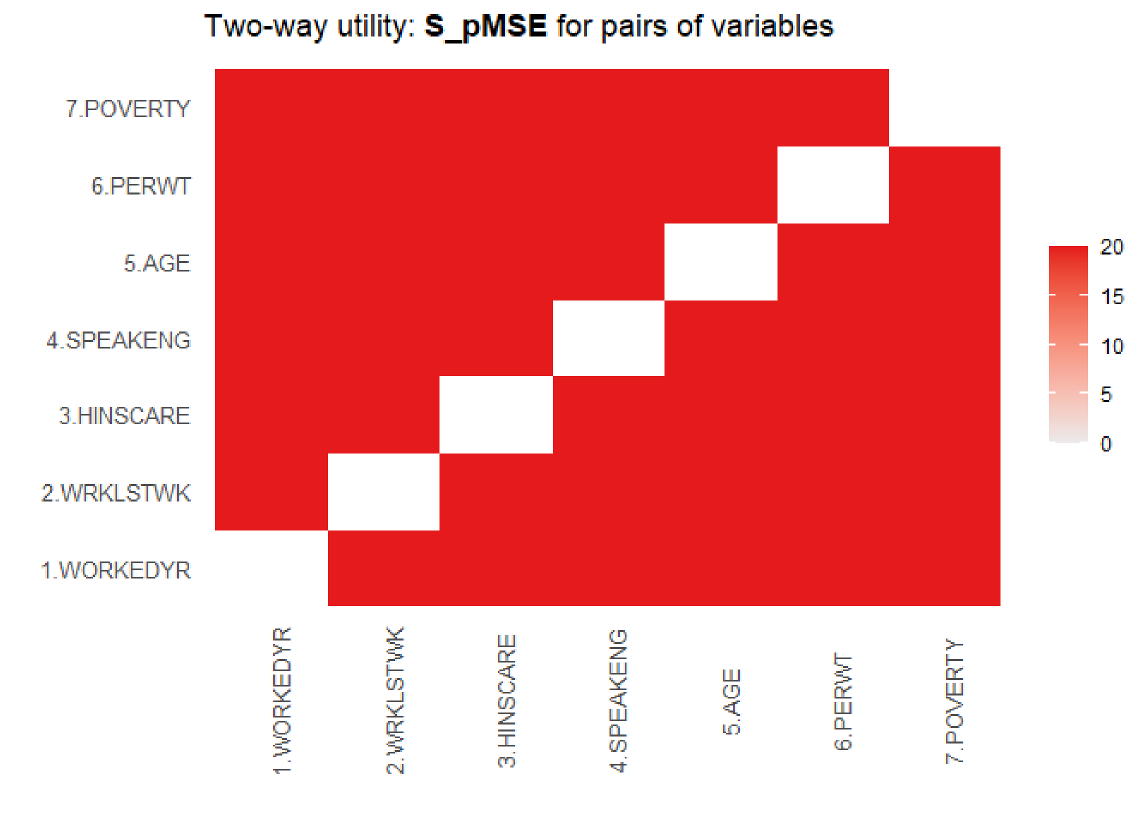
	pMSE	S_pMSE	df

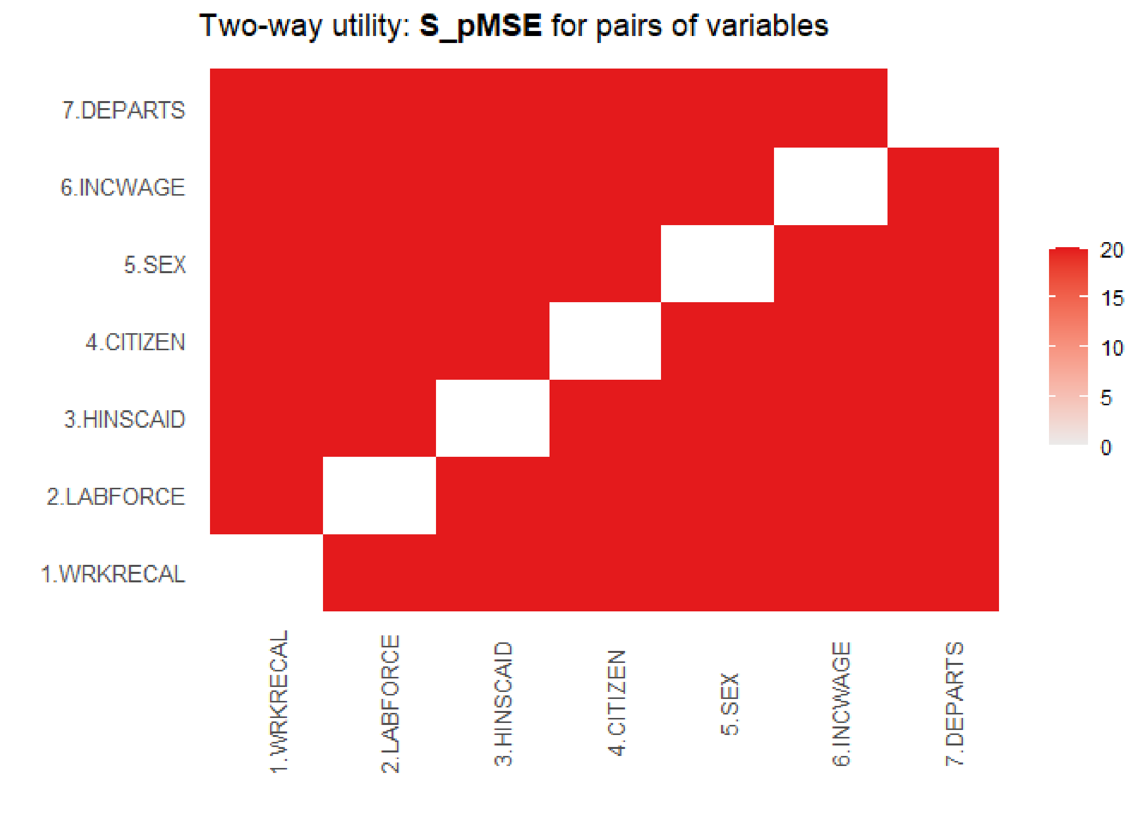
	pMSE	S_pMSE	df
ABSENT	0.0124855	103400.0692	2
EDUC	0.0004760	788.3904	10
HCOVANY	0.0001541	2552.1508	1
MARST	0.0002868	950.0797	5
HHWT	0.0104755	43376.9275	4
INCEARN	0.0729040	301881.2943	4

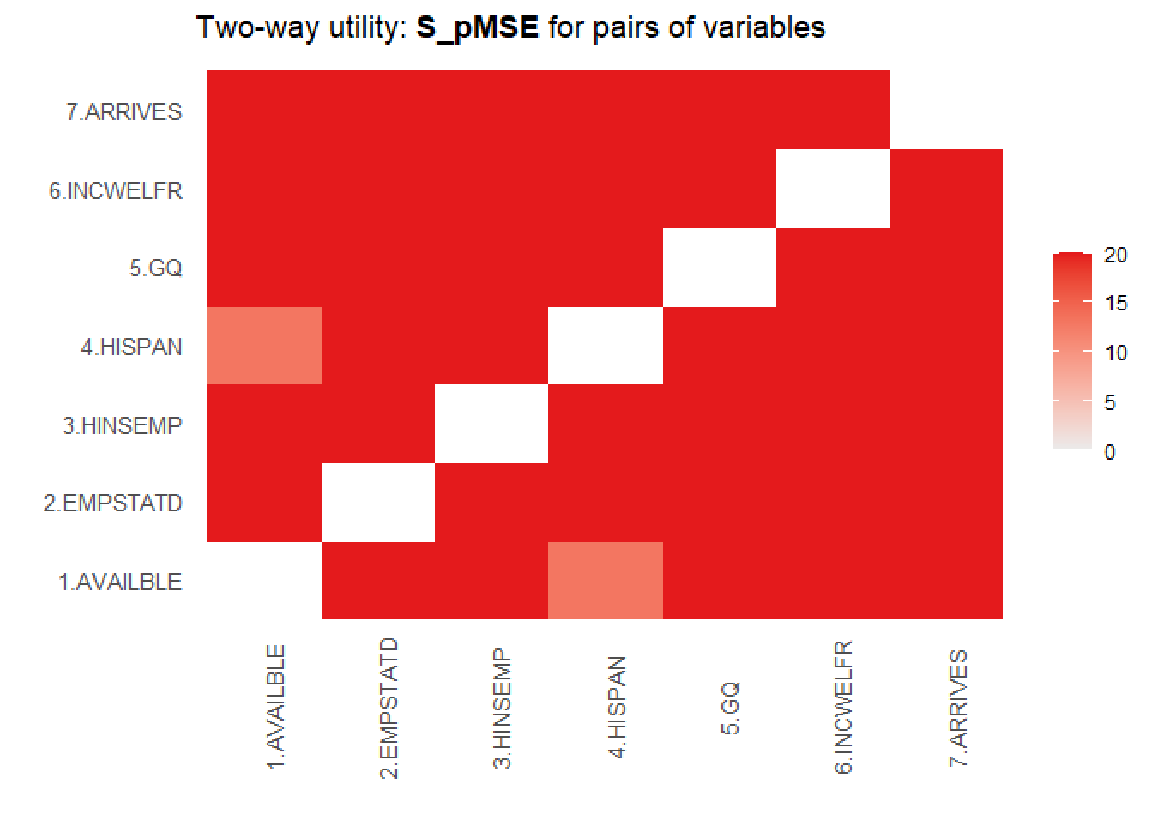
pMSE	S_pMSE
0.1159474	107.7212

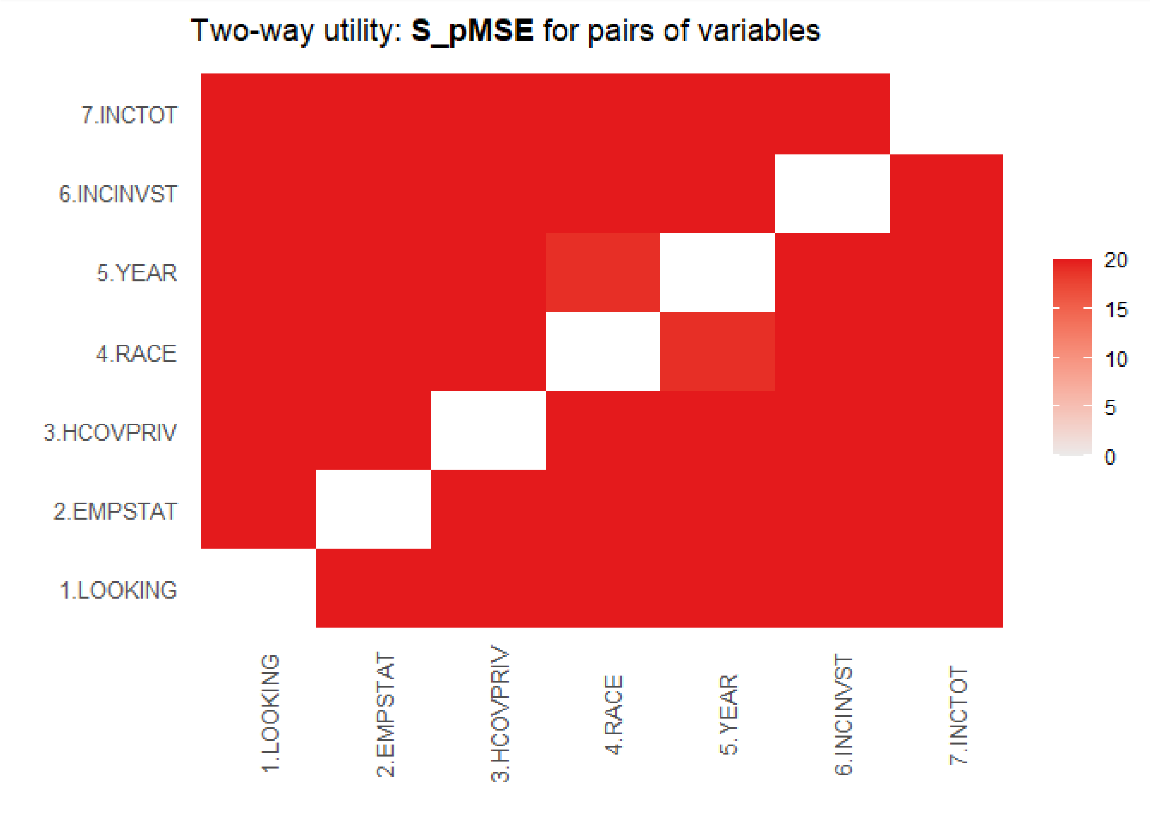
Two-way Tables Comparison of Synthesised Data (R-Package: synthpop) by (S_)pMSE

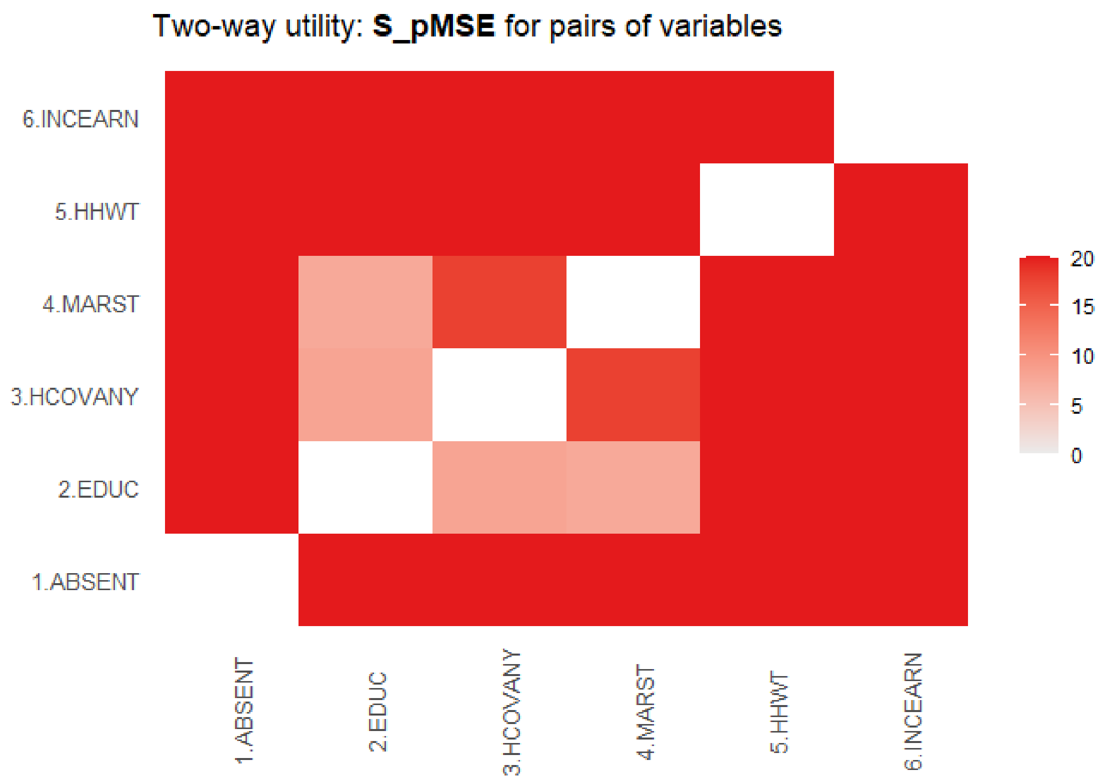
Two-way tables are evaluated based on the original and the synthetic dataset based on S_pMSE (see above). We also present the results for the mean absolute difference in densities (MabsDD) and the Bhattacharyya distance (dBhatt).

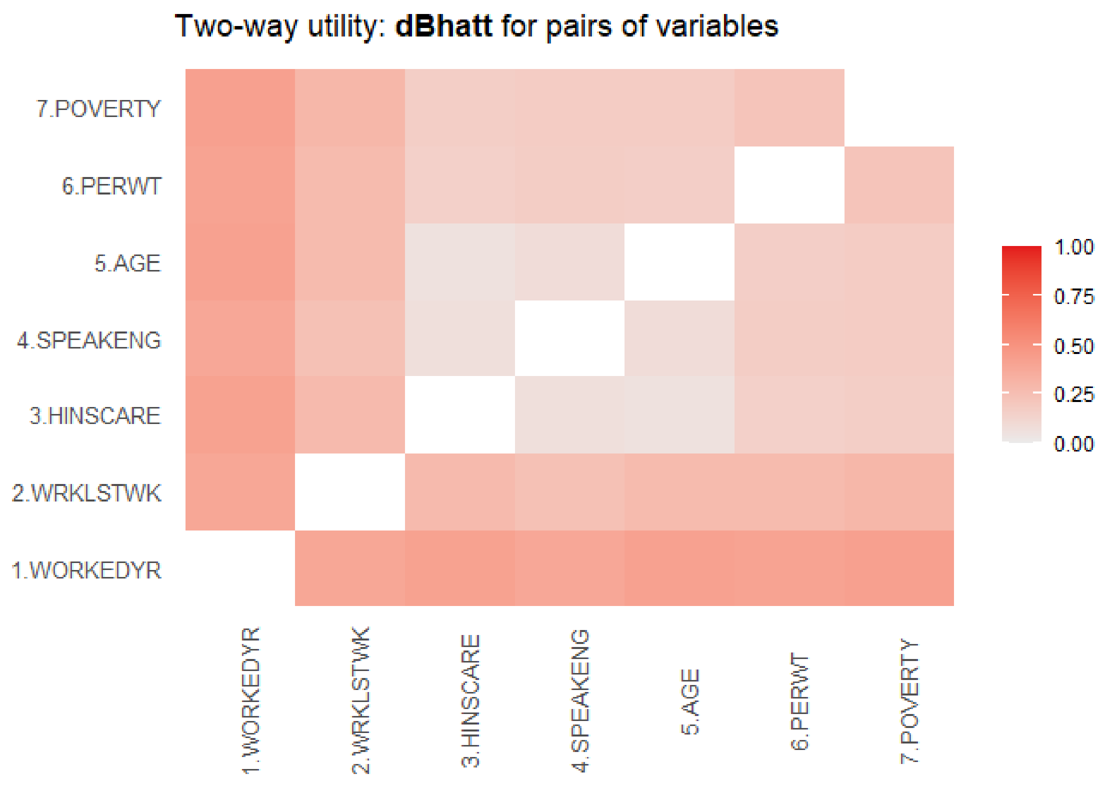

Two-way utility: S_pMSE for pairs of variables
7.POVERTY
6.PERWT
5.AGE
4.SPEAKENG
3.HINSCARE
2.WRKLSTWK
1.WORKEDYR
1.WORKEDYR
2.WRKLSTWK
3.HINSCARE
4.SPEAKENG
5.AGE
6.PERWT
7.POVERTY
20
15
10
5
0

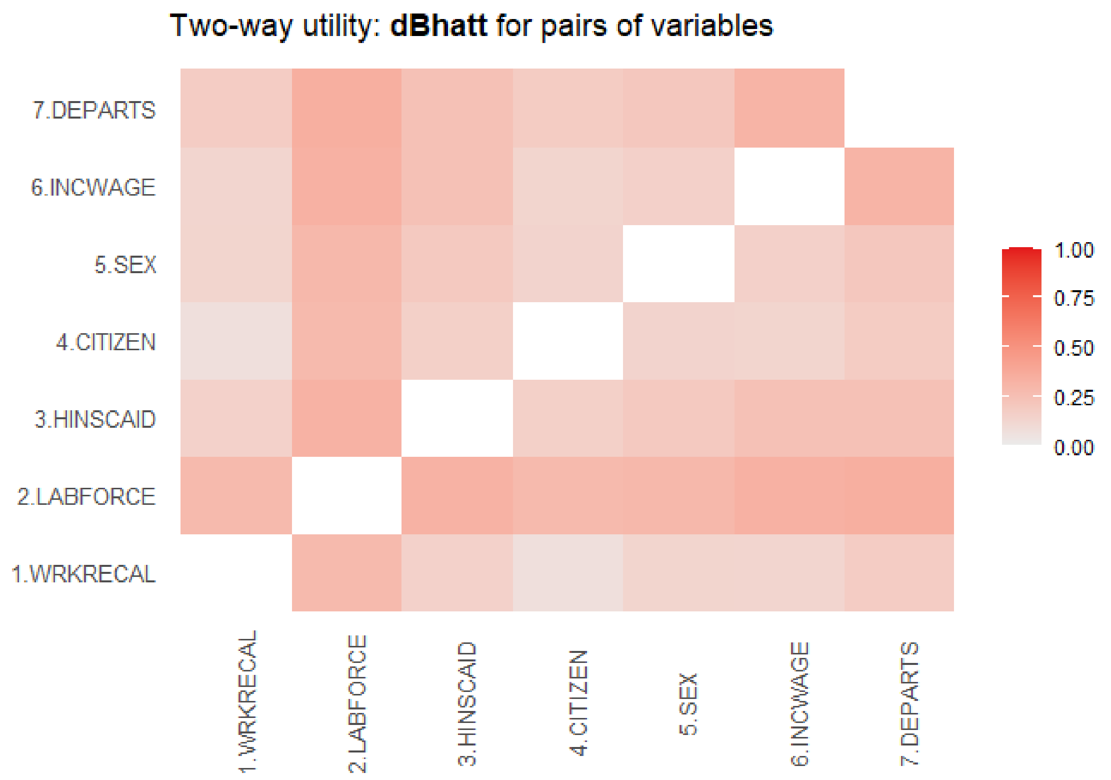

Two-way utility: S_pMSE for pairs of variables
7.DEPARTS
6.INCWAGE
5.SEX
4.CITIZEN
3.HINSCAID
2.LABFORCE
1.WRKRECAL
1.WRKRECAL
2.LABFORCE
3.HINSCAID
4.CITIZEN
5.SEX
6.INCWAGE
7.DEPARTS
20
15
10
5
0

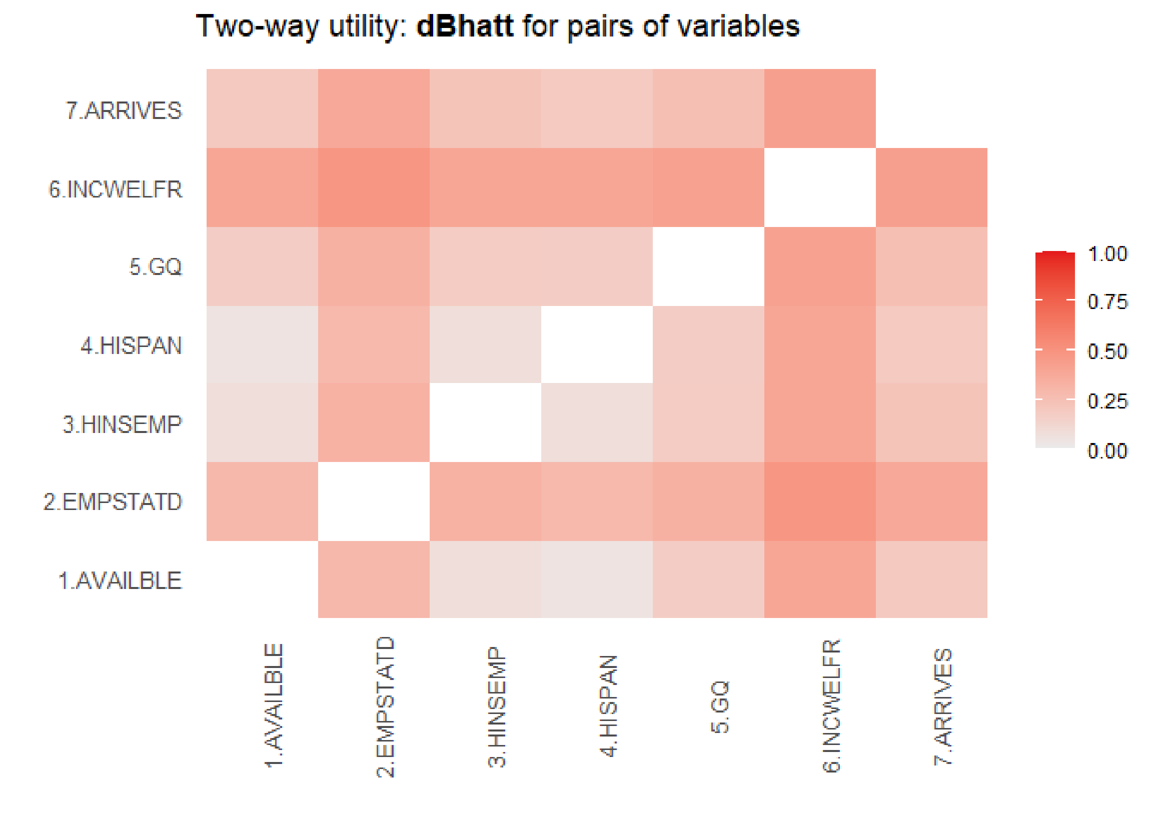

Two-way utility: S_pMSE for pairs of variables
7.ARRIVES
6.INCWELFR
5.GQ
4.HISPAN
3.HINSEMP
2.EMPSTATD
1.AVAILBLE
1.AVAILBLE
2.EMPSTATD
3.HINSEMP
4.HISPAN
5.GQ
6.INCWELFR
7.ARRIVES
20
15
10
5
0

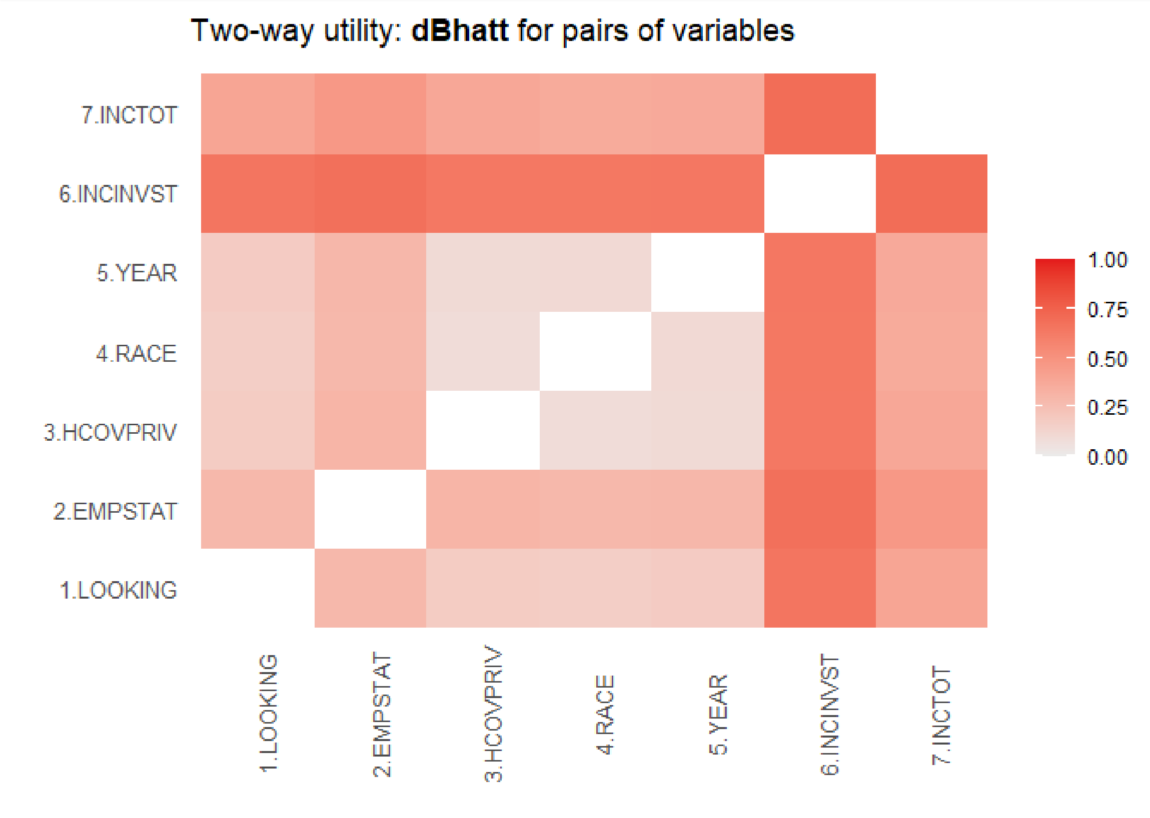

Two-way utility: S_pMSE for pairs of variables
7.INCTOT
6.INCINVST
5.YEAR
4.RACE
3.HCOVPRIV
2.EMPSTAT
1.LOOKING
1.LOOKING
2.EMPSTAT
3.HCOVPRIV
4.RACE
5.YEAR
6.INCINVST
7.INCTOT
20
15
10
5
0

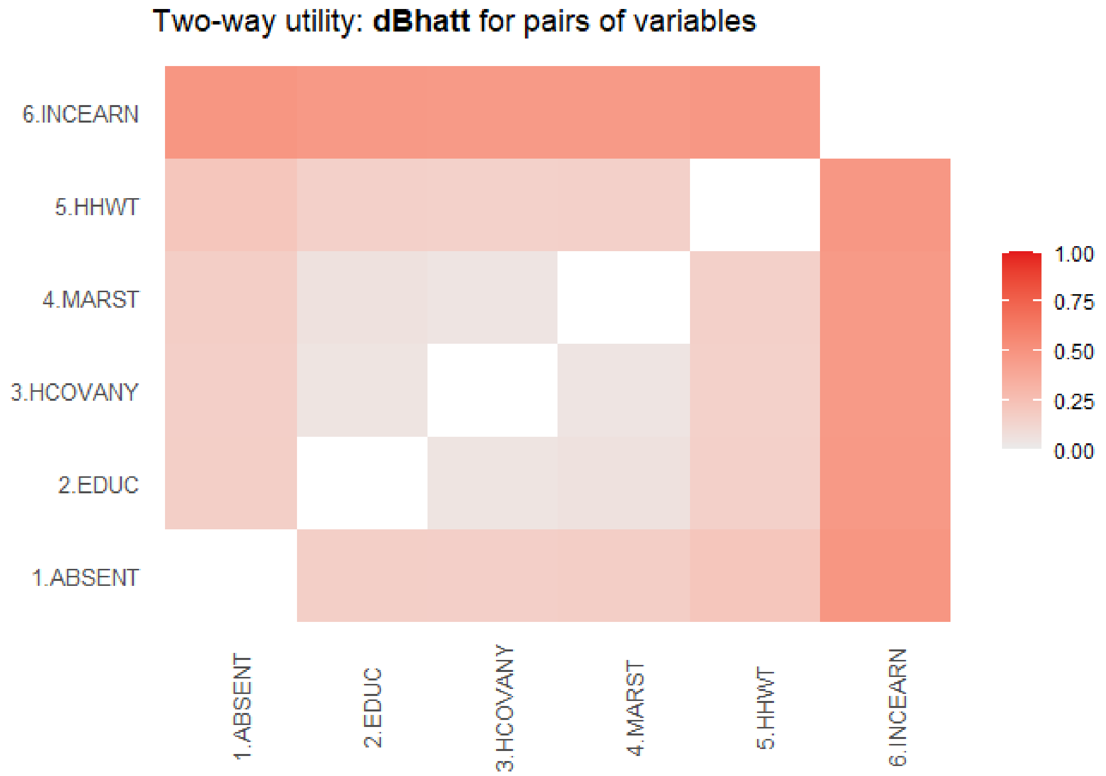

Two-way utility: S_pMSE for pairs of variables
6.INCEARN
5.HHWT
4.MARST
3.HCOVANY
2.EDUC
1.ABSENT
1.ABSENT
2.EDUC
3.HCOVANY
4.MARST
5.HHWT
6.INCEARN
20
15
10
5
0

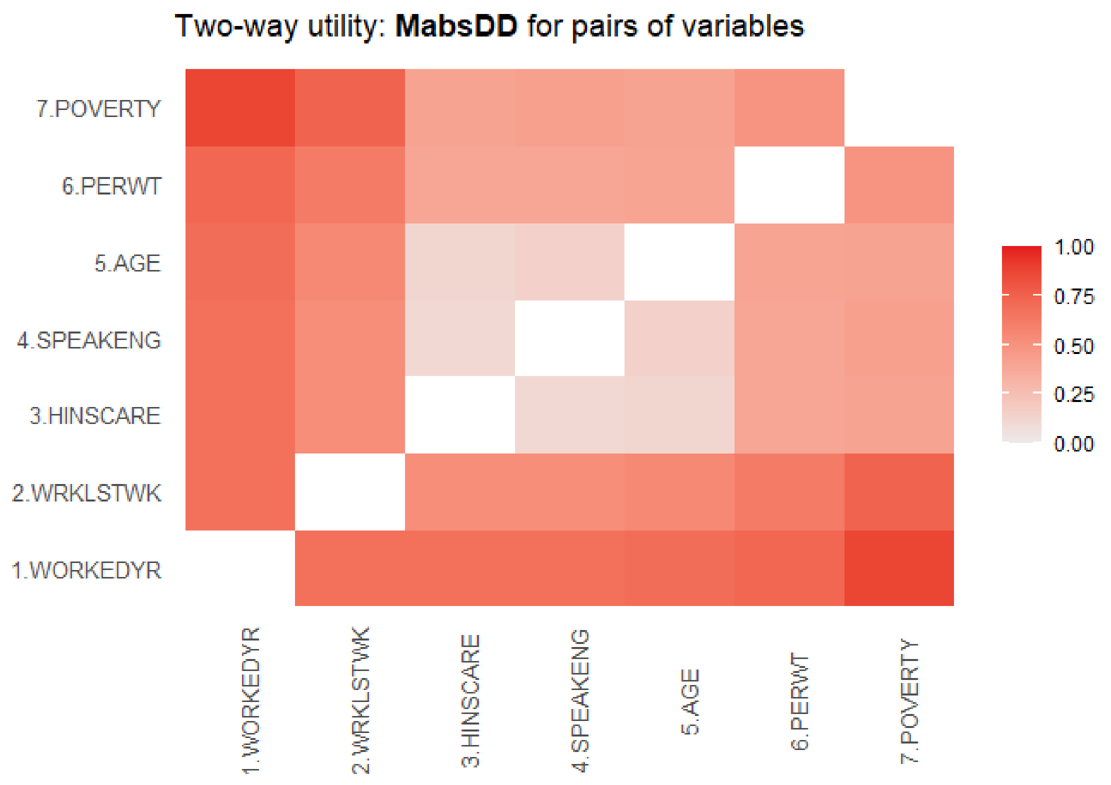

Two-way utility: **dBhatt** for pairs of variables
7.POVERTY
6.PERWT
5.AGE
4.SPEAKENG
3.HINSCARE
2.WRKLSTWK
1.WORKEDYR
1.WORKEDYR
2.WRKLSTWK
3.HINSCARE
4.SPEAKENG
5.AGE
6.PERWT
7.POVERTY
1.00
0.75
0.50
0.25
0.00


Two-way utility: **dBhatt** 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

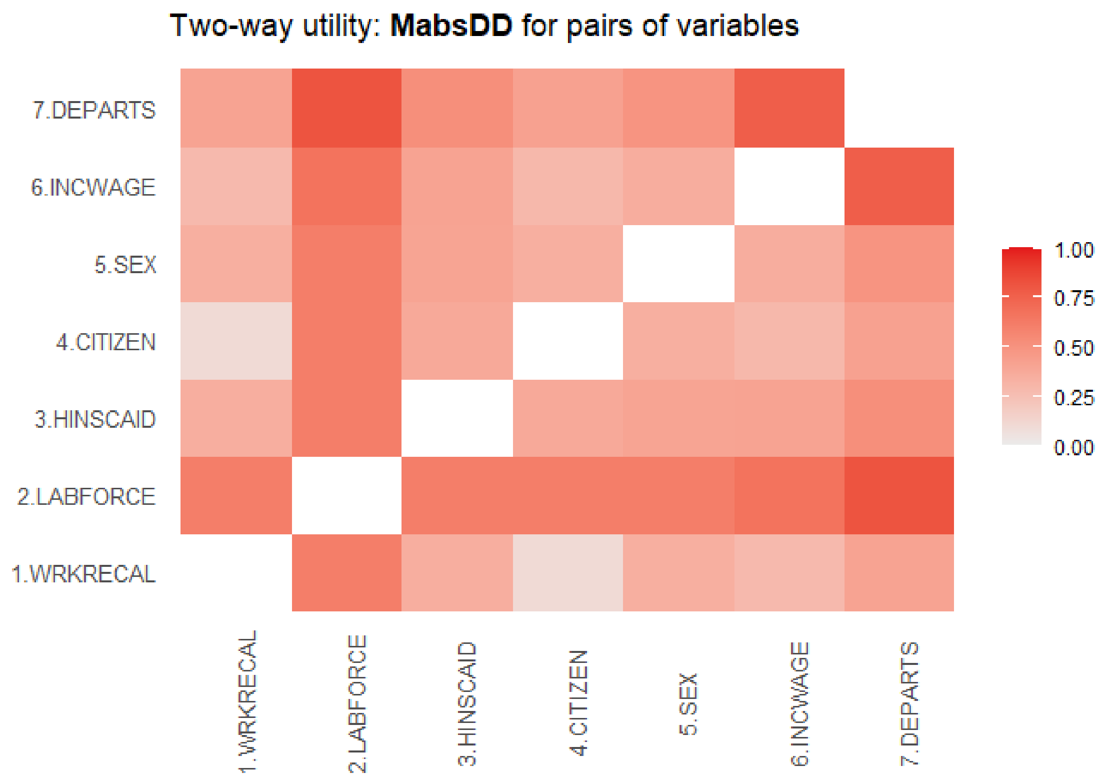

Two-way utility: dBhatt for pairs of variables
7.ARRIVES
6.INCWELFR
5.GQ
4.HISPAN
3.HINSEMP
2.EMPSTATD
1.AVAILBLE
1.00
0.75
0.50
0.25
0.00


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

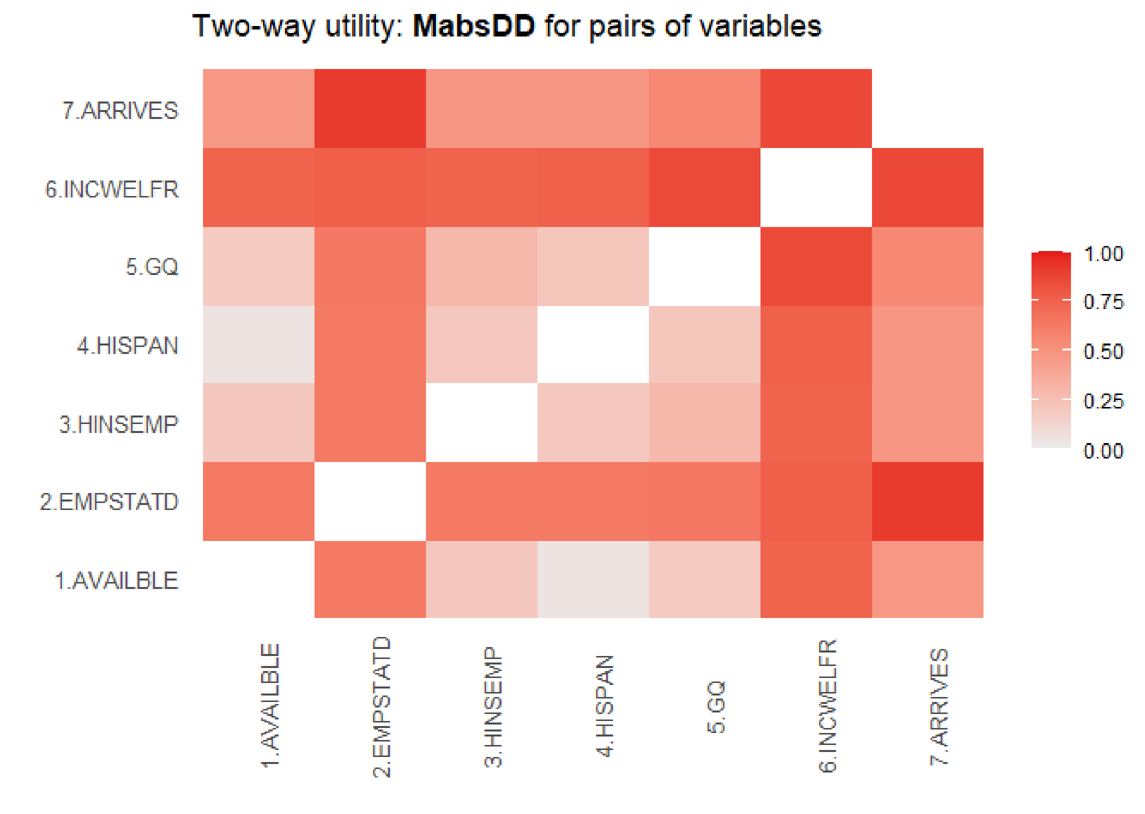

Two-way utility: dBhatt for pairs of variables
6.INCEARN
5.HHWT
4.MARST
3.HCOVANY
2.EDUC
1.ABSENT
1.ABSENT
2.EDUC
3.HCOVANY
4.MARST
5.HHWT
6.INCEARN
1.00
0.75
0.50
0.25
0.00

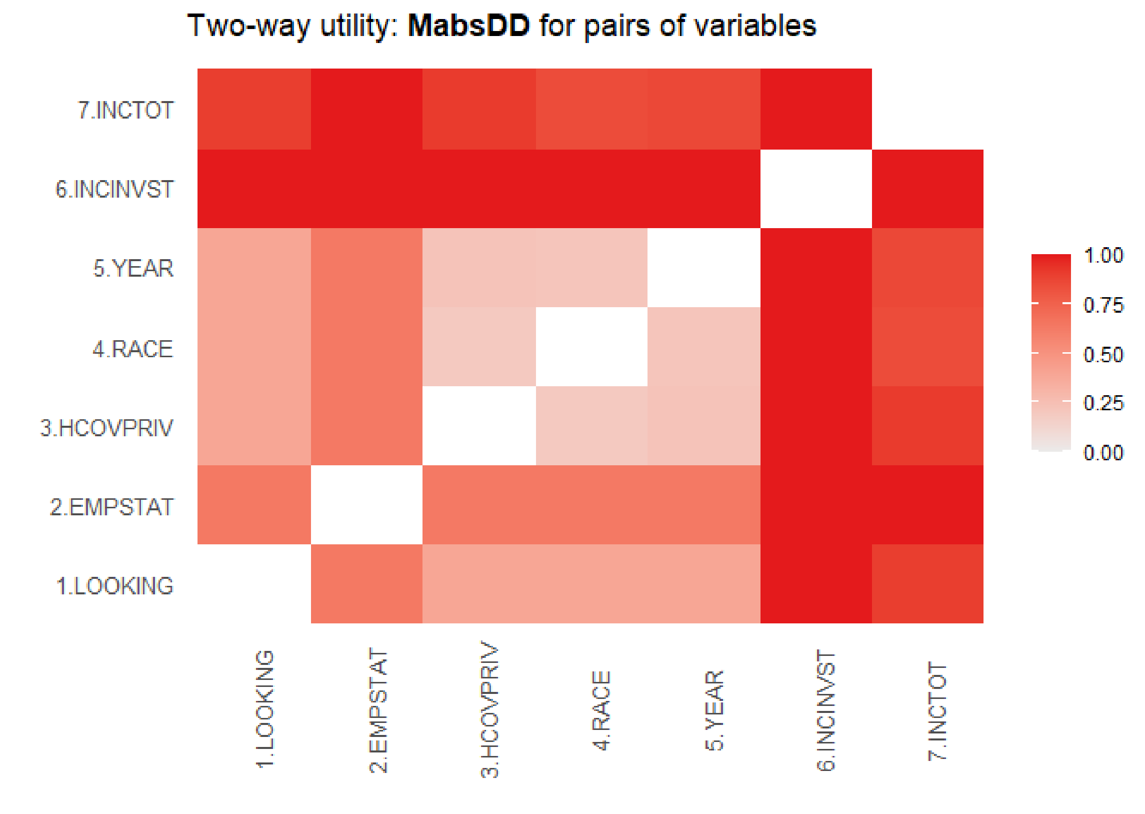


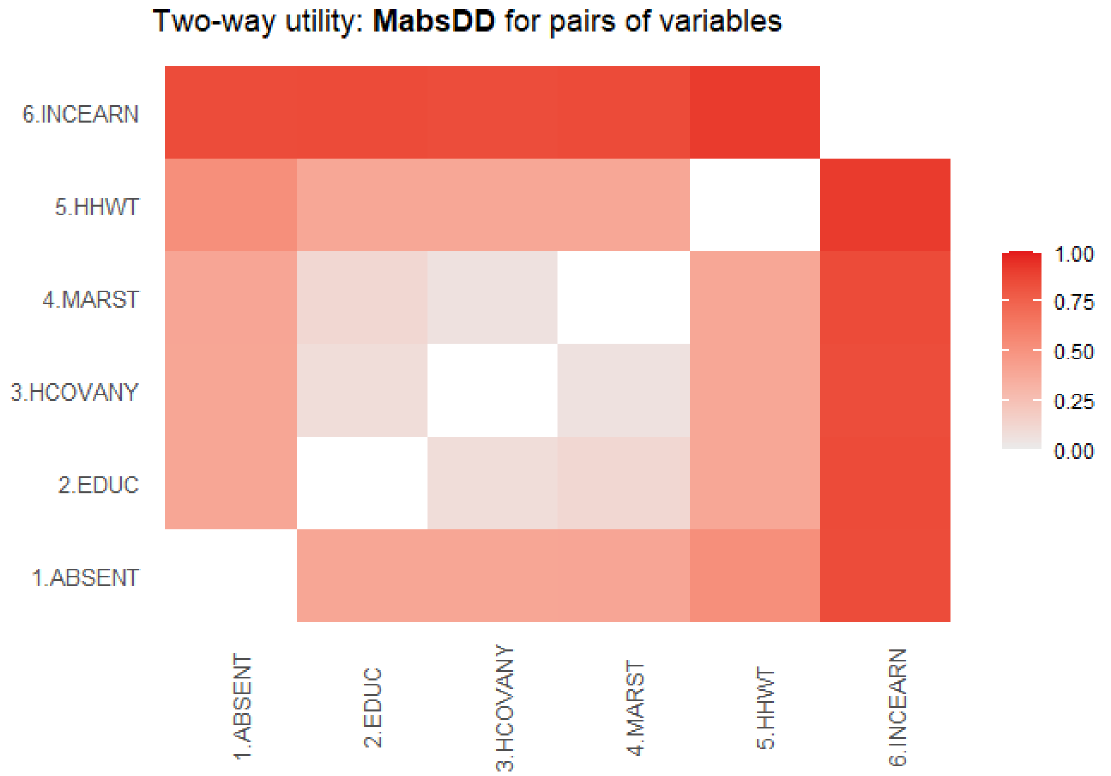
Two-way utility: MabsDD for pairs of variables
7.POVERTY
6.PERWT
5.AGE
4.SPEAKENG
3.HINSCARE
2.WRKLSTWK
1.WORKEDYR
1.WORKEDYR
2.WRKLSTWK
3.HINSCARE
4.SPEAKENG
5.AGE
6.PERWT
7.POVERTY
1.00
0.75
0.50
0.25
0.00



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.