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:

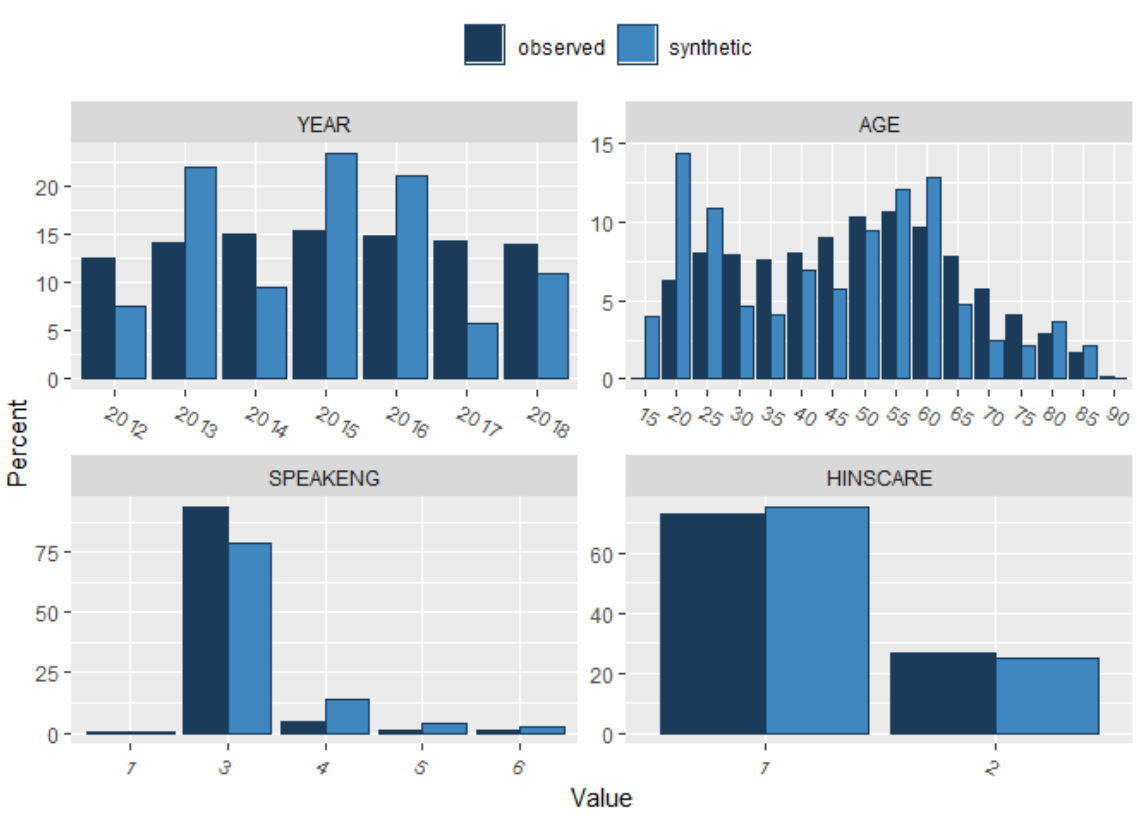
Metric	Number.Uniques	Number.Replications	Percentage.Replications
Perceived Risk	1035201	0	0

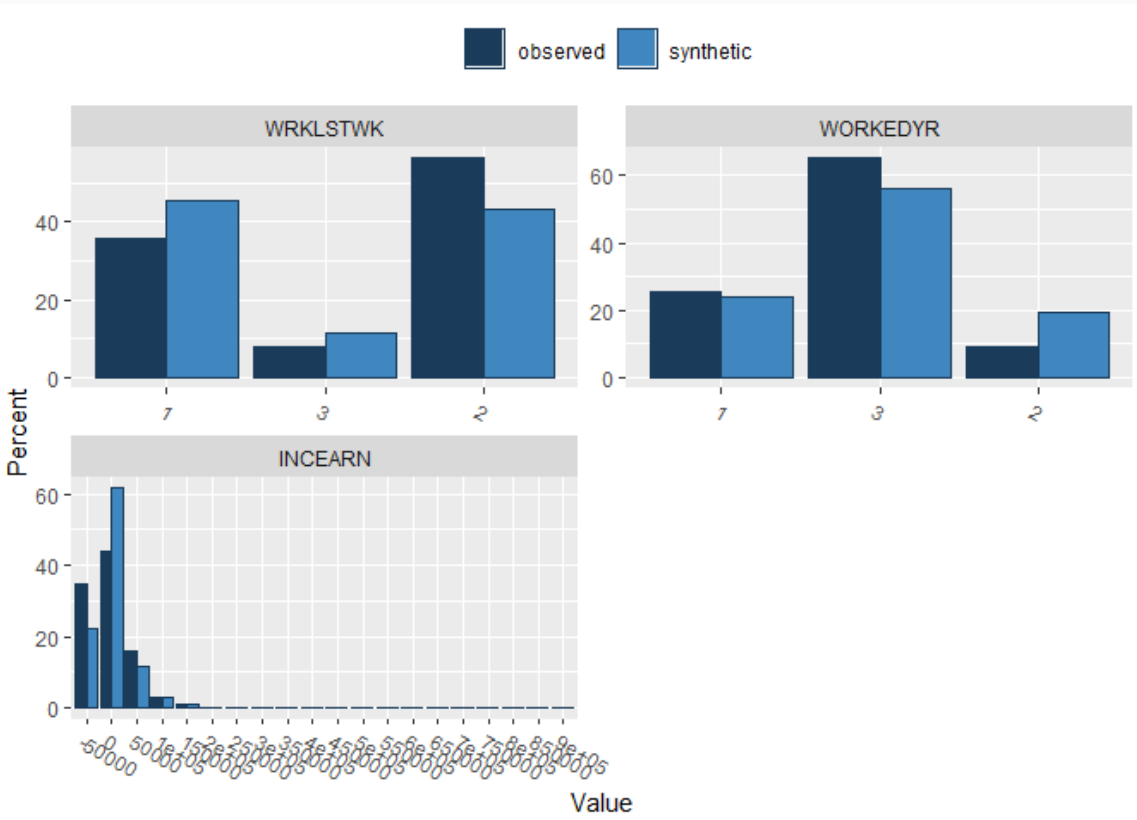
Utility Evaluation

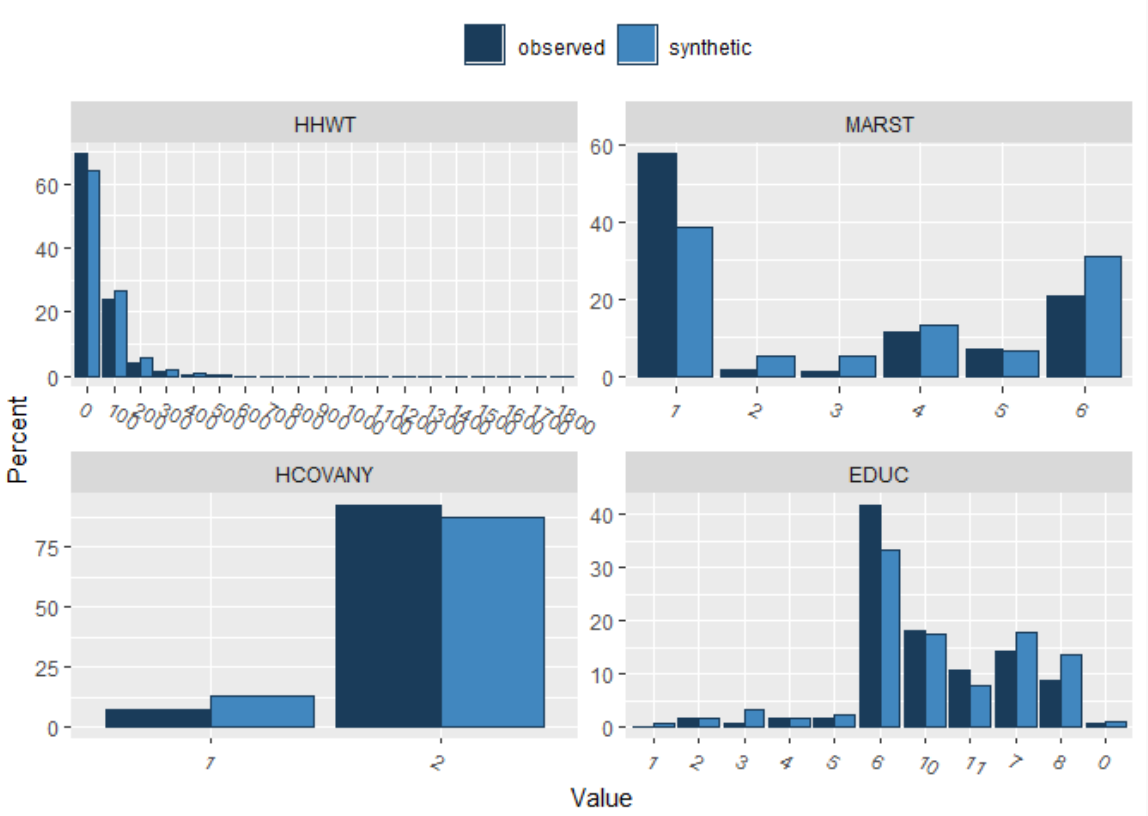
Different utility measures are applied in this section. These utility measures are the basis of utility evaluation for the generated synthetic dataset. The R packages synthpop, sdcMicro and corrplot were used to compute the following metrics. We do not use tests incorporating significance here. Confidence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

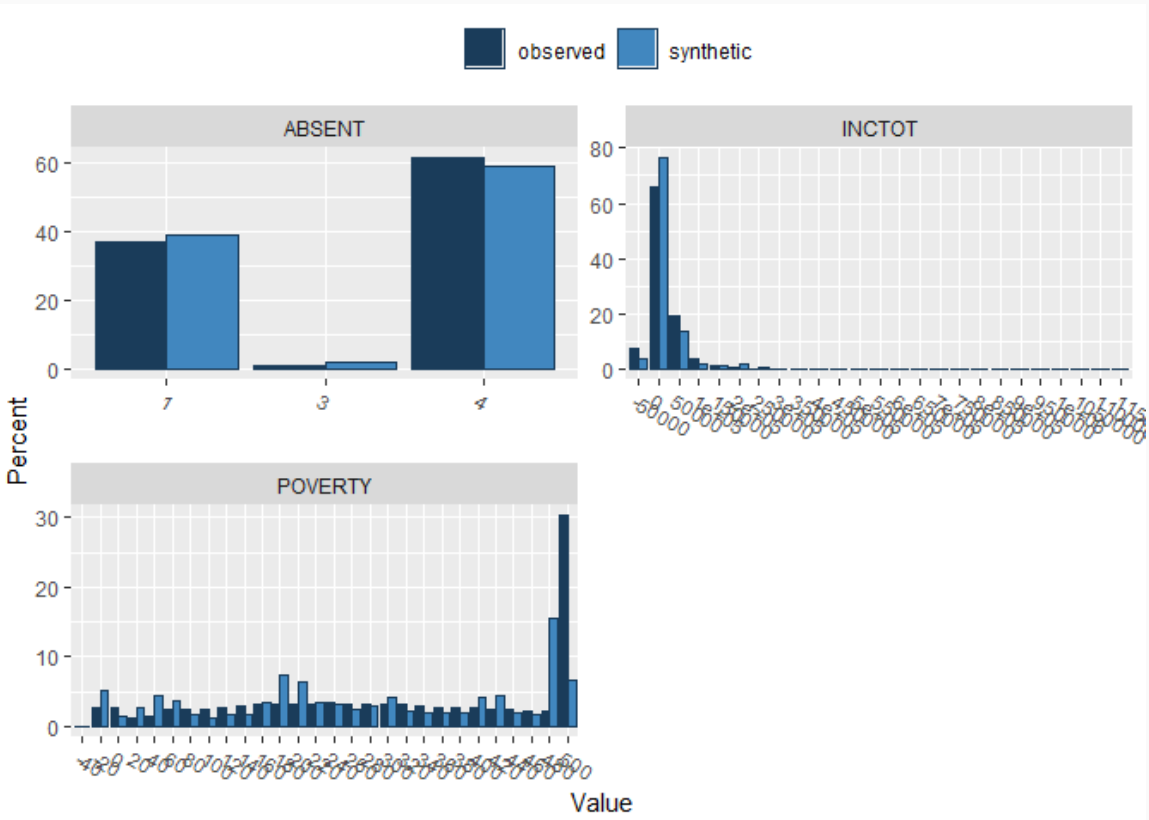
Graphical Comparison for Margins (R-Package: synthpop)

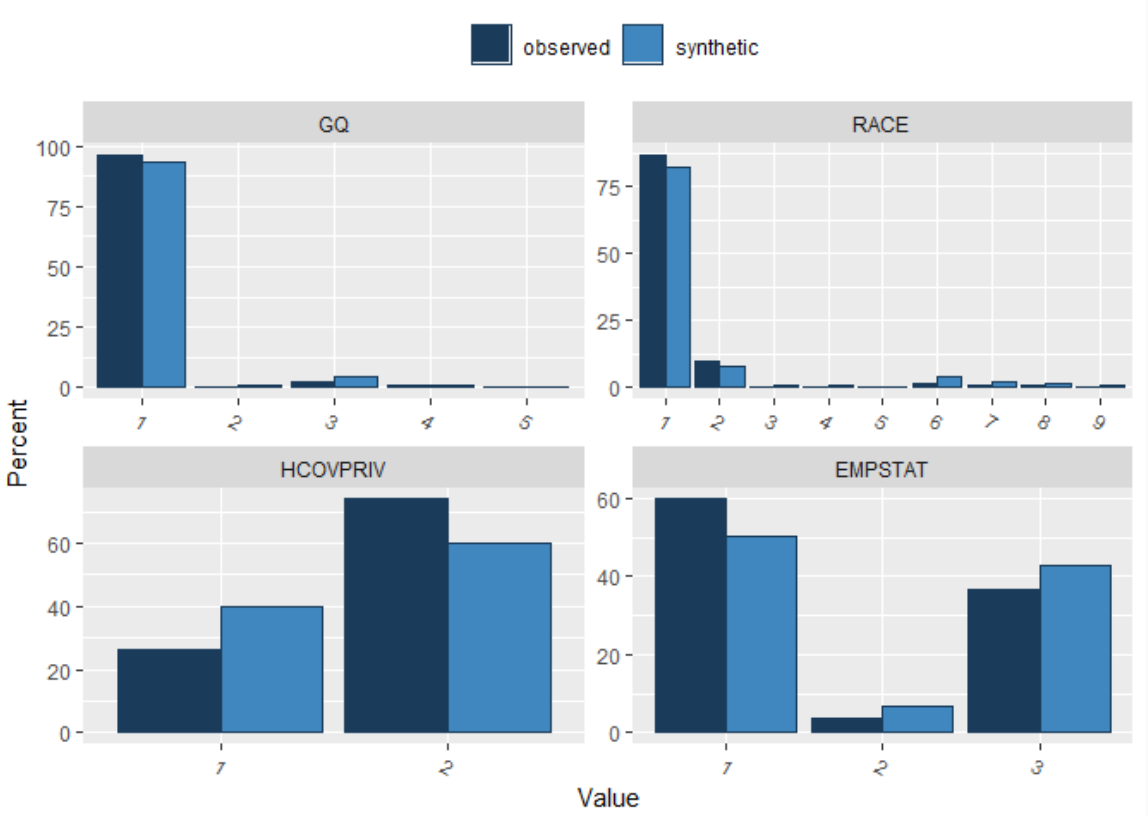
The following histograms provide an ad-hoc overview on the marginal distributions of the original and synthetic dataset. Matching or close distributions are related to a high data utility.

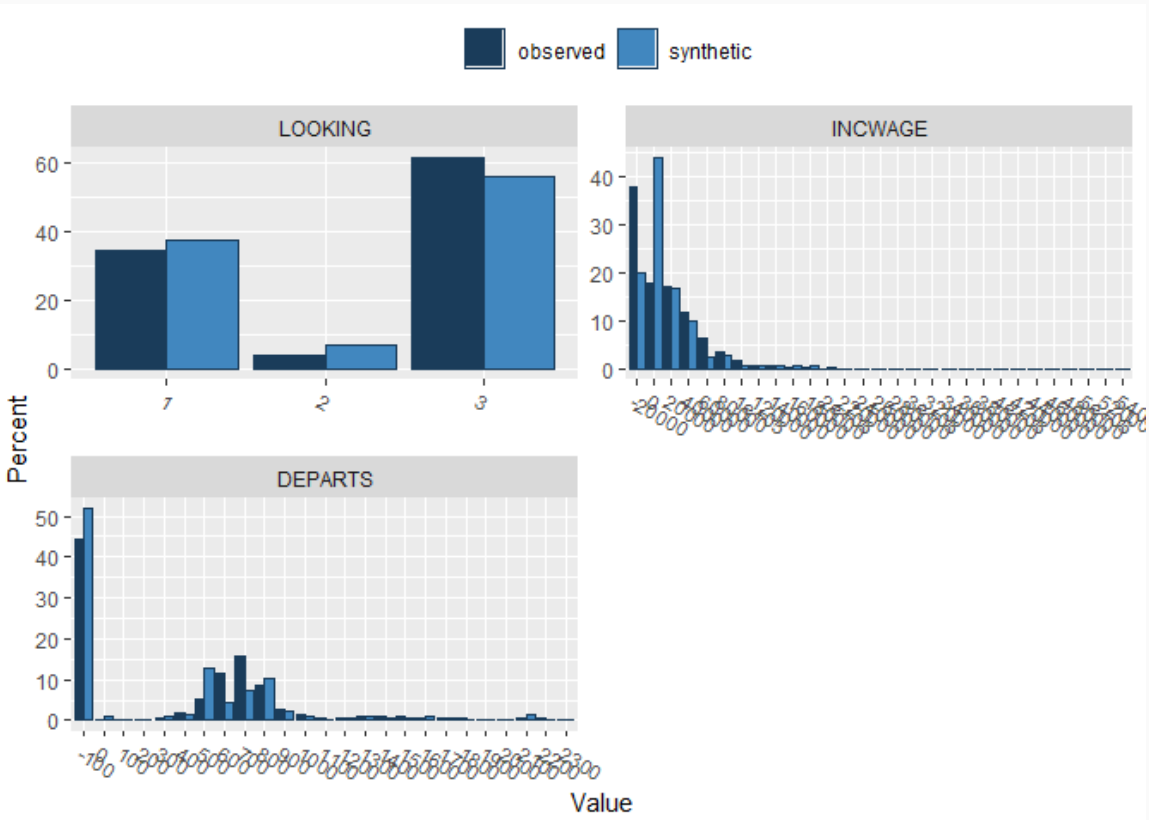

observed
synthetic
YEAR
AGE
SPEAKENG
HINSCARE
Percent
Value

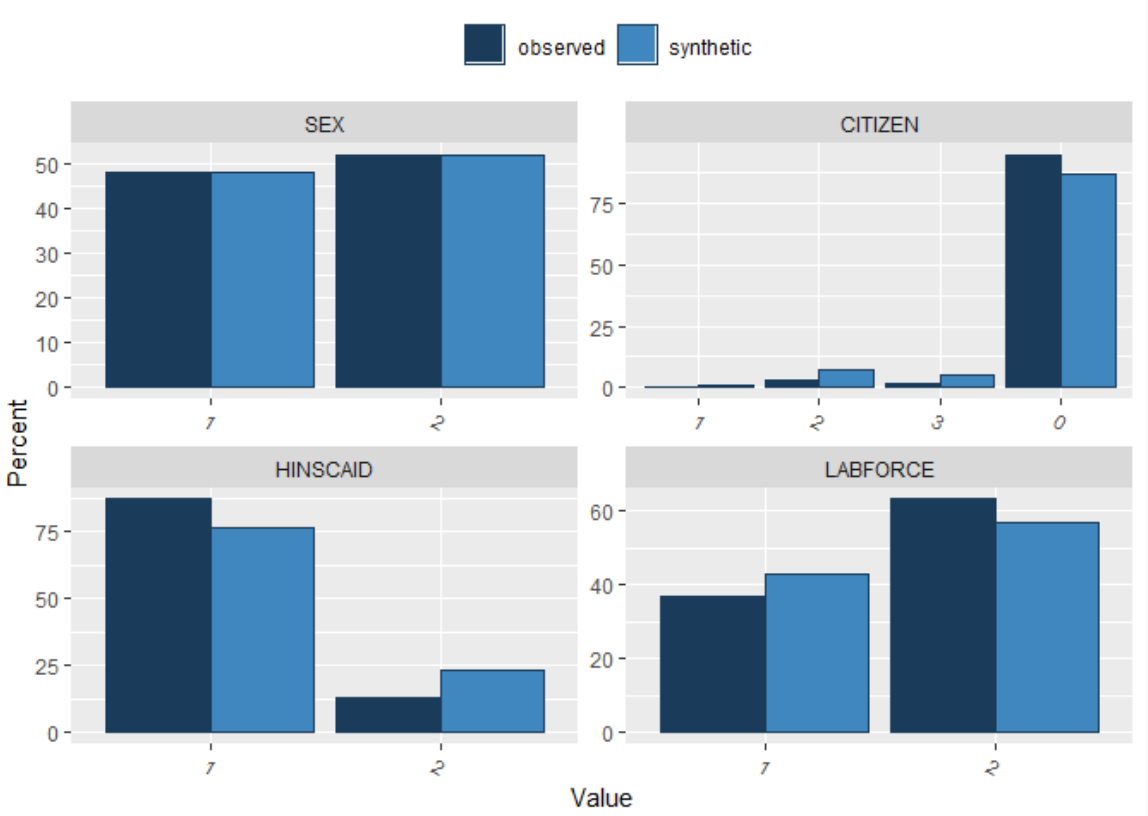

observed
synthetic
WRKLSTWK
WORKEDYR
INCEARN
Percent
Value

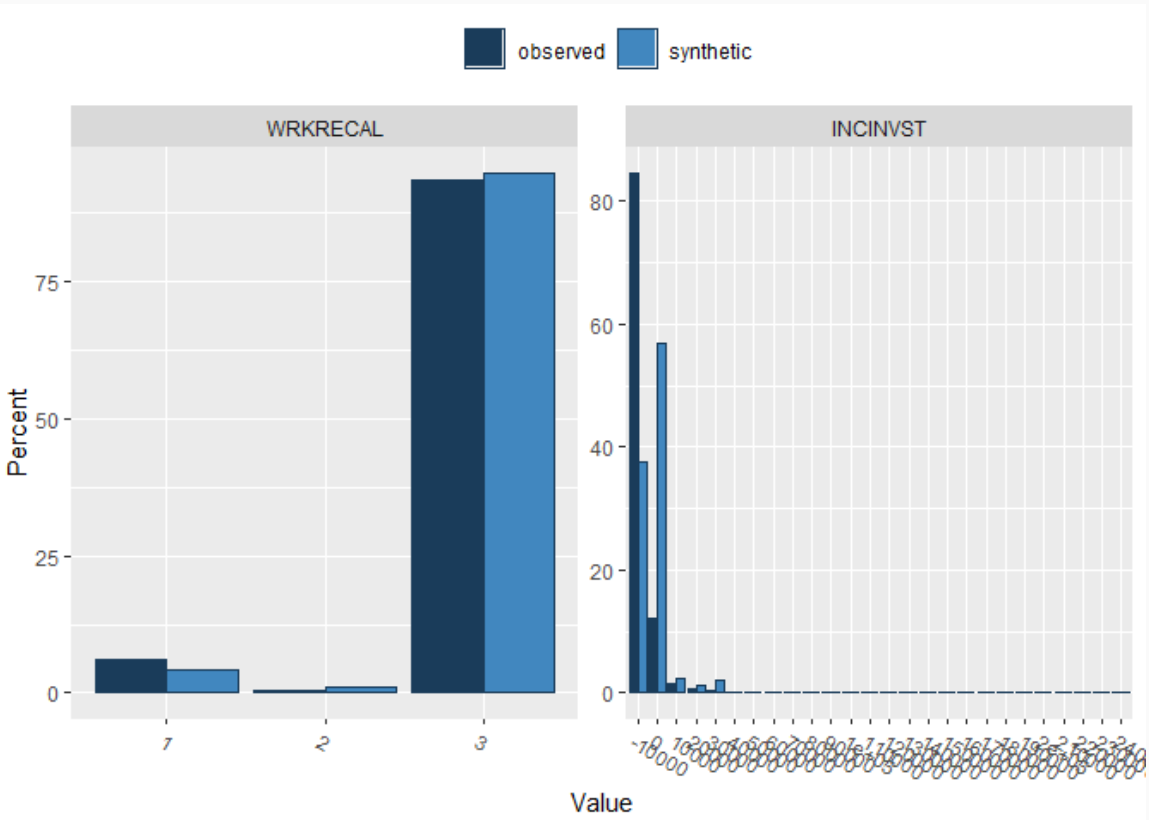

observed
synthetic
HHWT
MARST
HCOVANY
EDUC
Percent
Value


observed
synthetic
ABSENT
INCTOT
POVERTY
Percent
Value


observed
synthetic
GQ
RACE
HCOVPRIV
EMPSTAT
Percent
Value

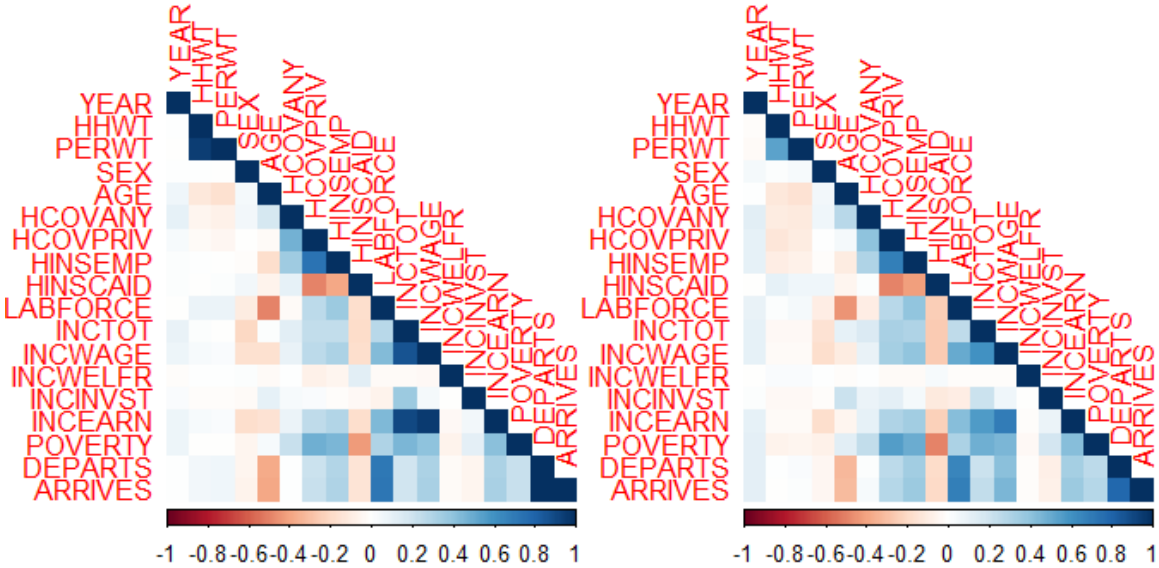

observed
synthetic
LOOKING
INCWAGE
DEPARTS
Percent
Value





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

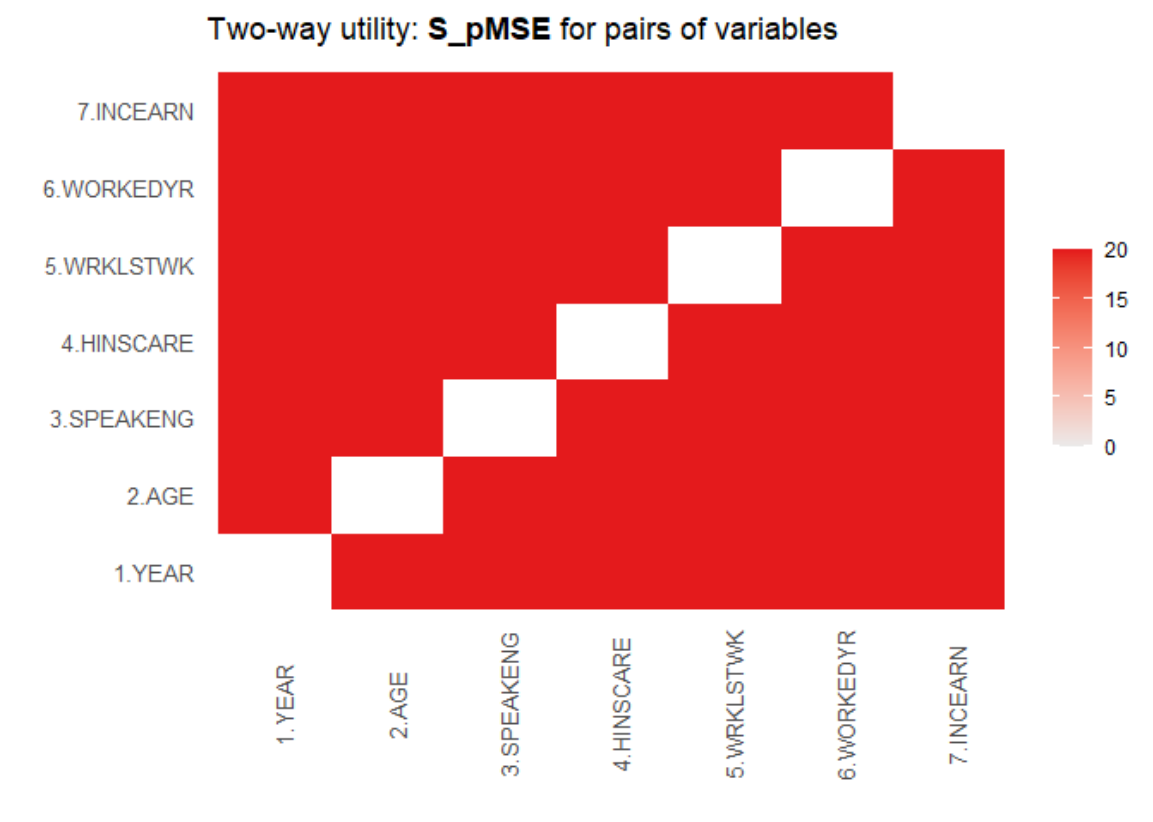
pMSE	S_pMSE
0.1917524	321.7915

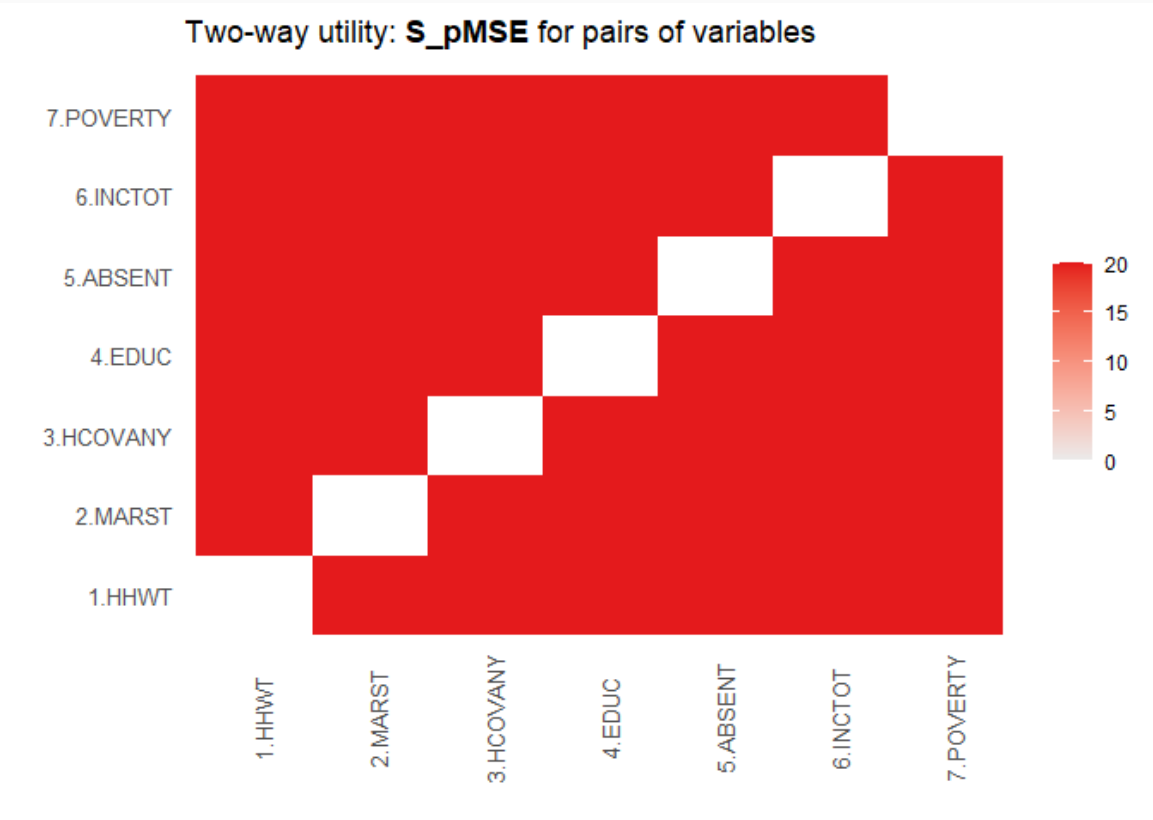
	pMSE	S_pMSE	df
SEX	0.0000001	9.521136e-01	1
CITIZEN	0.0046015	2.540531e+04	3
HINSCAID	0.0044901	7.437076e+04	1
LABFORCE	0.0009779	1.619639e+04	1
WRKRECAL	0.0007581	6.278456e+03	2
INCINVST	0.0553913	4.587289e+05	2

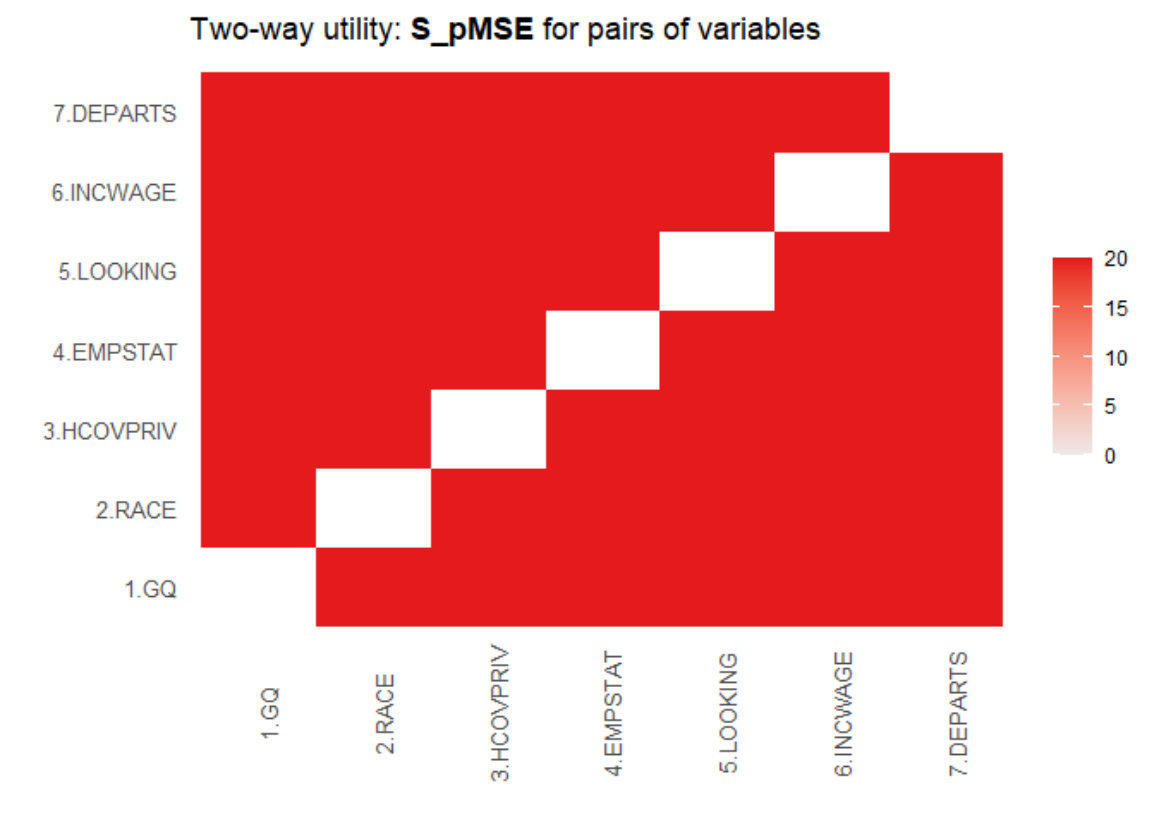
pMSE	S_pMSE
0.1884497	490.0153

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

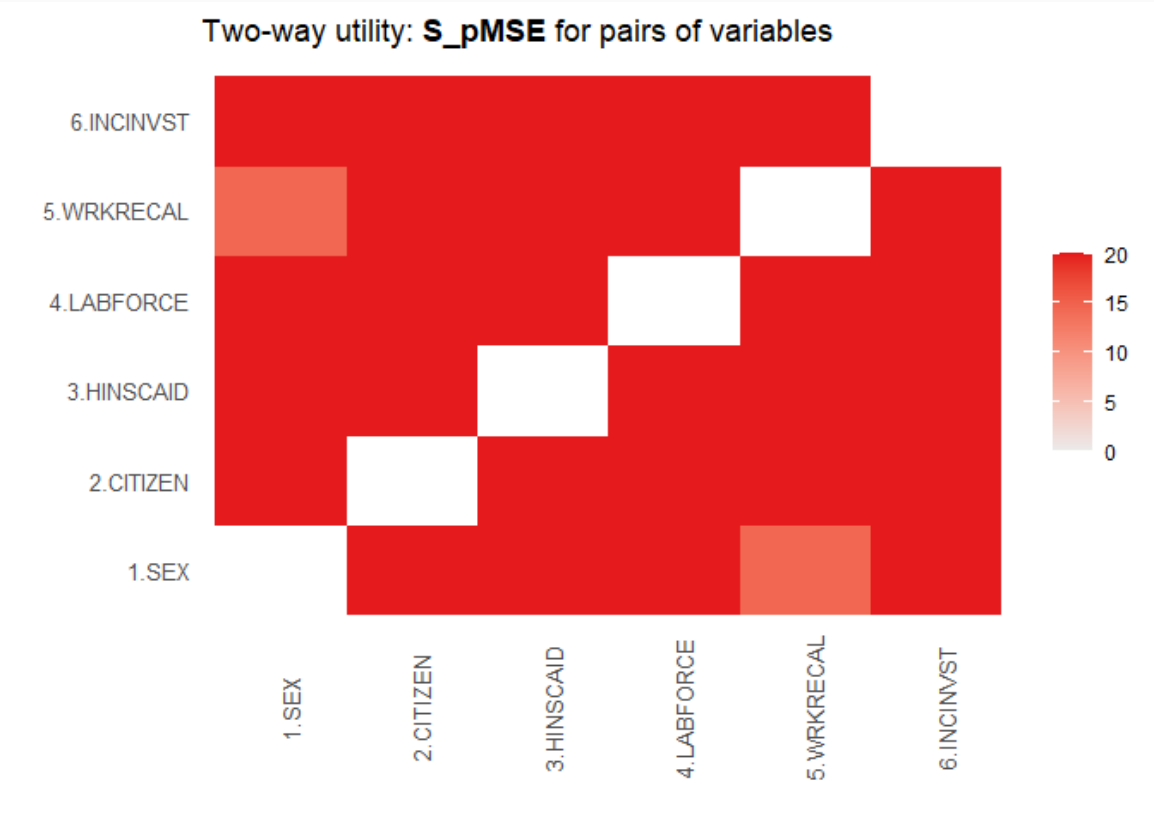
Two-way tables are evaluated based on the original and the synthetic dataset based on S_pMSE (see above). We also present the results for the mean absolute difference in densities (MabsDD) and the Bhattacharyya distance (dBhatt).

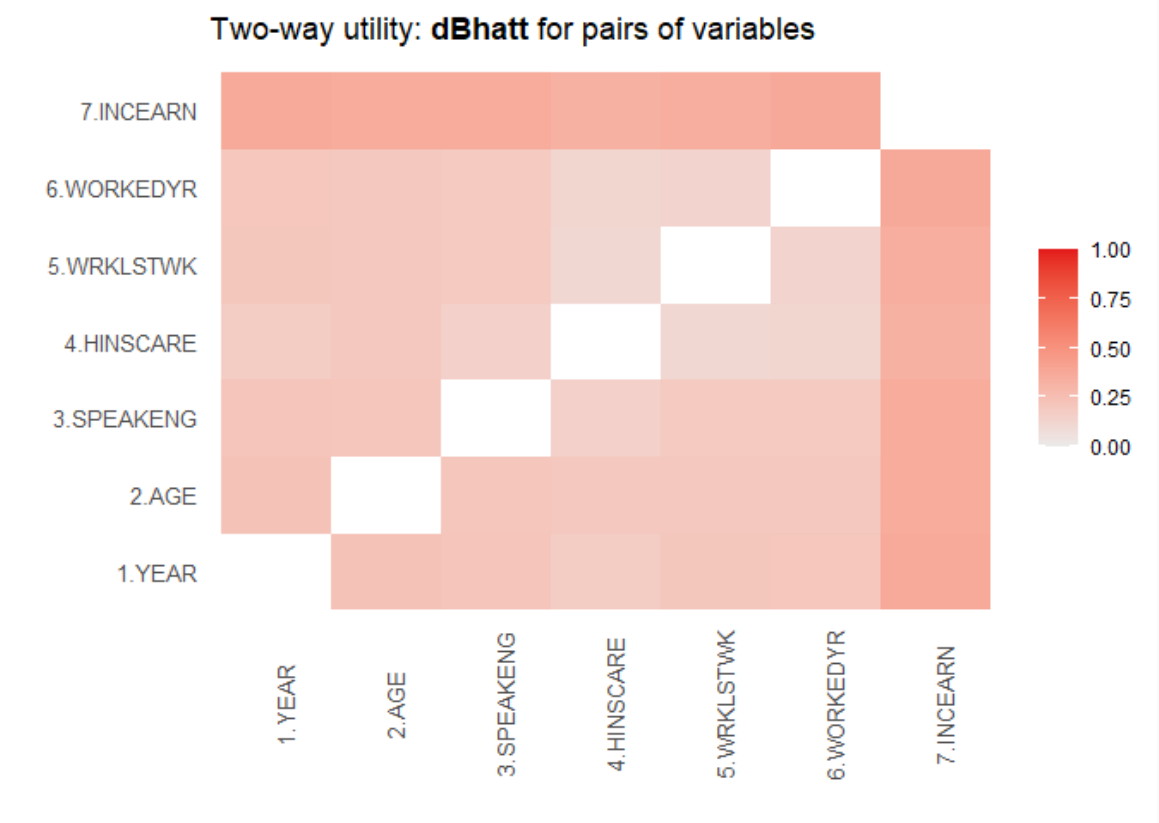


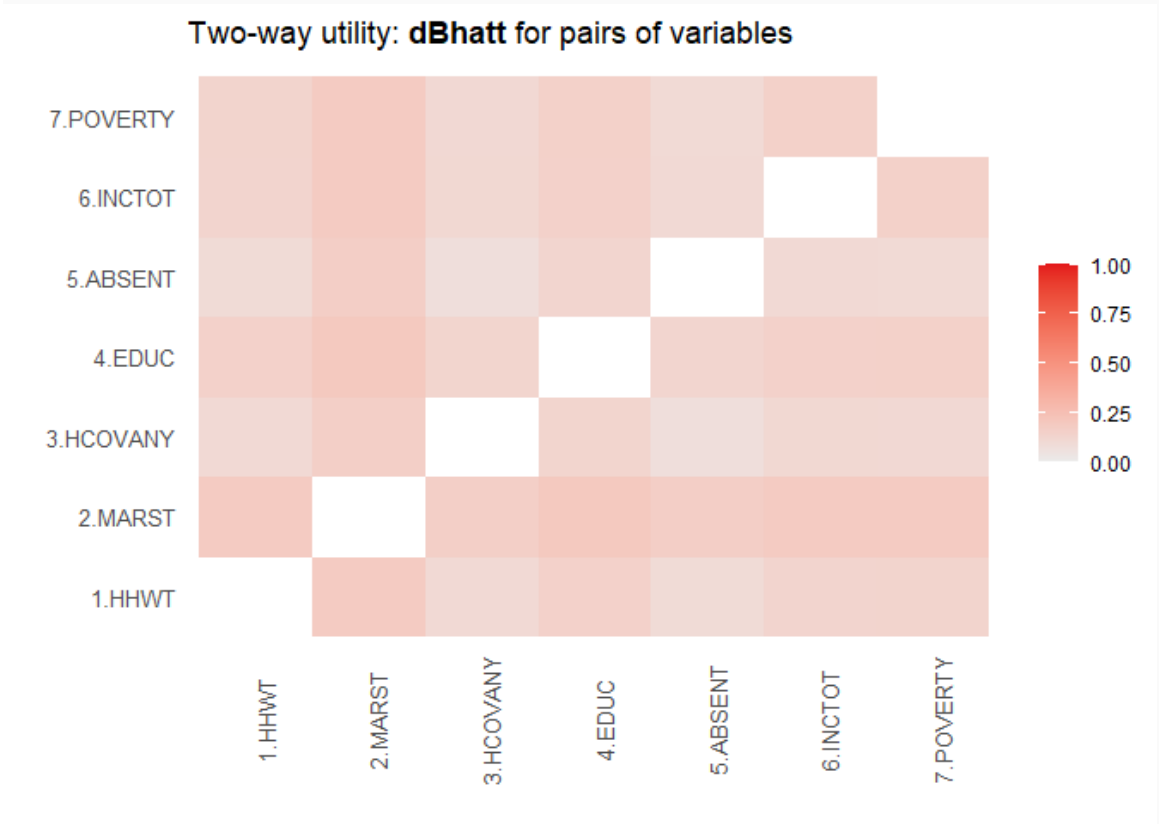


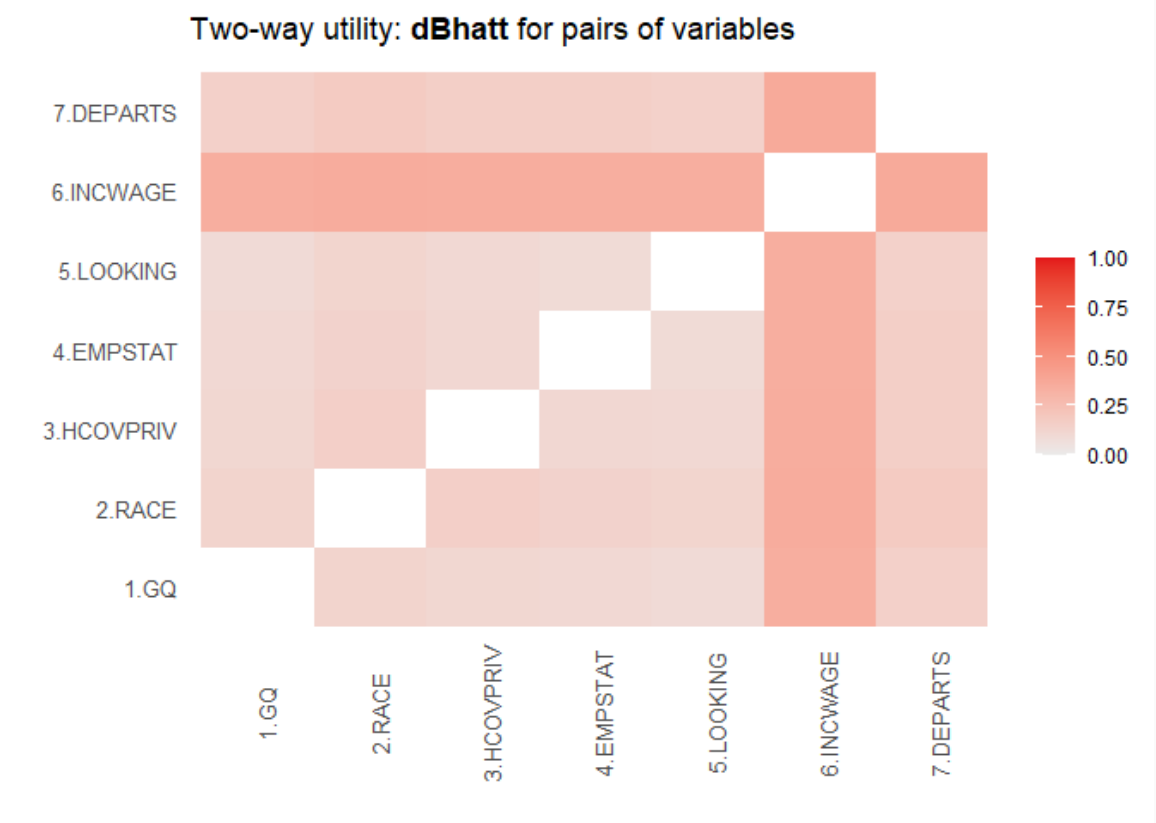

Two-way utility: S_pMSE for pairs of variables
7.DEPARTS
6.INCWAGE
5.LOOKING
4.EMPSTAT
3.HCOVPRIV
2.RACE
1.GQ
1.GQ
2.RACE
3.HCOVPRIV
4.EMPSTAT
5.LOOKING
6.INCWAGE
7.DEPARTS
20
15
10
5
0

```
## NULL
```

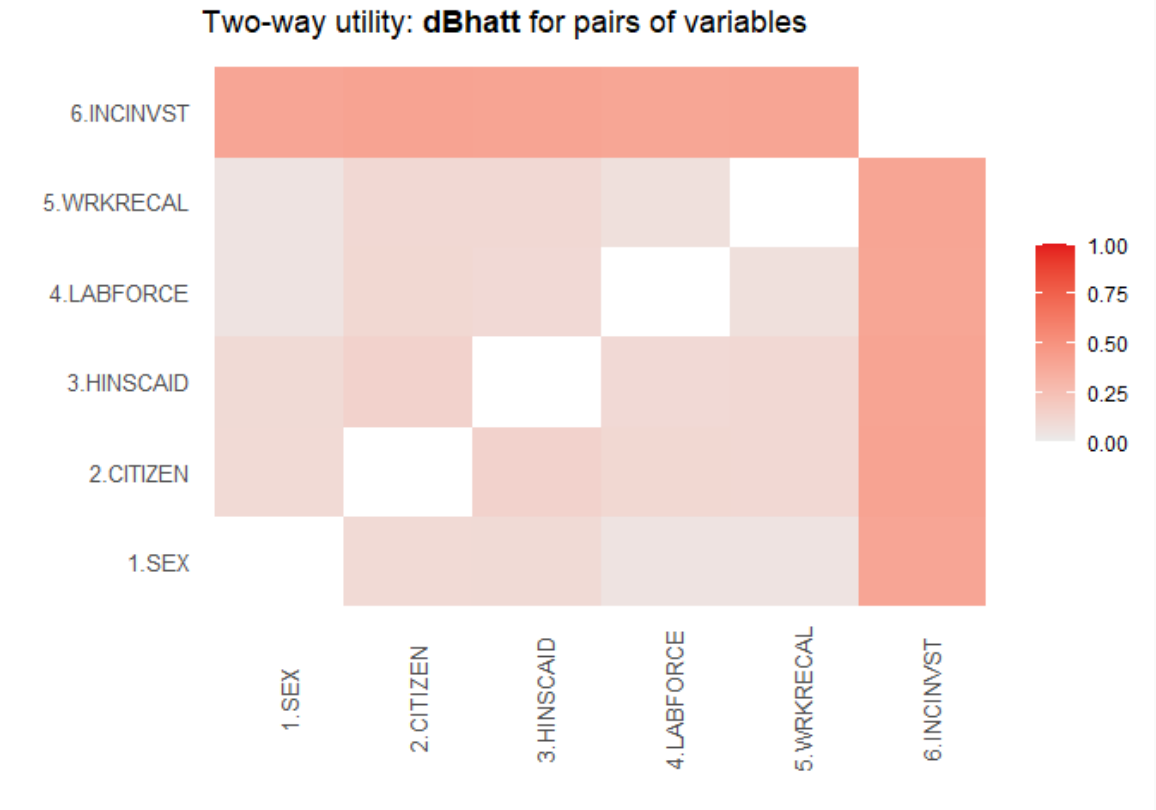

Two-way utility: S_pMSE for pairs of variables
6.INCINVST
5.WRKRECAL
4.LABFORCE
3.HINSCAID
2.CITIZEN
1.SEX
1.SEX
2.CITIZEN
3.HINSCAID
4.LABFORCE
5.WRKRECAL
6.INCINVST
20
15
10
5
0

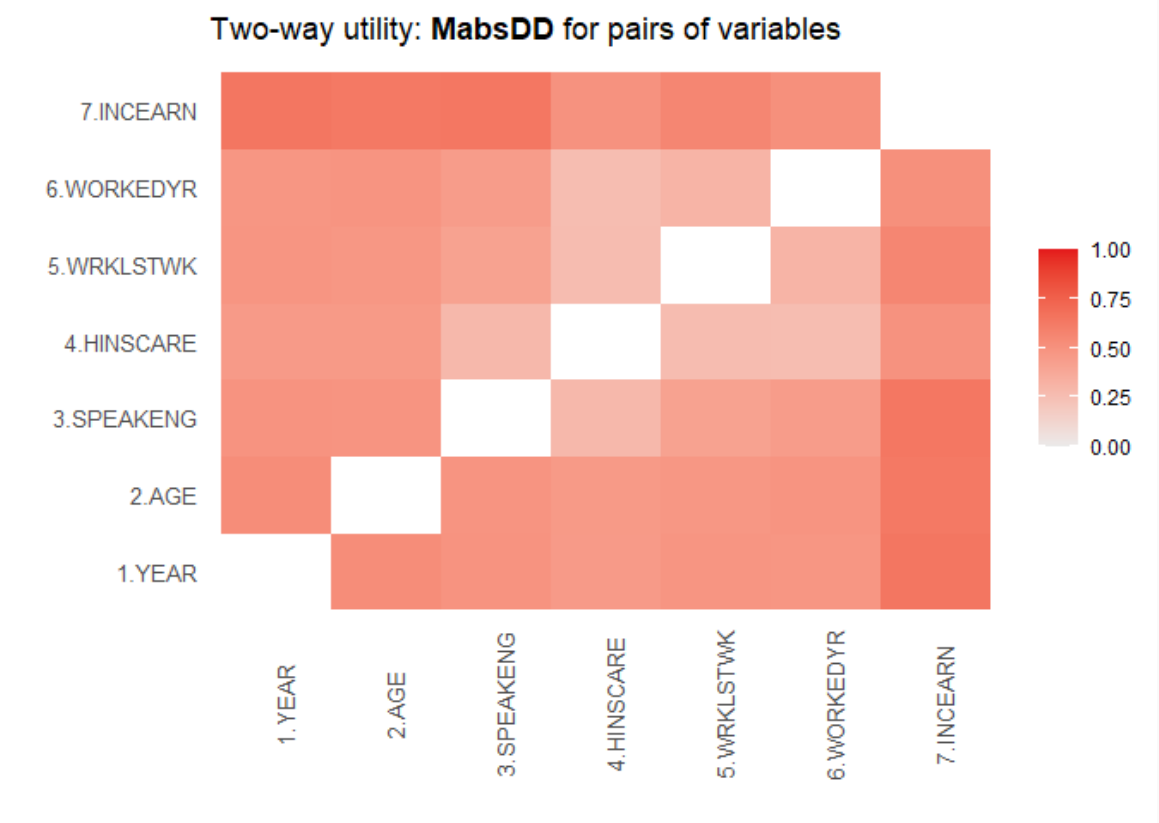

Two-way utility: dBhatt for pairs of variables
7.INCEARN
6.WORKEDYR
5.WRKLSTWK
4.HINSCARE
3.SPEAKENG
2.AGE
1.YEAR
1.00
0.75
0.50
0.25
0.00
1.YEAR
2.AGE
3.SPEAKENG
4.HINSCARE
5.WRKLSTWK
6.WORKEDYR
7.INCEARN

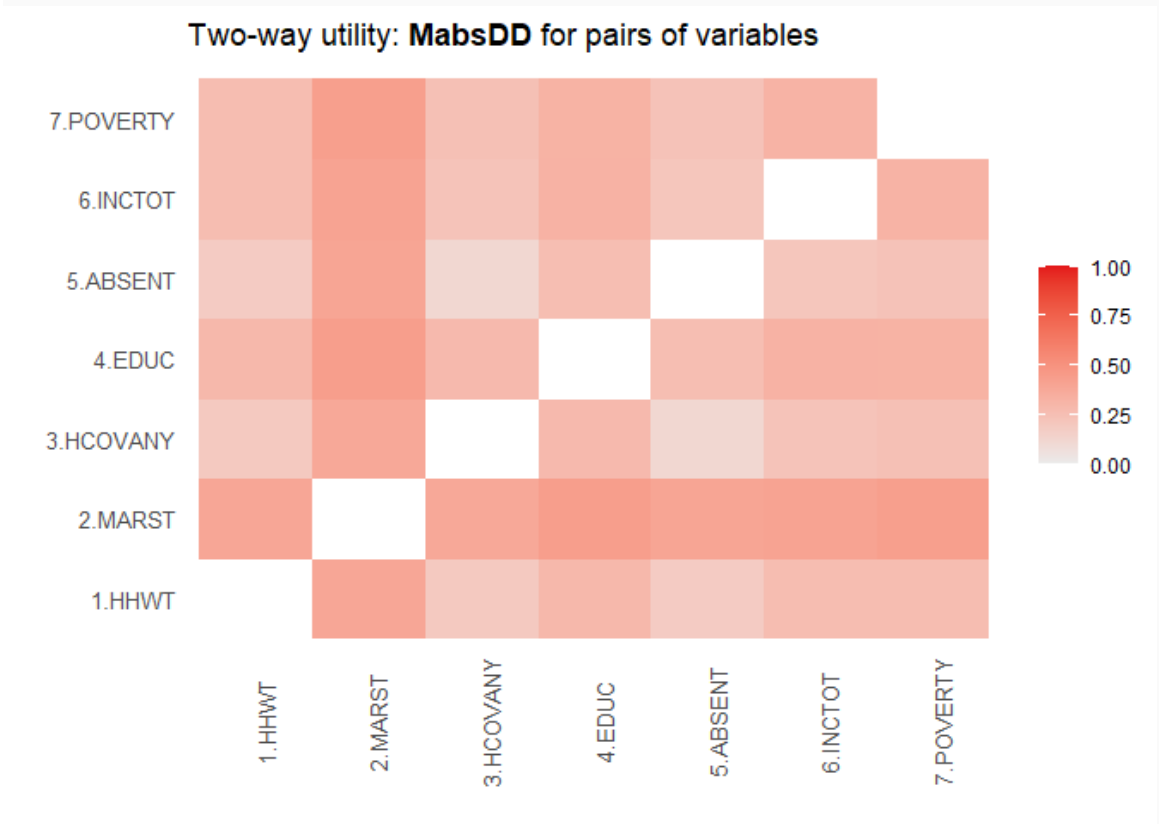

Two-way utility: dBhatt for pairs of variables
7.POVERTY
6.INCTOT
5.ABSENT
4.EDUC
3.HCOVANY
2.MARST
1.HHWT
1.00
0.75
0.50
0.25
0.00
1.HHWT
2.MARST
3.HCOVANY
4.EDUC
5.ABSENT
6.INCTOT
7.POVERTY

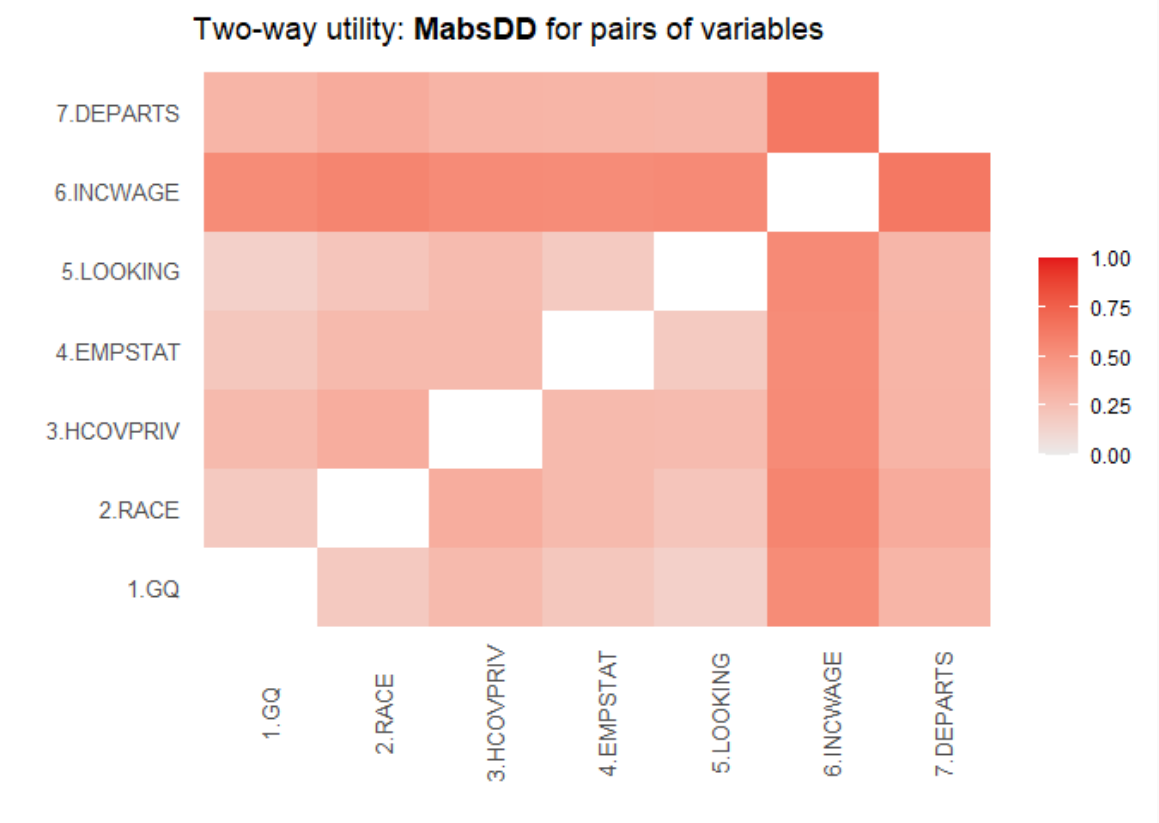

Two-way utility: dBhatt for pairs of variables
7.DEPARTS
6.INCWAGE
5.LOOKING
4.EMPSTAT
3.HCOVPRIV
2.RACE
1.GQ
1.GQ
2.RACE
3.HCOVPRIV
4.EMPSTAT
5.LOOKING
6.INCWAGE
7.DEPARTS
1.00
0.75
0.50
0.25
0.00

```
## NULL
```

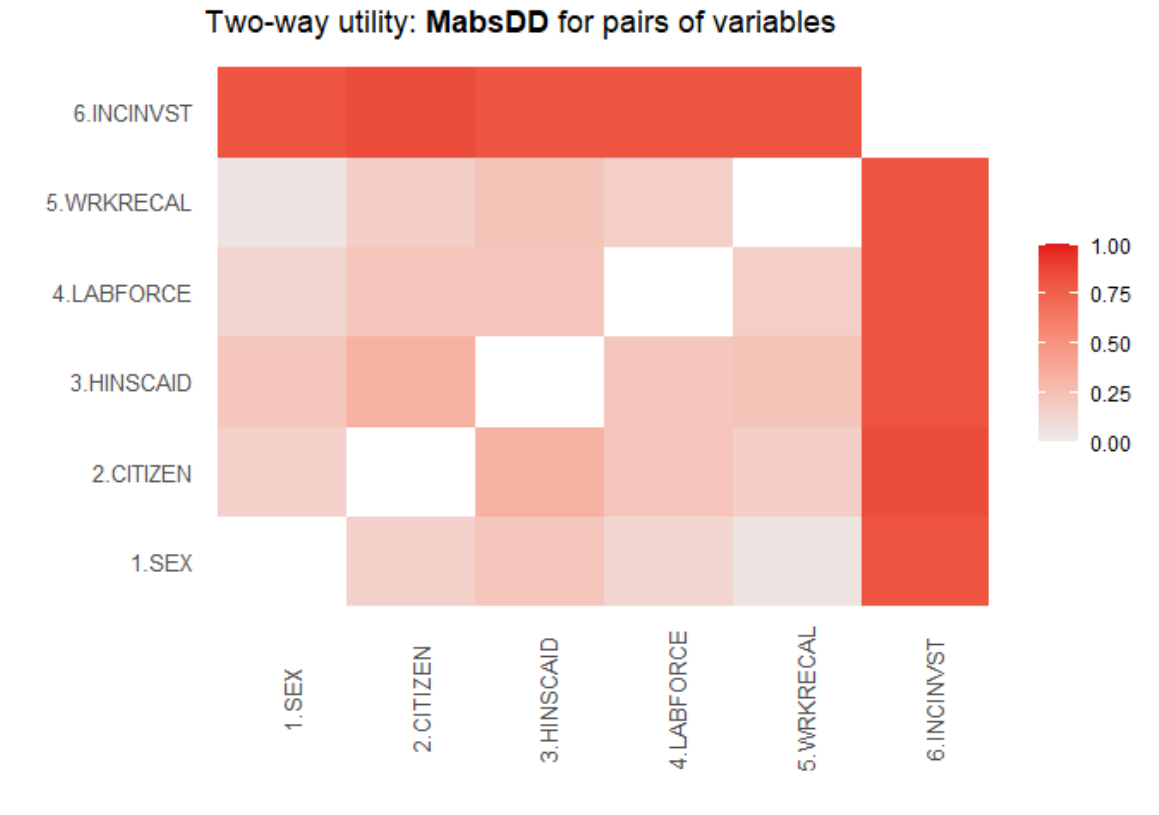

Two-way utility: dBhatt for pairs of variables
6.INCINVST
5.WRKRECAL
4.LABFORCE
3.HINSCAID
2.CITIZEN
1.SEX
1.SEX
2.CITIZEN
3.HINSCAID
4.LABFORCE
5.WRKRECAL
6.INCINVST
1.00
0.75
0.50
0.25
0.00


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


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




```
## NULL
```



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

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

Information.Loss
0.5616973

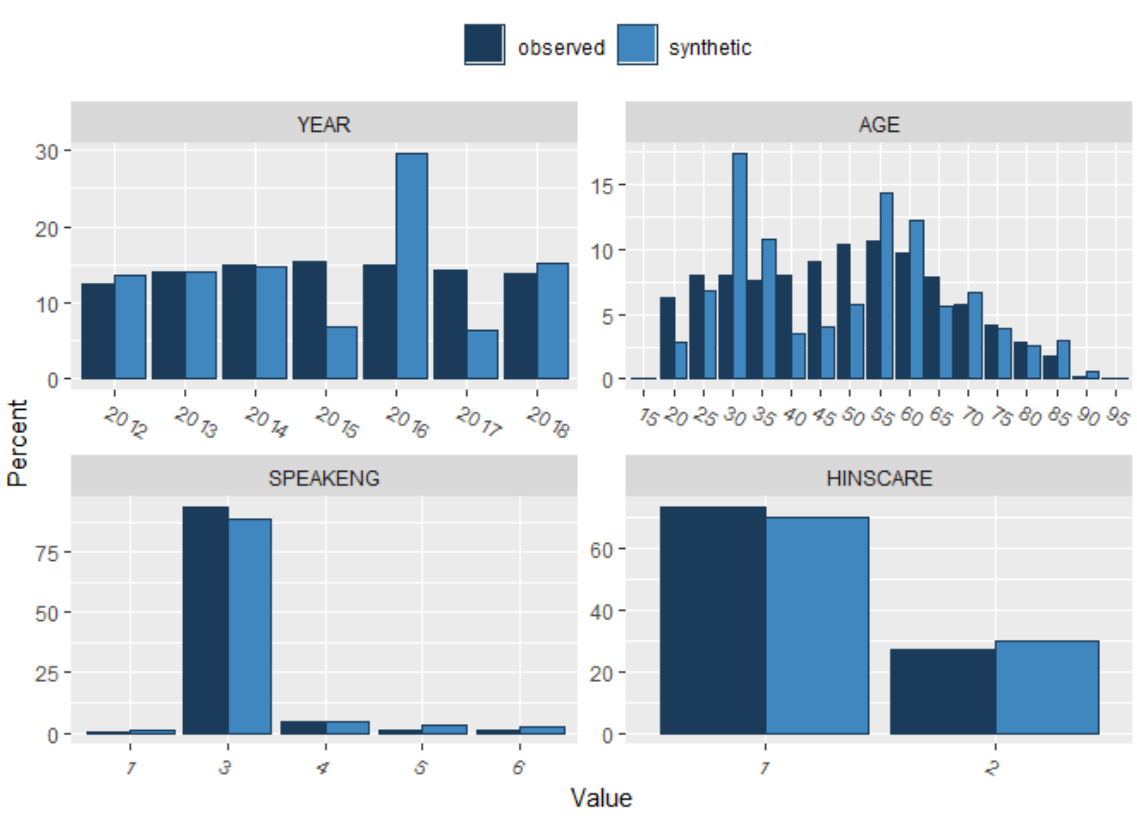
Individual Distances for Information Loss:

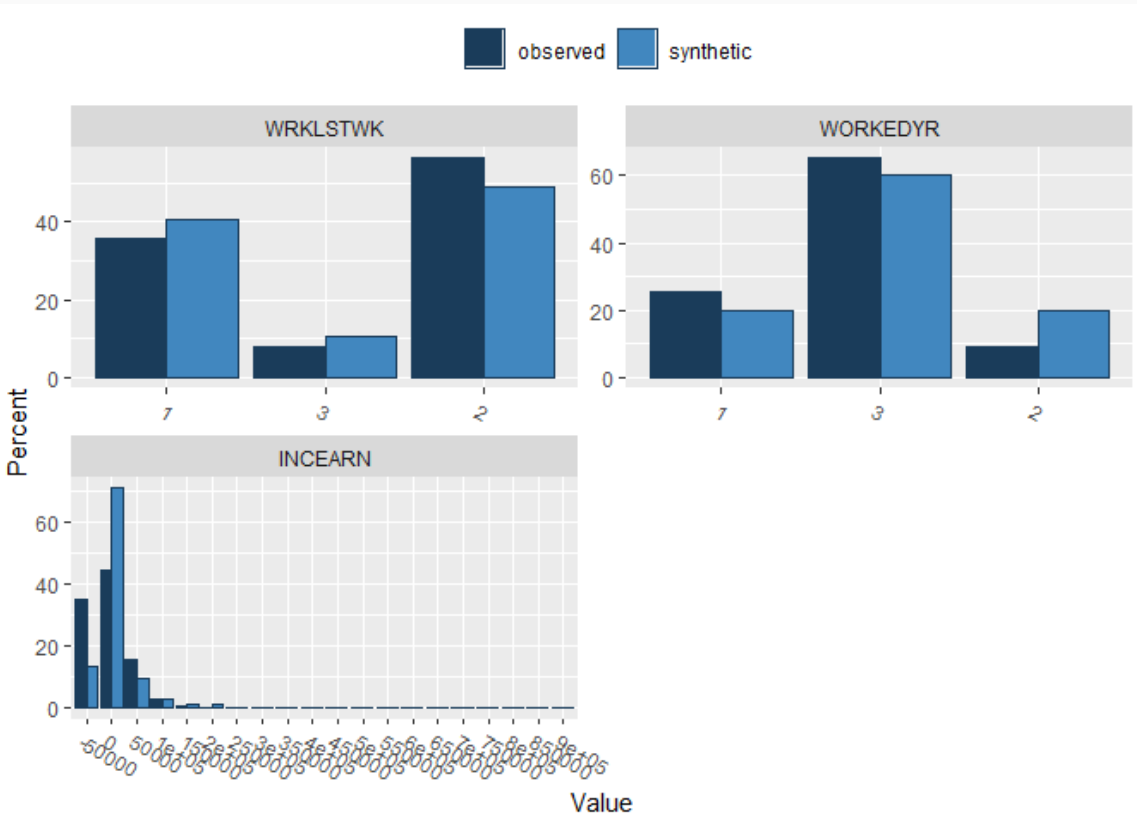
```
##       YEAR       HHWT         GQ      PERWT        SEX        AGE      MARST
## 0.85508418 0.95945851 0.09829975 0.95905994 0.49935230 0.91471230 0.69000996
##       RACE     HISPAN    CITIZEN   SPEAKENG    HCOVANY   HCOVPRIV    HINSEMP
## 0.27595704 0.11088185 0.17048380 0.25596575 0.18490612 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
```

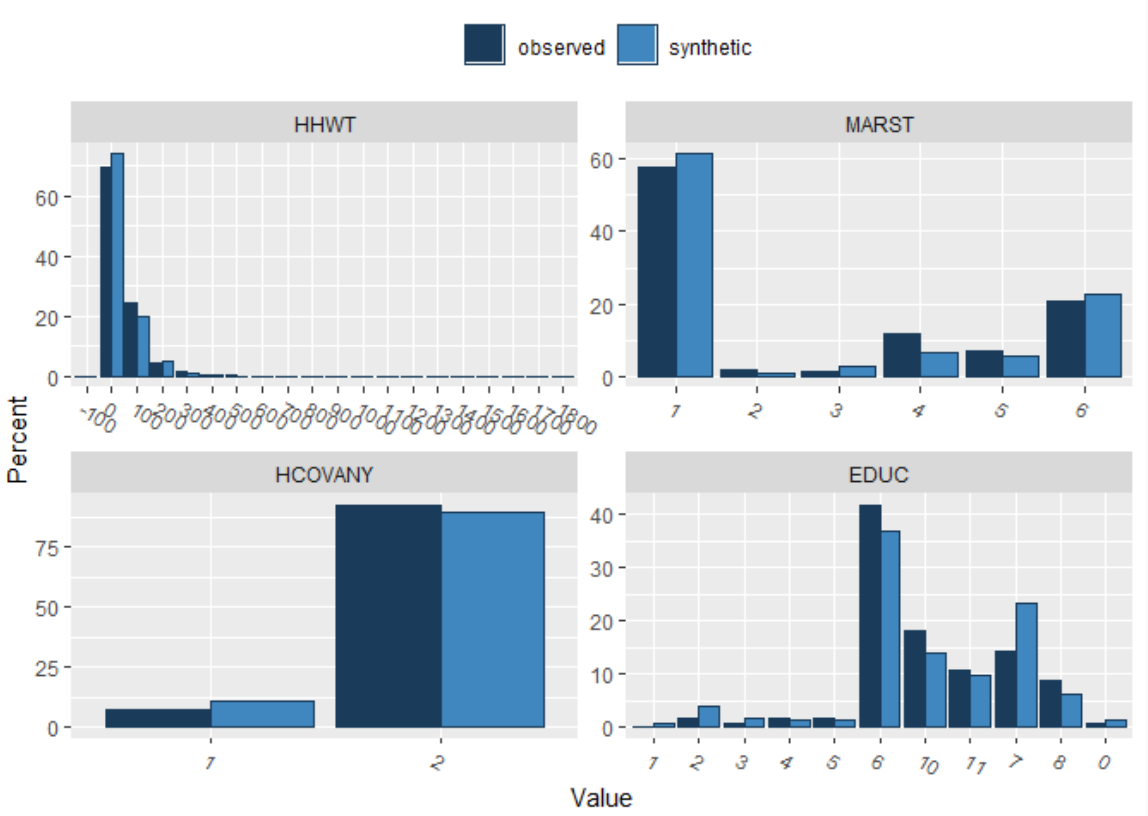
Tuning and Optimizations

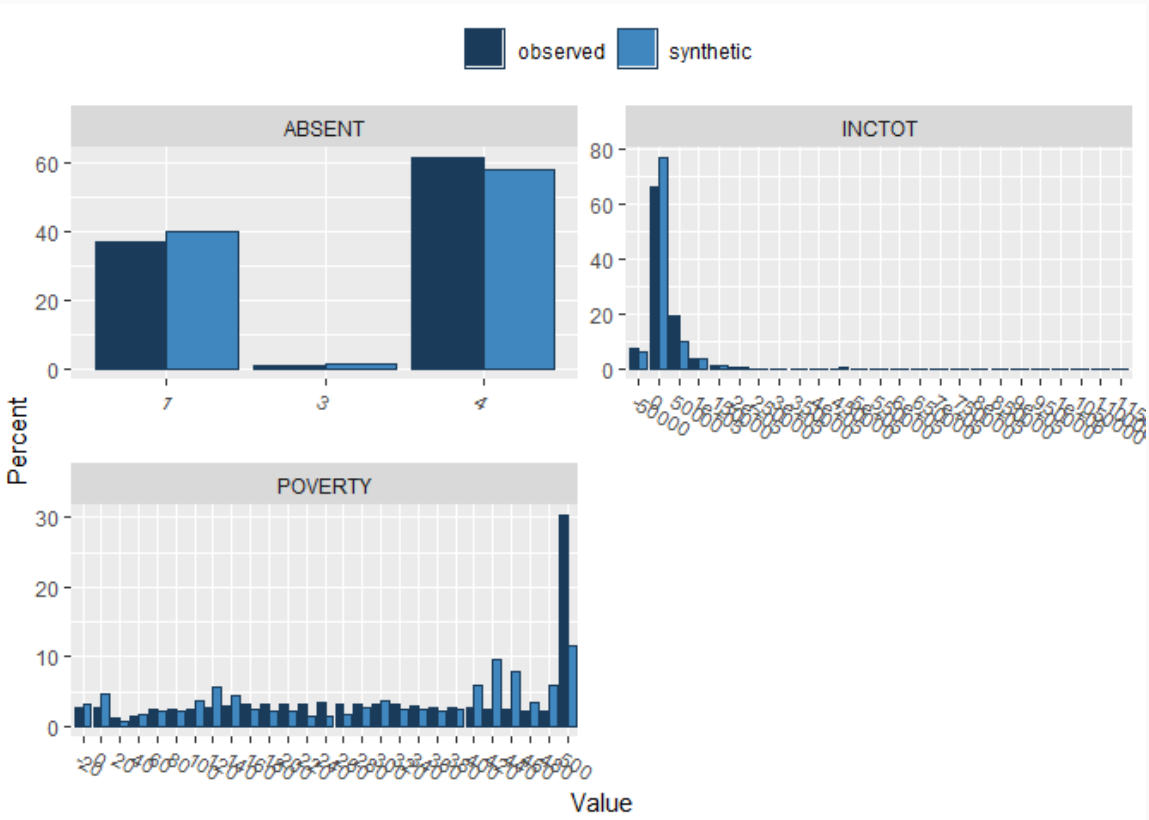
We also tried to optimize parameters and settings for the **GAN** methods on the ACS dataset. Our main problem here was our **limited computing time**. We tried using `CopulaGAN`, which we stopped (without result) after 8h computing time. Also for `ctgan` computing time was an issue. Our first try with `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.

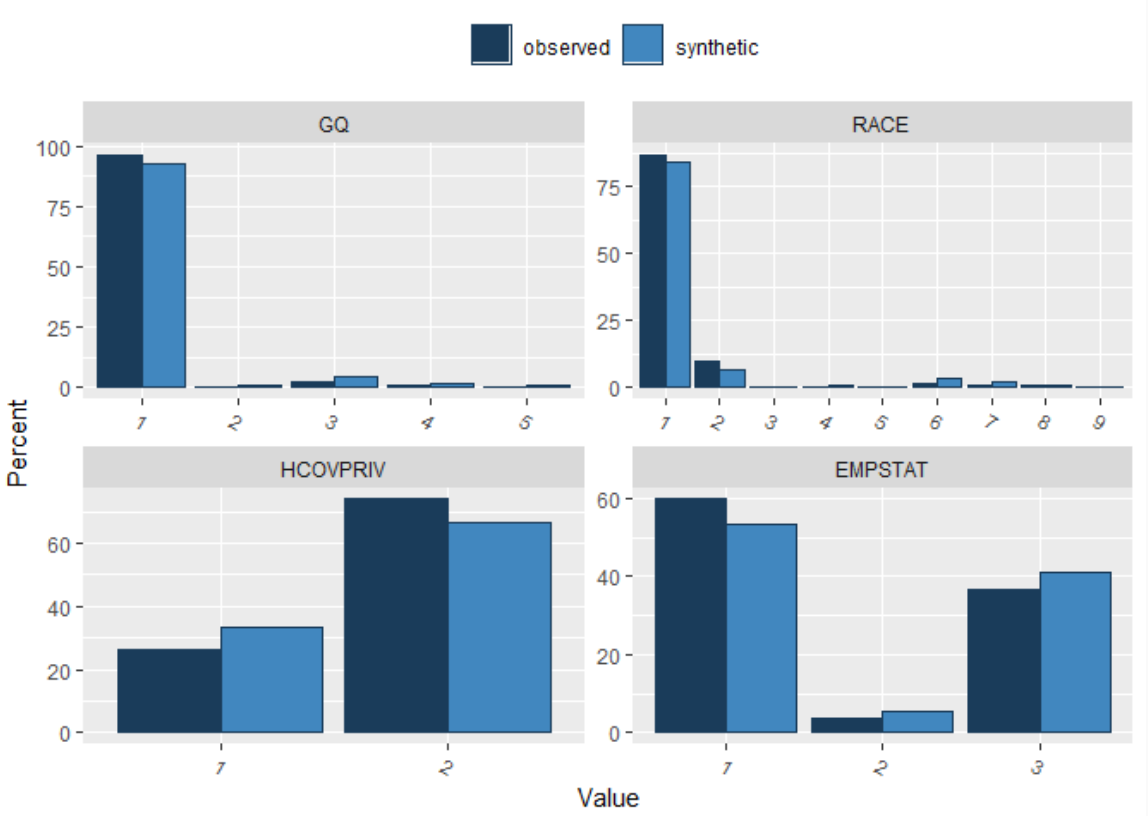
Here are some measures and plots for `epochs = 10`. As can be seen with lower usability results than our final model.

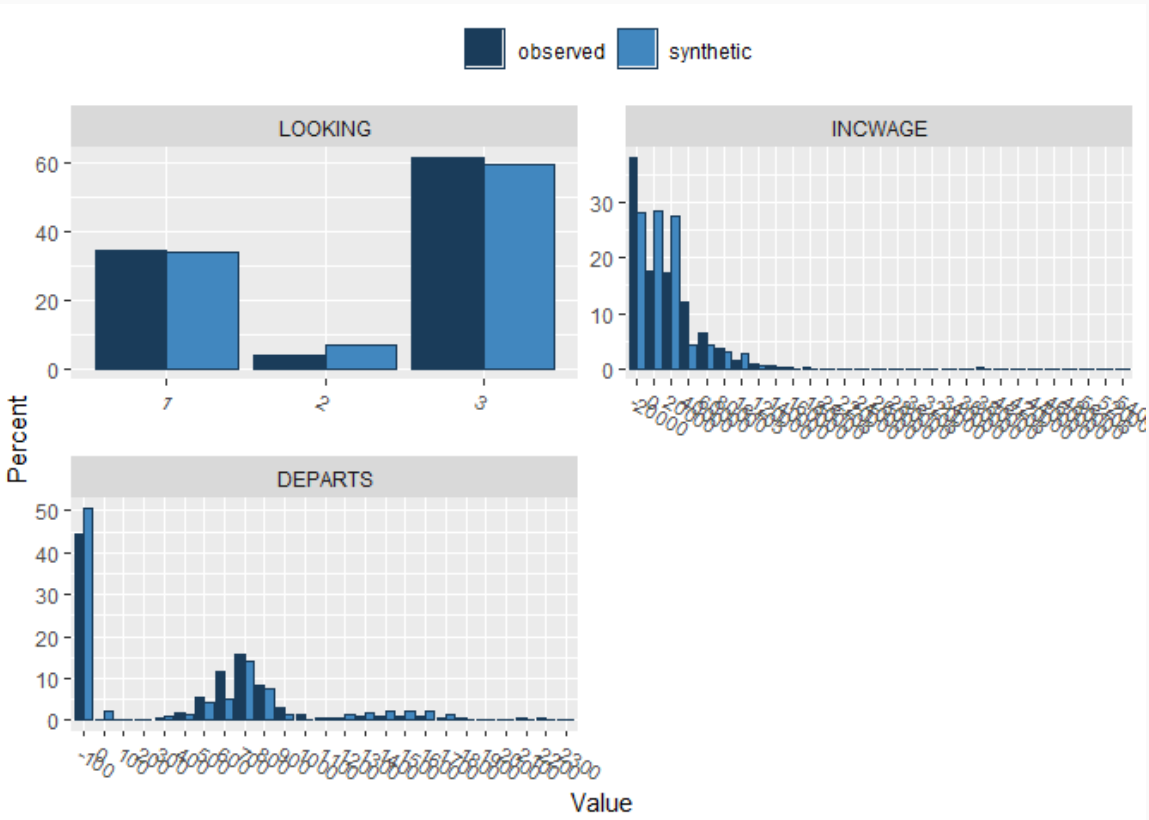

observed
synthetic
YEAR
AGE
SPEAKENG
HINSCARE
Percent
Value

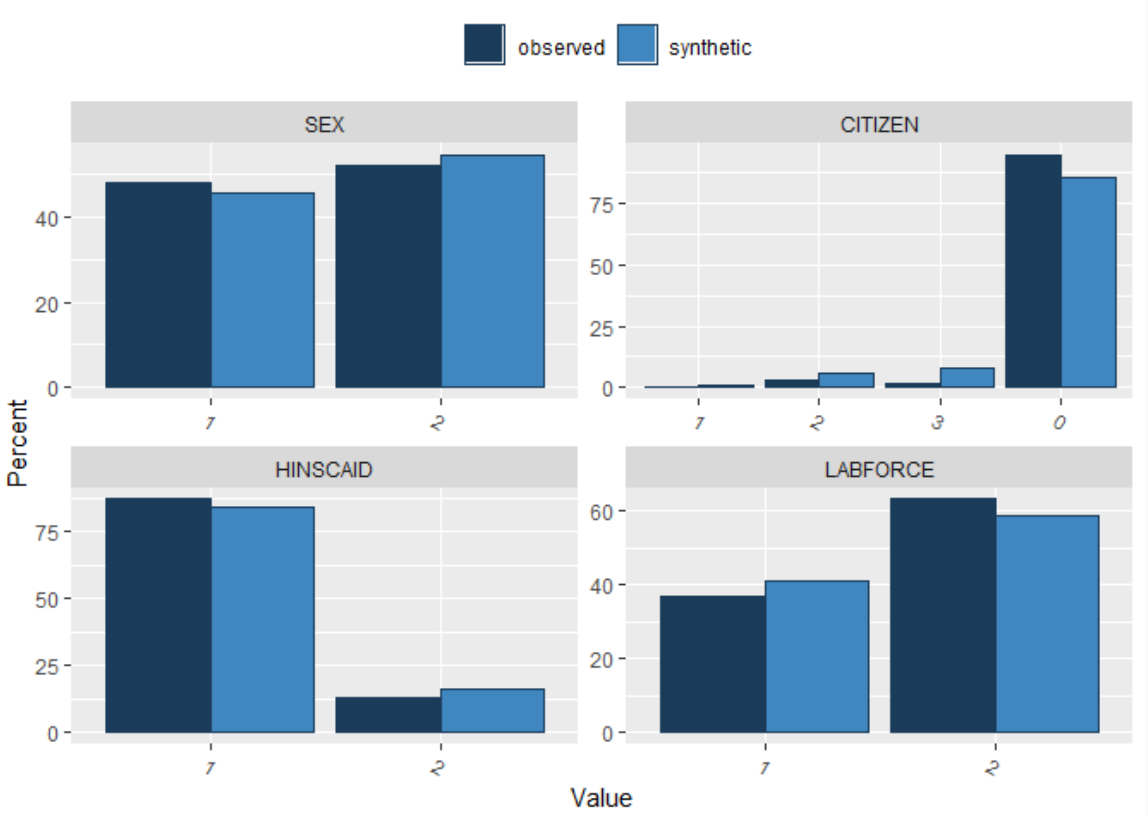

observed
synthetic
WRKLSTWK
WORKEDYR
INCEARN
Percent
Value

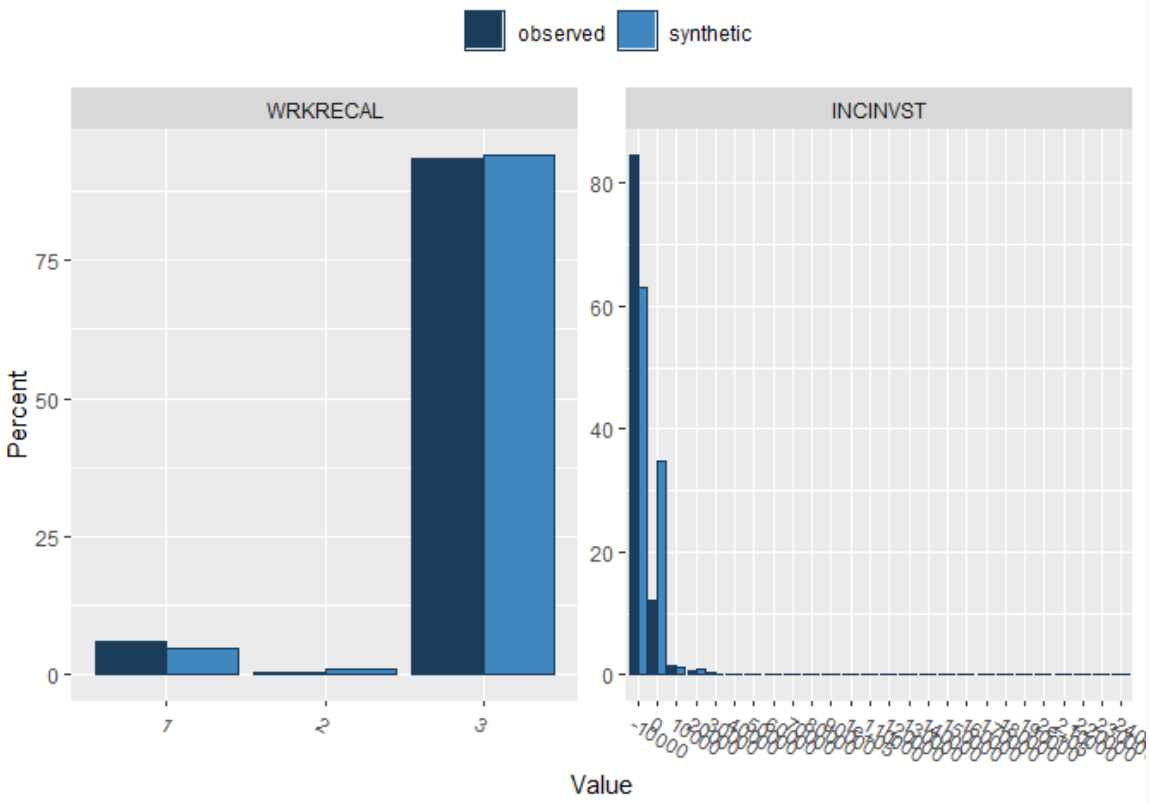

observed
synthetic
HHWT
MARST
HCOVANY
EDUC
Percent
Value


observed
synthetic
ABSENT
INCTOT
POVERTY
Percent
Value


observed
synthetic
GQ
RACE
HCOVPRIV
EMPSTAT
Percent
Value


observed
synthetic
LOOKING
INCWAGE
DEPARTS
Percent
Value





	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