SYNDATA

Evaluation report of synthetic dataset

recovered_data_train.csv

Introduction 2
Model 1 2
Features similarity 2
a. Individual distributions 3
Wasserstein Distance 3
Individual distributions visuals 3
Overall WD distribution 5
b. Pairwise distributions 5
Pearson's correlation 5
Mutual Information 6
Missing Values 7
a. Missing Values Matrices 7
Pearson's Correlation 7
Mutual Information 8
Risk of Re-Identification 8
Appendix: Model 2 8
Features similarity 9
Missing Values 13
Risk of Re-Identification 14
Models' comparison 14

Introduction

The evaluation reports assess the quality of the synthetic datasets produced by Syndata based on the original datasets provided by Region Västerbotten (RVB). Syndata produced 2 evaluation reports: 1 evaluation report covering model 1 and 2 for *recovered* dataset and 1 report for *discharged* dataset.

Syndata concluded that 2 models deem good synthetic datasets. These 2 models can be used by RVB to sample synthetic datasets of their preferred size. Syndata concludes that both models have the potential for quality synthetisation. Given good quality synthetic datasets, RVB can achieving its goal of predicting patients' recovery.

The current report evaluates the characteristic of the synthesized **recovered_data_train.csv** dataset sampled with **Model 1**. This report considered general statistics and visuals as comparison tools between the original and synthesized dataset. The size of the synthetic datasets evaluated are of the same size as the original (e.g. "1x").

Comparison datasets:

- recovered_data_train.csv
- recovered_synthetic_model1__1x_dataset.csv

The datasets and the evaluation framework (as jupyter notebooks) are available on the RVB server.

Model 1

Model 1 uses CTGAN networks, a collection of Deep Learning based Synthetic Data Generators for single table data, which are able to learn from real data and generate synthetic clones with high fidelity. The CTGAN model is available in **sdv** library.

Features similarity

Once we have created a synthesized datasets of the same size as the original, the next step is to visualize how well the properties of each feature have been preserved. A naive method is to evaluate the individual distributions one by one. As a secondary method we will be looking at pairwise distributions to understand how well the relations between features are preserved in the synthetic datasets.

a. Individual distributions

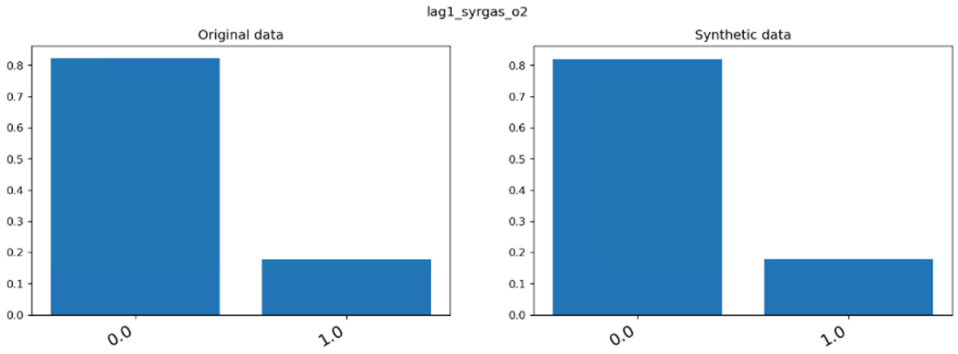
Wasserstein Distance

We compare individual distributions one-to-one using a distance metric, namely Wasserstein Distance (WD). A value of 0 means that the two distributions are identical.

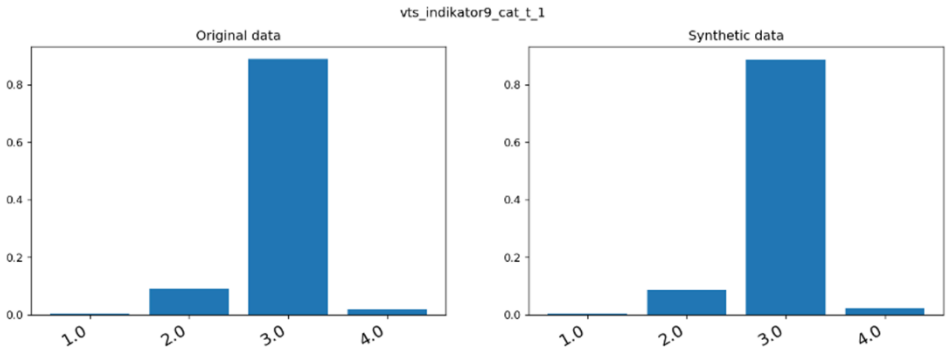
The distance is also known as the earth mover's distance, since it can be seen as the minimum amount of "work" required to transform one distribution into another one, where "work" is measured as the amount of distribution weight that must be moved, multiplied by the distance it has to be moved.

Individual distributions visuals

Examples of similar distributions



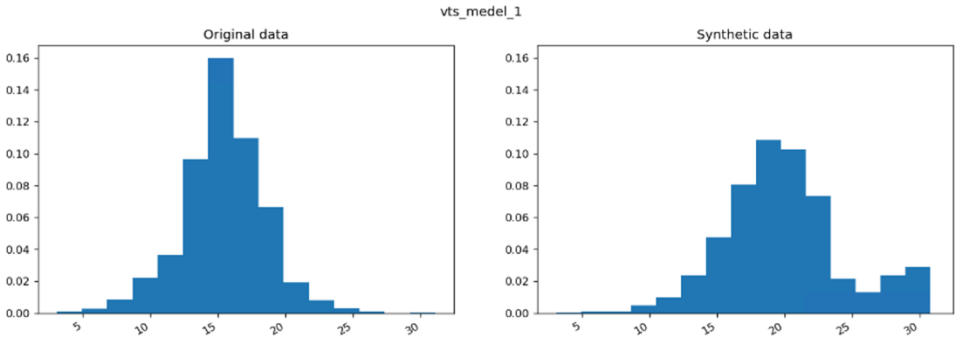
WD: 0.0019



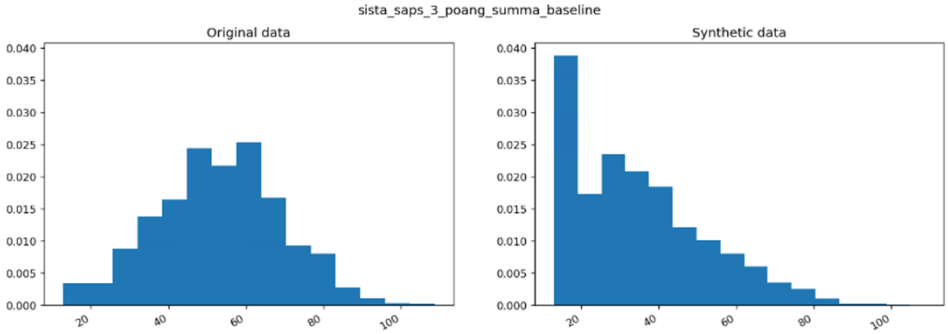
WD: 0.0032

Interpretation: The most similar 5 features are categorical features such as *lag1_syrgas_o2, vts_indikator9_cat_t_1 ,oral_tub_t_1, vts_indikator3_cat_t_1, vts_indikator9_txt_t_1* . Their distributions are almost identical between original and synthetic datasets. Distribution shape is preserved indicating a close replica to the original.

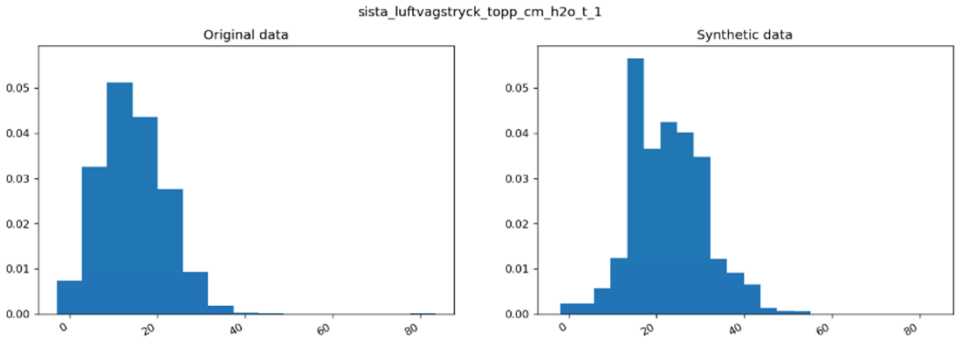
Examples of less similar distributions:



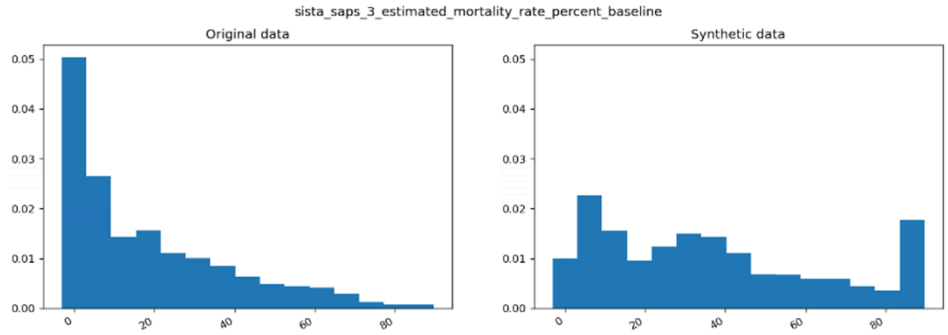
WD: 1.46



WD: 1.1615



WD: 1

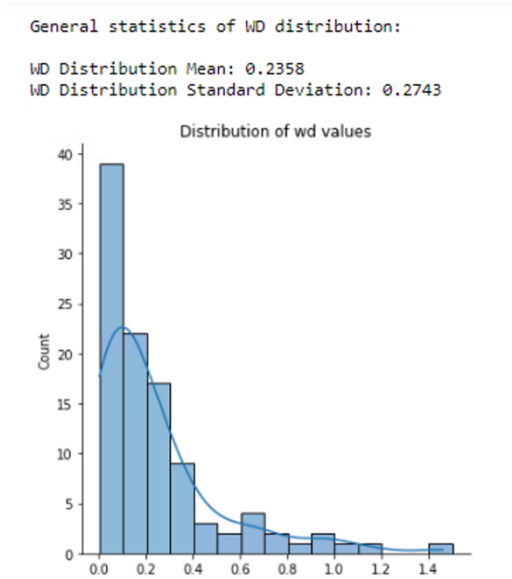


WD: 0.91

Interpretation: The least 4 similar features are numerical values. Distribution shape is similar even though the higher WD distance.

Overall WD distribution

As a way to summarize the individual distributions quality, we use a distribution plot of Wasserstein Distances. The closer the mean to 0 and the smaller the standard deviation, the more similar are the features to the original dataset.



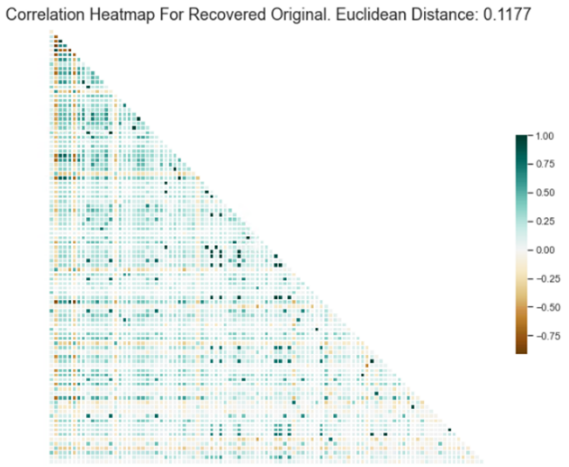
Interpretation: Most of the fields show identical distributions. There are few outliers, however the distance for the 4 outliers does not go above 1.4.

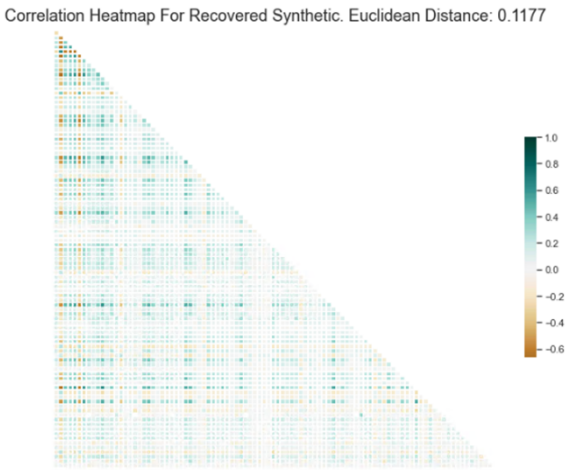
b. Pairwise distributions

Pearson's correlation

Pearson's correlation coefficient measures the strength of a linear association between two variables. Its value range spans over [-1,1]. A value of 0 indicates that there is no association between the two variables. We compare the distance between these 2 matrices by the Euclidean distance. A point with an increased colour intensity means a pair of 2 features have high correlations.

Finally, we compare the distance between these 2 matrices with a Euclidean distance. The Euclidean Distance for the correlation matrix ranges on a **[0, 2] scale.**



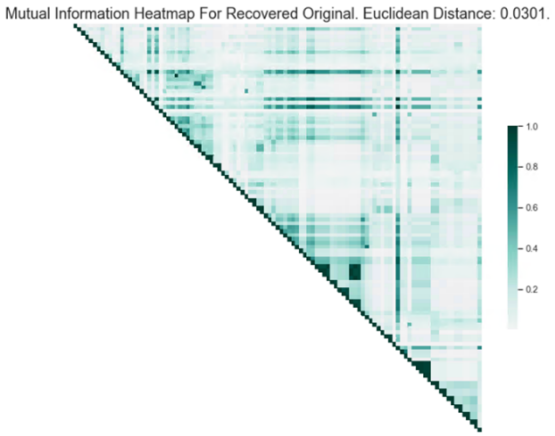


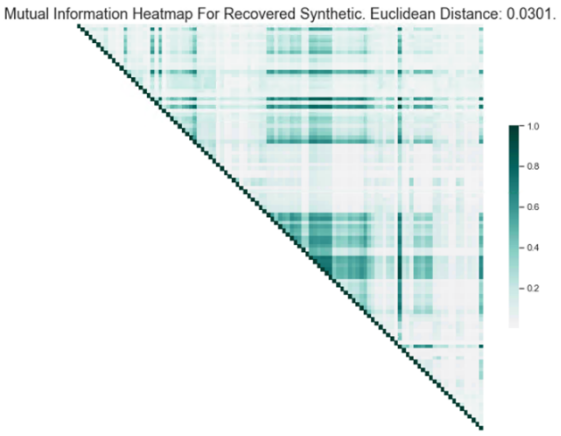
Euclidean distance: 0.1177

Interpretation: Overall patterns of correlations are preserved. Some strong positive correlations show a lower intensity in the synthetic. The overall distance **(0.1177)** is small indicating a high association of pairwise distributions.

Mutual Information

Mutual information measures the relatedness between two random variables. It has a value range of [0,1]. A value of 0 indicates that there is no relatedness between the two variables. We use the Mutual Information metric to plot a heatmap indicating the level of relatedness between any 2 features. A higher intensity of the point indicates a higher relatedness between 2 specific features.





Euclidean distance: 0.0301

Finally, we compare the distance between these 2 matrices with a Euclidean distance. The Euclidean Distance for the mutual information matrix ranges on a **[0, 1] scale.**

Interpretation: Overall patterns are well preserved with similar intensities. The overall distance is small **(0.03)** indicating a high relatedness between fields.

Missing Values

We evaluate the relation between missing values with Mutual Information (to capture non-linear relations) and Pearson's Correlation (to capture linear relations).

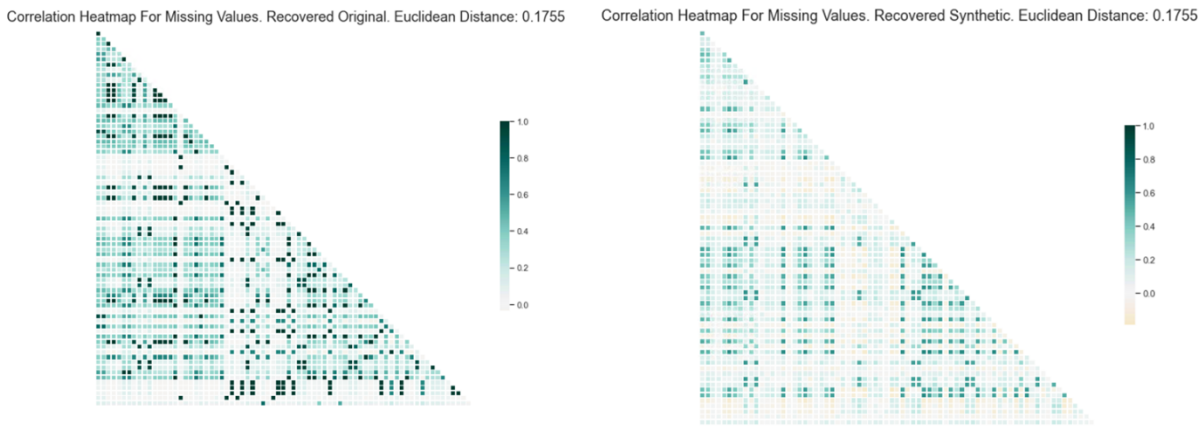
Euclidean Distance is used to measure the similarity between the 2 matrices. It compares the differences between two Pearson's Correlation (or Mutual Information) matrices. The closest to 0, the more similar the 2 matrices are against each other.

We have a different distance scale for each metric. The Euclidean Distance for the correlation matrix ranges on a **[0, 2]** scale, while the distance for the mutual information matrix ranges on a **[0, 1]** scale.

a. Missing Values Matrices

Pre-steps: We map the original and synthetic datasets values to matrices that indicate null or not null values. Later, we remove the columns without any variation. Finally, we compute the mutual information/correlation matrix.

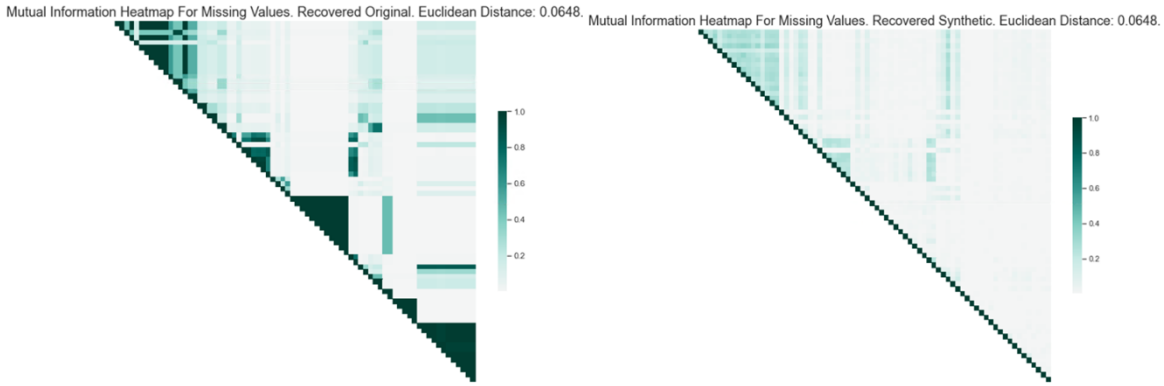
Pearson's Correlation



Euclidean distance: 0.1755

Interpretation: Overall patterns are preserved in the synthetic dataset. Some of the positive correlations are blurred in the synthetic. Distance is low **(0.1755)** indicating a high similarity between the two datasets.

Mutual Information



Euclidean distance: 0.0648

Interpretation: Overall patterns are preserved, with few exceptions. The distance is small **(0.0648)** indicating the missing values patterns are preserved in the synthetic dataset.

Risk of Re-Identification

Risk of re-identification is a relative metric to the original dataset characteristics. It indicates the minimum and maximum risk of the individuals from the synthetic dataset to be re-identified.

The score is influenced by the number of: (1) unique values in each column (2) number of identical values per individual with any point in the original dataset.

Statistic	Measurement
RIR min	0.270
RIR max	0.566
RIR mean	0.422
RIR scale	[0,2]

Interpretation: All datapoints of the synthetic dataset are evaluated at a risk between 0.27 and 0.57. A synthetic individual with a risk of 0 means there is no datapoint in the synthetic with an identical value. All synthetic individuals have a risk in the lower 28% of the overall risk. Based on our experience, the synthetic dataset has a relatively low risk.

Appendix: Model 2

Model 2 is based on copula functions, available in the **sdv** library.

Comparison datasets:

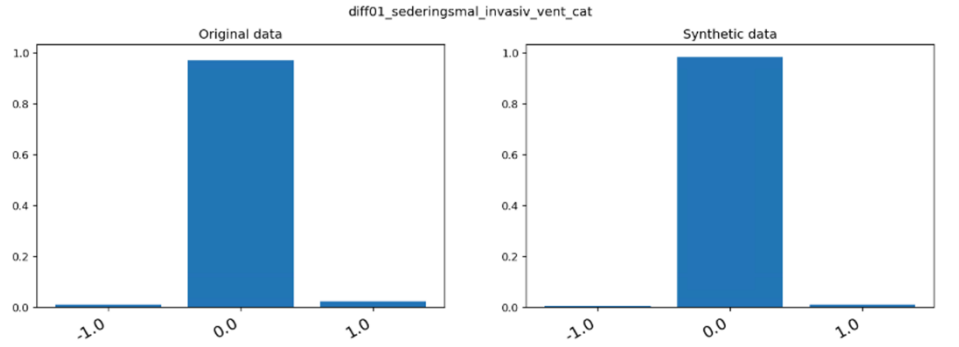
- recovered_data_train.csv
- recovered_synthetic_model2__1x_dataset.csv

Features similarity

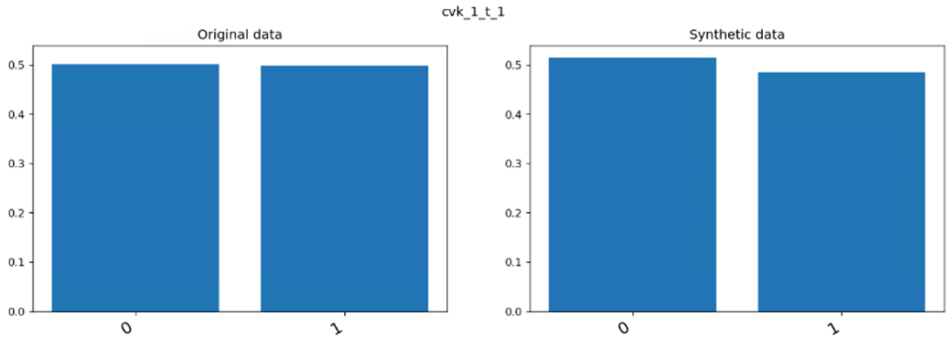
a. Individual distributions

Individual distributions visuals

Examples of similar distributions



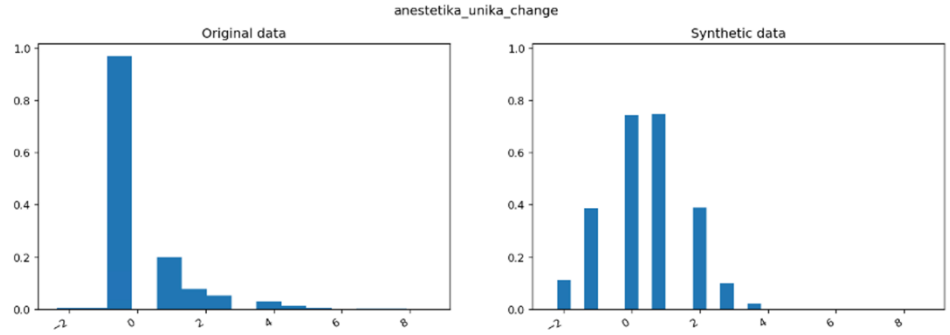
WD: 0.0073



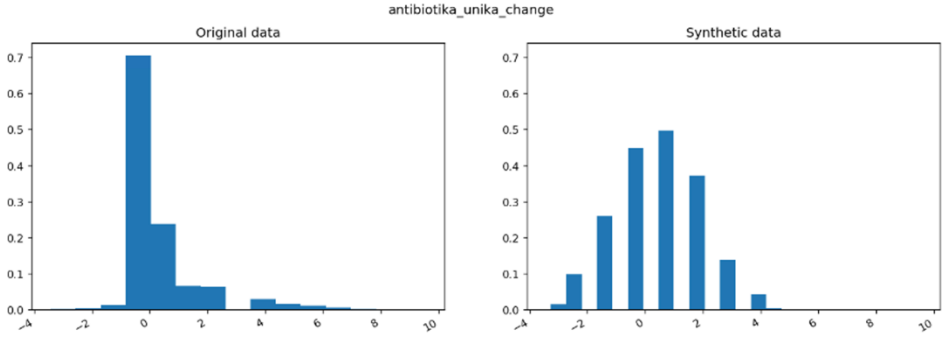
WD: 0.022

Interpretation: The most similar 5 features are categorical features. This include *diff01_sederingsmal_invasiv_vent_cat , cvk_1_t_1, vts_indikator3_cat_t_1 , vts_indikator3_txt_t_1, vts_indikator6_txt_t_1.* Each model produces different top 5 fields.

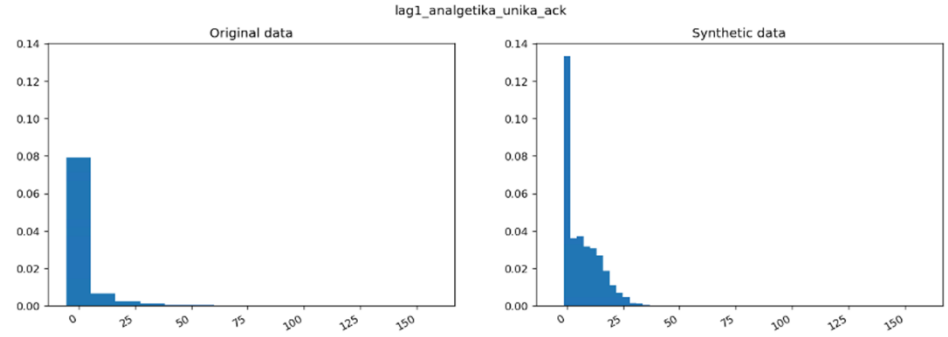
Examples of less similar distributions:



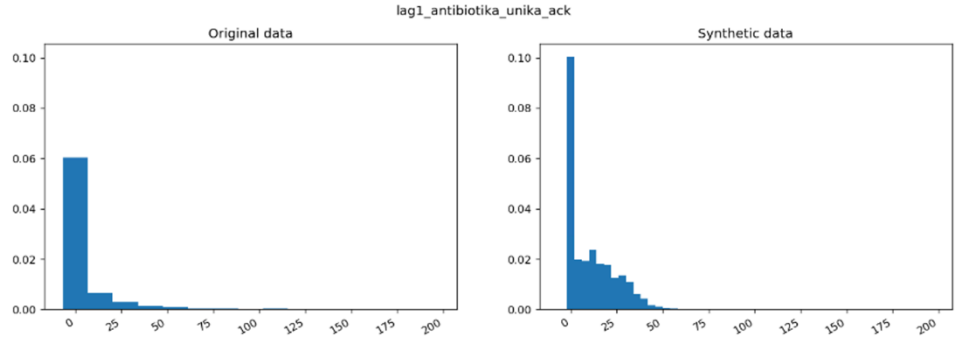
WD: 1



WD: 0.518



WD: 0.478



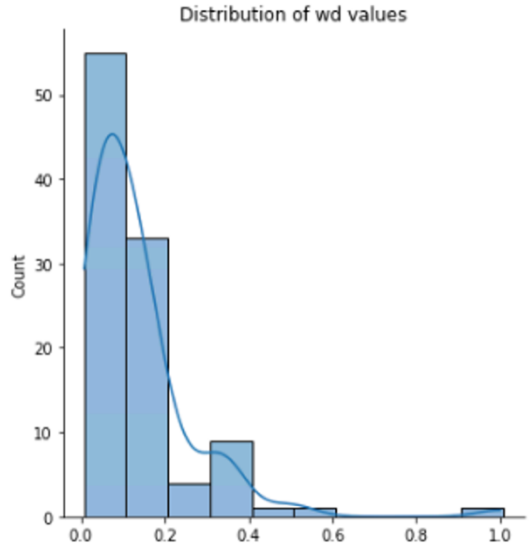
WD: 0.3622

Interpretation: The least similar individual distributions are non-categorical fields. Even though a higher WD, the distribution shape is still preserved. Model 2 is able to correct the worse distributions produced by model 1 and thus, the overall similarity of individual distributions is increased.

Overall WD distribution

```
General statistics of WD distribution:

WD Distribution Mean: 0.1325
WD Distribution Standard Deviation: 0.1348
```

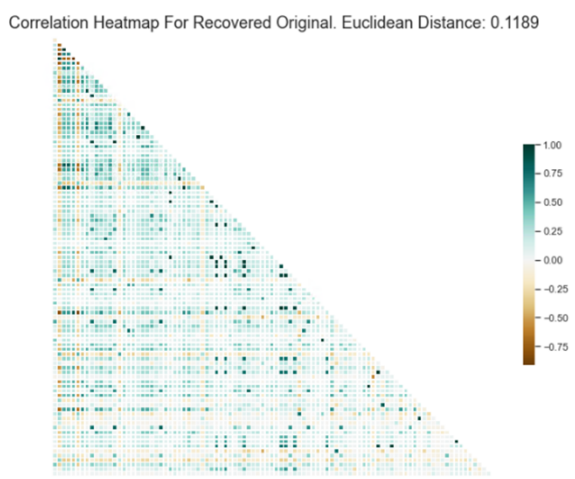


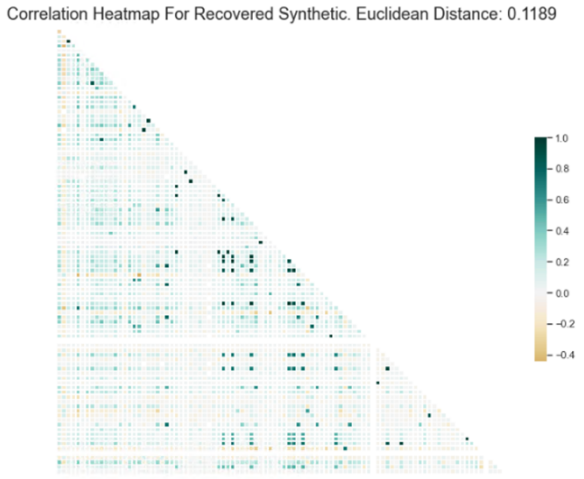
Interpretation: The mean and the standard deviation of model 2 show a big improvement of 0.1 in mean difference. Model 2 produces above 50 features with a WD distance between 0 and 0.1 while model 1 has less than 40.

Overall, model 2 outperforms model 1 in synthesising individual distributions of higher similarity.

b. Pairwise distributions

Pearson's correlation

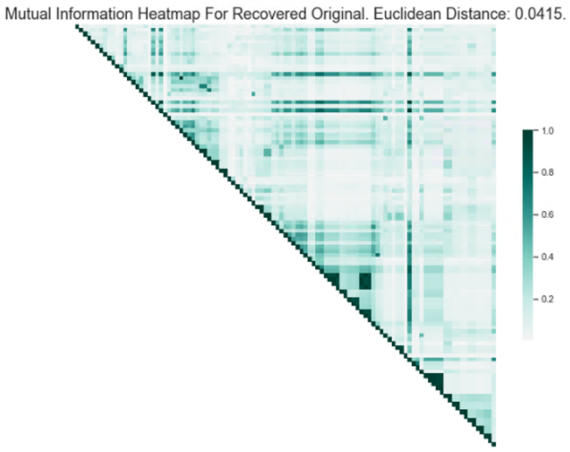


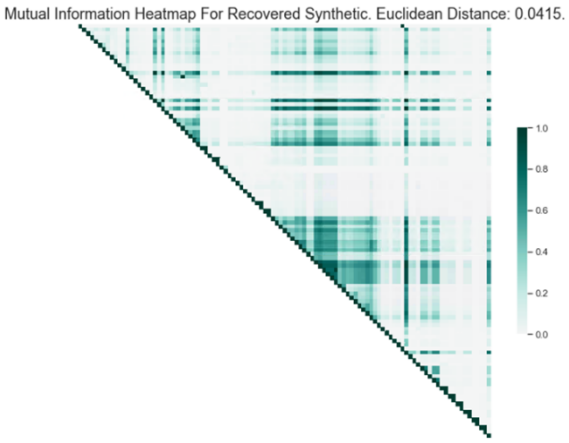


Euclidean distance: 0.1189

Interpretation: High positive correlations are preserved in the synthetic dataset. The overall distance is small indicating a high correlation between fields of synthetic versus original dataset. There is no significant improvement in the Euclidean distance between the 2 models.

Mutual Information





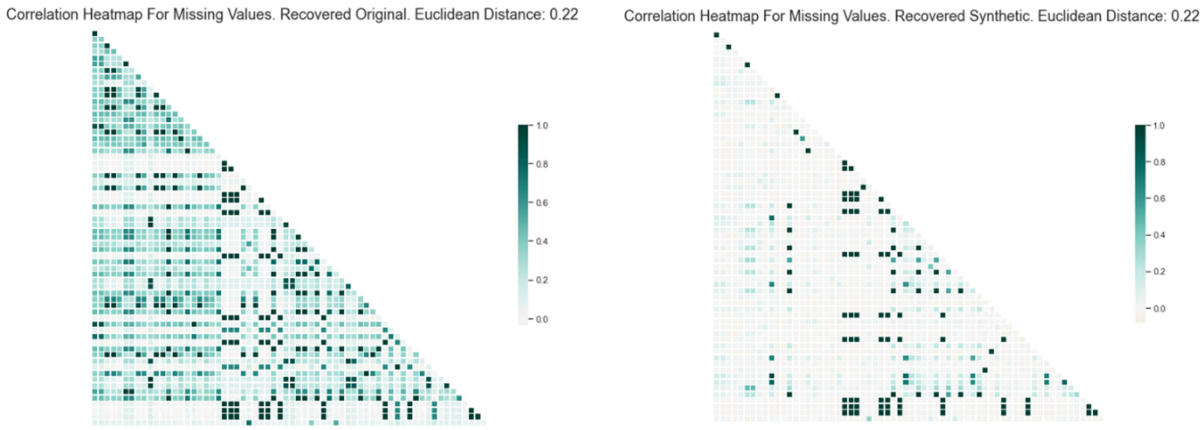
Euclidean distance:0.0415

Interpretation: Overall patterns are well preserved with similar intensities. The overall distance is small indicating a high relatedness between fields. In terms of Euclidean distance for the Mutual Information, model 2 (**0.0415**) performs worse than model 1 (**0.0301**).

Missing Values

a. Missing Values Matrices

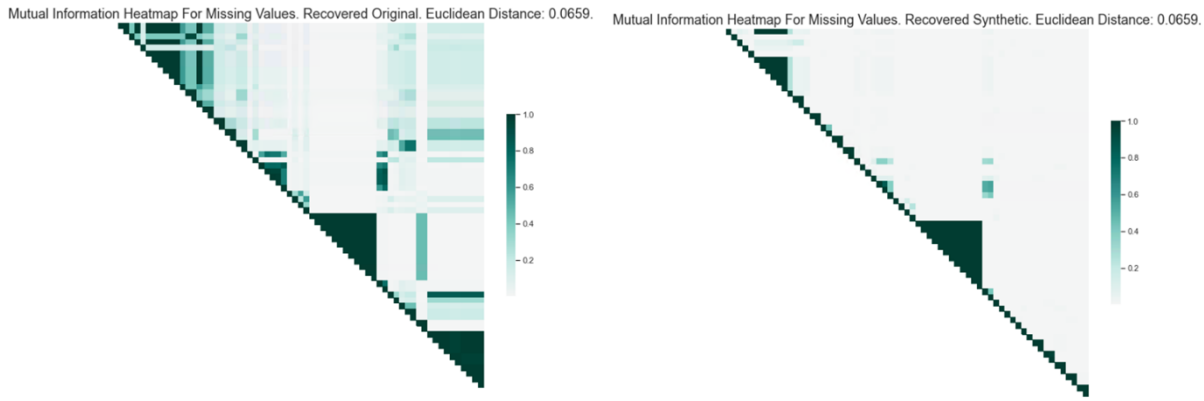
Pearson's Correlation



Euclidean distance: 0.22

Interpretation: The fields with high positive correlations are preserved in the synthetic dataset. Some points lose the correlation sharpness. The overall distance is small indicating the missing values patterns are preserved between the datasets. Model 2 shows an increase in Euclidean distance indicating a loss in the correlation of missing values. The behaviour is attributed to a tendency of model 2 to replace a higher rate of missing values.

Mutual Information



Euclidean distance: 0.0659

Interpretation: Overall patterns are preserved, with few exceptions (eg. lower triangle). High intensity Mutual Information values are preserved in the synthetic dataset, while others are lost. The distance is small indicating the missing values patterns are preserved in the synthetic dataset. Model 2 does not show any improvement in terms of preserving missing values patterns.

Risk of Re-Identification

Risk of re-identification is a relative metric to the original dataset characteristics. It indicates the minimum and maximum risk of the individuals from the synthetic dataset to be re-identified.

The score is influenced by the number of: (1) unique values in each column (2) number of identical values per individual with any point in the original dataset.

Statistic	Measurement
RIR min	0.327
RIR max	0.518
RIR mean	0.418
RIR scale	[0,2]

Interpretation: All datapoints of the synthetic dataset are evaluated at a risk between 0.21 and 0.54. A synthetic individual with a risk of 0 means that there is no datapoint in the original that hold at least one similar value. All synthetic individuals have a risk in the lower 27% of the overall risk. Based on our experience, the synthetic dataset has a relatively low risk.

Models' comparison

Model 1 (based on CTGAN networks) outperforms model 2 in better preserving missing values patterns. This behaviour is explained by a lower rate of replacing missing values. Thus, the patterns are better preserved.

In addition, model 1 outperforms model 2 in terms of pairwise correlations. We see an improvement in the Euclidean score for both metrics (Pearson's correlation and Mutual information).

Model 2 (based on Gaussian Copula) outperforms model 1 in producing more fields with similar individual distributions. On the other hand, model 2 shows a loss in the features' relatedness and correlation.