Luogo, giorno mese anno

TITOLO DEL CONVEGNO

Telephony Synthetic Data Generation via Generative Adversarial Networks

NOME COGNOME

Istat | Direzione

The experimentation (1/2)

Dataset

○ A random sample of WIND's CDR

○ Dimension: 10,000 rows – 4 attributes

○ Attributes: SIM_CODE
CALL_DATE
TIME_CALL
ANTENNA_CODE

Pre-proceesing (steps)

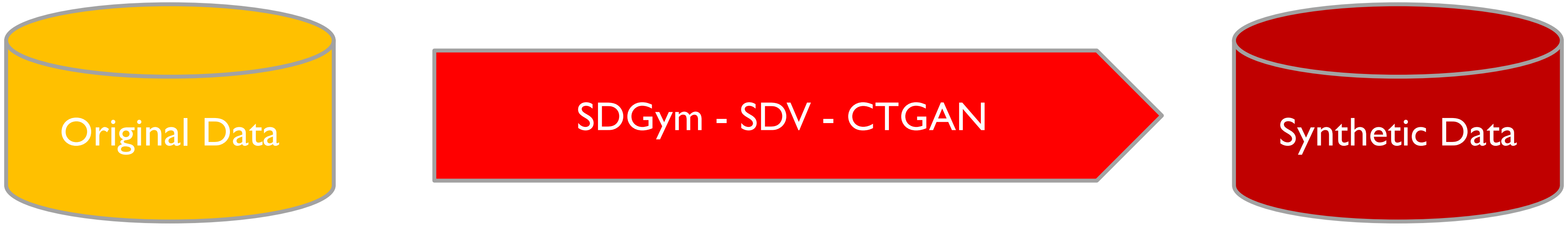
○ Pseudo-anonymization

○ Setting of data-type: ***categorical*** (SIM code, Antenna codes) or ***continuous***(date and time of calls)

The experimentation (2/2)

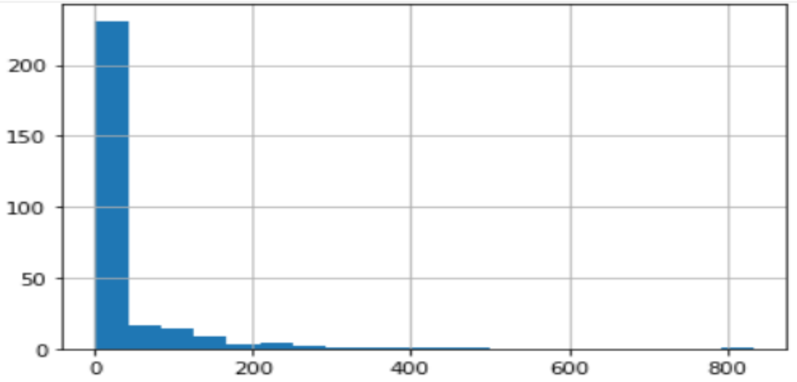
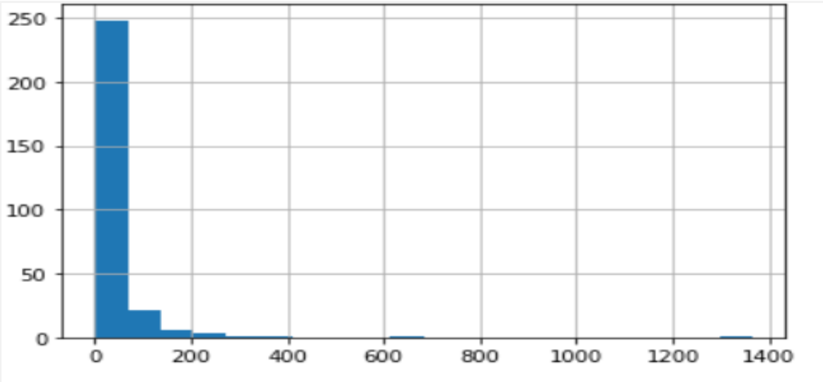
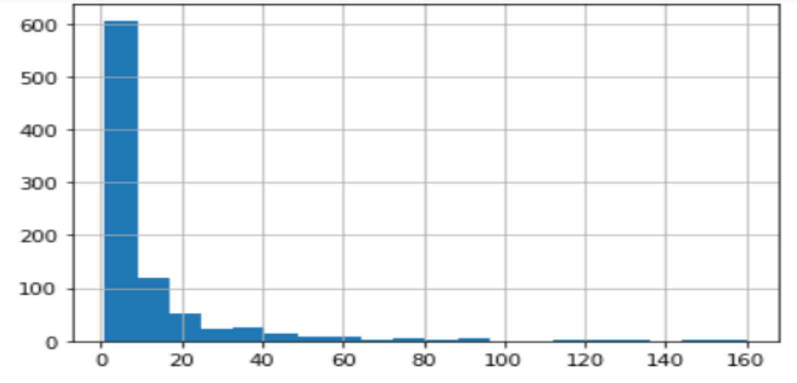
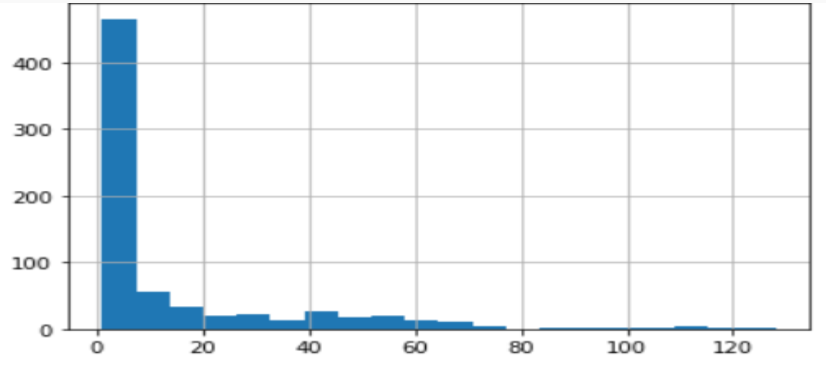
Process

- Input : a random sample of WIND's CDR (***original dataset***)
- Type of process: generate synthetic data
- Framework used: SDGym - Synthetic Data Gym Metrics Evaluation (https://github.com/sdv-dev/SDGym)
- Algorithm used to generate synthetic data: Synthetic Data Vault (SDV) – based on CTGAN
- Output: synthetic dataset of WIND's CDR (***synthetic dataset)***



Original and Synthetic Data Univariate Distribution Visualization

Graphical analysis with **Univariate Distributions Comparisons** on the **Categorical** Variables from **Real Data** and **Synthetic Data**.

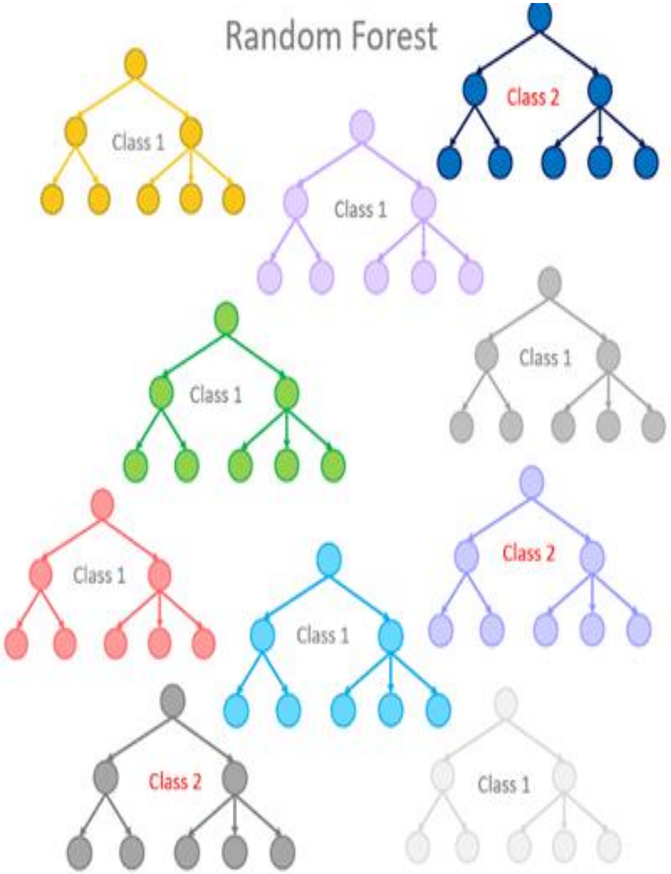
	ORIGINAL DATASET	SYNTHETIC DATASET
SIM_CODE		
ANTENNA_CODE		

Utility Metrics - Machine Learning Test: Random Forest Classifier Accuracy

- The **Random Forest (RF)** is a classification (there is also a Regression version) algorithm consisting of many **Decisions Trees**. It uses **bagging** and **feature randomness** when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

- Estimating the **Accuracy** of a Classifier such as **RF** on a column taken as the a **Target Original** Data and a column as a **Target Synthetic** Data Column we can argue that Original Data and Synthetic Data have the same **properties** and **characteristics**, so they are similar, but not the same.

- In our test, we chose **“SIM_CODE”** as Target Column and the other columns as features (input). **Random Forest Accuracy** on the Original Data was **0.576** while **RF Accuracy** on the Synthetic Data was **0.0526** so we can state that Original Data and Synthetic Data have the same characteristics and properties, that is they are comparable.

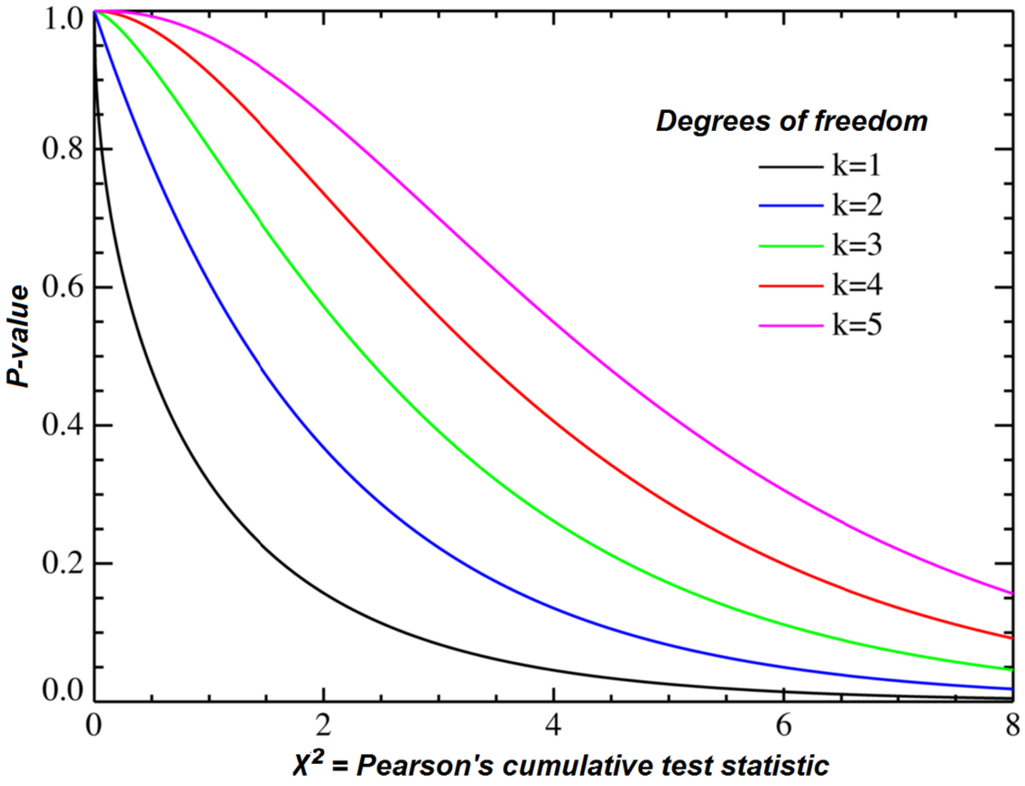


Utility Metrics - Model Evaluation via SDGym Tools: Statistical Metrics

- The **metrics** of the **Synthetic Data Gym** of the **Statistical Metrics Family** compare the tables by running different types of statistical tests on them. In the simplest scenario, these metrics compare individual columns from the real table with the corresponding column from the synthetic table.

- **sdv.metrics.tabular.CSTest:** This metric make use of the **Chi-Squared Test** to compare the distributions of two discrete columns. The output for each column is the **CSTest p-value**, which indicates the probability of the two columns having been sampled from the same distribution.

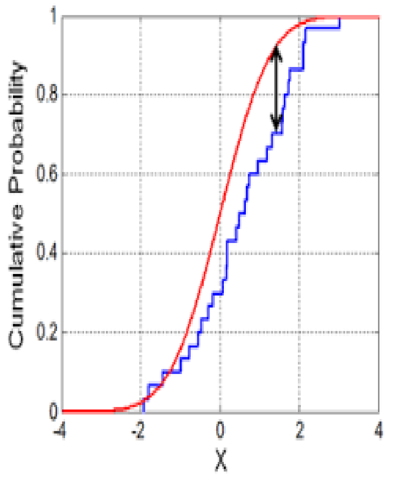
- **Chi-Squared Test p-value** must be between 0 and 1. Since we achieved **1.0** in this test, it means that our distributions (original and synthetic) are sampled from the same distribution of data.



Utility Metrics - Model Evaluation via SDGym Tools: Statistical Metrics

- **sdv.metrics.tabular.KSTest:** This metric uses the two-sample **Kolmogorov–Smirnov Test** to compare the distributions of continuous columns using the empirical **CDF (Cumulative Distibution Function: P(X≤x)**). The output of **KSTest** for each column is 1 minus the **KS Test D** statistic, which indicates the maximum distance between the expected CDF and the observed CDF values.

- The letter **"D"** stands for "distance." Geometrically, D measures the maximum vertical distance between the empirical cumulative distribution function (ECDF) of the sample and the cumulative distribution function (CDF) of the reference distribution.



- If the two samples were randomly sampled from identical populations, what is the probability that the two cumulative frequency distributions would be as far apart as observed? More precisely, what is the chance that the value of the **Komogorov-Smirnov D statistic** would be as large or larger than observed? If the P value is small, conclude that the two groups were sampled from populations with different distributions. The populations may differ in median, variability or the shape of the distribution.

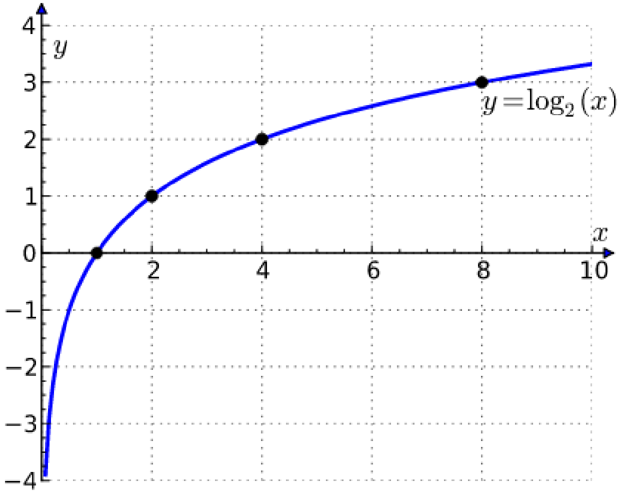
- So our result **0.95375** which is 1-D, means that the distance between CDF of original data and CDF of Synthetic Data is low, hence the two distribution are very close.

Utility Metrics - Model Evaluation via SDGym Tools: Likelihood Metrics

- The **metrics** of this family compare the tables by fitting the real data to a probabilistic model and afterwards compute the likelihood of the synthetic data belonging to the learned distribution.

- **sdv.metrics.tabular.BNLikelihood:** This metric fits a **Bayesian Network** to the real data and then evaluates the average likelihood of the rows from the synthetic data on it.

- **Bayesian Networks Likelihood** is the Error calculated on Synthetic data after fitting the model on Real Data. Very low error (likelihood) like **0.00013183** means the two datasets are very close in terms of probabilistic models.



- **sdv.metrics.tabular.BNLikelihood:** This metric fits a **Bayesian Network** to the real data and then evaluates the average likelihood of the rows from the synthetic data on it. With very low error (close to 0) according to the Log function, the output value must be negative, therefore a score like this one: **-17.415473543655498**, it is a very good result.

Privacy Metrics - Model Evaluation by Matching Common Values

- **Privacy Metric:** In this kind of metric we merge the Original Dataset and the Synthetic Dataset doing and Inner Join over 1 or 2 columns. And we get the list of rows merged. After this step, we search for all the matches between the Original Dataset and the Synthetic Dataset over 1 other column. These matches will be depicted on a bar plot via histogram. The idea behind this Privacy Metric is looking for **“values”** within the Synthetic Dataset which are also within the Original Dataset. If there are many of these values, this means that the GAN model is not able to generate a Synthetic Dataset similar to the Original one, but preserving all the values within the Original Dataset, and so preserving its privacy.

- We performed **2 Privacy Metrics Tests**: In the **Privacy Metrics Test 1** we took the field **ANTENNA_CODE** as first field for the merge and **SIM_COSE** as second feature for the final match. Vicevers, in the **Privacy Metrics Test 2** we chose **SIM_COSE** as first field and **ANTENNA_CODE** as second field.

- **Aggregated Privacy Metric (APM)**: Eventually, in order to have a final aggregate measure of privacy (APM) in the interval [0, 1] we calculated a normalized sum of all matches. More matches means less privacy, if 1 - (normalized sum) is equal to 1 means high privacy, 0 means low privacy. In our first test we achieved a value close to **0.9876543209876543,** which means very high privacy. In the second test we achieved **1.0** as aggregated value which means maximum privacy. The formula we adopted is reported in the next slide with all the results.

Privacy Metrics - Model Evaluation by Matching Common Values

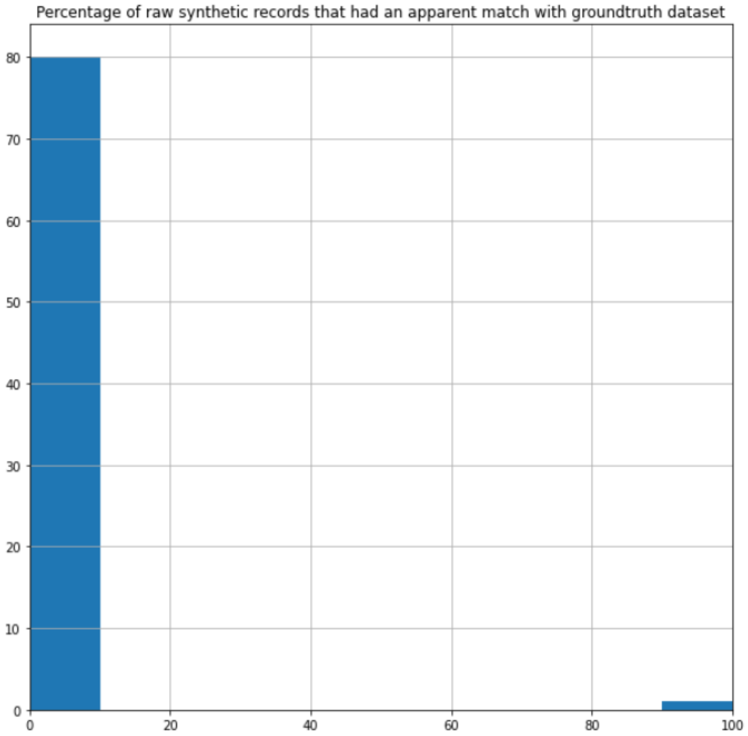
$$\mathbf{APM} = 1 - \frac{\sum_{k=0}^{n} \frac{S_k}{100}}{N}$$

The First Chart (**Test 1**) means that within the 81 rows obtained with the merge on **ANTENNA_CODE** there are 80 with 0% of matches on **SIM_CODE** and 1 with 100% of matches. So there is a small **failure** in the **Privacy Preserving Process**.

In the second chart there are no matches and failures.

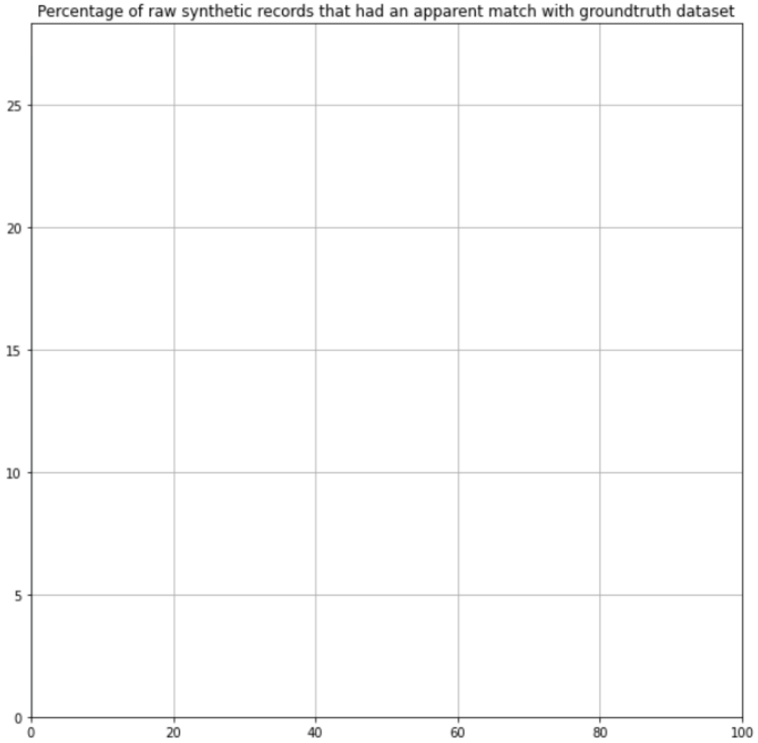
Privacy Metrics Test 1

Aggregate Data: 0.9876543209876543



Privacy Metrics Test 2

Aggregate Data: 1.0



Thank You
for your attention

FRANCESCO PUGLIESE | francesco.pugliese@istat.it

MASSIMO DE CUBELLIS | decubell@istat.it

ROBERTA RADINI | radini@istat.it

Istat Istituto Nazionale di Statistica