FEDERAL STATE AUTONOMOUS EDUCATIONAL INSTITUTE
OF HIGHER EDUCATION
ITMO UNIVERSITY

Report on learning practice # 3

Sampling of multivariate random variables

Performed by:

Igor Vernyy, j4134c

Kirill Mukhin, j4134c

Alexander Petrov, j4134c

Bogdan Chertkov, j4132c

Saint-Petersburg

2022

Substantiation of chosen sampling

We have chosen “total cases per million”, “total deaths per million” and “reproduction rate” as target variables because it has a lot of meaning to predict these values in real world problems. “positive rate”, “new people vaccinated smoothed per hundred”, “DGP per capita”, “population density”, “life expectancy”, “extreme poverty” and “hospital beds per thousand” are predictors.

Sampling of chosen target variables using two methods

The first method is a so called “inverse transform sampling”. Figures 1-3 contain plots that describe results obtained by this approach for each target variable. As you can see distributions are nearly identical which is confirmed by QQ biplot.

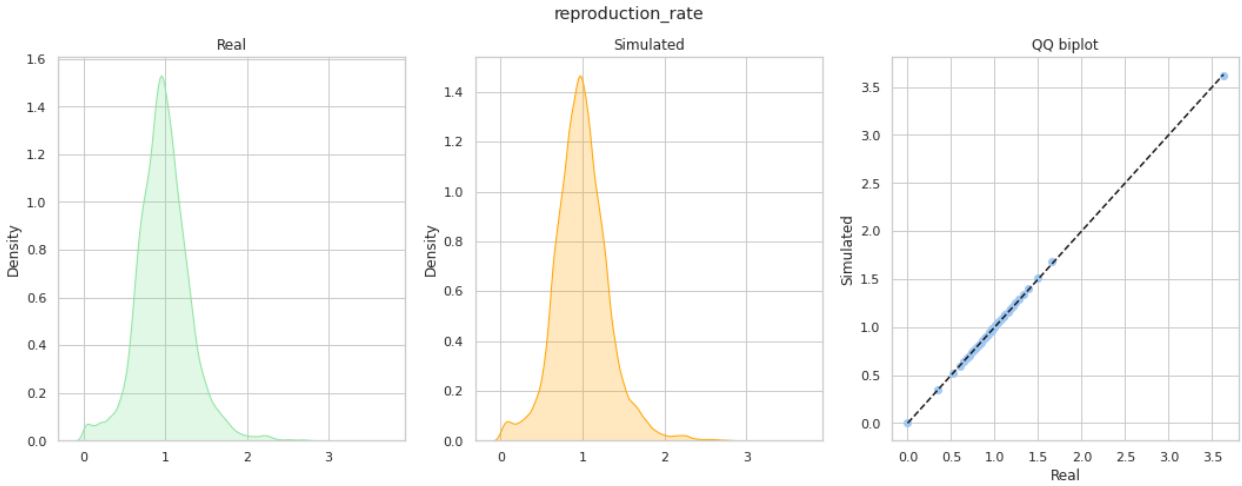

Figure 1 – Comparison of distributions obtained by inverse transform sampling for “reproduction rate” variable

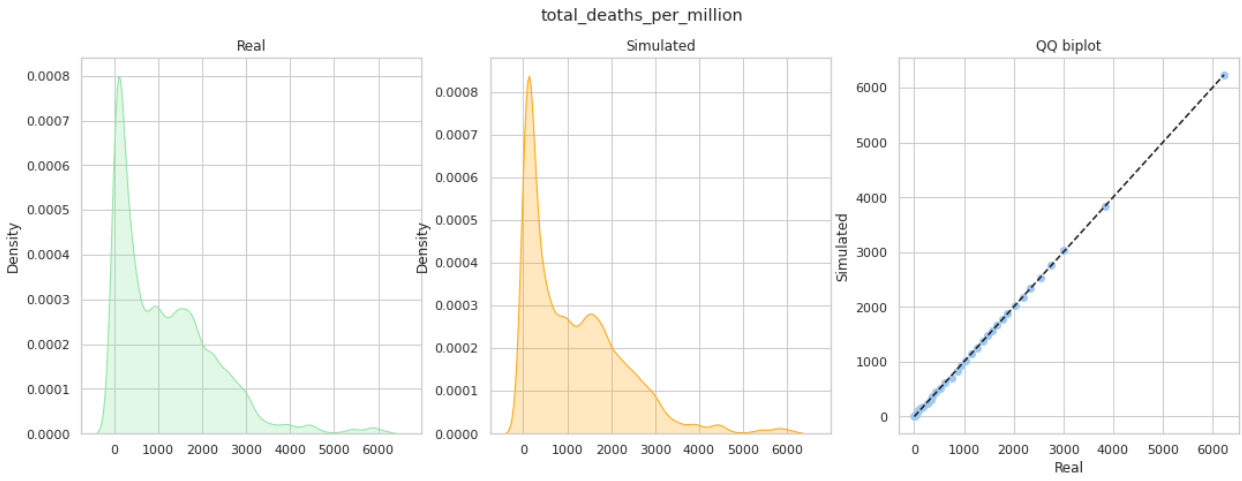

Figure 2 – Comparison of distributions obtained by inverse transform sampling for “total deaths per million” variable

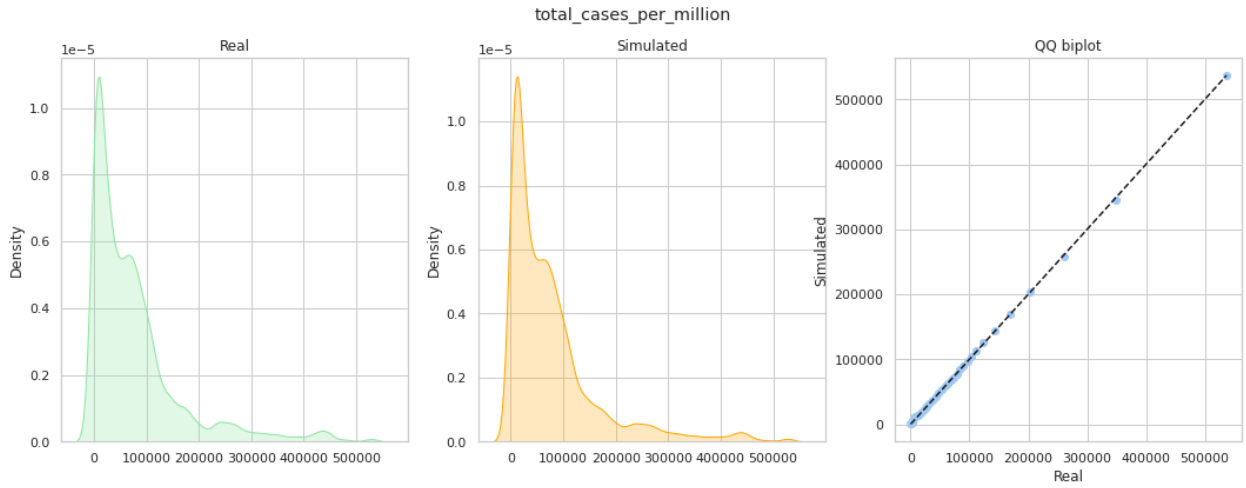

Figure 3 – Comparison of distributions obtained by inverse transform sampling for "total cases per million" variable

The second method is a geometric sampling. Figures 4-6 depicts these distributions. Results are very similar to inverse transform sampling.

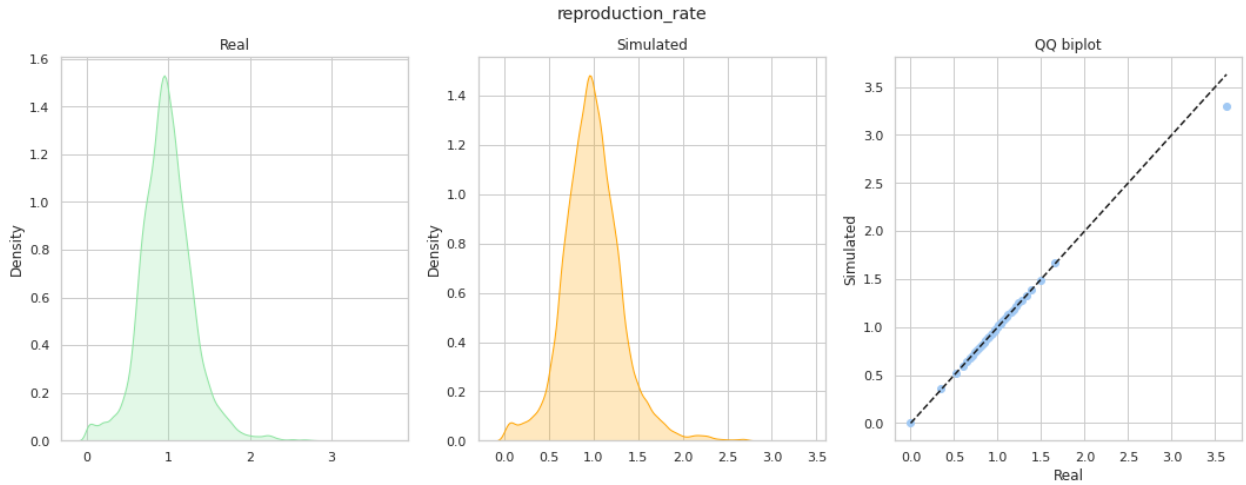

Figure 4 – Comparison of distributions obtained by geometric sampling for "reproduction rate" variable

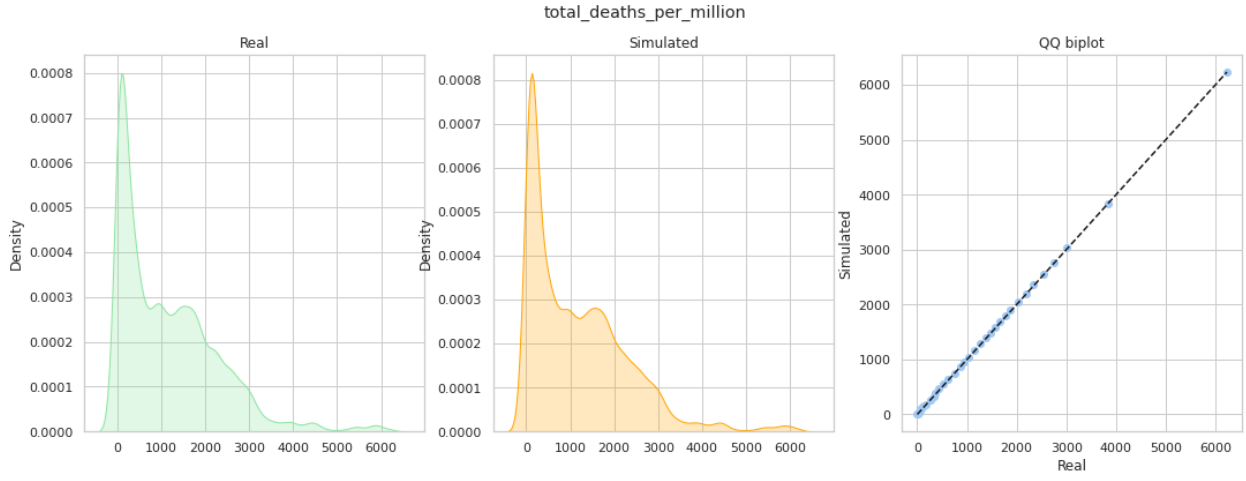

Figure 5 – Comparison of distributions obtained by geometric sampling for "total deaths per million" variable

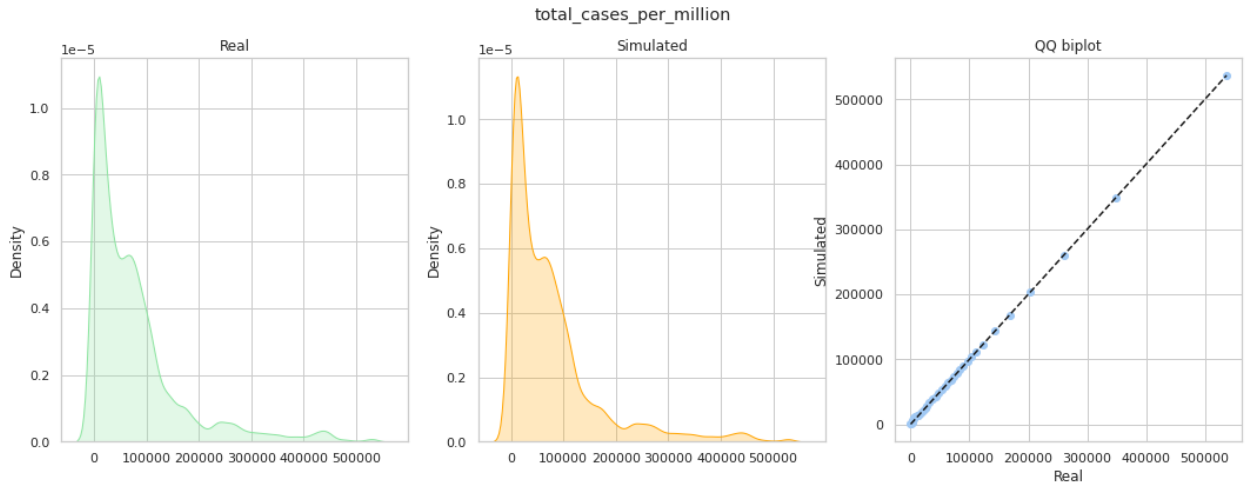

Figure 6 – Comparison of distributions obtained by geometric sampling for “total cases per million” variable

Estimation of relations between predictors and chosen target variables

To estimate relations between predictors and target variables we have to build correlation matrix (Figure 7).

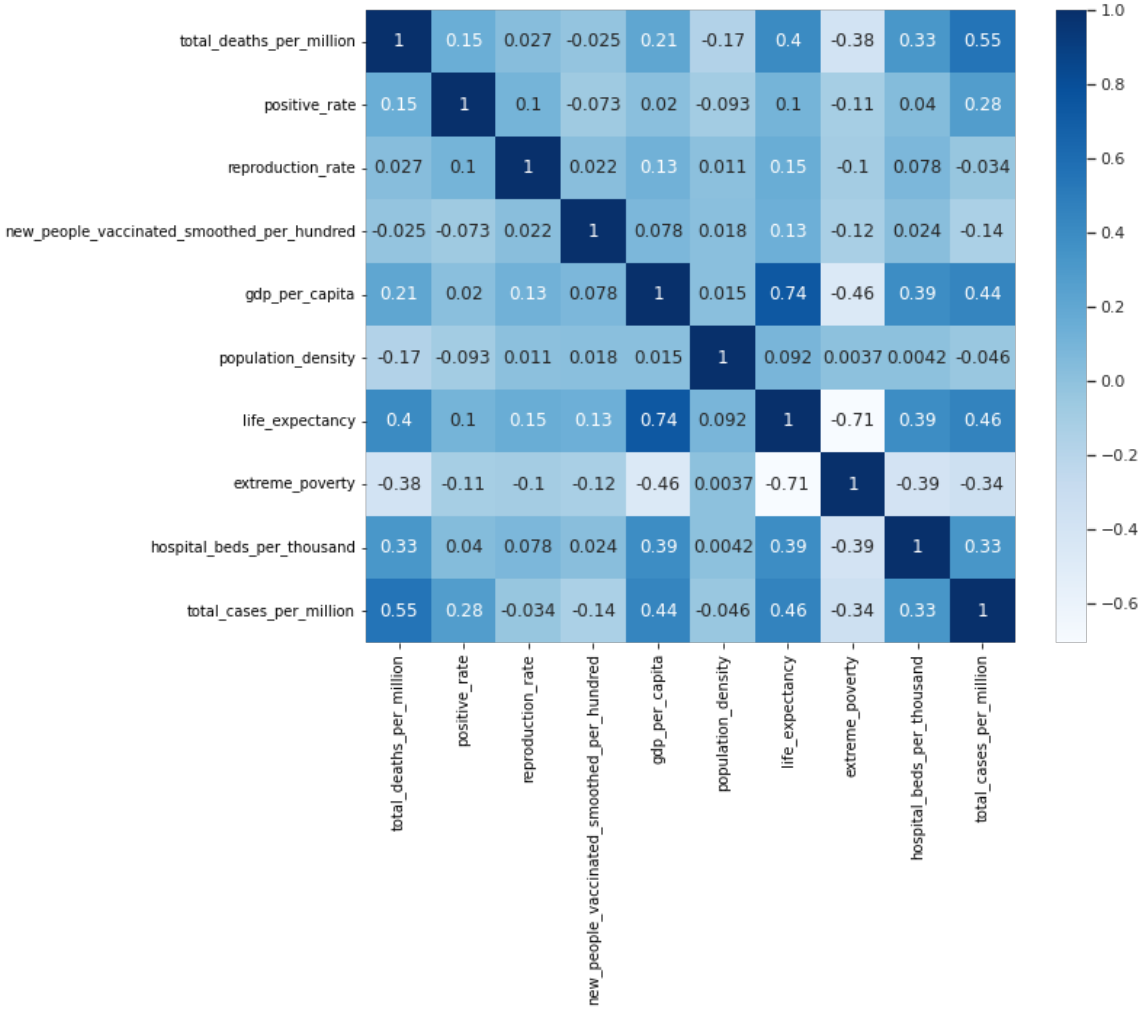

Figure 7 – Correlation matrix for chosen variables

Now it is required to drop out weakly dependent pairs (let threshold be 0.2, so correlation below 0.2 is negligible). The dependencies are:

$$dgp\_per\_capita \rightarrow total\_deaths\_per\_million$$

$$life\_expectancy \rightarrow total\_deaths\_per\_million, \quad gdp\_per\_capita$$

$$extreme\_poverty \rightarrow total\_deaths\_per\_million, \quad gdp\_per\_capita, \quad life\_expectancy$$

$$hospital\_beds\_per\_thousand \rightarrow total\_deaths\_per\_million, \quad dgp\_per\_capita, \quad life\_expectancy, \quad extreme\_poverty$$

$$total\_cases\_per\_million \rightarrow total\_deaths\_per\_million, \quad positive\_rate, \quad dgp\_per\_capita, \quad life\_expectancy, \quad extreme\_poverty, \quad hospital\_beds\_per\_thousand$$

Bayesian network

Based on known dependencies between variables, we have built Bayesian network with the structure shown at Figure 8. This model was fitted by Maximum Likelihood Estimator.

Then the models were built and trained using Hill Climb and Tree Search algorithms (Figures 9-10).

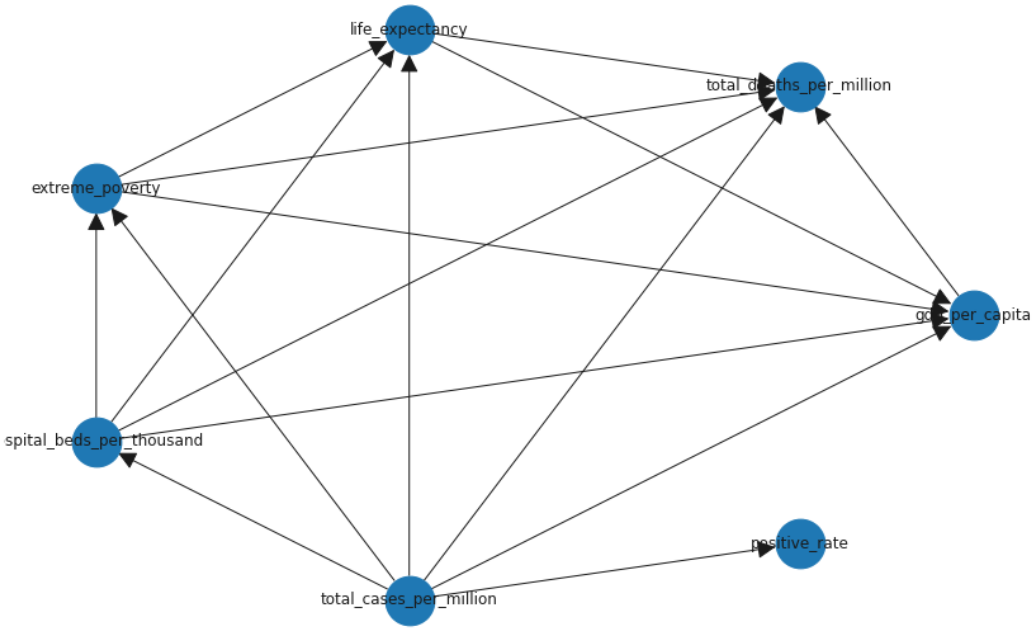

Figure 8 – Manually-created architecture of Bayesian network

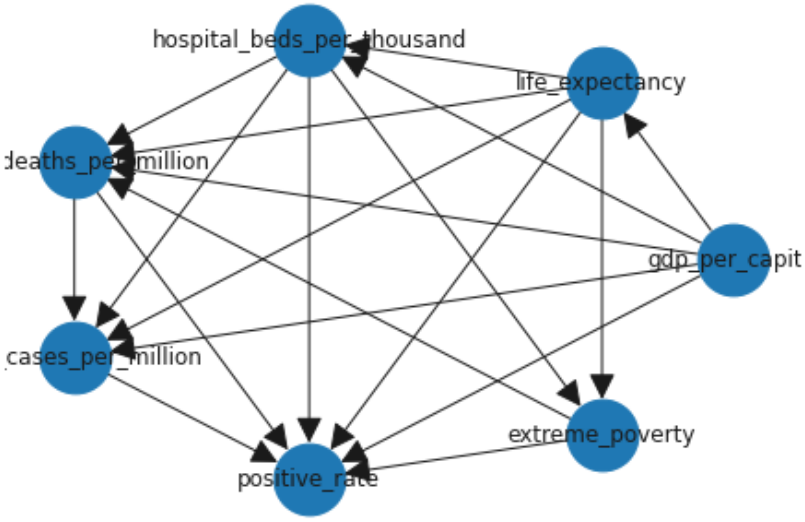

Figure 9 – Hill Climb learned architecture of Bayesian network

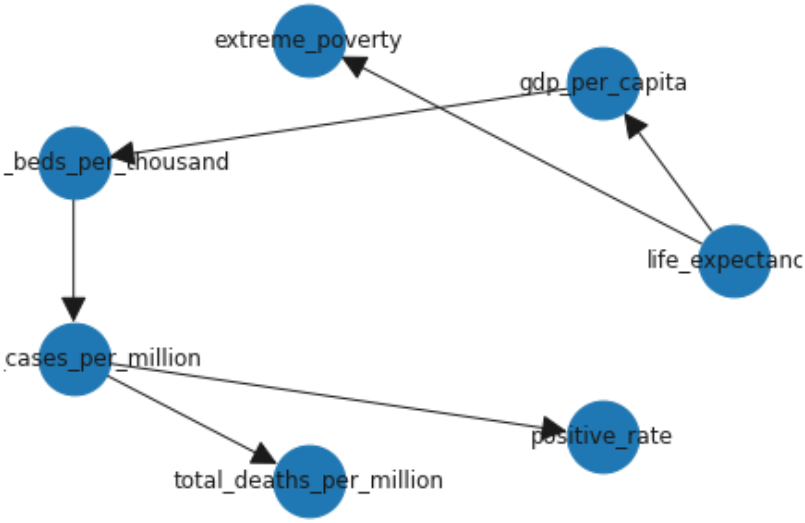

Figure 10 – Tree Search learned architecture of Bayesian network

Quality analysis

In order to compare the quality of the generated sequences, a diagram of the distribution of each model was constructed in comparison with real data. You can see these plots at figures 11 and 12. Quantity quality analysis was conducted for discretized data. Table 1 contains accuracy metrics for target variables and different BN. Table 2 contains RMSE for non-discrete data.

Random variable	BN	Accuracy
Total deaths per million	Manually-created	0.22
	Hill Climb (K2)	0.22
	Tree Search	0.22
Total cases per million	Manually-created	0.26
	Hill Climb (K2)	0.26
	Tree Search	0.27

Table 1 – Quantity quality analysis for discretized data

Random variable	BN	RMSE
Total deaths per million	Manually-created	1621
	Hill Climb (K2)	3785
	Tree Search	1814
Total cases per million	Manually-created	118806
	Hill Climb (K2)	117967
	Tree Search	144053

Table 2 – Quantity quality analysis

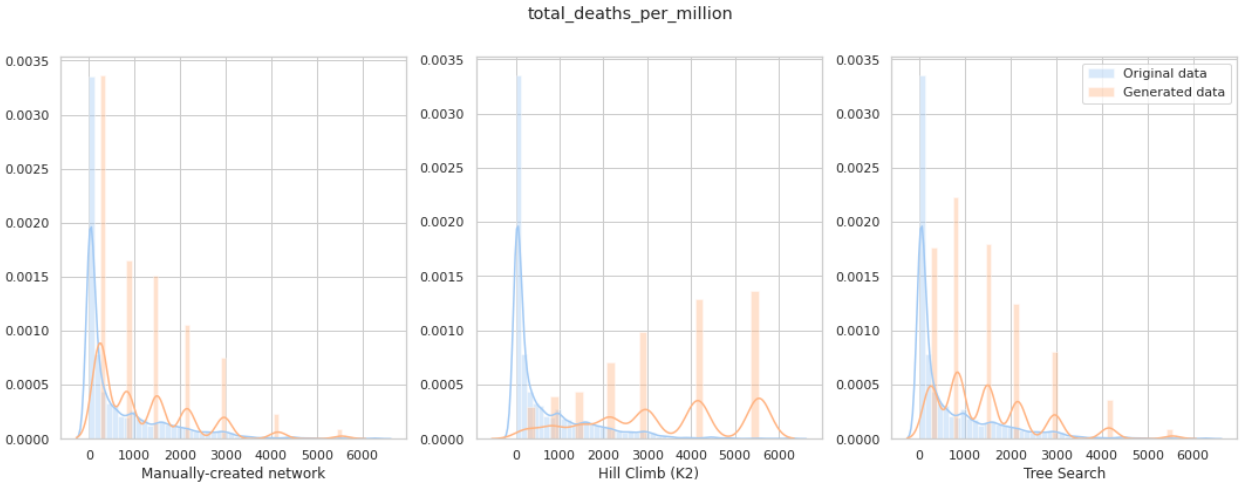

Figure 11 – Comparison distributions of generated data by different models for “total deaths per million” variable

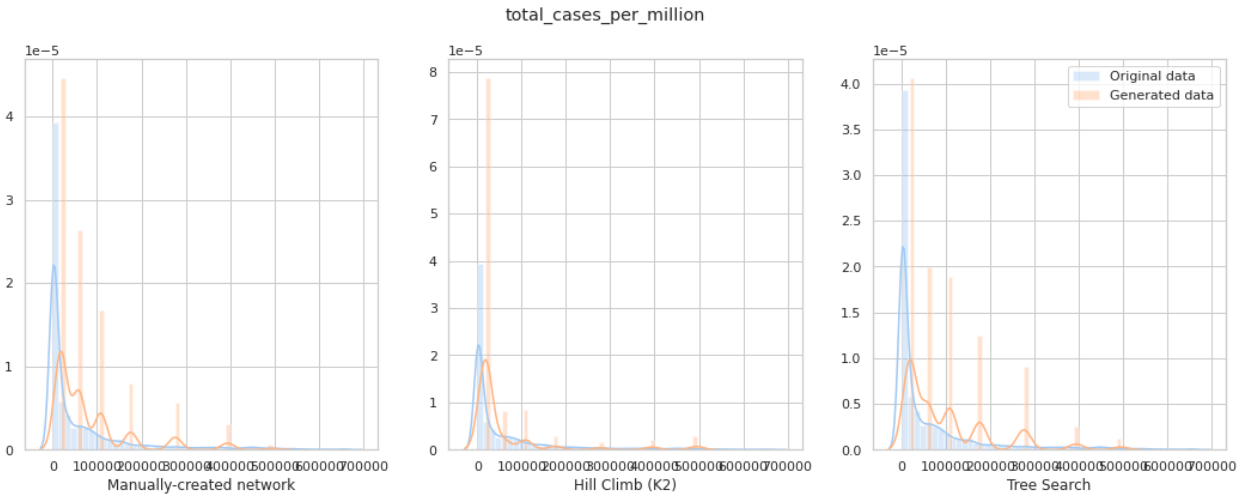

Figure 12 – Comparison distributions of generated data by different models for “total cases per million” variable

Appendix

https://github.com/D3lph1/methods-and-models-for-multivariate-data-analysis/blob/master/Lab%203/lab_3_notebook.ipynb