FEDERAL STATE AUTONOMOUS EDUCATIONAL INSTITUTION
OF HIGHER EDUCATION
ITMO UNIVERSITY

Report on learning practice #3

Sampling of multivariate random variables

Performed by
Dmitry Grigorev,
Eugenia Khomenko,
Efim Podkovirkin,
Arina Syrchenko

St. Petersburg
2022

Contents

1. Data description

Let D be the modified dataset on Narvik roads. The features here are:

- lat_ — latitude
- lon_ — longitude
- State_ — word description of road state (1: 'dry', 2: 'moist', 3: 'wet', 4: 'icy', 5: 'snowy', 6: 'slushy')
- Ta_mean,Ta_min,Ta_max — atmosphere temperature
- Tsurf_mean,Tsurf_min,Tsurf_max — surface temperature
- Water_mean,Water_min,Water_max — water layerw width (0 – 3 mm)
- Speed_mean,Speed_min,Speed_max — wind speed (in knots, 5 $knots \approx 9.3\ km/h$)
- Height_mean,Height_min,Height_max — height of location above mean sea level
- Tdew_mean,Tdew_min,Tdew_max — dew point ($Celsius$)
- Friction_mean,Friction_min,Friction_max — friction value (0 – 1, 0 means no friction)
- Date,Time, date_time, FullDate — time and date
- Direction_min,Direction_max — wind direction ($degrees$)
- ClosestCity, location
- maxtempC,mintempC — day maximum and minimum of temperature ($Celsius$)
- totalSnow_cm — total snowfall (cm)
- sunHour — passed sun energy in $Sun-Hours$ (A $Sun-Hour$ is "1000 watts of energy shining on 1 square meter of surface for 1 hour")
- uvIndex — ultraviolet index
- moon_illumination — moon phase ($percents$)
- moonrise — time of Moon rise
- moonset — time of Moon set

- sunrise — time of Sun rise
- sunset — time of Sun set
- DewPointC — hourly dew point measurement ($Celsius$)
- FeelsLikeC — hourly Feels-like temperature ($Celsius$)
- HeatIndexC — hourly heat index ($Celsius$)
- WindChillC — hourly wind-chill index ($Celcius$)
- WindGustKmph — hourly wind gust measure (km/h)
- cloudcover — hourly cloud cover index ($percents$)
- humidity — hourly humidity ($percents$)
- precipMM — hourly precipitation (mm)
- pressure — hourly atmosphere pressure ($mbar$)
- tempC — hourly atmosphere temperature ($Celsius$)
- visibility — hourly visibility (0–10, 0 means poor visibility)
- winddirDegree — hourly wind direction ($degrees$)
- windspeedKmph — hourly wind speed (km/h)

2. Substantiation of chosen sample

In this lab work Friction_mean, windspeedKmph and FeelsLikeC are chosen as targets and the predictors are humidity, State_, pressure, Height_mean, totalSnow_cm, Water_mean, moon_illumination. Further we test the significance of dependence between chosen targets and predictors. We have chosen 1000 observations as a subsample.

3. Additional research on appropriate distribution

As a result of Lab #1 it was obtained that

- the most appropriate distribution for FeelsLikeC is normal with $\mu \approx -8.867$ and $\sigma \approx 5.874$ (it minimizes least squares and maximizes log-likelihood);

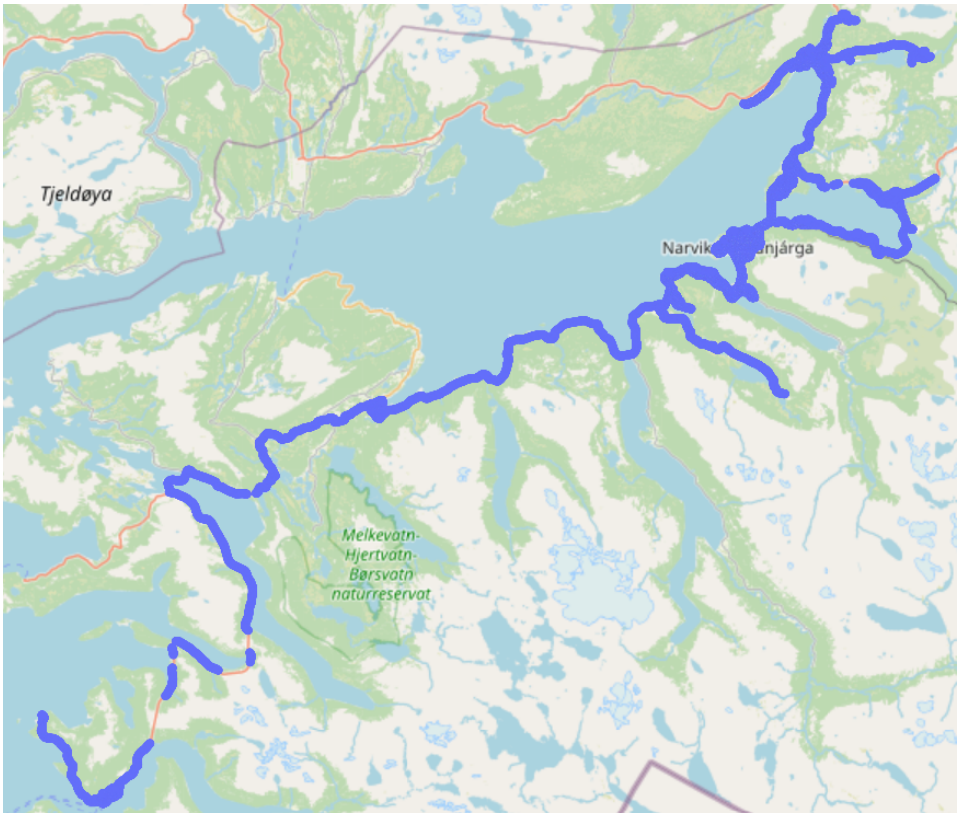

Figure 1: Geography of the data

- the most appropriate distribution for windspeedKmph is Cauchy's with $loc \approx 11.637$ and $scale \approx 2.830$ (it minimizes least squares and maximizes log-likelihood).

Feature Friction_mean was under focus in Lab #2 where linear regression model for this feature has been built. Its distribution is bimodal (see fig. 2) so any simple continuous distribution would not fit here for estimation. As an example, it is proposed to find the mixture of two normal distributions $\mathcal{P} = (1-\alpha)N(\mu_1, \sigma_1^2) + \alpha N(\mu_2, \sigma_2^2)$.

To find the unknown parameters we have applied EM algorithm which we implemented for the two components normal distribution mixture separation (see [3]). As a result, we obtained the following parameters:

$$\alpha \approx 0.463,$$

$$\mu_1 \approx 0.353,$$

$$\sigma_1 \approx 0.065,$$

$$\mu_2 \approx 0.694,$$

$$\sigma_2 \approx 0.063,$$

and the corresponding distribution is illustrated in figure 3.

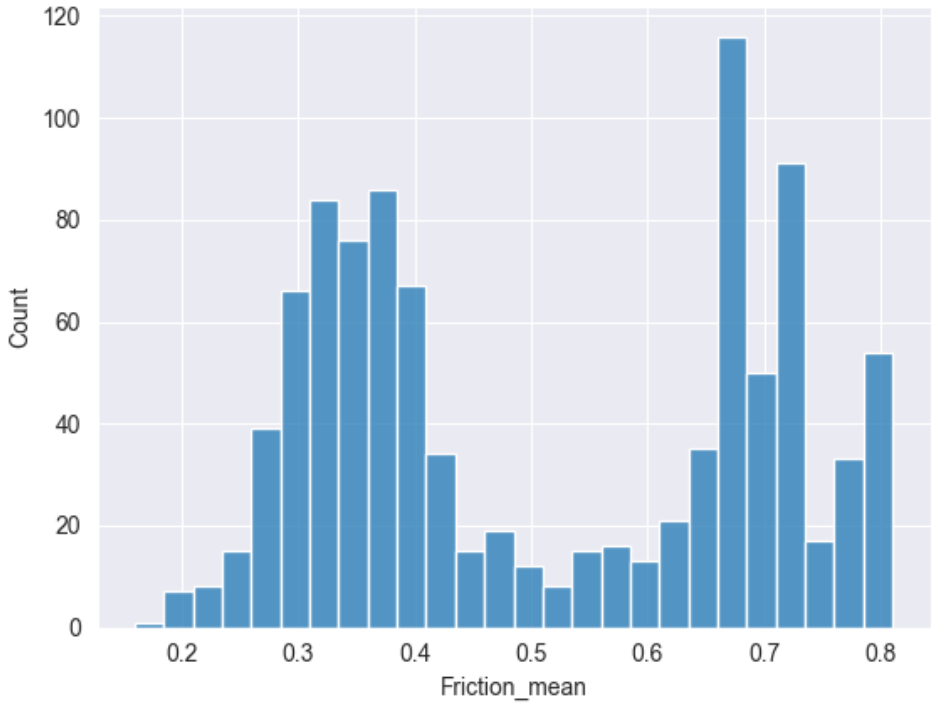

Figure 2: Histogram of Friction_mean

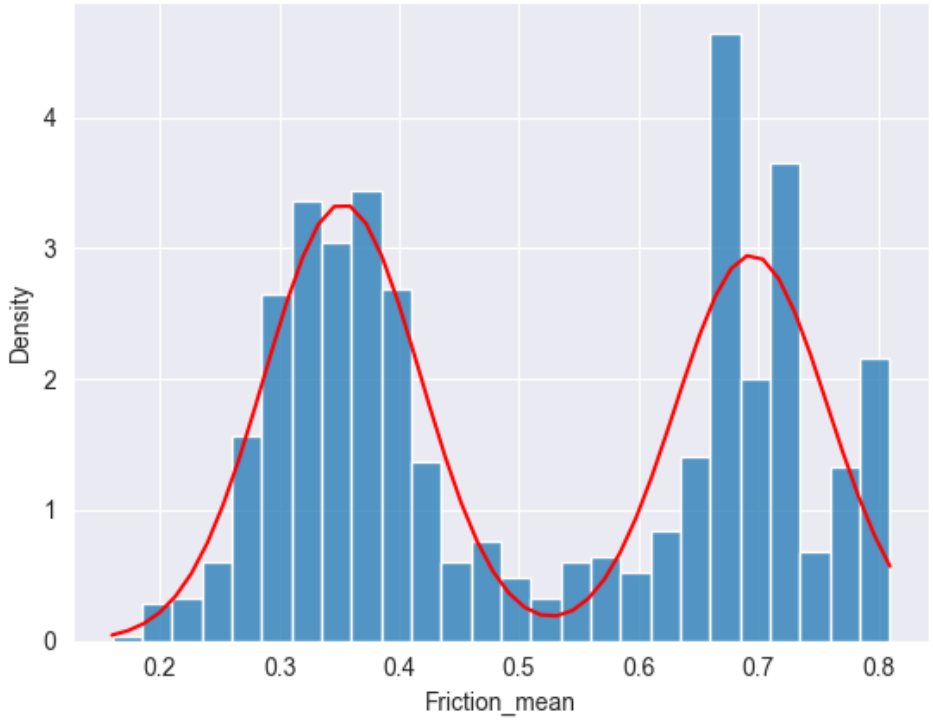

Figure 3: The found mixture of two gaussians which fits the data of Friction_mean

As one can see from the figure 3, the obtained mixture fits the data quite well. The QQ-plot of data vs the mixture in picture 4 also allows us to conclude the same. However, Kolmogorov-Smirnov test under significance level $\widetilde{\alpha} = 0.05$ with p-value≈ 0.009 rejects the hypothesis that our sample came from the distribution described by the mixture. That is why we do not expect that a synthesized

sample will follow the same distribution as Friction_mean.

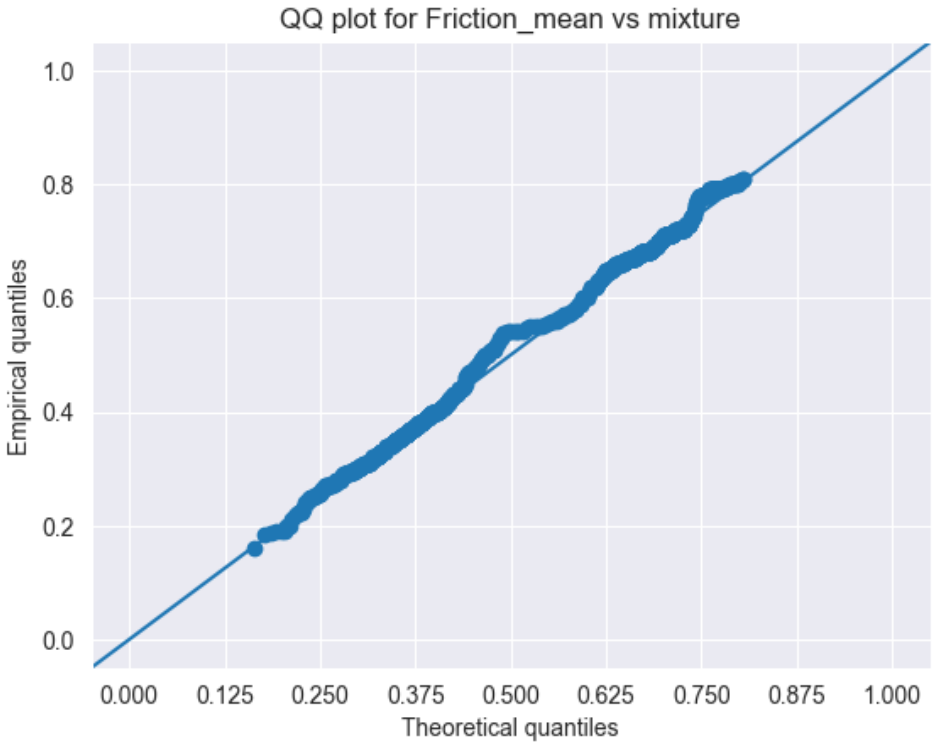

Figure 4: QQ-plot for Friction_mean versus the found mixture

4. Sampling of chosen target variables using univariate parametric distributions (from practice #1) with 2 different sampling methods

Friction_mean

To generate a sample for Friction_mean by means of the fitted distribution, we used Box-Muller algorithm which generates normal distribution given a mixture component and the mixture component number probability is determined by α, and acceptance-rejection sampling (ARS) with bins which inner distribution is uniform across a bin and probability is proportional to the bin area.

The idea behind Box-Muller algorithm is as follows: given i.i.d $\alpha_1, \alpha_2 \sim U(0,1)$, the variables

$$\xi_1 = \sqrt{-2\ln(\alpha_1)}\sin(2\pi\alpha_2), \xi_2 = \sqrt{-2\ln(\alpha_1)}\cos(2\pi\alpha_2)$$

are i.i.d. and $\xi_i \sim N(0,1)$. The number of mixture component is a random variable with values 1 and 2 and probabilities $(1-\alpha)$ and α correspondingly. To obtain $\xi \sim N(\mu, \sigma^2)$ from $\eta \sim N(0,1)$ one has to perform linear transformation: $\xi \leftarrow \sigma\eta + \mu$.

As soon as a sample of 1000 elements was obtained from the found distribution by the Box-Muller algorithm, two sample Kolmogorov-Smirnov test was performed under significance level $\widetilde{\alpha} = 0.05$. Its

p-value≈ 0.164 so there is no evidence to reject the hypothesis that the synthesized sample came from the same distribution as our sample. Two figures in fig. 6 demonstrate to what extent the generated sample is similar to the theoretical distribution and the empirical distribution of Friction_mean.

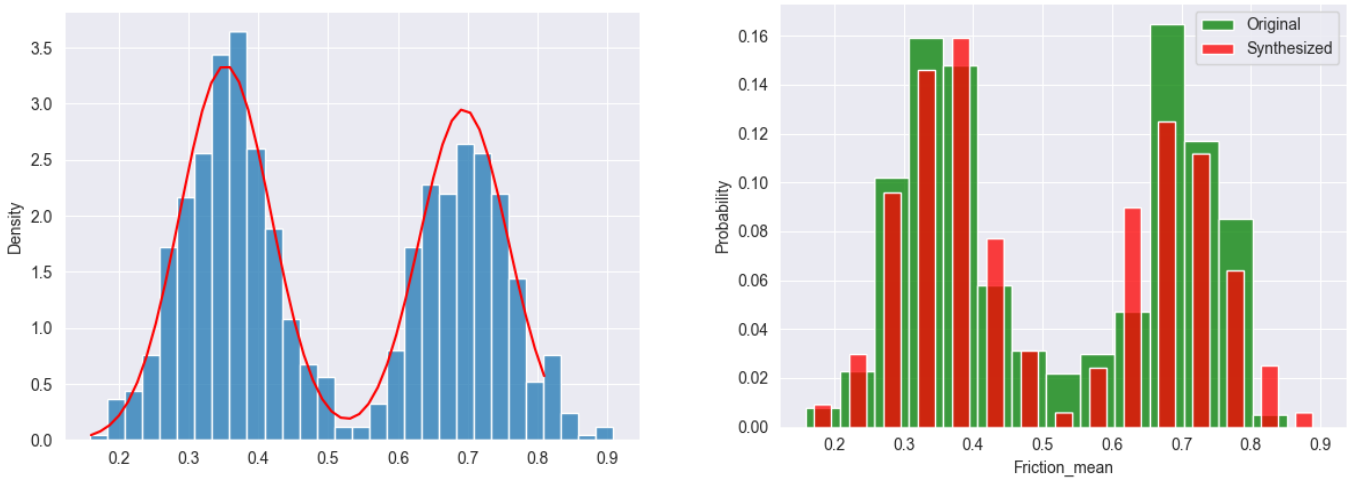

Figure 5: The histogram of the generated sample vs the mixture's density and the histogram of observed Friction_mean

As for the second algorithm of sampling, we performed partition of the range of Friction_mean variable into 6 bins and built piece-wise constant majorant depicted in figure 7. It is worth mentioning that we restricted the mixture on the Friction_mean range interval with the corresponding normalization so that the integral of the density is equal to 1. Further we use this majorant in ARS to make sampling efficient. Each bin determines its own rectangle in which ARS is performed and its own sampling distribution which is uniform by our choice. The sampling scheme is presented in [4].

As a result of sampling by the aforementioned method, we obtained a sample of 1000 observations and two sample Kolmogorov-Smirnov test under significance level $\widetilde{\alpha} = 0.05$ executed $p - value \approx 0.006$ rejects the hypothesis that the real data sample and the synthesized sample came from the same distribution. Two figures in fig. 7 demonstrate to what extent the generated sample is similar to the theoretical distribution and the empirical distribution of Friction_mean.

FeelsLikeC

In Lab #1 it was obtained that the appropriate distribution for FeelsLikeC is normal $N(\mu, \sigma^2)$ with $\mu \approx -8.867$, $\sigma \approx 5.873$. To simulate the normal distribution we used Box-Muller algorithm and ARS with sampling from Laplace distribution.

As soon as a sample of 1000 elements was obtained from the found distribution by the Box-Muller algorithm, two sample Kolmogorov-Smirnov test was performed under significance level $\widetilde{\alpha} = 0.05$.

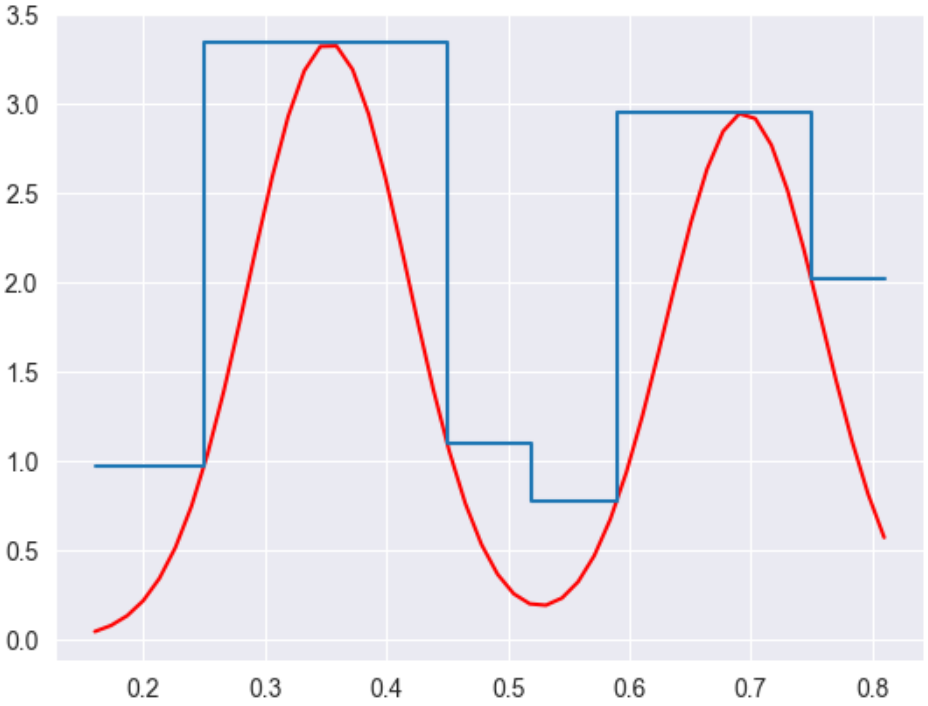

Figure 6: The majorant of the mixture's density

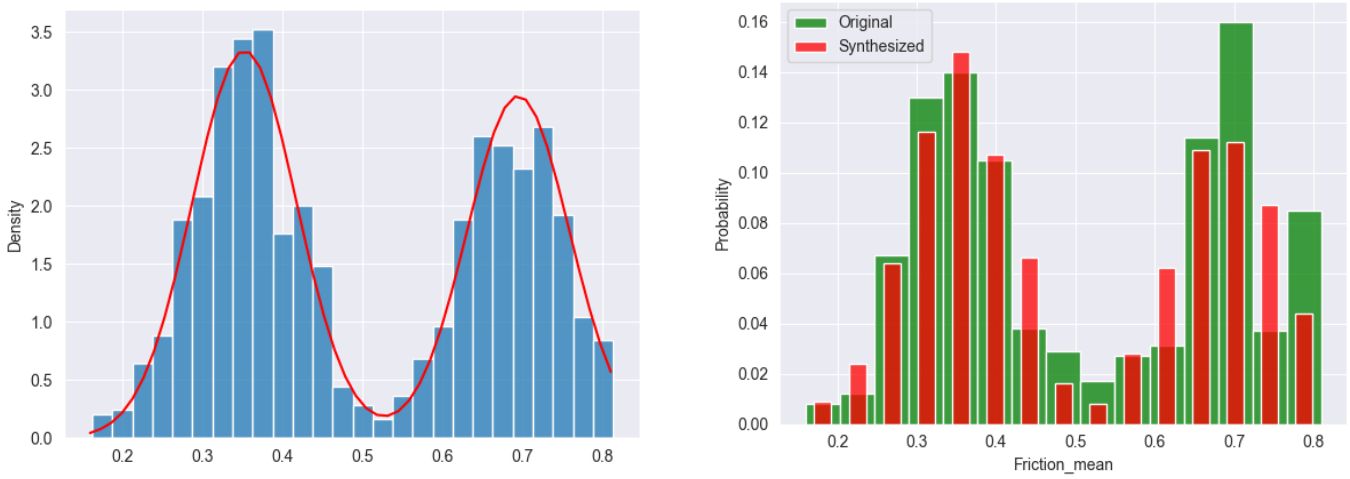

Figure 7: The histogram of the ARS-generated sample vs the mixture's density and the histogram of observed Friction_mean

Its p-value$\approx$ 0.00028 so it rejects the hypothesis that the synthesized sample came from the same distribution as our sample. Two figures in fig. 12 demonstrate to what extent the generated sample is similar to the theoretical distribution and the empirical distribution of FeelsLikeC.

As for the second sampling algorithm, we used ARS using Laplace standard distribution to simulate normal one. Two sample Kolmogorov-Smirnov test was performed under significance level $\widetilde{\alpha} = 0.05$. Its p-value$\approx 4.4 \cdot 10^{-6}$ so it rejects the hypothesis that the synthesized sample came from the same distribution as our sample. Two figures in fig. 13 demonstrate to what extent the generated

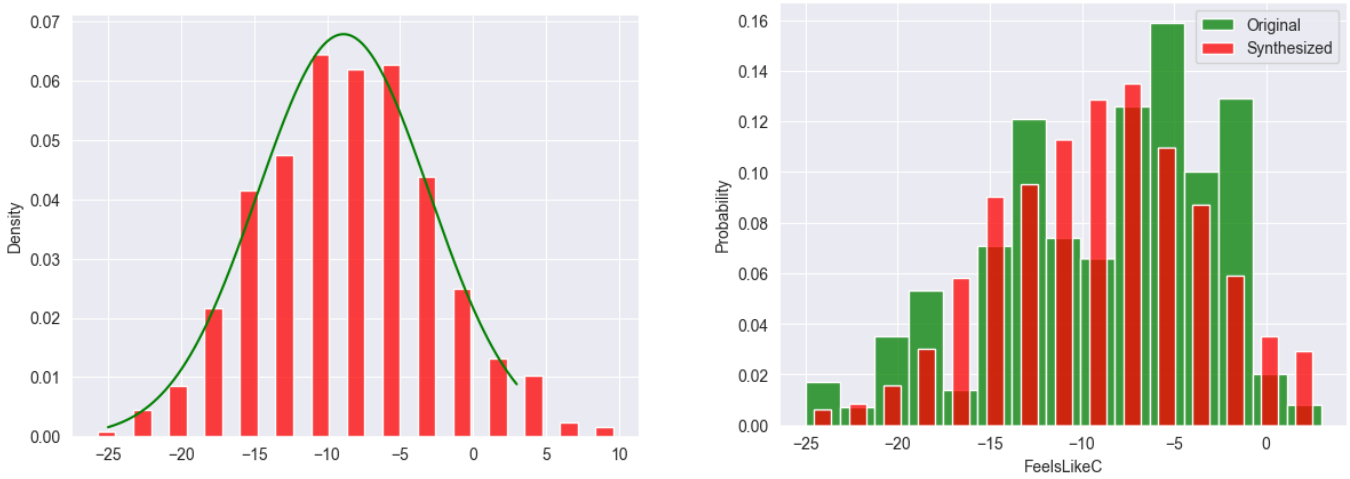

Figure 8: The histogram of the Box-Muller-generated sample vs the reference density and the histogram of observed FeelsLikeC.

sample is similar to the theoretical distribution and the empirical distribution of FeelsLikeC.

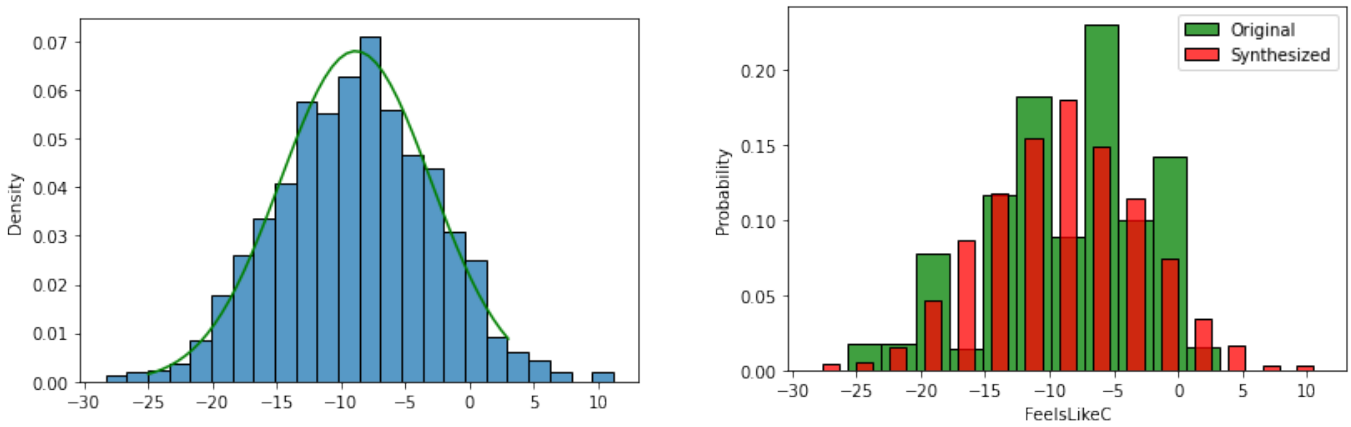

Figure 9: The histogram of the ARS-generated sample vs the reference density and the histogram of observed FeelsLikeC.

windspeedKmph

In Lab #1 it was obtained that the appropriate distribution for windspeedKmph is $Cauchy(loc, scale)$ with $loc \approx 11.637$, $scale \approx 2.830$. The only correction here is that we modified the distribution so that it is restricted on the range of the variable (at least, its values are non-negative). To simulate this corrected distribution we used both inverse transform sampling (ITS) and ARS using uniform distribution.

As soon as a sample of 1000 elements was obtained from the found distribution by the ITS algorithm, two sample Kolmogorov-Smirnov test was performed under significance level $\widetilde{\alpha} = 0.05$. Its p-value$\approx$ 0.001 so it rejects the hypothesis that the synthesized sample came from the same

distribution as our sample. Two figures in fig. 14 demonstrate to what extent the generated sample is similar to the theoretical distribution and the empirical distribution of windspeedKmph.

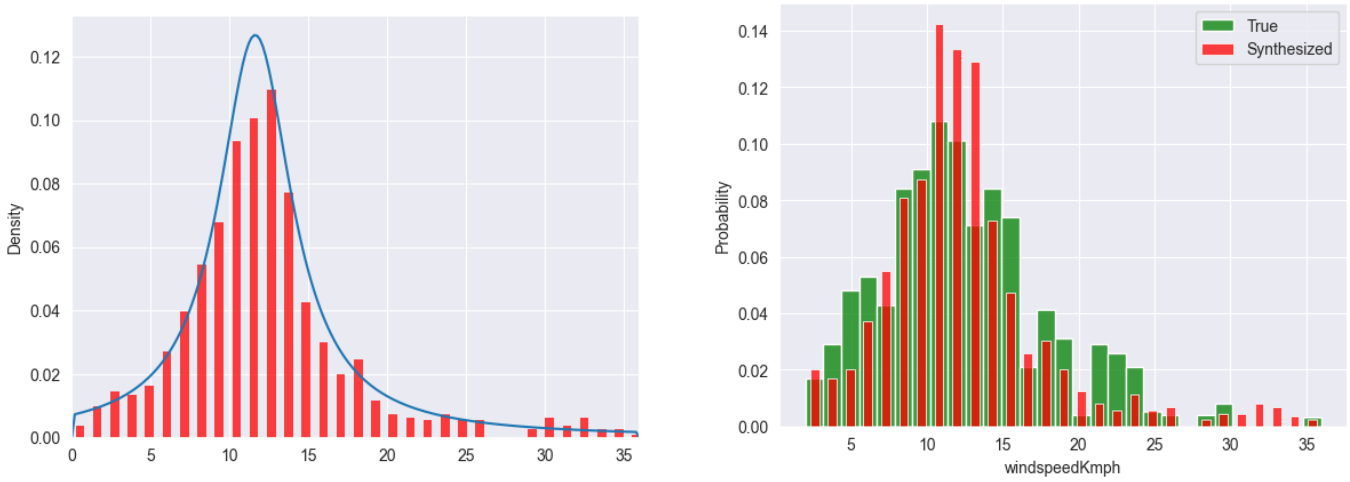

Figure 10: The histogram of the ITS-generated sample vs the reference density and the histogram of observed windspeedKmph.

As for the second sampling algorithm, we used ARS using uniform distribution over the range of the variable to simulate normal one. Two sample Kolmogorov-Smirnov test was performed under significance level $\widetilde{\alpha} = 0.05$. Its p-value$\approx$ 0.0004 so it rejects the hypothesis that the synthesized sample came from the same distribution as our sample. Two figures in fig. 15 demonstrate to what extent the generated sample is similar to the theoretical distribution and the empirical distribution of FeelsLikeC.

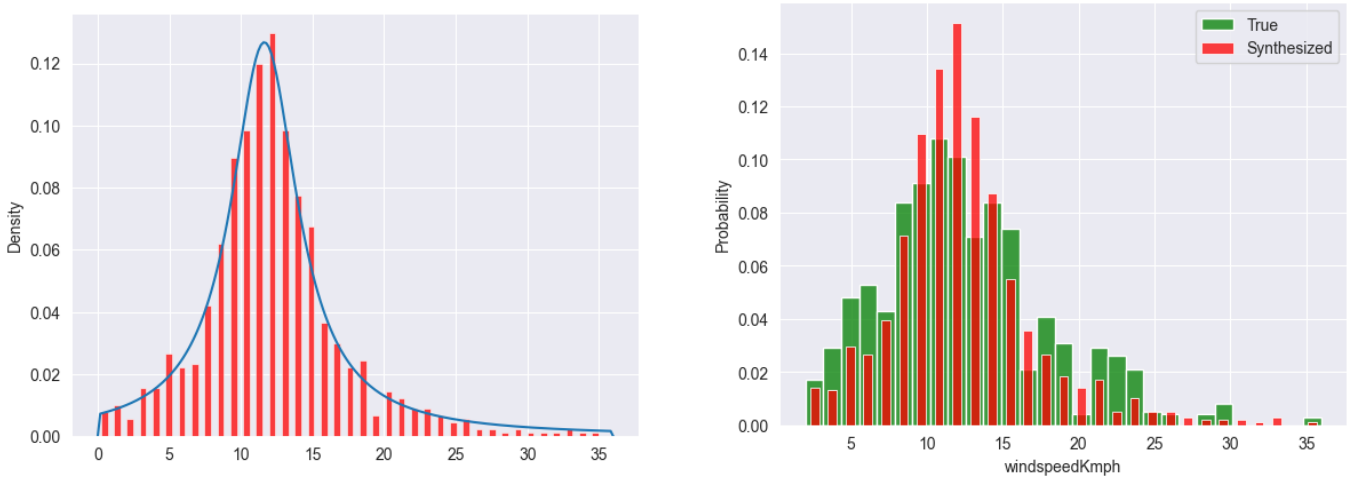

Figure 11: The histogram of the ARS-generated sample vs the reference density and the histogram of observed windspeedKmph.

5. Estimation of relations between predictors and chosen target variables

In this section we study significance of relations between targets and chosen predictors. As was obtained in Lab #2, variable State_ is able to determine bimodality in the distribution of Friction_mean so here we conclude that the former variable is worth using in our future study as a predictor.

As for other variables, we performed correlation analysis to find the variables among the predictors that significantly correlate with at least one target. The results of the correlation significance test for windspeedKmph, FeelsLikeC and Friction_mean are provided in table 1, 2 and 3 correspondingly. As one can see, each of the predictors has significant correlation with at least one target. So we further use the variables moon_illumination, Water_mean, Height_mean, pressure, State_, totalSnow_cm, humidity as predictors.

windspeedKmph	$\widehat{\rho}_n$	left	right	p-value
humidity	0.076246	0.014321	0.137589	1.588231e-02
pressure	-0.315770	-0.370511	-0.258844	1.366632e-24
Height_mean	-0.033959	-0.095751	0.028093	2.833401e-01
totalSnow_cm	0.314295	0.257316	0.369097	2.297680e-24
Water_mean	-0.030320	-0.092140	0.031732	3.381442e-01
moon_illumination	0.217900	0.158042	0.276163	3.264125e-12

Table 1: Correlation significance of windspeedKmph vs predictors test results. Predictors highlighted by pale blue have significant correlation with the variable.

FeelsLikeC	$\widehat{\rho}_n$	left	right	p-value
humidity	0.153972	0.092865	0.213923	9.974678e-07
pressure	-0.079740	-0.141036	-0.017835	1.165389e-02
Height_mean	0.122230	0.060697	0.182837	1.066483e-04
totalSnow_cm	0.032002	-0.030050	0.093809	3.120183e-01
Water_mean	0.397541	0.344027	0.448481	3.321284e-39
moon_illumination	0.110915	0.049260	0.171727	4.415945e-04

Table 2: Correlation significance of FeelsLikeC vs predictors test results. Predictors highlighted by pale blue have significant correlation with the variable.

Friction_mean	$\widehat{\rho}_n$	left	right	p-value
humidity	-0.027756	-0.089595	0.034296	3.806060e-01
pressure	-0.281915	-0.338001	-0.223834	1.001731e-19
Height_mean	-0.102377	-0.163333	-0.040642	1.187311e-03
totalSnow_cm	-0.087782	-0.148965	-0.025930	5.472678e-03
Water_mean	0.071716	0.009767	0.133118	2.333191e-02
moon_illumination	0.125759	0.064267	0.186300	6.671204e-05

Table 3: Correlation significance of Friction_mean vs predictors test results. Predictors highlighted by pale blue have significant correlation with the variable.

6. Bayesian network

To build and fit Bayesian networks (BN) in this lab we use BAMT toolkit [1]. Firstly, we determined the types of targets and predictors. The variables State_, FeelsLikeC and humidity are set to be discrete. The latter should be discrete since often humidity is measured in integers. The second has been set as discrete since the number of unique states of this variable is one of the lowest. The others are continuous.

6.1. BN built on the multivariate analysis results

By the usage of the multivariate analysis results provided in the tables 1, 2 and 3 and some physics reasoning, we have built BN which structure is presented in figure 12. The usage of the mixture of gaussian distributions is allowed. Since the built network has inner discrete nodes, logits are allowed.

We performed sampling by means of the constructed BN (10000 synthetic observations) and then did two sample statistical tests, i.e. Kolmogorov-Smirnov and chi-square [2], under significance level $\widetilde{\alpha} = 0.05$.

- FeelsLikeC: chi-square test rejects the hypothesis that synthetic and original data came from the same distribution with p-value ≈ 0.00060;
- windspeedKmph: Kolmogorov-Smirnov test rejects the analogous hypothesis with p-value $\approx 3 \cdot 10^{-10}$;
- Friction_mean: Kolmogorov-Smirnov test rejects the analogous hypothesis with p-value $\approx 1.2 \cdot 10^{-10}$.

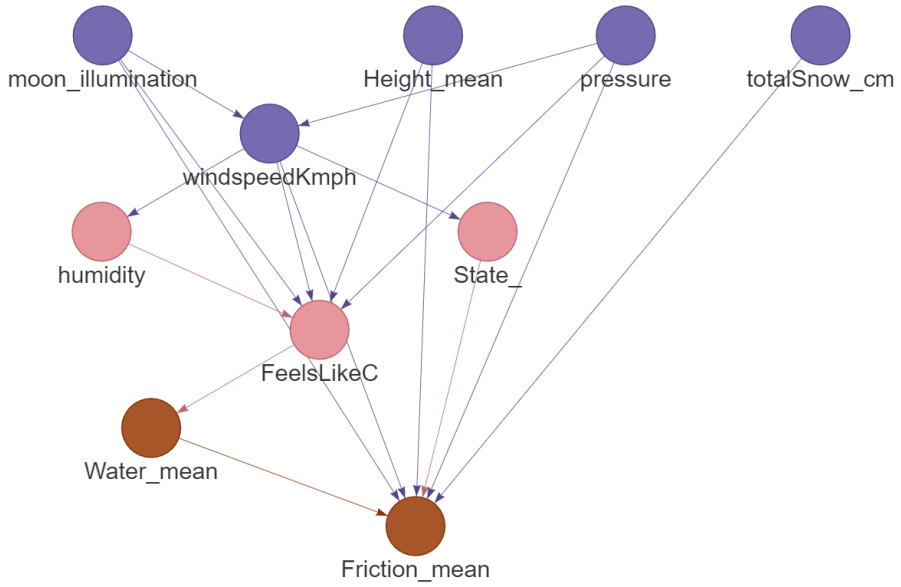

Figure 12: BN built on the correlation analysis results. Green nodes are MixtureGaussian, brown ones are Discrete, and green ones are ConditionalMixtureGaussian, pale pink ones are Logit and pink ones are ConditionalLogit

To be fair, the results are non-satisfactory so we hope that score-constructed BNs will provide good results.

6.2. BN built by score optimization and quality analysis

We have built two BNs with gaussian mixtures allowed and logits denied by means of BIC and K2-score maximization. Their structures are provided in figures 13 and 14. To build the structure one needs to discretize the data.

As for the BN built on BIC, we sampled 10000 synthetic observations and performed the same tests. The results are as follows:

- FeelsLikeC: chi-square test failts to reject the hypothesis that synthetic and original data came from the same distribution with p-value ≈ 0.915;
- windspeedKmph: Kolmogorov-Smirnov test rejects the analogous hypothesis with p-value $\approx 8 \cdot 10^{-6}$;
- Friction_mean: Kolmogorov-Smirnov test rejects the analogous hypothesis with p-value ≈ 0.00047.

Then we did the same for the BN built on K2-score:

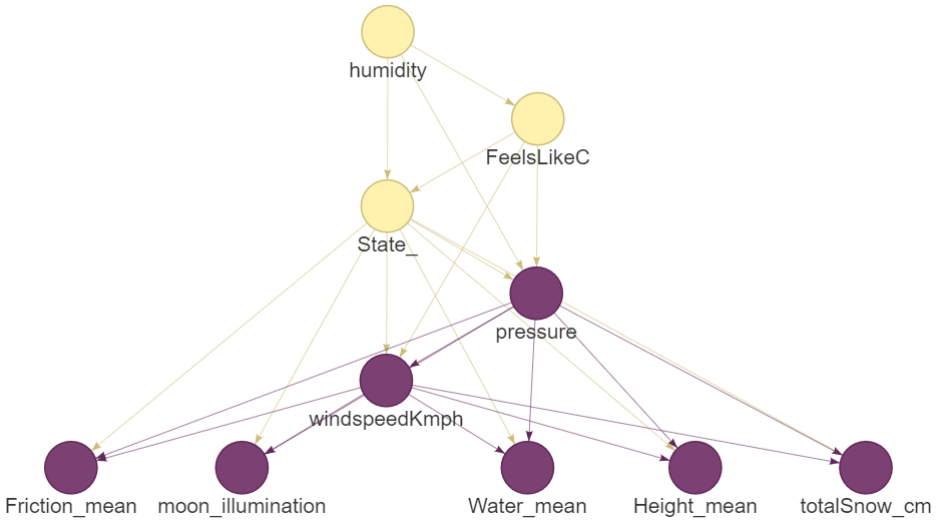

Figure 13: BN built from BIC maximization. Orange nodes are Discrete, and blue ones are ConditionalMixtureGaussian

- FeelsLikeC: chi-square test failts to reject the hypothesis that synthetic and original data came from the same distribution with p-value ≈ 0.999;
- windspeedKmph: Kolmogorov-Smirnov test rejects the analogous hypothesis with p-value ≈ 0.0001;
- Friction_mean: Kolmogorov-Smirnov test fails to reject the analogous hypothesis with p-value ≈ 0.08.

As one can see, the second BN exhibits good quality of synthetic data. Histograms of the generated target variables are presented in figure 15

Also we have checked the quality of synthetic data in the sense of missing data imputation. The experiment was conducted as follows: we obtained another subsample with 1000 observations from the original dataset and assign NA-value to the target variables of 400 randomly chosen individuals. Next, we have predicted missed target values given the evidence of predictors. Unfortunately, 236 observations were left with NA so we discard them from further consideration and the ones corresponding from the original subsample. At last, we compared original data and recovered one by means of statistical tests and analyzing of histograms. The results of tests are as follows:

- FeelsLikeC: chi-square test failts to reject the hypothesis that recovered and original data came from the same distribution with p-value ≈ 0.991;

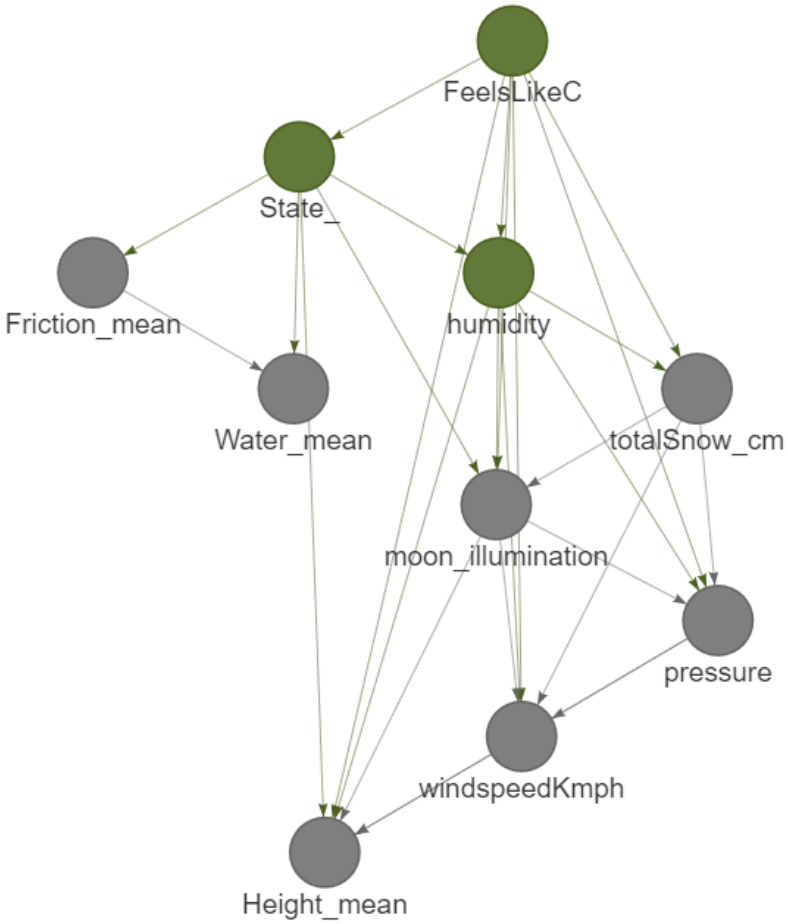

Figure 14: BN built from K2-score maximization. Green nodes are Discrete, and blue ones are ConditionalMixtureGaussian

- Friction_mean: Kolmogorov-Smirnov test fails to reject the analogous hypothesis with p-value ≈ 0.910.

The histograms are illustrated in figure 16.

We may conclude that the built BN is worth using in missed data imputation.

7. Appendix

The Python notebook related to the aforementioned calculations is presented in Github [3].

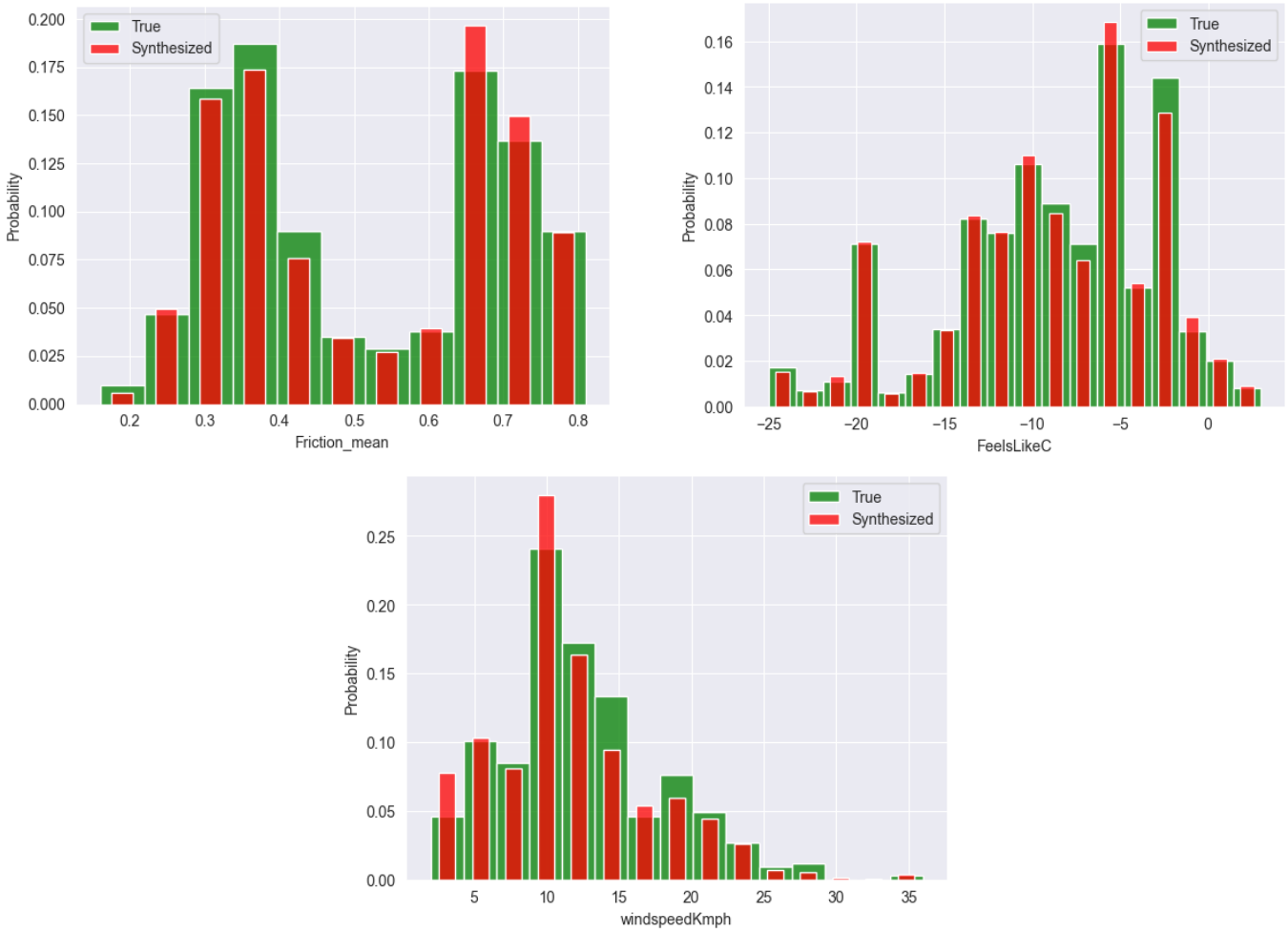

Figure 15: The histograms of targets: synthetic generated by BN based on K2-score vs original

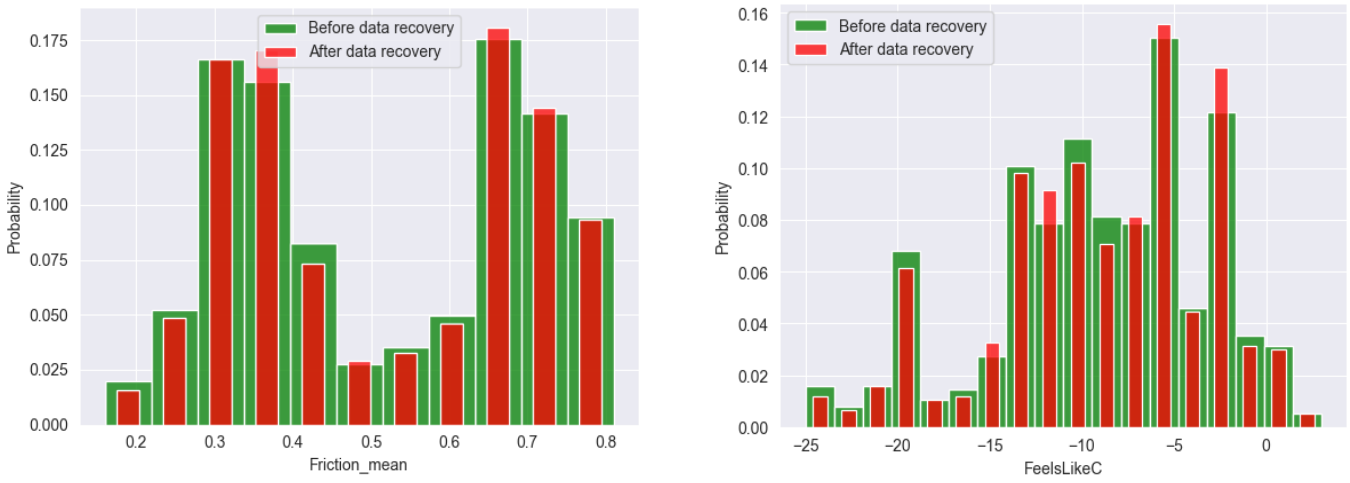

Figure 16: The histograms of targets: recovered by BN based on K2-score vs original

Bibliography

1. BAMT - Bayesian Analytical and Modelling Toolkit. — https://github.com/ITMO-NSS-team/BAMT.
2. CHI SQUARE TWO SAMPLE TEST. — https://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/chi2samp.htm.
3. Grigorev Dmitry. Code repository. — https://github.com/dmitry-grigorev/MultivarAnalysis/blob/master/Lab3/Lab3notebook.ipynb. — 2022.
4. A Novel Method for Circuits of Perfect Electric Conductors in Unstructured Particle-in-Cell Plasma-Object Interaction Simulations / Marholm Sigvald, Darian Diako, Mortensen Mikael, Marchand Richard, and Miloch Wojciech // IEEE Transactions on Plasma Science. — 2020. — 07. — Vol. 48. — P. 1–17.