Presentation 4

Optimization of Parameters

Grid Search

Datasets

- Based on the previous discussion on amount of the whole dataset, I tried to tune the parameters on two different datasets:
- One of the is the Credit Card dataset from the Kaggle website
- Second one is the Simulated dataset which the link was mention in the GitHub

Credit Card dataset

- The parameters that I have tried to tune are:

1. N_0 : Number of references
2. N0 : Number of expected background
3. M : Number of centers
4. NS: Number of Fraudulent usage

- Fixed parameters are:

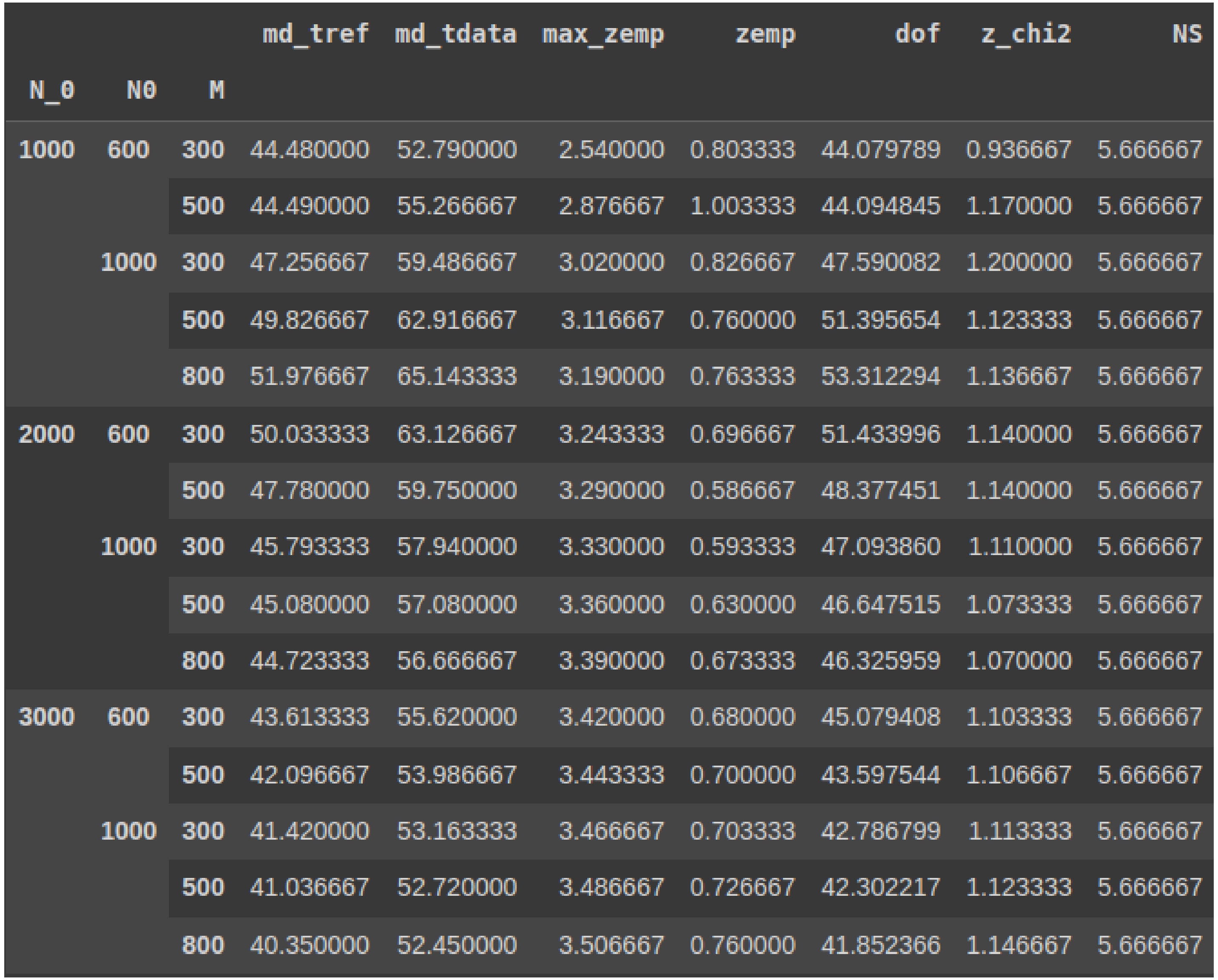
1. Lam = 1e-10
2. N_toys = 300 (reference), 100 (data)
3. Flk_sigma = 3

Grid Search

- In this search, I changed the parameters and the variables that were saved on different scenarios are as follows:

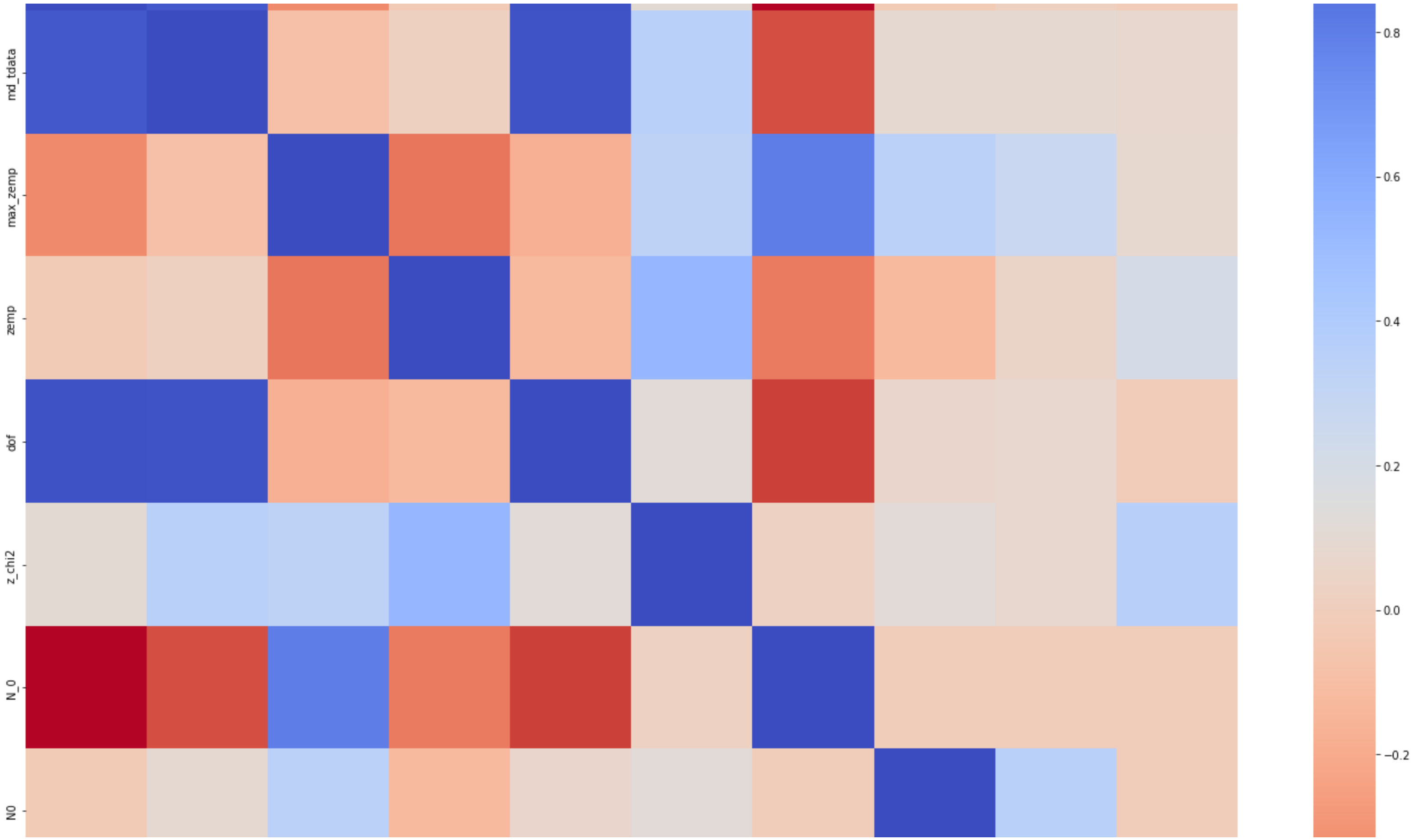
1. Md_tref
2. Md_tdata
3. Max_zemp
4. Zemp
5. DoF
6. Z_chi2

Results

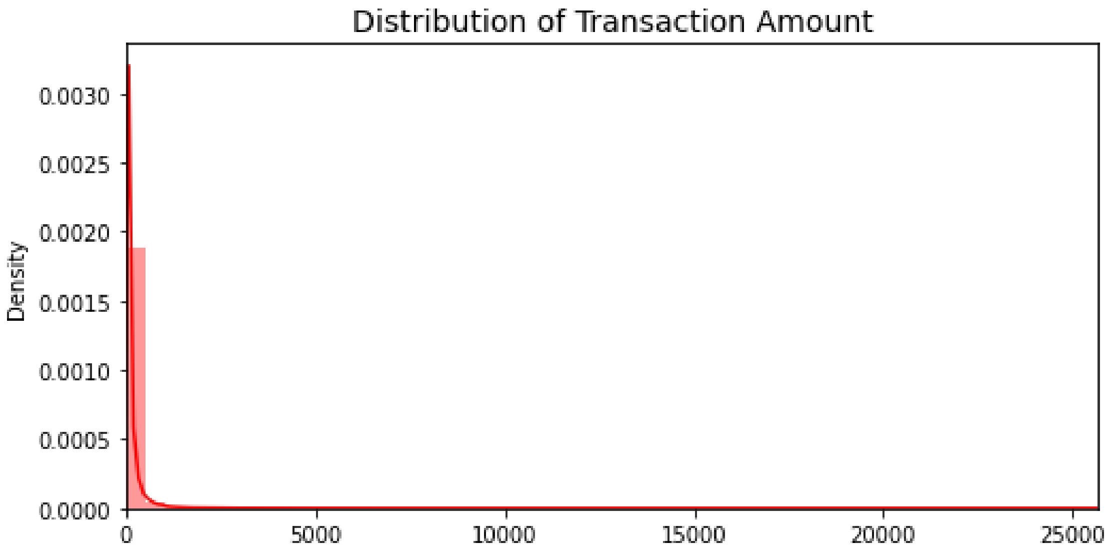


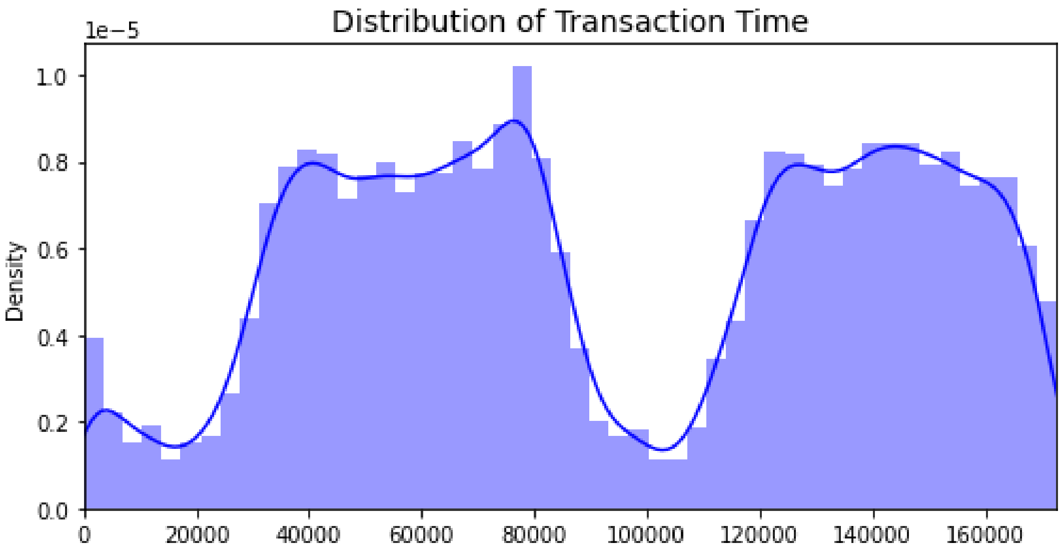
			md_tref	md_tdata	max_zemp	zemp	dof	z_chi2	NS
N_0	N0	M							
1000	600	300	44.480000	52.790000	2.540000	0.803333	44.079789	0.936667	5.666667
		500	44.490000	55.266667	2.876667	1.003333	44.094845	1.170000	5.666667
	1000	300	47.256667	59.486667	3.020000	0.826667	47.590082	1.200000	5.666667
		500	49.826667	62.916667	3.116667	0.760000	51.395654	1.123333	5.666667
		800	51.976667	65.143333	3.190000	0.763333	53.312294	1.136667	5.666667
2000	600	300	50.033333	63.126667	3.243333	0.696667	51.433996	1.140000	5.666667
		500	47.780000	59.750000	3.290000	0.586667	48.377451	1.140000	5.666667
	1000	300	45.793333	57.940000	3.330000	0.593333	47.093860	1.110000	5.666667
		500	45.080000	57.080000	3.360000	0.630000	46.647515	1.073333	5.666667
		800	44.723333	56.666667	3.390000	0.673333	46.325959	1.070000	5.666667
3000	600	300	43.613333	55.620000	3.420000	0.680000	45.079408	1.103333	5.666667
		500	42.096667	53.986667	3.443333	0.700000	43.597544	1.106667	5.666667
	1000	300	41.420000	53.163333	3.466667	0.703333	42.786799	1.113333	5.666667
		500	41.036667	52.720000	3.486667	0.726667	42.302217	1.123333	5.666667
		800	40.350000	52.450000	3.506667	0.760000	41.852366	1.146667	5.666667

Correlation Matrices



Distribution of Time and Amount





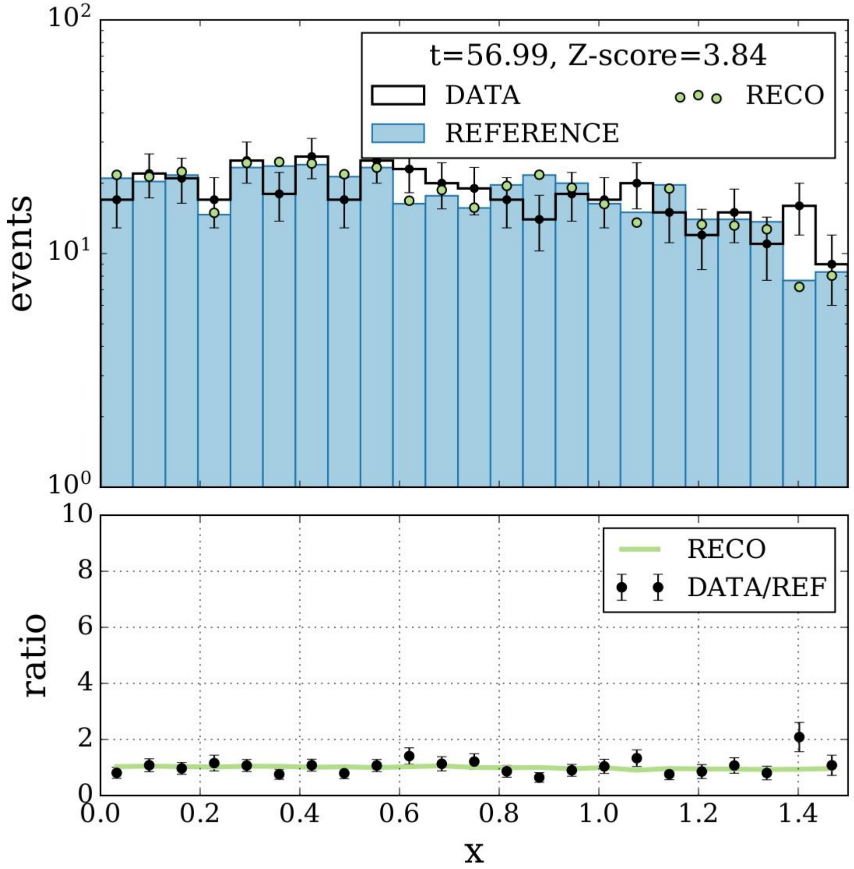
Discussion

- Based on the result and the correlation matrix, N_0 has the most negative effect on z_chi2 and NS has the most positive effect on the z_chi2(if we set z_chi2 as our results measure).

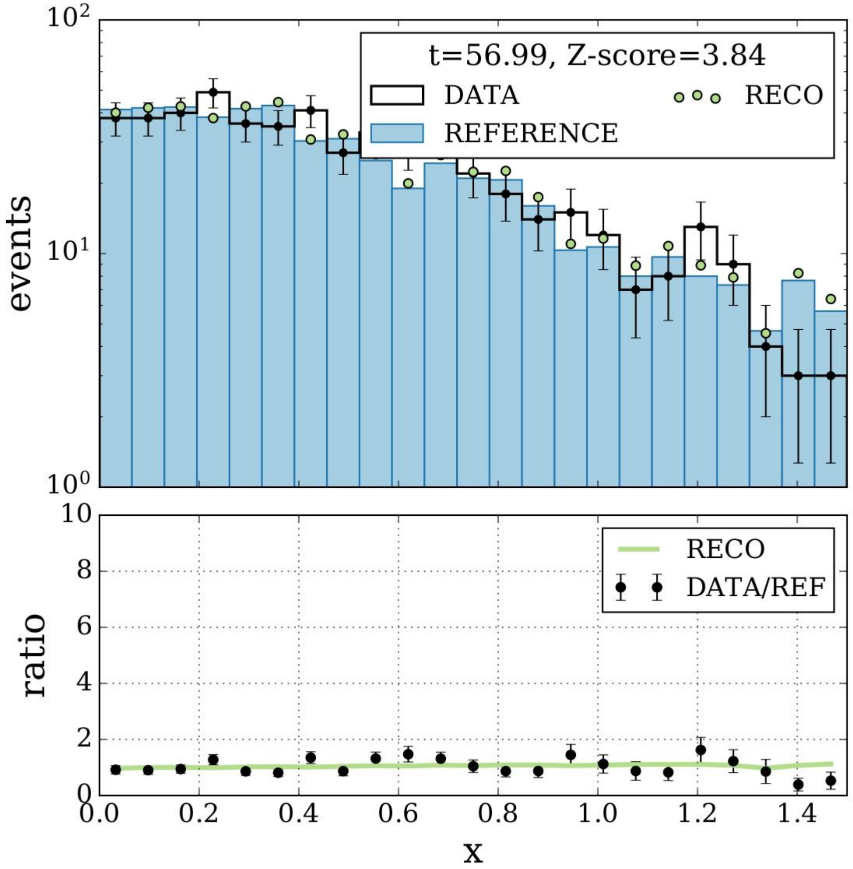
Optimized parameters used in the codes

- Parameters:
- N_0 = 3000
- N0 = 1000
- Lam = 1e-10
- Flk_sigma = 3
- N_toys = 300, 100
- M = 800
- NS = 10

Reference

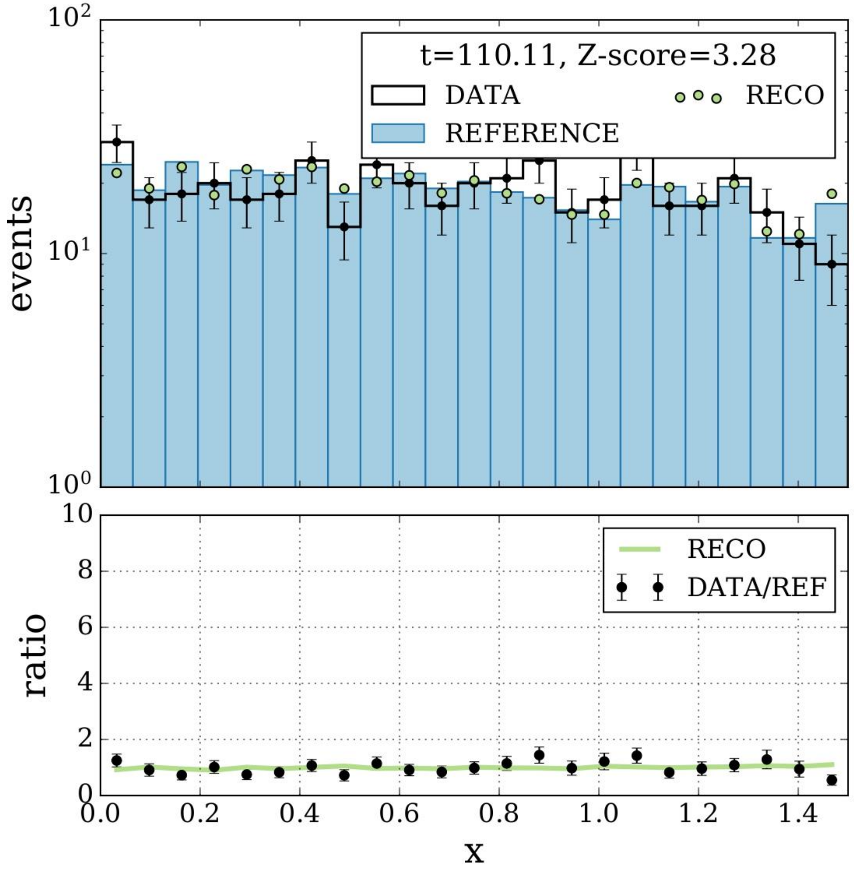


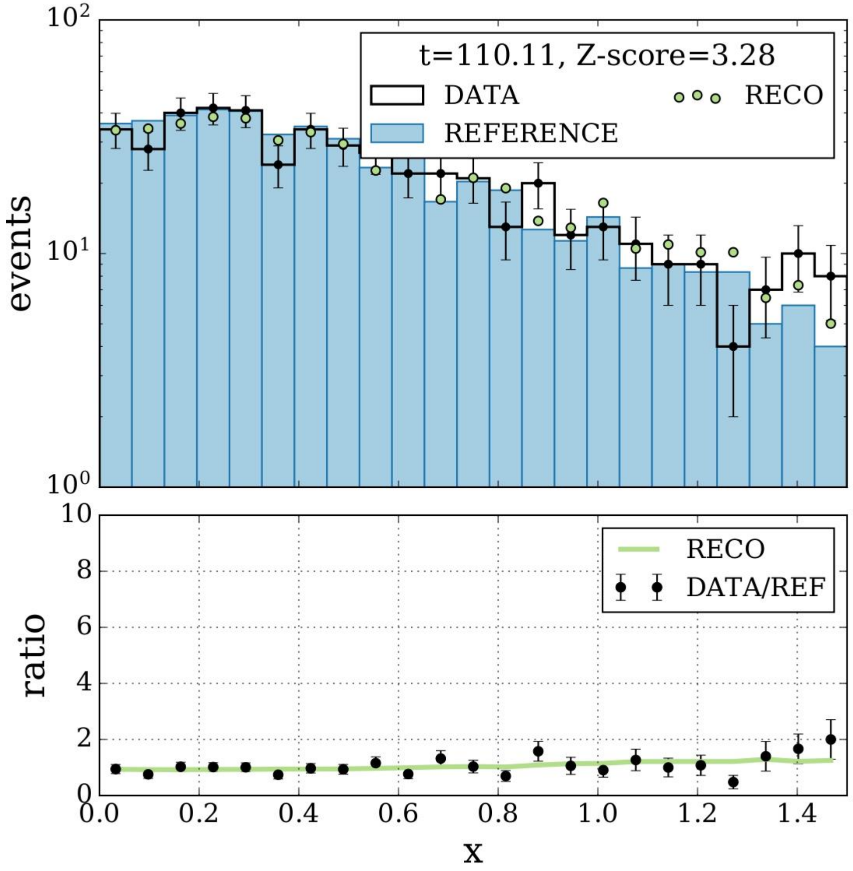
t=56.99, Z-score=3.84
DATA
RECO
REFERENCE
events
10^2
10^1
10^0
RECO
DATA/REF
ratio
10
8
6
4
2
0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
x



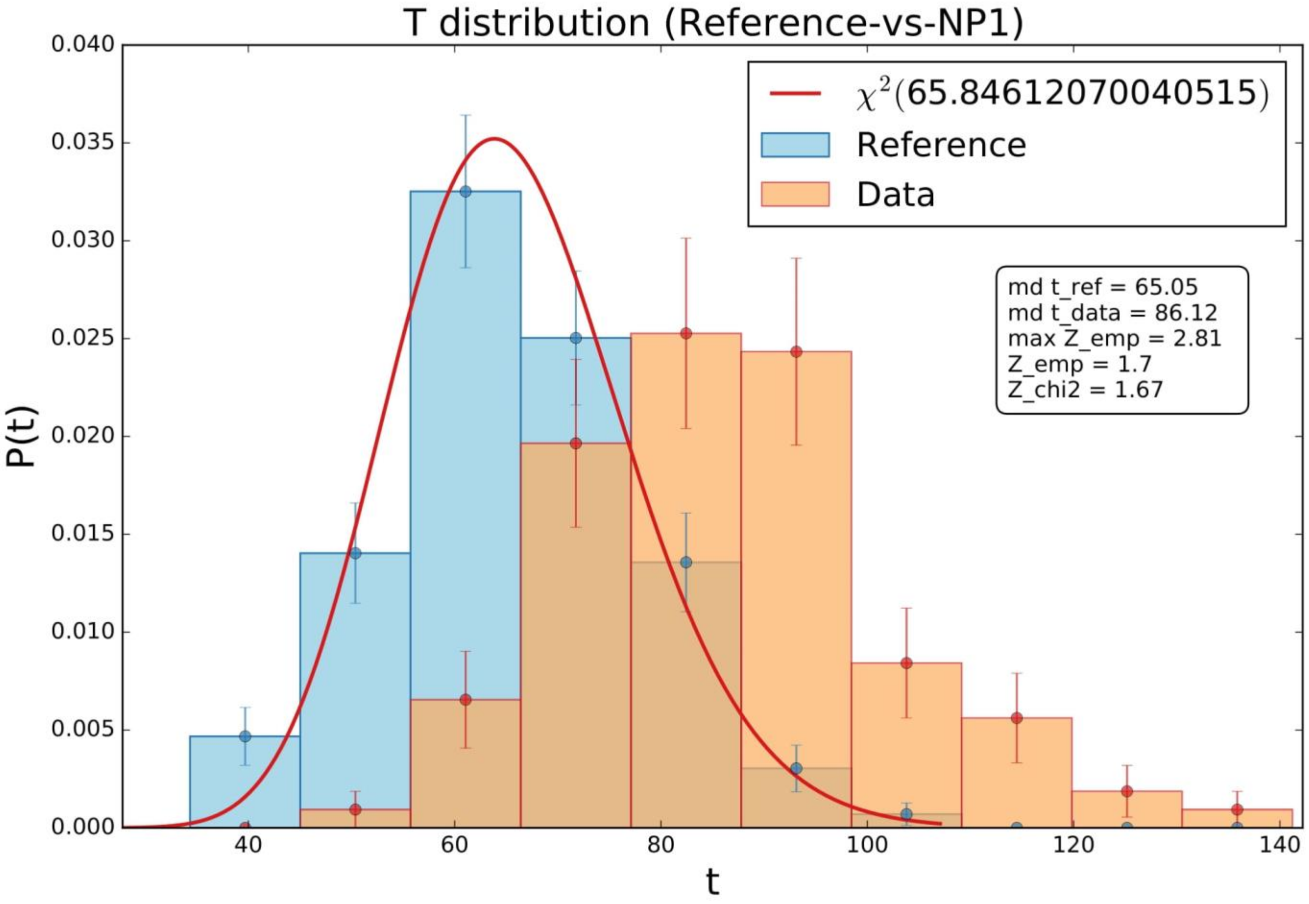
t=56.99, Z-score=3.84
DATA
RECO
REFERENCE
events
10^2
10^1
10^0
RECO
DATA/REF
ratio
10
8
6
4
2
0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
x

Data

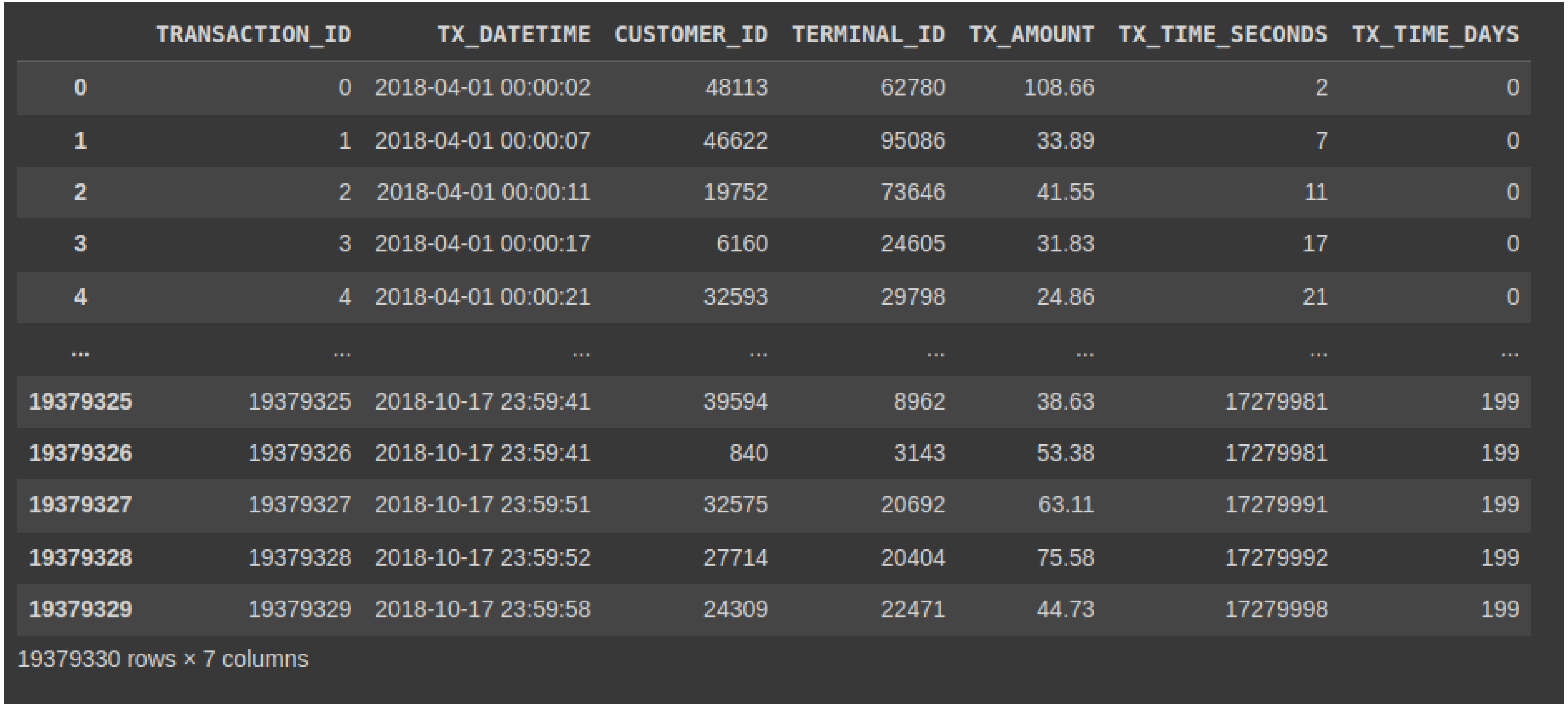

t=110.11, Z-score=3.28
DATA
REFERENCE
RECO
events
10^2
10^1
10^0
RECO
DATA/REF
ratio
10
8
6
4
2
0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
x


t=110.11, Z-score=3.28
DATA
REFERENCE
RECO
events
10^2
10^1
10^0
RECO
DATA/REF
ratio
10
8
6
4
2
0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
x

T_Distribution



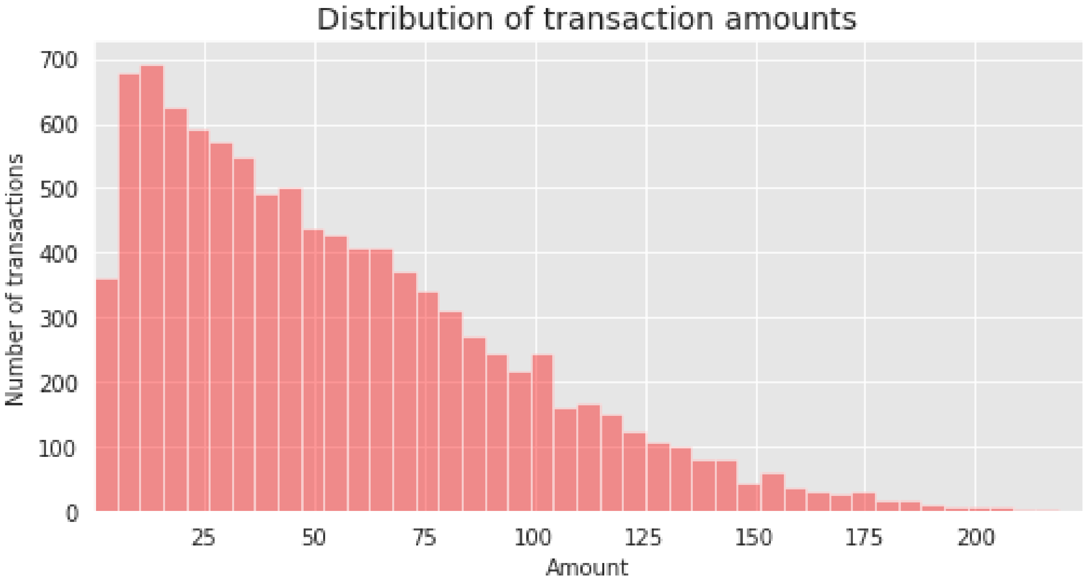
Simulated dataset

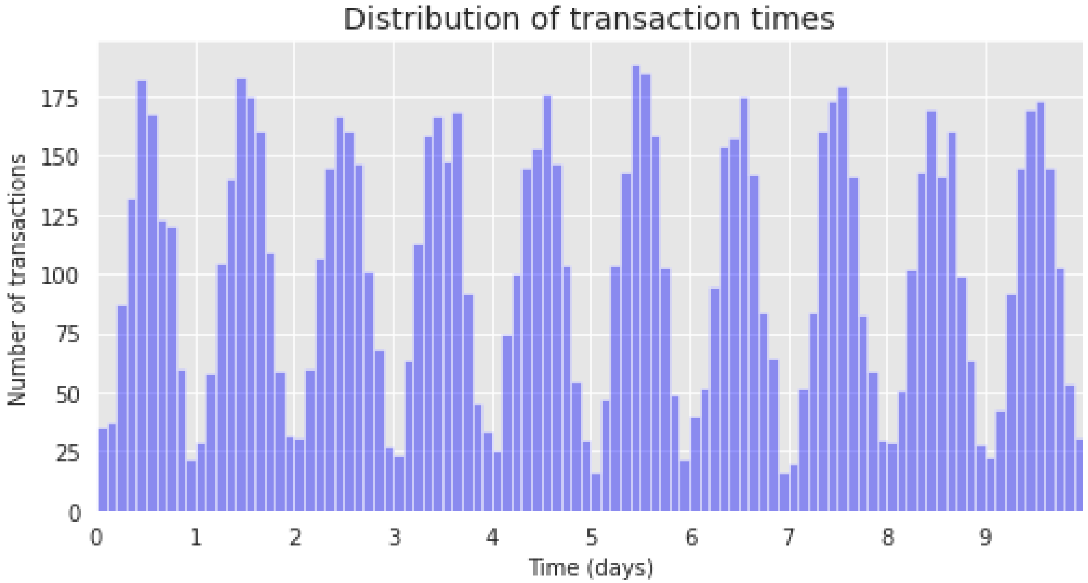


	TRANSACTION_ID	TX_DATETIME	CUSTOMER_ID	TERMINAL_ID	TX_AMOUNT	TX_TIME_SECONDS	TX_TIME_DAYS
0	0	2018-04-01 00:00:02	48113	62780	108.66	2	0
1	1	2018-04-01 00:00:07	46622	95086	33.89	7	0
2	2	2018-04-01 00:00:11	19752	73646	41.55	11	0
3	3	2018-04-01 00:00:17	6160	24605	31.83	17	0
4	4	2018-04-01 00:00:21	32593	29798	24.86	21	0
...	...	...	...	...	...	...	...
19379325	19379325	2018-10-17 23:59:41	39594	8962	38.63	17279981	199
19379326	19379326	2018-10-17 23:59:41	840	3143	53.38	17279981	199
19379327	19379327	2018-10-17 23:59:51	32575	20692	63.11	17279991	199
19379328	19379328	2018-10-17 23:59:52	27714	20404	75.58	17279992	199
19379329	19379329	2018-10-17 23:59:58	24309	22471	44.73	17279998	199

19379330 rows × 7 columns

Simulated dataset





Simulated dataset

- The parameters that I have tried to tune are:

1. N_0 : Number of references
2. N0 : Number of expected background
3. M : Number of centers
4. NS: Number of Fraudulent usage

- Fixed parameters are:

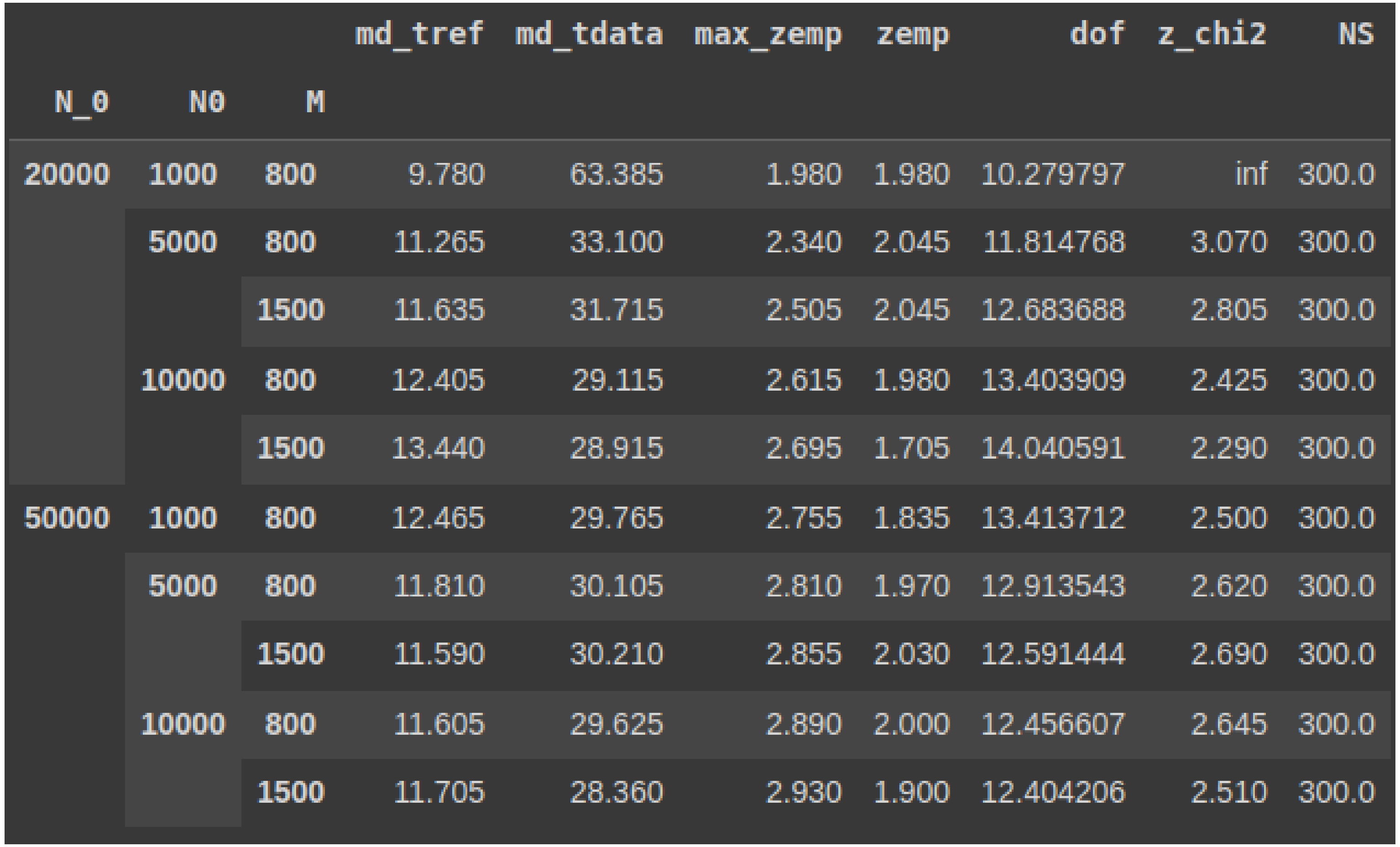
1. Lam = 1e-7
2. N_toys = 30 (reference), 30 (data)
3. Flk_sigma = 3

Grid Search

- In this search, I changed the parameters and the variables that were saved on different scenarios are as follows:

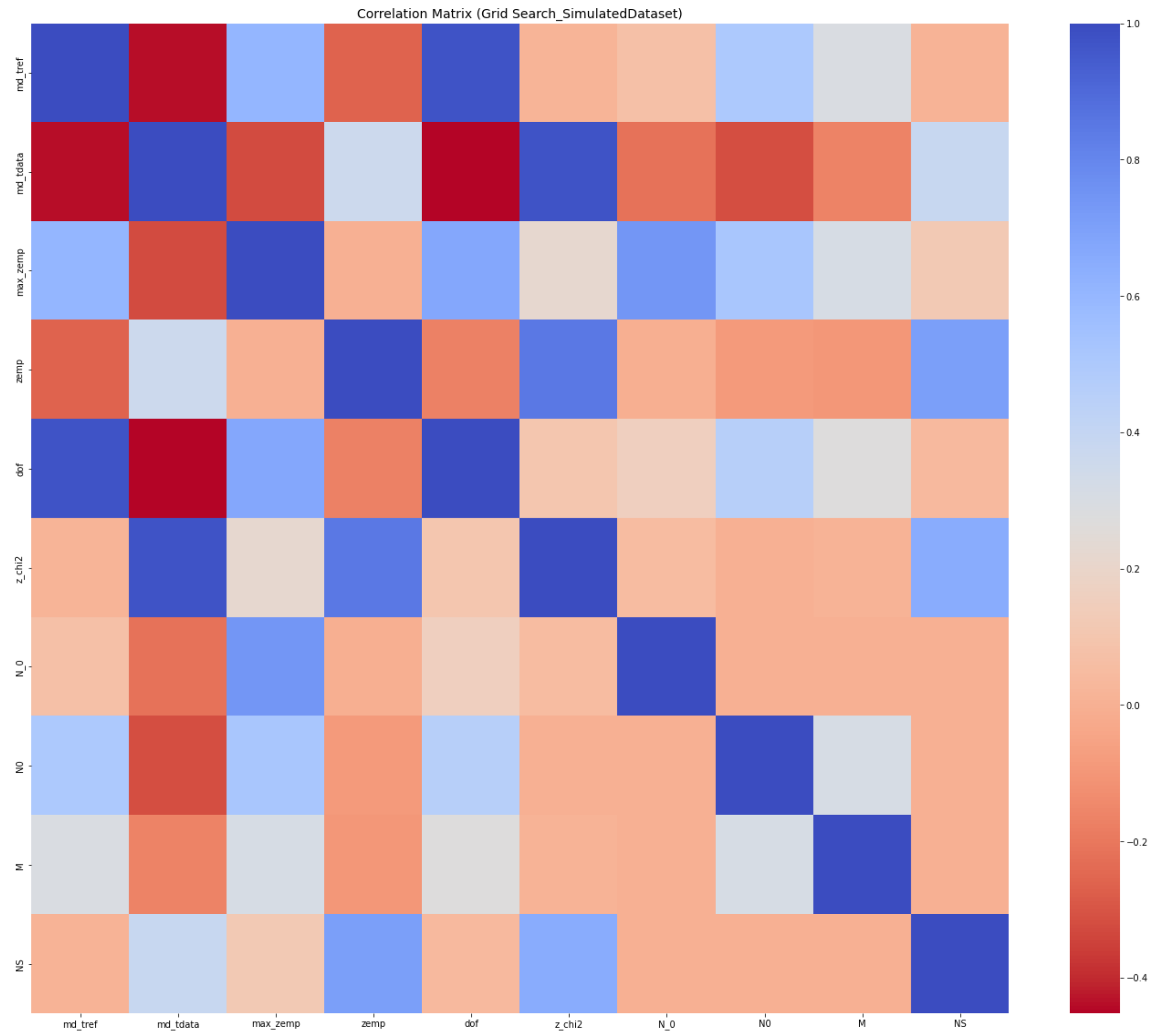
1. Md_tref
2. Md_tdata
3. Max_zemp
4. Zemp
5. DoF
6. Z_chi2

Results



			md_tref	md_tdata	max_zemp	zemp	dof	z_chi2	NS
N_0	N0	M							
20000	1000	800	9.780	63.385	1.980	1.980	10.279797	inf	300.0
	5000	800	11.265	33.100	2.340	2.045	11.814768	3.070	300.0
		1500	11.635	31.715	2.505	2.045	12.683688	2.805	300.0
	10000	800	12.405	29.115	2.615	1.980	13.403909	2.425	300.0
		1500	13.440	28.915	2.695	1.705	14.040591	2.290	300.0
50000	1000	800	12.465	29.765	2.755	1.835	13.413712	2.500	300.0
	5000	800	11.810	30.105	2.810	1.970	12.913543	2.620	300.0
		1500	11.590	30.210	2.855	2.030	12.591444	2.690	300.0
	10000	800	11.605	29.625	2.890	2.000	12.456607	2.645	300.0
		1500	11.705	28.360	2.930	1.900	12.404206	2.510	300.0

Correlation Matrices



Discussion

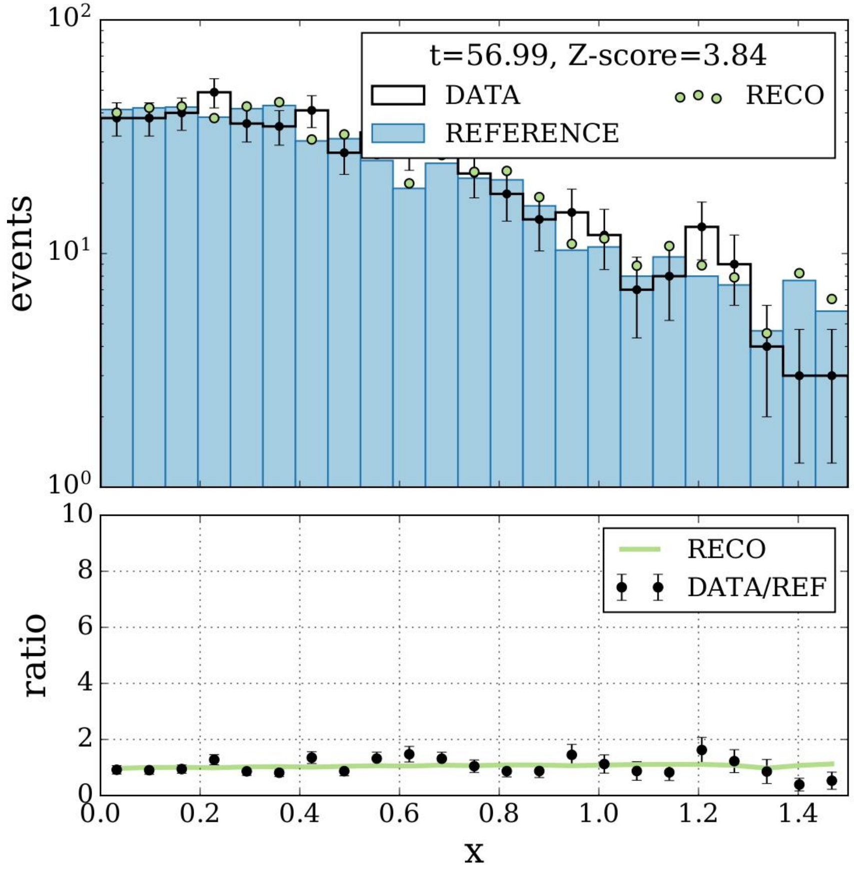
- Based on the results and the total number of simulated data the optimum parameters are:

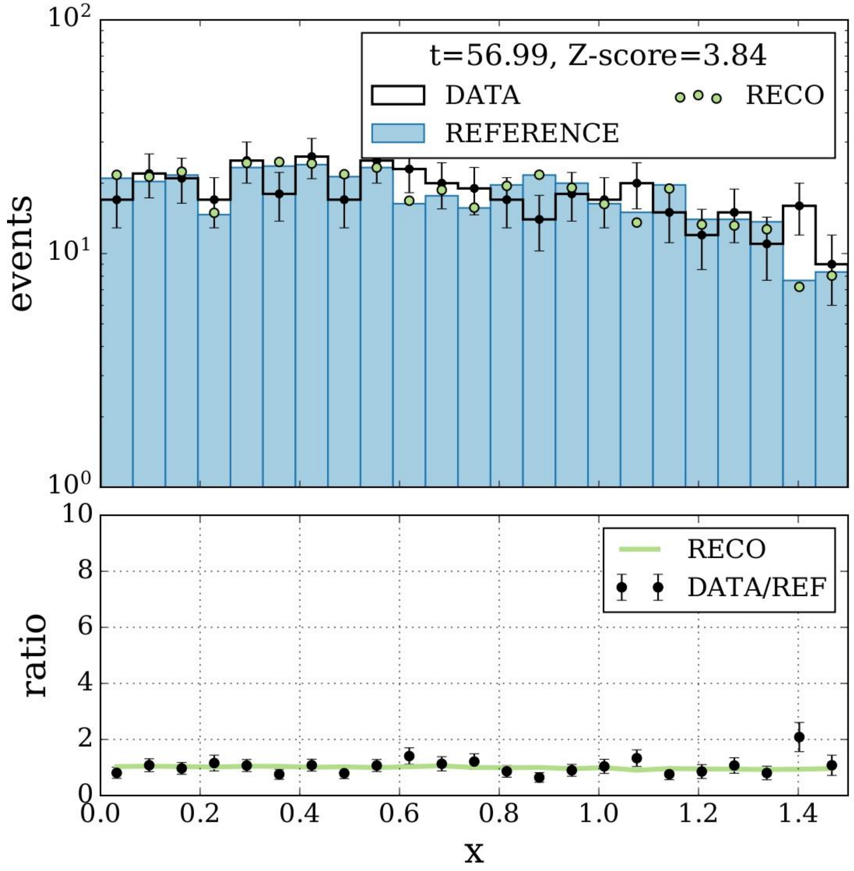
1. N_0 = 20000
2. N0 = 10000
3. M = 800
4. Ns = 500

Optimized parameters used in the codes

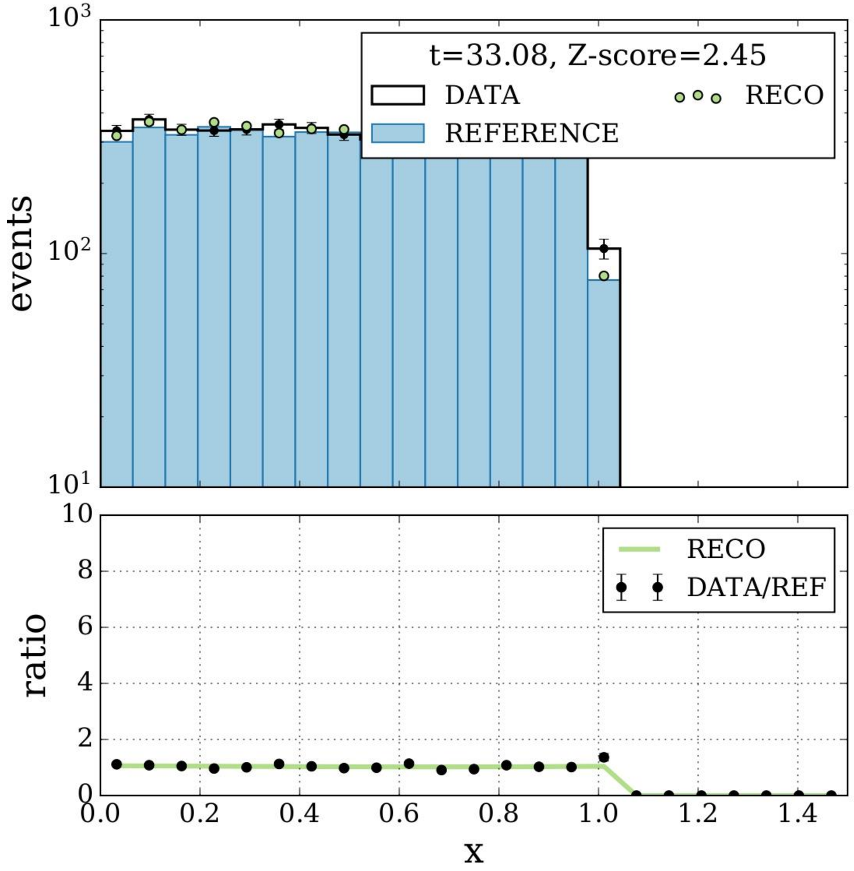
- Parameters:
- N_0 = 20000
- N0 = 10000
- Lam = 1e-7
- Flk_sigma = 3
- N_toys = 30, 30
- M = 800
- NS = 500

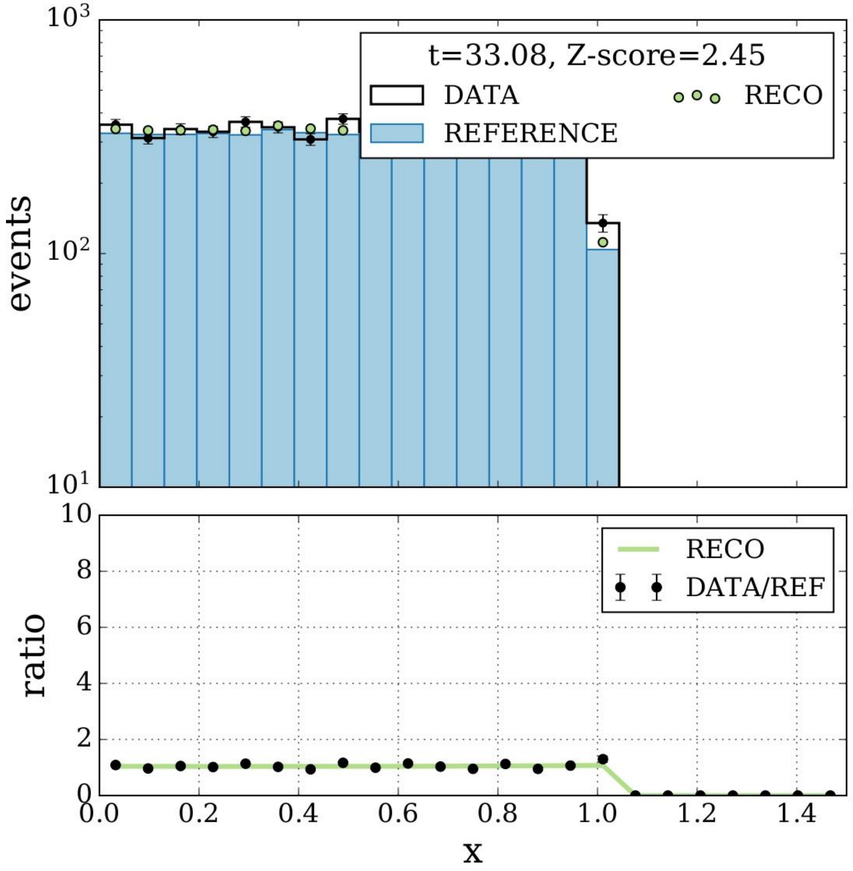
Reference


t=56.99, Z-score=3.84
DATA
RECO
REFERENCE
events
10^2
10^1
10^0
ratio
10
8
6
4
2
0
RECO
DATA/REF
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
x

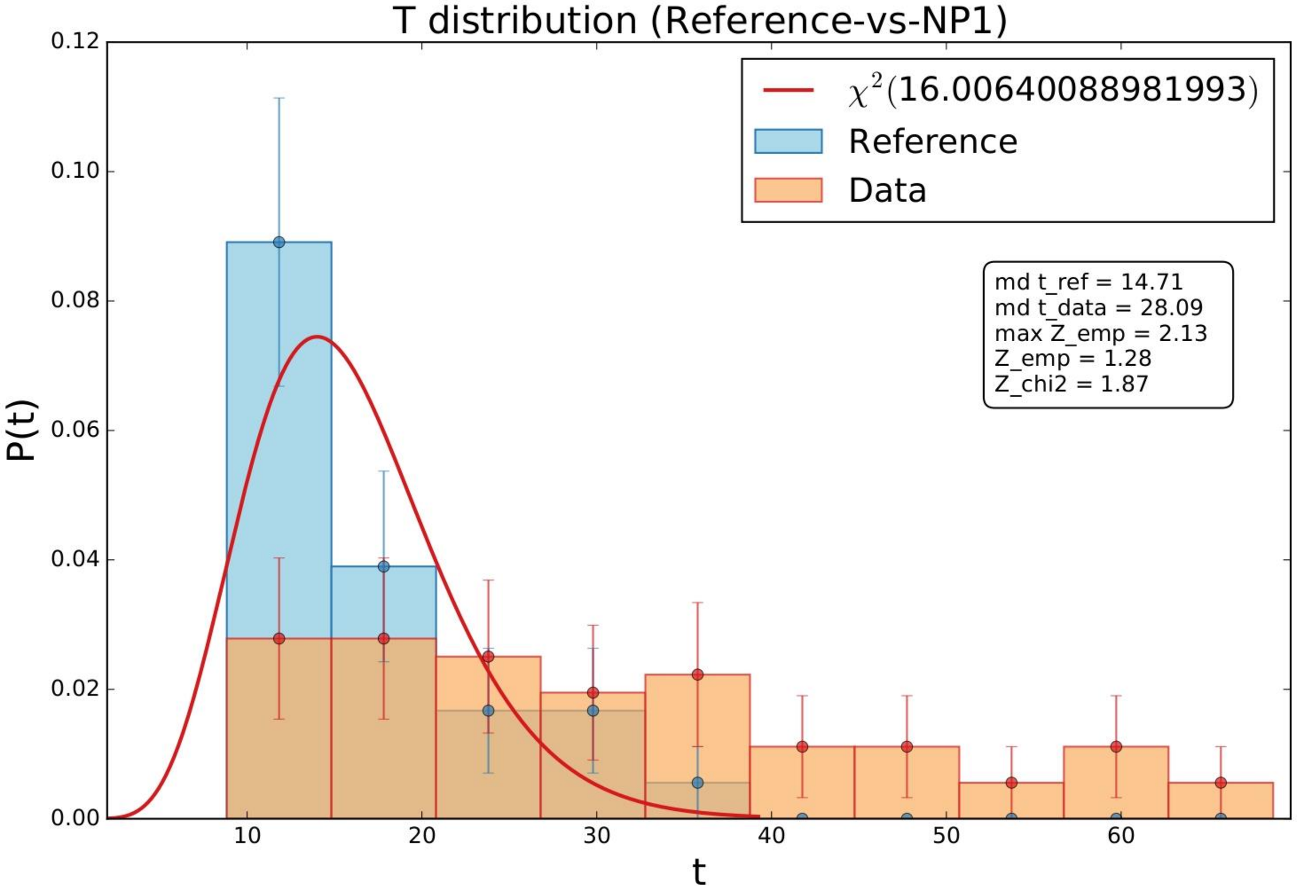

t=56.99, Z-score=3.84
DATA
RECO
REFERENCE
events
10^2
10^1
10^0
ratio
10
8
6
4
2
0
RECO
DATA/REF
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
x

Data


t=33.08, Z-score=2.45
DATA
RECO
REFERENCE
events
10^3
10^2
10^1
ratio
10
8
6
4
2
0
RECO
DATA/REF
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
x


t=33.08, Z-score=2.45
DATA
RECO
REFERENCE
events
10^3
10^2
10^1
ratio
10
8
6
4
2
0
RECO
DATA/REF
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
x

T_Distribution


T distribution (Reference-vs-NP1)
$\chi^2(16.00640088981993)$
Reference
Data
md t_ref = 14.71
md t_data = 28.09
max Z_emp = 2.13
Z_emp = 1.28
Z_chi2 = 1.87
P(t)
t