TÜRKİYE CUMHURİYET MERKEZ BANKASI

Data Sharing Under Confidentiality: CBRT Case

Timur Hülagü, Ph. D.

AUGUST 30, 2018 | BASEL

Disclaimer

This is a joint project between CBRT and METU. All views expressed here are those of authors and do not necessarily reflect those of the two institutions.

Motivation

- **Main Goal**: Address the growing need of accessing micro data for academic research

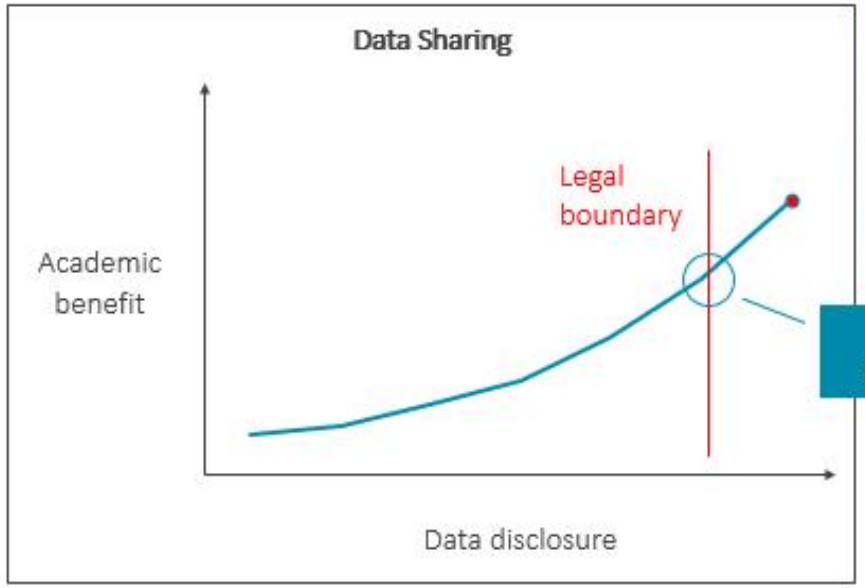
- **G-20 Data Gaps Initiative 2, Recommendation II.20:** Promotion of Data Sharing by G-20 Economies

Share information and ideas on ways to apply confidential rules/arrangements in a manner that would allow sharing of more granular data

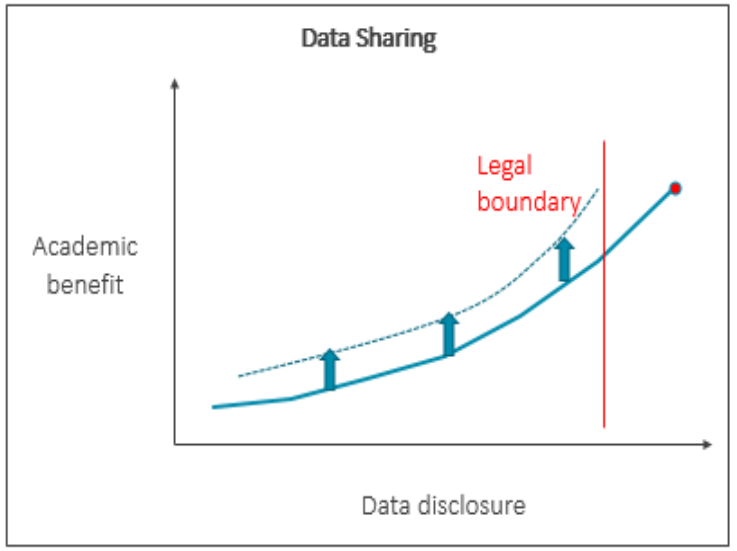
- **Eurostat Peer review report on the compliance with the Code of Practice and the coordination role of the National Statistical Institute in Turkey**

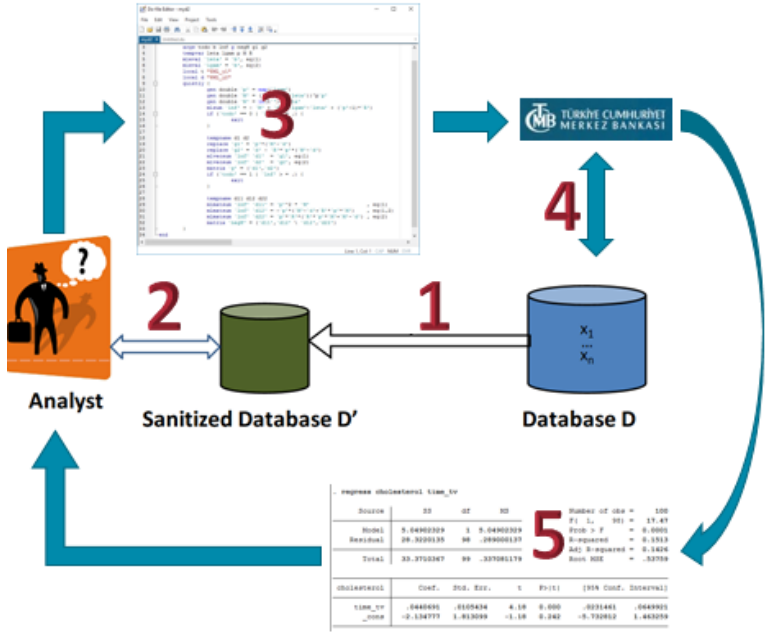
Recommendation 22: TurkStat should introduce remote access facilities for researchers, who are permitted to use its anonymized microdata for research purposes (European Statistics Code of Practice, indicator 15.4)

Data Sharing Trade-off

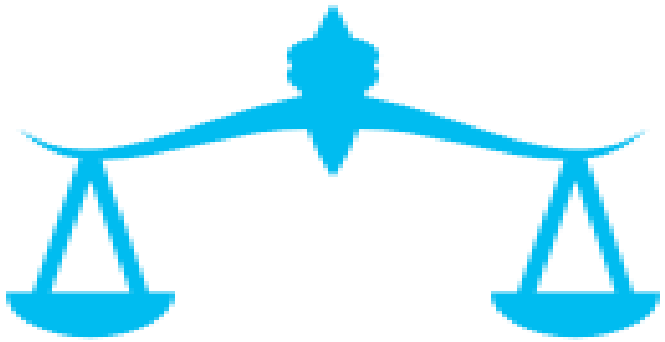


CBRT Practice





Main Aspects



Accuracy **Privacy**

Accuracy

- Descriptive Analysis ☑
- Univariate Regression Analysis ☑
- Multivariate Regression Analysis ☑
- Logistic Regression ☑
- Logarithmic Regression ☑

Accuracy

Descriptive analysis

Variable	B30	B50	B15	G600
Seed	123000	234000	345000	456000
Mean (O/M)	0.9756	0.9756	0.9756	0.9756
Standart Deviation (O/M)	0.9997	0.9995	0.9999	0.9998
Skewness (O/M)	1.001	1.003	1.002	1.002
Kurtosis (O/M)	1.002	1.008	1.004	1.004

Multiple Linear Regression Analysis

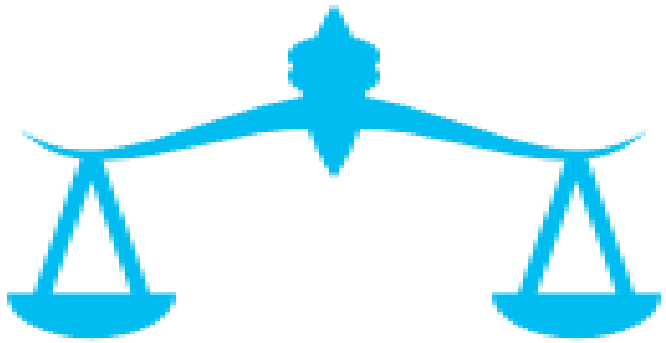
Original

```
Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept)  8.08888    0.11373   71.13 < 0.0000000000000002 ***
B2L          0.04716    0.00616    7.66    0.000000000000022 ***
B10L         0.21006    0.00486   43.22 < 0.0000000000000002 ***
B32L         0.32332    0.00583   55.44 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.856 on 7342 degrees of freedom
Multiple R-squared:  0.511,     Adjusted R-squared:  0.511
F-statistic: 2.56e+03 on 3 and 7342 DF,  p-value: <0.0000000000000002
```

Masked

```
Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept)  8.28657    0.11670   71.01 < 0.0000000000000002 ***
B2L_M        0.04777    0.00617    7.75    0.000000000000011 ***
B10L_M       0.20994    0.00486   43.16 < 0.0000000000000002 ***
B32L_M       0.32314    0.00584   55.36 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.857 on 7342 degrees of freedom
Multiple R-squared:  0.51,      Adjusted R-squared:  0.51
F-statistic: 2.55e+03 on 3 and 7342 DF,  p-value: <0.0000000000000002
```

Main Aspects



Accuracy **Privacy**

Privacy

- Deeper Focus on Privacy and Security

i. Spectral Filtering
ii. Singular Value Decomposition
iii. Principal Component Analysis

Attacks

$$\mathrm{m} = \left[\frac{d(O,E)}{d(O,M)}\right] \quad \text{where} \quad d(A,B) = \sum_{i=1}^{n} |a_i - b_i|$$

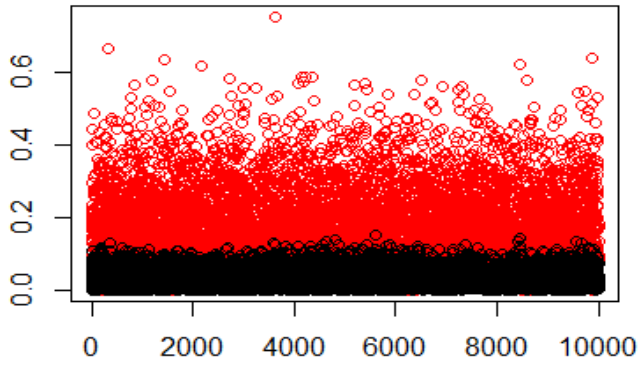
Privacy

Spectral Filtering

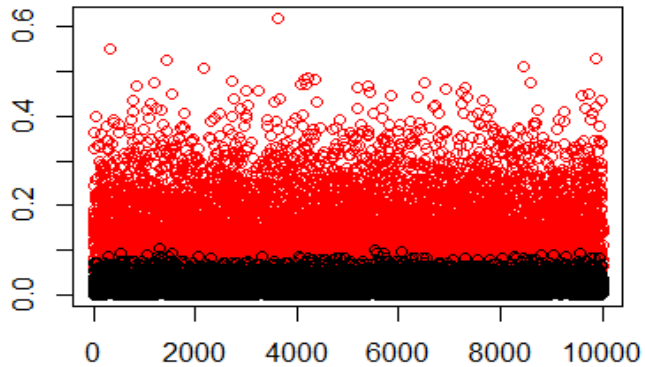
	Sinusoidal	Triangular
λ_{min}	$0,01121$	$0,007627$
λ_{max}	$0,06104$	$0,04152$
m x n	250x40	250x40
k	2	1
$d(O,M)$	1410	1163
$d(O,E)$	$319,6$	$205,2$
m	$0,227$	$0,176$

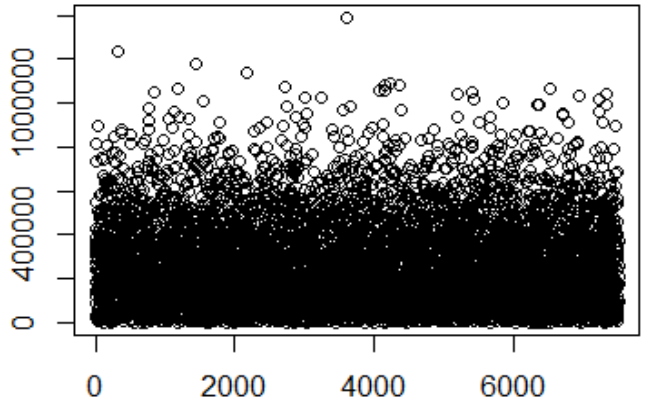
	G69
λ_{min}	45870034800
λ_{max}	194658444176
m x n	250x30
k	30
$d(O,M)$	1964402616
$d(O,E)$	1964402616
m	1

Sinusoidal



Triangular





No disclosure

Privacy

Random Number Generation In Our System

We can use 2^{30} different seed for generating random numbers. So that, brute force attacks may be a threat. To solve this problem we offer the following algorithm.

Original Data	Seed	Noise	Masked Data
x_1	$s_1 = f(\mathrm{IV} + x_1) mod\ 2^{30}$	$\varepsilon_1 = \mathrm{RNG}(s_1)$	$x'_1 = x_1 + \varepsilon_1$
x_2	$s_2 = f(x_2 + x'_1) mod\ 2^{30}$	$\varepsilon_2 = \mathrm{RNG}(s_2)$	$x'_2 = x_2 + \varepsilon_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_n	$s_n = f(x_n + x'_{n-1}) mod\ 2^{30}$	$\varepsilon_n = \mathrm{RNG}(s_n)$	$x'_n = x_n + \varepsilon_n$

We choose nonlinear f(x), such that : $f(x) = \mu x^3 + \sigma$

Review

What is done

- We achieve, measurable accuracy on masked data.
- We observed that our system is secure for the attacks we mentioned.

Future Works

- We will study two new attacks called map estimation and distribution analysis.
- In the masked data set, we will check the accuracy of other statistical functions.
- Finally, we will produce a user friendly software product developed by using Java and R.