Fake it 'till you make it
Generating synthetic data

Erik-Jan van Kesteren
Thom Volker

Utrecht University
ODISSEI Social Data Science team

The privacy-utility tradeoff

Utility vs. privacy

Utility

- How close is my synthetic data to my real data? Can I distinguish synthetic and real samples?
- How can we define ‘close’?

Privacy

When I have the synthetic data generated by $p(\boldsymbol{X}|\theta)$, how well can I

- Reproduce the original data? (model inversion attack)
- Determine whether a person was part of the original data? (differential privacy)
- Estimate a specific person’s income within certain bounds? (attribute disclosure)

Utility and **privacy** are opposites

How much does the synthetic data look like the real data?

I don't know what I'm looking at

Perfect imitation

Utility **Privacy**

In principle, you always lose information when creating synthetic data. The question is 'how much information do you need to sacrifice in order to protect privacy?'

Utility vs. privacy

- Every parameter in the data-generating model contains **information** about the observations in the real data
- The more parameters (information) you use to generate synthetic data, the more utility it will have
- When the information in the parameters equals the information in the real data, we have just recreated the real data
- At that point, there is no more privacy / disclosure control

What can we do with the synthetic data?

Anything you can do with real data

Nothing

Utility

Privacy

What can we do with the synthetic data?

- Getting to know the data
- Use the data as a toy example
- Develop & validate data analysis scripts and pipelines
- …

Nothing

Find out how much your colleagues earn

Investigate & answer all your research questions

Anything you can do with real data

Basic correlation analysis

Utility

Privacy

Estimate parameters with low simulation error

Visualisation of association

Visualisation of variation

How to evaluate utility?

General vs. specific utility

General utility

Assessing **similarity** between the real and synthetic data

- Comparing the multivariate distributions of both data sets
- Predicting whether observations are real or synthetic

Specific utility

Assessing **similarity** between results of analyses on real and synthetic data

Quantitative measures

Comparing descriptive statistics

	Mean	Median	Standard deviation	Skewness
X^{true}	3.060	3.000	1.426	-0.033
X^{syn}	3.063	3.066	1.439	-0.13

Predicting whether observations are "real" or "synthetic" ($pMSE$)?

1. Stack the real and synthetic data.
2. Train a prediction model (e.g., logistic regression, CART) to distinguish "real" from "synthetic" data
3. Compare predictions ($pMSE$) with a reference distribution to see if our predictions are better than chance ($pMSE$-ratio; < 3 is considered good; < 10 acceptable).

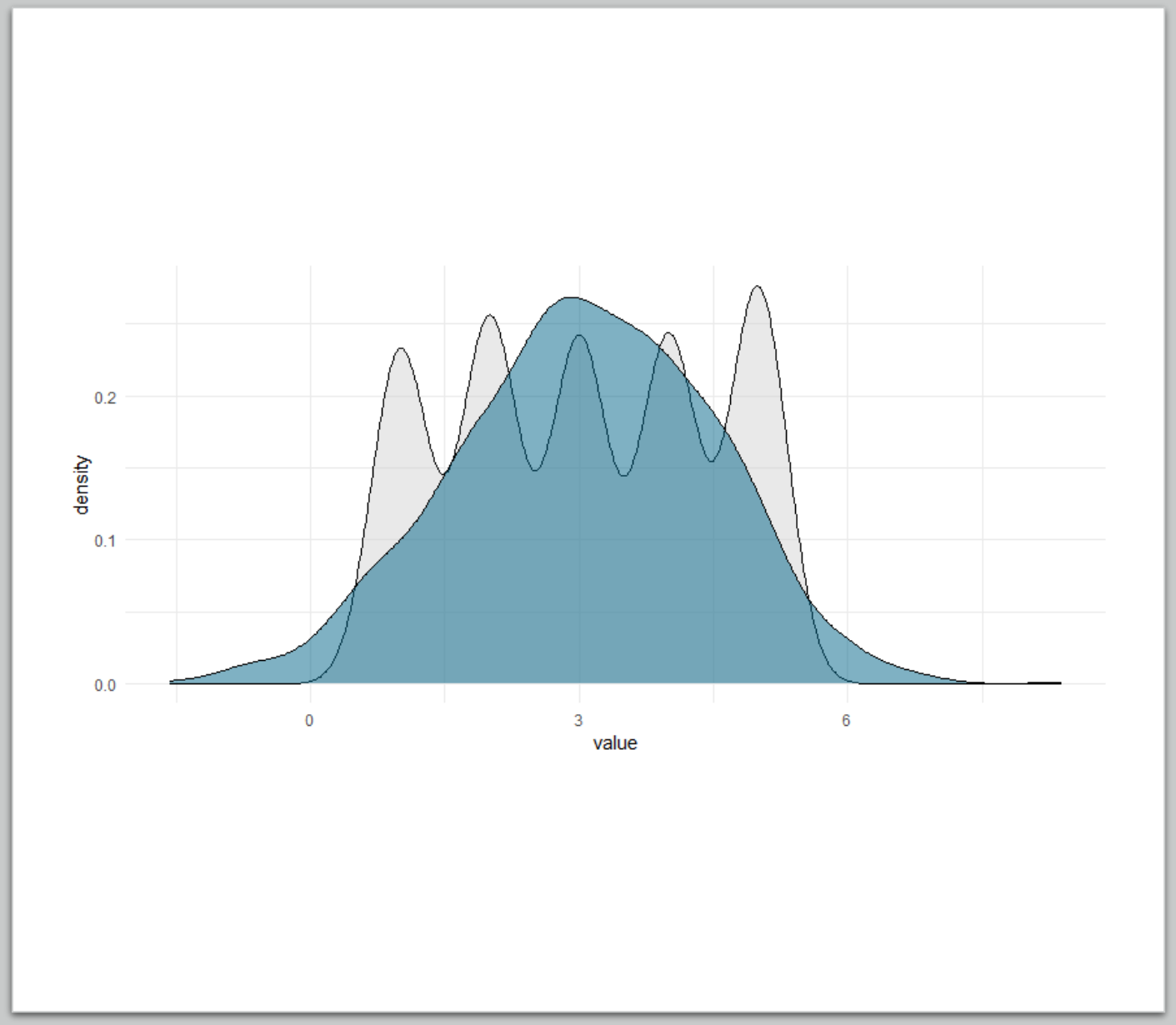
$pMSE$-ratio(X^{true}, X^{syn}) = 0.958

Visual measures

Quantitative measures can be misleading

Visual inspection is almost always superior

But difficult on more than 2 dimensions



How to evaluate privacy?

Evaluating privacy

Much more difficult than evaluating utility

Context-dependent

- How sensitive is the data?
- What are the disclosure risks in the original data?

Check whether the original observations are not accidentally reproduced in the synthetic data.

Use common sense!

Practical 2

Evaluating privacy and utility of synthetic data

Icons from the noun project

Scientist by Justicon
Idea by Icon
house tree by LUTFI GANI AL ACHMAD
Euro by Larea
people by Alice Design
Table by Alex Burte
Hacking by Alfredo
Paper by Egi Maulana
Scientist 2 by Justicon
Question by Anggara Putra

https://thenounproject.com/