Synthetic Data

Broadening the scope of replication

Joachim K. Rennstich, CVJM- Hochschule, @digprof

24 – 25 Sep 2020 | Workshop *Teaching Replication in the Social Sciences* | MZES, Mannheim, Germany

Synthetic data

- **What** they are
- **Why** that matters
- **How** they work
- Possible **uses**

What are synthetic data?

New data(set) that **mimics** original data(set) by preserving **statistical properties** & **relationship between variables.**

Synthetic data

Methodology - Basic concept (Drechsler & Reiter, 2019)

- Idea is closely related to **multiple imputation for nonresponse**
- Generate synthetic datasets by drawing from a **model fitted to the original data**
- **Not the missing values** but the **sensitive values** are replaced with a set of plausible values given the original data
- Generate **multiple draws** to be able to obtain valid variance estimates from the synthetic data

Synthetic data

Properties

- **Replicated** sets from the original data-values
- **Extreme** values
- So: high **general** as well as **specific** utility

Why should I care?

Utility vs. disclosure protection

- **Replication** / **Open Data** / FAIR-principles (Findable, Accessible, Interoperable, Reusable) > Verify results, generate new knowledge, form new hypotheses
- **Problems**:
  - Ethics, privacy, legal, “data-guarding” > common remedies: remove identifiers (tricky); aggregate (not reproducible)
  - **Utility** vs. **disclosure protection**

Why should I care?

Utility vs. disclosure protection with SDL

- **Statistical disclosure limitation (SDL)** techniques for microdata (Drechsler, 2011)
    - **Categorizing** continuous variables
    - **Top coding**: setting values above a certain threshold equal to the threshold
    - **Coarsening categorical variables**: coarsening to a reduced number of categories
    - **Dropping variables**

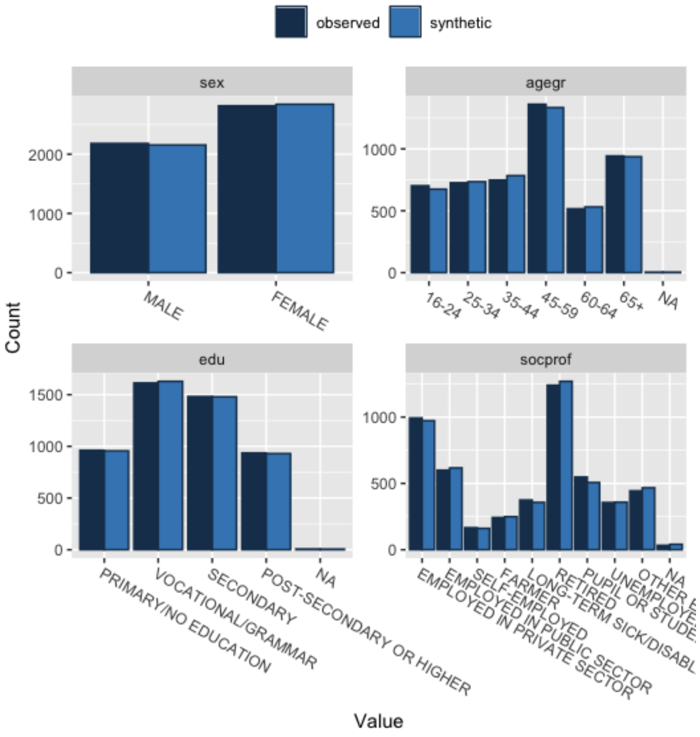
Why should I care?

Utility vs. disclosure protection with synthetic data

- With **synthetic datasets**:
    - Possible to mimic original dataset **statistical properties** and **variable relationships** w/o revealing the underlying original data
    - Avoid **ethics**, **privacy**, **legal** (GDPR, rights) & **data-use** issues

How do I create synthetic datasets?

Method



- `R` package `synthpop` (Nowok et al., 2016)
- Excellent first guide is Quintana (2020), in-depth Drechsler (2011)
- Great overall introduction to topic, status of research and pros and cons of synthetic data in Drechsler & Reiter (2019)

Examples

So, what’s in it for me?

- **Open data support** > possible to make datasets 📈 available even with common constraints attached (ethics, rights, data-use) to original data
- **Visibility** > dataset availability increases visibility 👓 in scientific community 💪
- **Better science** > publishers ❤️ to publish dataset(s) along with papers (even in 🇩🇪 and the social sciences…)

References

Drechsler, J. (2011). *Synthetic datasets for statistical disclosure control: Theory and implementation*. Springer. https://doi.org/10.1007/978-1-4614-0326-5

Drechsler, J., & Reiter, J. (2019). *Synthetic data: Balancing data confidentiality & quality in public use files*.

Nowok, B., Raab, G. M., & Dibben, C. (2016). Synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software*, *74*(11). https://doi.org/10.18637/jss.v074.i11

Quintana, D. S. (2020). A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis-generation. *eLife*, *9*, e53275. https://doi.org/10.7554/eLife.53275