Data quality after disclosure limitation: A density-ratio perspective on utility

Target outlet

- JPC
- JOS
- JRSSA

Introduction

Openly accessible research data can accelerate scientific progress tremendously, by allowing a wide audience of researchers to evaluate their theories and validate existing ones. Additionally, making research data available in combination with analysis code allows others to evaluate and replicate results reported in journal articles, improving the credibility of science. In many circumstances, sharing research data bears a risk of disclosing sensitive attributes of the individuals that comprise the data. In fact, privacy constraints have been named among the biggest hurdles in the advancement of computational social science (Lazer et al. 2009), and among top reasons for companies to not share their data with researchers (Future of Privacy Forum 2017). To overcome these obstacles, data providers can employ a suite of different disclosure limitation techniques before sharing data, for example top-coding, record-swapping or adding noise (e.g., Hundepool et al. 2012; Willenborg and De Waal 2012).

Recently, synthetic data as a means to disclosure limitation has gained substantial traction. The idea of synthetic data is to replace some, or all, of the observed values in a data set by synthetic records that are generated from some model (e.g., Little 1993; Rubin 1993; Drechsler 2011). Essentially, the idea is to approximate the data-generating distribution with a model from which new values are generated. These techniques were originally closely related to methods used for multiple imputation of missing data, such as sequential regression procedures techniques (e.g., Nowok, Raab, and Dibben 2016) or fully conditional specification (Volker and Vink 2021), also knows as Multivariate Imputation by Chained Equations (MICE; van Buuren and Groothuis-Oudshoorn 2011). Recently, the computer science community added a great deal of novel methods for data synthesis based on deep learning techniques (e.g., Xu et al. 2019; Park et al. 2018; Yoon, Drumright, and Van Der Schaar 2020).

@Peter-Paul, vraag voor jou: denk je dat het beter is om ons alleen op synthetische data te richten, of dat we het breder kunnen trekken naar statistical disclosure limitation techniques in het algemeen? Karr et al. (2006) hebben het bijvoorbeeld over utility measures voor disclosure limitation (en nog niet voor synthetische data); misschien kunnen we daarbij aansluiten.

All methods for statistical disclosure limitation alter the data before these are provided to the public. By doing so, the utility of the provided data is always lower than the utility of the original data, because some of the information in the data is sacrificed to protect the privacy of the respondents. The questions that naturally arise are how much information in the original data is actually sacrificed, and how useful the provided data are? Answering this question allows researchers to decide what the altered data can and cannot be used for, and to evaluate the worth of conclusions drawn on the basis of these data. After all, inferences from the altered data are valid only up to the extent that the perturbation methods approximate the true data-generating mechanism. For data providers, a detailed assessment of the quality of the altered data can guide the procedure of altering the data. Statistical disclosure limitation is often an iterative process: some disclosure limitation technique is applied on the data, after which the result is investigated and modifications are made to applied process. Good measures of data quality are essential to determine the appropriate mechanisms used to protect the data, and can help to improve the utility of the data that will be disseminated.

In the statistical disclosure control literature, two different branches of utility measures have been distinguished: specific utility measures and general utility measures. *Add one/two sentences on the merit of visualization when assessing utility of altered data.* Specific utility measures focus on similarity of results obtained from analyses performed on the altered data and the original data. For example, after fitting the same analysis model on both data sets, one can calculate the confidence interval overlap of the estimated parameters (Karr et al. 2006). Alternative measures are ellipsoidal overlap (Karr et al. 2006), which extends to confidence interval overlap to a measure that addresses the joint distribution of all model parameters simultaneous, the standardized absolute difference between estimates (Snoke et al. 2018), and the ratio of estimates for tabular count data (Taub, Elliot, and Sakshaug 2020). As these measures quantify similarity between estimates from analyses performed on the observed and altered data, they are informed only to the extent that data users will recreate those analyses. This can be highly useful if the data is provided for reproducibility purposes (e.g., for third parties to evaluate analysis scripts). However, the goal of distributing the protected data is often to allow researchers to do novel research with the data. In many practical situations, data providers thus have have only limited knowledge on the analyses that will be performed with the altered data. Covering the entire set of potentially relevant analyses is therefore not feasible. If it was, the data providers could simply report the (potentially privacy-protected) results of those analyses performed on the real data, so that access to the (perturbed) data no longer yields additional benefits (for a similar argument, see Drechsler 2022). Additionally, similarity between results on the analyses that have been performed gives no guarantee that the results will also be similar for other analyses. Hence, when determining how useful the altered data is for novel research, specific utility measures are only of limited use.

General utility measures attempt to capture how similar the multivariate distributions of the observed and altered data are. This can be done by, for instance, estimating the Kullback-Leibler divergence between the distributions of the observed and altered data (Karr et al. 2006). An alternative strategy is to try to discriminate between the observed and altered data, as is done with the $pMSE$ (Snoke et al. 2018; Woo et al. 2009). In essence, the $pMSE$ quantifies how well one can predict whether observations are from the observed or the altered data. If better one can do this, the more pronounced the differences between the observed and altered data ought to be. However, various authors have criticized general utility measures for being too broad. That is, important discrepancies between the real and altered data might be missed, and an altered data set that is good in general (i.e., has high global utility) might still provide results that are far from the truth for some analyses (see, e.g., Drechsler 2022). Additionally, it is not straightforward to determine which prediction model to use for calculating the $pMSE$. Specifying a good prediction model in itself may be a challenging task, especially when the number of variables is large. When good models are available, different models, or even different choices of hyperparameters, may yield different results, potentially rendering ambiguity with respect to which altered data set is best. Lastly, the output of global utility measures can be hard to interpret, and say little about the regions in which the synthetic data do not resemble the true data accurately enough. That is, they give little guidance on how the quality of the altered data can be improved.

Section 6: our contribution

Moet nog verder uitgewerkt worden

1. We introduce density ratio estimation to the field of statistical disclosure control. Short remark on the idea that density ratio estimation is a complicated endeavor, especially if the goal is to compare distinct densities. Having to estimate just a single density (ratio) is generally much easier.

2. Note that density ratio estimation can capture specific and general utility measures into a common framework by being applicable on the level of the entire data, but also on the subset of variables that is relevant in an analysis. Additionally, note that confidence interval overlap, ellipsoidal overlap, but also $pMSE$ and Kullback-Leibler divergence, are closely related to density ratio estimation, and can be considered from this perspective.

3. Create a new utility metric based on density ratio estimation (probability with respect to some reference distribution as in permutation testing).

4. Because density ratio estimation can be difficult when there are many variables, we use dimension reduction techniques to capture most of the variability in the data in fewer dimensions on which density ratio estimation can be applied. A by-product of this is that the lower-dimensional subspace allows to create visualizations of deviations from the observed data.

5. Perform a simulation study to give indications about which methods to use (think about how to do this).

6. Implement all this in an R-package

Section 6: outline article

In the next section, we describe density ratio estimation and discuss how this method can be used as to measure utility. Subsequently, we provide multiple examples that show how density ratio estimation works in the context of evaluating the quality of synthetic data. Hereafter, we show in multiple simulations that the method is superior (HOPEFULLY) to current global utility measures as the $pMSE$. Lastly, we discuss the advantages and disadvantages of density ratio estimation as a utility measure.

Methodology

@Peter-Paul: Eventueel een korte beschrijving van data perturbation techniques/synthetic data generation hier. Denk je dat dit wat toevoegt hier?

Section 1: density ratio estimation

In essence, the goal of utility measures is to quantify the similarity between the multivariate distribution of the observed data with the distribution of the altered data. If the used data perturbation techniques, or synthetic data generation models, approximate the distribution of the real data sufficiently, these distributions should be highly similar, and analyses on the two data sets should give similar results. However, estimating the probability distribution of a data set is known to be one of the most complicated challenges in statistics [E.G. Vapnik 1998]. Estimating the probability distribution for both observed and altered data can lead to errors in both, artificially magnifying discrepancies between the two. Hence, subsequent comparisons will be affected by these errors. The procedure can be simplified by using density ratio estimation, because this only requires to estimate a single density.

Introduce density ratio estimation as a utility measure. What does this measure mean/how to interpret it. How to make decisions based on this measure.

Say something on whether (and if so, how) categorical variables can be incorporated as well.

Section 3: theoretical comparison with conventional approaches for general utility assessment

Relate density ratio estimation to specific and general utility measures. Pick one/two specific utility measures and relate these to density ratio estimation (ratio of estimates seems straightforward, as well as confidence interval overlap and ellipsoidal overlap).

Relate density ratio estimation to $pMSE$ and KL divergence (to some extent, both are generalizations of density ratio estimation, or at least are conceptually similar). Give some more information on the $pMSE$, describe what it shortcomings are. The quality of the $pMSE$ highly depends on the model used to calculate the propensity scores. Perhaps give an example of logistic regression, which basically estimates whether the conditional mean of the observed and altered data is the same. Explain how density ratio estimation can overcome the shortcomings of the previously mentioned methods.

Section 4: Dimension reduction for utility

The difficulty of density ratio estimation increases with the dimensionality of the data. Therefore, we follow previous recommendations to incorporate dimensionality reduction techniques in density ratio estimation.

Shortly name examples of dimensionality reduction techniques (i.e., PCA; LFDA or UMAP).

A useful by-product of dimension reduction is that it allows to create visualizations, and these visualizations can be used to get more insight in discrepancies between observed and altered data. Show what such visualizations can look like, and how they can help.

Simulations

Small illustration / example with multivariate Gaussian distributions.

1. Simple, multivariate normal simulation (e.g., two correlation structures, two sample sizes, so 2 × 2 full factorial design); basically similar to what we did already.

More complex simulation, more variables, non-linearities, perhaps using real data.

2. More advanced simulation (e.g., some non-linearities, different sample sizes)

Have to think about this in more detail still.

Real data example

Clinical records heart-failure data? Misschien ook niet, nog over nadenken.

Exemplify how utility measures could (should!) be used to improve the quality of the altered data (e.g., illustrate how models can be adjusted iteratively based on utility assessment).

Some notes to self

Current ways to assess the utility?

- pMSE - logistic, regression, CART models (Snoke, Raab, Nowok, Dibben & Slavkovic, 2018; General and specific utility measures for synthetic data AND Woo, Reiter, Oganian & Karr, 2009; Global measures of data utlity for microdata masked for disclosure limitation)

- Kullback-Leiber divergence (Karr, Kohnen, Oganian, Reiter & Sanil, 2006; A framework for evaluating the utility of data altered to protect confidentiality).
- According to multiple authors, both specific and general utility measures have important drawbacks (see Drechsler Utility PSD; cites others). Narrow measures potentially focus on analyses that are not relevant for the end user, and do not generalize to the analyses that are relevant. Global utility measures are generally too broad, and important deviations in the synthetic data might be missed. Moreover, the measures are typically hard to interpret.
- See Drechsler for a paragraph on fit for purpose measures, that lie between general and specific utility measures (i.e., plausibility checks such as non-negativity; goodness of fit measures as χ^2 for cross-tabulations; Kolmogorov-Smirnov).
- Drechsler also illustrates that the standardized $pMSE$ has substantial flaws, as the results are highly dependent on the model used to estimate the propensity scores, and unable to detect important differences in the utility for most of the model specifications. Hence, it is claimed that a thorough assessment of utility is required.

Methodology

TO DO

Simulations

TO DO

Real data example

TO DO

Results

TO DO

Discussion and conclusion

TO DO

Drechsler, Jörg. 2011. *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*. Springer. https://doi.org/10.1007/978-1-4614-0326-5.

———. 2022. "Challenges in Measuring Utility for Fully Synthetic Data." In *Privacy in Statistical Databases*, edited by Josep Domingo-Ferrer and Maryline Laurent, 220–33. Springer International Publishing. https://doi.org/10.1007/978-3-031-13945-1_16.

Future of Privacy Forum. 2017. "Understanding Corporate Data Sharing Decisions: Practices, Challenges, and Opportunities for Sharing Corporate Data with Researchers."

Hundepool, Anco, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer, and Peter-Paul De Wolf. 2012. *Statistical Disclosure Control*. John Wiley & Sons. https://doi.org/10.1002/9781118348239.

Karr, A. F, C. N Kohnen, A Oganian, J. P Reiter, and A. P Sanil. 2006. "A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality." *The American Statistician* 60 (3): 224–32. https://doi.org/10.1198/000313006X124640.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, et al. 2009. "Computational Social Science." *Science* 323 (5915): 721–23. https://doi.org/10.1126/science.1167742.

Little, Roderick J. A. 1993. "Statistical Analysis of Masked Data." *Journal of Official Statistics* 9 (2): 407–7.

Nowok, Beata, Gillian M. Raab, and Chris Dibben. 2016. "synthpop: Bespoke Creation of Synthetic Data in R." *Journal of Statistical Software* 74 (11): 1–26. https://doi.org/10.18637/jss.v074.i11.

Park, Noseong, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. 2018. "Data Synthesis Based on Generative Adversarial Networks." *arXiv*. https://doi.org/10.48550/arXiv.1806.03384.

Rubin, Donald B. 1993. "Statistical Disclosure Limitation." *Journal of Official Statistics* 9 (2): 461–68.

Snoke, Joshua, Gillian M. Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. 2018. "General and Specific Utility Measures for Synthetic Data." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 181 (3): pp. 663–688. https://www.jstor.org/

stable/48547509.

Taub, Jennifer, Mark Elliot, and Joseph W Sakshaug. 2020. "The Impact of Synthetic Data Generation on Data Utility with Application to the 1991 UK Samples of Anonymised Records." *Transactions on Data Privacy* 13 (1): 1–23. http://www.tdp.cat/issues16/abs.a306a18.php.

van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in r." *Journal of Statistical Software* 45 (3): 1–67. https://doi.org/10.18637/jss.v045.i03.

Volker, Thom Benjamin, and Gerko Vink. 2021. "Anonymiced Shareable Data: Using Mice to Create and Analyze Multiply Imputed Synthetic Datasets." *Psych* 3 (4): 703–16. https://doi.org/10.3390/psych3040045.

Willenborg, Leon, and Ton De Waal. 2012. *Elements of Statistical Disclosure Control*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4613-0121-9.

Woo, Mi-Ja, Jerome P. Reiter, Anna Oganian, and Alan F. Karr. 2009. "Global Measures of Data Utility for Microdata Masked for Disclosure Limitation." *Journal of Privacy and Confidentiality* 1 (1). https://doi.org/10.29012/jpc.v1i1.568.

Xu, Lei, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. "Modeling Tabular Data Using Conditional GAN." In *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/254ed7d2de3b23ab10936522dd547b78-Paper.pdf.

Yoon, Jinsung, Lydia N Drumright, and Mihaela Van Der Schaar. 2020. "Anonymization

Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN).” *IEEE Journal of Biomedical and Health Informatics* 24 (8): 2378–88. https://doi.org/10.1109/JBHI.2020.2980262.