Journal of Official Statistics, Vol. 30, No. 3, 2014, pp. 563–565, http://dx.doi.org/10.2478/JOS-2014-0035

Book Review

Peter-Paul de Wolf[1]

Jörg Drechsler. *Synthetic Datasets for Statistical Disclosure Control, Theory and Implementation.* 2013. NY: Springer, ISBN 9-781461-403258, $99USD.

Nowadays, (national) statistical institutes increasingly feel pressure from the research community to provide datasets with high utility. Researchers want to be able to perform their analyses on the rich datasets that are available at those statistical institutes. The use of tabular data, traditionally provided by statistical institutes, often no longer suffices for their research purposes. They prefer to have microdata available at their own desk to which they can then apply their analyses.

Even though national statistical institutes are often quite willing to meet the wishes of the researchers to some extent, their national statistical laws are often a complicating factor. Among other things, these laws oblige national statistical institutes to safeguard the confidentiality of their data providers' responses. Keeping the response confidential also helps to build the trust that is needed between statistical institutes and their respondents. Only when institutes are trustworthy are respondents willing to provide detailed and sensitive information.

This book, based on a PhD thesis, deals with a method of producing detailed microdata, maintaining utility whilst respecting the confidentiality issues just mentioned. The book deals specifically with a method for producing synthetic datasets, using multiple imputation techniques.

In Chapter 2, the author discusses the history of the use of multiply imputed datasets within the statistical disclosure control setting. Multiple imputation is an approach wherein multiple datasets are created, each with newly imputed missing values. Multiple imputation retains the advantages of imputation, while allowing the uncertainty due to imputation to be directly assessed. Multiply imputed datasets are also used in the setting of nonresponse. At present the method is used more often in the United States than in Europe in both of these settings. Chapter 3 acts as a background chapter on multiple imputation techniques. This is very convenient as it allows readers to follow the rest of the book with ease.

In Chapter 4, a specific dataset is discussed that is used throughout the book as the major example: the German IAB establishment panel. This panel is based on the German employment register and is used to produce official statistics. It is considered one of the most important business surveys in Germany and is very popular among researchers. Chapter 5 is used to show how multiple imputation can be used to address nonresponse, and at the end of the chapter it is explicitly applied to the German IAB establishment

[1] Department of Process Development, IT and Methodology, Statistics Netherlands. Email: pp.dewolf@cbs.nl.

panel. It is shown how estimates can be constructed based on multiple datasets, each being a separate instance of applying the same imputation technique to the original dataset.

Chapter 6 deals with the fully synthetic dataset approach. This means that for a sample, all variables of all nonsampled units in the sample frame are imputed. Then multiple samples are drawn from that fully imputed sample frame. As a special case, one might even go further, fully synthesizing all units in the sample frame and "only" using the original sample to construct a proper imputation model. That way, no original "real" data is left in any of the sampled units.

At present no agency has adopted the fully synthetic approach. One version adopted called partially synthetic datasets is discussed in Chapter 7: here, only some of the variables are replaced using multiple imputation techniques. Although the author speculates that the variables being replaced could be sensitive variables as well as key identifiers, the disclosure risk measures he deals with in this chapter are only related to identification disclosure. Identification disclosure refers to the situation that a single record in a dataset can be linked to a known individual, that is, that individual can be identified in the dataset. These risk measures are thus only influenced by replacing values of key identifiers. At the end of this chapter, the author discusses some pros and cons of fully synthetic datasets versus partially synthetic datasets. The main conclusion is that fully synthetic data sets are harder to produce, but reduce the disclosure risk more effectively. On the other hand, partially synthetic datasets often have a higher utility because fully synthetic datasets are completely determined by the imputation model, whereas partially synthetic datasets still contain 'original' (nonsensitive) data. Moreover, partially synthetic datasets are usually easier to produce.

The first seven chapters deal with multiple imputation in relation to disclosure control. However, imputation techniques are also used to correct for nonresponse. In Chapter 8, the author discusses a way to combine multiple imputation techniques to correct for nonresponse and to limit the disclosure risk at the same time. This is applied to partially synthetic datasets only. The major part of this chapter is devoted to the IAB establishment panel example. At the end, an interesting issue is touched upon. It is stated that "Since the intruder never knows if her match is correct . . . the data are well protected from these kinds of attacks." This poses the question of how well a dataset is protected by uncertainty about the information supposedly disclosed. Indeed, in some cases "disclosing" untrue information might do more harm than disclosing true information about an individual unit.

In the case of multiple imputation, multiple datasets (m) are being released. The larger the m, the higher the utility of the dataset (the additional variance introduced by the imputation decreases with the number of released datasets), but at the same time the higher the disclosure risk. In Chapter 9, a two-stage imputation process is discussed to deal with the trade-off between utility and risk. Finally, in Chapter 10, some arguments are given to promote the use of multiple imputation techniques in deriving synthetic datasets, over 'standard' SDC techniques, such as local suppression, global recoding or top-coding. These arguments try to deal with the scepticism about the method, the tendency to stick with 'known' methods and the reluctance to use new methods before they are implemented in known statistical software.

This book gives a good overview of recent developments within the area of multiple imputation as a technique of deriving synthetic datasets. It is not an easy book, however.

The notation throughout the book is not always consistent and contains some minor typos. Specifically, Chapters 8 and 9 have some formulas that are very much alike, but are difficult to compare because slightly different notation is used. This makes reading the book and using it for reference somewhat difficult. On the other hand, the use of a single example throughout all chapters (the IAB establishment panel) is very beneficial. It shows the effects of the different methods on the same "real-life" dataset.

In the area of statistical disclosure control, the term "transparency" has received a lot of attention again recently. This term is related to the advantages and disadvantages of revealing information about the statistical disclosure control methods used to produce safe data. In this book, it is evident that transparency is crucial to improve the use of synthetic datasets. Obviously, not all information of the imputation model needs to be released along with the synthetic datasets. However, some information, for example which variables are included in the imputation model, can be used by a researcher to determine whether his or her analysis is still likely to be valid.

Transparency should also apply to the methods themselves. This book is a good example of providing insight into the methods that can be used to produce datasets that are useful to researchers while at the same time limiting the disclosure risk.

The views expressed in this review are those of the author and do not necessarily reflect the policies of Statistics Netherlands.