UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL, HUMAN AND MATHEMATICAL SCIENCES

Social Statistics and Demography

On the use of hierarchical models for multiple imputation and synthetic data generation

by

Sana Rashid

Thesis for the degree of Doctor of Philosophy

Supervisors: Dr. Robin Mitra, Dr. Nikolaos Tzavidis

Examiners: Dr. Peter Smith, Dr. Enrico Fabrizi

April 4, 2017

To Baba and Mama

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL, HUMAN AND MATHEMATICAL SCIENCES
Social Statistics and Demography

Thesis for the degree of Doctor of Philosophy

ON THE USE OF HIERARCHICAL MODELS FOR MULTIPLE IMPUTATION AND SYNTHETIC DATA GENERATION

by Sana Rashid

Missing data are often imputed with plausible values when various analyses are performed. One popular approach employed to impute data is multiple imputation, which requires specification of a suitable imputation model. This thesis investigates the impact on multiply imputed hierarchical datasets when the imputation model is misspecified. The first issue studied is the presence of omitted variable bias. The same issue is then studied with a focus on the use of multiple imputation for creating synthetic data to protect data confidentiality. Here, the quality of multiply imputed datasets is studied not only through performance of various analysis models, but also, risks of disclosure for sensitive data. With the help of simulation studies and a longitudinal dataset from establishments in Germany, the detrimental effect of such model misspecification is evaluated, and recommendations are made for users of multiple imputation for both missing and synthetic data. The second issue investigated is model misspecification due to incorrect modelling of the shape of the error term. Existing methods for robust regression and alternatives to the normal distribution are compared within the synthetic data context only. Results from simulation studies and data on household wealth in the UK are used to identify appropriate methods for multiple imputation in such a scenario.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Sana Rashid , declare that the thesis entitled *On the use of hierarchical models for multiple imputation and synthetic data generation* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission

Signed:.....

Date:*04/04/2017*.....

Acknowledgements

Firstly, thanks to Allah; His innumerable bounties surprise me each day. I also wish to thank my supervisors, Robin and Nikos for their care, support and expert advice throughout the period of this research. A very special thanks to my collaborator, Joerg Drechsler, who also acted as a guide, friend and mentor on numerous occasions. This thesis could not have been completed without his support.

The ESRC DTC at the University of Southampton has also played a vital role in the completion of this thesis. From their team, I thank Glenn, Amos and Pauline, for the various ways in which they have helped me develop as a researcher and tackle many obstacles on the way. Additionally, many thanks to the Institute for Employment Research (IAB), Germany for hosting my visit to their office for the use of data extracted from the IAB Establishment Panel. Likewise, thanks to the Office for National Statistics for providing access to data from the Wealth and Assets survey online. I would also like to extend my gratitude to my examiners, Dr Peter Smith and Dr Enrico Fabrizi, for their valuable comments on this research and the thesis.

A very special thanks to my parents, Rashid and Yasmeen, and sisters, Nida and Hira for their love, motivation and belief in me, that helped me progress throughout my life and continues to do so. I would like to express my most sincere gratitude to many friends whose support has been crucial at various stages of the PhD, the list of whose names may not be possible to produce. To mention a few, thanks for the love from Tayyaba, Wasifa, Zunaira, Wanpen, Emily, Sean, Mark, Melike, Dilek and many others with whom I have shared accommodation, office space and grievances.

Finally, I will be eternally grateful to Dr Kennedy and Dimitris, both of whom acted as my backbone on countless occasions within the last few years; the importance of their role in both the completion of this thesis and my life cannot be described in words.

Chapter 1

Introduction

Data are often used to make inferences regarding a population using a sample of observations. However, when some observations are missing, the inferential procedure may not always be carried out as for fully observed data. Researchers commonly use adjustments to incorporate uncertainty arising from the missing data within their inferences. One such popular approach is multiple imputation (MI). Using MI, one may impute the values of the missing observations, given certain assumptions, several times. This results in multiple copies of the same dataset with different imputed values. Final inferences are then made on each of the completed datasets, and combined using certain rules.

When using MI, data are imputed under the assumption that a certain model describes the data. This model is used to create the imputations, and is called the imputation model. The imputation model may or may not be complex. No matter what the imputation model, there is always a chance that the assumed model is misspecified for the given data. Subsequently, analyses carried out on data imputed using a misspecified model may be unreliable. With real data, it is not possible to define an imputation model that is universally suitable for any type of analysis carried out on the completed datasets. Therefore, imputation models are assessed according to their suitability to certain common types of analyses. Where possible, researchers must use the most general imputation model, i.e., that which helps analysts to produce valid inferences for a variety of analysis models.

When data contain certain correlation structures, the imputation model must take these structures into account as well. An example is clustered data, such as data collected from students nested within classrooms, or data collected on the same individuals over time. Analysts often make assumptions regarding the correlation structures present in data clustered in either of these two ways. In the first part of the thesis, we explore the models available for the imputation of such hierarchical data. We then study these models in combination with various analysis models and identify issues when the imputation and

analysis models make different assumptions regarding the correlations present in the data.

In the second part of this thesis, we study the same issue within the context of protection of confidential data. Open access to data is a contentious issue. Privacy concerns discourage data keepers to release data for public use, but this hinders research and innovation. Data collecting multinational organisations and various governments have put efforts to provide the data they collect for free use. For instance, the Organisation for Economic Co-operation and Public Development (OECD) decided to make all the data it collected from public funds required to be publicly available. Other well-known sources of open data include the World Bank and the US government. The UK government also accelerated its efforts in the area with the launch of the *data.gov.uk* website in 2010. A comprehensive list of open data sources is available through the *datacatalogs.org* project.

The concept of open data entails the free use, re-use and redistribution of data for any commercial or non-commercial purposes - the exact attributes that are in conflict with data confidentiality requirements. Private information on individuals and valuable internal data from organisations are usually collected under confidentiality pledges, which restricts public release of such data. There are several benefits of open access to data. Freely available data can open channels for communication and collaboration between various groups, encourage public participation and innovation, assist academics in generating knowledge, and help governments ensure transparency and improve political decision-making. Yet data confidentiality limits the extent to which such benefits can be reaped. Even a single disclosure can disrepute the data collection agency and the data collection process, discouraging further participation in public surveys and access to data for other researchers. The UK legislation, as in the US and other European countries, handles this complex issue through the Freedom of Information Act 2000, Data Protection Act 1998 and Public Records Act 1958, tackling the delicate balance between open access to data and the right to privacy.

Confidential data may be protected by providing restricted access to users. Under this arrangement, researchers are allowed to travel to designated data centres, where they can run analyses on specific data on site, or remotely, by providing the software code to the agency. The output that is allowed to be released to the user is checked by authorities to ensure that there is a limited chance of disclosure of confidential information. Services like these are normally available for a charge, are inconvenient for the researcher, and a logistical burden for the data centre itself.

A possible alternative is the release of microdata to the public. Statistical disclosure control (SDC) methods are applied to alter the real data before public release. Some ad-hoc measures of SDC include swapping observations, aggregating records and recoding sensitive variables. Recently, it has been proposed that multiple imputation (MI) methods can also be used for SDC. The idea is to replace sensitive observations in a given

data set by synthetic values generated from a plausible model. The key requirement from this model is that it is able to preserve relationships between the various variables in the data set. Multiple copies of the synthetic data can then be released to the public. This thesis focuses on using MI for SDC for a multilevel data set. When applying SDC measures, one of the main question to be addressed is the balance between the quality of the synthetic data and the disclosure risks associated with it.

Statistical disclosure occurs if confidential information about units in a dataset is disclosed using data released for public use. Statistical Disclosure Control (SDC) is the research area covering limitation of such misuse of data. It does not cover disclosure by other means, such as stolen databases. The two main branches of SDC are 1) methods used to modify data before release and 2) the assessment of disclosure risks for these methods and particular datasets. Hundepool et al. (2012) and Willenborg and De Waal (2012) give a comprehensive overview of SDC methods and disclosure risk assessment. Techniques such as categorising continuous variables, making sensitive values missing, or replacing sensitive values with imputed values, releasing only a sample of values from a given dataset, adding noise to the existing data and swapping observed values may be used to protect confidential microdata. Alternative techniques may be used for tabular data, which is beyond the scope of this project. Naturally, any perturbation has direct implications for analyses run on such a dataset. Data perturbation techniques under SDC are therefore developed, although motivated by protection of sensitive data, with a focus on preservation of final analyses.

Disclosure limitation for sensitive data has traditionally focussed on tabular data as information collected by authorities was only ever released as tables. An example would be the disclosure assessment for data released from the US Economic Census (Jewett, 1993). A comprehensive overview of early tabular data SDC methods can be found in Cox et al. (1986a,b). Nowadays, collection of microdata through governmental and non-governmental organisations is more common. Information collected from social media, store cards and mobile phone operators are common examples of such data. As microdata is more likely to contain personal information, confidentiality concerns have now become more serious. There are two main ways to protect privacy for microdata. One is to simply not release the sensitive information, for instance, remove selected variables or observations from a released dataset. Another way is to modify the released observations, for instance by, censoring observations at certain maximum or minimum values or categories. The advantage of the latter approach is clear; the remaining data is released unaltered, but overall, there is some loss of information. This information loss, however, can be harmful for data analyses. Ad hoc data altering methods may not only affect the shape and characteristics of certain variables, but also their relationship with other variables in the dataset.

The perturbation approach is used to modify the data to help protect data confidentiality while preserving some characteristics of the data themselves. We briefly discuss some of

the earliest methods proposed to modify data in the literature. One of the perturbation techniques is data swapping, i.e. swapping records for observations between individuals; see Delanius (1984) and Delanius and Reiss (1982) for the earliest literature on data swapping. Swapping may be done individually for variables or in a multivariate fashion, referred to as shuffling (Muralidhar et al., 2006). Standard univariate swapping can have detrimental effects on data quality as associations between variables may be distorted. To ensure a better quality of data, only a small proportion of the dataset may be subject to swapping. In shuffling, the entire mean and covariance structure in a dataset or sub domains may be preserved. However, the approach does not aim to preserve other characteristics of the data, such as higher order moments, data structures or design effects.

Another standard technique to perturb data is adding noise to the observations. This is added as a function of the original data and it may still negatively effect the relationship between the original and noisy data. The method inevitably increases variability in each perturbed variable and thus, any overall analysis. The noise added can be additive or multiplicative. Additive noise may not be best for variables with skewed distributions which may not be preserved in this fashion. While multiplicative noise may solve this problem, it still may lead to deteriorated relationships between variables, which is undesirable. Nevertheless, noise addition may also be designed to preserve means, variances and covariances between important variables and marginal cumulative distributions (Kim, 1986; Sullivan and Fuller, 1989).

Another popular technique is called microaggregation. The basic idea of the method is to preserve totals within groups of observations created according to various criterion. Once the groups have been defined, observations within each group may be modified such that the total value of the variable in the group does not change. Naturally, the smaller the groups, the better the data utility. In a simple approach, if observations within a group are replaced by the mean of the group, variability in the perturbed data set will be lower than in the original data. Also, the relationships between variables may be distorted. Detailed information on these techniques and many others can be found in Willenborg and De Waal (2012). Many of these methods are still used in practice (Cleveland et al., 2012) because of mistrust in more complicated perturbation techniques to protect data utility and confidentiality.

Ad-hoc methods for SDC are known to bias analyses run on the perturbed data set. Winkler (2007) and Purdam and Elliot (2007) illustrate the detrimental effects of traditional SDC methods on applied analyses. Besides these methods, techniques from the Computer Sciences literature may also be used to ensure privacy. These methods, such as the privacy-preserving data publishing technique (Fung et al., 2010), can guarantee formal levels of privacy but do not focus on data utility. SDC methods on the other hand are motivated by the preservation of statistical inferences.

Releasing synthetic data is one of the most recent solutions designed to address the utility-privacy trade-off. Rubin (1993) proposed using MI as an SDC technique. The advantage of this approach over the earlier masking approaches is the ease of computation on the analyst's part. After MI, an analyst only needs to run their own analysis as usual without requiring special adjustments because of the perturbation techniques. In the following sections, we review the literature in MI methods for both missing data and SDC.

1.1 Literature Review

In this section, we review the literature most relevant to the research documented in this thesis. The focus is on multiple imputation (MI) for hierarchical data for both missing and synthetic data scenarios.

1.1.1 Multiple imputation framework

MI (Rubin, 1987) for missing data aims to account for the variability in the analysis of incomplete data by explicitly modelling the variability coming from the missingness in the dataset. Let D represent an $n \times p$ data set with p variables, some incompletely observed and let D_{mis} represent the units with missing observations in the data set and D_{obs} represent the observed data.

We first introduce the concepts of *MCAR*, *MAR* and *MNAR*. Let R be an $n \times p$ matrix of 0's and 1's indicating whether an observation is observed or not, where 0 represents a missing value. We assume that each observation has a probability of being missing, and the process that governs these probabilities is called the *missing data mechanism* and the model that describes the mechanism is called the *missing data model*, $P(R = 0|D_{obs}, D_{mis}, \psi)$. If the probability of being missing is same for all cases, i.e. $P(R = 0|D_{obs}, D_{mis}, \psi) = P(R = 0|\psi)$, then the data are said to be MCAR. If this probability depends on observed data instead, i.e. $P(R = 0|D_{obs}, D_{mis}, \psi) = P(R = 0|D_{obs}, \psi)$, then the data are said to MAR. For MNAR, the probability depends on unobserved observations as well, i.e. $P(R = 0|D_{obs}, D_{mis}, \psi)$. The parameters ψ govern the values in R. The missing data mechanism is said to be ignorable if the distribution of ψ is independent of the distribution of other scientific quantities of interest belonging to the complete data. Finally, the matrix R also represents the *missing data pattern*. The missing data pattern is called *monotone* if the p variables in D can be arranged in such a way that if a variable Y_j is missing, then all variables Y_k with $k > j$ are also missing.

In a typical MI process, a researcher will attempt to approximate the posterior predictive distribution of the missing data given the observed data, $P(D_{mis}|D_{obs})$. From this distribution, the missing values are predicted and imputed several, m, times; therefore

completing the data set m times. The m completed datasets can then be used for analysis. An analysis involving a quantity of interest, Q, such as means or regression coefficients, has to be repeated on the m completed datasets and the results combined using the following simple rules:

$$\bar{Q} = \frac{1}{m}\sum_{k=1}^{m}\hat{Q}_k, \tag{1.1}$$

$$\begin{aligned} T &= \frac{1}{m}\sum_{k=1}^{m}U_k + \frac{m+1}{m(m-1)}\sum_{k=1}^{m}(\hat{Q}_k - \bar{Q})(\hat{Q}_k - \bar{Q})' \\ &= \bar{U}_m + \frac{m+1}{m}B, \end{aligned} \tag{1.2}$$

where, $\hat{Q}_k$ is the estimate for Q from the k^{th} completed data set, U_k is the estimate for the variance of $\hat{Q}_k$ from the k^{th} completed data set, and B is the variance of $\hat{Q}_k$ for $k = 1, ..., m$.

The total variance estimate, T is a sum of the within and between imputation variances. The added factor of $\frac{B}{m}$ corrects the estimate of T for a finite number of imputations. These combining rules have been formulated using certain approximations described by Rubin (1987) (Chap. 3), and under the assumption that the posterior distribution of Q is asymptotically Normal. Inferences regarding Q can be made using a Student's t distribution with degrees of freedom, $\nu = (m-1)(1+r_m^{-1})^2$, where $r_m = \frac{(1+m^{-1})B}{\bar{U}}$. Barnard and Rubin (1999) proposed a small sample degrees of freedom, $\nu_2 = \left(\frac{1}{\nu} + \frac{1}{\frac{(a+1)(a)}{a+3}\cdot\frac{\bar{U}}{\bar{U}+(1+1/m)B}}\right)^{-1}$, where a is the complete-data degrees of freedom. We also note that the combining rules have been derived assuming that the Bayesian process of synthesis uses noninformative priors.

For data that have been partially synthesised, Reiter (2003) proposed a modified variance estimator:

$$T^* = \bar{U}_m + \frac{B}{m}. \tag{1.3}$$

As in the missing data context, $\bar{U}_m$ estimates the within variance of Q, while $\frac{B}{m}$ is the added variability in Q due to a finite number of imputations. The additional variability B is not required in partially synthetic data, as this is to accommodate the uncertainty in the nonresponse mechanism. For partially synthetic data, data to be synthesised are selected by the data keeper, and the process is not considered to be stochastic. After combining the results from the multiple copies, further inferences for Q can be based on a Student's t distribution with degrees of freedom, $\nu^* = (m-1)\left(1+\frac{\bar{U}_m}{B/m}\right)^2$.

1.1.2 Multiple imputation for synthetic data

Multiple imputation for synthetic data follows the procedures designed for multiple imputation to handle missing data (Rubin, 1987) by considering the sensitive data the 'missing' values. The idea is to fit a model to the sensitive data given the full dataset, repeatedly draw new synthetic observations from this model and make m completed datasets. These m datasets can then be released to the public.

Data may be *fully* or *partially* synthesised. Rubin (1993) suggested the use of multiple imputation (MI) to create fully synthetic data. Rubin's idea was to define an appropriate model that describes a complete data set and generate synthetic values from this model to produce multiple synthetic datasets. Raghunathan et al. (2003) and Reiter (2005a) further developed the MI theoretical framework supporting inferences from a fully synthetic dataset. Applications illustrating the approach can be found in Reiter (2004a) and Drechsler et al. (2008b). A similar approach was proposed by Fienberg (1994) and extended for categorical variables by Fienberg et al. (1998). The American Community Survey has also been subject to another fully synthetic data approach (Rodríguez and Creecy, 2007), with extensions to preservation of small area statistics by incorporating detailed geographical information within the SDC technique (Sakshaug and Raghunathan, 2011).

Little (1993) developed the idea further, proposing that only sensitive observations or variables need to be synthesised. Reiter (2003, 2005a) provided basis for inferences using partially synthetic data. Partially synthetic data have been developed in the US (Abowd et al., 2006; Hawala, 2008; Kennickell, 1997) and Germany (Drechsler, 2012).

The advantages of fully synthesising data are that no real observations are released and this provides better confidentiality protection to the dataset. An intruder will therefore be discouraged to identify information about units on an individual level. For an analyst, this means that all variables can be released to the public, aiding a variety of analyses. Nevertheless, fully synthetic data heavily depends on the correctness of the synthesis model, which may also be very difficult to specify, given the full range of variables in the dataset. Moreover, it is not always required to synthesise all variables in a dataset. The concept of partially synthetic data recognises this and, therefore, proposes synthesising only the variables and observations that may be considered sensitive or lead to disclosure of sensitive information. This implies that partially synthetic data can potentially offer better analytical validity as original observations in the dataset may be retained. Also, data utility, is then, relatively less dependent on the quality of the synthesis model and the models themselves are comparatively easier to define. The approach, however, calls for more detailed investigation into possible disclosure risks as these increase because of the presence of original data in the released versions. Drechsler et al. (2008a) gives a detailed comparison for both the approaches along with their associated data utility and disclosure risks. For both types of synthesis, literature also guides the synthesis process

when facing missing data in the original dataset. For more details, see Reiter (2004b) and Kinney and Reiter (2010).

Some other synthesis procedures have also been suggested in the literature. Reiter and Drechsler (2010) suggest a two-stage synthesis process. In the two-stage process, the idea is to create more synthetic copies for some variables than for others, as having more copies of synthetic data may increase disclosure risks (Mitra and Reiter, 2006). This is the direct result of increasing data utility with increasing number of multiply synthesised datasets (Reiter, 2003). The motivation behind the two-stage process is the ability to control the trade-off between data utility and disclosure risks. The approach requires its own new combining rules as derived by Reiter and Drechsler (2010).

Another synthesis scheme using MI is called *sampling with synthesis*. The approach was proposed by Drechsler and Reiter (2010) specifically for Census data. This is an extension to the idea that only samples may be released from an original dataset to protect confidentiality. The idea is to synthesise the full dataset using MI and then release samples from the multiple synthetic datasets. Two approaches to select samples for release are proposed. For the first one, the same samples of observations are released from each of the multiple datasets; for the second approach, different samples may be selected. There is a trade-off between data utility and disclosure risks between the various sampling schemes as well. Further rules to create inferences after synthesis and more details about the approach can be found in Drechsler and Reiter (2010). An extension to the idea was provided by Drechsler and Reiter (2012), applying the approach to sample data rather than a Census.

MI for nonresponse and MI for SDC differ in a few aspects. In MI for missing data, the procedure by which the data have gone missing, i.e. the missing mechanism, has important implications for the imputation process. For nonresponse, the missingness mechanism can be ignored as long as it is *missing at random*, MAR. For fully synthetic data MAR always holds, while for partially synthetic data the 'missingness' can be justified as MAR by using all observations in the dataset for modelling purposes. We also consider missingness patterns. For fully synthetic data, the 'missingness' pattern is monotone by definition. For partially synthetic data, the same records are usually chosen for synthesis for the different variables, and a monotone pattern is achievable. Therefore, several iterations for imputation, here synthesis, are not required and there is no need to monitor convergence.

Moreover, for partially synthetic data, posterior draws for the model parameters are not necessary. In the nonresponse case, model parameters need to be estimated through posterior distributions as only part of the dataset is observed. This is not required for the partially synthetic data case, where the dataset is fully observed and the parameters can be directly estimated. This implies that the parameter estimates from the observed sample, such as those obtained using maximum likelihood methods, may directly be

used to generate partially synthetic data. Reiter and Kinney (2012) establish this both analytically and empirically.

Nevertheless, MI for SDC shares some challenges with MI for missing data. There is still a need to define an appropriate imputation, here synthesis, model. Moreover, the risk of uncongeniality (Meng, 1994), i.e. potential bias in the data analysis stage caused by the use of synthesis models different from the analysts' models, still needs to be addressed. An added consideration, particular to MI for SDC is the risk of disclosure of sensitive data. This adds another layer of requirement to the synthesis process, i.e. ensuring the selection of observations to synthesise and the synthesis models do not help intruders obtain sensitive information, usually by means of identifying particular units or individuals in the released dataset.

Literature from both MI for missing data and SDC methods can be used for our research on MI for SDC. In particular, modelling strategies from the extensive MI for missing data literature can be utilised to define appropriate models, and disclosure risk measures from the SDC literature can be used to assess the disclosure risks arising from the particular synthesis models. The aim of this research is to produce high quality multilevel synthetic data with the ability to handle various types of analyses while protecting the identities of units in the data set. Below, we discuss existing research around modelling and uncongeniality issues relevant to the current research topic, as well as measurement of disclosure risks.

1.1.3 Modelling

There are two distinct modelling strategies in the literature that can be used to describe a particular dataset. A joint modelling approach assumes that data can be described by a multivariate distribution. An appropriate multivariate distribution, conventionally the multivariate normal, can then be fitted to the data and used to predict new values to replace the observations. The alternative approach aims to define, instead, a sequence of univariate distributions that can be used to impute/synthesise one variable at a time.

Proposals for joint models to describe datasets include the multivariate normal, the log-linear and the general location model; details for these approaches can be found in Schafer (1997a). If all variables are categorical, a Dirichlet process of mixtures of products of multinomials (DPMPM) can provide the framework for the joint model (Si and Reiter, 2013).

For hierarchical structures, joint random effects models for imputation were first proposed by Liu et al. (1995) with a focus on longitudinal data. Applications of this model are found in Coffey et al. (2003) and Mumtaz et al. (2007). Schafer (1997b) extended this model to allow for missing covariates. Subsequently, Schafer and Yucel (2002) modelled multivariate missing variables under a multivariate mixed-effects model. Yucel

(2008) illustrates application of such a model for imputation and Yucel et al. (2008) extend the strategy to incorporate multiple classification and membership issues in the model. Yucel and Demirtas (2010) study the implications of imputing under a random effects model with random effects assumed to be normally distributed incorrectly. Yucel (2011a) further extended the random effects model to incorporate random covariances into the imputation strategy.

With regards to software, the R package *pan* by Zhao and Schafer (2013) provides the multivariate approach to handling multilevel missing data, provided missingness only occurs at the first level of the model. MLwiN (Browne, 2009) also offers the procedure, with added features that can handle missingness at various levels of the model and for categorical variables. A review on software available for MI is provided by Yucel (2011b).

On the other hand, the approach using a sequence of multivariate models has been proposed by Raghunathan et al. (2001) under the name, Sequential Regression Multiple Imputation (SRMI), and van Buuren et al. (1999, 2006) and van Buuren (2007), called the Fully Conditional Specification (FCS). The basis of the idea relies on, ideally, being able to write the joint distribution of the variables in a dataset as a product of univariate conditional distributions. Then each of the conditionals may be modelled separately. The approach works as long as the missingness pattern is monotone. A monotone missingness pattern implies that the conditional distributions of the observed data do not change as each variable is imputed. However, with non-monotone missingness, each next draw of imputations may depend on imputed values at a previous step. This means that a number of iterations may be required to update new draws of imputed values based on previous draws of imputations. The iterations must run until convergence to a joint distribution is believed to be achieved. We note that the sequential regression approach requires that the series of conditional distributions must converge to the joint distribution of the various variables, but it is not always clear if a joint distribution for the data actually exists. Liu et al. (2013) propose that as long as the conditional distributions are valid, even if not compatible, consistent estimates for various common estimands may still be obtained.

In terms of hierarchical data, a chained equations approach for multilevel data is implemented by Jacobusse (2005). Yucel et al. (2006) work with the sequential regression approach for multilevel data with an approach named SHRIMP. Performance of SRMI for multilevel structures has also been studied by Zhao and Yucel (2009). The package *mice* in R (van Buuren and Groothuis-Oudshoorn, 2011) provides the sequential regression approach to multilevel missing data. The limitation, however, is that only two levels of data, and continuous variables may be modelled.

There are various advantages and disadvantages for both the modelling approaches. A joint modelling approach guarantees valid imputations. Survey data may, however, contain a large number of variables, potentially of different types, such as continuous,

ordinal or binary. Many of these variables may also have logical constraints to satisfy. It is often not possible to define a multivariate distribution complicated enough to describe such a dataset, let alone predict plausible values from such a model. Therefore, for most practical purposes the sequential regression approach is the only viable option for imputations. A similar argument can be applied to MI for SDC.

A number of different univariate models may be used in the imputation/synthesis process using MI. For the parametric approach, regression models may be used to model each variable. Naturally, this would imply using the simple linear regression for continuous variables and generalised linear models for binary and nominal variables. An alternative method for categorical variables is the multinomial-dirichlet model, which assumes a Dirichlet distribution for the priors and the multinomial distribution for the variable itself. There are some non-parametric approaches proposed in the literature as well. These include the Bayesian bootstrap (Lo, 1987) and Classification and Regression Trees (CART) models. CART models can handle synthesis for data with irregular distributions and capture non-linear and interaction effects between variables, issues which sometimes may not be easily handled in parametric approaches (Reiter, 2005c). Other related methods such as bagging, random forests and support vector machines have also been suggested to replace the CART approach. However, Drechsler and Reiter (2011) find the CART method superior to the alternatives for generating synthetic data.

In addition to choosing the form of the model, several other adjustments may be required within the models to account for special structures within the dataset. Drechsler (2012) introduce an imputation model that inflates the variance for synthesis of highly sensitive records, such as large businesses. The author shows that in practice, several models may be required to synthesise even one variable, modelling units separately by strata, region or other characteristics that make them differ considerably. Real datasets may also pose other challenges. Specific strategies must be employed to deal with semi-continuous variables, variables with linear constraints and questions with skip patterns. Drechsler has written extensively about some of these practical issues while dealing with a complex business data set in Drechsler (2011a), Drechsler et al. (2008b), Drechsler (2012) and Drechsler and Raghunathan (2008).

The synthetic data literature has so far not directly addressed multilevel data structures. The two released synthetic hierarchical datasets, i.e. Longitudinal Business Database (Kinney et al., 2011) and the Survey of Income and Program Participation (Abowd et al., 2006), use synthesis models that condition on available variables at various levels using a series of non-hierarchical models. For large longitudinal surveys, for instance, it may not always be possible to include all observations from all waves for each variable in the synthesis model. In the existing hierarchical synthetic data examples mentioned above, some kind of model selection was necessary to condition on at least some of the available observations from previous time points. It is important, in the interest of reducing the chance of uncongeniality, that the synthetic data represent the correlation

within clusters correctly. Therefore, there is a strong motivation to explore the use of multilevel models to synthesise hierarchical datasets.

1.1.4 Uncongeniality

Accompanying the modelling choices, mentioned above, is the issue of congeniality. For a statistic of interest, Q, Meng (1994) describes congeniality in terms of the mean and variance of Q. If congeniality holds, the estimate for the mean and variance of Q under the imputer's and the analyst's models should match asymptotically, given they are both utilising the same form of the dataset, whether incomplete or complete. When the specific relationship within or between variables explored by the analyst have not been preserved by the imputation model, uncongeniality may occur. We note that congeniality is specific to the analysis; an imputation model may be congenial for a particular analysis model, but not another.

This leads us to the suggestions for imputation models in the literature to avoid uncongeniality. The most general and saturated models are usually recommended (Meng, 1994; Rubin, 1980, 1996; Clogg et al., 1991; Schenker et al., 1993, 1988). It is proposed that even if certain predictors are not significant in the imputation model, they may still be retained if it is suspected that the analyst may use them. This is the most important difference between modelling for analysis and modelling for imputation purposes. While it is undesirable for analysis, overfitting the data is arguably the best possible strategy for imputation models to ensure congeniality to a wide range of analyses.

Nevertheless, as Machanavajjhala et al. (2008) note, there is no perfectly fitting model for each and every observation which would also be adequate in a predictive sense to provide synthetic data. Both the bias and noise coming from the fitted model are sources of differences between the original and the synthetic data. These differences are required. If it was not for these differences, data confidentiality would be impossible to achieve.

In the course of our research. we focus on driving the synthesis strategy for multilevel data based on the modelling preferences of prospective analysts. We tackle two distinct issues. The first is related to omitted variable bias and the second regarding the shape of the error distribution.

1.1.5 Omitted variable bias

Researchers from different disciplines may prefer different analysis models for the same dataset. When it comes to hierarchical data models, analysts often choose between random effects (RE) and (FE) models. The research paper by Hausman (1978) popularised the Hausman test, which brought about a wave of change in how models were chosen, especially in the discipline of Econometrics. The Hausman test (Wu, 1973; Hausman,

1978) is a specification test that can be used to test the consistency of an estimator. One of the areas where the Hausman test is widely used to date, is to compare FE and RE models. The null hypothesis is that an RE estimator is consistent; failure to reject the null hypothesis implies that an RE model will give an estimate which is as 'good' as the FE model estimate, under the assumption that the FE model estimator is always consistent.

The FE versus RE debate has been given considerable attention in the literature; Allison (2009) provides a comprehensive review. The two types of models have different approaches towards dealing with errors correlated within clusters. In an FE model, the effect of all unobserved variables, that are constant within a given cluster, is absorbed in *fixed* effects, added as extra covariates to a simple linear regression model. Because of this, an FE model may not be used to estimate the effect of any other covariate that is constant within a cluster, even if it observed, as the *fixed* effects take those into account too. As Allison (2009) states, the aim is to discard the 'contaminated' variation, as this may be confounded with other unobserved characteristics of the units, sacrificing efficiency to reduce potential bias. In an RE type model, the cluster effects are usually assumed to be normally distributed and uncorrelated with any other covariate. Both within and between-cluster variations are utilised for the estimation of the cluster effects in an RE model, while only the within-cluster variation is used to estimate the parameters in an FE model, which makes an RE model more efficient in most applications. The efficiency advantage of an RE model over an FE model is most pronounced when the between-cluster variation is smaller than the within-cluster variation. As the between-cluster variability becomes larger, the RE model is not significantly more efficient than the FE model.

The reason why a standard RE model may not be the preferred choice of model in some disciplines is its assumption regarding the independence of cluster effects from other covariates. There is an alternative; Mundlak (1978) showed that an RE model can provide similar regression estimates to an FE model, if an RE model allowed for the correlations between the cluster effects and the covariates. In this sense the RE model is a special case of the FE model, with an added assumption of independence of covariates and cluster effects. As long as the assumption holds true, an RE model can provide unbiased parameter estimates with more efficiency than an FE model. However, assume there exist unobserved variables, that are constant within each cluster, correlated to the observed covariates. The effect of these unobserved covariates is absorbed by the cluster effects and this makes the cluster effects correlated with the observed covariates in the model. In such a case, the assumption of independence of cluster effects and covariates of an RE model is violated, and biases may be observed in the regression coefficient estimates for the covariates that are correlated with these unobserved variables. Such a bias is commonly known as the Omitted Variable Bias (OVB) in the Econometrics literature. The FE model is protected from OVB as it discards the between-cluster variation. Note

that the FE model does not protect against biases resulting from omitting a variable that changes within a cluster over several measures (cluster-varying), or an effect that is related to a characteristic constant within a cluster related to a cluster-varying variable, for example an interaction term between gender and salary.

Clark and Linzer (2015) presented simulation results which show that OVB increases with increasing values of the correlation between the unit effects and the covariate of interest. The number of units in a data set do not affect the magnitude of OVB, but as the number of observations per unit/cluster increases, biases in an RE model parameter estimates decrease. The simulation also addressed the power of the Hausman test. The authors illustrated that for a small number of units, the Hausman test fails to reject the RE model when it should, i.e. even when the correlation between the omitted variable and a modelled covariate is high.

For our research, we are interested in exploring the effects of OVB on the synthesis process. As we do not wish to bias the final analysis by choosing the FE model over RE or vice versa, it is important to investigate the effect of model assumptions (here, independence of unit effects and covariates) on the synthesis procedure.

1.1.6 Fixed or random effects?

There is scarce literature directly addressing fixed effects versus random effects models when applied to multiple imputation. Firstly, Reiter et al. (2006) proposed incorporating sampling design within imputation models to avoid biases in analyses that use a design-based approach. In an attempt to control for clustering effects because of sampling, the authors fit both fixed and random effects type models for imputation of missing data. They address the problem of excluding clustering and stratification effects from imputation models when the data in fact contain such features by design. In this thesis, we work with the assumption that the imputer has decided to use imputation models that incorporate clustering features. Our focus is on the possibility of model misspecification for hierarchical models, and the choice between them.

The second most relevant piece of research on MI using fixed and random effects models is by Andridge (2011). As random effects models are commonly used to analyse cluster randomised trials, the author studied the effects of using MI with fixed effects models in such a context. The factors explored include the intra-class correlation, the cluster size and the missing data mechanism. The author showed that biased variance estimates may be obtained using the fixed effects imputation models when the cluster sizes and the intra-class correlation are small. The simulation focused on the estimation of the fixed effects within an RE model only.

Drechsler (2015) extended this research with an emphasis on education research, where the random effects themselves are of interest in an analysis model. He showed that,

if missing data are imputed using an FE model, when the intra-class correlation and cluster sizes are small, the final random effects themselves can be biased. As these are important in the educational context, such as school effects, this suggests caution in the choice of imputation models.

All three articles above agree on the simplicity of running fixed effects models, and therefore the practical convenience associated with them. The research by Reiter et al. (2006), although addresses correlation within clusters, explores the necessity of using hierarchical models. The two most recent articles focus on the use of only random effects models as analysis models. As analysts using the RE model, they suggest that imputation using a random effects approach is a safer option. We note that the prospect of using the more efficient random effects models with the ability to model time-constant covariates make the random effects model an attractive choice for imputation models as well.

In the first part of this thesis, we aim to fill a number of gaps in the literature regarding MI for hierarchical data which may arise from repeated measures for the same cluster which may or may not be spread over time. First, we hope to cover all ground in terms of analysis models and do not wish to assume working only in a specific context, although still driven by the Econometrics literature, we do not make any assumptions about which of the two, fixed or random effects models, an analyst may use. Second, we specifically work in the Synthetic Data scenario, where no such research currently exists and, therefore, the added subject of comparative disclosure risks can be studied. Third, we address the issue of the need for hierarchical models in the first place. Simple linear regressions may be used, as currently done in the synthetic data literature, to impute or synthesise datasets with a hierarchical structure. While this is done for convenience, we ask how much more value can be added by using the hierarchical models, given that the variable to be imputed or synthesised does depend on the clustering in the data. Fourth, again although driven by the Econometrics literature, we recognise that the omitted variable bias problem is simply a problem of correlation between the cluster effects and the covariates. This correlation may be induced by design effects or any other source, and therefore, the applications of the current work can be extended to other areas of research. Fifth, we also hope to explore novel modelling approaches that may not involve hierarchical modelling but can be used to synthesise hierarchical data.

1.1.7 Shape of the error distribution

Data often contain variables with skewed distributions that should ideally be modelled appropriately for their shape. The most popular regression models, whether in non-hierarchical or hierarchical data, assume normality of the residuals. Non normality of the error term may be studied through the residuals of a model. Skewed distributions

of the residuals are often dealt through transformations in the dependent variable. Ideally, if the distribution of the variable is known in advance, it could be used to find a suitable transformation for the data. If not, then plots of residuals can help identify the transformations required or one could be selected by trial and error. Box and Cox (1964) suggested selecting an appropriate power transform depending on maximising the likelihood for the model at hand with various trial transformations. For long-tailed errors, analysts may use a different distribution to base inferences on. As in the case of analysts, data keepers also need to look out for unusual error distributions when synthesising data for SDC. Below, we discuss the literature covering MI models adjusted for unusual observations in the data.

Robust regression is a class of regression models that are designed to be comparatively insensitive to the failure of standard linear regression assumptions. Some authors have proposed the use of robust estimation methods such as the M-estimator for model coefficients (Rana et al., 2012). A number of semi and non-parametric imputation approaches, such as predictive mean matching (PMM) or local residual draw (LRD), have also been proposed (Lazzeroni et al., 2011; Little, 1988; Schenker and Taylor, 1996).

In standard application of MI, transformations are generally recommended to satisfy the normality assumption (Drechsler, 2011a; Schafer and Graham, 2002; Allison, 2001; Raghunathan et al., 2001). White et al. (2011) discuss using the Box-Cox or shifted log transformation to deal with skewness. In case of different kurtosis and other non-normal behaviours, they suggest exploring Johnson S_u family (Johnson, 1949) and the modulus-exponential-Normal (MEN) and modulus-power-Normal (MPN) transformations (Wright and P., 1996). It is useful to note that certain transformations such as the log may not handle tail behaviour properly (He and Raghunathan, 2006). When the imputed values are transformed back, the inverse log can inflate certain imputations to implausible values.

As a common solution to this problem, softwares such as Stata, SAS and SPSS offer users to impute within certain bounds. For instance, to impute the exponential distribution, all negative imputed values, can be rounded to 0 (censoring) or the procedure repeated several times until a positive value is obtained (truncation). However, it has been shown that using transformation or truncation for MI models may bias results for imputed data (von Hippel, 2013).

A number of articles have been published proposing the use of more flexible distributions for MI. Tukey's g-and-h family of distributions (Tukey, 1997) has been shown to perform well for imputation of variables with various skewness and elongation properties. The idea is proposed by He and Raghunathan (2006) and further extended for multivariate data by He and Raghunathan (2012). The authors have further developed their techniques for the imputation of error terms through the g-and-h distributions (He and Raghunathan, 2009). This is the only article in our knowledge that directly addresses

multiple imputation for the non normal error term in regression analysis using a flexible parametric family.

Further ideas for flexible distributions for MI of univariate data have been proposed by Demirtas and Hedeker (2008) and Demirtas (2009). They propose the use of Fleishman's power polynomials, and the generalised lambda distribution. Generalised additive models for location, scale and shape (GAMLSS) have also been shown to perform well in certain scenarios (de Jong et al., 2014; Rigby and Stasinopoulos, 2005; Stasinopoulos and Rigby, 2007).

In terms of hierarchical data, errors are usually studied separately for the various levels of the model. In a two-level random intercepts model for instance, these would be the observation level and cluster level error terms. Usually, both are assumed to be independently and normally distributed. Just as in the case of non-hierarchical models, there are several articles that discuss misspecification of the error distributions for random effects models.

For the within-variation errors, we found various methods to model heterogenous error terms (Foulley et al., 1992; Chinchilli et al., 1995; Lin et al., 1997), variances related to the mean (Davidian and Giltinan, 1993; Vonesh, 1992) and skewed error terms (Arellano-Valle et al., 2005). There are a number of articles suggesting that misspecifying the shape of the random effects in a linear mixed model may not have any significant consequences for the usual targets of inference, especially the fixed effects (McCulloch and Neuhaus, 2011; Ghidey et al., 2010; Verbeke and Lesaffre, 1997; Butler and Louis, 1992; Neuhaus et al., 1992, 1994). However, analysts may often be interested in the random effects and the accuracy of model-based standard errors for coefficients, both of which may be affected by misspecified shape of the within-error term. Methods to work around misspecification of the between-variance term include use of nonparametric likelihoods (Agresti et al., 2004), flexible parametric distributions (Zhang et al., 2008; Lesaffre and Molenberghs, 1991; Piepho and McCulloch, 2004), marginalized models (Heagerty et al., 2000) and h-likelihood (Lee et al., 2004). Some approaches have been developed to estimate the actual shape of the random effects. These include use of mixtures of normal distributions (Magder and Zeger, 1996; Ghidey et al., 2004; Caffo et al., 2007), the heterogeneity approach (Verbeke and Lesaffre, 1996), non parametric alternatives (Zhang and Davidian, 2001) and smoothening by roughening approach (Shen and Louis, 1999). However, these approaches are mainly inference-driven, rather than motivated by prediction or imputation.

For MI, Yucel and Demirtas (2010) studied the effects of using random effects models when the assumption of normality does not hold true for the random effects. They concluded that the estimation of fixed effects and individual error variances are quite robust to the failure of non-normality of random effects. However, the estimates for variances of the random effects themselves may be biased. They indicate interest in

exploring the use of less restrictive distribution assumptions for MI of hierarchical data, such as those utilising marginal models. Furthermore, Yucel (2011a) proposed use of random covariance structures to model random effects with heterogenous variances while using multilevel models for MI. Other works by the authors also offer developments to support the use of random effects models for MI in real data scenarios (Yucel, 2008; Schafer and Yucel, 2002; Yucel et al., 2008; Demirtas et al., 2008).

A further alternative to flexible parametric approaches to model unusual error distributions is quantile regression (Koenker and Bassett Jr, 1978). Quantile regression is becoming an increasingly popular procedure to handle various quantiles of the data more flexibly in a number of applications (Koenker, 2005; Yu et al., 2003). An imputation scheme for using quantile regression for MI has also been developed by Bottai and Zhen (2013).

Some of the issues related to the application of quantile regression to MI have been discussed by Geraci (2016). Geraci (2016) ran simulations and a presented a real data application to compare the performance of various imputation methods when the data are subject to complex survey design and have various skewed and bounded variables. For data generated with heteroscedastic errors, the quantile regression imputation outperformed the normal imputation, normal imputation with transformation and even quantile imputation with transformations. When the errors were homoscedastic, the quantile regression approach performed as well as the other methods studied. The method used by the author is available in MICE (van Buuren and Groothuis-Oudshoorn, 2011), one of the MI packages in R. The package also allows for transformations before the imputation, as proposed by Geraci and Jones (2015).

For missing independent covariates, Wei et al. (2012) develop a suitable estimator for parameters in a quantile regression after MI. Their procedure uses quantile regression for the imputation process itself and further adjustments to get the final estimates. Wang and Feng (2012) developed an MI procedure for missing covariates that are also censored using censored quantile regression. They show that their procedure has efficiency gains comparing to the usual normal imputation methods in certain situations. Methods covering quantile regression for longitudinal data with missingness are addressed by Lipsitz et al. (1997) and Yuan and Yin (2010).

The Ph.D. thesis by Chen (2014) provides a detailed account of MI using both parametric and semi-parametric quantile regressions. The research, however, focuses on the using quantile regression MI in combination with analysis procedures that use either weighted generalised method of moments (GMM) or empirical likelihood (EL) methods to estimate parameters using generalised estimating equations. A two-stage MI procedure using nonparameteric quantile regression has also been proposed (Hu et al., 2014b).

In the second part of this thesis, we contribute to the synthetic data literature in a number of ways. First, we assess the performance of transformations for synthesis of

data containing outliers. Second, we propose the use of quantile regression as a synthesis model. Third, we explore the synthesis strategies that are congenial to a quantile regression analysis. Fourth, we extend the use of flexible distributions for MI to synthetic data generation. Fifth, we propose how flexible error distributions may be utilised for synthesis of hierarchical data with outliers. Finally, we compare the risk profiles for all the proposed procedures in both non-hierarchical and hierarchical data settings.

1.1.8 Disclosure risks

We first introduce common terminology used in the literature on disclosure risk assessment for synthetic data. Variables in a dataset can be classified according to the role they play for data disclosure. *Key* variables, as defined by Bethlehem et al. (1990), are the variables that allow an intruder to match a record in the dataset with an individual/business/unit. Obvious key variables in a dataset are names and addresses of the units. However, these are normally not released. We can think of key variables as either identifiers and quasi-identifiers, where identifiers can be names, addresses, national insurance numbers, and quasi-identifiers include variables that may not be unique to each unit but can help uniquely identify a unit, such as a combination of demographic and geographic information. The variables which contain sensitive information that need protection are called *sensitive* variables. The two sets of key and sensitive variables need not be disjoint.

An intruder is defined to be an ill-intentioned user of the released dataset who attempts to identify individuals, known as targets, or information about them, from a released dataset. Therefore, there are two main types of disclosures, 1) identification disclosure and 2) attribute disclosure. Correctly identifying an individual in a synthetic dataset is called identification disclosure, while attribute disclosure occurs when an intruder learns the values of sensitive variables contained within the data. Identification disclosure is usually assessed using key variables and combinations of them, while attribute disclosure is mostly studied through the sensitive variables. Manipulating the values of key variables can help reduce identification risks, and further attribute disclosure risks can be controlled by perturbing the values of sensitive variables. We discuss these in more detail along with risk evaluation frameworks below.

Identification disclosure may occur and not lead to any attribute disclosure, for instance, an intruder may identify an individual in a dataset, but does not learn anything new about that unit. Also, attribute disclosure may occur without identification, for instance, if the water bill in postcode SO17 is same for all houses in the area, the release of the average water bill, is also the bill for any one house and this can be known without identifying that house in the dataset (for more, see Duncan and Lambert (1986) and Skinner (1992)). The two types of disclosures are not always independent either. According to Reiter (2012), organisations mainly focus on identification disclosure risks, as

an intruder who can identify a unit in the dataset, then also knows sensitive information about that unit, i.e. attribute disclosure risk.

There are other types of disclosures studied in the literature; these include, but are not limited to, *inferential*, *model* and *population* disclosure. These disclosure measures may or may not be concerned with a particular unit and can be measured using similar frameworks as for identification or attribute disclosures. When a certain relationship between variables in the population or information on a particular model is confidential itself, population or model disclosure risks may be evaluated. These were first studied by Palley and Simonoff (1987) by considering how much information on certain models or relationships differed in the real and synthetic data. OECD and other governmental organisations are also concerned about *residual* disclosure risk - a confidentiality breach that occurs by combining outputs from two different data sources. We do not discuss these types of risk any further.

An important distinction is made between a *true* and *false* disclosure. It is a matter of debate whether wrongly identifying an individual or obtaining wrong sensitive values for them amounts to a disclosure, as this may harm individuals in a different way. Papers such as Blien et al. (1992) focus on true disclosure only, while the research by Duncan and Lambert (1986, 1989) does not make any distinction between true and false disclosures. Lambert (1993) describes the concept of *perceived* disclosure, which constitutes of both true and false disclosure, whether identification or attribute. Duncan and Lambert (1989) argue that limiting the perceived disclosure risk is equivalent to preventing an intruder to believe that a disclosure has occurred regardless of whether it is true or not, as false disclosure may also harm. However, there are other reasons as to why false disclosure risks can be of concern. As Duncan and Lambert (1986) explain, if the disclosure risk can be seen as a density of probability of matches, we would like this function to be smooth, i.e. not have peaks for certain individuals. However, equally important is the location of these peaks, i.e. whether a true or false risk is more likely. The problem with peaks at false matches is that these are possibly a result of lower data utility and are, therefore, undesirable.

We now discuss how an identification can be made, and therefore, assessed by a data keeper. From here onwards, we only discuss methods for partially synthetic data and also ignore the uncertainty from sampling, i.e. the possibility that the intruder's target unit may not be in the sample released for public use. For an application of intruder uncertainty on whether the target is in the sample, see Drechsler and Reiter (2008).

Identification disclosure risks are mainly studied by calculating probabilities of identification for each unit in the dataset. A higher probability of match is more risky than a lower one. However, thresholds of what is considered 'risky' are usually arbitrarily chosen.

One of the ways to identify someone is through *uniqueness*. This is related to the *key*, i.e. the value of the combination of key variables that identifies an individual. For instance, if there is only one asian female neurosurgeon living in Northumberland, she is a unique record in the dataset, making identifying her a trivial task. The more commonly occurring a key is, the less is the identification risk for each individual having that key, as the probability of identifying that unit in the dataset is spread across all possible units having that key. The concept of uniqueness is well-studied (see Bethlehem et al. (1990); Greenberg and Zayatz (1992); Skinner (1992); Skinner et al. (1994); Chen and Keller-McNulty (1998); Fienberg and Makov (1998); SM (1998); Pannekoek (1999) and Dale and Elliot (2001)) and we do not go into further detail regarding uniqueness here. Elamir and Skinner (2006) and Skinner and Shlomo (2008) review research on estimating the probability that a unit is unique in the population, given they are also unique in the sample released to study such disclosure risks.

Uniqueness is important because uniques in a dataset have higher risks of identification as compared to non-uniques. However, as a measure of disclosure risk, uniqueness on its own has some limitations. Wherever continuous variables are recorded, it is possible to have several sample uniques. Also, the number of uniques in a dataset alone is not a strong measure of disclosure risks. These can remain unchanged before and after the data are perturbed. Most importantly, the concept of uniqueness does not incorporate the intruder's knowledge when assessing disclosure risks. An intruder may have strong prior knowledge for a unit who may not be a unique and studying uniqueness alone will disregard disclosure risks for such a unit. There are other measures of disclosure risks that overcome some of these problems. There are three main streams of research in the literature in this regard, and we organise them as 1) record-linkage, 2) differential privacy and 3) model-based approaches.

For record-linkage, we assume that the intruder may have another source of information regarding the units of interest and tries to match the units in the released synthetic data to this information. The number of correct matches may then be used in various ways to quantify the risk, for example, either as a percentage of the total number of units, or as compared to false matches. The data keeper may use the original data instead of an external dataset to assess such a risk. This is conservative, but an easy-to-implement approach that can provide an upper bound on the amount of risk such an approach can pose. Spruill (1982, 1983, 1984) use such an approach. Other research on record-linkage include work by Paass (1988), Yancey et al. (2002), Domingo-Ferrer (2002), Domingo-Ferrer and Torra (2003) and Skinner (2008).

The record-linkage approach is probably one of the best ways to replicate intruder behaviour. However, it is completely uninformed by the synthesis process - a source of information that may help an intruder learn more about the underlying original sensitive data. Secondly, the approach lacks a universal underpinning theory that may cover a

variety of assumptions regarding the intruder's external information, as assuming that the intruder has the original dataset, even as a benchmark, may not be realistic.

There are some examples of the use of record linkage for the evaluation of disclosure risks in the literature. Abowd et al. (2006) use both probabilistic and distance-based measures to perform the linking for their synthetic public-use files for the Census Bureaus Survey of Income and Program Participation (SIPP), the Internal Revenue Services (IRS) individual lifetime earnings data, and the Social Security Administrations (SSA) individual benefit data in the US. In their case, there existed already public use data for the SIPP, and so trying to link the original SIPP data to the new synthetic data is not so much of an unrealistic exercise anymore, as intruders would have had access to the previously released SIPP data on the public domain. Their research aims to satisfy the standards set by the Census Disclosure Review Board, which regulates the process by setting caps on the true match vs. number of units, and the true match vs. total match rates. We note that linking incorrectly may also be damaging to units; Duncan and Lambert (1989) propose that in case of incorrect linkage, it is the data keeper's job to ensure that an intruder considers it unwise to link incorrectly. They achieve this through a study of loss functions that quantify the loss incurred when an incorrect link is made.

One of the earlier definitions of disclosure is, called *inferential* disclosure was first proposed Delanius (1977). The idea is that if an intruder can learn more about a unit by the release of a dataset or statistics associated with it than without their release, then a disclosure has taken place. Delanius (1977) explains inferential disclosure in detail along with its 64 sub-types in his paper. Dwork (2006) explains why achieving absolute privacy by such a criterion is not possible. Instead such a disclosure may be limited using the concept of *differential* privacy - a concept borrowed from the computer science literature. This is measured by the criterion called ε-differential privacy:

$$max \left| ln \left(\frac{Pr(D|N^1)}{Pr(D|N^2)} \right) \right| \leq \varepsilon. \tag{1.4}$$

Here, N^1 and N^2 are two copies of the original dataset differing only by one row. Note that to find this maximum, all possible N^1 and N^2 must be generated. D is the released dataset, and therefore, $Pr(D|N^1)$ represents the probability that the released dataset comes from the underlying population N^1 and ε is a threshold defined by the data keeper.

Differential privacy does not require any particular assumptions regarding an intruder's prior knowledge or probability distributions to represent these as priors. As a Bayesian understanding of differential privacy, we can say that the prior information held by an intruder is information on all other units except one. Then the criterion, (1.4), must be less than a predefined threshold, ε, to assure that despite knowing about all

units in the data other than that one, the released dataset does not add substantially (where substantial is determined by ε) to the intruder's knowledge about that unit. This implies that the differential privacy criterion protects against all possible sets of prior information (Chap. 6, Drechsler (2011b)). There are some other advantages of the differential privacy criterion. It is the only method that is associated with the synthesis process and not the actual data themselves. Therefore, an SDC method has a differential privacy level, independent of the values in the dataset it is applied to. This provides a strong basis for legal and policy requirements where a method alone can be specified beforehand, as it would be known in advance the level of differential privacy it guarantees. Also, a small value of ε is a tough criterion, which is not satisfied by several standard methods of synthesis, such as swapping, deletion, shuffling or sampling (Abowd and Vilhuber, 2008). This makes the differential privacy criterion very strict, as may be desirable for certain datasets.

Dwork et al. (2006) demonstrate an application of the ε-differential privacy criterion if the aim is to release all possible tables from a census. Barak et al. (2007) revisit the problem and simplify it with the use of Fourier basis. Machanavajjhala et al. (2008) use synthesisers satisfying differential privacy criterion for the US Census Bureaus On-TheMap (OTM) micro-data, which documents the commuting patterns of the population of the United States. They use the publicly available synthetic data files and this is not an official assignment from the Census Bureau.

With sparse data and a large number of variables, it may not be possible to create a synthetic dataset, every observation of which may not be reached by every observation from the original dataset with a positive probability. In fact, Machanavajjhala et al. (2008) argue that for a synthesis procedure to satisfy the differential privacy criterion for a complex dataset, it must be completely random in nature, as a deterministic algorithm would certainly fail. This may leave the synthetic dataset with a very low utility. The authors, instead, introduced the probabilistic version of differential privacy, called (ε, δ)-differential privacy, which ensures that (1.4) is achieved with a probability $1 - \delta$, where δ, and thus the probability of failing to achieve (1.4), is small. Sarathy and Muralidhar (2011) also note how the amount of noise added to the data to satisfy such a criterion may result in a low-utility synthetic dataset for continuous variables that may be unbounded. Amongst some of the disadvantages noted in the paper, one particularly questions the privacy offered by the approach. The authors note that the additional knowledge acquired by an intruder is only $exp(\varepsilon)$ for the first few records he is looking for. After some of the records are identified, the relative knowledge gain increases exponentially for a Laplace-based noise addition SDC procedure, which was first suggested in Dwork (2006). Sramka (2012) also presents examples of how noise from datasets that may satisfy ε-differential privacy criterion can be removed and poses threat of significant disclosure risks.

So far, differential privacy and data synthesis through imputation have only managed to come together for the dirichlet/multinomial synthesiser for categorical data. Examples of these include articles by Abowd and Vilhuber (2008), Machanavajjhala et al. (2008) and Charest (2010). Certain assumptions, such as the use of non-informative priors in data synthesis that allow the MI combining rules for synthetic data, are not satisfied by the synthesisers that attain differential privacy. Charest (2010) presents this argument with an example for fully synthetic data.

We now discuss in detail the model-based approaches to investigate disclosure risks - our chosen method for this piece of research. The basic idea is to calculate the probability of identification for each record using a model-based setup. The origins of the idea can be mapped back to the concept laid out by Duncan and Lambert (1986), again following from Delanius's (1977) idea of inferential disclosure. Duncan and Lambert (1986) identified the need to model an intruder's prior knowledge and update these models using any released data or statistics to obtain a Bayesian posterior predictive distribution for the information on the target record. Again, the authors argue that complete protection using Delanius's (1977) concept is not possible, but the amount of inferential disclosure can be restricted by investigating the difference between the prior and posterior knowledge of an intruder. This is termed as *knowledge gain* in the paper, while terms such as *knowledge* and *relative knowledge gain* are also defined as alternative measures to assess disclosure and limit it.

The model-based approach was further extended and applied by Fienberg et al. (1997), Reiter (2005b), Reiter and Mitra (2009) and Drechsler and Reiter (2008). The aim of the model-based approach is to replicate the synthesis process, and create a large number of further synthetic datasets. Then, the synthesised variable, Y_{syn}, can be used to reconstruct the posterior predictive distribution of Y used for synthesising the variable in the first place. This distribution can provide the intruder some information on the range of the original values for an observation, given the value of the synthetic observation. Disclosure risk assessment with this understanding then requires some assumptions regarding the intruder's knowledge about the synthesis model, i.e. whether he has no information regarding the synthesis model, or he knows the model but not the original parameter estimates, or that full information on the model, including its parameter estimates is released to assist analysis procedures. The risk assessment can then be applied under these different scenarios. In a Bayesian setup, the prior information held by the intruder can also be incorporated using the same procedure.

There are a number of advantages to the model-based approach. Firstly, information from the synthesis models is incorporated while calculating the probability of matches. Secondly, the probabilities are calculated at a unit-level. This means that if certain records or groups of units seem to have more disclosure risks than others, these can be evaluated and rectified. In fact, this might be an important exercise as the summary measures explained may not demonstrate high risks to a certain record. Thirdly, intruder

knowledge can be explicitly modelled in this framework. However, this also means that certain assumptions need to be made before calculating the probabilities and according to different threat scenarios. Moreover, the approach also takes into the account the release of multiple synthetic datasets.

1.2 Gaps in knowledge

Research in different disciplines with specific preferences of analysis models provide some insight into concerns for the correct model for imputation of hierarchical data. However, there are no comprehensive studies in the literature so far that address the choice between FE and RE models in consideration for a variety of analyses. Additionally, the subject of omitted variable bias along with data imputation has not been studied at all.

As for synthetic data, practitioners have so far relied on series of non-hierarchical models for data generation. Therefore, the use of hierarchical models has not been studied within the synthetic data literature. This also implies that research on the implications of omitted variable bias, or disclosure risks for competing hierarchical models has not yet been conducted.

For non normality of data residuals, there are some studies that address the problem within a MI framework. However, several methods have been proposed and compared with the use of normal distribution but not against each other. Moreover, none of this research has been undertaken for synthesis of confidential data. Furthermore, it is more common to find solutions for a non-hierarchical data setting, rather than for hierarchical data.

1.3 Aims and objectives

In the course of this research, we focus on imputation and synthesis of hierarchical data. In this thesis, we aim to:

1. identify imputation models that are congenial to both fixed and random effects type analyses, as well as non-hierarchical models fitted to hierarchical data,
2. advise researchers regarding the role of omitted variable bias when using multiple imputation,
3. assist data keepers in controlling disclosure risks by using flexible modelling strategies when synthesising data using multiple imputation,
4. investigate the performance of conventional synthesis models for hierarchical data,

5. enrich the synthetic data literature with modelling strategies not currently employed in the field,

6. encourage data keepers to explore more flexible synthesis procedures, and

7. fill the gap between research in multiple imputation for missing data and multiple imputation for synthetic data generation.

To achieve these aims, we run extensive simulation studies to study imputation and synthesis of hierarchical data. Our simulations cover different features in the data with a hierarchical structure; the structure may arise in real data applications when information is collected for observations grouped in clusters in either a cross-sectional or longitudinal setting. We explore a variety of imputation/synthesis models. We further investigate our topics of interest through real data applications. We propose new methods for data synthesis and test their performance against existing models. Moreover, our investigations include two sources of model misspecification, i.e. omitted variables and non normal error distributions. Our first study is for missing data. Next, we address generation of partially synthetic data. In studies for synthetic data generation, we assess disclosure risks between the modelling strategies under consideration, to form a well-rounded view of the risk-utility trade off.

1.4 Contributions of the thesis

The original contributions following from the investigations performed in this thesis are:

1. Imputation strategies that are suited to both fixed and random effects analyses have been identified, both in point and variance estimation of regression coefficients of interest.

2. The effects of omitted variable bias on multiple imputation have been quantified. Moreover, methods to control propagation of omitted variable bias through the imputation procedure have been proposed.

3. A novel strategy for synthesis of confidential data has been proposed that provides data keepers control over disclosure risks.

4. More flexible synthesis strategies for non normal error terms have been proposed and tested.

5. Existing imputation methods have been extended from multiple imputation for missing data to multiple imputation for synthetic data.

6. Existing imputation methods have been extended from multiple imputation for non-hierarchical data to multiple imputation for hierarchical data.

7. Congeniality to quantile regressions in the context of data synthesis has been studied.

1.5 Organisation of the thesis

This thesis is organised as follows. The current Chapter provides an introduction to the thesis, reviews the literature and outlines the aims and objectives of this thesis. In Chapter 2, we study the effects of omitted variable bias for MI for missing hierarchical data through simulation studies. This is followed by a similar study with a focus on synthetic data instead, in Chapter 3, where we present results from extensive simulations and a real data application using a German Establishment survey. In Chapter 4, we propose a range of methods to synthesise confidential non normal data. We test the effectiveness of each of the methods in both a non-hierarchical and hierarchical data setting within simulations studies. We also illustrate our methodology using the Wealth and Assets survey data from Great Britain. Finally, in Chapter 5, we summarise the findings of this thesis, and point out directions for future research.

Chapter 2

Multiple imputation for missing data and omitted variable bias

In this Chapter, we study multiple imputation (MI) of missing data in the presence of Omitted Variable Bias (OVB). Section 2.1 provides the motivation and Section 2.2 outlines the statistical methods and models employed in this Chapter and Chapter 3. This is followed by a simulation study in Section 2.3 and we provide concluding remarks in Section 2.4.

2.1 Introduction

In our research, we explore both fixed (FE) and random (RE) models as imputation models. We expect that a fixed effects imputation model will be congenial to a fixed effects analysis model and a random effects imputation model to a random effects analysis. It is unlikely that the cross combination is a congenial pair for all coefficients of interest, unless the dataset is ideally suited to both. Unfortunately, in real data examples it is not possible to identify which set of assumptions is correct, i.e. those under random or fixed effects models. Therefore, we are also interested in exploring imputation models that may be congenial to both fixed and random effects types of analyses. These models may not necessarily be hierarchical in structure. Identifying such models can help us provide more general advice for analysts across various disciplines.

Some of the issues that may render one of the fixed or random effects approaches as unsuitable may not be observable, as in the case of OVB. This is the bias in a model's parameter estimates caused by the absence of a variable constant within a cluster in the dataset which may be correlated to the covariates in the model. A fixed effects model controls for such bias by separating out the unit effects from the model's other covariates, and in the process, separating out all unit specific characteristics. Whereas, a random

effects model makes the assumption that such an unobserved variable is independent of the other covariates in the model, thereby allowing the model to estimate coefficients for the unit specific covariates in a given dataset. In this Chapter, we explore the effect of OVB on the quality of imputation of hierarchical data.

As mentioned in the Chapter 1, Andridge (2011) and Drechsler (2015) showed that the variance estimates for the fixed and random effects obtained using a random effects analysis model may be biased, if the missing data imputation is carried out using a fixed effects model. The three most important factors affecting this bias identified by the authors are 1) the intra-class correlation, 2) the cluster size and 3) the amount of missing data.

However, existing literature only focuses on analysts interested in using a random effects model, such as those interested in cluster randomised trials or school effects in the educational research context. In many Econometric analyses, analysts may want to control for unobserved time invariant effects through fixed effects regression analysis (Chap. 2, Allison (2009)). In this Chapter, we attempt to cater for both RE and FE models for analysis.

2.2 Methods

In this section, we introduce the main statistical concepts and methods that are employed throughout the thesis. The discussed methods, especially models, are the essential components for all the simulation studies and analyses in the current Chapter and in Chapter 3.

2.2.1 Omitted variable bias

In the course of this research, we bring together MI and OVB. An introduction to MI has been provided in Section 1.1.1. This section further expands on the introduction to OVB provided in Section 1.1.5. Analytical expressions to quantify OVB can be derived for the simple linear regression. Suppose that the true model for a given set of variables, y, x_1, x_2 and x_3 is:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i,$$

where, $i = 1, \ldots, n$ and $\varepsilon_i \sim N(0, \sigma^2)$. But, we estimate the relevant coefficients for the explanatory variables using,

$$y_i = \beta_0^* + \beta_1^* x_{1i} + \beta_2^* x_{2i} + \varepsilon_i^*,$$

then,

$$E(\hat{\beta}_1^*) = \beta_1 + \beta_3 \left[\frac{r_{x_1x_3} - r_{x_1x_2} r_{x_2x_3}}{1 - r_{x_1x_2}^2} \sqrt{\frac{v_{x_3}}{v_{x_1}}} \right],$$

where, v_{x_1} and v_{x_3} are the sample variances of x_1 and x_3. The r's are the correlations, so for example, $r_{x_1x_3}$ is the sample correlation between x_1 and x_3.

Now suppose we omitted two variables instead and estimated:

$$y_i = \beta_0^{**} + \beta_1^{**} x_{1i} + \varepsilon_i^{**}.$$

Then, the expected value of $\hat{\beta}_1^{**}$ is:

$$E(\hat{\beta}_1^{**}) = \beta_1 + \beta_2 \left[r_{x_1x_2} \sqrt{\frac{v_{x_3}}{v_{x_1}}} \right] + \beta_3 \left[r_{x_1x_3} \sqrt{\frac{v_{x_3}}{v_{x_2}}} \right], \tag{2.1}$$

where, v_{x_1}, v_{x_2} and v_{x_3} are the sample variances of the relevant variables and the r's represent the sample correlation between two variables (Clarke, 2005).

Although the analytical form for OVB (2.1) is known (in the simple linear regression case), it is impossible to actually calculate the bias itself as information about the omitted variable, by its nature, is unknown. Furthermore, OVB may come from not only one but several omitted variables. Sometimes the effect of two omitted variables may even cancel each other's contribution towards the bias of an estimate of a parameter of interest. The problem is thus unquantifiable, even in a simple linear regression setup. Nevertheless, from (2.1), some of the factors that affect the magnitude of OVB can be identified. These are the strength of correlation between the omitted and included variables, and the variances of the two correlated variables.

As we move on to generalised least squares for multilevel models, predicting the effect of an omitted variable becomes an even more complicated problem. However, we can expect that the sample correlations between the omitted and included variables, r's and the variances of these variables, v's contribute towards OVB in a multilevel model as well.

In a similar spirit of Clark and Linzer (2015), we ran simulations to observe the effect on bias of an estimate from an RE model for a parameter of interest, given changes in the variances of the relevant covariate and its correlation with the unit effects through an omitted variable. We found that the bias increases with the increasing correlation between the unit effect and the covariate and decreases with increasing variance of the covariate. Moreover, the bias increases with the increasing variance of the omitted variable. Increasing the value of the true coefficient of the omitted variable increases the bias as well (results not shown).

To quantify OVB, some information regarding the omitted variable must be known, which is not possible, as noted before. Therefore, we can only aim to protect a chosen set of analysis from OVB in the most efficient way possible. For the purposes of current research, an understanding of OVB has two important implications. Firstly, we would like to explore the adverse effects of OVB on the imputation process to protect the quality of the imputed data, wherever OVB may effect the parameter estimates from the imputation model. Secondly, respecting the analysts' preference for either of the FE or RE models, we would like to impute a given data set with the aim of avoiding uncongeniality, regardless of which of the two models an analyst prefers to use.

2.2.2 Imputation/Synthesis for hierarchical data

This section describes conventional models used to describe the relationships between variables in a hierarchical data set. These models can also be used to predict missing values or generate synthetic values for sensitive data through the posterior predictive distribution of the missing/sensitive observations given the full data set. Given the hierarchical structure of the data, there are two distinct choices for modelling the missing/sensitive observations, call these y_{ij}. More specifically, the cluster effects themselves can be modelled as either 'random' or 'fixed'. Therefore, we divide all models under consideration broadly into two groups: FE type, where the cluster effects are considered 'fixed' and RE type models, where the cluster effects are considered 'random'. Before considering the models, we first set the basis for the Bayesian framework that can be used to predict missing/synthetic values for both FE and RE type models below.

Generating imputations/synthetic values using FE type models

Suppose we wish to impute/synthesise y_{ij}, for $i = 1, \ldots, n$ and $j = 1, \ldots, t$, from an FE type model. The predicted y_{ij}, call these y_{new}, can be generated using the following setup. Consider, the ordinary linear regression:

$$y|B, \sigma^2, \boldsymbol{X} \sim N(\boldsymbol{X}B, \sigma^2 \boldsymbol{I}_{nt}),$$

where, $\boldsymbol{X}$ is the design matrix, B is a p-dimensional vector of parameters and $\boldsymbol{I}_{nt}$ is an $nt \times nt$ identity matrix, i.e. the errors are distributed independently and identically (iid).

We can use the non-informative prior, distributed uniformly on $(\beta, \log(\sigma))$:

$$p(B, \sigma^2|\boldsymbol{X}) \propto \sigma^{-2},$$

which results in the following conditional posterior distributions for the parameters, (B, σ^2):

$$\begin{aligned} \sigma^2|y &\propto \chi^{-2}(nt-p, \frac{1}{nt-p}(y-\boldsymbol{X}\hat{B})^T(y-\boldsymbol{X}\hat{B}), \\ B|y,\sigma^2 &\propto N(\hat{B}, (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2), \end{aligned} \tag{2.2}$$

where, $\hat{B} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T y$. The values for y_{new} can then be generated from the following distribution:

$$y_{new} \sim N(\boldsymbol{X}_{new}B, (\boldsymbol{I}_{nt} + \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\sigma^2),$$

after generating B and σ^2 from (2.2).

In the case of missing data, $\boldsymbol{X}$ contains the completely observed records only, while $\boldsymbol{X}_{new}$ contains the observed values of predictors for missing Y. For synthetic data, we assume that no data are missing, and therefore, $\boldsymbol{X} = \boldsymbol{X}_{new}$.

Generating imputations/synthetic values using RE type models

Imputations or synthetic data from a conventional RE type model can be generated using a Gibbs sampler. To run a Gibbs sampler, we require the conditional distributions for each of the parameters involved in the model we wish to predict the values from. The number of iterations run depend on when convergence to the target joint distribution is believed to be achieved. The interested reader may refer to Chapter 15, Gelman et al. (2003) for a comprehensive coverage of the topic.

We can express a standard RE (here, random intercepts only) model in vector notation as follows:

$$y|B, b, \sigma^2, \sigma_b^2, \boldsymbol{X}, \boldsymbol{Z} \sim N(\boldsymbol{X}B + \boldsymbol{Z}b, \sigma^2 I_{nt}), \tag{2.3}$$

where, $\boldsymbol{X}$ is the design matrix for all 'fixed' covariates, $\boldsymbol{Z}$ is the design matrix for all random effects, B is a p-dimensional vector, where p is the number of 'fixed' parameters to be estimated, and b is an n-dimensional vector of random effects, $b \sim N(0, \sigma_b^2 I_n)$ for n clusters.

We can derive the conditional distributions for the Gibbs sampler for an RE model using conjugate prior distributions for all the parameters of interest. If the priors are:

$$B \sim N_p(\mu_B, \boldsymbol{\Sigma}_B), \sigma^2 \sim \Gamma(h, c), \text{and } \sigma_b^2 \sim \Gamma(h_b, c_b),$$

the joint posterior distribution of all the parameters of interest can be expressed as:

$$\begin{aligned} f(b, B, \sigma^{-2}, \sigma_b^{-2}|y, \boldsymbol{X}, \boldsymbol{Z}) &\propto f(y|b, B, \sigma^2) f(b|\sigma_b^2) f(B) f(\sigma^2) f(\sigma_b^2), \\ &\propto \phi_{nt}(y|\boldsymbol{X}B + \boldsymbol{Z}b, \sigma^2 I_{nt}) \phi_n(b; 0, \sigma_b^2 I_n) \times \\ &\phi_p(\beta; \mu_B, \boldsymbol{\Sigma}_B)(\sigma^{-2})^{h-1} \exp(-c\sigma^2) \times \\ &(\sigma_b^{-2})^{h_b - 1} \exp(-c_b \sigma_b^2), \end{aligned}$$

where, $\phi_n(y; \mu, \sigma^2)$ is the probability density function of a $N_n(\mu, \sigma^2)$ random variable.

The posterior conditional distributions, therefore, are:

$$\begin{aligned} &B|y, \boldsymbol{X}, \boldsymbol{Z}, b, \sigma^{-2}, \sigma_b^{-2} \sim N(\boldsymbol{\Sigma}^*[\boldsymbol{X}^T(y - \boldsymbol{Z}b)\sigma^{-2} + \boldsymbol{\Sigma}_B^{-1}\mu_B], \boldsymbol{\Sigma}^*), \\ &b|y, \boldsymbol{X}, \boldsymbol{Z}, B, \sigma^{-2}, \sigma_b^{-2} \sim N(\boldsymbol{\Sigma}_b^*[\boldsymbol{Z}^T(y - \boldsymbol{X}B)\sigma^{-2}], \boldsymbol{\Sigma}_b^*), \\ &\sigma^{-2}|y, \boldsymbol{X}, \boldsymbol{Z}, B, b, \sigma_b^{-2} \sim \Gamma(h + nt/2, c + (y - \boldsymbol{X}B - \boldsymbol{Z}b)^T(y - \boldsymbol{X}B - \boldsymbol{Z}, b)/2), \\ &\sigma_b^{-2}|y, \boldsymbol{X}, \boldsymbol{Z}, B, \sigma^{-2}, b \sim \Gamma(h_b + n/2, c_b + b^T b/2), \end{aligned} \tag{2.4}$$

where,
$\boldsymbol{\Sigma}^* = (\boldsymbol{\Sigma}_B^{-1} + \sigma^{-2}\boldsymbol{X}^T\boldsymbol{X})^{-1}$,
$\boldsymbol{\Sigma}_b^* = (\sigma_b^{-2}\boldsymbol{I}_n + \sigma^{-2}\boldsymbol{Z}^T\boldsymbol{Z})^{-1}$.

Using (2.4), we can generate values for B, b, σ^{-2} and σ_b^{-2} by Gibbs sampling, which can be used to generate y_{new} from the conditional distribution of y_{ij} (2.3). In the case of missing data, missing observations in Y may be filled in with sensible values to initiate the Gibbs sampler.

For all simulations involving the above setup in the current Chapter and Chapter 3, we use vague priors. More specifically, we set each element of μ_B equal to zero, each diagonal element of $\boldsymbol{\Sigma}_B$ equal to 1000000, $h = 0.001$, $c = 0.001$, $h_b = 0.001$ and $c_b = 1$.

Next, we describe a set of possible models that can be used to generate imputations or synthetic values for a variable within a multilevel data set. Suppose that we are interested in imputing or synthesising y_{ij}, where $i = 1, \ldots, n$ and $j = 1, \ldots, t$, implying that there are n clusters in total and therefore a total of $n \times t$ records. Also suppose we fully observe variable z_{ij}, with varying values within a cluster. We avoid involving variables constant within a cluster in our research to allow for a fair comparison between the FE snd RE models, as FE models cannot be used to estimate effects for these. Sections 2.2.3 and 2.2.4 provide a description of the FE and RE type models that we consider in this study, adapted to the described dataset.

2.2.3 FE type models

The conventional fixed effects model, FE:

$$y_{ij} = \alpha_i + \beta z_{ij} + \varepsilon_{ij}, \tag{2.5}$$

where the unit effects α_i are fixed, not random. Note that as the number of units gets larger, the number of parameters in the model will also additively increase. Nevertheless, we would expect that the estimate for β will be unbiased even if there is an omitted variable, i.e. a variable constant within a cluster, used in the data generation process, correlated with z_{ij}, which is not available to the data keeper.

To generate new y_{ij}, we can follow the procedure described in Section 2.2.2 with:

$$y = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1t} \\ y_{21} \\ \vdots \\ y_{2t} \\ \vdots \\ y_{nt} \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & 0 & \dots & 0 & z_{11} \\ 1 & 0 & \dots & 0 & z_{12} \\ \vdots & \vdots & \dots & \vdots & \vdots \\ 1 & 0 & \dots & 0 & z_{1t} \\ 0 & 1 & \dots & 0 & z_{21} \\ \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 1 & \dots & 0 & z_{2t} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & z_{nt} \end{pmatrix} \text{ and } B = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \\ \beta \end{pmatrix}.$$

The non-hierarchical simple regression model which ignores the unit effects, IGN:

$$y_{ij} = \alpha_0 + \beta z_{ij} + \varepsilon_{ij}, \tag{2.6}$$

where, α_0 is a fixed intercept for all individuals. To produce y_{new} from the IGN model, we use the procedure in Section 2.2.2 with:

$$y = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1t} \\ \vdots \\ y_n t \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & z_{11} \\ 1 & z_{12} \\ \vdots & \vdots \\ 1 & z_{1t} \\ \vdots & \vdots \\ 1 & z_{nt} \end{pmatrix} \text{ and } B = \begin{pmatrix} \alpha_0 \\ \beta \end{pmatrix}.$$

It is not completely unreasonable to use IGN to generate values for y_{ij}. In a simulation study for MI of missing data, van Buuren (2010) notes that as long as the model is used to impute values of the response variable only, the IGN model (or flat file approach, as he names it) should give unbiased results. Moreover, analysts may be interested in

using non-hierarchical model to analyse hierarchical data. In our research, we would like to observe the implications of using the IGN approach for imputation/synthesis on an analysis models' parameter estimates and their variability. Given that this model is computationally easier to run, the real question is whether it is worth using more complicated models for synthesis of hierarchical data.

The IGN model with an extra unit-specific covariate, IGN2:

$$y_{ij} = \alpha_0 + \beta z_{ij} + \gamma \bar{z}_i + \varepsilon_{ij}, \tag{2.7}$$

where, α_0 is a fixed intercept for all individuals. To produce y_{new} from the IGN model, we use the procedure in Section 2.2.2 with:

$$y = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1t} \\ \vdots \\ y_{nt} \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & z_{11} & \bar{z_1} \\ 1 & z_{12} & \bar{z_1} \\ \vdots & \vdots & \vdots \\ 1 & z_{1t} & \bar{z_1} \\ \vdots & \vdots & \vdots \\ 1 & z_{nt} & \bar{z_n} \end{pmatrix} \text{ and } B = \begin{pmatrix} \alpha_0 \\ \beta \\ \gamma \end{pmatrix}.$$

The IGN2 model includes in the covariates the $\bar{z}_i$ variable to help decompose the between and within effects for z_{ij}. In a dataset where OVB can exist, this extra variable can remove the effect of OVB from β, giving an unbiased estimate of β. We would like to see if using this as a strategy to protect the synthesis process from OVB can help improve data utility.

A series of simple linear regressions, synthesising y_{ij} for each j sequentially, WIDE:

This strategy makes more sense if the hierarchical data is longitudinal, for example, time points clustered within individuals, rather than individuals within schools. Therefore, if the longitudinal data is organised in a 'wide' format (and all the individuals share the same time points), y_{ij} for each time point j can be modelled separately, hence using information from each of the other time points to impute/synthesise the one being modelled. The WIDE approach avoids the need for hierarchical models. The series of t models for y_{ij} utilised are;

$$\begin{aligned} y_{i1} &= \zeta^1_{0,0} + \zeta^1_{1,1} y_{i2} + \ldots + \zeta^1_{1,t-1} y_{it} + \zeta^1_{2,1} z_{i1} + \ldots + \zeta^1_{2,t} z_{it}, \\ &\vdots \\ y_{it} &= \zeta^t_{0,0} + \zeta^t_{1,1} y_{i1} + \ldots + \zeta^t_{1,t-1} y_{i,t-1} + \zeta^t_{2,1} z_{i1} + \ldots + \zeta^t_{2,t} z_{it}. \end{aligned} \tag{2.8}$$

The basic idea of the WIDE model is to run the series of models in the spirit of a Gibbs sampler. Each of the models in (2.8) is run sequentially to predict new values of the their respective response variables. To impute/synthesise y_{i1}, for instance, we follow the

procedure in Section 2.2.2 with:

$$y = \begin{pmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{n1} \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & y_{12} & \dots & y_{1t} & z_{11} & \dots & z_{1t} \\ 1 & y_{22} & \dots & y_{2t} & z_{21} & \dots & z_{2t} \\ \vdots & \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ 1 & y_{n2} & \dots & y_{nt} & z_{n1} & \dots & z_{nt} \end{pmatrix} \text{ and } B = \begin{pmatrix} \zeta^j_{0,0} \\ \zeta^j_{1,1} \\ \vdots \\ \zeta^j_{1,t-1} \\ \zeta^j_{2,1} \\ \vdots \\ \zeta^j_{2,t} \end{pmatrix}.$$

After predicting new y_{i1}, we update the records of y_{i1} and then repeat the procedure with $y_{i2}, \dots, y_{it}$, updating each set of records at each step. Then, we repeat the whole procedure again, and iterate through the steps until convergence to the target distribution is believed to be achieved. Note that, there may not exist a well defined joint distribution for the data, based on the conditional models within the WIDE model and there is no guarantee that the sampler will converge to a target joint distribution. Nevertheless, it is not uncommon to see multilevel data being synthesised through ordinary linear regressions by conditioning on the response variable for some or all of the remaining $j-1$ observations (Kinney et al., 2011; Abowd et al., 2006).

2.2.4 RE type models

The random intercepts (we focus on random intercepts only) model, RE:

$$y_{ij} = \beta_0 + b_i + \beta z_{ij} + \varepsilon_{ij}, \tag{2.9}$$

where,
$b_i \overset{iid}{\sim} N(0, \sigma_b^2)$,
$\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$.

In its conventional form, the random effects, b_i, are modelled using a Normal distribution either centered at 0 or if expressed differently, at β_0. Correlations between b_i and any covariate z_{ij}, either through an omitted variable or otherwise, can introduce bias in the estimates for β. The random effects b_i and other coefficients are estimated as a weighted average of the within unit and between unit coefficients, where the weights themselves are the within and between unit variances. Therefore, parameter estimation in the RE model borrows information from across the clusters. If the within cluster variability is high as compared to the between cluster variability, naturally, the RE model will produce parameter estimates that are closer to the estimates for the fixed effects model.

To generate new y_{ij} using an RE model, we follow the Gibbs sampler approach explained in Section 2.2.2, with:

$$y = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{it} \\ y_{21} \\ \vdots \\ y_{2t} \\ \vdots \\ y_{nt} \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & z_{11} \\ 1 & z_{12} \\ \vdots & \vdots \\ 1 & z_{1t} \\ 1 & z_{21} \\ \vdots & \vdots \\ 1 & z_{2t} \\ \vdots & \vdots \\ 1 & z_{nt} \end{pmatrix}, \boldsymbol{Z} = \begin{pmatrix} 1 & 0 & & 0 \\ 1 & 0 & & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & 1 \end{pmatrix}, B = \begin{pmatrix} \beta_0 \\ \beta \end{pmatrix}.$$

The hybrid model, HYB:

$$y_{ij} = \beta_0 + b_i + \beta z_{ij} + \gamma \bar{z}_i + \varepsilon_{ij}, \tag{2.10}$$

where,
$b_i \overset{iid}{\sim} N(0, \sigma_b^2)$,
$\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$.

The HYB model is another RE model that combines some of the features of the FE and RE models (Chap. 2, Allison (2009)). For HYB, the covariates that change within units may or may not be centred on their means and the means are separately added to the model with their own covariate, here γ. The inclusion of $\gamma \bar{z}_i$ is an alternative way to control for OVB by introducing a time invariant covariate that can absorb any effects of a time invariant omitted variable which may be correlated with z_{ij}. The purpose of this practice is to avoid the weighted average parameter estimates from the RE model, by splitting the between cluster contribution (here, β) from the averaged contribution (here, γ) for any one estimate. Therefore, HYB produces the same estimates for β as in an FE model (as long as the data are balanced, i.e. j is equal in all clusters). However, the estimate will not be any more efficient than the β estimate from an FE model (Schunck, 2013). If the estimates for β and γ are different, this is an indication that such an omitted variable may exist and that the β estimates from a conventional RE model will produce biased estimates in that case. If no such correlated omitted variable exists, γ should be equal to β and may not be very enlightening on its own. This comparison is an alternative form of the Hausman test. The use of an RE type approach to estimation of parameters allows the inclusion of unit-fixed covariates on their own as well. The HYB model can also be used to include random slopes and more complex error structures for hierarchical data (Chap. 2, Allison (2009)).

To generate y_{new} from the HYB model, we can use the Gibbs sampler in Section 2.2.2, with:

$$y = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{it} \\ y_{21} \\ \vdots \\ y_{2t} \\ \vdots \\ y_{nt} \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & z_{11} & \bar{z_1} \\ 1 & z_{12} & \bar{z_1} \\ \vdots & \vdots & \vdots \\ 1 & z_{1t} & \bar{z_1} \\ 1 & z_{21} & \bar{z_2} \\ \vdots & \vdots & \vdots \\ 1 & z_{2t} & \bar{z_2} \\ \vdots & \vdots & \vdots \\ 1 & z_{nt} & \bar{z_n} \end{pmatrix}, \boldsymbol{Z} = \begin{pmatrix} 1 & 0 & & 0 \\ 1 & 0 & & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & 1 \end{pmatrix}, B = \begin{pmatrix} \beta_0 \\ \beta \\ \gamma \end{pmatrix}.$$

For the HYB model, the posterior conditionals remain the same as for the RE model, (2.4). As apparent from $y, \boldsymbol{X}, \boldsymbol{Z}$ and β above, the Gibbs sampler for the HYB model can be run as for the RE model with added covariates in $\boldsymbol{X}$. The form of $\boldsymbol{Z}$ does not alter and random effects are estimated just as in the RE model.

The model proposed by Bafumi and Gelman (2006), BG:

$$y_{ij} = \beta_0 + b_i + \beta z_{ij} + \varepsilon_{ij}, \tag{2.11}$$

where,
$b_i \overset{iid}{\sim} N(u, \sigma_b^2)$,
$\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$.

Here, the random effects b_i are no longer centred at 0 as in the RE and HYB models, but at u which is a linear function of covariates at the second level of the random effects model. Here we use, $u = \eta \bar{z}_i$.

Bafumi and Gelman (2006) note that if the drawback of the RE model relates to the independence of b_i and the covariates, the issue can be countered by simply allowing this dependence. Therefore, they modified the RE model to include covariates at the second level of the RE model as shown in (2.11). This is similar to the idea shown by Mundlak (1978), where he proves that a conventional RE model can produce FE parameter estimates (for the covariates that change within clusters, such as z_{ij}) if the unit effects, b_i were modelled under a normal model with covariates. Mundlak (1978) calls it the 'Correlated Random Effects' model. The equivalence between HYB and BG is apparent. The only difference is that the $\bar{z}_i$ covariate is now at level 2, rather than level 1, which makes posterior prediction from such a model a slightly more complicated task.

For the BG model, let $\boldsymbol{X_2}$ be the design matrix for the model for the random effects, so that $b \sim N(\boldsymbol{X_2}\eta, \sigma_b^2 \boldsymbol{I}_k)$. Therefore, we require a prior on the set of parameters η, which we can specify as a $N(\mu_\eta, \boldsymbol{\Sigma}_\eta)$. The posterior conditional distribution for b_i in the BG model is different from that of the RE and HYB models; there is also an additional posterior conditional distribution for η. Therefore, in addition to (2.4), we require:

$$b|y, \boldsymbol{X}, \boldsymbol{X_2}, \boldsymbol{Z}, B, \sigma^{-2}, \sigma_b^{-2}, \eta \sim N(\boldsymbol{\Sigma}_b^*[\boldsymbol{Z}^T(y - \boldsymbol{X}B)\sigma^{-2} + \boldsymbol{X_2}\eta], \boldsymbol{\Sigma}_b^*), \tag{2.12}$$

$$\eta|\boldsymbol{X}, \boldsymbol{X_2}, \boldsymbol{Z}, B, \sigma^{-2}, \sigma_b^{-2}, b \sim N(\boldsymbol{\Sigma}_\eta^*[\boldsymbol{X}_2^T b\sigma_b^{-2} + \boldsymbol{\Sigma}_\eta^{-1}\mu_\eta], \boldsymbol{\Sigma}_\eta^*), \tag{2.13}$$

where,
$\boldsymbol{\Sigma}_\eta^* = (\boldsymbol{\Sigma}_\eta^{-1} + \sigma_b^{-2}\boldsymbol{X_2}^T\boldsymbol{X_2})^{-1}$,
$\boldsymbol{\Sigma}_b^* = (\sigma_b^{-2}\boldsymbol{I}_k + \sigma^{-2}\boldsymbol{Z}^T\boldsymbol{Z})^{-1}$.

To generate new values from the BG model, $y, \boldsymbol{X}, \boldsymbol{Z}$ and B remain the same as in an RE model. For the particular data set in this report, the additional components of the Gibbs sampler, $\boldsymbol{X_2}$ and η take the form:

$$\boldsymbol{X_2} = \begin{pmatrix} \bar{z}_1 \\ \bar{z}_2 \\ \vdots \\ \bar{z}_n \end{pmatrix} \text{ and } \eta = \begin{pmatrix} \eta_{\bar{\boldsymbol{Z}}} \end{pmatrix}.$$

For the additional priors required for BG, we set each element of μ_η equal to zero, and each diagonal element of $\boldsymbol{\Sigma}_\eta$ equal to 1000. In the following sections, we use simulations studies to evaluate the performance of our various models in a missing data scenario.

2.3 Simulation study

We now set up a simulation study to explore the possibility of uncongeniality while imputing multilevel datasets. We also investigate the suitability of various imputation models in the presence of an omitted variable. At this stage, we focus on univariate imputation only. We generate 2500 datasets for each of the following 8 scenarios, which belong to either Case 1 (without an omitted variable) or Case 2 (with an omitted variable):

1. Case 1, Large ICC, 30% missing: data are generated from a random effects model for 1000 clusters and 5 observations in each cluster such that:

$$y_{ij} = \alpha_i + \beta z_{ij} + \varepsilon_{ij}, \tag{2.14}$$

where, y_{ij} is the response variable measured for the i^{th} cluster at the j^{th} observation, α_i are the individual intercepts generated from a $N(12.5, 4)$ distribution, z_{ij} are values for a time-varying variable generated as a combination of a cluster specific and a random component, $z_{ij} = \zeta_i + v_{ij}$, where $\zeta_i \sim N(0, 1)$ and $v_{ij} \sim N(0, 1)$ distribution, ε_{ij} are the independent errors generated from a $N(0, 4)$ distribution and $\beta = 3$. We are aware that the addition of a variable constant within a cluster within z_{ij} adds to the sluggishness of the variable, which can effect the bias and efficiency of a fixed effects model estimate for β (Plümper and Troeger, 2007; Beck and Katz, 2001), nevertheless we choose to keep the ζ_i element as it is more realistic that a cluster may have similar values for a varying characteristic over repeated measures, for instance, the magnitude of total wages may be similar for a firm over a period of five years.

We have chosen the particular parameter values for data generation above to result in a large conditional ICC, here, $4/(4+4) = 0.5$. We keep the variability explained by the covariate Z relative to the variance of the error term, equal to about half, fixed throughout our simulation. This is an additional factor that can be studied in the future. We expect that as the predictive strength of Z in the model decreases, uncongeniality between the cross combinations of FE and RE models may also be more pronounced as the predicted values will then be more dependent on the form of the model. Furthermore, we use medium sized data (5000 observations) and do not address issues particular to very small or very large datasets. One important parameter in our study is the size of the individual clusters, as these make a significant impact on the differences between the FE and RE models. We keep our repeated observations restricted to 5, as larger cluster sizes may imply that the differences between FE and RE models are not noticeable.

Data are made missing before imputation using a MAR mechanism. The missingness in Y depends on $\boldsymbol{Z}$ such that:

$$\log\left(\frac{P(R_{ij}=0)}{1-P(R_{ij}=0)}\right) = a + bz_{ij}, \tag{2.15}$$

where R_{ij} is the missingness indicator for observation y_{ij}. The coefficients a, b in (2.15) were chosen to ensure a missingness rate of around 30% on average.

2. Case 1, Small ICC, 30% missing: same as 1, with α_i generated from a $N(12.5, 0.25)$ distribution. Here we focus on generating a small conditional ICC, $0.25/(4 + 0.25)$ 0.06; this may affect the uncongeniality between various models as suggested in the literature;

3. Case 2, Large ICC, 30% missing: same as 1, with an additional covariate, W:

$$y_{ij} = \alpha_i + \beta z_{ij} + \delta w_i + \varepsilon_{ij}, \tag{2.16}$$

where, w_i is generated from a $N(3, 2)$ distribution, $\delta = 4$, and there is a correlation of $\rho = 0.7$ between w_i and z_{ij}. Here, we keep the ratio between the variance of Z and variance of W equal to 1. The OVB we expect will then be mostly a function of $\rho = 0.7$. Lesser variability in Z can further pronounce omitted variable bias, as noted in Section 2.2.1;

4. Case 2, Small ICC, 30% missing: same as 3, with α_i generated from a $N(12.5, 0.25)$ distribution;
5. Case 1, Large ICC, 70% missing: same as 1, with 70% missingness in y_{ij};
6. Case 1, Small ICC, 70% missing: same as 2, with 70% missingness in y_{ij};
7. Case 2, Large ICC, 70% missing: same as 3, with 70% missingness in y_{ij};
8. Case 2, Small ICC, 70% missing: same as 4, with 70% missingness in y_{ij}.

We then impute the missing observations of the variable y_{ij} using each of the following models:

1. the fixed effects model, FE (2.5);
2. the simple linear regression, IGN (2.6);
3. the simple linear regression with additional covariate, IGN2 (2.7);
4. the random intercepts model, RE (2.9);
5. the random intercepts model with additional covariate at level-2 of the model, BG (2.11);
6. the random intercepts model with additional covariate at level-1 of the model, HYB (2.10);
7. the sequence of simple linear regressions, WIDE (2.8).

We use the posterior predictive distribution of y_{ij} to create $m = 10$ copies for missing data.

After imputing the data, we consider how an analyst may intend to use the imputed datasets. Analysts from different disciplines may use either fixed or random effects type models depending on their aim of study and preferences. Therefore, we consider the following models as potential analysis models:

1. the fixed effects model, FE (2.5);
2. the simple linear regression, IGN (2.6);

3. the simple linear regression with additional covariate, IGN2 (2.7);
4. the random intercepts model, RE (2.9);
5. the random intercepts model with additional covariate at level-1 of the model, HYB (2.10).

The only difference between fitting the models in the imputation and the analysis stages is that imputation of y_{ij} requires a Bayesian setup while we assume that the analyst is more likely to fit a frequentist model. Given the multiply-imputed datasets, an analyst is expected to run the analysis model on each of the 10 datasets and combine the results using Rubin's rules as outlined in Section 1.1.1. For each of the analysis models, we assess the quality of our imputed data through:

1. the absolute percentage bias for the final estimate for β, the coefficient for z_{ij};
2. VR for β, where VR is defined as the median of the estimate of the variance over 2500 datasets divided by the true variance of the estimate over the same datasets. Ideally, VR must equal 1, implying that a given T^* (the total estimated MI variance, (1.3)) produces an unbiased estimate for the true variability of $\hat{\beta}$. Noticeable departures from 1 will indicate that the variance estimates are biased for the analysis model. Equivalently, the mean of the estimate of variances may be used in the numerator as well;
3. length of confidence intervals (CI) for β relative to the length of CI from the original data;
4. coverage of the true β provided by the CIs over 2500 datasets;
5. final estimates for the within error variance, σ_e^2 and the between error variance, σ_b^2, wherever RE type models are used for analysis.

All the simulations are conducted in software R (R Core Team, 2016). Data are generated using base functions such as *rnorm*. Similarly, imputation code has been written according to the descriptions in Section 2.2.2 using functions such as *rgamma* and *rmvnorm* utilising the *MASS* (Venables and Ripley, 2002) and *mvtnorm* (Genz et al., 2016) packages. We have also used the package *lme4* (Bates et al., 2015) to fit linear mixed models at the analysis stage.

Results

We use a form of 'Imputation-Analysis' expression to denote combinations of various imputation and analysis models. For example, the term HYB-FE stands for the procedure where the imputation model is HYB and the analysis model is FE. We also add the abbreviation 'imp' to denote a model as an imputation model, as opposed to an analysis model for which we use 'an'. For instance, REimp is an imputation model, while REan

is an analysis model. We also report results when each of the analysis models is run on the complete data without missingness called ORIG, and a further set of results for models run on the observed data, i.e. a complete case analysis, called CC.

Table 2.1 shows the final results for both large and small ICC scenarios for the properties of $\hat{\beta}$ in Case 1. Table 2.2 shows the estimates obtained for σ_e^2 and σ_b^2, the observation level and subject level error variance estimates for random effects analysis models. The rows represent the imputation models and the columns, the analysis models. The table is split in half to separate the large and small ICC scenarios. We observe no significant biases, as expected, not even from CC (complete-case) analysis. This is because although CC results are not only valid when the missingness is MCAR, they are also valid if the missingness is MAR when missingness only occurs in the response variable conditional on the covariates, and the missingness is independent of the response variable itself (Bartlett et al., 2014). The disadvantage of CC analysis is that it can be inefficient when missingness occurs in more than one variable and a considerable set of data are discarded (Chap. 1, Rubin (1987)). Here, we do not deal with any of these situations, and therefore, have valid CC analysis results, i.e. unbiased and efficient.

We now discuss the VRs. In our results, we may not achieve a VR of exactly 1 because of various factors that may affect the simulation study; therefore, we use the ORIG VR as benchmark for the synthetic data to achieve. In the article by Andridge (2011) about clustered randomised trials, the statistic of interest was the mean of Y. In the author's findings, the VR for the FEimp-REan combination is reported to be considerably higher than 1 for certain very small ICC scenarios (less than 0.01). For $\hat{\beta}$, we find that the VR's are generally close to 1 over 2500 datasets with large ICC, for most imputation-analysis model combinations, unless the IGN models are involved at either the imputation or the analysis stage. The VR for FEimp-REan in the small ICC scenario is equal to 1.07, which is in line with Andridge (2011)'s results. When the ICC is small, the REan variance estimates for $\hat{\beta}$ are higher than the actual variability in $\hat{\beta}$. In contrast, the VR for FEimp-FEan, and FEimp-HYBimp are slightly low at 0.92.

When Y are imputed using REimp instead, we observe slight overestimation of the variability in $\hat{\beta}$ when FEan or HYBan are used (VR $\sim$ 1.06). Given the RE model is comparatively more efficient than the FE model, the $\hat{\beta}$ used for imputations have less variability than that estimated using FEan or HYBan. In the small ICC scenario, this effect is even more pronounced; the VR for REimp-FEan and REimp-HYBan recorded is equal to 1.19. Using other models that employ random effects do not result in positively biased variance estimates even in the small ICC scenario, i.e. BGimp (VR = 0.92) and HYBimp (VR=0.94). These VR are in fact slightly low, and match the VR when FEimp is used in combination with FEan or HYBan.

The VR resulting from using WIDEimp do not seem to be significantly affected by the value of the ICC in the data, and the two sets of results are close and slightly

underestimated as compared to ORIG VR's. VR observed for CC analysis are all close to 1.

VR's for the IGNimp-FEan or IGNimp-HYBan combinations are high (~ 1.45). When IGN2imp is used instead, the VR are slightly lower but still considerably high (~ 1.22). However, VR for the IGNimp-REan and IGN2imp-REan combination are close to 1. When IGNan or IGN2an are used for analysis, the VR are either too low (~ 0.50) or too high (~ 1.92), as can be seen for the VR for ORIG as well. IGN and IGN2 are generally not suitable for analysis of hierarchical data. Nevertheless, we study these non-hierarchical analysis models, as a number of real data analysts ignore the hierarchical structure in the data. We provide a real data example in Section 3.4. Although, we note that we have not employed robust standard error estimates that are often used in combination with simple linear regressions for clustered data (Rogers et al., 1994; White, 1980).

We now compare the length of 95% CI for all combinations. The CI are always longer for FEimp (~ 1.46), as compared to imputations created using the random effects type models (~ 1.39) for all the analysis models. This is a consequence of the relative inefficiency of a fixed effects model as compared to a random effects model. The CI lengths are the longest when data are imputed using WIDEimp and the shortest for complete-case analysis (CC). We consider the length of CI jointly with the coverages provided by them, which we observe to be close to nominal throughout the table unless the IGN models are used. With coverages close to nominal, one may choose the shortest possible interval, in which case CC analysis would be preferred based alone on this criterion. Nevertheless, we note that many real datasets will have missingness in more than one variable, and a CC analysis on such a dataset can result in substantial loss of information and inefficiency.

In Table 2.2, we observe that all imputation models, with the exception of the IGNimp, IGN2imp and WIDEimp, result in unbiased estimates for σ_e^2 ($= 4$). The σ_e^2 for IGNimp and IGN2imp are biased (~ 5.89) when the analysis models are FEan, REan or HYBan; when the models are IGNan or IGN2an we expect to observe a value of 8 as the two analysis models do not separate out the within and between variances. The σ_e^2 are also also positively biased for WIDEimp (~ 4.07). However, the σ_b^2 estimates resulting after imputation using WIDEimp are unbiased in both large and small ICC scenarios. A bias in the estimate for σ_b^2 for REan and HYBan is observed when FEimp is used for imputations, as previously noted by Drechsler (2015) as well. In the small ICC scenario, most imputation models return slightly biased σ_b^2 estimates. The significant overestimation by the FEimp model is also observed, as in the large ICC scenario. The estimated ICCs have also been reported in Table 2.2; we observe that the ICC estimated are similar to those estimated using the complete data for most imputation models, with the exception of FEimp, IGNimp and IGN2imp.

Overall, we find that when the ICC is small, the variance estimation for REan may be biased when FEimp is used to the impute the data. For both large and small ICC, VR for the REimp-FEan combination is also biased upwards. Regardless of the ICC, the use of FEimp biases results for σ_b^2. WIDEimp results in positive bias for σ_e^2. So far, results using BGimp or HYBimp look suitable for all the analysis models under consideration, although some of the VR observed for BGimp are slightly low. The use of WIDEimp results in longer CI as compared to any other approach; these are generally shorter when the imputation model employs random effects.

Imputation Model		Analysis Model									
		ICC = 0.5					ICC = 0.06				
		FE	IGN	IGN2	RE	HYB	FE	IGN	IGN2	RE	HYB
ORIG	Bias	0.03	0.00	0.03	0.02	0.03	0.03	0.01	0.03	0.01	0.03
	VR	0.96	0.50	1.92	0.98	0.96	0.96	0.90	1.02	1.00	0.96
	Len	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Cov	94.48	83.04	99.28	94.72	94.48	94.48	93.84	95.16	95.04	94.48
FE	Bias	0.00	0.01	0.00	0.00	0.00	0.02	0.03	0.02	0.02	0.02
	VR	0.97	0.70	1.63	1.00	0.97	0.92	1.07	1.11	1.07	0.92
	Len	1.46	1.43	1.31	1.47	1.46	1.47	1.75	1.54	1.75	1.47
	Cov	95.04	90.76	98.76	95.36	95.04	94.80	96.16	96.40	95.36	94.80
IGN	Bias	0.00	0.00	0.00	0.00	0.00	0.03	0.02	0.03	0.02	0.03
	VR	1.45	0.66	1.73	0.97	1.45	1.24	0.92	1.26	0.94	1.24
	Len	1.71	1.51	1.31	1.60	1.71	1.34	1.50	1.31	1.45	1.34
	Cov	97.96	89.48	98.92	95.04	97.96	97.12	94.40	97.20	94.60	97.12
IGN2	Bias	0.01	0.00	0.01	0.01	0.01	0.03	0.02	0.03	0.02	0.03
	VR	1.22	0.66	1.44	1.06	1.22	0.95	0.91	0.96	0.93	0.95
	Len	1.77	1.51	1.35	1.62	1.77	1.37	1.50	1.34	1.45	1.37
	Cov	96.72	90.04	97.92	96.04	96.72	94.96	94.16	95.16	94.36	95.00
RE	Bias	0.01	0.00	0.01	0.00	0.01	0.03	0.02	0.03	0.02	0.03
	VR	1.06	0.65	1.66	0.97	1.06	1.19	0.94	1.24	0.98	1.19
	Len	1.39	1.34	1.20	1.39	1.39	1.32	1.51	1.31	1.47	1.32
	Cov	96.04	89.40	99.08	95.24	96.04	96.80	94.60	97.12	95.28	96.84
BG	Bias	0.01	0.01	0.01	0.01	0.01	0.04	0.03	0.04	0.03	0.04
	VR	0.96	0.63	1.48	0.97	0.96	0.92	0.92	0.95	0.97	0.92
	Len	1.42	1.34	1.22	1.41	1.42	1.36	1.50	1.34	1.47	1.36
	Cov	95.04	89.64	98.12	95.36	95.04	94.16	94.68	94.72	95.24	94.16
HYB	Bias	0.01	0.01	0.01	0.01	0.01	0.04	0.03	0.04	0.03	0.04
	VR	0.97	0.64	1.49	0.99	0.97	0.94	0.93	0.98	0.97	0.94
	Len	1.42	1.35	1.22	1.41	1.42	1.36	1.50	1.34	1.47	1.36
	Cov	94.80	89.16	98.36	95.28	94.80	94.68	94.60	95.24	95.16	94.68
WIDE	Bias	0.01	0.01	0.01	0.01	0.01	0.06	0.04	0.06	0.04	0.06
	VR	0.94	0.69	1.34	0.95	0.94	0.94	0.95	0.97	0.98	0.94
	Len	1.61	1.66	1.33	1.65	1.61	1.54	1.85	1.51	1.78	1.54
	Cov	94.92	89.84	97.72	94.72	94.92	94.48	94.64	94.96	95.04	94.48
CC	Bias	0.00	0.00	0.00	0.01	0.00	0.02	0.02	0.02	0.02	0.02
	VR	0.99	0.67	1.98	1.01	0.99	0.96	0.96	1.02	1.00	0.96
	Len	1.39	1.43	1.39	1.33	1.39	1.39	1.42	1.39	1.39	1.39
	Cov	94.92	89.68	99.32	95.36	94.92	94.16	94.48	94.64	95.00	94.08

Table 2.1: Properties of $\hat{\beta}$ over 2500 datasets, Case 1, Large and Small ICC, 30% data missing. True value: $\beta = 3$; sample size is 5000, 1000 clusters, 5 observations per cluster.

Imputation Model		Analysis Model									
		ICC = 0.5					ICC = 0.06				
		FE	IGN	IGN2	RE	HYB	FE	IGN	IGN2	RE	HYB
ORIG	σ_e^2	4.00	7.99	7.99	4.00	4.00	4.00	4.25	4.25	4.00	4.00
	σ_b^2	-	-	-	3.99	3.99	-	-	-	0.25	0.25
	ICC	-	-	-	0.50	0.50	-	-	-	0.06	0.06
FE	σ_e^2	4.00	9.31	9.30	4.00	4.00	4.00	5.56	5.56	4.00	4.00
	σ_b^2	-	-	-	5.31	5.31	-	-	-	1.57	1.57
	ICC	-	-	-	0.57	0.57	-	-	-	0.28	0.28
IGN	σ_e^2	5.89	7.99	7.99	5.89	5.89	4.12	4.25	4.25	4.14	4.14
	σ_b^2	-	-	-	2.10	2.10	-	-	-	0.11	0.11
	ICC	-	-	-	0.26	0.26	-	-	-	0.03	0.03
IGN2	σ_e^2	5.89	7.99	7.99	5.89	5.89	4.12	4.25	4.25	4.14	4.14
	σ_b^2	-	-	-	2.10	2.10	-	-	-	0.11	0.11
	ICC	-	-	-	0.26	0.26	-	-	-	0.03	0.03
RE	σ_e^2	4.00	7.99	7.99	4.00	4.00	3.98	4.25	4.25	3.98	3.98
	σ_b^2	-	-	-	3.99	3.99	-	-	-	0.27	0.27
	ICC	-	-	-	0.50	0.50	-	-	-	0.06	0.06
BG	σ_e^2	4.00	7.99	7.99	4.00	4.00	3.99	4.25	4.25	3.98	3.98
	σ_b^2	-	-	-	4.00	4.00	-	-	-	0.27	0.27
	ICC	-	-	-	0.50	0.50	-	-	-	0.06	0.06
HYB	σ_e^2	4.00	7.99	7.99	4.00	4.00	3.98	4.25	4.25	3.98	3.98
	σ_b^2	-	-	-	4.00	4.00	-	-	-	0.27	0.27
	ICC	-	-	-	0.50	0.50	-	-	-	0.06	0.06
WIDE	σ_e^2	4.07	8.06	8.06	4.07	4.07	4.06	4.31	4.31	4.06	4.06
	σ_b^2	-	-	-	3.99	3.99	-	-	-	0.24	0.24
	ICC	-	-	-	0.50	0.50	-	-	-	0.06	0.06
CC	σ_e^2	4.00	7.99	7.99	4.00	4.00	4.00	4.25	4.25	4.00	4.00
	σ_b^2	-	-	-	4.00	4.00	-	-	-	0.24	0.24
	ICC	-	-	-	0.50	0.50	-	-	-	0.06	0.06

Table 2.2: Average of $\hat{\sigma}_e^2$ and $\hat{\sigma}_b^2$ over 2500 datasets and resulting ICC, Case 1, Large and Small ICC, 30% data missing. True Values: For large ICC, $\sigma_e^2 = 4$, $\sigma_b^2 = 4$, for small ICC, $\sigma_e^2 = 4$, $\sigma_b^2 = 0.25$; sample size is 5000, 1000 clusters, 5 observations per cluster.

We now study the results under Case 2. Table 2.3 shows the new set of results when the datasets are designed such that the RE model results in biased $\hat{\beta}$ estimates. The magnitude of the biases can be changed by altering the simulation setup, so this is not our primary concern. We first discuss the large ICC results. Results from ORIG show that the RE model has roughly 5% bias when used to analyse the complete data without missingness, and the CC analysis shows the bias observed when only the observed data are analysed ($\sim 8.50\%$). When REimp is fitted to the data, this bias is passed on to the final analyses through the imputed values. As a result, we observe a bias of 8.50% for the REimp-REan combination. In contrast, data imputed using FEimp or HYBimp does not result in any more bias than that which is observed for ORIG for all analysis models. Similar results follow for WIDEimp and BGimp. We note that, once the bias is imputed through the use of REimp, none of the analysis models can provide reasonable estimates and coverages for $\hat{\beta}$. The bias observed for REimp-FEan, REimp-IGN2an and REimp-HYBan is 3.31%. IGN2imp does not result in biased estimates for $\hat{\beta}$ itself, and in combination with FEan and HYBan, the estimates are still unbiased. Nevertheless, the use of IGN2imp biases the REan $\hat{\beta}$ estimates even further ($\sim 13\%$). There is a slight increase in the bias for REan estimates when WIDEimp is used ($\sim 0.07\%$).

When the ICC is small, we observe similar results. FEimp, HYBimp, and BGimp do not further bias the estimates. While IGNimp and REimp bias the results for all analysis models under consideration. IGN2imp increases the bias in REan estimates as does WIDEimp, but not significantly so.

We now consider the variance ratios. Results from ORIG suggest that using REan on such data, results in negatively biased variance estimates for $\hat{\beta}$ (VR ~ 0.89). Nevertheless, the VR resulting from the use of most imputation models are less than 1 (~ 0.90) for both small and large ICC cases. The CC VR are generally closest to the ORIG VR. The second closest values are observed when HYBimp is used, or from WIDEimp when ICC is small. The VR computed after using REimp and BGimp are further away from 1 than those observed for other imputation models, if we do not consider the IGNimp and IGN2imp models.

In terms of the length of CI relative to the ORIG length of CI, the shortest CIs are provided from the CC analysis (~ 1.36), and longest from the WIDEimp approach (~ 1.55), if we don't take the IGNimp approaches into account (~ 2). There is very little difference between the length of CI for FEimp (~ 1.43) and the length of CI resulting from the use of models with random intercepts (~ 1.40). Coverages provided by the 95% CI for combinations involving any of IGNimp, IGNan, REimp and REan are unacceptably low or close to 0. Coverages resulting from the use of FEimp, BGimp, HYBimp and WIDEimp are generally close to the nominal level. The small ICC results follow similar patterns.

In Table 2.4, we report the final variance estimates for the error terms in Case 2. The variance estimates for σ_e^2, level-1 errors, are unbiased when FEimp, BGimp or HYBimp are used (~ 4.02). There is slight overestimation if REimp, WIDEimp or CC analysis are used instead (~ 4.07). The unit level variance estimates are similar for most models, except when IGNimp or IGN2imp are used, in which case, they are lower than the variances observed for the original data. There is also slight overestimation (as compared to analysis on the original data), if FEimp, REimp or CC analysis are used in combination with HYBan. As such, in this case, estimates for σ_b^2 are not noticeably more biased for FEimp, than for REimp. Variance estimates resulting from BGimp or HYBimp are generally closest to those observed for ORIG, as are the estimates from CC analysis. We find similar trends in the small ICC scenario. The ICCs estimated are also similar to those observed for ORIG for most imputation strategies, except for the use of the IGNimp and IGN2imp models.

Overall, we have observed that the HYBimp, BGimp and WIDEimp approaches work well under both the cases for the analysis models under consideration, with or without OVB, and for the different ICC values. We also know that FEimp can bias variance estimates when REan is used and REimp can be problematic when OVB exists. The WIDEimp approach, although results in longer confidence intervals than those for HYBimp or BGimp, provides reasonable results overall. In our simulation, the CC analysis also works well. However, we understand that data are not usually missing for one variable in a dataset and a CC analysis can result in waste of useful data when there are many variables with missing values.

We now turn to the problem if the rate of missingness is higher; the results can be found in Appendix A. We note that we obtain similar results as when 30% data are missing but all approaches result in comparatively longer CIs. This is extremely problematic for the WIDEimp approach, which results in nominal coverage values for interval lengths about 6 times longer than those from the original data. Similarly, biases for the σ_b^2 estimates for the FEimp-REan combination and σ_e^2 whenever WIDEimp is applied, are considerably larger.

In Case 2, the effect of OVB is even more pronounced than in the 30% missingness results. However, the trends are the same. REimp-REan results in biased $\hat{\beta}$ equal to the bias observed for the CC-REan combination. The combination of FEimp-REan biases the between variance estimates. The CI intervals from WIDEimp are substantially longer than the ORIG CI. The best results according to our criteria, are obtained when either of HYBimp or BGimp are used. As the two models are equivalent ways of modelling the effect of β, this is expected.

2.4 Conclusions

In this Chapter, we investigated the choice of models for imputation of hierarchical data. Two key concerns formed the basis of our study. Firstly, we addressed the choice between FE and RE imputation models. We confirmed that the use of FEimp in combination with random effects analysis may not be suitable, as suggested in the literature. We also discovered that using REimp may bias variance estimates for regression parameters for a fixed effects analysis as well. Secondly, we studied omitted variable bias. In the presence of OVB, REimp is not a suitable imputation model, even if the analysis model employs random effects.

Additionally, we observed that the use of non-hierarchical models, such as IGNimp or IGN2imp, for imputation may also harm analyses. The choice of the WIDEimp approach results in minor losses in data utility, and the variability in data imputed by WIDEimp can become unacceptable with high levels of missing data. Across all scenarios, the best performing models were HYBimp and BGimp, which provide a balance between the FEimp and REimp models.

One limitation of our study is that we have not compared FE and RE models when there are several observations per cluster in the hierarchical data. In such a situation, we expect the performance of the two models to be comparable and we do not know whether the use of HYBimp would be beneficial. However, we do expect that the WIDE model in this case will not perform as well, as with the increase in number of time points, the number of covariates for WIDE will additively increase. This may add unnecessary variability in the imputed data, which may be detrimental for analyses. We have also not studied very small or very large datasets. The amount of information in a given dataset may significantly alter the performance of certain models and the effects of model misspecification may also vary.

In this Chapter, we dealt with two unknown sources of variability, the omitted variable and the missingness. How each of these play a part in the imputation process and the interaction between them is difficult to quantify. In Chapter 3, we study fully observed datasets instead, while synthesising hierarchical data using multiple imputation.

Imputation Model		Analysis Model									
		ICC = 0.5					ICC = 0.06				
		FE	IGN	IGN2	RE	HYB	FE	IGN	IGN2	RE	HYB
ORIG	Bias	0.03	93.30	0.03	5.19	0.03	0.03	93.30	0.03	5.83	0.03
	VR	0.96	0.38	3.32	0.89	0.96	0.96	0.39	2.43	0.87	0.96
	Len	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Cov	94.48	0.00	99.92	0.24	94.48	94.48	0.00	99.72	0.04	94.48
FE	Bias	0.02	93.28	0.02	5.02	0.02	0.04	93.28	0.04	5.62	0.04
	VR	0.92	0.44	2.28	0.88	0.92	0.92	0.47	1.81	0.87	0.92
	Len	1.43	1.14	1.17	1.45	1.43	1.44	1.17	1.24	1.47	1.44
	Cov	94.24	0.00	99.72	9.60	94.24	94.04	0.00	98.84	4.92	94.04
IGN	Bias	26.99	93.29	26.99	48.68	26.99	27.70	93.34	27.70	50.23	27.70
	VR	1.29	0.51	1.57	0.89	1.29	1.25	0.53	1.44	0.88	1.25
	Len	2.71	1.46	1.58	2.88	2.71	2.54	1.47	1.70	2.78	2.54
	Cov	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IGN2	Bias	0.04	93.30	0.04	12.89	0.04	0.08	93.33	0.08	11.26	0.08
	VR	1.40	0.49	1.82	1.24	1.40	1.26	0.48	1.56	1.12	1.26
	Len	2.19	1.28	1.32	2.28	2.19	1.91	1.24	1.32	1.98	1.91
	Cov	97.88	0.00	99.08	0.00	97.88	97.16	0.00	98.52	0.00	97.16
RE	Bias	3.31	92.57	3.31	8.50	3.31	3.90	92.49	3.90	9.77	3.90
	VR	0.88	0.42	2.14	0.83	0.88	0.87	0.45	1.68	0.80	0.87
	Len	1.43	1.12	1.16	1.46	1.43	1.44	1.14	1.22	1.48	1.44
	Cov	41.76	0.00	76.88	0.00	41.76	27.92	0.00	51.52	0.00	27.92
BG	Bias	0.04	93.30	0.04	5.21	0.04	0.05	93.30	0.05	5.86	0.05
	VR	0.88	0.42	2.10	0.84	0.88	0.89	0.44	1.65	0.84	0.89
	Len	1.39	1.11	1.12	1.41	1.39	1.39	1.14	1.16	1.41	1.39
	Cov	94.08	0.00	99.32	6.68	94.08	93.68	0.00	98.56	3.00	93.68
HYB	Bias	0.03	93.29	0.03	5.20	0.03	0.04	93.29	0.04	5.85	0.04
	VR	0.93	0.42	2.18	0.88	0.93	0.93	0.44	1.71	0.88	0.93
	Len	1.41	1.12	1.13	1.43	1.41	1.40	1.13	1.17	1.43	1.40
	Cov	94.44	0.00	99.44	6.92	94.44	94.64	0.00	98.68	3.16	94.64
WIDE	Bias	0.01	93.29	0.01	5.27	0.01	0.04	93.30	0.04	5.95	0.04
	VR	0.92	0.46	1.99	0.90	0.92	0.96	0.49	1.62	0.92	0.96
	Len	1.55	1.22	1.18	1.58	1.55	1.57	1.26	1.24	1.60	1.57
	Cov	94.32	0.00	99.28	11.24	94.32	94.76	0.00	98.28	5.68	94.76
CC	Bias	0.02	93.30	0.02	8.49	0.02	0.04	93.34	0.04	9.76	0.04
	VR	0.96	0.51	3.84	0.82	0.96	0.95	0.53	2.95	0.79	0.96
	Len	1.36	1.39	1.46	1.35	1.36	1.37	1.39	1.51	1.36	1.37
	Cov	94.64	0.00	100.00	0.00	94.64	94.40	0.00	99.88	0.00	94.52

Table 2.3: Properties of $\hat{\beta}$ over 2500 datasets, Case 2, Large and Small ICC, 30% data missing. True value: $\beta = 3$; sample size is 5000, 1000 clusters, 5 observations per cluster.

Imputation Model		Analysis Model									
		ICC = 0.5					ICC = 0.06				
		FE	IGN	IGN2	RE	HYB	FE	IGN	IGN2	RE	HYB
ORIG	σ_e^2	4.00	24.31	13.85	4.02	4.00	4.00	20.56	10.11	4.03	4.00
	σ_b^2	-	-	-	34.30	9.86	-	-	-	30.34	6.12
	ICC	-	-	-	0.90	0.71	-	-	-	0.88	0.60
FE	σ_e^2	4.00	25.52	15.06	4.02	4.00	4.00	21.81	11.36	4.03	4.00
	σ_b^2	-	-	-	35.57	11.08	-	-	-	31.66	7.37
	ICC	-	-	-	0.90	0.73	-	-	-	0.89	0.65
IGN	σ_e^2	13.34	24.31	19.02	13.76	13.34	11.77	20.56	15.38	12.22	11.77
	σ_b^2	-	-	-	14.15	5.69	-	-	-	11.71	3.62
	ICC	-	-	-	0.51	0.30	-	-	-	0.49	0.24
IGN2	σ_e^2	8.56	24.31	13.86	8.71	8.56	6.89	20.56	10.11	7.00	6.89
	σ_b^2	-	-	-	27.29	5.30	-	-	-	25.73	3.22
	ICC	-	-	-	0.75	0.38	-	-	-	0.79	0.32
RE	σ_e^2	4.04	24.27	14.69	4.07	4.04	4.06	20.52	11.08	4.09	4.06
	σ_b^2	-	-	-	32.97	10.66	-	-	-	28.79	7.04
	ICC	-	-	-	0.89	0.73	-	-	-	0.88	0.63
BG	σ_e^2	4.00	24.31	13.85	4.03	4.00	4.00	20.56	10.11	4.03	4.00
	σ_b^2	-	-	-	34.29	9.86	-	-	-	30.34	6.12
	ICC	-	-	-	0.89	0.71	-	-	-	0.88	0.60
HYB	σ_e^2	4.00	24.31	13.85	4.02	4.00	4.00	20.56	10.11	4.03	4.00
	σ_b^2	-	-	-	34.30	9.87	-	-	-	30.34	6.12
	ICC	-	-	-	0.90	0.71	-	-	-	0.88	0.60
WIDE	σ_e^2	4.07	24.38	13.92	4.09	4.07	4.07	20.63	10.17	4.10	4.07
	σ_b^2	-	-	-	34.28	9.86	-	-	-	30.31	6.12
	ICC	-	-	-	0.89	0.71	-	-	-	0.88	0.60
CC	σ_e^2	4.00	24.30	16.08	4.07	4.02	4.00	20.56	12.38	4.09	4.04
	σ_b^2	-	-	-	32.97	13.04	-	-	-	28.79	9.23
	ICC	-	-	-	0.89	0.76	-	-	-	0.88	0.70

Table 2.4: Average of $\hat{\sigma}_e^2$ and $\hat{\sigma}_b^2$ over 2500 datasets and resulting ICC, Case 2, Large and Small ICC, 30% data missing. True Values: For large ICC, $\sigma_e^2 = 4$, $\sigma_b^2 = 4$, for small ICC, $\sigma_e^2 = 4$, $\sigma_b^2 = 0.25$; sample size is 5000, 1000 clusters, 5 observations per cluster.

Chapter 3

Multiple imputation for synthetic data and omitted variable bias

In Chapter 2 we studied the effects of OVB when imputing missing data. In this Chapter, we extend our study to the synthesis of confidential data, i.e. for SDC methods. We evaluate whether observing OVB at the synthesis stage is detrimental for analysts who use released partially synthetic data. In addition to data utility, we are now also concerned about the level of protection synthetic data provides for confidential data. Section 3.1 introduces the problem. We study the problem as in Chapter 2. Additionally, in concern for risks, we propose a new method for synthesising hierarchical data in Section 3.2. Here, we also also outline the methods for computing disclosure risks. In Section 3.3, we present a simulation study, followed by a real data application in Section 3.4. We end with concluding remarks in Section 3.5.

3.1 Introduction

The key aim of applying SDC measures to sensitive data is balancing data utility, i.e. the preservation of inferences obtained from the data, against disclosure risk, i.e., the probability of learning sensitive information about units in the data. Any high level perturbations to the data may be effective in reducing disclosure risks, but can prove detrimental for data utility. Different synthesis strategies may provide different levels of data utility and disclosure risks for the synthetic datasets. These characteristics are considered, more or less, a consequence of the synthesis model, and measures, in addition to the modelling, may be taken to reduce disclosure risks. In terms of MI techniques for SDC, the risk-utility trade-off also plays an important role in deciding whether to fully or partially synthesise the data. Details and comparisons of these are documented in various articles (Reiter, 2002; Drechsler et al., 2008a; Reiter, 2004a; Raghunathan et al., 2003). In this Chapter, we focus on partially synthetic data only. Our main interest is in

exploring this risk-utility trade off in terms of the choice of hierarchical models that we proposed in Section 2.2.2. As before, we are interested in noting how the OVB affects analysts who use released data that has been protected for confidentiality purposes. A significant addition to Chapter 2 is that we are also concerned about the risk profile for each of our synthesis models.

3.2 Methods

In this Chapter, we utilise all the models described in Section 2.2.2. However, we add to the collection of FE type models presented in Section 2.2.3 by proposing a new idea for synthesis of sensitive data:

The differences model, DIFF:

$$(y_{ij} - \bar{y}_i) = \beta(z_{ij} - \bar{z}_i) + (\varepsilon_{ij} - \bar{\varepsilon}_i). \tag{3.1}$$

The DIFF model is essentially an equivalent form of the FE model with cluster specific means subtracted from each side of the equation for each variable. If one is not interested in the cluster effects, a DIFF model can be used to estimate β with a reduction in computational burden. The DIFF estimate for β will be exactly the same as in an FE model, as are the degrees of freedom used to make inferences. Therefore, even though the unit effects do not appear in the DIFF model, the number of parameters implicitly included in the model are equal to those in an FE model, leading to the same number of residual degrees of freedom as in an FE model.

Equation (3.1) can be used to predict new values for the response variable, i.e. $(y_{ij} - \bar{y}_i)$ through the setup described in Section 2.2.2. However, the errors in (3.1) can no longer be assumed to be iid, i.e. $Var(\varepsilon_{ij} - \bar{\varepsilon}_i) \neq \sigma^2 I_{nt}$. In fact, the variance of $(\varepsilon_{ij} - \bar{\varepsilon}_i) = (1 - 1/t)\sigma^2$ and covariance between $(\varepsilon_{ij} - \bar{\varepsilon}_i, \varepsilon_{ij'} - \bar{\varepsilon}_i) = -\sigma^2/t$. For a variance-covariance matrix structure of this type, we can use generalised least squares (GLS) to predict new values for $(y_{ij} - \bar{y}_i)$. However, another way to deal with GLS is to express $Var(\varepsilon_{ij}) = \sigma^2\boldsymbol{\Sigma}$, GLS minimises $(y - \boldsymbol{X}\beta)^T\boldsymbol{\Sigma}^{-1}(y - \boldsymbol{X}\beta)$ to give an estimate of $\beta, \hat{\beta} = (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}y)$. We can also write $\boldsymbol{\Sigma} = \boldsymbol{S}\boldsymbol{S}^T$, such that $\boldsymbol{S}$ is a triangular matrix obtained through the Choleski Decomposition, which implies that we aim to minimise:

$$(y - \boldsymbol{X}\beta)^T\boldsymbol{S}^{-T}\boldsymbol{S}^{-1}(y - \boldsymbol{X}\beta) = (\boldsymbol{S}^{-1}y - \boldsymbol{S}^{-1}\boldsymbol{X}\beta)^T(\boldsymbol{S}^{-1}y - \boldsymbol{S}^{-1}\boldsymbol{X}\beta),$$

which is equivalent to regressing $\boldsymbol{S}^{-1}\boldsymbol{X}$ on $\boldsymbol{S}^{-1}y$ (Chap. 5, Faraway (2002)). This implies that we can replace $\boldsymbol{X}$ with $\boldsymbol{X}' = \boldsymbol{S}^{-1}\boldsymbol{X}$ and y with $y' = \boldsymbol{S}^{-1}y$ and simply follow the procedure in Section 2.2.2. Here,

$$y = \begin{pmatrix} y_{11} - \bar{y}_1 \\ y_{12} - \bar{y}_1 \\ \vdots \\ y_{1t} - \bar{y}_1 \\ y_{21} - \bar{y}_2 \\ \vdots \\ y_{2t} - \bar{y}_2 \\ \vdots \\ y_{nt} - \bar{y}_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} z_{11} - \bar{z}_1 \\ z_{12} - \bar{z}_1 \\ \vdots \\ z_{1t} - \bar{z}_1 \\ z_{21} - \bar{z}_2 \\ \vdots \\ z_{2t} - \bar{z}_2 \\ \vdots \\ z_{nt} - \bar{z}_n \end{pmatrix} \text{ and } B = \begin{pmatrix} \beta \end{pmatrix}.$$

As we know the exact form of $\boldsymbol{\Sigma}$ in this case, this method allows an easy to use computational setup for GLS. It also means that once the predictions for y' are obtained, these must be scaled back to y by multiplying the new y' with $\boldsymbol{S}$, as $\boldsymbol{S}(\boldsymbol{S}^{-1}y) = y$. We note that $\boldsymbol{\Sigma}$ is not a full rank matrix, and hence cannot be inverted. We work around this problem by using a generalised inverse instead (Moore, 1920).

The above procedure will give us a set of new values for $y_{ij} - \bar{y}$. To be able to complete the process, we need to add either new or the original values of $\bar{y}_i$ to the predicted values from (3.1). We consider two different ways to find $\bar{y}_i$ for synthesis here:

1. modelling the $\bar{y}_i$, option DIFF1, such that the MEAN model is;

$$\bar{y}_i = \gamma_0 + \gamma_1 \bar{z}_i + \bar{\varepsilon}_i. \tag{3.2}$$

   Again, to predict new values for $\bar{y}_i$, we can use the setup in Section 2.2.2 with:

$$y = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & \bar{z}_1 \\ 1 & \bar{z}_2 \\ \vdots & \vdots \\ 1 & \bar{z}_n \end{pmatrix} \text{ and } B = \begin{pmatrix} \gamma_0 \\ \gamma_1 \end{pmatrix};$$

2. or, preserving the original $\bar{y}_i$ from the data set, option DIFF2. These can be obtained by calculating the observed values of $\bar{y}_i$ from the original data set:

$$\bar{y}_i = \sum_j y_{ij}/t, \tag{3.3}$$

   where, t is the number of repeated measures or time points per cluster.

The drawback of using the DIFF1 approach is the inability of the model for $\bar{y}_i$ to estimate different intercepts for each unit. This may affect the quality of the imputations, therefore affecting the utility of the synthesised datasets. On the other hand, the DIFF2 approach may be expected to preserve the utility of the synthesised data set better, but by compromising on its disclosure risks, i.e. keeping the means in the real data set non-synthesised.

We also note that the DIFF approach, although more complicated to implement than other approaches for synthesis, also has the advantage of providing the flexibility to model $\bar{y}_i$ independently of $y_{ij} - \bar{y}_i$. Since we are focusing on a data structure where only the unit intercepts are different, the synthesis of $\bar{y}_i$ is an important feature of the synthesis process. Our strategy is motivated by one of the reasons that make hierarchical data more disclosive than non-hierarchical data. The release of $Y_1 \dots Y_5$ is not the equivalent of releasing information of five variables, but even more. A base intercept or slope determined from $Y_1 \dots Y_5$ could be an important piece of information for an intruder, and synthesising these carefully can promise better privacy. In this simulation, we only considered two approaches that could be used to obtain $\bar{y}_i$; we expect that more strategies can be developed in the future.

Existing research has so far not considered the DIFF approach as a strategy to synthesise sensitive data, or even as an imputation model for missing data. We expect that the flexibility of the DIFF approach can help data keepers balance the trade-off between utility and disclosure risks by directly modifying the way $\bar{y}_i$ is modelled.

3.2.1 Disclosure risks

To evaluate disclosure risks, we follow the model-based approach used by Reiter (2005b); Reiter and Mitra (2009) and Drechsler and Reiter (2008). We use various risk measures to evaluate the relative risks posed by synthesising data using our synthesis models.

The approach assumes that the intruder possesses information on some quasi-identifiers for a target of his interest, call this information vector t. Let the multiple synthetic datasets be $\boldsymbol{D} = \{\boldsymbol{D}^1, \boldsymbol{D}^2, \dots, \boldsymbol{D}^m\}$. Let t_0 be a unique identifier for the record with the intruder, and d_{j0} the unique identity (a name or ID) of a record j for $j = 1, \dots, n$ in the original dataset, which is not released. Let M denote any information on how the synthetic datasets were created, for instance, the model used.

The intruder's aim is to declare a match between the j^{th} record in $\boldsymbol{D}$ to his target when $d_{j0} = t_0$ and not to match when $d_{j0} \neq t_0$. Consider J a random variable that equals j when $d_{j0} = t_0$. The intruder is interested in calculating $Pr(J = j|t, \boldsymbol{D}, M, \boldsymbol{D}_S)$ for $j = 1, \dots, n$, where $\boldsymbol{D}_S$ contains the unreleased original values of all observations that are synthesised. The intruder then has to decide whether the maximum of the probabilities for all $j = 1, \dots, n$ is large enough to declare to a match. As the intruder

does not know the values within $\boldsymbol{D}_S$, he must integrate out its distribution to compute the match probabilities:

$$Pr(J = j|t, \boldsymbol{D}, M) = \int Pr(J = j|t, \boldsymbol{D}, M, \boldsymbol{D}_S) Pr(\boldsymbol{D}_S|t, \boldsymbol{D}, M) d\boldsymbol{D}_S. \tag{3.4}$$

A Monte Carlo approach can be used to calculate the value of (3.4). This implies that the intruder first needs a draw from $Pr(\boldsymbol{D}_S|t, \boldsymbol{D}, M)$, call this $\boldsymbol{D}_{new}$ and then use this draw to calculate $Pr(J = j|t, \boldsymbol{D}, M, \boldsymbol{D}_S = \boldsymbol{D}_{new})$ by matching D_{new} to his information t. For discrete data, the matches can be made exactly; for continuous data, some distance-based measures can be used. Essentially, the intruder repeats the synthesis process. A draw of $\boldsymbol{D}_{new}$ represents a possible original value in D_S that resulted in the synthetic data $\boldsymbol{D}$ using the synthesis process. If the intruder creates many, say K, such plausible values for a synthetic observation in the dataset, he can calculate how near his new draws are to t for each of the K times.

For our research, we assume that M is empty. Therefore, the intruder does not have any information to recreate the synthesis process. In this instance, the intruder can generate $\boldsymbol{D}_{new}$ by fitting a plausible model to the released data. Here, we use a simplified approach instead, assuming that the best draws for $\boldsymbol{D}_{new}$ are the multiple copies of the synthetic data themselves and the prior information held by the intruder, t, are the original data themselves. The intruder can then attempt to match t to each of the m copies of the data released, and average the probabilities of matches over m.

For example, if N^i is the number of records in $\boldsymbol{D}^i$ that are declared matches for a record j in one of the m datasets, then:

$$Pr(J = j|t, \boldsymbol{D}, M, \boldsymbol{D}_S) = \frac{1}{m} \sum_i \frac{1}{N^i} \mathrm{I}(\boldsymbol{D}^i_{new,j} = t), \tag{3.5}$$

where, $\mathrm{I}(\boldsymbol{D}^i_{new,j} = t) = 1$, when record j is a match and 0 otherwise. The probability (3.5) is averaged over the m copies.

If the exercise is repeated for all n units in a dataset, an $n \times n$ matrix can be constructed that displays each record's probability of match against each target record. In our simulation, we assume that all n units are targets. A number of summary measures from this matrix can then be used as disclosure risk measures. Following Reiter (2005b), Drechsler and Reiter (2008) and Chap. 7, Drechsler (2011b), we discuss three measures here: 1) the expected match risk, 2) the true match rate and 3) the false match rate. Let c_j be the number of records with the highest match probability for the target t_j where $j = 1, \ldots, n$. Let $I_j = 1$ if the true match is among the c_j units and $I_j = 0$ otherwise. Let $K_j = 1$ when $c_j I_j = 1$ and $K_j = 0$ otherwise. The expected match risk equals $\sum_j (1/c_j) I_j$. The true match rate equals K_j/n. Let $F_j = 1$ when $c_j(1 - I_j) = 1$

and $F_j = 0$ otherwise and, let u equal the number of records with $c_j = 1$. The false match rate equals F_j/u.

The expected match risk is the sum of probabilities obtained when the correct target and synthetic records are matched. Summed over all the n units in consideration, this value represents the average number of correct matches an intruder can obtain, by randomly guessing the correct match. To explain the true match rate, we first assume that it is prudent for an intruder to declare a match for the record with the highest value for (3.5). If we sum the number of times the highest probabilities are uniquely obtained for the correct targets, we obtain the number of correct matches the intruder actually makes. The number of correct matches divided by the total number of units in the dataset is called the true match rate. Instead, if we sum the number of times the highest matching probabilities are uniquely obtained for the wrong targets, we obtain the false match risk. This can be converted into a rate by dividing the number of false matches by the number of times the intruder declares a unique match.

We now list the step by step calculations involved in measuring disclosure risk using the process described above applied to our simulation studies. We note that there are at least two different ways to see our application. Firstly, if Y is a sensitive variable, our assessment of disclosure risks is a measure of how well an intruder knows about the sensitive values in the real dataset, i.e. attribute disclosure risk. Secondly, if Y is instead used as a key variable, i.e. used to identify an individual, the disclosure risk measures identification risk.

- Treat $Y_1, \ldots, Y_5$ as 5 different variables belonging to the same unit. First, we calculate the Euclidean distance between the real and synthetic datasets, i.e. differences between all the values of the various variables, squared and added and then, we take square root of the total.
- For each real unit, calculate this distance against each synthetic unit.
- Set a calliper, such that the Euclidean distance less than this calliper is a 'match'.
- Record for each real unit, the probability of match, so if three synthetic units match a real unit, the probability of match for each of these synthetic units is 1/3.
- Repeat the calculation for $m = 10$ datasets, and average the probabilities across the 10 synthetic datasets.
- Create a matrix of the averaged probabilities for all units in the data, such that the columns represent the real units and the rows represent the synthetic units.
- Investigate disclosure risks through various measures using this matrix; we utilise:
  - the trace of the matrix, the *expected match rate*,

  - the number of correct matches as a proportion of the total number of individuals, the *true match rate*, and
  - the number of false matches as a proportion of the total number of unique matches, the *false match rate*.

- Repeat the above calculation for various callipers (we use the integers $1, \ldots, 20$).

We also consider that it is prudent for an intruder to declare a true match not only on the basis of a unique maximum probability of match, but also on the size of this probability. Throughout our calculations, we assume that the intruder searches for probabilities which are at least equal to 0.1. In some of our analyses, we also consider another risk measure, the *perceived match risk*; this is the number of unique matches made by the intruder, whether correct or false, given a probability threshold of 0.2. Hence, it is a measure of the number of matches the intruder believes he has made, if he relies on unique matches alone.

3.3 Simulation study

We utilise a simulation setup similar to the one described in Section 2.3. However, here we use data that are fully observed. We are interested in replacing all observations for a variable, Y. Data are simulated using the following four scenarios:

1. Case 1, Large ICC: data are generated from a random effects model for 1000 clusters and 5 observations in each cluster such that:

$$y_{ij} = \alpha_i + \beta z_{ij} + \varepsilon_{ij}, \tag{3.6}$$

   where, y_{ij} is the response variable measured for the i^{th} cluster at the j^{th} observation, α_i are the individual intercepts generated from a $N(12.5, 4)$ distribution, z_{ij} are values for a time-varying variable generated as a combination of a cluster specific and a random component, $z_{ij} = \zeta_i + v_{ij}$, where $\zeta_i \sim N(0, 1)$ and $v_{ij} \sim N(0, 3)$ distribution, ε_{ij} are the independent errors generated from a $N(0, 4)$ distribution and $\beta = 3$.

2. Case 1, Small ICC: same as 1, with α_i generated from a $N(12.5, 0.25)$ distribution;

3. Case 2, Large ICC: same as 1, with an additional covariate, W:

$$y_{ij} = \alpha_i + \beta z_{ij} + \delta w_i + \varepsilon_{ij}, \tag{3.7}$$

   where, w_i is generated from a $N(3, 2)$ distribution, $\delta = 4$, and there is a correlation of $\rho = 0.7$ between w_i and z_{ij};

4. Case 2, Small ICC: same as 3, with α_i generated from a $N(12.5, 0.25)$ distribution.

The imputation, here, synthesis, procedure is the same as in the case of missing data. A Bayesian setup can be used for each of the synthesis models to generate synthetic values for Y multiple times. An analyst repeats his analysis procedure on the multiple synthetic datasets and uses the combining rules for synthetic data, as described in Section 1.1.1, to obtain the final results. However, Reiter and Kinney (2012) showed that generating synthetic data from posterior predictive distributions is unnecessary for partially synthetic data. Using a simulation study, the authors demonstrated that point and variance estimates, and thus coverage, of a statistic of interest remain close to the inferences observed using posterior distributions, if plug-in estimates are used in the synthesis stage instead. The motivation is practical. Not only is this computationally easier and quicker, but also facilitates using synthesis models that are more complicated to set up as Bayesian models. Reiter and Kinney (2012) note that not using posterior predictive distributions makes the posterior draws of the parameters less variable, and consequently, the variance estimates of a statistic of interest are smaller. We note that this, although good for data utility, may also have adverse consequences for data confidentiality. To the best of our knowledge, the confidentiality implications of using plug-in estimates have not been investigated in the literature. In our simulation, we will put this to test for a hierarchical data structure recording how using maximum likelihood estimates as plug-in estimates affects data confidentiality.

Therefore, for all data generation mechanisms described above, and all synthesis strategies considered, we synthesise data using two approaches:

1. the posterior predictive approach: predictions are made using the posterior predictive distribution of the sensitive data;
2. the MLE approach: predictions are made by using plug-in MLE estimates for the model parameters.

We consider the following models for generating synthetic data:

1. the fixed effects model, FE (2.5);
2. the differences model with means synthesised using a simple linear regression model, DIFF1 (3.1) and (3.2);
3. the differences model with original means preserved, DIFF2 (3.1) and (3.3);
4. the simple linear regression, IGN (2.6);
5. the simple linear regression with additional covariate, IGN2 (2.7);
6. the random intercepts model, RE (2.9);

7. the random intercepts model with additional covariate at level-2 of the model, BG (2.11);
8. the random intercepts model with additional covariate at level-1 of the model, HYB (2.10);
9. the sequence of simple linear regressions, WIDE (2.8).

For each of the models above, we generate $m = 10$ copies for Y. We then analyse the data using the following models:

1. the fixed effects model, FE (2.5);
2. the simple linear regression, IGN (2.6);
3. the simple linear regression with additional covariate, IGN2 (2.7);
4. the random intercepts model, RE (2.9);
5. the random intercepts model with additional covariate at level-1 of the model, HYB (2.10).

We explore congeniality by studying various combinations of synthesis-analysis models. We summarise our results for the following properties averaged over 2500 datasets:

1. the absolute percentage bias for the final estimate for β, the coefficient for z_{ij};
2. VR for β, where VR is defined as the median of the estimate of the variance over 2500 datasets divided by the true variance of the estimate over the same datasets;
3. length of confidence intervals (CI) for β relative to the length of CI from the original data;
4. coverage of CI for β over 2500 datasets;
5. final estimates for the within error variance, σ_e^2 and the between error variance, σ_b^2, wherever RE type models are used for analysis.

As in Chapter 2, all simulations are run in software R (R Core Team, 2016) with the packages mentioned earlier. For the MLE approach to synthesis, the exact expressions for the maximum likelihood parameter estimates are coded in for the fixed effects type of models. For the models employing random effects, the package *lme4* (Bates et al., 2015) is utilised to obtain MLE estimates.

3.3.1 Data utility

Case 1 results

We set our expectations from the various models by first testing them out in Case 1, where the data generation process is a straightforward random intercepts model and no omitted variable exists. Table 3.1 shows the absolute percentage bias observed in the estimation of β. We observe that none of the combinations result in biased estimates in Case 1, not even the IGN model, which completely ignores the hierarchical structure of the data. As far as the bias is concerned, one can argue that IGNsyn may be simplest and quickest to run amongst all the models studied, and the use of more complicated models may be unnecessary.

We now move on to the variance estimate for $\hat{\beta}$. We compare our final estimates of variance of $\hat{\beta}$ against the variance estimates produced by the ORIG dataset using the same analysis model. We expect that the $\bar{u}$ component of the final variance, T^*, matches the variance estimates obtained at the synthesis stage, while the B component is small, as data are fully observed. As expected, we find that the B estimated are nearly 0; we plot $\bar{u}$ estimates from the 2500 datasets against the variances estimated from the original data for all combinations of synthesis-analysis models. Here, we only discuss some of these plots for our synthesis-analysis combinations of interest. Figure 3.1 shows the $\bar{u}$ estimates obtained for four different analysis models when the data have been synthesised using FEsyn for the large ICC case. We observe that FEsyn results in an inflation of variance for the IGNan and REan analysis models. This inflation is even more serious when the between variability amongst sample units is very small (see Figure 3.2 for the small ICC case).

When REsyn is used instead, none of the variance estimates for $\hat{\beta}$ suffer from inappropriate variability (see Figure 3.3). The plots for DIFF1syn, DIFF2syn, HYBsyn and BGsyn also look similar to that of REsyn's (plots not shown).

Perhaps the most interesting case is the WIDEsyn model. We already know that WIDEsyn performs well in terms of bias in exchange for variability, and is a popular approach in the literature. Here, we see that the WIDEsyn model always increases the variability in $\hat{\beta}$, no matter what the analysis model (see Figure 3.4). This is most probably due to a lack of model selection at the synthesis stage, as this implies that all variables were used to synthesise each sensitive variable as proposed in the literature. Many of the estimated coefficients in each step of the WIDEsyn model were very small, contributing little to the model, but induced more variability in the synthetic Y overall. In practice, we would expect some model selection to take place, as real datasets may contain several variables. This is an example of a trade-off between keeping as many covariates in the model as possible to prevent uncongeniality, and controlling the variability in final parameter estimates.

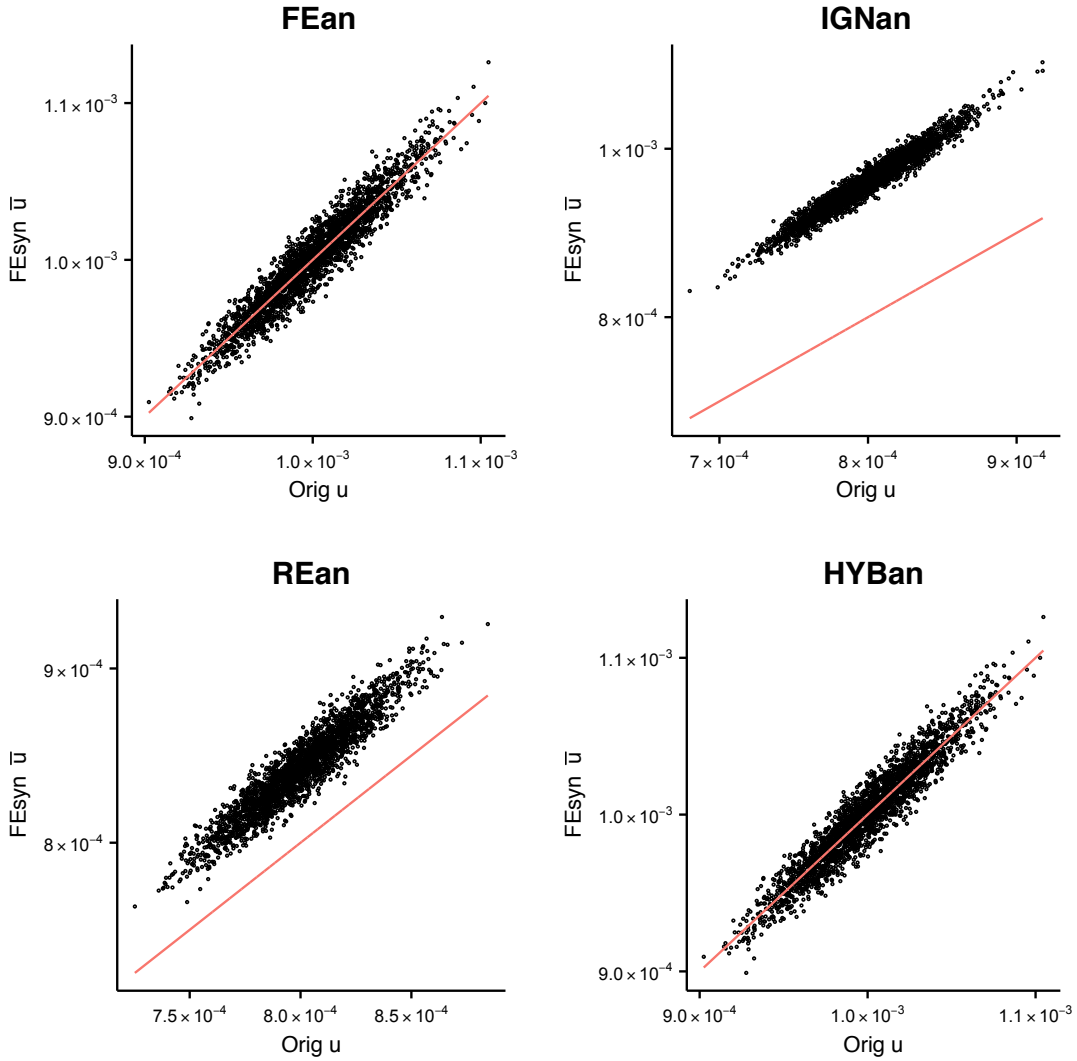

Figure 3.1: Posterior predictive approach: FEsyn $\bar{u}$ for various analysis models, Case 1, Large ICC. Plotted line (red) is $y = x$.

Overall, we have identified a few models that stand out as always having appropriate $\bar{u}$ estimate on average for the Case 1 data, for each of the analysis models in consideration. These include DIFF1syn, DIFF2syn, HYBsyn, BGsyn and REsyn. IGNsyn and IGN2syn inflate the variances considerably, unless the analysis model is also IGNan. We have also identified the price of using the WIDEsyn model as opposed to the rest of the hierarchical models in terms of added variability. The real question is whether this has serious consequences for analysts.

So far, we have addressed the issue of congeniality by comparing mean and variance estimates for our synthesis and analysis models. A related but separate concern is whether the final estimate of variance, T^* correctly represents the variability in estimates $\hat{\beta}$. We measure this by calculating *Variance Ratios*, abbreviated as VR for $\hat{\beta}$. As expected, the VR observed after fitting the various analysis models to the original data

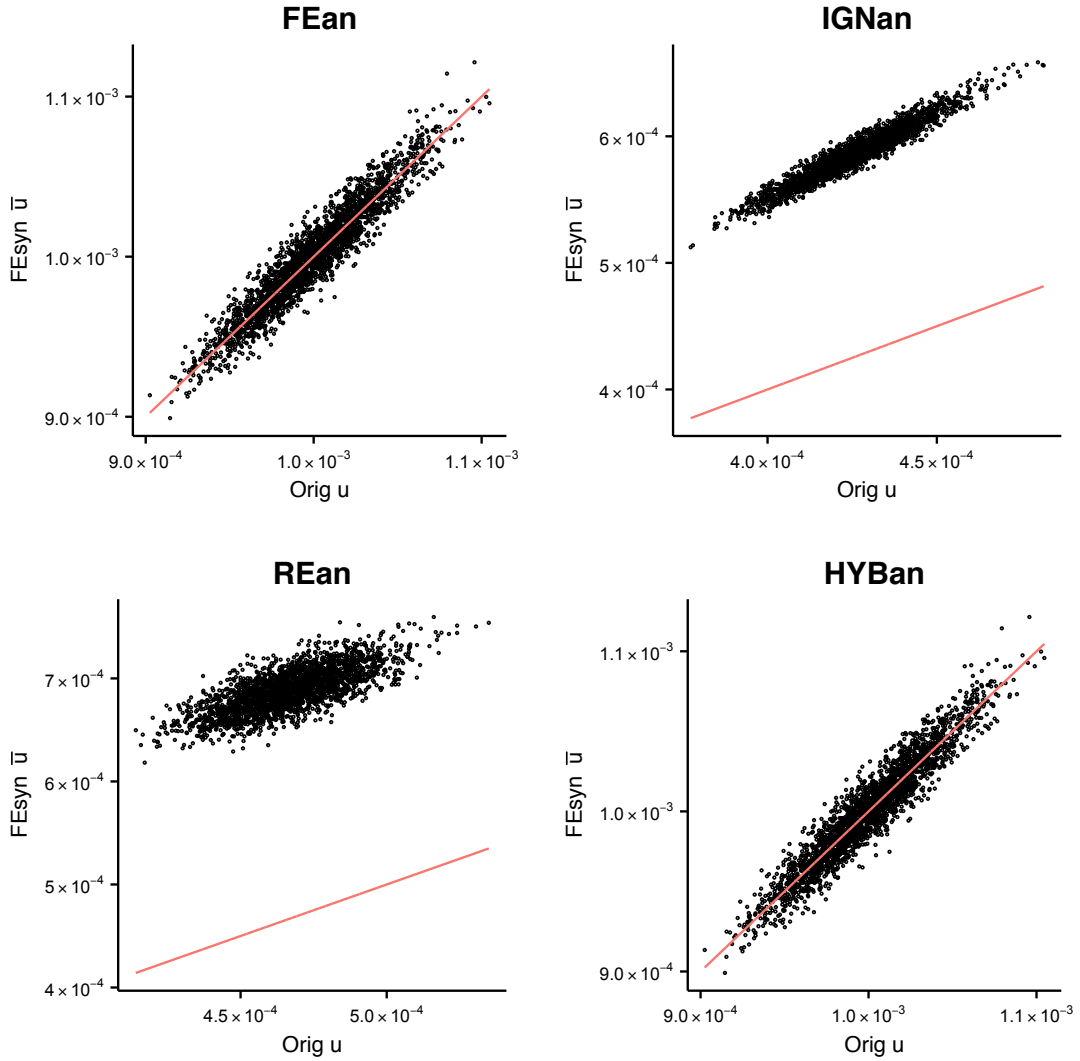

Figure 3.2: Posterior predictive approach: FEsyn $\bar{u}$ for various analysis models, Case 1, Small ICC. Plotted line (red) is $y = x$.

are generally close to 1, except for the IGNan and IGN2an models, which do not account for the hierarchical structure in the data. The VR for FEan and HYBan for both the large and small ICC cases are around 0.96. There could be two main reasons why these are not exactly 1: 1) the number of datasets, i.e., 2500 in this simulation are not enough to compute the true variability in $\hat{\beta}$, or 2) the analysis models do not match the data generation mechanism. We extended the simulation to 5000 datasets and did not observe different results. Therefore, we believe that this small discrepancy occurs because our data generation mechanism is an RE model, and not HYB or FE. Changing the data generation mechanism to FE solves this issue, but then the observed VR for REan are too small. Generating the covariate z_{ij} with a time-constant element also adversely effects the VR's. Currently, we keep a time-constant element in the generation of z_{ij}, as it is more realistic that a given unit in the data has similar observations for continuous variables over time. In our simulation studies, we have chosen to generate data from

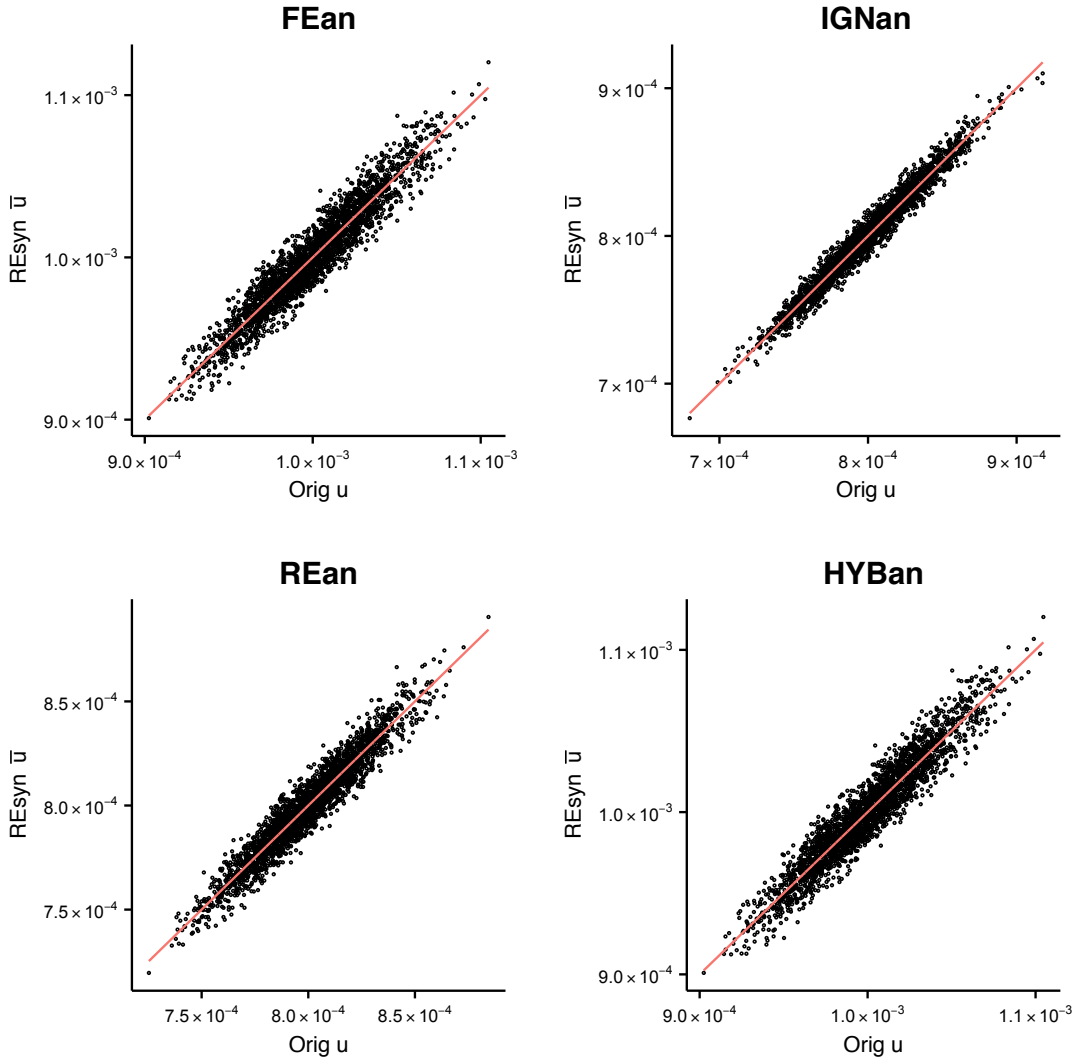

Figure 3.3: Posterior predictive approach: REsyn $\bar{u}$ for various analysis models, Case 1, Large ICC. Plotted line (red) is $y = x$.

the RE model, so that REsyn and REan are not disadvantaged any further than they would be when we introduce omitted variable bias (OVB) in the simulation. Hence, we consider 0.96 to be acceptable values for FEan and HYBan VR's in Case 1. Moreover, as our main interest is in replicating the results from the original data (ORIG), we assess the synthesis models relative to the results observed for ORIG, rather than absolutely.

From Table 3.1, we now consider the VR for the various synthesis-analysis combinations. We note that the use of FEsyn, inflates the variance estimates for REan, especially in the case of small ICC, when the between variability is very small. Given the $\bar{u}$ estimates from the Figures 3.1 and 3.2, this result is expected. We find that DIFF1syn, DIFF2syn, HYBsyn, BGsyn and WIDEsyn result in VR close to those observed for the original data. As in the missing data simulation, it is interesting to note that the use of REsyn inflates VR (1.16) for FEan and HYBan, and this is despite the fact that the $\bar{u}$ estimates from

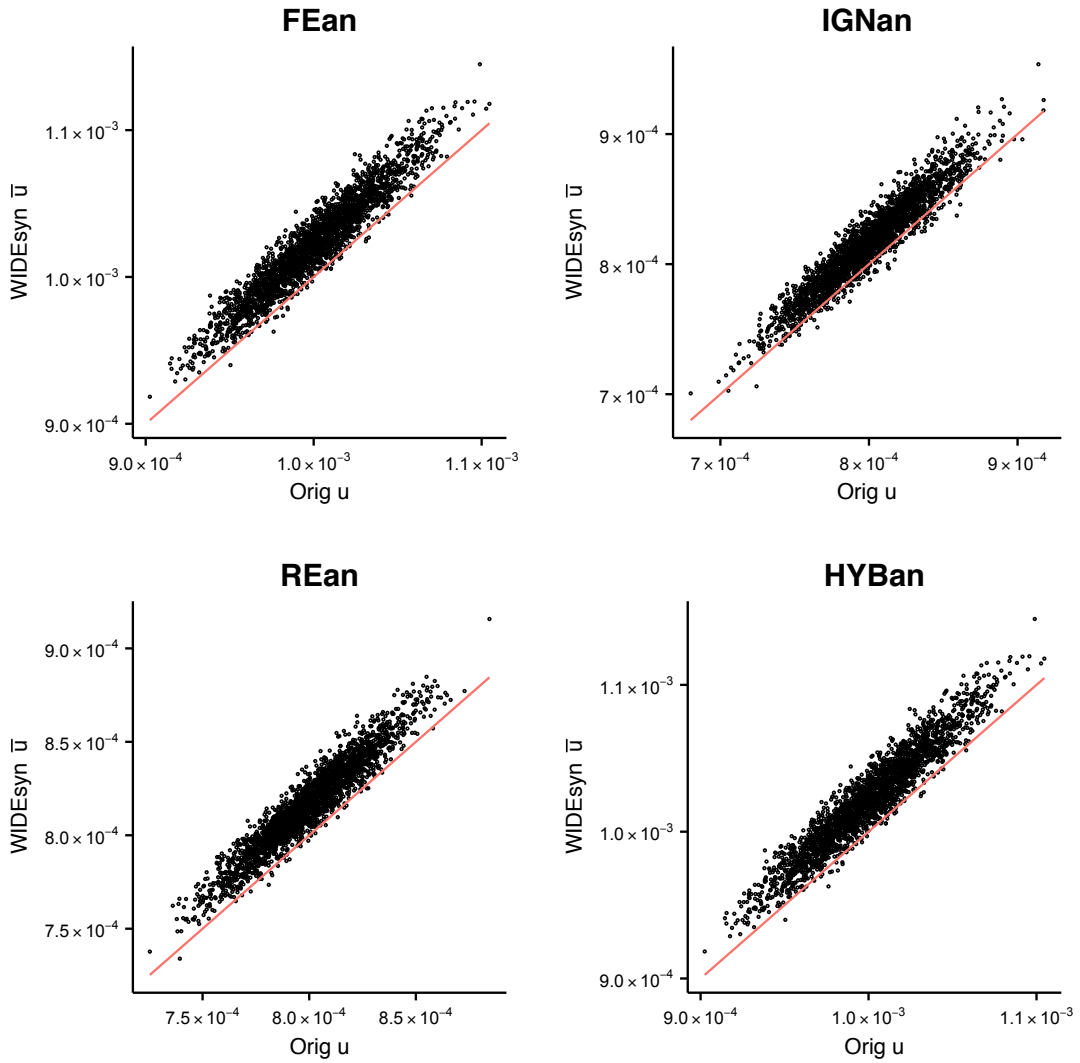

Figure 3.4: Posterior predictive approach: WIDEsyn $\bar{u}$ for various analysis models, Case 1, Large ICC. Plotted line (red) is $y = x$.

REsyn matched the original variance estimates on average, as observed in Figure 3.3. In fact, the results for REsyn VR match closely the results from IGNsyn VR for FEan and HYBan. The VR for REsyn-REan is observed to be 0.89, which is too low. We found that this is mainly affected by the time-constant element in z_{ij}. As mentioned before, it is reasonable to assume that z_{ij} are similar across time points/ repeated observations for a given unit. However, here, we see that this is detrimental for REsyn, especially when combined with REan in the case of large ICC.

We now discuss the final length of 95% confidence intervals (CI) for $\hat{\beta}$ relative to the length of CI observed from ORIG. As expected, after data synthesis, the variance estimates for $\hat{\beta}$ are slightly larger than those for ORIG, resulting in longer CI. For most of the synthesis models in Table 3.1, we observe CI about 1.10 times the length of ORIG

CI. This number is 1.11 for the WIDE model, as expected, given the observations in Figure 3.4. For REsyn, the length of CI are around 1.09 times the length of ORIG CI. We combine these observations with the coverage of CI. Here, FEsyn, DIFF1syn, DIFF2syn, BGsyn, HYBsyn and WIDEsyn, all provide about 95% coverage for appropriate analysis models. Using REsyn, however, results in slight overcoverage when HYBan or FEan are used. In the case of small ICC, large variance estimates, as noticed in the VR's, results in about 99% coverage for FEan and HYBan. For IGNsyn and IGN2syn the lengths of CI (1.56) are too large, resulting in high coverages (97.20). This is less of a problem in the case of small ICC, where the use of non-hierarchical models is arguably more justifiable.

Table 3.2 shows the final within and between error estimates from the random effects type models after applying various synthesis models. As expected, FEsyn inflates the between error estimate for REan and HYBan in both small and large ICC cases. The synthesis models employing random effects (REsyn, BGsyn, HYBsyn) also result in slightly inflated between error variance when the ICC is small. Analysis after using the IGNsyn and IGN2syn models result in the within error variance estimate of around 8, which is the sum of the original within error variance of 4 and the variance of the unit effects of 4. In the case of WIDEsyn, we see the expected inflation of within error estimates (4.09) and in the case of large ICC, the between error estimate is also slightly biased upwards (4.07). When data ICC is small, the between error estimates for WIDEsyn are unbiased on average. As observed in the missing data simulations in Chapter 2, the ICCs estimated are biased for FEsyn, IGNsyn and IGN2syn.

Case 2 results

We now consider the same set of results in Case 2, where we introduce OVB deliberately. Table 3.3 shows the biases we expect to observe for ORIG. Both the REan and IGNan show biased (93% and 5%) results. We do not delve into the size of the bias, as this is adjustable by changing the data generation process. We are more interested in understanding how the bias at the synthesis stage affects the final analyses. Using REsyn induces this bias within the synthetic data, and regardless of the analysis model, the $\hat{\beta}$ estimates are always biased at the analysis stage. When REsyn is used in combination with REan, the bias nearly doubles from the synthesis to the analysis stage (from 5% to 10%). The same cannot be said for the IGNan model, for instance, where the final bias stays around 90% regardless of the synthesis model. All the other models do not induce such a bias in the synthetic data. While IGN2syn does not result in biased estimates for FEan and HYBan, it does inflate the REan bias to about 27%.

As before, we now consider the $\bar{u}$ estimates to the variance estimates from the original data. We found these results largely unchanged from Case 1 to Case 2, except for FEsyn and REsyn. FEsyn no longer seems to inflate $\bar{u}$ estimates for REan (see Figure 3.5), and REsyn results in very slightly inflated $\bar{u}$ estimates for all the analysis models (see Figure

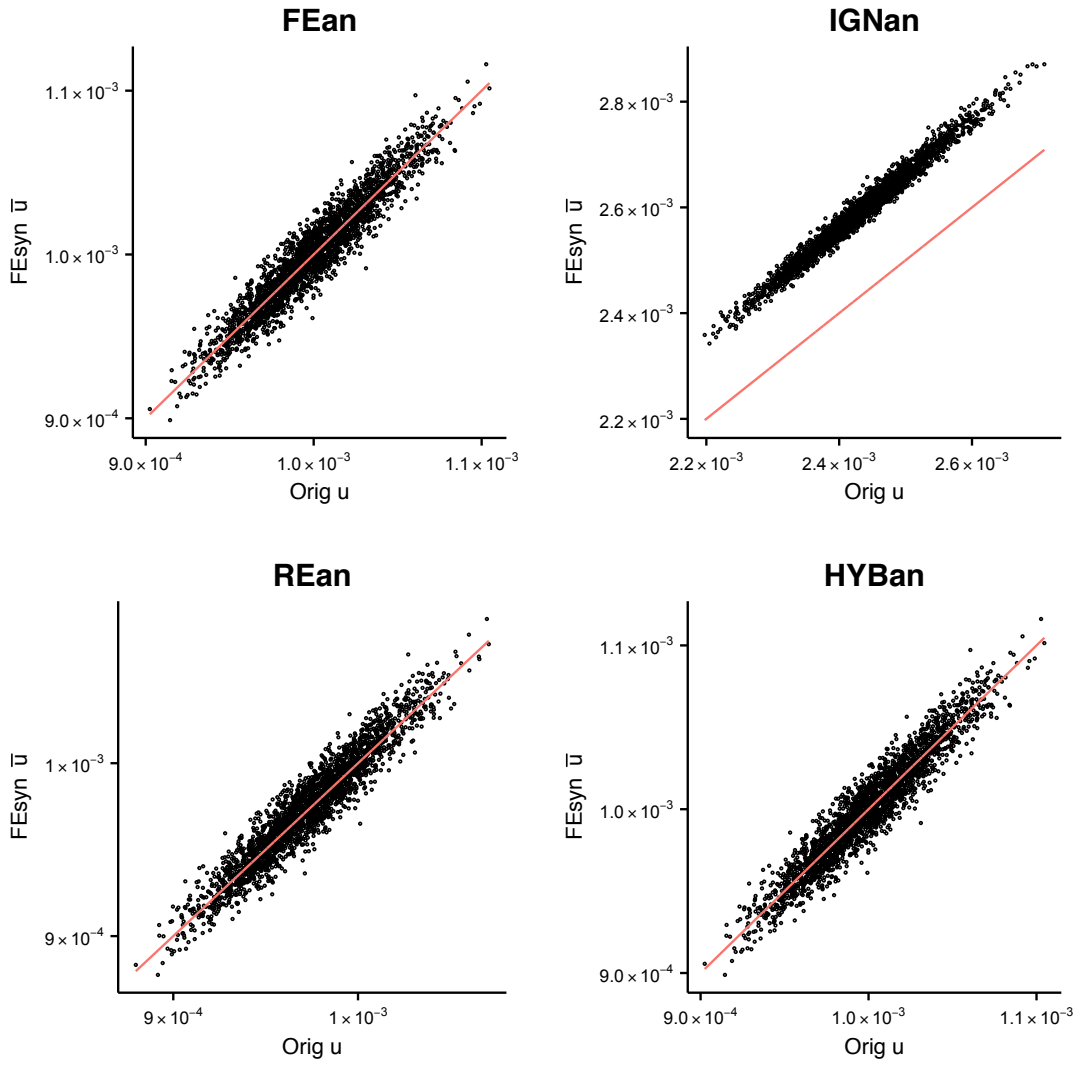

Figure 3.5: Posterior predictive approach: FEsyn $\bar{u}$ for various analysis models, Case 2, Large ICC. Plotted line (red) is $y = x$.

3.6). Again DIFF1syn, DIFF2syn, BGsyn and HYBsyn $\bar{u}$ estimates match the original variances on average, regardless of the choice of analysis models in this simulation.

We now discuss the VRs in Table 3.3 for Case 2. IGNsyn and IGN2syn show considerable overestimation, with ratios of around 2.3. REsyn eventually results in VRs which are slightly low, around 0.93 and even lesser when ICC is small. HYBsyn, WIDEsyn and all fixed effects type approaches result in ratios close to the VR from ORIG, as long as the analysis model takes the omitted variable into account. As an analysis model, REan always results in low VR, even when using ORIG data, implying tighter inferences than would be ideal. This level of confidence, combined with biased $\hat{\beta}$ can make the inferences for β even worse, and is undesirable. It is also interesting to note that VR for BGsyn model are also low, ranging from 0.77 for REan (small ICC) to about 0.92 for FEan and HYBan (large ICC).

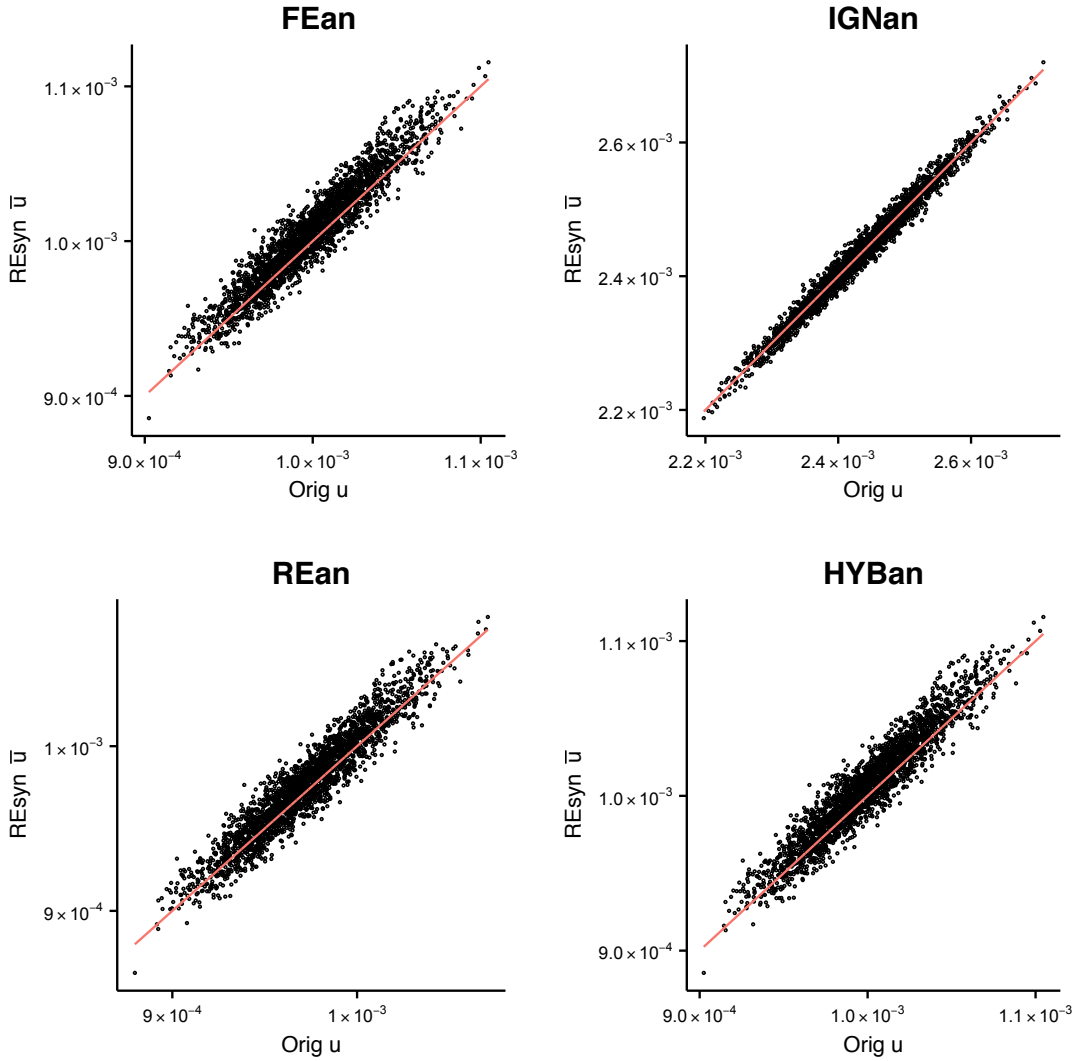

Figure 3.6: Posterior predictive approach: REsyn $\bar{u}$ for various analysis models, Case 2, Large ICC. Plotted line (red) is $y = x$.

Our observations regarding the length of CI in Case 1 also hold true in Case 2, with length of CI nearly 1.10 times the ORIG length of CI for most synthetic data, slightly wider for the WIDEsyn model at about 1.11 and considerably wider for the IGNsyn and IGN2syn model at about 2. The coverages are also as expected. REsyn fails to cover the true value for most of the datasets, while IGN2syn overcovers. IGNsyn fails to cover mainly due to the bias. The rest of the synthesis models result in close to nominal coverages, as long as the analysis model is suited to the Case 2 data generation mechanism.

Table 3.4 shows the estimates for the error terms for each of the models at the analysis stage. These are also important. We note that OVB does not only effect the parameter estimates for the β parameters but also the error terms. The between variance estimates for REan and HYBan are biased for the original data for both large and small ICC

scenarios. FEsyn inflates the between error estimates for the random effects type models, but this is hardly noticeable, given the bias coming from the analysis models themselves. Error variance estimates from FEsyn, DIFF1syn, DIFF2syn, BGsyn and HYBsyn are similar to those observed for ORIG. The within errors for WIDEsyn models are slighlty inflated as before (around 4.10 rather than 4), but the between error estimates are close to those observed for ORIG. When it comes to the classification of error terms into between and within components, the IGNsyn and IGN2syn models attribute all the error term variability to their only error term, which represents within variability. The use of REsyn slightly inflates the error term variance estimates, but this is hardly a concern, given the biases in $\hat{\beta}$ observed earlier. Throughout the table, we find that the estimated ICCs are close to those estimated for ORIG, except for IGNsyn and IGN2syn.

Overall, we conclude that although non-hierarchical models can be used to produce unbiased analysis whether omitted variables exist or not, the variance estimates from such a synthesis model are inappropriate unless the analysis model is also of the same form. We, therefore, consider the standard fixed effects model, which may also cause biased between error variance estimates if the analysis model used employs random effects. In the presence of OVB, the standard random effects model is unsuitable for synthesis purposes. The HYBsyn and BGsyn approaches solve this problem, although the variance estimates for model coefficients is slightly questionable for BG. Performance from the DIFF1syn and DIFF2syn approaches is very promising; both the models closely replicate results from the original data in all the cases considered above. The WIDEsyn approach works well in both in the presence and absence of omitted variables, albeit with a moderate increase in the variance of the final estimate. This is expected, as Kinney et al. (2011) also remark. The data are hierarchical and the model is not, with the synthesis of each single wave's data, some simulation error is propagated through the waves and variables, as all subsequent waves are synthesised using previous wave's synthetic data. We also note that the best performing models (DIFF1syn, DIFF2syn, HYBsyn, BGsyn, WIDEsyn) perform consistently across the two ICC scenarios. For other models, the change of ICC either improves or deteriorates their performance, as observed in the results for Case 1 (without OVB). Data with small ICC as opposed to large ICC, is detrimental for FEsyn-REan, as noted in the literature. We also find this true for all IGNsyn results and in fact REsyn-FEan - a combination not studied in the literature. The opposite is observed for IGN2syn, where results seem more favourable when the ICC is small.

Synthesis Model		Analysis Model									
		ICC = 0.5					ICC = 0.06				
		FE	IGN	IGN2	RE	HYB	FE	IGN	IGN2	RE	HYB
ORIG	Bias	0.03	0.00	0.03	0.02	0.03	0.03	0.01	0.03	0.01	0.03
	VR	0.96	0.50	1.92	0.98	0.96	0.96	0.90	1.02	1.00	0.96
	Len	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Cov	94.48	83.04	99.28	94.72	94.48	94.48	93.84	95.16	95.04	94.48
FE	Bias	0.02	0.01	0.02	0.01	0.02	0.04	0.01	0.04	0.02	0.04
	VR	0.98	0.63	2.13	1.03	0.98	0.94	1.20	1.30	1.15	0.94
	Len	1.10	1.14	1.14	1.12	1.10	1.10	1.25	1.25	1.31	1.10
	Cov	94.92	87.88	99.56	95.08	94.92	93.52	96.92	97.48	96.44	93.52
DIFF1	Bias	0.03	0.02	0.03	0.03	0.03	0.02	0.01	0.02	0.02	0.02
	VR	0.96	0.59	1.78	0.99	0.96	0.96	0.92	1.02	1.00	0.96
	Len	1.10	1.20	1.05	1.10	1.10	1.10	1.11	1.09	1.10	1.10
	Cov	94.60	88.16	98.72	94.44	94.60	94.76	94.28	95.68	95.20	94.76
DIFF2	Bias	0.03	0.00	0.03	0.02	0.03	0.02	0.01	0.02	0.01	0.02
	VR	0.96	0.51	1.78	0.97	0.96	0.96	0.90	1.02	0.99	0.96
	Len	1.10	1.02	1.05	1.08	1.10	1.10	1.04	1.09	1.05	1.10
	Cov	94.60	83.28	98.80	94.04	94.60	94.76	93.76	95.72	95.44	94.76
IGN	Bias	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00
	VR	1.22	0.54	1.22	0.55	1.22	1.97	0.92	1.97	0.93	1.96
	Len	1.51	1.10	1.07	1.10	1.51	1.10	1.10	1.07	1.05	1.10
	Cov	97.20	85.80	97.20	85.80	97.20	99.44	94.24	99.40	94.36	99.40
IGN2	Bias	0.02	0.01	0.02	0.01	0.02	0.03	0.01	0.03	0.01	0.03
	VR	1.64	0.54	1.64	0.55	1.63	1.01	0.91	1.01	0.91	1.01
	Len	1.56	1.10	1.10	1.10	1.55	1.13	1.10	1.10	1.05	1.13
	Cov	98.92	84.84	98.88	85.16	98.88	94.92	93.56	95.00	93.76	94.92
RE	Bias	0.01	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00
	VR	1.16	0.52	2.15	0.89	1.16	1.91	0.93	2.03	1.02	1.91
	Len	1.09	1.05	1.05	1.08	1.09	1.07	1.09	1.07	1.09	1.07
	Cov	96.04	84.28	99.56	93.68	96.04	99.32	94.48	99.60	95.68	99.32
BG	Bias	0.02	0.00	0.02	0.01	0.02	0.02	0.00	0.02	0.00	0.02
	VR	0.95	0.52	1.75	0.97	0.95	0.91	0.90	0.97	0.98	0.91
	Len	1.10	1.05	1.05	1.08	1.10	1.10	1.10	1.09	1.09	1.10
	Cov	94.56	83.80	99.28	94.96	94.56	93.92	93.60	94.64	94.80	93.92
HYB	Bias	0.02	0.00	0.02	0.01	0.02	0.05	0.02	0.05	0.03	0.05
	VR	0.96	0.52	1.77	0.98	0.96	0.94	0.93	1.00	1.02	0.94
	Len	1.10	1.05	1.05	1.08	1.10	1.10	1.09	1.09	1.09	1.10
	Cov	94.48	84.64	99.08	95.28	94.48	94.64	94.28	95.12	95.16	94.64
WIDE	Bias	0.02	0.01	0.02	0.01	0.02	0.02	0.00	0.02	0.01	0.02
	VR	0.97	0.59	1.78	0.99	0.97	0.95	0.90	1.00	0.98	0.95
	Len	1.11	1.20	1.06	1.11	1.11	1.11	1.12	1.10	1.11	1.11
	Cov	94.56	86.68	99.28	94.68	94.56	94.12	93.80	94.88	94.76	94.12

Table 3.1: Posterior predictive approach: Properties of $\hat{\beta}$ over 2500 datasets, Case 1, Large and Small ICC. True value: $\beta = 3$; sample size is 5000, 1000 clusters, 5 observations per cluster.

Synthesis Model		Analysis Model									
		ICC = 0.5					ICC = 0.06				
		FE	IGN	IGN2	RE	HYB	FE	IGN	IGN2	RE	HYB
ORIG	σ_e^2	4.00	7.99	7.99	4.00	4.00	4.00	4.25	4.25	4.00	4.00
	σ_b^2	-	-	-	3.99	3.99	-	-	-	0.25	0.25
	ICC	-	-	-	0.50	0.50	-	-	-	0.06	0.06
FE	σ_e^2	4.00	9.59	9.59	4.00	4.00	4.00	5.85	5.85	4.00	4.00
	σ_b^2	-	-	-	5.60	5.60	-	-	-	1.85	1.85
	ICC	-	-	-	0.58	0.58	-	-	-	0.32	0.32
DIFF1	σ_e^2	4.00	8.01	8.00	4.00	4.00	4.00	4.26	4.25	4.00	4.00
	σ_b^2	-	-	-	4.01	4.00	-	-	-	0.25	0.25
	ICC	-	-	-	0.50	0.50	-	-	-	0.06	0.06
DIFF2	σ_e^2	4.00	7.99	7.99	4.00	4.00	4.00	4.25	4.25	4.00	4.00
	σ_b^2	-	-	-	4.00	3.99	-	-	-	0.25	0.25
	ICC	-	-	-	0.50	0.50	-	-	-	0.06	0.06
IGN	σ_e^2	7.99	7.99	7.99	7.96	7.96	4.25	4.25	4.25	4.23	4.23
	σ_b^2	-	-	-	0.03	0.03	-	-	-	0.02	0.02
	ICC	-	-	-	0.00	0.00	-	-	-	0.00	0.00
IGN2	σ_e^2	7.99	8.00	7.99	7.97	7.96	4.25	4.25	4.25	4.24	4.23
	σ_b^2	-	-	-	0.03	0.03	-	-	-	0.02	0.02
	ICC	-	-	-	0.00	0.00	-	-	-	0.00	0.00
RE	σ_e^2	4.00	7.99	7.99	4.00	4.00	3.98	4.26	4.26	3.98	3.98
	σ_b^2	-	-	-	4.00	4.00	-	-	-	0.28	0.28
	ICC	-	-	-	0.50	0.50	-	-	-	0.07	0.07
BG	σ_e^2	4.00	7.99	7.99	4.00	4.00	3.98	4.26	4.26	3.98	3.98
	σ_b^2	-	-	-	4.00	3.99	-	-	-	0.28	0.28
	ICC	-	-	-	0.50	0.50	-	-	-	0.07	0.07
HYB	σ_e^2	4.00	7.99	7.99	4.00	4.00	3.98	4.26	4.26	3.98	3.98
	σ_b^2	-	-	-	4.00	3.99	-	-	-	0.28	0.28
	ICC	-	-	-	0.50	0.50	-	-	-	0.07	0.07
WIDE	σ_e^2	4.09	8.16	8.15	4.09	4.09	4.09	4.34	4.34	4.09	4.09
	σ_b^2	-	-	-	4.08	4.07	-	-	-	0.25	0.25
	ICC	-	-	-	0.50	0.50	-	-	-	0.06	0.06

Table 3.2: Posterior predictive approach: Average of $\hat{\sigma}_e^2$ and $\hat{\sigma}_b^2$ over 2500 datasets and resulting ICC, Case 1, Large and Small ICC. True Values: For large ICC, $\sigma_e^2 = 4$, $\sigma_b^2 = 4$, for small ICC, $\sigma_e^2 = 4$, $\sigma_b^2 = 0.25$; sample size is 5000, 1000 clusters, 5 observations per cluster.

Synthesis Model		Analysis Model									
		ICC = 0.5					ICC = 0.06				
		FE	IGN	IGN2	RE	HYB	FE	IGN	IGN2	RE	HYB
ORIG	Bias	0.03	93.30	0.03	5.19	0.03	0.03	93.30	0.03	5.83	0.03
	VR	0.96	0.38	3.32	0.89	0.96	0.96	0.39	2.43	0.87	0.96
	Len	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Cov	94.48	0.00	99.92	0.24	94.48	94.48	0.00	99.72	0.04	94.48
FE	Bias	0.03	93.31	0.03	4.97	0.03	0.04	93.30	0.04	5.56	0.04
	VR	0.96	0.42	3.27	0.90	0.96	0.95	0.43	2.48	0.87	0.95
	Len	1.10	1.05	1.08	1.11	1.10	1.10	1.06	1.11	1.11	1.10
	Cov	94.48	0.00	99.84	1.32	94.48	94.56	0.00	99.68	0.44	94.56
DIFF1	Bias	0.03	93.31	0.03	5.20	0.03	0.03	93.30	0.03	5.84	0.03
	VR	0.96	0.44	2.94	0.89	0.96	0.94	0.43	2.16	0.87	0.94
	Len	1.10	1.14	1.03	1.11	1.10	1.10	1.11	1.04	1.11	1.10
	Cov	94.48	0.00	99.96	0.72	94.48	93.88	0.00	99.68	0.16	93.88
DIFF2	Bias	0.03	93.30	0.03	5.20	0.03	0.03	93.30	0.03	5.84	0.03
	VR	0.96	0.39	2.94	0.89	0.96	0.94	0.39	2.16	0.87	0.94
	Len	1.10	1.01	1.03	1.11	1.10	1.10	1.01	1.04	1.11	1.10
	Cov	94.48	0.00	99.92	0.76	94.48	93.88	0.00	99.68	0.16	93.88
IGN	Bias	93.32	93.31	93.32	93.31	93.32	93.28	93.28	93.28	93.28	93.28
	VR	0.98	0.43	0.98	0.43	0.98	0.96	0.43	0.96	0.43	0.96
	Len	2.64	1.10	1.42	1.74	2.63	2.42	1.10	1.52	1.61	2.42
	Cov	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IGN2	Bias	0.01	93.30	0.01	27.21	0.01	0.01	93.29	0.01	19.36	0.01
	VR	2.33	0.40	2.33	0.89	2.32	1.93	0.41	1.94	1.03	1.93
	Len	2.05	1.06	1.10	2.08	2.05	1.75	1.05	1.10	1.77	1.74
	Cov	99.68	0.00	99.68	0.00	99.68	99.24	0.00	99.24	0.00	99.24
RE	Bias	5.18	93.30	5.18	10.32	5.18	5.84	93.29	5.84	11.65	5.84
	VR	0.93	0.39	3.07	0.83	0.93	0.91	0.40	2.30	0.79	0.91
	Len	1.11	1.02	1.07	1.11	1.11	1.10	1.02	1.10	1.11	1.10
	Cov	0.76	0.00	19.20	0.00	0.76	0.24	0.00	3.96	0.00	0.24
BG	Bias	0.02	93.30	0.02	5.20	0.02	0.03	93.30	0.03	5.84	0.03
	VR	0.92	0.39	2.85	0.86	0.92	0.84	0.40	1.94	0.77	0.84
	Len	1.10	1.02	1.03	1.10	1.10	1.09	1.02	1.04	1.10	1.09
	Cov	94.32	0.00	99.88	1.16	94.32	93.00	0.00	99.20	0.40	93.00
HYB	Bias	0.01	93.30	0.01	5.18	0.01	0.02	93.29	0.02	5.83	0.02
	VR	0.94	0.39	2.89	0.88	0.94	0.95	0.40	2.19	0.87	0.95
	Len	1.10	1.02	1.03	1.11	1.10	1.10	1.02	1.04	1.11	1.10
	Cov	94.84	0.00	99.72	0.96	94.84	95.16	0.00	99.48	0.16	95.16
WIDE	Bias	0.02	93.27	0.02	5.28	0.02	0.02	93.28	0.02	5.94	0.02
	VR	0.96	0.43	2.95	0.89	0.96	0.98	0.44	2.24	0.90	0.98
	Len	1.11	1.14	1.04	1.12	1.11	1.11	1.11	1.05	1.12	1.11
	Cov	95.04	0.00	99.92	0.60	95.04	94.96	0.00	99.60	0.20	94.96

Table 3.3: Posterior predictive approach: Properties of $\hat{\beta}$ over 2500 datasets, Case 2, Large and Small ICC. True value: $\beta = 3$; sample size is 5000, 1000 clusters, 5 observations per cluster.

Synthesis Model		Analysis Model									
		ICC = 0.5					ICC = 0.06				
		FE	IGN	IGN2	RE	HYB	FE	IGN	IGN2	RE	HYB
ORIG	σ_e^2	4.00	24.31	13.85	4.02	4.00	4.00	20.56	10.11	4.03	4.00
	σ_b^2	-	-	-	34.30	9.86	-	-	-	30.34	6.12
	ICC	-	-	-	0.90	0.71	-	-	-	0.88	0.60
FE	σ_e^2	4.00	25.91	15.45	4.02	4.00	4.00	22.16	11.71	4.03	4.00
	σ_b^2	-	-	-	35.98	11.46	-	-	-	32.04	7.72
	ICC	-	-	-	0.90	0.74	-	-	-	0.89	0.66
DIFF1	σ_e^2	4.00	24.34	13.87	4.03	4.00	4.00	20.58	10.12	4.03	4.00
	σ_b^2	-	-	-	34.35	9.88	-	-	-	30.37	6.13
	ICC	-	-	-	0.89	0.42	-	-	-	0.88	0.61
DIFF2	σ_e^2	4.00	24.31	13.85	4.03	4.00	4.00	20.56	10.11	4.03	4.00
	σ_b^2	-	-	-	34.31	9.86	-	-	-	30.34	6.12
	ICC	-	-	-	0.89	0.71	-	-	-	0.88	0.60
IGN	σ_e^2	24.32	24.32	24.32	24.22	24.22	20.57	20.57	20.57	20.49	20.49
	σ_b^2	-	-	-	0.10	0.10	-	-	-	0.08	0.08
	ICC	-	-	-	0.00	0.00	-	-	-	0.00	0.00
IGN2	σ_e^2	13.86	24.32	13.86	14.52	13.80	10.11	20.57	10.11	10.45	10.07
	σ_b^2	-	-	-	17.72	0.06	-	-	-	20.01	0.04
	ICC	-	-	-	0.55	0.00	-	-	-	0.66	0.00
RE	σ_e^2	4.02	24.31	14.97	4.05	4.02	4.03	20.56	11.37	4.06	4.03
	σ_b^2	-	-	-	32.70	10.97	-	-	-	28.54	7.35
	ICC	-	-	-	0.89	0.73	-	-	-	0.88	0.65
BG	σ_e^2	4.00	24.31	13.85	4.03	4.00	4.00	20.57	10.11	4.03	4.00
	σ_b^2	-	-	-	34.31	9.86	-	-	-	30.35	6.12
	ICC	-	-	-	0.90	0.71	-	-	-	0.88	0.60
HYB	σ_e^2	4.00	24.31	13.85	4.02	4.00	4.00	20.57	10.11	4.03	4.00
	σ_b^2	-	-	-	34.31	9.86	-	-	-	30.35	6.12
	ICC	-	-	-	0.90	0.71	-	-	-	0.88	0.60
WIDE	σ_e^2	4.09	24.57	14.10	4.11	4.09	4.09	20.76	10.3	4.12	4.09
	σ_b^2	-	-	-	34.46	10.03	-	-	-	30.42	6.22
	ICC	-	-	-	0.89	0.71	-	-	-	0.88	0.60

Table 3.4: Posterior predictive approach: Average of $\hat{\sigma}_e^2$ and $\hat{\sigma}_b^2$ over 2500 datasets and resulting ICC, Case 2, Large and Small ICC. True Values: For large ICC, $\sigma_e^2 = 4$, $\sigma_b^2 = 4$, for small ICC, $\sigma_e^2 = 4$, $\sigma_b^2 = 0.25$; sample size is 5000, 1000 clusters, 5 observations per cluster.

MLE approach

We now change the procedure, and do not draw parameters from their predictive distributions for data synthesis. Instead, we use the maximum likelihood estimates of the parameters at the synthesis stage to generate all 10 copies of the synthetic data. We also do not consider the BGsyn model further, because of the computational burden associated with it. The rest of the simulation is the same as before.

We start from Case 1. Our results show that this change makes no difference to point estimation, as expected. Therefore, we look more closely at the variance estimates of $\hat{\beta}$ and the error terms. We repeat our assessment of the accuracy of the $\bar{u}$ estimates. Overall, we find the $\bar{u}$ estimates are less variable than in the previous results. Nevertheless, the patterns are similar to the ones observed when we sampled synthetic data using the posterior distributions of the parameters. This helps the WIDEsyn model variance estimates, making them closer to the original variances and also less variable; see Figure 3.7.

Table 3.5 shows that VRs are similar to those observed before, where most models replicate the ORIG VR except the IGNsyn and IGN2syn models. The VR for REsyn-REan is now close to the ORIG VR at 0.96, unlike before.

We now investigate the 95% CI. The average length of CI for synthetic data, is larger than for ORIG, but less so than before, replacing the 1.10 in Table 3.1 by 1.05, resulting in tighter inferences. However, the coverages of the true value of β are overall similar to those observed before, staying at nominal level for most models that take the hierarchical structure into account. The large REsyn VR and coverages of CI in the small ICC case are also similar to before.

Table 3.6 shows the final estimates for the within and between error components for the various analysis models. As in the results earlier, we observe the overestimation of the between error variance (4.80 instead of 4) when using FEsyn and the overestimation of the within error variance (4.04 instead of 4) from the WIDEsyn model. In both cases, however, the overestimation is lesser than before, again helped by the decrease in variation of the parameter estimates as fixed MLE estimates are used. This is also reflected in the ICC estimates, which are now closer to the ORIG ICCs, as compared to the simulation using the posterior predictive approach.

We now discuss results for Case 2, where omitted variable bias is expected. We first discuss the percentage biases. While the results exactly match for all synthesis-analysis combinations from Table 3.3, the biases observed for REsyn are noticeably different. Before, we noted that synthesising using REsyn doubled the biases from the synthesis to the analysis stage. In the case of MLE estimation, this is not true. In fact, the biases remain the same as in the synthesis stage, here 5.19% for the large ICC and 5.84% for the small ICC case. This doesn't change the fact using REsyn-HYBan or REsyn-FEan

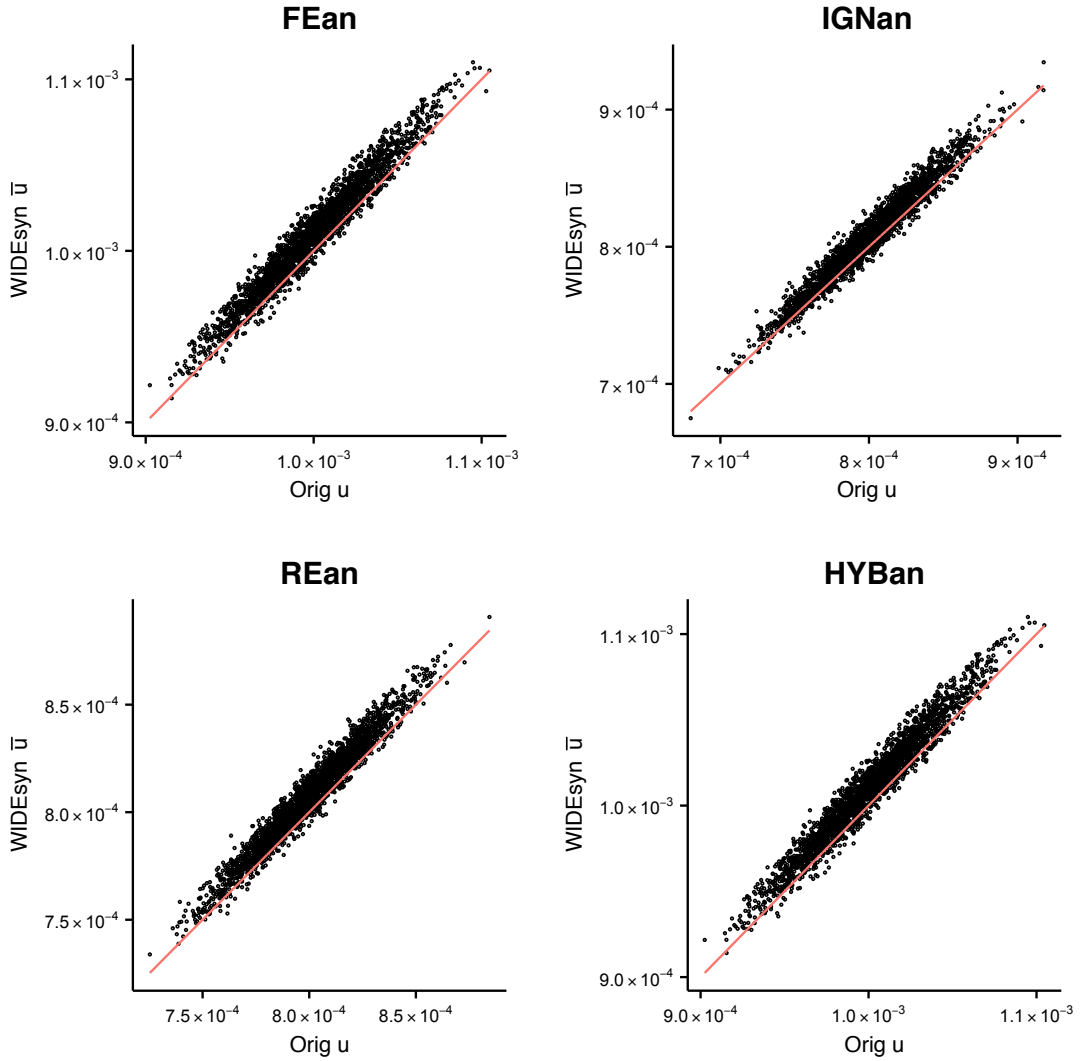

Figure 3.7: MLE approach: WIDEsyn $\bar{u}$ for various analysis models, Case 1, Large ICC. Plotted line (red) is $y = x$.

now results in biased estimates for $\hat{\beta}$, which is undesirable. However, it does make the implications of using a biased REsyn model for synthesis less severe. This raises the question why using draws from the posterior draws for parameters makes matters worse for the REsyn model biases. Upon investigation, we found that the extra bias is mainly driven by the fact that the posterior distribution of the random effects themselves is biased. This results in biased draws of the random intercepts that in turn bias the estimates for $\hat{\beta}$. Using the MLE approach avoids this step as the random intercepts are not drawn from their posterior distribution.

Table 3.8 shows that all other results for VR, length of CI and coverages are similar to those before, with the exception that all CI are now shorter. The reduced variability caused by using MLE estimates for the parameters, reduces variance overestimation for

the FEsyn and WIDEsyn models for the variance of $\hat{\beta}$. This also holds true for the within and between variance estimates as documented in Table 3.8.

Overall, the comments from Reiter and Kinney (2012) hold true in our results as the MLE results match those using the posterior distributions for the parameters. The MLE approach makes FEsyn and WIDEsyn look slightly more favourable. In fact, REsyn also results in reduced biases. This observation does not necessarily imply that using the MLE approach is better than using the full posterior simulation, as we are using a misspecified REsyn model. However, it does raise the concern that using a misspecified model in a posterior simulation can in fact make inferences worse for multiply-imputed data than using the MLE approach. In the missing data scenario, using the MLE approach is not yet a choice, but it is so for partially synthetic data.

In our evaluations for data utility, we observed that we more or less replicated the results observed in the missing data simulation. Each of the FEsyn, REsyn, WIDEsyn and IGNsyn comes with its fair share of disadvantages. Results from HYBsyn appear to provide a good balance between the REsyn and FEsyn models, and we find both the DIFF1syn and DIFF2syn approaches work well in a variety of scenarios, which is promising.

Synthesis Model		Analysis Model									
		ICC = 0.5					ICC = 0.06				
		FE	IGN	IGN2	RE	HYB	FE	IGN	IGN2	RE	HYB
ORIG	Bias	0.03	0.00	0.03	0.02	0.03	0.03	0.01	0.03	0.01	0.03
	VR	0.96	0.50	1.92	0.98	0.96	0.96	0.90	1.02	1.00	0.96
	Len	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Cov	94.48	83.04	99.28	94.72	94.48	94.48	93.84	95.16	95.04	94.48
FE	Bias	0.02	0.01	0.02	0.01	0.02	0.03	0.01	0.03	0.02	0.03
	VR	0.95	0.56	2.00	1.00	0.95	0.96	1.05	1.19	1.17	0.96
	Len	1.05	1.07	1.07	1.06	1.05	1.05	1.13	1.13	1.18	1.05
	Cov	94.16	86.08	99.40	94.64	94.16	94.20	95.52	96.68	96.44	94.20
DIFF1	Bias	0.03	0.00	0.03	0.02	0.03	0.02	0.00	0.02	0.01	0.02
	VR	0.96	0.54	1.84	0.97	0.96	0.96	0.90	1.01	0.99	0.96
	Len	1.05	1.10	1.03	1.05	1.05	1.05	1.06	1.05	1.05	1.05
	Cov	94.28	84.80	99.28	94.92	94.28	94.72	94.16	95.16	94.92	94.72
DIFF2	Bias	0.03	0.00	0.03	0.02	0.03	0.02	0.00	0.02	0.01	0.02
	VR	0.96	0.51	1.84	0.98	0.96	0.96	0.90	1.01	1.00	0.96
	Len	1.05	1.01	1.03	1.04	1.05	1.05	1.02	1.05	1.02	1.05
	Cov	94.28	82.48	99.28	94.60	94.28	94.72	93.72	95.16	94.88	94.72
IGN	Bias	0.02	0.00	0.02	0.00	0.02	0.01	0.01	0.01	0.01	0.01
	VR	1.23	0.52	1.23	0.53	1.23	2.05	0.91	2.05	0.92	2.04
	Len	1.48	1.05	1.05	1.05	1.48	1.08	1.05	1.05	1.01	1.08
	Cov	96.84	84.96	96.84	85.08	96.80	99.72	93.92	99.72	94.08	99.72
IGN2	Bias	0.03	0.01	0.03	0.01	0.03	0.03	0.00	0.03	0.00	0.03
	VR	1.75	0.52	1.75	0.53	1.74	1.03	0.93	1.03	0.94	1.02
	Len	1.48	1.05	1.05	1.05	1.48	1.08	1.05	1.05	1.01	1.08
	Cov	98.96	84.32	98.96	84.72	98.92	95.16	94.24	95.16	94.36	95.12
RE	Bias	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	VR	1.18	0.95	2.26	0.96	1.18	1.92	0.93	2.03	1.01	1.92
	Len	1.05	1.10	1.02	1.05	1.05	1.05	1.06	1.05	1.05	1.05
	Cov	96.84	94.68	99.68	94.56	96.84	99.40	94.24	99.48	95.40	99.40
HYB	Bias	0.04	0.01	0.04	0.03	0.04	0.02	0.00	0.02	0.01	0.02
	VR	0.95	0.53	1.83	0.97	0.95	0.96	0.92	1.01	1.01	0.96
	Len	1.05	1.10	1.02	1.05	1.05	1.05	1.06	1.05	1.05	1.05
	Cov	94.28	84.68	99.16	94.88	94.28	94.52	93.84	95.24	95.16	94.52
WIDE	Bias	0.02	0.01	0.02	0.01	0.02	0.02	0.01	0.02	0.01	0.02
	VR	0.99	0.54	1.88	0.99	0.99	0.97	0.93	1.02	1.02	0.97
	Len	1.06	1.10	1.03	1.06	1.06	1.05	1.06	1.05	1.05	1.05
	Cov	94.84	85.16	99.48	94.64	94.84	94.48	94.48	95.20	95.64	94.48

Table 3.5: MLE approach: Properties of $\hat{\beta}$ over 2500 datasets, Case 1, Large and Small ICC. True value: $\beta = 3$; sample size is 5000, 1000 clusters, 5 observations per cluster.

Synthesis Model		Analysis Model									
		ICC = 0.5					ICC = 0.06				
		FE	IGN	IGN2	RE	HYB	FE	IGN	IGN2	RE	HYB
ORIG	σ_e^2	4.00	7.99	7.99	4.00	4.00	4.00	4.25	4.25	4.00	4.00
	σ_b^2	-	-	-	3.99	3.99	-	-	-	0.25	0.25
	ICC	-	-	-	0.50	0.50	-	-	-	0.06	0.06
FE	σ_e^2	4.00	8.79	8.79	4.00	4.00	4.00	5.05	5.05	4.00	4.00
	σ_b^2	-	-	-	4.80	4.79	-	-	-	1.05	1.05
	ICC	-	-	-	0.54	0.54	-	-	-	0.21	0.21
DIFF1	σ_e^2	4.00	7.99	7.99	4.00	4.00	4.00	4.25	4.25	4.00	4.00
	σ_b^2	-	-	-	4.00	3.99	-	-	-	0.25	0.25
	ICC	-	-	-	0.50	0.50	-	-	-	0.06	0.06
DIFF2	σ_e^2	4.00	7.99	7.99	4.00	4.00	4.00	4.25	4.25	4.00	4.00
	σ_b^2	-	-	-	4.00	3.99	-	-	-	0.25	0.25
	ICC	-	-	-	0.50	0.50	-	-	-	0.06	0.06
IGN	σ_e^2	7.99	7.99	7.99	7.96	7.96	4.25	4.25	4.25	4.23	4.23
	σ_b^2	-	-	-	0.03	0.03	-	-	-	0.02	0.02
	ICC	-	-	-	0.00	0.00	-	-	-	0.00	0.00
IGN2	σ_e^2	7.99	7.99	7.99	7.96	7.96	4.25	4.25	4.25	4.23	4.23
	σ_b^2	-	-	-	0.03	0.03	-	-	-	0.02	0.02
	ICC	-	-	-	0.00	0.00	-	-	-	0.00	0.00
RE	σ_e^2	4.00	7.98	7.98	4.00	4.00	4.00	4.25	4.25	4.00	4.00
	σ_b^2	-	-	-	3.99	3.99	-	-	-	0.25	0.25
	ICC	-	-	-	0.50	0.50	-	-	-	0.06	0.06
HYB	σ_e^2	4.00	7.98	7.98	4.00	4.00	4.00	4.25	4.25	4.00	4.00
	σ_b^2	-	-	-	3.99	3.99	-	-	-	0.25	0.25
	ICC	-	-	-	0.50	0.50	-	-	-	0.06	0.06
WIDE	σ_e^2	4.04	8.06	8.06	4.04	4.04	4.04	4.29	4.29	4.04	4.04
	σ_b^2	-	-	-	4.03	4.03	-	-	-	0.25	0.25
	ICC	-	-	-	0.50	0.50	-	-	-	0.06	0.06

Table 3.6: MLE approach: Average of $\hat{\sigma}_e^2$ and $\hat{\sigma}_b^2$ over 2500 datasets and resulting ICC, Case 1, Large and Small ICC. True Values: For large ICC, $\sigma_e^2 = 4$, $\sigma_b^2 = 4$, for small ICC, $\sigma_e^2 = 4$, $\sigma_b^2 = 0.25$; sample size is 5000, 1000 clusters, 5 observations per cluster.

Synthesis Model		Analysis Model									
		ICC = 0.5					ICC = 0.06				
		FE	IGN	IGN2	RE	HYB	FE	IGN	IGN2	RE	HYB
ORIG	Bias	0.03	93.30	0.03	5.19	0.03	0.03	93.30	0.03	5.83	0.03
	VR	0.96	0.38	3.32	0.88	0.96	0.96	0.39	2.43	0.87	0.96
	Len	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Cov	94.48	0.00	99.92	0.24	94.48	94.48	0.00	99.72	0.04	94.48
FE	Bias	0.01	93.30	0.01	5.06	0.01	0.04	93.30	0.04	5.70	0.04
	VR	0.95	0.40	3.26	0.89	0.95	0.97	0.41	2.50	0.88	0.97
	Len	1.05	1.02	1.04	1.05	1.05	1.05	1.03	1.06	1.05	1.05
	Cov	94.32	0.00	99.92	0.88	94.32	94.76	0.00	99.68	0.12	94.76
DIFF1	Bias	0.02	93.30	0.02	5.19	0.02	0.03	93.28	0.03	5.84	0.03
	VR	0.96	0.41	3.13	0.89	0.96	0.96	0.41	2.31	0.87	0.96
	Len	1.05	1.07	1.01	1.05	1.05	1.05	1.05	1.02	1.06	1.05
	Cov	94.40	0.00	99.92	0.76	94.40	94.48	0.00	99.64	0.04	94.48
DIFF2	Bias	0.02	93.30	0.02	5.19	0.02	0.03	93.30	0.03	5.84	0.03
	VR	0.96	0.39	3.13	0.89	0.96	0.96	0.39	2.31	0.87	0.96
	Len	1.05	1.00	1.01	1.05	1.05	1.05	1.00	1.02	1.06	1.05
	Cov	94.40	0.00	99.92	0.68	94.40	94.48	0.00	99.64	0.04	94.48
IGN	Bias	93.30	93.30	93.30	93.30	93.30	93.29	93.28	93.29	93.28	93.29
	VR	0.96	0.40	0.96	0.41	0.96	0.96	0.41	0.96	0.41	0.95
	Len	2.59	1.05	1.39	1.67	2.58	2.38	1.05	1.50	1.54	2.38
	Cov	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IGN2	Bias	0.03	93.31	0.03	27.20	0.03	0.03	93.31	0.03	19.35	0.03
	VR	2.75	0.39	2.75	0.89	2.74	2.20	0.40	2.20	1.11	2.19
	Len	1.95	1.03	1.05	1.92	1.95	1.67	1.02	1.05	1.65	1.66
	Cov	99.80	0.00	99.80	0.00	99.80	99.56	0.00	99.56	0.00	99.56
RE	Bias	5.19	5.14	5.19	5.19	5.19	5.84	5.85	5.84	5.84	5.84
	VR	0.91	2.18	7.94	0.88	0.91	0.92	2.02	7.20	0.88	0.92
	Len	1.05	1.43	1.67	1.05	1.05	1.05	1.46	1.85	1.05	1.05
	Cov	0.64	36.68	85.84	0.56	0.64	0.08	17.08	57.68	0.08	0.08
HYB	Bias	0.03	93.30	0.03	5.20	0.03	0.03	93.28	0.03	5.84	0.03
	VR	0.96	0.41	3.12	0.89	0.96	0.95	0.41	2.28	0.86	0.95
	Len	1.05	1.07	1.01	1.05	1.05	1.05	1.05	1.02	1.05	1.05
	Cov	94.60	0.00	99.88	0.72	94.60	94.24	0.00	99.60	0.16	94.24
WIDE	Bias	0.03	93.29	0.03	5.23	0.03	0.02	93.31	0.02	5.88	0.02
	VR	0.98	0.41	3.17	0.91	0.98	0.97	0.41	2.32	0.88	0.97
	Len	1.05	1.07	1.02	1.06	1.05	1.05	1.06	1.02	1.06	1.05
	Cov	93.92	0.00	99.96	0.40	93.92	94.36	0.00	99.52	0.12	94.36

Table 3.7: MLE approach: Properties of $\hat{\beta}$ over 2500 datasets, Case 2, Large and Small ICC. True value: $\beta = 3$; sample size is 5000, 1000 clusters, 5 observations per cluster.

Synthesis Model		Analysis Model									
		ICC = 0.5					ICC = 0.06				
		FE	IGN	IGN2	RE	HYB	FE	IGN	IGN2	RE	HYB
ORIG	σ_e^2	4.00	24.31	13.85	4.02	4.00	4.00	20.56	10.11	4.03	4.00
	σ_b^2	-	-	-	34.30	9.86	-	-	-	30.34	6.12
	ICC	-	-	-	0.90	0.71	-	-	-	0.88	0.60
FE	σ_e^2	4.00	25.11	14.65	4.02	4.00	4.00	21.36	10.91	4.03	4.00
	σ_b^2	-	-	-	35.15	10.67	-	-	-	31.19	6.92
	ICC	-	-	-	0.90	0.73	-	-	-	0.89	0.63
DIFF1	σ_e^2	4.00	24.31	13.85	4.02	4.00	4.00	20.56	10.11	4.03	4.00
	σ_b^2	-	-	-	34.32	9.86	-	-	-	30.34	6.12
	ICC	-	-	-	0.90	0.71	-	-	-	0.88	0.60
DIFF2	σ_e^2	4.00	24.31	13.85	4.02	4.00	4.00	20.56	10.11	4.03	4.00
	σ_b^2	-	-	-	34.31	9.86	-	-	-	30.34	6.12
	ICC	-	-	-	0.90	0.71	-	-	-	0.88	0.60
IGN	σ_e^2	24.31	24.31	24.31	24.21	24.21	20.57	20.57	20.57	20.49	20.49
	σ_b^2	-	-	-	0.10	0.10	-	-	-	0.08	0.08
	ICC	-	-	-	0.00	0.00	-	-	-	0.00	0.00
IGN2	σ_e^2	13.85	24.31	13.85	14.52	13.79	10.1	20.56	10.11	10.44	10.06
	σ_b^2	-	-	-	17.71	0.06	-	-	-	20.01	0.04
	ICC	-	-	-	0.55	0.00	-	-	-	0.66	0.00
RE	σ_e^2	4.02	38.24	38.23	4.02	4.02	4.03	34.30	34.29	4.03	4.03
	σ_b^2	-	-	-	34.26	34.26	-	-	-	30.31	30.31
	ICC	-	-	-	0.89	0.89	-	-	-	0.88	0.88
HYB	σ_e^2	4.00	24.29	13.83	4.02	4.00	4.00	20.55	10.09	4.03	4.00
	σ_b^2	-	-	-	34.29	9.84	-	-	-	30.33	6.10
	ICC	-	-	-	0.90	0.71	-	-	-	0.88	0.60
WIDE	σ_e^2	4.04	24.43	13.97	4.06	4.04	4.04	20.66	10.19	4.07	4.04
	σ_b^2	-	-	-	34.38	9.94	-	-	-	30.39	6.16
	ICC	-	-	-	0.89	0.71	-	-	-	0.88	0.60

Table 3.8: MLE approach: Average of $\hat{\sigma}_e^2$ and $\hat{\sigma}_b^2$ over 2500 datasets and resulting ICC, Case 2, Large and Small ICC. True Values: For large ICC, $\sigma_e^2 = 4$, $\sigma_b^2 = 4$, for small ICC, $\sigma_e^2 = 4$, $\sigma_b^2 = 0.25$; sample size is 5000, 1000 clusters, 5 observations per cluster.

3.3.2 Disclosure risks

We now discuss the results from our disclosure risks assessment for the simulation study. We consider each property from our disclosure risk matrix one by one. The trace of the matrix represents the average number of matches an intruder would get right just by random matching. Figure 3.8 follows the value of the trace through various integer values of the calliper for Case 1. Each model follows its own path represented by the different coloured lines and markers. For each model, it is easy to see that there is an optimal value for the calliper, although these are not very different for the different models. In our case this lies around the Euclidean distance of 5 for most of the models. The plot shows that the best privacy is generally offered by the IGNsyn and IGN2syn models, the lines for which lie below all others. For these, the highest number of randomly guessed correct matches is around 15. The DIFF1syn and WIDEsyn models are close behind with the highest value of the trace around 22. All other models, i.e. FEsyn, REsyn, BGsyn and HYBsyn, show a higher risk with the highest trace value of around 50. The most risky, not surprisingly, is DIFF2syn with the highest trace of over 90. We also find that changing the value of $m = 2, \ldots, 10$, the number of synthetic datasets did not change these results significantly (results not shown).

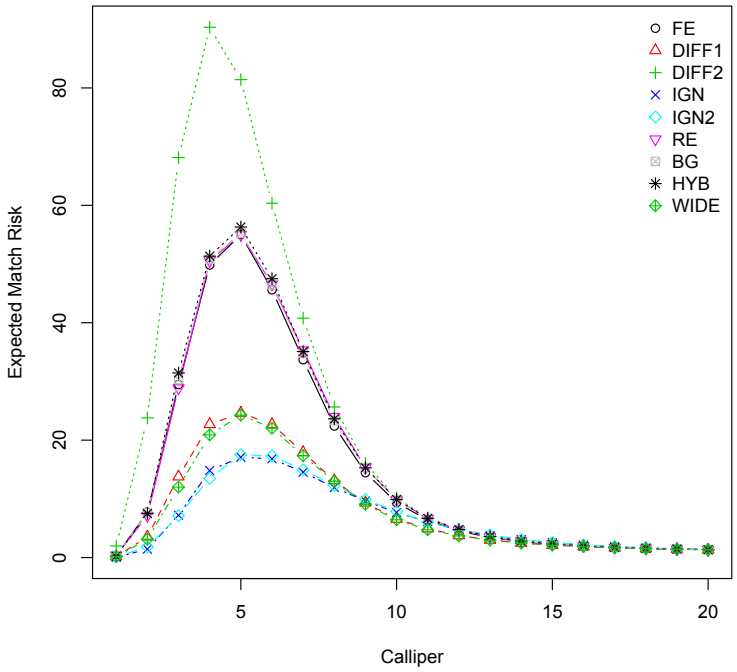

Figure 3.8: Posterior predictive approach: Expected match risk, Case 1, $m = 10$, ICC $= 0.5$.

We now discuss the *true* match rate. This is the number of observations for which the correct match is achieved with the highest probability, subject to the minimum 0.1

probability requirement, divided by the total number of units in the dataset, here, 1000. For the true risk measure, we find that changing m does make a difference. More number of copies help the intruder in shortlisting the right candidates for matches. We first look at the true match risk for $m = 2$ in Figure 3.9. As expected the highest number of true matches is observed for the DIFF2syn approach with a rate of about 0.25. The trends from the expected match risk follow here too, with the REsyn, BGsyn, HYBsyn and FEsyn approaches close together with a true match rate of about 0.13. The lowest values, of about 0.04, are for the IGNsyn, IGN2syn, WIDEsyn and DIFF1syn models. Figure 3.9 shows the true match rate when $m = 2$. When $m = 10$ instead, we notice that with the exception of IGN2syn, IGNsyn, DIFF1syn and WIDEsyn, all other models have slightly higher rates of true matches (see Figure 3.10).

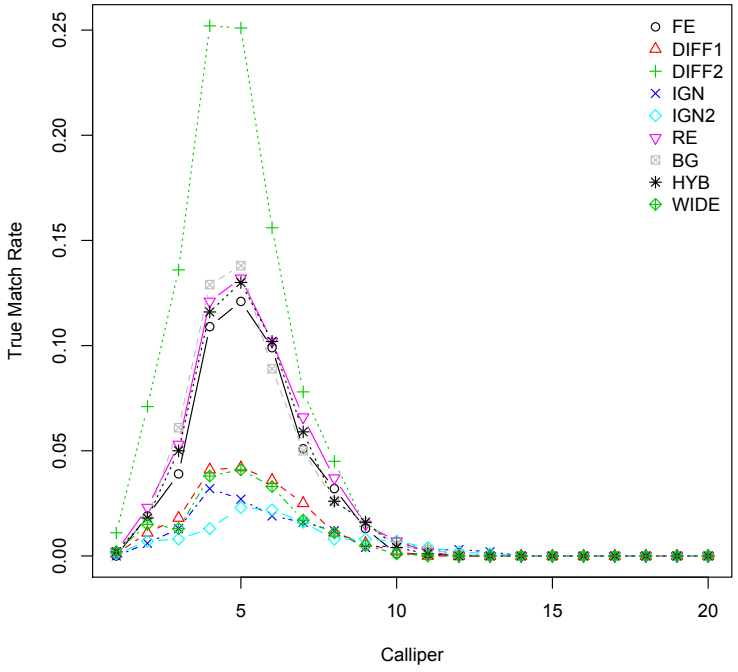

Figure 3.9: Posterior predictive approach: True match rate, Case 1, $m = 2$, ICC $= 0.5$.

We now discuss the *false* match rate. This is the number of times the intruder uniquely identifies a match for the wrong unit using the highest probability of matching as a criterion, divided by the number of unique matches identified. High false rates mean that the intruder believes he has made a match when it is not correct. However, limiting the false match rate may also be of interest, where a false match may cause harm and perhaps also indicate lack of data utility.

Figure 3.11 shows the false match rate for Case 1 with $m = 2$. We find that most of the unique matches declared by the intruder based on 2 copies of the data are false. However, as m increases, the probability of declaring a false match decreases around

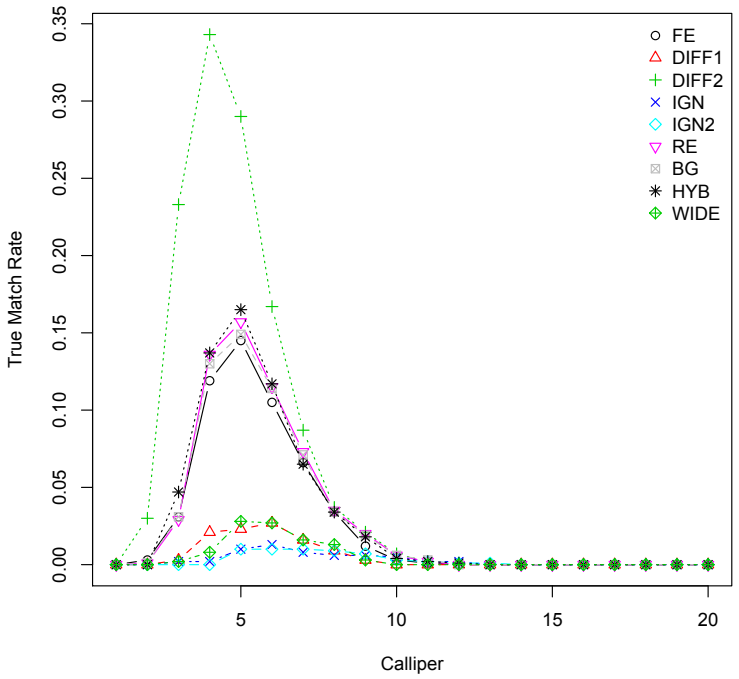

Figure 3.10: Posterior predictive approach: True match rate, Case 1, $m = 10$, ICC $= 0.5$.

the optimal range of calliper of 5. Figure 3.12 shows the results for 10 copies; here, the proportion of false matches amongst the unique matches falls to less than 0.4 for most models around the Euclidean distance of 5. This is a dramatic improvement for the intruder. Combined with the increase in true match rate from $m = 2$ to $m = 10$, we observe that increasing m increases the probability for an intruder to make the correct matches.

Moving from Case 1 to Case 2 does not change the trend in the disclosure risk results. In Case 2, the trends in expected and true risks do not change, but the magnitude increases. Compared to Figure 3.10, in Figure 3.13 we observe almost double the rate of true matches. Nevertheless, the risks in both cases are not directly comparable, as the Case 2 datasets differ substantially from Case 1 datasets with the addition of an extra time constant variable (the omitted variable), that adds to the mean of each unit, making matching using Euclidean distances easier. We focus on the fact that the trends in the risk measures remain the same. The results for the false match rate are quite similar to those observed in Case 1.

As in the data utility results, we measure risks of identification when the ICC of the generated data is small ($= 0.06$). Having a small ICC means that the units in the data are now less different from each other based on the intercept alone. Figure 3.14 shows that differences between the risk profiles of the IGNsyn, IGN2syn, DIFF1syn, and

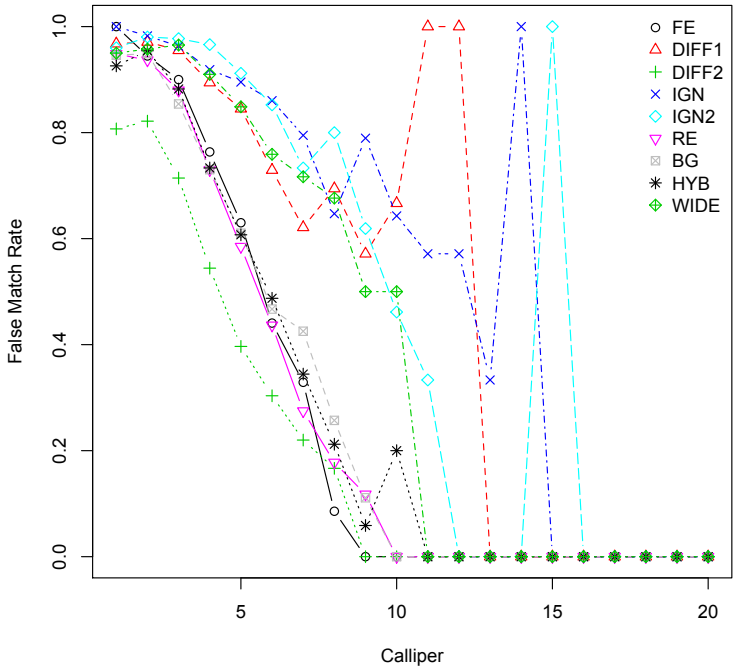

Figure 3.11: Posterior predictive approach: False match rate, Case 1, $m = 2$, ICC $= 0.5$.

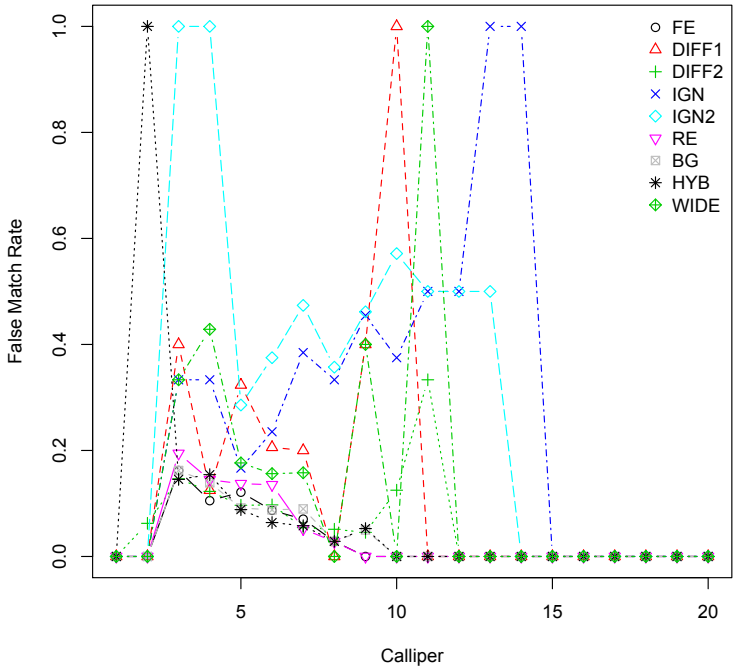

Figure 3.12: Posterior predictive approach: False match rate, Case 1, $m = 10$, ICC $= 0.5$.

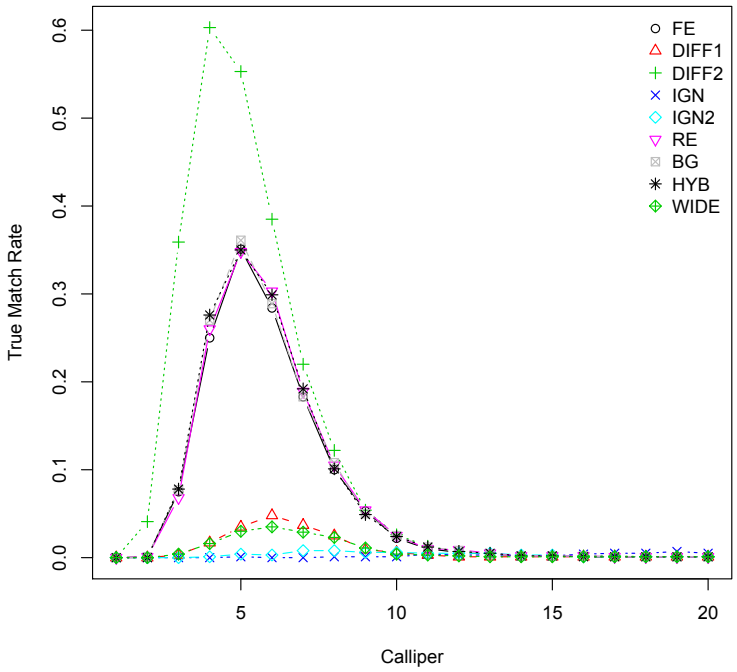

Figure 3.13: Posterior predictive approach: True match rate, Case 2, $m = 10$, ICC $= 0.5$.

WIDEsyn models as compared to the FEsyn, REsyn, HYBsyn and BGsyn models are now eliminated. All these models result in similar risks, except DIFF2, which stands out for the highest values as before. Results for the true and false match rates also show similar patterns where the difference in risks for the hierarchical and non-hierarchical synthesis models decreases.

In the small ICC scenario, moving from Case 1 to Case 2 results in the same consequences as in the large ICC scenario. Models that separate out individual intercepts, such as FEsyn, REsyn, HYBsyn and BGsyn, are more risky to use when the individual units differ more from each other. All other models, remain unaffected.

We also consider the consequences of using the maximum likelihood method on disclosure risks. We find that in this scenario disclosure risks generally decrease for the synthesis models with random effects, such as REsyn and HYBsyn. On the other hand, results for FEsyn show more risks. Figure 3.15 shows that the expected match risk for REsyn and HYBsyn now matches the risk profile of WIDEsyn and DIFF1syn, with the highest number around 22 as compared to 50 before. In contrast, the expected match risk for FEsyn increases from around 55 in the posterior simulations, to about 65 for the MLE approach. Results for DIFF2syn and IGNsyn and IGN2syn remain similar. Changing the number of copies m does not impact the expected match risk, as before.

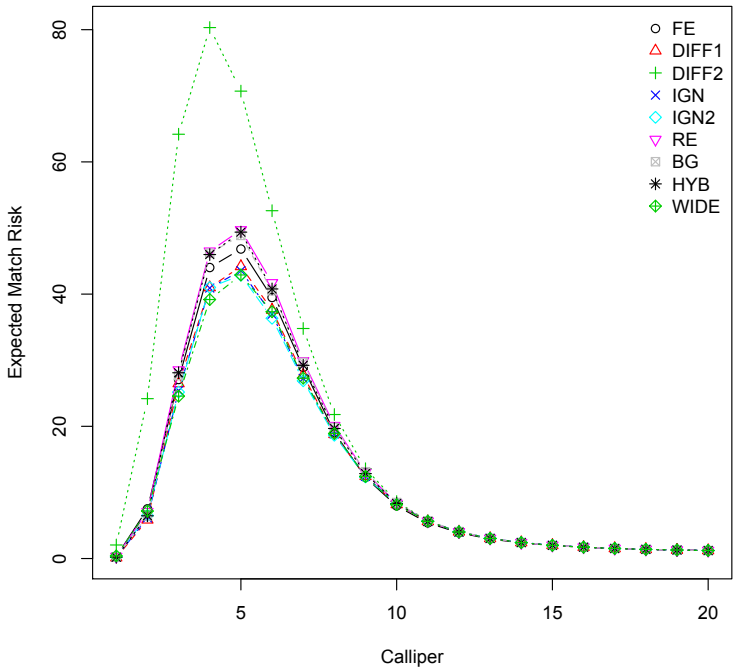

Figure 3.14: Expected match risk, Case 1, $m = 10$, ICC $= 0.06$.

In Figure 3.17, we make similar observations for the true match rate. The models employing random effects now seem less risky with true match rates under 0.03. Instead, the true match rate for FEsyn increases from 0.15 in the posterior simulation, to 0.20 for the MLE approach. In the posterior simulation approach, we observed that the true match risk increased with m. Here, we find that this is true for FEsyn and DIFF2syn only. For REsyn, HYBsyn, DIFF1syn and WIDEsyn, the observed rate actually slightly decreases when m goes from 2 to 10. In Figure 3.16, we see that the random effects type models have a highest true match rate of nearly 0.05, which is almost double of that observed for $m = 10$.

We now discuss the false match rate. The results when $m = 10$ remains similar to those observed for the posterior simulations. However, when $m = 2$, we note that REsyn now results with higher false match rates over a range of callipers (see Figure 3.18), as would be in line with the observed lower expected and true match risks.

In the simulation where the ICC is small (= 0.06), we again note that the differences between the hierarchical and non-hierarchical models diminish. The random effects type models show reduced risks than before, but not dramatically so (results not shown). When data are generated with an omitted variable, we also observe results as in the posterior simulation and closely match the results in the large ICC case (results not shown).

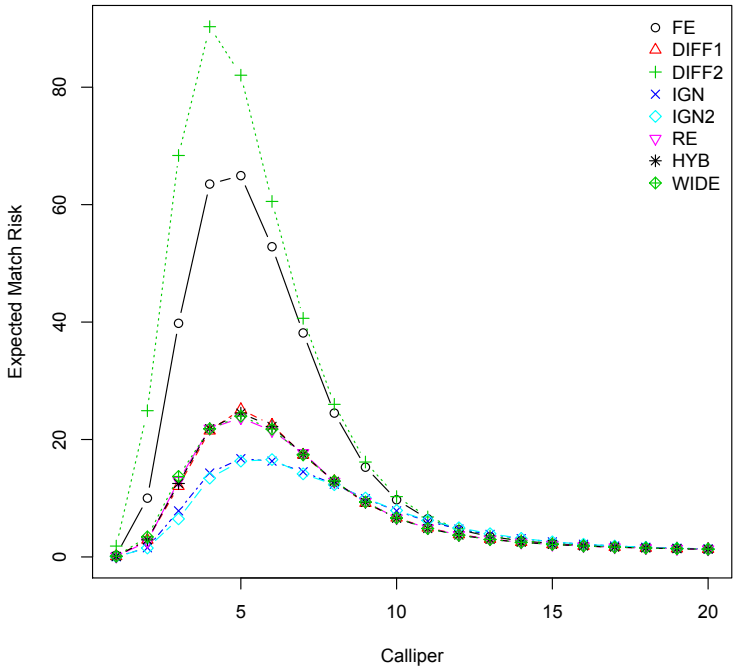

Figure 3.15: MLE approach: Expected match risk, Case 1, $m = 10$, ICC $= 0.5$.

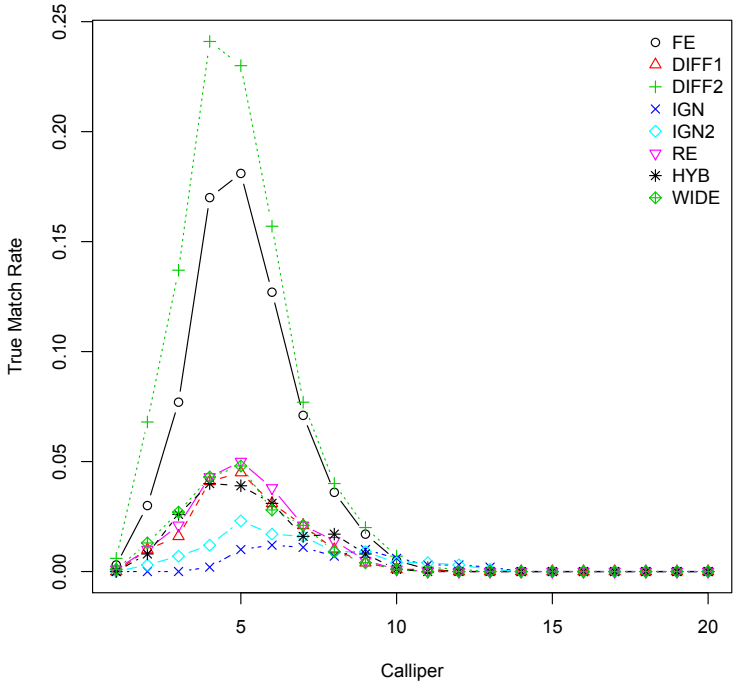

Figure 3.16: MLE approach: True match rate, Case 1, $m = 2$, ICC$= 0.5$.

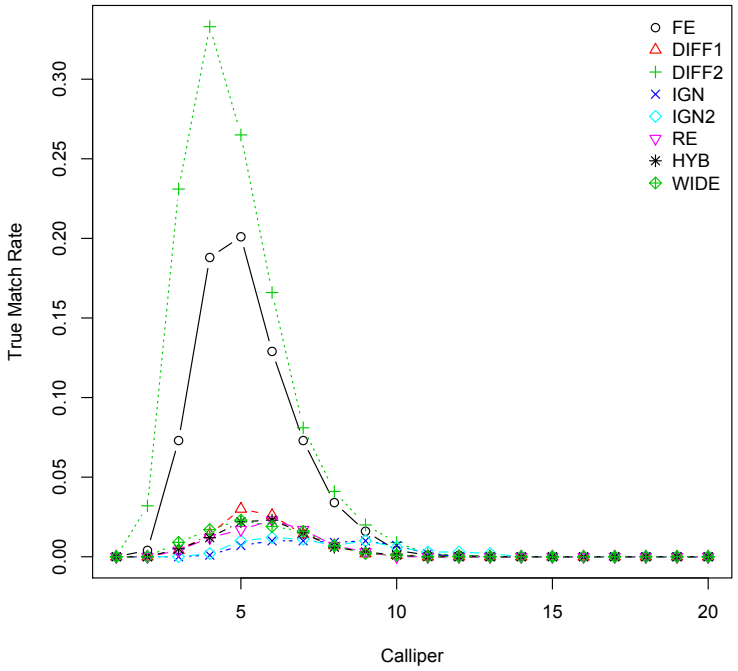

Figure 3.17: MLE approach: True match rate, Case 1, $m = 10$, ICC $= 0.5$.

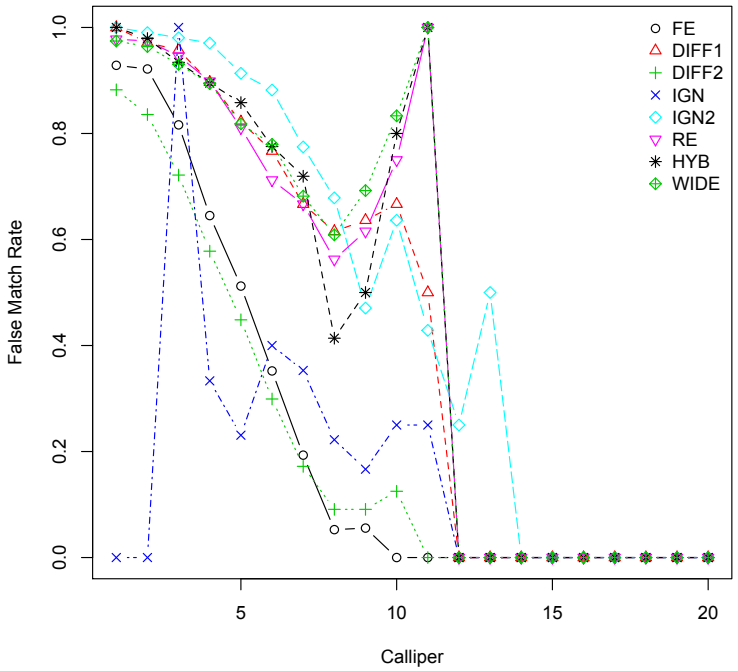

Figure 3.18: MLE approach: False match rate, Case 1, $m = 2$, ICC$= 0.5$.

Overall, we find that the best protection, in our limited investigation, is provided by the IGNsyn and IGN2syn models. These are followed by WIDEsyn and DIFF1syn. The rest of the models, then show higher risks, followed by DIFF2 which is the riskiest overall. There is some value in keeping the number of released datasets, m, small by increasing the chances of false matches and decreasing the true match rate. The results are different when comparing the Bayesian setup to the MLE approach, when risks for FEsyn increase when using the plug-in estimator, while for the RE type models, they decrease. Models with low disclosure risks, such as IGNsyn or WIDEsyn, do not seem to be affected by the number of datasets, m, released. This may be a desirable property if data analyses also improve with more copies of data released. However, we haven't studied this in our current research. The change of ICC brings the disclosure risk profile of all models close to each other when no OVB exists, except for DIFF2syn. However results across the two ICC scenarios were similar when we ran simulations with OVB.

It is clear that in terms of absolute numbers, it seems that the disclosure risks are very high for our simulation in general. However, we also note that this is a very conservative estimate of the possible number of identifications for several reasons. Firstly, we are assuming that the intruder has perfect knowledge of the target units as we match our synthetic datasets to original data. Furthermore, we also assume that the intruder knows that the target is in the sample. This is generally unknown, and there will be an associated probability with the target not being in the sample at all, which we have ignored. Moreover, we worked with several calliper levels; in practice the intruder has to make some arbitrary decision on what level of calliper to use to obtain the least number of incorrect matches and the most correct ones. At all occasions that an intruder makes a unique match, he is unaware whether the match is correct or wrong. As shown by our results on the false match risk, many of the unique matches made may be incorrect, the chances of which increase by keeping m small.

Nevertheless, we are aware of the shortcomings of the disclosure risk procedure we have applied. Firstly, we have not explicitly taken into account the fact that this is a hierarchical dataset and $Y_1, \ldots, Y_5$ are, in fact, the same variable. Research in risk measures specifically for hierarchical data is sparse, and generally poorly understood (Elliot, 2005). Secondly, we have made the assumption that the intruder knows nothing about the synthesis process, and therefore, as a best guess uses the synthetic datasets to find the probabilities of identification. In reality, some information about the synthesis process must be released to aid the analysts, unfortunately, this information will also aid an intruder at the same time.

The simulation studies provide us a basic understanding of the interactions between various synthesis and analysis models for hierarchical data. As expected we find that the synthesis models offer either high utility with high risks, or low utility with low risks. A surprise package is DIFF1syn, which displays data utility results at par with DIFF2syn results, yet disclosure risks equivalent to those of IGNsyn. It is not possible

to identify whether this is an artefact of the study design, and further investigation is needed to understand the DIFF1syn model better. Our results showed that just as FEsyn may not be congenial to REsyn, REsyn may also not be congenial to FEsyn. Data utility can be improved with the use of the HYBsyn, BGsyn or DIFFsyn models, but if risks are a major concern, WIDEsyn results in much lower risks than any of these models in exchange for added variability.

3.4 Real data application

Given the empirical evaluation in Section 3.3, we now turn to our real world application of interest. The dataset in question is the IAB Establishment Panel administered by the Institute for Employment Research of the German Federal Employment Agency. This study uses the IAB Establishment Panel, Waves 2005 and 2007. Data access was provided via on-site use at the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB).

The IAB Establishment Panel data contain information from a survey of a random sample of plants stratified by employment size, industry and federal states, that employ at least one worker covered by social security at the 30th of June each year. The survey has been undertaken annually since 1993 in West and 1996 in East Germany. Currently, approximately 16000 plants are interviewed in each wave. The sample is updated every year and response rates between $63\% - 73\%$ for written surveys and over 80% for interviews have been reported (Fischer et al., 2008).

In light of the interests of the Federal Employment Agency, the IAB Establishment Panel has a focus on employment related questions. However, areas such as volume of sales, investments, R&D, innovations, business policy and organisational change are also covered. Each year, a special topic is also introduced, for instance the 2003 and 2010 waves provide information on temporary agency work. With the wealth of information and over a decade of annual observations, it is not surprising that the IAB Establishment Panel is a popular dataset with analysts. Over 4000 research articles based on analysis carried on the dataset or its linked versions are available on the Research Data Centre (FDZ) website. Currently, there are two ways to access the data, either through a research visit to the FDZ or through a request for remote access where the FDZ runs the user's code on the original data and returns the output to the user after checking it for disclosure risks. An open access to samples of various versions of the dataset will be of immense value to the community of researchers. Nevertheless, as ample information on establishment characteristics can be accessed online, data confidentiality is a challenge. Moreover, given the sample stratification on size, some of the largest organisations would be represented in the data and these are easier to identify as the number of large organisations is relatively small compared to the small or medium sized establishments.

In our evaluation, we follow the simulation setup and focus on data utility, while recording the disclosure risks for each method used. To demonstrate our methodology, we choose analyses carried out in the literature using the IAB Establishment Panel. The aim is to replicate the selected analyses using synthetic data created by our models.

Selected paper

The article by Ellguth et al. (2014) explores the effect of opening clauses, combined with the existence of German works councils, on wages. In Germany, collective bargaining agreements are industry level arrangements, between the trade unions and employer organisations, around pay and working conditions for German workers. The German works council, an establishment level worker representative body, usually does not have much say in the collective bargaining agreements. However, opening clauses provide some flexibility to employers to be able to act differently from the collective bargaining agreements. The implementation of opening clauses, however, is an establishment level decision, which implies that works councils have more say when such decisions come by. This article studies this interaction between works councils and opening clauses with respect to the effect on wages. The data used are extracted from two years of the IAB Establishment Panel survey, as these are the only two available waves with information on opening clauses.

The purpose of the analysis is to compare the wage bills of establishments who apply opening clauses to those which do not, taking into account whether such an opening clause exists and a works council exists for each establishment. The dataset contains information on both, the existence and application of, opening clauses.

This article is ideal for our study, as the authors employed a number of sensitivity checks to verify their conclusions. Firstly, they fit various forms of models; these correspond directly to our IGN, FE and RE models. The paper also includes the same analysis on various subsets of the data, which we also run to assess the quality of our synthesis procedures. Most interestingly, the FE and RE models disagree on the significance of certain key variables of interest, and these could potentially be a result of omitted variable bias. The focus of our research is not to identify the best model for the data. As we take the role of a data keeper, we would like to ensure that the authors obtain the same conclusions using the synthesised version of the dataset. If there are substantial differences between the conclusions of the various models, we aim to preserve these too.

Data synthesis

We first focus on the selection and modification of the variables in the data to prepare for synthesis. We gather all the variables used in the various regression analyses, in their untransformed state, for the 2005 and 2007 waves. We do not include records for establishments that discontinued operation in the survey year or earlier. As in the analysis procedure, we treat the variable identifying ‘application of opening clauses’ as

non-missing when no opening clauses exist, and mark the 'NA' observations as 'opening clauses not applied'. One of the covariates included in the analyses is the establishment's churning rate. This is a function of the number of new hires and terminations for the establishment in the first half of the year. As in the case of application of opening clauses, if the establishment has been recorded not interested in hiring new employees or has not terminated any employees, we treat 'new hires' or 'terminations' = 0, rather than missing. The analysis focuses only on establishments that have industry-level bargaining agreements. We do not make this selection for synthesis, however, we include the variable that classifies bargaining agreements for the firms, during synthesis. We treat the variable 'state of technical equipment' as a factor rather than a continuous variable as in the analysis. We do not change the levels of a factor given in the real dataset. For example, 'ownership' of plants has several categories, but the analyst is only interested in whether the ownership is foreign or not. We do not re-define the 'ownership' variable as in the analysis and use all the categories. We use the 16-level industry classification for the establishments as in the analysis. For the 'federal state' variable, in the 2005 wave, it is possible to separate East Berlin from West Berlin, but not in the 2007 wave. We choose to combine East and West Berlin, into Berlin which is treated as West Germany in the 2007 wave. Therefore, we treat East Berlin as West Germany in the 2005 wave as well. We also remove establishments with records of more number of casual workers than the number of total employees, as this may be an error. This selection process left us with 12522 observations.

One of the main challenges for data synthesis in this scenario is that only two waves of data are used, as the questionnaire includes questions on opening clauses only for two waves. The data used in the article is an unbalanced panel, which implies that some firms may only have one observation. This is problematic for the use of FEsyn, where we cannot estimate fixed effects for establishments with only one observation over the two years. Therefore, we choose to work with a balanced data set only, to be able to compare all of our synthesis models on equal footing. Our final selected balanced data have 7230 observations. Note that we use this selection for the synthesis procedure; a further subset of this data is used for final analyses as the article does not consider small establishments or establishments without certain collective bargaining agreements.

Given our selection of the data, we note that we cannot replicate the analyses in the article perfectly. Therefore, we choose to run the analyses from the article but our benchmark results come from our selection of the data. We do find that our results closely match those in the article, with the exception of a few covariates. We will make note of any important differences as we describe each of the various analyses in Section 3.4.1.

We treat the wages variable as sensitive and create $m = 10$ synthetic copies for observed wages using each of the synthesis models. We take a number of steps to improve the synthesis process. We log-transform heavily skewed covariates, such as the number of

fixed-term, part-time or casual employees. For the few establishments that change their industry classification or federal state between 2005 and 2007, we assign them the federal state and industry code recorded in 2007. The existence of these few establishments make the federal state and industry variables time-varying, but the FEsyn model does not handle time-varying variables which are nearly time-constant very well, as the effect gets confounded with the fixed effects. Therefore, we choose to treat these variables as time-constant for all our synthesis models. We also divide our data into four sets of observations based on the average magnitude of wages over the two years, and synthesise each set separately. This is to constrain the range of synthesised data within the various quartiles of the observed wages. For each synthesis model and subset of data, we also search for the optimal Box-Cox transformation (Box and Cox, 1964; Gurka et al., 2006) for the response variable, wages, and back-transform the synthetic values after running the synthesis models. We only choose to use the MLE approach to synthesise the real data. Due to the normality assumption of the various synthesis models, if a generated synthetic value turns out to be negative, we replace this with 0. Additionally, the two interaction terms of interest in the analysis models were not included in our synthesis models. In practice, it would be challenging for the data keeper to predict what models analysts may be interested in. For the data keeper, there could possibly be several two-way, or higher, interaction terms that could be included in the synthesis models, and not all can possibly be included; using synthesis models with just the main effects is often what is achievable and observed in real data applications.

Having synthesised wages, we now turn to the analysis of our synthetic data. We first assess the utility provided by our synthesis models for all the analyses presented in the original paper.

3.4.1 Data utility

Summary statistics

We start by preparing the data as in the original article. The analyses carried out in the article do not include establishments that have either fewer than four employees, or do not have industry level bargain agreements or do not know whether, for their plant, opening clauses exist or not. Many of the covariates are also transformed, for instance, the number of employees in various categories are changed to proportions of the total number of employees, and certain categorical variables are recoded with only two levels (0/1) such as ownership, investment last year, and the establishment's position within the organisation. Table 3.9 presents the summary of our selection of data; we call this ORIG or original data from here onwards. This closely matches the summary provided in the chosen article (see Appendix B, Table B.1 for comparison). The mean and standard deviation for the response variable, log wages per full-time equivalent employee are also included. As these change for each of synthesis models, we include the figures from

synthetic data averaged over multiple copies. We observe that the mean and standard deviation of the response variable from all our synthesis models are close to that from the original data. We observe that the variability in wages synthesised using the random effects type models, i.e. REsyn, HYBsyn and BGsyn is slightly less ($sd \sim 0.42$) than that in the original data ($sd \sim 0.45$).

We now turn to the 4-way classification table presented in the article (Table 2, page 101). This is a summary of numbers and proportions of establishments with or without works councils, with or without opening clauses and with or without application of opening clauses, given the establishments have total wages below or above the average total wages. As we have synthesised total wages, we check whether we preserve the classification table as in the original data. Figure 3.19 shows the numbers and proportions for each classification cell under the original and synthesised datasets for establishments with wages below the average. For establishments with above average wages, Figure 3.20 shows the corresponding results. Amongst the establishments with wages below the mean, we find that across all synthesis models, the proportion of establishments with works councils are overestimated (at around 38% rather than 36%) while those without works councils are slightly underestimated (at around 60% rather than 63%). This is most noticeable when DIFF1syn, IGNsyn, IGN2syn and WIDEsyn are used. However, overall, we do not find the sample proportions too different from those observed in the original data, ORIG. For establishments with wages above the mean, the proportions of establishments with or without opening clauses and application of opening clauses are generally well-preserved. There is very slight underestimation for the number of establishments without works councils, with the exception of results for FEsyn. For both, establishments below and above the mean, we note that DIFF2syn most closely matches the numbers and proportions observed for the original data, while DIFF1syn, IGNsyn, IGN2syn and WIDEsyn are further from the true numbers than all other models. Nevertheless, the numbers are very close to each other for all the synthesis models.

Variable	Mean (SD)
Log(wages/full-time equivalent)	**7.752(0.432)**
FEsyn - Log(wages/full-time equivalent)	7.750(0.454)
DIFF1syn - Log(wages/full-time equivalent)	7.749(0.448)
DIFF2syn - Log(wages/full-time equivalent)	7.752(0.432)
IGNsyn - Log(wages/full-time equivalent)	7.750(0.448)
IGN2syn - Log(wages/full-time equivalent)	7.750(0.448)
REsyn - Log(wages/full-time equivalent)	7.753(0.423)
BGsyn - Log(wages/full-time equivalent)	7.753(0.422)
HYBsyn - Log(wages/full-time equivalent)	7.754(0.424)
WIDEsyn - Log(wages/full-time equivalent)	7.750(0.448)
Works council (yes = 1)	0.615(-)
Existence of an opening clause (yes = 1)	0.320(-)
Application of an opening clause (yes = 1)	0.153(-)
Existence of an opening clause and works council (yes = 1)	0.240(-)
Application of an opening clause and works council (yes = 1)	0.118(-)
Proportion of qualified employees	0.742(0.234)
Proportion of employees with fixed-term contracts	0.063(0.124)
Proportion of casual workers	0.019(0.053)
Proportion of part-time employees	0.211(0.234)
Proportion of trainees	0.049(0.074)
Churning rate	0.045(0.130)
Establishment not part of larger enterprise (Single = 1)	0.603(-)
Technical state of the establishment (1 = very good, ..., 5 = bad)	2.151(0.723)
Invested in physical capital within the previous year (Invest = 1)	0.792(-)
Establishment is under foreign ownership (Foreign = 1)	0.074(-)
5 - 9 employees	0.116(-)
10 - 19 employees	0.114(-)
20 - 49 employees	0.167(-)
50 - 99 employees	0.150(-)
100 - 199 employees	0.141(-)
200 - 499 employees	0.166(-)
500 - 999 employees	0.083(-)
1000 - 4999 employees	0.059(-)
5000 or more employees	0.004(-)
Number of observations	5195

Table 3.9: Sample description - mean and standard deviation of key variables in the original and synthetic datasets.

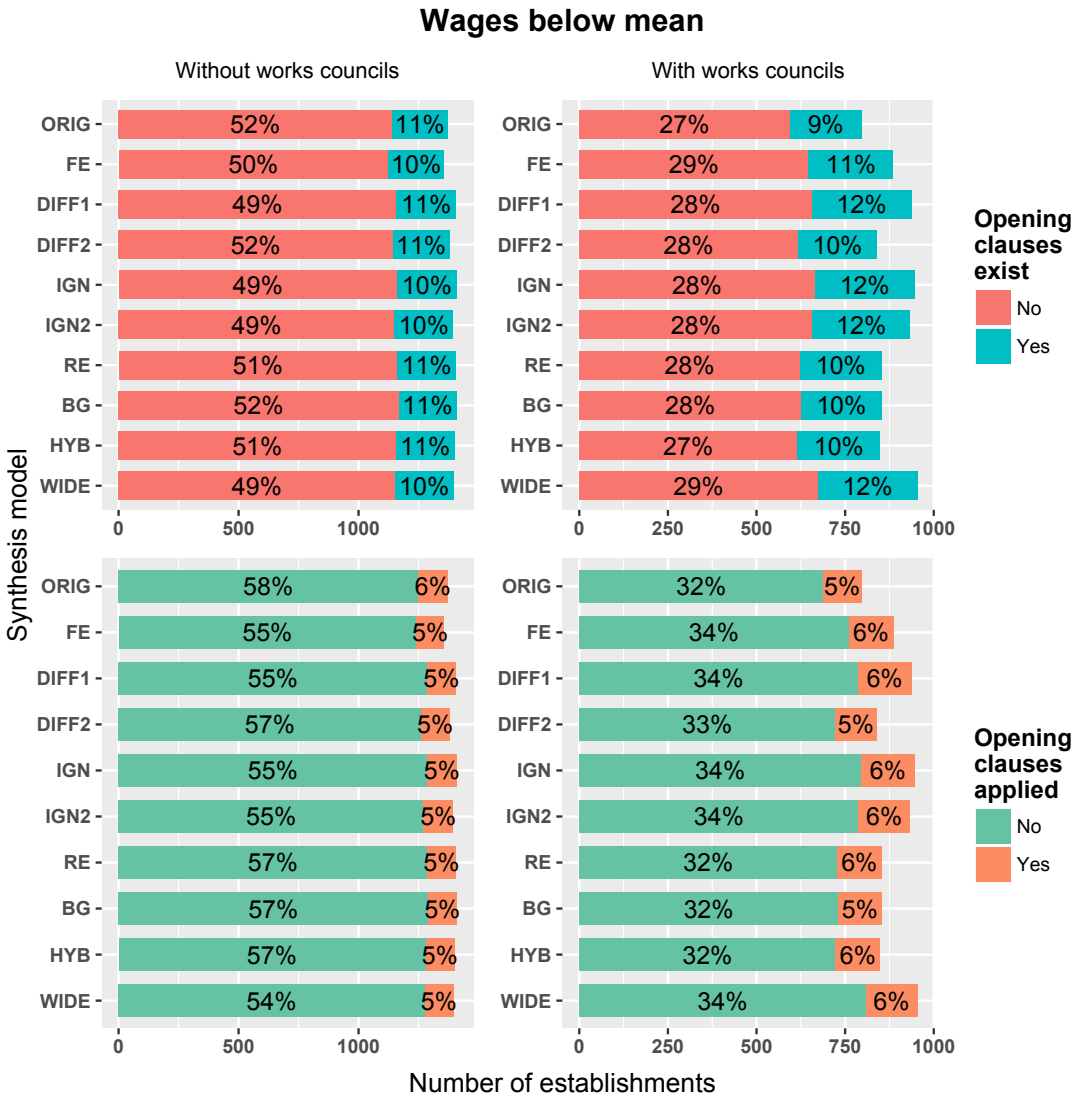

Figure 3.19: Number and proportions of establishments with or without work councils, opening clauses and application of opening clauses for establishments with wages below the mean

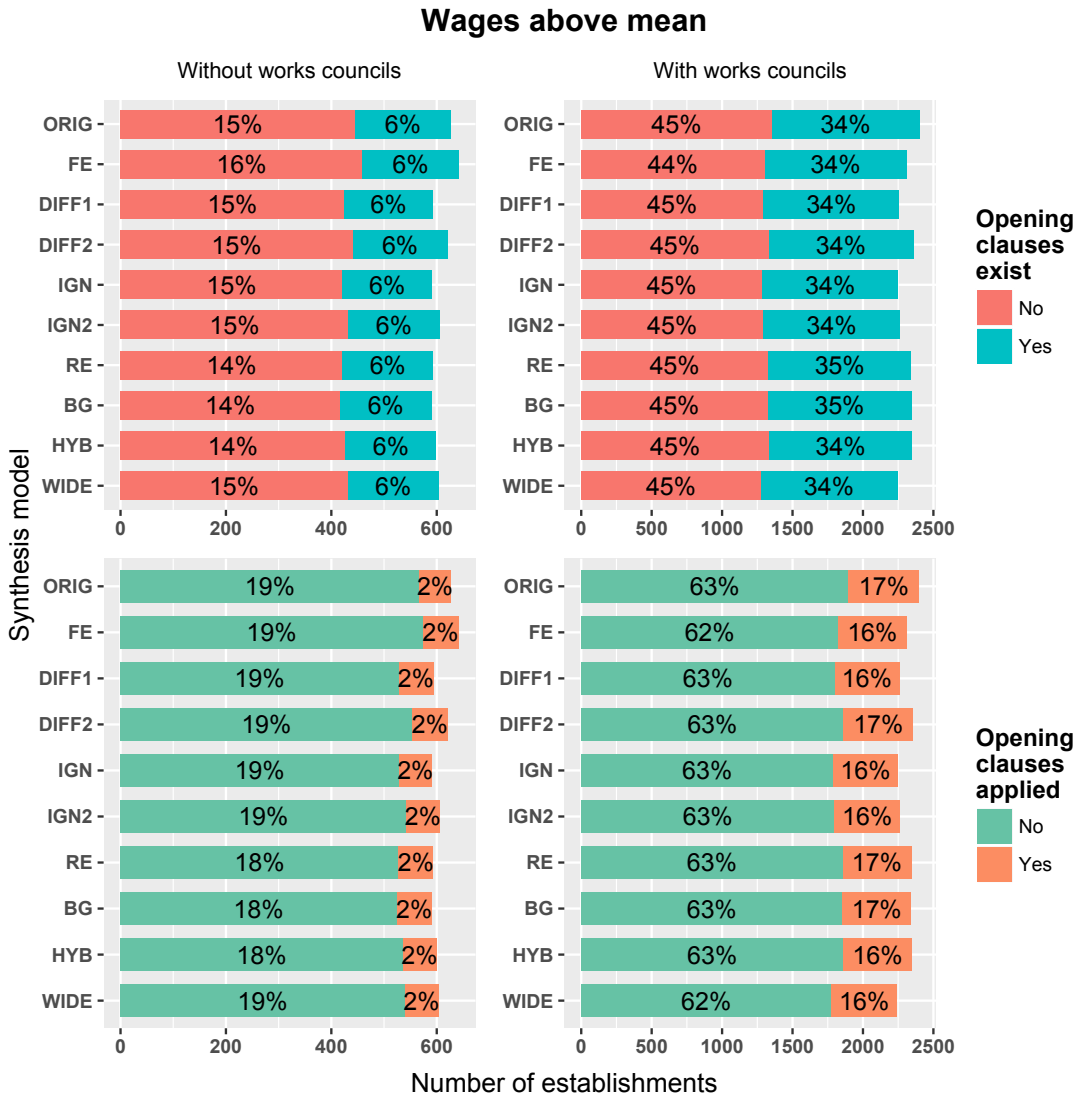

Figure 3.20: Number and proportions of establishments with or without work councils, opening clauses and application of opening clauses for establishments with wages above the mean

OLS regression without application of opening clauses (OLS1)

Following from Ellguth et al. (2014), the first model of interest is:

$$\ln(Y) = \beta_0 + \beta_1 WOCO + \beta_2 OC + \beta_3 OC \times WOCO + \mathbf{x}^{'}\gamma + \varepsilon, \tag{3.8}$$

where, Y is the establishment's total monthly wage bill per full-time equivalent employee, OC is a dummy variable indicating the existence of opening clauses, $WOCO$ is the dummy variable indicating the existence of a works council and $OC \times WOCO$ is the interaction term between the existence of opening clauses and works councils. The term $\mathbf{x}^{'}$ is a vector of potential confounders. These include establishment characteristics such as the proportion of qualified employees, the proportion of employees with fixed-term contracts, the proportion of casual workers, the proportion of part-time employees, the proportion of trainees, the churning rate, a dummy variable indicating if the establishment is single or part of a firm, the technical state of equipment, a dummy variable for investment activities, a dummy variable for foreign ownership, industry dummy variables, establishment size dummy variables, state dummy variables and the year dummy variable.

In Figure 3.21, we report some of the estimates for the coefficients in the model, with bars indicating the 95% confidence interval (CI). The coefficients are divided into four sets for clarity of display, we discuss one set here and the rest of the results can be found in Appendix B. The coefficient names are accompanied by stars to indicate significance of the coefficient at 1% (***), 5%(**) and 10%(*) levels in the original data (ORIG) results. We flag these so that we make particular note of the synthetic data results if they differ noticeably for these covariates.

We find results from our selection of the original data close to those observed in the article. We find the WOCO coefficient significant at 1% level in the original data with a wage premium of having works councils equal to $[\exp(\beta) - 1] = 21.2\%$. We find that all synthetic data estimates fall within the confidence interval from the original data. However, there is some variation in the overlap between the CI from ORIG and the synthetic data. Results from DIFF2syn most closely match the ORIG results. The FEsyn CI also closely match the ORIG CI, and have slightly longer lengths, as we observed in the simulations. The point estimates from the random effects type models (REsyn, BGsyn and HYBsyn) are close the ORIG point estimates as well, with CI of almost the same length as for ORIG CI but the region covered is slightly different. Results that differ the most for both point estimates and CI overlap come from the DIFF1syn, IGNsyn, IGN2syn and WIDEsyn models. The CI for WIDEsyn, FEsyn and DIFF1syn are the longest. While the CI from IGNsyn and IG2syn are not as long, they only overlap about 65% with the ORIG CI.

We observe similar patterns amongst the synthesis models for the coefficients for proportion of qualified employees, trainees and churning rate and single enterprise. Across

the synthetic data models, the results differ less so for the rest of the coefficients in Figure 3.21, apart from the non-significant coefficient for proportion of part-time employees. For the proportion of part-time employees, besides FEsyn and DIFF2syn, the rest of the synthesis models have noticeably larger point estimates. The random effects type models have CIs that contain 0, as does the ORIG CI. In contrast, we observe that synthesising under DIFF1syn, IGNsyn, IGN2syn and WIDEsyn would result in a significant coefficient for the variable in the final analysis model, whose estimates and CI are centred around 0.1.

We find a positive wage premium for establishments with opening clauses, at 8.0%, and an insignificant interaction between WOCO and OC, at -2.2% wage premium. Using IGNsyn, IGN2syn, DIFF1syn and WIDEsyn result in slightly higher wage premiums for OC ($\sim 9.6\%$). Point estimates for FEsyn and DIFF2syn almost match the ORIG results but the FEsyn CI is slightly longer.

Overall, we observe that DIFF2syn and FEsyn estimates consistently match the ORIG estimates and their CI overlap with the ORIG CI but FEsyn CI are always slightly longer than the ORIG CI. The three random effects models (REsyn, HYBsyn and BGsyn) result in similar inferences for most covariates. The WIDEsyn CI are often the longest amongst all those compared, as observed in the simulation study. Point estimates from DIFF1syn, IGN2syn and WIDEsyn closely match each other, except in the case of the year effect, for which IGN2syn estimate is further away from the estimates for each of the other models. Amongst the three, the DIFF1syn and WIDEsyn CI are consistently longer than those for IGN2syn. We find that even though IGNsyn is the closest model to the analysis model, most of the coefficient estimates after synthesising data using IGNsyn do not match the estimates from ORIG and the two CI only partially overlap. This is also true for coefficients significant at 1% level such as WOCO, proportion of qualified employees and trainees and the churning rate (Figure 3.21).

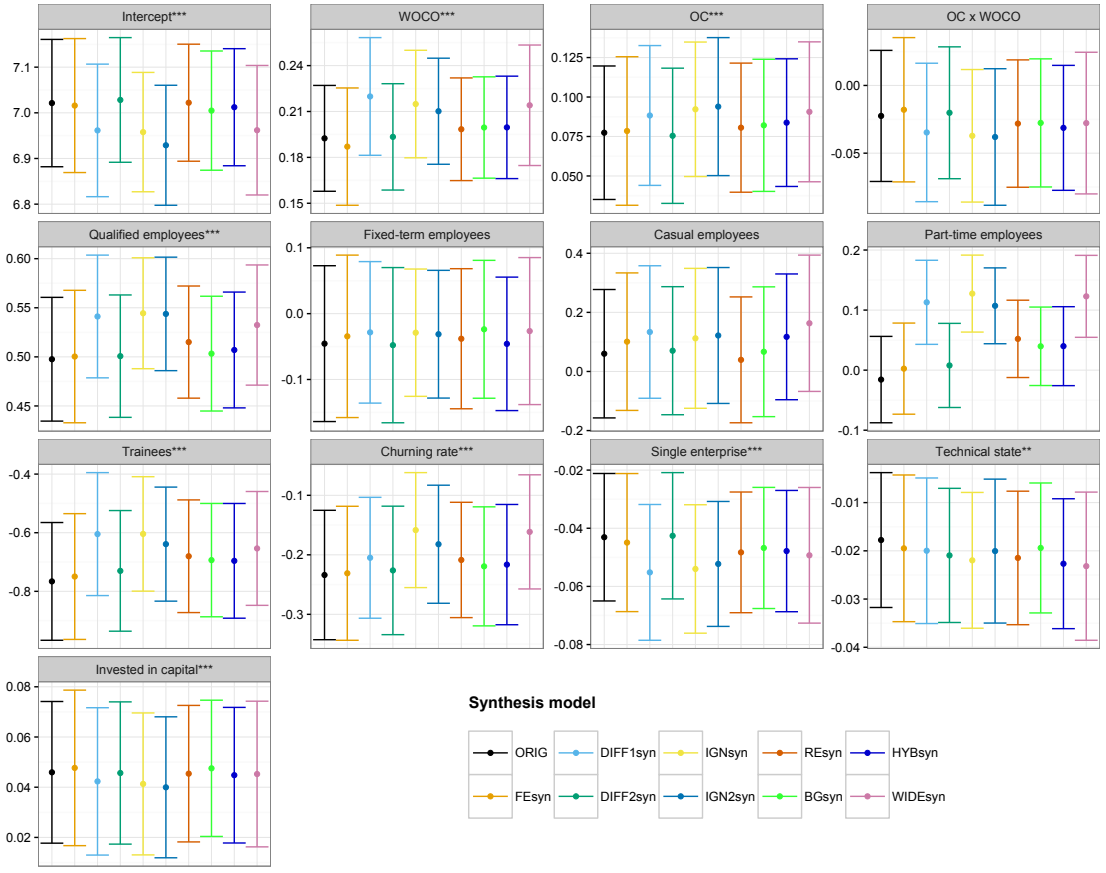

Figure 3.21: Set 1 of the OLS1 coefficient estimates and confidence intervals for both original and synthetic data.

OLS regression with application of opening clauses (OLS2)

The main model for analysis in the article is an extension of OLS1. This model includes information on whether establishments apply opening clauses, OC2, and the interaction of this term with WOCO:

$$\ln(Y) = \beta_0 + \beta_1 WOCO + \beta_2 OC + \beta_3 OC \times WOCO + \beta_4 OC2 + \beta_5 OC2 \times WOCO + \mathbf{x}'\gamma + \varepsilon, \tag{3.9}$$

Figure 3.22 shows one set of coefficient estimates for OLS2. Our observations for OLS1 generally hold true here as well. The main difference is the addition of two new terms related to OC2. For OLS2, our ORIG results do not exactly match the results in the article. In contrast to the article, we did not find the terms OC2 and the interactions OC× WOCO and OC2 × WOCO significant (comparison in Appendix B, Table B.2). However, our ORIG results for the interactions are more or less replicated by all the synthesis models.

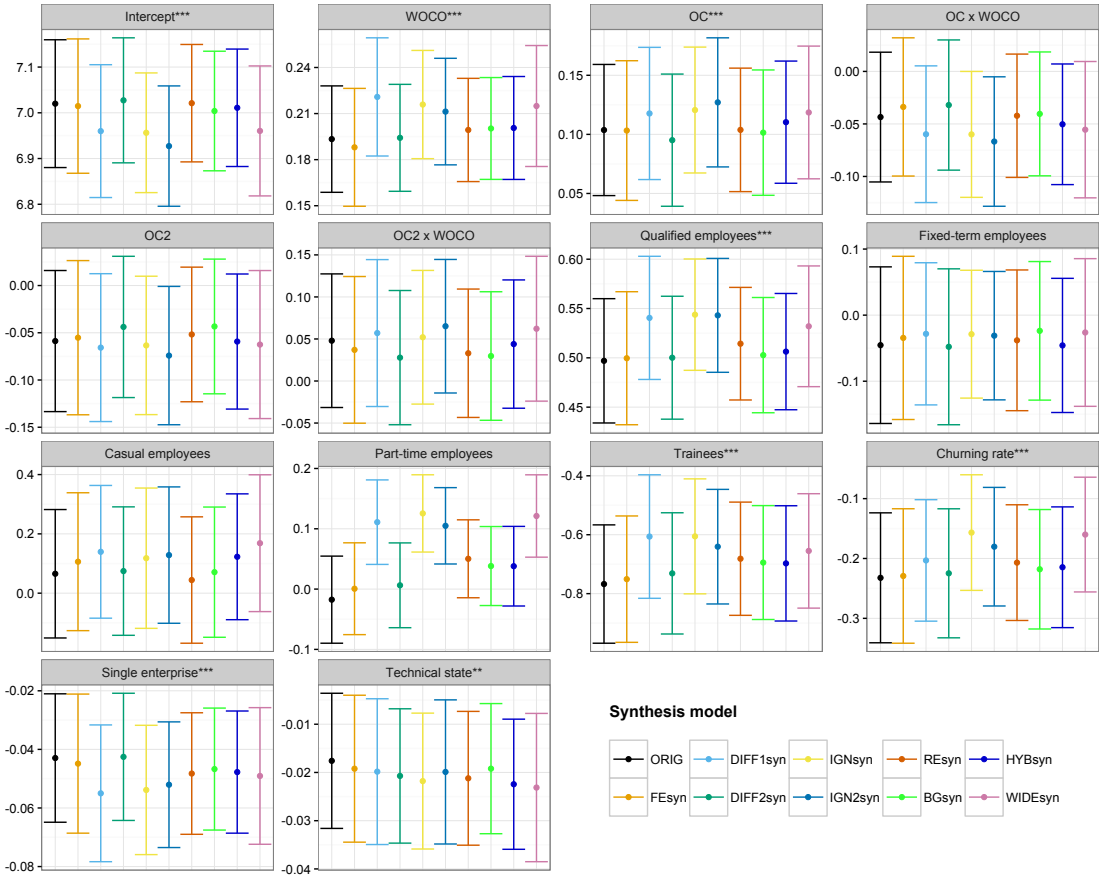

Figure 3.22: Set 1 of the OLS2 coefficient estimates and confidence intervals for both original and synthetic data.

OLS2 for only 21-100 employees (OLS2a), and OLS2 for establishments not belonging to the public sector (OLS2b)

As part of sensitivity checks, the article reports results for five coefficients of interest, after fitting OLS2 for a number of subsets of their selected data. Here, we repeat their analyses and display results for these coefficient estimates and their CIs. We would ideally like our synthesis models to be congenial to each of the subset analyses as well. Figure 3.23 shows the results for OLS2a (OLS2 for only 21-100 employees). Again, we observe that the WOCO results are best replicated by data synthesised using FEsyn, DIFF2syn or any of the REsyn, BGsyn and HYBsyn models. We observe similar trends for the OC and OC2 terms. Results for the two interaction terms are also not too different for the various synthesis models and generally close to those observed for the original data.

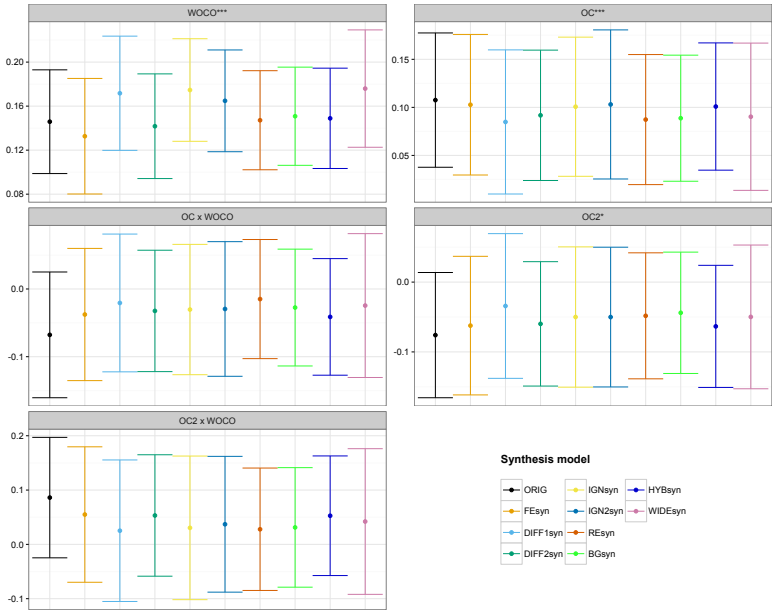

Figure 3.23: Key OLS2a coefficient estimates and confidence intervals for both original and synthetic data.

OLS2b is fitted to establishments outside the public sector only. As in OLS2a, the results closest to the ORIG results come from data synthesised using FEsyn, DIFF2syn or HYBsyn. Other models are close behind. The CIs for the DIFF1syn, IGNsyn and WIDEsyn models are generally longer than those obtained from other models.

OLS2 for year 2005 (OLS2c), and OLS2 for year 2007 (OLS2d)

Figures 3.25 and 3.26 display results when OLS2 is fit separately for each of the two waves of the data, 2005 and 2007. This implies that these regressions are fitted to non-hierarchical data only. We find that the results are preserved quite well for all the covariates. The synthesis models differ slightly in their estimates and CI for the WOCO variable, for which the patterns from results mentioned above are also observed here.

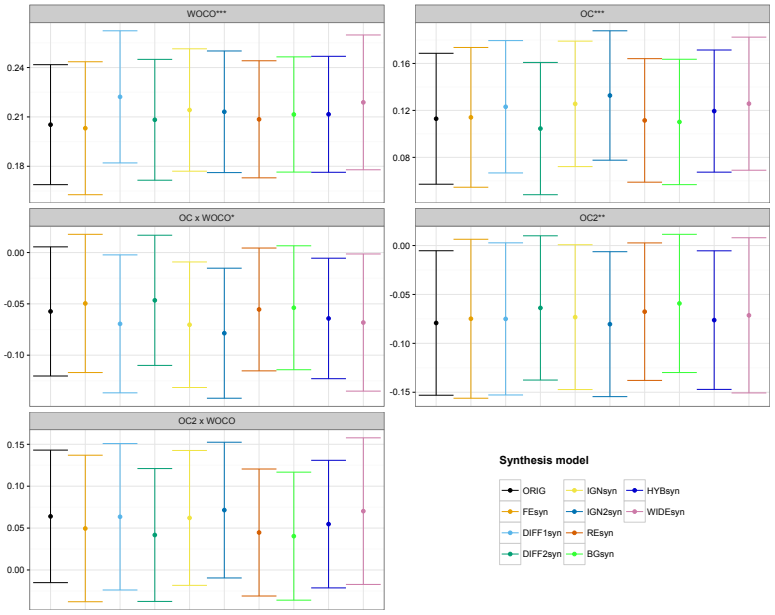

Figure 3.24: Key OLS2b coefficient estimates and confidence intervals for both original and synthetic data.

The estimates are biased away from the ORIG results more for DIFF1syn, IGNsyn and WIDEsyn. Nevertheless, their CI still overlap with over 60% of the ORIG CI.

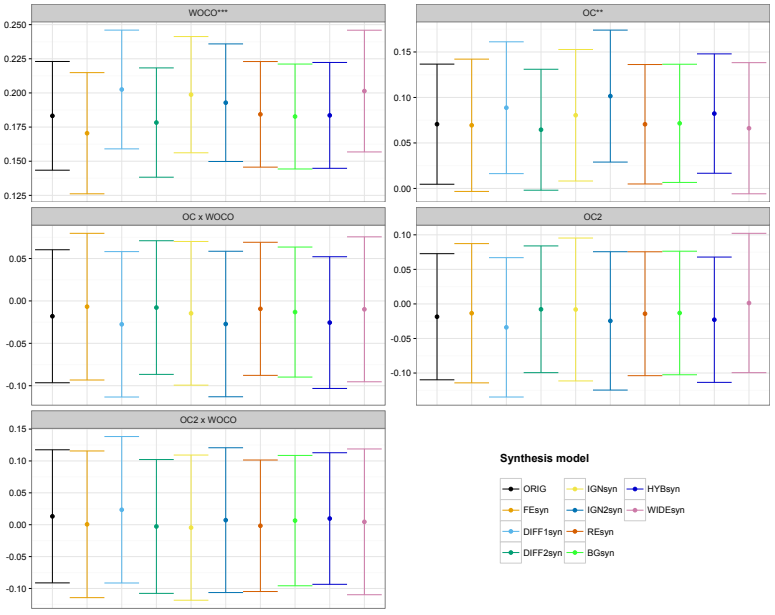

Figure 3.25: Key OLS2c coefficient estimates and confidence intervals for both original and synthetic data.

OLS2 for East Germany (OLS2e), and OLS2 for West Germany (OLS2f)

The next two OLS2 are fitted separately for East and West Germany. We find that for East Germany (Figure 3.27) all the results are closely replicated by the synthesis models. However, the discrepancy in results for DIFF1syn, IGNsyn and WIDEsyn become more

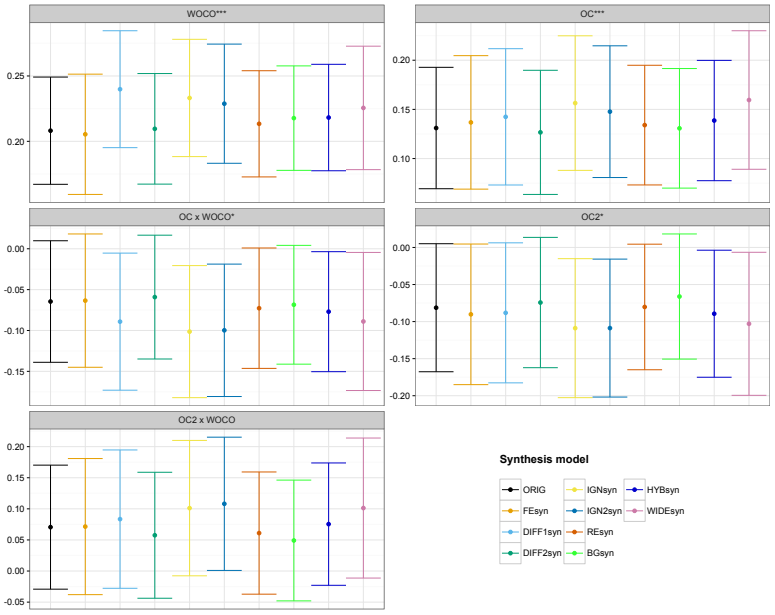

Figure 3.26: Key OLS2d coefficient estimates and confidence intervals for both original and synthetic data.

obvious in the case of West Germany (Figure 3.28), as seen in the slightly biased results for the coefficients of WOCO and OC for the particular models. Again, FEsyn and DIFF2syn stand out for the most closely matched results with ORIG.

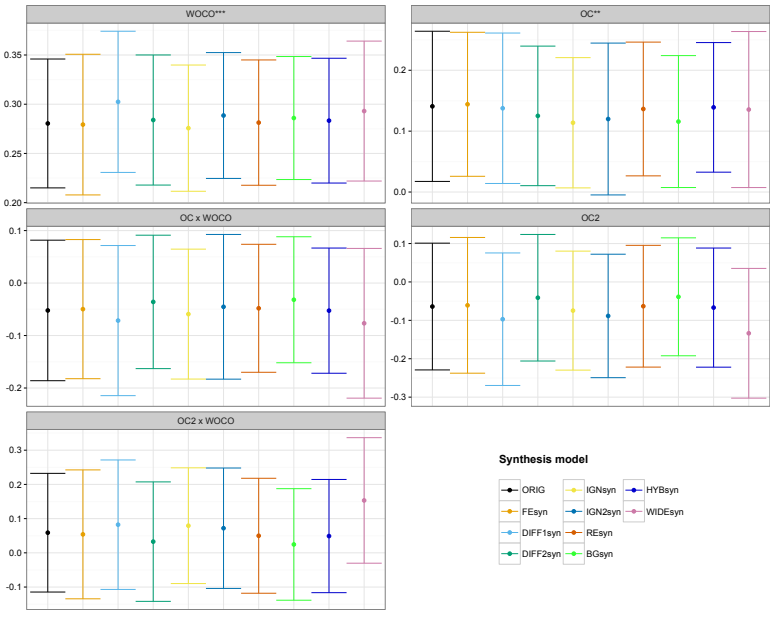

Figure 3.27: Key OLS2e coefficient estimates and confidence intervals for both original and synthetic data.

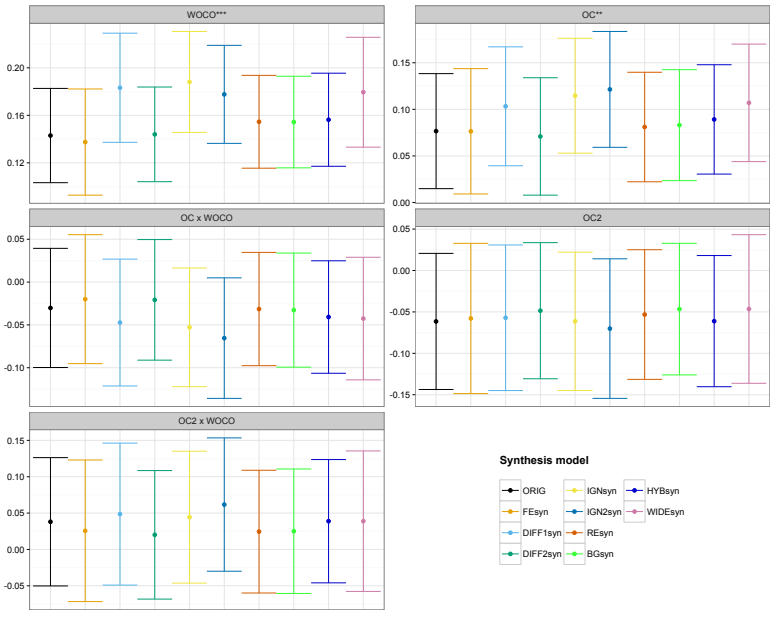

Figure 3.28: Key OLS2f coefficient estimates and confidence intervals for both original and synthetic data.

FE and RE analysis models

We now focus on the models employing fixed and random effects that have been fit in the article to explore the problem in question. As with OLS2a-f, we focus on the estimates for five key coefficients of interest from the article. The authors have fit three models taking into account the hierarchical structure of the data. The first is a fixed effects model (FE) fitted on a balanced data subsetted from their selected data. In direct comparison, there is also a random intercepts model on the balanced dataset (RE (balanced)) and a further random intercepts model for the unbalanced panel (RE (unbalanced)). In the balanced RE model, the authors find four of the five terms (WOCO, OC, WOCO×OC, WOCO×OC2) significant. However, none of the five key coefficients are found significant for the FE model. As such, the two models provide very different inferences, and it is therefore of interest to us whether the use of our various synthesis models preserve the two different sets of analyses. In the unbalanced RE model, all five coefficients are found significant, as in the case of OLS2. As mentioned earlier, our choice of synthesis models implied that we could only use a balanced panel to synthesise and so our results on the ORIG data, do not exactly match the results in the paper. The most significant differences occur in the use of the RE (unbalanced) model, where we do not find the coefficients for OC2 and the two interaction terms significant (Appendix B, Table B.2).

Figure 3.29 shows the results obtained when an FE model is fitted to the ORIG and synthetic data. As in the article, we do not find any of the covariates significant using the ORIG data. However, when we repeat the analysis on data synthesised with REsyn and IGNsyn, the point estimates for WOCO and OC for the models are not only biased, but in fact become significant. As mentioned earlier, the authors of the selected article found the FE model results different to those obtained from all other models. We suspect that the difference in the FE and RE models may stem from omitted variable bias. As expected, we find that there is a noticeable discrepancy between results from the original data and those obtained when data are synthesised using IGNsyn or REsyn. While IGN2syn does not result in biased point estimates, the resulting CI for the model are much longer than those for the other models - observations in line with our simulation studies.

Another noticeable difference in our results for FE analysis, as compared to results from the various OLS regressions above, is that the DIFF1syn and WIDEsyn approaches provide inferences that replicate the ORIG analysis as well as FEsyn, DIFF2syn, HYBsyn or BGsyn. In our simulation studies, we did not find the performance of DIFF1syn or WIDEsyn too variable, as they performed well across all analysis models studied. We believe the results observed in the real data application reflect the consequences of suitability of the model to the dataset, analyses of various subsets of the synthetic data and addition of several covariates in the model. Furthermore, as expected, it is difficult to separate the effects of incompatibility between the data and the analysis model, and synthesis model and analysis model when working with real data.

In Figure 3.30, we display the results when using the RE (balanced) model for analysis. As in the article, results from ORIG data suggest that the WOCO and OC coefficients are significant at 1% level, and the two interactions are also significant at 5% and 10% levels. While the two interaction estimates are generally well-preserved by all the synthesis models, results are more varied for the WOCO and OC coefficients. For WOCO, data synthesised using FEsyn and DIFF2syn most closely match the ORIG results. Following behind are slightly biased results from DIFF1syn, HYBsyn, BGsyn and WIDEsyn. WOCO estimates from IGNsyn, IGN2syn and REsyn models are most biased compared to the rest of the models, and their CI overlap less than half the CI of ORIG, i.e. the ORIG estimate does not lie in the CI for synthetic data estimates for these models. We find similar results for the OC variable, although the biases are slightly smaller. Overall, we observe that the use of IGNsyn, IGN2syn and REsyn inflates the wage premium calculated for the existence of works councils ($\sim$ 22.4% as opposed to 17.1%) and opening clauses ($\sim$ 11.1% as opposed to 6.6%).

Using the RE (unbalanced) model for analysis, we only find WOCO and OC variables significant in the model, as opposed to all five of the key variables under consideration, i.e. WOCO, OC, OC $\times$ WOCO, OC2 and OC2 $\times$ WOCO. The results (Figure 3.31) show trends similar to those observed for the RE (balanced) model, with slightly more variation in the CIs for the interaction terms. The WOCO coefficient estimates are larger than the ORIG estimates while using the REsyn, DIFF1syn and WIDEsyn models, and smaller for FEsyn. DIFF2syn results most closely match the ORIG results and estimates for IGNsyn, IGN2syn and REsyn are most biased, with their CIs not overlapping the ORIG CI as well as for other models. Similar results are also observed for the OC coefficient.

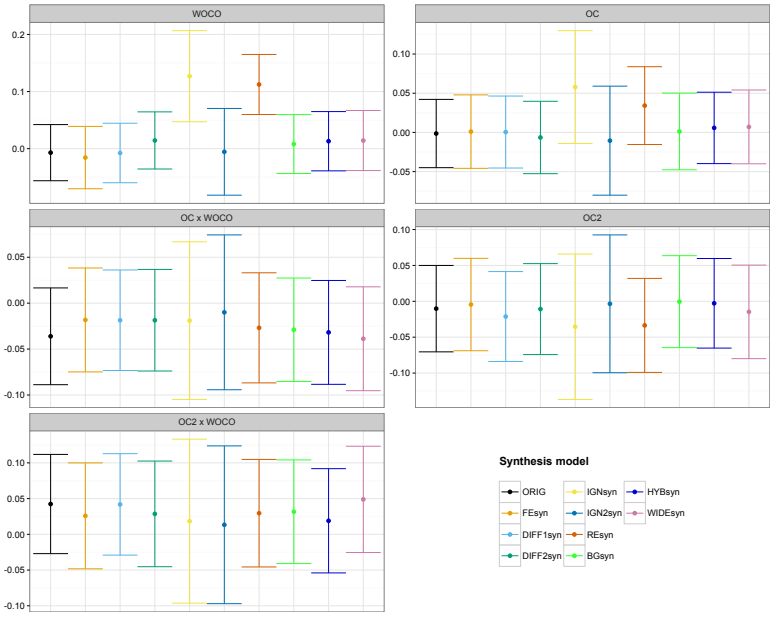

Figure 3.29: Key FE coefficient estimates and confidence intervals for both original and synthetic data.

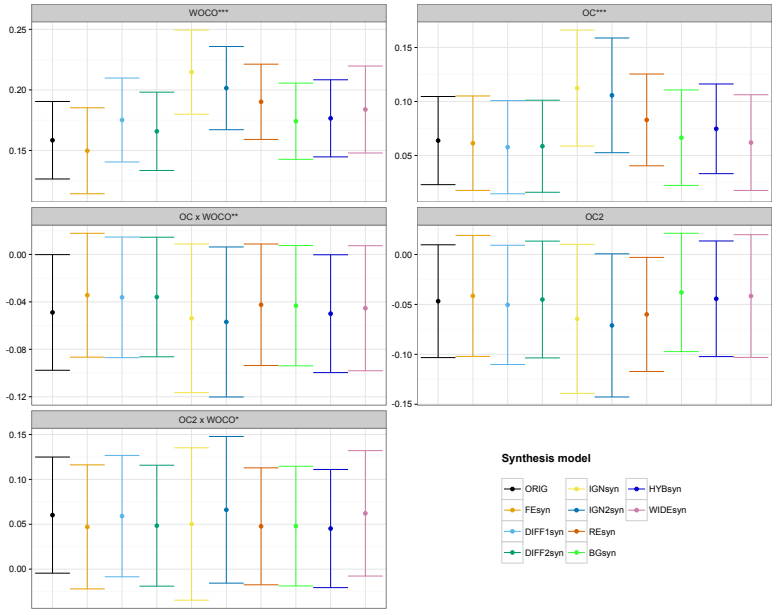

Figure 3.30: Key RE (Balanced) coefficient estimates and confidence intervals for both original and synthetic data.

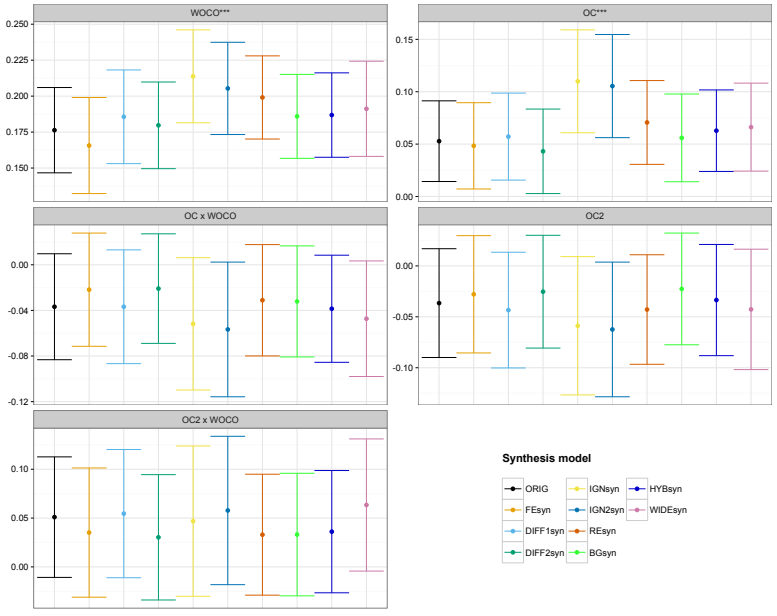

Figure 3.31: Key RE (Unbalanced) coefficient estimates and confidence intervals for both original and synthetic data.

CI Overlap

We also present a summary of the CI overlaps between the CI resulting from the original and synthetic data. To compute the overlaps, we follow the measure documented in Karr et al. (2006). Let (L_{orig}, U_{orig}) be the CI for a regression estimate obtained using the original data, and (L_{syn}, U_{syn}) the equivalent CI obtained using the synthetic data. Let D_{over} be the overlap between the original and the synthetic CI, defined as $\max(0, (U_{\max} - L_{\min}) - (U_{\max} - U_{\min}) - (L_{\max} - L_{\min}))$, where $U_{\max} = \max(U_{orig}, U_{syn})$, $U_{\min} = \min(U_{orig}, U_{syn})$, $L_{\max} = \max(L_{orig}, L_{syn})$ and $L_{\min} = \min(L_{orig}, L_{syn})$. Then the relative overlap in the two CI can be measured using the overlap index (OI):

$$\text{OI} = \frac{1}{2}\left[\frac{D_{over}}{U_{orig} - L_{orig}} + \frac{D_{over}}{U_{syn} - L_{syn}}\right] \tag{3.10}$$

Therefore, the measure OI takes the length of the overlapping section of the two CI, measures this length as a proportion of the length of the two CI and averages this proportion. An OI of 0 means that the two CI do not overlap at all, while the maximum value of 1 indicates that the two CI perfectly match.

We computed the CI overlaps for each covariate as per the definition (3.10) for all parameters in each of the analysis models and present their summary as box plots in Figure 4.6; however, the fixed effects from the FE model have been excluded. For the OLS1 analysis, we observe that the CI overlaps for FEsyn and DIFF2syn are closest to 1 as compared to other models and none of the CI overlaps for these models drop below 0.75. Close behind are the overlaps observed for the random effects type models, REsyn, BGsyn and HYBsyn with the performance of HYBsyn and BGsyn slightly better than REsyn. The overlaps for these models only drop below 0.75 but not below 0.50 for a few coefficients. All other synthesis models, although on average not far behind the best overlap statistics, have a few covariates the CI for which do not overlap much with the ORIG CI. These include, DIFF1syn, IGNsyn, IGN2syn and WIDEsyn. We observe similar trends for all the synthesis models for OLS2 to OLS2f. So far, when considering CI overlap, DIFF2syn again, stands out as a clear winner, closely followed by FEsyn, HYBsyn and BGsyn.

If we consider the FE analysis model instead, we find slightly different results. Here the use of FEsyn and BGsyn seem to provide the best results, although both of them also result in a few CI overlaps less than 0.50. In this case, WIDEsyn outperforms REsyn, with higher CI overlaps on average as in the case of HYBsyn and DIFF2syn. While all synthesis models results in a few overlaps of nearly or equal to 0, it is apparent that the analyst is worse off when data are synthesised using IGNsyn, IGN2syn, REsyn and DIFF1syn. This is in line with our discussion above where we considered the covariates individually.

When the analyst is interested in an RE model instead (whether for balanced or unbalanced data), we find that most models have a similar performance. In fact, FEsyn, HYBsyn and BGsyn have similar results. REsyn, DIFF2syn and WIDEsyn follow with an average overlap of 0.875, although the REsyn overlaps have more variability as the length of the box plots indicate. As expected, we don't find that the IGNsyn and IGN2syn models perform as well.

When studying CI overlap, we note that as long as the point estimates are accurate, FEsyn CI have a greater chance to perform well as the CI for FEsyn are generally longer than those for models with random effects. When making inferences, analysts may prefer CIs obtained when using HYBsyn or BGsyn, as we expect these to be shorter because the models with random effects are more efficient than FEsyn. In the results for the regression coefficients, and for the CI overlap, we found the performance of DIFF1syn and WIDEsyn more variable over the different analysis models, as opposed to the simulation results. We suspect this is a result of lack of covariate selection at the synthesis stage. The REsyn CI overlaps indicate that the model performed generally well for the various OLS analysis models but the results are more variable when the analyst is interested in either of FE or RE models.

Figure 3.32: Boxplots of overlap between original and synthetic data CI for all analysis models under consideration.

Overall, there is a clear indication of bias induced from the use of IGNsyn and REsyn as synthesis models for the WOCO and OC coefficients. These might be a result of

OVB. We find in general that the use of IGNsyn model for the data results in the most biased coefficient estimates. While IGN2syn improves upon IGNsyn's performance when the analysis model used is also non-hierarchical, the gains are not so obvious when the analysis models employs establishments' fixed or random effects. In our analyses of interest, the article reported the difference in results when using either an FE or an RE model. We find these differences are passed on from the synthesis to analysis stage. For all the hierarchical analysis models, REsyn data fails to replicate the ORIG analyses for certain key covariates of interest, while results from FEsyn still look promising, regardless of the choice of the analysis model. HYBsyn and BGsyn performed generally well for the various analysis models. The CI resulting from using either of these models are generally tighter than those obtained from FEsyn, and point estimates are close to the ORIG results for most analyses. Performance from WIDEsyn is generally more variable, where results match the ORIG results but not as closely as for FEsyn, and the CI are longer. Throughout all our analyses, we found DIFF2syn best replicated the ORIG results, for all significant and non-significant, time-constant and time-varying variables, and models for non-hierarchical and hierarchical data. As in the simulation, we expect this will also impact disclosure risks for such synthetic data, which we address below. In our simulation study, we did not model time-constant covariates. The real data application shows that this has not changed our conclusions dramatically as we modelled several variables in the data. Our application also serves as an interesting, and realistic, example where the IGN model is not suitable for an IGN model, and an RE model is also detrimental for RE analysis.

We caution against the use of the above results as an absolute guarantee of data utility. In the course of research, it is most sensible to assess data utility with models and analyses that are suited to the data. In a simulation study, we ensure that we have at least one analysis model which is the 'correct' model for the data. With real data, however, it is not possible to determine which of the analysis models is the 'correct' model for the data. It is then a philosophical concern whether the synthesis process must preserve all analyses, even if they are not the correct for analysis purposes. The fact that an analysis model is not the correct model for the data may be the reason behind some of the discrepancies between the original and synthetic data analysis in the results above. Our view is that as a data keeper, we must deal with as many problems as we can for failure of model assumptions. For instance, we recommend using HYBsyn when dealing with OVB. This ensures that even if the data analyst is using an REan model which will suffer from OVB, this bias will be no worse than what would be obtained using the original data.

3.4.2 Disclosure risks

We now utilise the disclosure risk measures described in Section 3.2 to assess the relative risks of identification when using data synthesised from the various models. We assume that the intruder knows the federal state for each establishment before trying to match records with the true wages. It is not realistic to assume that the intruder only knows the federal state for each establishment, and more information may be publicly known. Nevertheless, it is also not realistic that only one variable is synthesised when releasing the data. As such, we do not draw any conclusions regarding the magnitude of risk measures reported here, and only compare the results between the synthesis models. Our measures include the expected match risk, the perceived match risk, the true match rate and the false match rate.

Table 3.10 shows the results for each of the synthesis models. We find that the risk profile for each of the synthesis models is in line with the data utility obtained for that model. As such DIFF2syn stands out as the most risky model, with an expected match risk of nearly 287. This is approximately the number of matches the intruder can make by random guessing. The perceived match risk is the total number of unique matches obtained by the intruder subject to a minimum probability requirement (here, 0.2). For the DIFF2syn model, the intruder makes 951 unique matches, although only 50% of them are correct as indicated by the false match risk. The true match risk indicates that the intruder can identify 25% of the establishments correctly from the released data.

Risks from all other models are relatively smaller. The next most risky model is FEsyn with about 90 expected match risk and only 5% of true match rate. Following behind are REsyn, BGsyn and HYBsyn, with nearly 200 unique matches each, out of which more than 85% are incorrect and less than 5% of establishments are identified correctly. The rest of the models, i.e. DIFF1syn, IGNsyn, IGN2syn and WIDEsyn, are the least risky overall with approximately 2% true match rate.

Synthesis model	Expected	Perceived	True rate	False rate
FEsyn	91.369	296	0.053	0.843
DIFF1syn	45.639	171	0.024	0.926
DIFF2syn	286.964	951	0.250	0.501
IGNsyn	43.698	168	0.020	0.932
IGN2syn	44.125	154	0.017	0.944
REsyn	83.296	273	0.044	0.869
BGsyn	82.812	273	0.040	0.882
HYBsyn	86.607	269	0.048	0.861
WIDEsyn	45.250	139	0.020	0.933

Table 3.10: Real data disclosure risks: expected risk, perceived risk, true rate and false rate.

We find that although DIFF2syn provided the best utility results, it appears to be the most risky as well. As before, we find the use of the DIFF model an attractive idea, with the potential of having different risk profiles. In our real data illustration, a good option is the use of FEsyn, which produces excellent analysis results. However, WIDEsyn provides about half the risks of FEsyn while slightly compromising on data utility. Although it is not possible to make any clear recommendations, the overall advantage of using WIDEsyn or DIFF1syn as opposed to IGNsyn or IGN2syn is quite clear; these models provide much better data utility with a similar risk profile. Data keepers, however, may want to preserve the analysis even better by using one of HYBsyn or BGsyn, but not REsyn, the disadvantages for which we have discussed in the simulation studies and the real data results above. We have identified three groups of models according to their risk profiles: 1) DIFF2syn, 2) FEsyn, REsyn, BGsyn, HYBsyn and 3) DIFF1syn, WIDEsyn, IGNsyn, IGN2syn. A data keeper can first select the risk profile that is acceptable to the agency, and then choose the model with the best data utility within the chosen group. We present a risk-utility (R-U) map for our application that can help data keepers with this decision. This is a plot displaying data utility against disclosure risks for all the synthesis models. As an illustration, we choose to plot the average CI overlap (utility) against the true match rate (risk) in Figure 3.33.

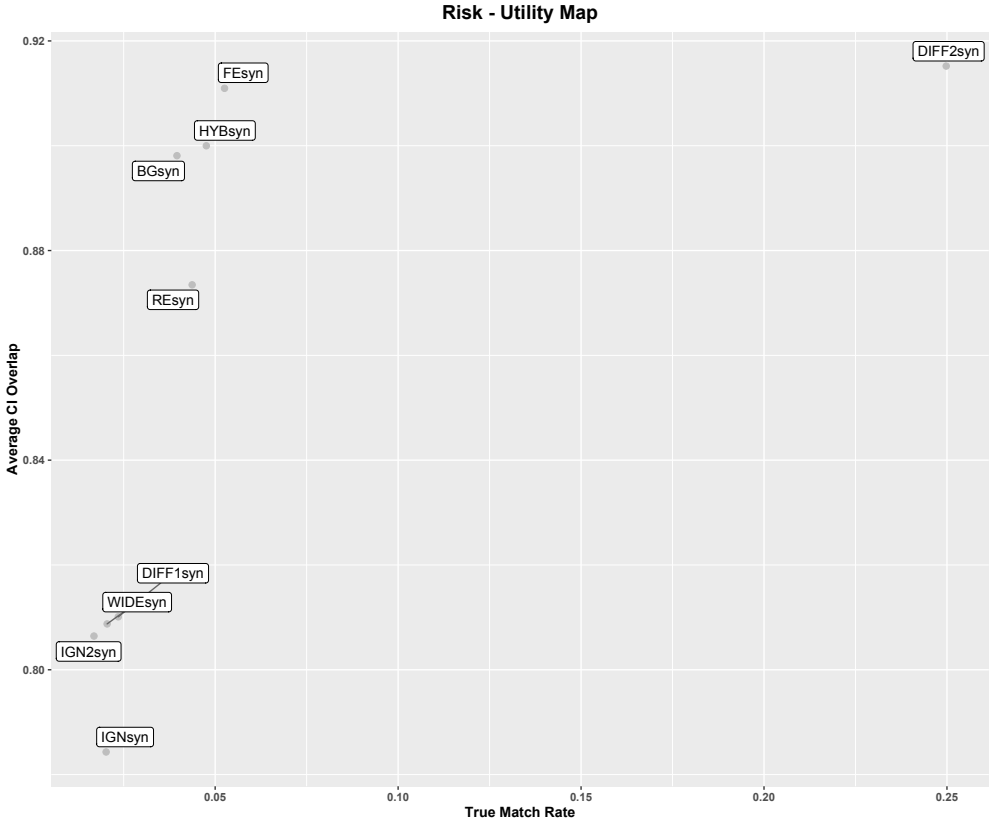

Figure 3.33: Plot of data utility (CI overlap) against disclosure risks (True match rate) for various synthesis models.

Figure 3.33 provides a visual aid with regards to the relative utility and risks associated with each of the synthesis models; it is easy to identify the three groups of models mentioned earlier in the graph. Similar plots with various utility and risk measures can be used to identify the cost associated with each of the synthesis models under consideration, and the choice of models may be made using these.

3.5 Conclusions

In this Chapter, we have studied combinations of a number of factors with a focus on omitted variable bias to investigate synthesis models that can be used for hierarchical data. We ran simulation studies with and without OVB, with large and small ICC, using posterior predictive distribution of the sensitive data and plug-in estimates, and studied both data utility and disclosure risks. Within each of the combinations of these factors, we studied nine synthesis models and their performance using five analysis models. A real data illustration provided us further insights. The main findings from this Chapter are summarised below.

When no omitted variable exists:

- the use of a single simple linear regression (IGNsyn, IGN2syn) is unsuitable for synthesis, as it results in biased estimates for variance of regression coefficients. For both models, results vary slightly with changing data ICC and the use of MLE estimates instead of the posterior predictive distribution. The models result in low disclosure risks that do not generally vary across the number of copies of the data released;

- both FE and RE models are unsuitable for synthesis based on the assumption that the analyst may use any of the two models to study the partially synthetic data. Some of the variance estimates for regression parameters or the error term are biased for the cross combinations FEsyn-REan and REsyn-FEan. Results for both the models are worse when data ICC is small. The disclosure risks associated with both the models appear to be similar but the use of the MLE approach increases disclosure risks for FEsyn and reduces these for REsyn;

- the use of WIDEsyn is more suitable; this is already implemented in the literature and known to add extra variability in the synthetic data. The use of the MLE approach helps reduce this variability. For moderate sized datasets, WIDEsyn may be one of the best options for the data keeper, given that the disclosure risks associated with it are small;

- the DIFF approach is a new approach that allows the data keeper the flexibility to model the data means separately from the differences. The DIFF1syn approach

has results comparable to the WIDEsyn approach, both in utility and risks. The DIFF2syn approach is associated with higher disclosure risks. Extensions to the DIFF approach may provide further options for data keepers interested in more complicated models such as those with random slopes or of non-linear form;

- the HYBsyn and BGsyn approaches are mathematically equivalent. However, there may be slight differences in their results as HYBsyn has a faster convergence rate compared to BGsyn. In this sense, HYBsyn may always be preferred over BGsyn. HYBsyn provides excellent results for data utility but the risks associated with its use are higher than those for WIDEsyn, similar to those for FEsyn or REsyn, and lower than those for DIFF2syn. When data ICC is small, the risks for HYBsyn are similar to those for WIDEsyn or DIFF1syn. The use of the MLE approach also results in reduced risks for HYBsyn.

When there is an omitted variable correlated with observed covariates:

- the performance of IGNsyn suffers because of OVB. While IGN2syn provides unbiased point estimates for the regression coefficient, the final variance estimates are still biased;

- FEsyn appears to be one of the best options when OVB exists, the slight increase in between variance estimates of REan is hardly noticeable given the consequences of the omitted variable. Although the fixed effects approach has performed well in the real data illustration, it has some limitations. With increasing number of clusters, the number of parameters will increase. Also, no variables constant within a unit can be modelled using the approach;

- the performance of REsyn suffers from OVB at the synthesis stage, although the size of the bias that goes through to the analysis stage may be controlled by using the MLE approach to synthesis as opposed to the posterior predictive. In the real data illustration, REsyn failed to preserve the values of key coefficients of interest. We note that once a synthesis process induces OVB into the data, no amount of bias correction in the analysis models can remove this bias. This bias is always a possibility, because whether any omitted variables exist for a given set of analysis or not, is unknown. The synthesis process, therefore, should be undertaken in consideration of this possibility. If the synthesis model itself has OVB, attempting to match synthesis and analysis models for congeniality can only be secondary concern;

- the use of WIDEsyn, HYBsyn, DIFF1syn or DIFF2syn resolves the biases that may result from using REsyn in point estimates, and results are similar to those obtained from the use of FEsyn. The real data application confirmed that data utility for DIFF1syn and WIDEsyn is lower than for DIFF2syn, HYBsyn or FEsyn and the disclosure risks are in line with results for data utility;

- disclosure risks for all models follow similar trends as in the case of simulations without an omitted variable.

The results in this Chapter highlight three synthesis strategies with a potential to provide acceptable data utility in various scenarios. These are HYBsyn, WIDEsyn and DIFFsyn. HYBsyn is found congenial to many different analysis models and computationally efficient. Whether an omitted variable exists or not, HYBsyn consistently performs well. However, it does result in potentially high disclosure risks. If the data keeper is prepared to sacrifice on data utility, the WIDE approach may be a suitable option. The WIDE approach may be a preferred choice of synthesis not only because of convenience as in the previous literature, but specifically because it performs well for utility for a very small price of added variability. In addition, it is simpler than using hierarchical models. However, we only conducted studies with a limited number of covariates. WIDEsyn can be difficult to implement when either the number of waves in the data, or the number of variables is large. Nevertheless, as our simulations show, the variability from the WIDE approach can be reduced by using the MLE approach to synthesis instead of using the posterior distribution, which provides further convenience. Moreover, the WIDE approach performs well for disclosure risks. In this unexpected set of results, we find that the WIDE approach is a competitive approach to data synthesis. It is fast and convenient, preserves utility and controls for disclosure risks better than other hierarchical data synthesis models.

Finally, we note the potential in novel synthesis models such as the DIFF approach which can provide flexibility in the synthesis procedure and control over disclosure risks. However, the setup for these requires more attention than for other models. The DIFF1 approach illustrates how the DIFF approach may be used to control disclosure risks, while the DIFF2 approach consistently displayed excellent results for data utility.

It will also be useful to study the WIDE approach more carefully and use model selection to reduce each sub model within the approach, as would be undertaken in a real data scenario. Each of our models may also perform differently when the data are unbalanced. Furthermore, we have overall not employed any model selection criterion at the synthesis stage. It would be interesting to compare the various synthesis models in terms of various model fit criteria and explore what these imply in terms of selecting a synthesis model.

There is also more that could be done to investigate disclosure risks. So far, we have employed simple Euclidean distance based measures and have not considered other measures. Moreover, we have left the question of choice of a suitable calliper measure unanswered. It is also possible to investigate other non-model based methods to measure disclosure risks. As noted earlier, it would be useful to formulate disclosure risk assessment criterion specific to variables measured over time and easily identifiable observations in the data, such as big businesses.

We note that one of the disadvantages of using hierarchical models for data synthesis, as opposed to non-hierarchical models, is that additional model assumptions may need to be satisfied. This implies that there is a greater chance of model misspecification. For example, the use of HYBsyn requires that the between error term is distributed normally; this may easily be violated. Most social sciences data contain outliers and unusual observations that may not be easily handled by standard strategies, and require special strategies. This is something we investigate further in Chapter 4.

Chapter 4

Multiple imputation for synthetic data and distribution of the residuals

So far we have studied that, if the correlations between model covariates and the error term are not accounted for, synthetic data may not be fit for release. There are a number of other ways in which regression models for synthesis may be misspecified. We now turn our focus towards the shape of the distribution of the model residuals. This Chapter is organised as follows. Section 4.1 provides an introduction to the problem. In Section 4.2 we describe both existing and proposed modelling procedures for residuals with non normal distributions within a multiple imputation (MI) framework. We then compare the described methods in Section 4.3 through simulation studies. A real data application of interest follows in Section 4.4 and we conclude the Chapter with a discussion in Section 4.5.

4.1 Introduction

We consider data synthesis through regression modelling as a combination of two parts: 1) the preservation of the fitted mean, i.e. the model coefficients, $\hat{\beta}$, and 2) the preservation of the errors, i.e. the model residuals. In this Chapter, we assume that the data keeper has employed the most appropriate statistical techniques to estimate $\hat{\beta}$ and do not address the bias and efficiency of these. Our focus, instead, is on the synthesis of what is left over given the synthesis model and its $\hat{\beta}$ estimates, i.e. the residuals. More specifically, we focus on the preservation of the shape of the distribution of the residuals. When standard regression models are used to synthesise data, new residuals are often generated using the normal distribution. This may not be appropriate if the data do not follow a normal distribution or contain outliers. Generating the residuals from a normal

distribution, therefore, may not preserve the shape of the original data in the synthetic copies.

We argue that the shape of the residuals matters and deserves more attention. There are number of statistical estimates and models of interest that make use of the shape of variables involved. From simple estimates such as the quantiles of the data, to more sophisticated analyses, for example a study of the characteristics of the top 5% of the population through quantile regression. In this Chapter, we compare some alternatives to the use of the normal distribution to generate synthetic residuals. Through the synthetic data, we aim to deliver the most important features of the data to the analyst. This will include reproducing outliers, non normality of residuals, and correlations within the error terms if these appear in the data. However, no synthesis strategy is complete without the study of corresponding disclosure risks. Unusual observations may be important features of the data involved but they also make easy targets for intruders, as they are rarer. As always, here we face the utility-risk trade off.

This Chapter has two aims. Firstly, we extend some of the existing proposals in the literature that help deal with non normal residuals, namely, the use of Box-Cox transformations, quantile regression and flexible parametric distributions, to the synthetic data scenario. Secondly, we further extend these methods from a non-hierarchical to hierarchical data setting. We present a comparison of these methods though simulation studies, and a real data application.

We achieve our two aims by utilising advances in the field of Statistics, which have opened up a plethora of possibilities for agencies planning to synthesise confidential data. These include:

1. the ability to apply multiple imputation for several variables without having to specify a joint distribution (van Buuren et al., 2006; Raghunathan et al., 2001),
2. the elimination of the requirement of draws to be obtained from posterior predictive distributions for partially synthetic data (Reiter and Kinney, 2012) and,
3. the advances in random data generation from sets of univariate non-normal distributions through the multivariate normal distribution, for instance NORmal To Anything (NORTA) (Cario and Nelson, 1997).

Essentially, 1) allows agencies to be able to model different types of variables within the same dataset using different models, 2) facilitates the use of models which may be too complicated to run in a Bayesian setting and 3) makes it easy to take into account non-normal distributions without compromising on the joint variance-covariance structure of the data involved. While it is difficult to provide universal advice on what models are best for different data, there are more modelling options for agencies than ever before. We provide only a hint of this sentiment through this Chapter.

4.2 Methods

In this section, we describe the theory and computation behind a few selected methods from the literature that we compare later in illustrative simulations in Section 4.3, where we set up two different scenarios: one for non-hierarchical and another for hierarchical data. Therefore, each of the methods described in this section must be applicable in both the cases. Where a regression model is required, we use the simple linear regression throughout the non-hierarchical case. For the hierarchical data scenario, we restrict our focus to a random intercepts model only. In both the scenarios, we describe the methods for a dataset with two variables, the sensitive response variable Y and the covariate $\boldsymbol{X}$, where capital $Y, \boldsymbol{X}$ denote vectors and individual observations are indexed as y_i, x_i, where $i = 1, \ldots, n$ denotes the number of observations. With hierarchical data, there are two indices, y_{it}, x_{it}, where $i = 1, \ldots, n$ indicates the unit/cluster number and $t = 1, \ldots, T$ the time point/repeated measurement number.

4.2.1 Transformations

Transformation of response variables may remedy heterogenous variances or non normality of errors. It is not necessary that both of the problems may be solved by a single transformation, and therefore, one of the two must take precedence. However, the two issues usually go hand-in-hand and it is common to be able to deal with both with a single transformation. The Box-Tidwell (Box and Tidwell, 1962) method can be used to transform the independent variables. Here, we focus on transformations for the dependent variable only, which change the distributional properties of the dependent variable, and hence, the distributional characteristics of the errors. Box and Cox (1964) presented the power transformations that address both the problems of heterogenous variances and non normality. Their proposed transformation for the dependent variable Y is of the form:

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^{\lambda}-1}{\lambda \dot{y}^{\lambda-1}} & \text{if } \lambda \neq 0; \\ \dot{y} \ln(y) & \text{if } \lambda = 0, \end{cases}$$

$$\text{where, } \dot{y} = exp \sum \ln(y_i)/n. \tag{4.1}$$

The Box-Cox method assumes that the residuals for the linear model follow the standard assumptions of normality, homogeneity and independence for a given value of λ. With this assumption, the maximum likelihood estimates for the parameters of interest and λ, $\hat{\lambda}$, are obtained. Confidence intervals around the $\hat{\lambda}$ can also be constructed. We note that the value of $\hat{\lambda}$ is dependent on the form of the model. The method attempts to serve three purposes simultaneously: 1) linearising the relationship between the response and predictors, 2) making the variance of errors constant and 3) improving the normality

of the residuals. Box and Cox (1964) showed how the effect of λ may be allocated to the three purposes. The Box-Cox procedure may make the distribution of the residuals closer to a normal distribution than on the original scale but this may still not be very close, and therefore, the usual normality checks are required after the model is fitted. Large values of λ are considered suspect, and may be caused by big outliers in the dataset. Additionally, not all shapes of data can be made to look normally distributed with power transformations alone, for instance, variables with a large peak at 0.

The Box-Cox method is valid for positive observations only. To deal with negative values, an appropriate shift value may be used to first convert all the observations to positive values and then apply the method. Alternatively, Yeo and Johnson (2000) proposed a different set of power transformations that can be applied to any range of data. A different method for transforming a response variable, but only to achieve linearity, was proposed by Cook and Weisberg (1994). This involves plotting the fitted values for a linear model without transforming the response, against the observed values, and judging which transformation suits best. The transformations can be of any form and are not required to be power transformations. In this Chapter, we study the standard Box-Cox transformation only.

For the random intercepts model, we utilise the extension to the Box-Cox transformation proposed by Gurka et al. (2006). The authors note that using the scaled transformation for mixed models, as in (4.1), results in a standard data likelihood for linear mixed models and may be used as such. To find the optimal transformation, we maximise this likelihood over a grid of values of λ using the Restricted Maximum Likelihood method.

The following steps are required for partial data synthesis using the Box-Cox transformations in the case of non-hierarchical data:

1. search for the optimal λ, $\hat{\lambda}$, over a grid of values for a regression of Y on $\boldsymbol{X}$;
2. regress $Y^{(\hat{\lambda})}$ on $\boldsymbol{X}$, and obtain the estimate for the regression coefficients, $\hat{\beta}$, and the error variance, $\hat{\sigma_e^2}$;
3. generate M synthetic copies of Y, $y_{i,syn} = x_i\hat{\beta} + e_i$, where e_i are obtained as draws from a $N(0, \hat{\sigma_e^2})$ distribution,
4. transform back the Y_{syn} using $\hat{\lambda}$.

For hierarchical data, the process is essentially the same, with a random intercepts model instead:

1. search for the optimal λ, $\hat{\lambda}$, over a grid of values for a random intercepts regression of Y on $\boldsymbol{X}$;

2. regress $Y^{(\hat{\lambda})}$ on $\boldsymbol{X}$, and obtain the estimate for the regression coefficients, $\hat{\beta}$, and the error variances, $\hat{\sigma_e^2}$ and $\hat{\sigma_b^2}$;

3. generate M synthetic copies of Y, $y_{ij,syn} = x_{ij}\hat{\beta} + e_{it} + e_i^b$, where e_{it} are obtained as draws from a $N(0, \hat{\sigma_e^2})$ distribution and e_i^b are generated independently from a $N(0, \hat{\sigma_b^2})$ distribution,

4. transform back the Y_{syn} using $\hat{\lambda}$.

For both the simulation studies we search for an optimal λ by testing 30 values equally spaced between -2 and 2.

4.2.2 Quantile regression

The quantile function is defined as the inverse of the cumulative distribution function (CDF). For a variable Y, the quantile function, $Q_Y(\tau)$ is defined as $F_Y^{-1}(\tau) = \inf\{y : F(y) > \tau\}, \tau \in [0, 1]$. For strictly increasing and continuous $F(.)$, $Q_Y(\tau)$ is a unique real number y such that $F(y) = \tau$ (Chap. 1, Gilchrist (2000)). In the case of regressing Y on $\boldsymbol{X}$, the conditional quantile regression is defined as,

$$\hat{q}_Y(\tau, \boldsymbol{X}) = \underset{Q_Y(\tau, \boldsymbol{X})}{\arg\min}\, E[\rho_\tau(Y - Q_Y(\tau, \boldsymbol{X}))], \tag{4.2}$$

where, $Q_Y(\tau, \boldsymbol{X}) = Q_\tau[Y|\boldsymbol{X} = x]$ is the conditional quantile function and the loss function being minimised, $\rho_\tau(y) = [(1-\tau)I(y \leq 0) + \tau I(y > 0)]|y|$. We currently focus on linear models, so that:

$$\hat{\beta}(\tau) = \underset{\beta}{\arg\min}\, E[\rho_\tau(Y - \boldsymbol{X}\beta)] \tag{4.3}$$

are the τ-specific coefficient estimates for the covariates involved. The $\hat{\beta}(\tau)$ are invariant to certain transformations. These are summarised by Koenker and Bassett Jr (1978) as follows:

$$\begin{aligned}
&\hat{\beta}(\tau; aY, \boldsymbol{X}) = a\hat{\beta}(\tau; Y, \boldsymbol{X}), && a \in [0, \inf),\\
&\hat{\beta}(\tau; -aY, \boldsymbol{X}) = -a\hat{\beta}(1-\tau; Y, \boldsymbol{X}), && a \in (-\inf, 0],\\
&\hat{\beta}(\tau; Y + \gamma\boldsymbol{X}, \boldsymbol{X}) = \hat{\beta}(\tau; Y, \boldsymbol{X}) + \gamma, && \gamma \in \mathbb{R}^K,\\
&\hat{\beta}(\tau; Y, \boldsymbol{X}A) = A^{-1}\hat{\beta}(\tau; Y, \boldsymbol{X}), && A_{K\times K} \text{ nonsingular.}
\end{aligned}$$

The quantiles are also invariant to monotone transformations, i.e. for any monotonic function $h(.), q_{h(Y)}(\tau) = h(q_Y(\tau))$ as $P(Y \leq y) = P(h(Y) \leq h(y))$.

For MI using quantile regression, we follow the method proposed by Bottai and Zhen (2013) and Geraci (2016). Following is the basic algorithm to generate a single synthetic observation:

1. for the confidential observation, y_i, generate $\tau_i \sim U(0,1)$;
2. determine $\hat{\beta}(\tau_i) = \arg\min_\beta E[\rho_{\tau_i}(Y - \boldsymbol{X}\beta)]$;
3. generate synthetic observation, $y_i^* = \hat{q}_{\tau_i}(\boldsymbol{X}) = \boldsymbol{X}\hat{\beta}(\tau_i)$.

Repeating this procedure for each confidential observation gives us the complete synthetic data to be released.

We note that quantile regression lines may cross and researchers may be interested in avoiding such a phenomenon. This may occur when the model is misspecified or data are scarce in some regions of the predictor space. Various researchers have proposed methods to ensure proper ordering of the quantile lines (Koenker, 1984; Cole, 1988; Cole and Green, 1992). The restricted regression quantiles (RRQs) (He, 1997) method was applied by Geraci (2016) to impute missing data in a simulation study. However, the authors did not find substantial differences in their analyses, as compared to the standard quantile regression imputation. This is beyond the focus of this Chapter, so we do not discuss this further.

Quantile regression is robust to outliers in the dependent variable, as compared to least squares regression. The primary advantage of quantile regression is that no assumptions are made regarding the distribution of the error term. Moreover, the equivariance property of quantiles ensures that transformations can be made to the data before imputation if necessary and the quality of imputations will not be affected. This is especially helpful if the variable is within certain bounds and these need to be respected in the imputation procedure, or the variable is censored. However, the property may not be applicable to certain transformations, such as the expected value operator (Su et al., 2011). Quantile regression offers a natural way to model error terms which may have various distributional properties (such as skewness or kurtosis) conditional on the covariates - features that can not be easily captured by simple transformations. Moreover, quantile regression automatically adjusts for altering variability along the x-direction, making it a plausible choice of model for data with heteroscedasticity. Quantile regression methods have been developed for parametric, semi-parametric (Heagerty and Pepe, 1999), and non-parameteric (Yu and Jones, 1998; Koenker et al., 1994; Thompson et al., 2010) models, in addition to linear and non-linear models for discrete (Machado and Silva, 2005; Kordas, 2006; Lee and Neocleous, 2010) and survival (Koenker and Geling, 2001; Peng and Huang, 2008) data. Quantile regressions have also been used for robust small area estimation (Fabrizi et al., 2014a,b; Tzavidis et al., 2016).

Throughout this Chapter, we obtain our quantile regression parameter estimates by utilising the connection between minimising weighted absolute deviations and maximising a Laplace likelihood (Koenker and Machado, 1999; Yu and Moyeed, 2001; Geraci and Bottai, 2007, 2014). Built-in functions to run these models in software R can be found by installing the packages *quantreg* (Koenker, 2016) for the non-hierarchical model and *lqmm* (Geraci, 2014) for the random intercepts model. We note that there are several other proposed approaches to fitting quantile regressions, especially in the Bayesian framework. As the data likelihood for quantile regressions is unknown, the Laplace likelihood is popularly used as a working likelihood. However, it has been noted that the posterior inferences based on the method cannot be assumed to be valid (Yang et al., 2015; Kottas and Krnjajić, 2009; Tokdar et al., 2012). Alternative working likelihoods include the use of the Dirichlet process mixture model (Kottas and Krnjajić, 2009), semi-parameteric models on the quantile process (Reich and Smith, 2013; Tokdar et al., 2012) and empirical likelihood (Yang et al., 2012). We note that our construction of inferences using quantile regressions may not be appropriate due to our choice of the fitting method. This may affect the quality of our synthesised data. However, as this is not the focus of the current research, we consider the choice of fitting methods an interesting topic for future research and do not discuss it any further.

4.2.3 Flexible distributions

An alternative strategy is to generate the model residuals with the help of more flexible distributions that allow for different shapes of the data. The three specific methods we explore include: Tukey's g and h distribution (GH), the Generalised Lambda distribution (GL) and transformations through Fleishman's Power Polynomials (PP). Working with the likelihood for the three approaches is either extremely complicated or computationally intensive. Instead, we use certain alternatives developed in the literature to estimate the parameters for the distributions. We note that by using more flexible distributions, we are proposing the use of parametric methods, and the standard restrictions of applying defined parametric forms of distributions to data apply. Nevertheless, the three methods we have chosen here are known for their flexibility and we expect them to handle various distributional shapes in the data.

The MI procedure proposed by He and Raghunathan (2009) to accommodate for non-normal error terms involves two main steps. Firstly, the imputation model is fitted to the data using least squares estimation. Thereafter, the residuals are modelled using a flexible distribution, here, GH. We note that the approach splits the imputation, in our case, synthesis, procedure into two parts. The first is obtaining the fitted value for an observation of interest, and the second is the generation of a non normal residual for this observation. It is possible to deal with the two parts separately without affecting the inferential quality of the data in terms of the model coefficients. This is because,

least squares estimation for the model coefficients does not require assuming normality of the error term. Hence, the coefficient estimates that are used for generating synthetic data are unbiased, as long as the mean structure of the synthesis model is correctly specified. As such, to model errors distributed with, for instance a GH distribution, we do not have to model the data using the likelihood from a GH distribution, to obtain unbiased coefficient estimates. In this Chapter, we assume that the data keeper has taken appropriate measures to avoid bias in the point estimates themselves, for example in the presence of OVB, as illustrated in Chapter 3.

Equivalently, a maximum likelihood estimation (MLE) procedure with the assumption of normally distributed errors may be used to obtain the same coefficient estimates as least squares. MLE estimates for the slope coefficient are also known to be robust to the misspecification of the shape of the error distribution. When MI is used for missing data, the missing data are often generated from their posterior distribution assuming normality of the residuals. Under misspecification of the error term for imputation of missing data, it has been shown that the final coefficient estimates obtained are still unbiased and efficient (Schenker and Welsh, 1988; Schenker and Taylor, 1996).

To model hierarchical data with non normal errors, while we have a number of approaches mentioned in the literature, as discussed in Section 1.1, we concern ourselves more with replicating the data, rather than the inference of the parameters. Therefore, the approach proposed by He and Raghunathan (2009) to generate non normal errors is sufficient for our purpose, but this is implemented only in the missing data setting, for the GH distribution, and for non-hierarchical data. We extend the methodology proposed by He and Raghunathan (2009) to generate residuals: 1) for MI for partially synthetic data, 2) from the GL and PP distributions and, 3) for hierarchical data.

MI for missing data is only 'proper' (Chap. 4, Rubin (1987)) when the uncertainty around all parameter estimates involved in the imputation process is taken into account either through the use of sampling from posterior distributions or repeated bootstrap sampling of the observed data. By applying the method proposed by He and Raghunathan (2009) to partially synthetic data, we note that we do not need to synthesise with additional uncertainty around the parameter estimates, and therefore, skip their bootstrap step applied to the observed data. Secondly, the use of GL and PP have only been proposed in the literature for MI for univariate data. We use these distributions to synthesise the residuals instead. For non normal errors in hierarchical data, we mainly rely on multivariate data generation from various flexible, non-normal distributions. There are a number of methods to generate data from non-normal multivariate distributions, such as using the Pearson system (Nagahara, 2004) or copulas (Genest and MacKay, 1986). The computations applied in this Chapter are greatly facilitated by developments in the NORmal To Anything (NORTA) methodology (Cario and Nelson, 1997) and power method transformations (Vale and Maurelli, 1983).

Tukey's g-and-h Family (GH)

The use of GH (Tukey, 1997; Hoaglin, 1985; Jorge and Boris, 1984) for MI of skewed variables in the missing data scenario was proposed by He and Raghunathan (2006). If Z is a standard normal variable, the GH distributed variable R can be described with the following transformation of Z:

$$R_{gh}(Z) = \mu + \sigma \frac{e^{gZ} - 1}{g} e^{hZ^2/2}, \tag{4.4}$$

where, μ determines the location parameter, σ, the scale parameter and g and h are scalar values that control the skewness and elongation of R, respectively. For positive values of g, the distribution of R is right skewed, for negative, left and for $g = 0$, symmetric. When $g = 0$, positive values of h create positive elongation. The lognormal and normal families are special cases of the GH distribution. Estimates for the parameters $\{\mu, \sigma, g, h\}$ can be obtained by using the empirical quantiles of the observed data (Hoaglin, 1985). He (2005) showed that using the much more complicated maximisation of a GH likelihood may not result in significant efficiency gains, as compared to results from the quantile approach. Following the quantiles approach, the parameters of the GH distribution can be obtained from Y using the following result:

$$\begin{aligned} \hat{\mu} =& R_{0.5}, \\ \hat{g} =& -\frac{1}{Z_p} \log \frac{R_{1-p} - R_{0.5}}{R_{0.5} - R_p}, \end{aligned} \tag{4.5}$$

where, R_p and Z_p are the $100p^{th}$ quantiles of the GH and standard normal distributions. If $\hat{g}$ is positive, the estimates for σ and h are obtained by regressing:

$$\log \frac{g(R_{1-p} - R_{0.5})}{e^{-gZ_p} - 1} = \log \sigma + \frac{hZ_p^2}{2}, \tag{4.6}$$

for $p = \{0.005, 0.01, 0.025, 0.05, 0.10, 0.25\}$. If $\hat{g}$ is negative, the response variable is $\log \frac{g(R_{0.5} - R_p)}{1 - e^{gZ_p}}$ instead. Once the distribution parameters are estimated, imputed values can be generated by first generating random values for Z from the standard normal distribution and then transforming them using (4.4). In this Chapter, we are interested in non normal error terms instead. He and Raghunathan (2009) described the GH imputation process for the model error term for a simple linear regression. The imputation process involves bootstrapping the residuals to approximate the variability in the parameter estimates, as required by MI. For the variable to be imputed Y and covariates $\boldsymbol{X}$, their algorithm to generate imputations for missing Y involves the following steps:

1. fit the ordinary linear regression under the assumption of normality of the error term, $Y = \boldsymbol{X}\beta + \varepsilon$;

2. obtain a bootstrap sample for the model residuals, $R = Y - \boldsymbol{X}\hat{\beta}$;

3. utilising the sample in step 2, estimate the parameters for the GH distribution using the quantile approach;

4. generate standard normal deviates for each missing observation in Y, Z;

5. transform the Z using the parameter estimates in Step 3 and the transformation:

$$\varepsilon_{imp} = \sigma \frac{e^{gZ} - 1}{g} e^{hZ^2/2} - E_{GH}, \tag{4.7}$$

   where, $E_{GH} = \frac{1}{g\sqrt{1-h}}(e^{\frac{g^2}{2(1-h)}} - 1)$ is the mean of the standardised GH distribution $(\mu = 0; \sigma = 1)$;

6. generate $Y_{imp} = \boldsymbol{X}\hat{\beta} + \varepsilon_{imp}$.

Steps 2 to 6 can then be repeated M times to obtain M imputed copies for Y. We follow the above procedure for generating synthetic data with two changes. Firstly, we skip the bootstrap step. Secondly, we found the quantile approach (Hoaglin, 1985) for GH parameter estimation unstable. The results are sensitive to the choice of p and this becomes a more significant problem if sample sizes are small. Instead, we consider a different approach, method of percentiles (MOP), as laid out by Kuo and Headrick (2014). This method has closed form solutions for the $\{g, h\}$ parameters:

$$\begin{aligned} \hat{g} &= -\frac{\log \gamma_3}{z_{0.90}}, \\ \hat{h} &= \frac{2\log\left(\frac{\gamma_3^{1-z_{0.75}/z_{0.90}}(\gamma_3^{2z_{0.75}/z_{0.90}}-1)}{(\gamma_3^2-1)\gamma_4}\right)}{z_{0.90}^2 - z_{0.75}^2}, \end{aligned} \tag{4.8}$$

where, z_p is the standard normal percentile, $\gamma_3 = (\theta_{0.50} - \theta_{0.10})/(\theta_{0.90} - \theta_{0.50})$ and $\gamma_4 = (\theta_{0.75} - \theta_{0.25})/(\theta_{0.90} - \theta_{0.10})$, where θ_p is the empirical percentile for the data. Once, the parameters are determined, we generate a sample from the standard normal distribution and transform it according to equation (4.4).

We now turn to residuals for hierarchical data. Here, we only deal with a random intercepts (RE) model, $Y|\beta, b, \sigma^2, \sigma_b^2, \boldsymbol{X}, \boldsymbol{L} \propto N(\boldsymbol{X}\beta + \boldsymbol{L}b, \sigma^2\boldsymbol{I}_{nt})$, where, $\boldsymbol{X}$ is the design matrix for all 'fixed' covariates, $\boldsymbol{L}$ is the design matrix for all 'random' covariates, β is a p-dimensional vector, where p is the number of 'fixed' parameters to be estimated, and b is a vector of random effects, $b \sim N(0, \sigma_b^2\boldsymbol{I}_n)$ for n clusters. To generate synthetic data, we suggest fixing the linear mean component of the model $\boldsymbol{X}\hat{\beta}$ and generating the remaining errors from an appropriate distribution.

The model is first fitted assuming normal and independent errors (ε_b (between cluster) and ε_e (within cluster)) as per the usual RE model assumptions. In our setting it makes no difference whether the normality assumption is used for model fitting, as in the case

of a standard RE model, or the $\hat{\beta}$ estimates are obtained using generalised estimating equations (GEE) instead, through a marginal model with an exchangeable correlation structure. There are extensive discussions in the literature regarding the performance and robustness of RE models against marginal models (see Lee et al. (2004) and the references therein), but these mainly revolve around the validity of inferences. For synthetic data generation, we do not concern ourselves with inferential theory; the only important aspect of model fitting is the estimate of the model coefficients themselves. When the mean of the linear model is correctly defined, and errors have a compound symmetry structure, the two models will provide us the same $\hat{\beta}$ estimates. Our process only requires this $\hat{\beta}$ estimate from these models. We expect that in cases where the $\hat{\beta}$ estimates will differ across the RE and marginal models, as in the case of non-linear models or errors with a different structure, data keepers will choose to use the model that is best for the data in hand to obtain the $\hat{\beta}$ estimates. This is just a more specific example of a comment we made earlier; we expect the data keepers to use 'robust' estimation procedures wherever required but this is not the focus of this Chapter. Once the $\hat{\beta}$ are obtained, we propose arranging the model 'residuals' $\boldsymbol{\varepsilon} = Y - \boldsymbol{X}\hat{\beta}$ as an $n \times t$ matrix. As we expect the errors within clusters to be correlated, we consider generating synthetic residuals for hierarchical data, using a multivariate distribution fitted to the $n \times t$ matrix, $\boldsymbol{\varepsilon}$. For this we will require a multivariate distribution that preserves the correlations of the residuals across the t time points or repeated measures.

To generate errors from the multivariate GH, we follow the MOP procedure for multivariate data described by Kuo and Headrick (2014):

- specify the values for γ_3 and γ_4 for each variable separately;
- obtain the values for the GH parameters according to the MOP for each variable separately;
- specify the Spearman correlation matrix for the observed data;
- for each pair of variables, $\{k, l\}$, determine the intermediate correlations, ρ_{z_k,z_l}, using the observed correlations, ρ_{r_k,r_l} by numerically solving:

$$\rho_{r_k,r_l} = \left(\frac{6}{\pi}\right)\left[\left(\frac{n-2}{n+1}\right)\arcsin\left(\frac{\rho_{z_k,z_l}}{2}\right) + \left(\frac{1}{n+1}\right)\arcsin\left(\rho_{z_k,z_l}\right)\right] \tag{4.9}$$

- generate multivariate normal data, $\boldsymbol{Z}$, using the intermediate correlation matrix, ρ_{z_k,z_l};
- transform each component of $\boldsymbol{Z}$ by using equation (4.4) and the corresponding parameter estimates for that variable.

In both the non-hierarchical and hierarchical cases, we propose synthesising the standardised variables with mean 0 and variance 1, and transforming these back to the original scale after the synthesis.

Generalised Lambda Distribution (GL)

The GL distribution is a flexible class of distributions that can accommodate a wide range of skewness and kurtosis combinations (Ramberg and Schmeiser, 1972, 1974; Ramberg et al., 1979). The pdf and inverse cdf of the GL are known and this facilitates data generation by the inverse cdf method. If $p \sim U(0,1)$, then a GL distributed variable Y may be generated using the inverse cdf:

$$R = \lambda_1 + [p^{\lambda_3} - (1-p)^{\lambda_4}]/\lambda_2, \tag{4.10}$$

where, λ_1 and λ_2 are location and scale parameters and λ_3 and λ_4 define the shape of the distribution. The λ parameters may be determined by a number of different approaches, such as moment matching, least squares or numerical likelihood, but none of the approaches perform 'best' across the parameter space, and the support of the distribution, which is not fixed. The thesis by Dean (2013) provides a detailed discussion on the topic. Here, we follow the moment matching approach developed by Ramberg and Schmeiser (1974) and used by Demirtas (2009) for MI.

The mean (μ), variance (σ^2), third (μ_3) and fourth (μ_3) moments about the mean of the GL distribution are (Ramberg and Schmeiser, 1974):

$$\begin{aligned}
\mu &= \lambda_1 + \frac{A}{\lambda_2}, \\
\sigma^2 &= \frac{(B - A^2)}{\lambda_2^2}, \\
\mu_3 &= \frac{(C - 3AB + 2A^3)}{\lambda_2^3}, \\
\mu_4 &= \frac{(D - 4AC + 6A^2B - 3A^4)}{\lambda_2^4},
\end{aligned} \tag{4.11}$$

where,

$$\begin{aligned}
A &= \frac{1}{1+\lambda_3} - \frac{1}{1+\lambda_4}, \\
B &= \frac{1}{1+2\lambda_3} + \frac{1}{1+2\lambda_4} - 2\beta(1+\lambda_3, 1+\lambda_4), \\
C &= \frac{1}{1+3\lambda_3} - \frac{1}{1+3\lambda_4} - 3\beta(1+2\lambda_3, 1+\lambda_4) + 3\beta(1+\lambda_3, 1+2\lambda_4), \\
D &= \frac{1}{1+4\lambda_3} + \frac{1}{1+4\lambda_4} - 4\beta(1+3\lambda_3, 1+\lambda_4) - 4\beta(1+\lambda_3, 1+3\lambda_4) \\
&\quad + 6\beta(1+2\lambda_3, 1+2\lambda_4),
\end{aligned}$$

where, $\beta(a, b)$ represents the β-function with arguments a and b. The skewness, $\alpha_3 = \frac{\mu_3}{\sigma^3}$ and kurtosis (kurtosis proper, not kurtosis excess), $\alpha_4 = \frac{\mu_4}{\sigma^4}$ are functions of λ_3 and λ_4 only. This implies that one can match the empirical skewness and kurtosis to α_3 and α_4, and solve the two non linear simultaneous equations to find $\hat{\lambda}_3$ and $\hat{\lambda}_4$ which best fit the data. This requires the use of an appropriate root finding or optimisation routine, such as the Newton-Raphson. Thereafter, the values of λ_1 and λ_2 can be determined using the above system of equations.

Given the method of moments described to determine appropriate parameter estimates for the GL distribution, one can follow the MI algorithm described by Demirtas (2009) to impute missing data:

1. center and scale the data so that the mean is 0 and variance 1;
2. draw a non parametric bootstrap sample from the observed data;
3. estimate the GL parameters by the method of moments;
4. simulate independent $u_i \sim U(0, 1)$ variates for the missing observations;
5. transform the u_i using equation (4.10);
6. back transform the data to the original scale.

Steps 1-6 may be repeated M times to obtain M imputed copies of univariate data. Again, we utilise the above procedure to synthesise the error terms but without the bootstrapping step, as in the case of the GH distribution.

Next, we consider the synthesis of model residuals for the random intercepts model. First, we look at the computations involved to generate data from a multivariate GL distribution. Headrick and Mugdadi (2006) described a NORTA type approach to achieve this. The idea is to generate data from a multivariate normal distribution and a working correlation structure, called the intermediate correlation, and transform it into the multivariate GL. Consider the bivariate case, and let Z_k and Z_l have standard normal univariate and bivariate pdfs, and cdfs, defined as:

$$\begin{aligned}
f_k =& (2\pi)^{-1/2} \exp\left\{-z_k^2/2\right\}, \\
f_l =& (2\pi)^{-1/2} \exp\left\{-z_l^2/2\right\}, \\
f_{k,l} =& (2\pi\sqrt{1-\rho_{z_k,z_l}^2})^{-1} \exp\left\{-(2\sqrt{1-\rho_{z_k,z_l}^2})^{-1} \times (z_k^2 - 2\rho_{z_k,z_l} z_k z_l + z_l^2)\right\}, \\
\Phi(z_k) =& \int_{-\infty}^{z_k} (2\pi)^{-1/2} \exp\left\{-u_k^2/2\right\} du_k, \\
\Phi(z_l) =& \int_{-\infty}^{z_l} (2\pi)^{-1/2} \exp\left\{-u_l^2/2\right\} du_l,
\end{aligned} \tag{4.12}$$

where, $\Phi(z_k) \sim U_k(0,1)$ and $\Phi(z_l) \sim U_l(0,1)$ with correlation, $\rho_{\Phi(z_k),\Phi(z_l)} = (6/\pi)$ $\arcsin(\rho_{z_k,z_l}/2)$ (Pearson, 1907). Let $r_k(z_k, \lambda_{iw}) = \lambda_{k1} + [(\Phi(z_k))^{\lambda_{k3}} - (1-\Phi(z_k))^{\lambda_{k4}}]/\lambda_{k2}$ and $r_l(z_l, \lambda_{lw}) = \lambda_{l1} + [(\Phi(z_l))^{\lambda_{l3}} - (1-\Phi(z_l))^{\lambda_{l4}}]/\lambda_{l2}$ where $w = \{1,2,3,4\}$ be standardised GL distributions that have the form of (4.10). The correlation between $r_k(z_k, \lambda_{kw})$ and $r_l(z_l, \lambda_{lw})$ can then be expressed as:

$$\rho_{r_k,r_l} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} r_k(z_k, \lambda_{kw}) r_l(z_l, \lambda_{lw}) f_{k,l} dz_k dz_l. \tag{4.13}$$

The ρ_{z_k,z_l} term within $f_{k,l}$ is the intermediate correlation between the standard normal deviates that can be determined solving (4.13) numerically for a given correlation, ρ_{r_k,r_l} between the desired variables. Therefore to generate random observations from a multivariate GL distribution:

1. solve the system of equations (4.11) to compute the parameter estimates for each univariate GL distribution separately,
2. numerically determine the intermediate correlations for each pair of variables, using (4.13),
3. generate multivariate standard normal deviates with the intermediate correlation structure,
4. convert the standard normal deviates to uniform deviates, and
5. use the uniform deviates, the parameter estimates for the GL distributions and the transformation (4.10) to generate multivariate GL deviates with the desired correlation.

As in the case of the GH distribution, we conduct the synthesis of non normal error terms for both non-hierarchical and hierarchical data with the univariate and multivariate GL distributions.

Fleishman's Power Polynomials (PP)

For the MI of non normal missing data, Demirtas and Hedeker (2008) proposed the use of Fleishman's Power Polynomials, PP (Fleishman, 1978). Consider R, a standardised variable with 0 mean and variance equal to 1. With an emphasis on the first four moments of a distribution, Fleishman provided a polynomial transformation of a standard normal variable, Z to describe R:

$$R = a + bZ + cZ^2 + dZ^3. \tag{4.14}$$

Given empirical skewness, $v_1 = E(Y^3)$, and kurtosis, $v_2 = E(Y^4) - 3$ (kurtosis excess, not kurtosis proper), Fleishman showed that the constants $\{a, b, c, d\}$ can be determined

by solving the following system of equations by a suitable root finding method or optimisation routine:

$$\begin{aligned} a &= -c, \\ b^2 + 6bd + 2c^2 + 15d^2 - 1 &= 0, \\ 2c(b^2 + 24bd + 105d^2 + 2) - v_1 &= 0, \\ 24[bd + c^2(1 + b^2 + 28bd) + d^2(12 + 48bd + 141c^2 + 225d^2)] - v_2 &= 0. \end{aligned} \tag{4.15}$$

The PP system covers a large area in the skewness-elongation plane; restrictions on these can be found in Headrick and Sawilowsky (2000). The pdf and cdf of PP in general form were provided by (Headrick and Kowalchuk, 2007). Given standardised cumulants, the power method transformation can produce shapes for the normal, logistic or uniform distributions Headrick (2010). Using the PP, Demirtas and Hedeker (2008) proposed the following algorithm for MI of univariate missing data:

1. center and scale the data so that the mean is 0 and variance 1;
2. draw a non parametric bootstrap sample from the observed data;
3. using the empirical moments, solve the system of equations (4.15) to find the values of $\{a, b, c, d\}$;
4. generate independent standard normal variates;
5. transform the standard normal variates using (4.14);
6. back-transform the data to the original scale.

To obtain M copies of imputed data, steps 2-6 can be repeated M times. As in the cases of the GH and GL distributions, we omit the bootstrapping step and follow the procedure to synthesise the residuals of the regression model used for synthesis.

Considering our simulation requirements for hierarchical data, we explore data generation for the multivariate PP. Examples of data generation with the multivariate PP can be found in Headrick and Sawilowsky (1999) and Demirtas et al. (2012). Consider the bivariate case, where Z_k and Z_l are standard normal variables. Let the vectors $w_k = \{a_k, b_k, c_k, d_k\}$ and $w_l = \{a_l, b_l, c_l d_l\}$ contain the parameter estimates for the univariate PP for each of the two variables R_k and R_l, v_{R_k,R_l} be the correlation between R_k and R_l and v_{Z_k,Z_l} be the correlation between Z_k and Z_l. As R_k and R_l are centered and scaled, $E(R_k) = E(R_l) = 0$ and so, $v_{R_k,R_l} = E(R_k R_l) = E(w_k^{'} z_k z_l^{'} w_l) = w_k^{'} G w_l$,

where G is the expectation of the matrix product of $z_k z_l^{'}$:

$$G = E(z_k z_l^{'}) = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & v_{Z_k,Z_l} & 0 & 3v_{Z_k,Z_l} \\ 1 & 0 & 2v^2_{Z_k,Z_l}+1 & 0 \\ 0 & 3v_{Z_k,Z_l} & 0 & 6v^3_{Z_k,Z_l}+9v_{Z_k,Z_l} \end{pmatrix}.$$

This relationship can be algebraically manipulated to give:

$$v_{R_k,R_l} = v_{Z_k,Z_l}(b_1b_2 + 3b_1d_2 + 3d_1b_2 + 9d_1d_2) + v^2_{Z_k,Z_l}(2c_1c_2) + v^3_{Z_k,Z_l}(6d_1d_2), \quad (4.16)$$

which can be solved numerically between for any pair of variables $\{R_k, R_l\}$ and their correlation v_{R_k,R_l} to obtain the intermediate correlation v_{Z_i,Z_j}. The estimated v_{Z_i,Z_j} can then be used to generate deviates for a pair of standard normal variables, which can then be transformed to the desired PP form using (4.14). We use this algorithm to generate the multivariate residuals for hierarchical data using the PP transformation.

Next, we study the performance of the methods described in this section through simulation studies in Section 4.3.

4.3 Simulation studies

We set up two simulation studies to compare the performance of the various methods described in Section 4.2 for generating synthetic data in the presence of unusual observations: one for a non-hierarchical data scenario and one for the hierarchical data. We are interested in exploring:

- the consequences of using the normal distribution for generating synthetic residuals for non normal data;
- the extent to which transformations can improve upon the quality of the synthetic data over the normal distribution approach;
- the quality of synthetic data generated from the quantile approach;
- the flexibility of the various distributions presented in Section 4.2.3 in accommodating the non normal distributions;
- the synthesis models congenial to quantile regression as the analysis model;
- the relative disclosure risks associated with each of the synthesis models above.

Sections 4.3.1 and 4.3.2 present the simulation design and results for the two studies.

4.3.1 Non-hierarchical study

For non-hierarchical data, we generate 500 datasets with $n = 1000$ observations for a sensitive variable, Y dependent on covariates $\boldsymbol{X} = (1, X)$. Taking guidance from He and Raghunathan (2009), we note that if the synthesis model provides good prediction for Y, i.e. $\frac{var(\varepsilon)}{var(X)}$ is small, specification of the error term distribution will not impact the quality of the synthesis greatly, as the distribution of Y is largely determined by X. On the other hand, if the model is more imprecise, such that the error variance dominates the variance of X, the shape of newly generated Y, Y_{syn} will depend more on the shape of the residuals. We have chosen error variances such that the residuals do impact the shape of the distribution of Y_{syn}. This is not unrealistic, given that most linear regressions in Social Sciences report a low value of $R^2 = \frac{\text{variation explained by the model}}{\text{total variation in the response variable}}$, often 30% or less, indicating that errors form a vital component of predicted values. Data are generated under four different mechanisms:

1. Case 1: $Y = \boldsymbol{X}\beta + \varepsilon$, where, $X \sim N(4, 5), \beta = \{1, 2\}$ and $\varepsilon \sim N(0, 7)$;
2. Case 2: $Y = \boldsymbol{X}\beta + \varepsilon$, where, $X \sim N(4, 5), \beta = \{1, 2\}$ and $\varepsilon \sim LN(0, 1.39)$;
3. Case 3: $Y = \boldsymbol{X}\beta + \varepsilon$, where, $X \sim N(4, 5), \beta = \{1, 2\}$ and $\varepsilon \sim t_{1.46}$;
4. Case 4: $Y = \boldsymbol{X}\beta + \varepsilon$, where, $X \sim N(4, 5), \beta = \{1, 2\}$ and $\varepsilon \sim U(-12, 12)$.

Figure 4.1 displays the probability density functions for each of the four error distributions, Cases 1 - 4. Case 1 is the standard case, where data do not contain outliers and the residuals do satisfy the assumption of normality. In Cases 2 and 3, the error term is generated from either a lognormal distribution or a heavy tailed t distribution, i.e. outliers in either one or both tails of the distribution, respectively. Our synthesis methods have been deliberately chosen such that they may show some flexibility when fit to data with unusual and extreme values, i.e. the error distribution has long tails. However it is worth studying whether all the methods perform well if the residuals in fact have a more flat distribution than the normal, and hence, we consider data in Case 4, where the errors are generated using a uniform distribution.

We then synthesis Y using the following methods:

1. Norm: $Y_{syn} = \boldsymbol{X}\hat{\beta} + v$, where $v \sim N(0, \hat{\sigma}_e^2)$, $\hat{\beta}$ and $\hat{\sigma}_e^2$ are the least squares estimates;
2. Trans: same as Norm, with the transformed response variable, $Y_{syn}^T = \boldsymbol{X}\hat{\beta} + v$, where Y^T is obtained by transforming Y using a Box-Cox transformation, Y_{syn}^T is transformed back to obtain Y_{syn};
3. Quant: Y_{syn} is obtained by sampling from the various quantile regressions, as described in Section 4.2.2;

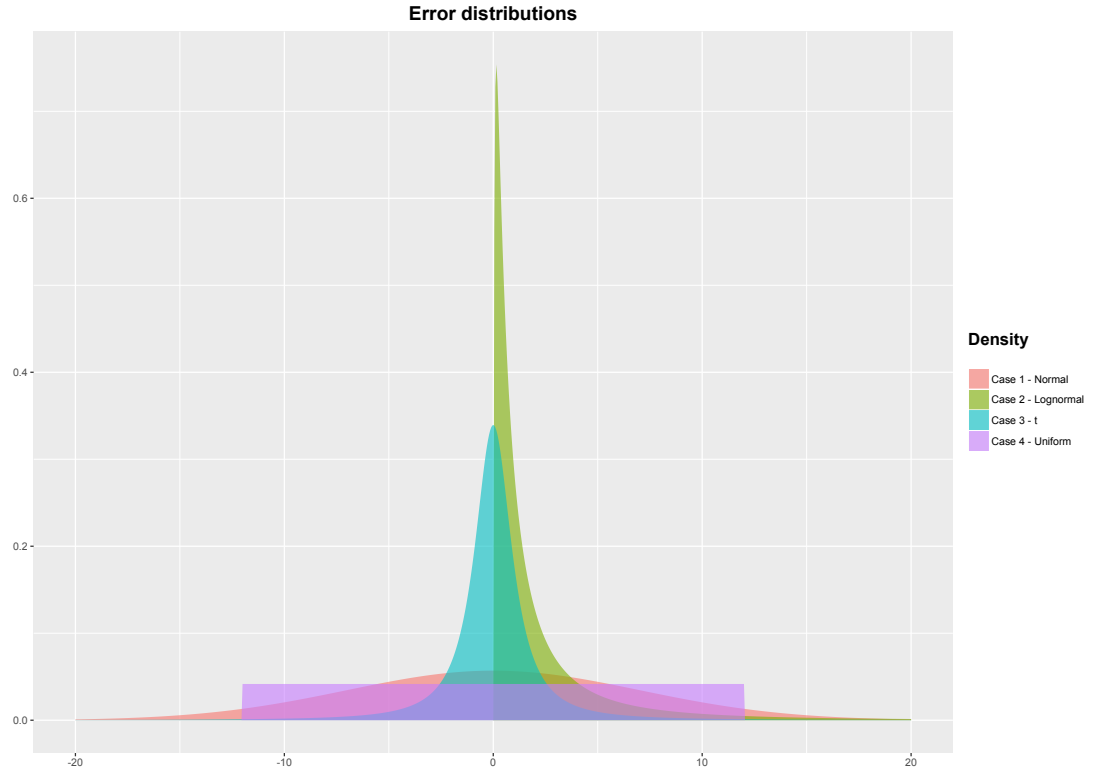

Figure 4.1: Cases 1-4, error distributions.

4. GH: $Y_{syn} = \boldsymbol{X}\hat{\beta} + v$, where $v \sim GH(0, 1, \hat{g}, \hat{h})$, $\hat{\beta}$ is obtained using least squares;
5. GL: $Y_{syn} = \boldsymbol{X}\hat{\beta} + v$, where $v \sim GL(\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3, \hat{\lambda}_4)$, $\hat{\beta}$ is obtained using least squares;
6. PP: $Y_{syn} = \boldsymbol{X}\hat{\beta} + v$, where $v \sim PP(\hat{a}, \hat{b}, \hat{c}, \hat{d})$, $\hat{\beta}$ is obtained using least squares.

For each method, we create $M = 10$ copies for Y_{syn}, and combine analyses over the 10 copies using the partial synthetic data rules in Section 1.1.1. The estimands of interest are:

1. the mean and variance, and sample quantiles of Y at levels $\{0.25, 0.50, 0.75\}$;
2. the estimate, standard errors, length and coverage of confidence intervals, CI, for the coefficients of quantile regressions of Y on $\boldsymbol{X}$, $\{\beta_0, \beta_1\}$ at selected levels $\tau = \{0.05, 0.50, 0.95\}$.

There are number of ways to estimate the standard error of coefficient estimates for quantile regression, and this is not the focus of our research. In our simulation, we choose the bootstrap method, as recommended by Koenker and Hallock (2001). We combine our bootstrap standard errors from the 10 copies using the standard combining rules. As the quantile regression estimator is asymptotically normally distributed, the MI combining rules are appropriate to use for point estimates. As for standard errors,

we note that there is some very recent discussion in the literature with regards to the combination of bootstrapping and MI (Schomaker and Heumann, 2016; Bartlett, 2016), and the most appropriate method is not yet fully established.

In a simulation study to evaluate synthetic data, it is useful if the 'correct' model for the data are known to compare results against such a model. While we use the lognormal, t and uniform distributions to generate the data, we note that the purpose of using these distributions was only to insert unusual observations in the data and not study data synthesis under these distributions. In real data applications, the distribution of the error term is unknown, and we treat our simulated data as such. This means that in this study, there is no gold standard to use for our evaluations. Therefore, to make a judgement regarding the utility of our synthetic data, we fit our analysis models to the original Y and use the coefficient and standard error estimates as the 'true' values, and compare our results against these. The length of CI are reported relative to the length of the CI from the original data, and the coverage of CI are computed assuming that the original data estimates averaged over 500 replications are the true values.

The simulations are conducted in R (R Core Team, 2016). Particular packages used include *nleqslv* (Hasselman, 2016) for solving non-linear equations to fit the GL distribution, *PoisNonNor* (Shi and Demirtas, 2015) to fit the PP to the data and *quantreg* (Koenker, 2016) for quantile regressions imputations and analyses.

Data utility

Figure 4.2 displays the graphs for Y against X for one chosen data set. The different coloured points indicate the values of Y_{syn} generated using the various synthesis models under evaluation. We observe that the Y_{syn} are contained within adequate range of Y for Case 1 for all the methods. In Cases 2 and 3, some of the strategies, such as GH may result in some extremely large or small values. In Case 4, all observations seem to be within reasonable range of Y. This gives us a rough idea of the shape of Y_{syn} for the various methods, but only for one data set.

We summarise the sample quantiles, mean and median averaged over the 500 datasets in Table 4.1. We find that all methods preserve the moments and quantiles of the distribution under consideration quite well for Case 1; this provides a validation for the adequacy of all the synthesis procedures for the simulation study. Considering Case 2, we see that the original Y quantiles indicate the presence of a right tail in the data. Again for most methods both the left and right tail quantile estimates are preserved very well in the synthetic data. The worst performing method in this scenario is the GH approach. We find that the shape of Y_{syn} for GH has a longer left tail than in the original data, but a shorter right tail. The median estimate is also nearly half of the original median estimate, although the mean is still preserved. We also find that the GH distribution is prone to generating a few extreme values during data generation. This is indicated in the variance estimate for Y_{syn}, which is almost 20 times larger than the

variance of the original Y. In practice, a few big values such as these can be detected and the synthesis process repeated to generate more plausible values. However, we have not undertaken any intervention in our simulation studies. This means that the quality of the GH synthesis will suffer from these few but very extreme values. Other observations include the slightly underestimated variability for the Trans approach with a variance estimate of ~ 137 rather than ~ 144. Sample quantile for the PP approach are well-preserved but the variability in Y is increased (~ 185).

In Case 3, the data have outliers in both tails of the distribution. From Table 4.1, we observe that only the Quant approach perfectly replicates the quantiles of the data, albeit with a variance estimate larger than that observed for Y (~ 916, as opposed to 827). Using Trans to synthesise data reduces the variability in Y considerably (~ 295, as opposed to 827). Although the mean and median are preserved well, the 5^{th} and 95^{th} quantile estimates are further out from the centre. We make a similar observation for the GL approach, but the variance of synthetic Y is much closer to the original variance estimate than the Trans approach. As expected, using the Norm approach results in slightly fatter tails, and the overall variability in Y is reduced slightly. Finally, both the GH and PP approaches result in a few very extreme observations, that cause the variance estimate to inflate; their quantile estimates are also pushed further out than that observed for the original data. In Case 4, we find that most of the methods handle the flat distribution of errors fairly well, the only exception being GL. We find that the summary statistics for the GL distribution are biased. We also observe that the variability in Y is lower when using the GH distribution in such a scenario.

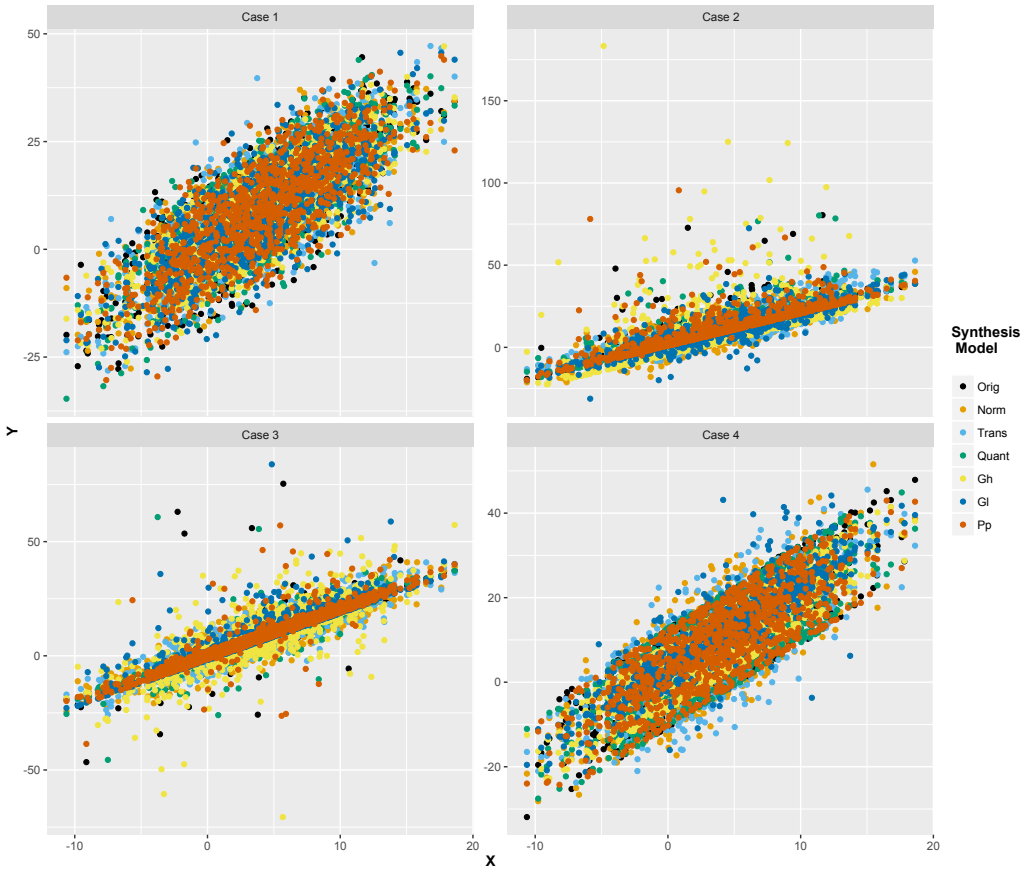

Figure 4.2: Plot of Y and Y_{syn} against X for one data set in each of the Cases 1-4.

Synthesis Model	Q.25	Q.50	Mean	Q.75	Var
Case 1					
ORIG	0.79	9.01	9.00	17.22	148.65
Norm	0.79	9.00	9.00	17.21	148.79
Trans	0.78	8.99	9.02	17.20	147.72
Quant	0.79	9.01	9.00	17.20	148.66
GH	0.77	9.00	9.00	17.23	150.50
GL	0.74	9.03	9.00	17.29	148.66
PP	0.79	9.01	9.00	17.22	148.69
Case 2					
ORIG	4.19	11.18	11.62	18.35	144.88
Norm	3.69	11.62	11.62	19.55	144.67
Trans	3.28	10.54	11.62	18.79	136.76
Quant	4.19	11.19	11.63	18.37	146.73
GH	-1.60	6.62	11.62	16.34	3161.79
GL	4.02	11.45	11.62	18.96	144.21
PP	3.71	10.99	11.62	18.56	185.28
Case 3					
ORIG	1.99	9.01	8.94	16.01	826.71
Norm	-1.45	8.95	8.95	19.35	825.85
Trans	0.37	8.98	8.98	17.60	294.91
Quant	1.99	9.01	8.93	16.02	915.72
GH	-2.78	8.93	8.95	20.65	5670.20
GL	0.31	9.09	8.94	17.76	859.31
PP	-0.63	9.40	8.96	19.13	2417.17
Case 4					
ORIG	0.67	9.02	9.00	17.33	147.22
Norm	0.84	9.00	9.00	17.17	147.20
Trans	0.78	8.94	9.02	17.15	146.46
Quant	0.70	9.00	9.00	17.31	147.28
GH	1.47	9.01	9.00	16.53	123.58
GL	4.01	11.45	11.62	18.95	144.62
PP	0.69	9.00	9.00	17.31	147.25

Table 4.1: Summary of Y compared to Y_{syn} over 500 datasets, for Cases 1-4 for the Norm, Trans, Quant, GH, GL and PP synthesis strategies. Sample size is 1000.

We now consider the intercept and slope estimates for each of our three quantile regressions, along with their standard errors, and confidence intervals. Table 4.2 displays the results computed over 500 datasets for Case 1. We find that all the methods preserve the estimate for β_1 for all three quantile regression analyses. However, we expect that the way the residuals are generated will impact the intercept and standard error estimates for the coefficients. In Table 4.2, we observe that most of the methods employed preserve the intercept estimates from the original data. The standard error estimates are also slightly higher, as we would expect to see from synthetic data, but close to the original standard errors, and the length of CI relative to the length of original data lengths are also only slightly larger. Most of the coverages for the 95% CI tend to be close to 1. As the construction of CI for quantile regression coefficients after MI is still a new area of research, we expect that future research may reveal more efficient CI combining rules for bootstrapped standard errors. The only exception in the table is the GL approach. In this case, we find that the standard errors are underestimated, resulting in very short CI that have considerably less than nominal coverage.

Results for Case 2 are displayed in Table 4.3. We first discuss β_1. We observe that while all the methods preserve the β_1 estimates and standard errors from the original data, the Trans approach results in biases at the two extreme quantiles. The transformation of the response variable results in altering the slope estimates, displaying a fanning effect, such that the slope for the regression at 5^{th} quantile is reduced, and the slope at the 95^{th} end is increased. Although the standard error estimates are close to those observed for the original data, this bias affects the coverage of the CI, reducing it to as low as 20%. We find that the CI for β_1 are nearly four times as long as the original CI for the median regression when the Norm, Trans or GH approaches are used and about 2.5 times longer for the GL and PP approaches. The most efficient inferences are obtained when the Quant approach is used. Again, overall coverages remain high, with the exception of those for the GL and PP approaches. At the more extreme quantiles, the relative length and coverage of CI are slightly more dramatic. This is mainly because the original CI are extremely short, with lengths averaging at very small values. We find that the CI for β_1 in the 5^{th} quantile regression are about 40 times the length of the original CI for the Norm, Trans and GL approaches. The GH approach results in about 10 times, while the Quant CI are consistently about 1.06 times long. An exception to these is the PP approach, the CI for which are far too short which impacts the coverages as well. For the 95^{th} regression line, the β_1 CI are about 5 times long for the GH approach, and about 1.6 times long for the PP method. All other methods, with the exception of Quant, result in CI which are slightly short, again impacting the coverages of the CI. As expected, we observe that the synthesis of the residuals have a direct impact on the standard errors and confidence intervals for the coefficient estimates.

For the intercept, we note that not all approaches are able to replicate the intercepts from the original data. The only method that almost perfectly reproduces the intercepts,

along with the standard errors, is the Quant approach. Its CI continue to cover the original estimate 100% of the time as before, for which we expect that more efficient inferential procedures will be developed in the future. The CI themselves are about 1.08 times longer than the original CI, which is expected from synthetic data. This, however, is far more efficient than any of the other approaches perform. The intercept estimate for the Norm and Trans approaches are biased at both the extreme quantiles, however, less severely so for the 95^{th} quantile regression. The standard error estimates for both the approaches are overestimated at the 5^{th} and underestimated at the 95^{th} quantile line. This is a direct consequence of the shape of the error distribution which has a very short tail at the left end but large positive outliers. With inappropriate estimates and standard errors, coverages of the true values remain low. The use of GH, GL and PP do not necessarily improve on the performance of Norm and Trans for β_0. We find that the GL and PP approaches perform reasonably for the 95^{th} regression line, but either the intercept or the standard error estimates, or both, remain biased at the 5^{th} quantile line.

We now consider Case 3. Table 4.4 shows the results when the outliers lie at both tails of the distribution of Y. We observe that the β_1 estimates are generally preserved across all synthesis methods and the three analysis models. However, there is some variation between the different approaches with regards to the standard error estimates for β_1. We find that the Trans approach has slightly smaller standard error estimates as compared to the Norm approach. This in turn shortens the lengths of the CI and impacts the coverage of the true value of β_1 for the quantile regressions on the edges, dropping them to about 70%. As in the other scenarios, the Quant approach almost exactly matches the results from the original data. The CI are acceptably slightly longer than the original CI, but coverages, as before are 100%. Finally, we discuss the performance of the various parametric families used to model the residuals. We find that the standard error estimates for β_1 for the GH approach are higher than any other approach considered in the simulation study, especially for the 5% and 95% regression quantiles, and this impacts the lengths of the CI, which are observed to be about 7 times longer than those for the original data. Standard errors and CI for the GL and PP approaches are comparatively shorter, but despite reporting nearly 3 times long CI, coverages tend to be less than 85%.

In the case of β_0, there are considerable differences between all the synthesis procedures. The Quant procedure replicates the original analysis very well. At the median quantile regression, most other models also preserve the intercept estimate, although some biases are observed for GL and PP. At the more extreme quantiles, the intercept estimates are directly affected by the ability of the procedures to replicate the tail behaviour of Y. We find that with intercept estimates at $\{-15.79, 17.77\}$, the Norm approach is far from the original intercepts estimates $\{-2.85, 4.84\}$. While Trans improves upon Norm, the standard error estimates are still too small to accommodate the original estimates in the

CI. The PP approach produces estimates close to those of Trans and higher standard error estimates imply that the true value is also covered about 40% of the time in the CI. Following behind is the performance from the GL and GH approaches which also report estimates much further away from the original intercepts. Overall, coverages for the intercepts at the more extreme quantiles tend to be nearly 0, unless the PP or Quant methods are employed.

In Case 4, we judge the performance of our methods when Y does not contain outliers, but in fact has a flat distribution. We find that all methods result in the correct β_1 estimates for the median regression line with appropriate coverages, with the exception of GL, for which the standard error estimates are too low. At the more extreme quantile regressions, the point estimates are generally preserved but the standard error estimates for GH are too small. We now consider β_0 for the median regression. For the Norm, Trans and GH approaches, point estimates are preserved, and although standard errors and CI lengths are smaller than those observed for the original data, CI coverages remain close to 100%. With slightly higher standard error estimates, Quant provides similar results. PP also results in appropriate estimates. GL is the only method for which the intercept estimates are biased for the median regression. At the more extreme quantiles, again, we find that Quant performs best, while estimates from GH and GL are very different from the original estimates at either one or both ends of the quantiles. The intercept estimates from Norm and Trans are close to the original ones, and coverages remain high, although at the price of nearly 3 times as long CI. Results for PP are also close, and the standard error estimates and lengths of CI are much shorter, which cover the true value nearly 80% of the time.

We note that most of our proposed synthesis approaches provide unbiased results for the estimate of β_1 across all error distributions and the various quantile regressions considered. The most noticeable discrepancy was the biased point estimates for the Trans approach in Case 2 for the more extreme quantile regressions. Slope estimates tend to be the main focus for data imputation and synthesis, and therefore, we show that Box-Cox transformations do not necessarily preserve the correct estimates when data contain outliers in one direction. The intercept estimates are generally not very well preserved for data with non normal errors, but the Quant approach performed well.

This simulation serves to look beyond the mean structure of the model into residuals as an important part of the synthesis procedure. We show that the shape of the residuals matters for standard error estimates for the coefficients. Our simulations show that the quantile regression approach is resistant to the varying shapes of the error term as compared to the usual simple linear regression or the use of a transformed response variable. We also tested some more flexible parametric approaches, but in our attempt to accommodate the shape of the residuals, we found that often the standard errors for the intercept or slope estimates were unnecessarily inflated or deflated. This has a direct impact on analysts hoping to derive inferences made from the synthetic data. We also

Synthesis Model		Analysis Model					
		$\tau = 0.05$		$\tau = 0.50$		$\tau = 0.95$	
		β_0	β_1	β_0	β_1	β_0	β_1
ORIG	Est	-10.51	2.00	1.02	2.00	12.48	2.00
	SE	0.63	0.10	0.36	0.06	0.61	0.10
Norm	Est	-10.49	2.00	1.00	2.00	12.49	2.00
	SE	0.66	0.10	0.38	0.06	0.66	0.10
	Len	1.05	1.05	1.07	1.06	1.10	1.08
	Cov	0.99	0.97	1.00	1.00	0.99	0.97
Trans	Est	-10.39	1.98	1.03	1.99	12.62	1.98
	SE	0.65	0.10	0.38	0.06	0.68	0.11
	Len	1.04	1.06	1.06	1.06	1.14	1.11
	Cov	0.99	0.98	1.00	1.00	0.99	0.99
Quant	Est	-10.51	2.00	1.01	2.00	12.48	2.00
	SE	0.67	0.10	0.38	0.06	0.65	0.10
	Len	1.07	1.05	1.06	1.05	1.08	1.05
	Cov	1.00	1.00	1.00	1.00	1.00	1.00
GH	Est	-10.57	2.00	0.99	2.00	12.57	2.00
	SE	0.68	0.11	0.38	0.06	0.68	0.11
	Len	1.08	1.08	1.07	1.06	1.13	1.10
	Cov	0.82	0.96	1.00	1.00	0.82	0.96
GL	Est	-10.09	2.00	1.29	2.00	11.46	2.00
	SE	0.41	0.07	0.48	0.08	0.39	0.06
	Len	0.66	0.66	1.35	1.35	0.65	0.64
	Cov	0.57	0.63	0.96	1.00	0.57	0.62
PP	Est	-10.46	2.00	1.01	2.00	12.45	2.00
	SE	0.64	0.10	0.39	0.06	0.64	0.10
	Len	1.03	1.03	1.08	1.08	1.06	1.03
	Cov	0.97	0.96	1.00	1.00	0.96	0.96

Table 4.2: Properties of $\hat{\beta}_0$ and $\hat{\beta}_1$ over 500 datasets, for quantile regressions with $\tau = \{0.05, 0.50, 0.95\}$, Case 1. Sample size is 1000.

report some very high coverages for our regression coefficients of interest, but believe that this is more a consequence of using bootstrapped standard errors for quantile regression, and not the application of MI combining rules.

We appreciate that it is extremely demanding to attempt to preserve the intercepts at various quantiles of the conditional distribution of Y. Nevertheless, as statistical analyses become finer and more sophisticated, the demands from synthetic data also become more challenging. For an analyst wishing to use quantile regressions at the extreme end of a conditional distribution, our simulations show, that amongst the approaches considered,

Synthesis Model		Analysis Model					
		$\tau = 0.05$		$\tau = 0.50$		$\tau = 0.95$	
		β_0	β_1	β_0	β_1	β_0	β_1
ORIG	Est	1.10	2.00	2.01	2.00	10.85	2.01
	SE	0.01	0.00	0.07	0.01	1.20	0.19
Norm	Est	-6.29	2.00	3.61	2.00	13.51	2.00
	SE	0.57	0.09	0.33	0.05	0.57	0.09
	Len	44.92	45.01	4.62	4.64	0.48	0.49
	Cov	0.00	1.00	0.00	0.98	0.27	0.73
Trans	Est	-3.52	1.78	3.24	2.04	11.32	2.29
	SE	0.40	0.08	0.24	0.05	0.45	0.07
	Len	31.29	41.34	3.40	4.15	0.38	0.39
	Cov	0.00	0.20	0.00	0.95	0.51	0.20
Quant	Est	1.10	2.00	2.01	2.00	10.95	2.00
	SE	0.01	0.00	0.08	0.01	1.31	0.19
	Len	1.08	1.06	1.07	1.05	1.11	1.05
	Cov	1.00	1.00	1.00	1.00	1.00	1.00
GH	Est	-8.32	2.00	-3.87	2.00	35.27	1.99
	SE	0.13	0.02	0.33	0.05	6.03	0.91
	Len	10.00	10.10	4.63	4.61	5.12	4.97
	Cov	0.00	0.55	0.00	0.99	0.00	1.00
GL	Est	-3.53	2.00	3.18	2.00	11.92	2.00
	SE	0.62	0.10	0.18	0.03	0.97	0.15
	Len	49.28	48.69	2.46	2.47	0.82	0.81
	Cov	0.00	0.95	0.01	0.83	0.63	0.88
PP	Est	-0.32	2.00	1.85	2.00	13.39	2.00
	SE	0.01	0.00	0.19	0.03	2.02	0.29
	Len	0.69	0.69	2.63	2.65	1.72	1.58
	Cov	0.01	0.04	0.69	0.79	0.82	1.00

Table 4.3: Properties of $\hat{\beta}_0$ and $\hat{\beta}_1$ over 500 datasets, for quantile regressions with $\tau = \{0.05, 0.50, 0.95\}$, Case 2. Sample size is 1000.

only a quantile regression synthesis appears to be congenial to a quantile regression analysis. We expect this would come with its fair share of disclosure risks. As data keepers interested in privacy, we do not wish that our synthetic data resemble the real data enough for intruders to identify the units. Measuring disclosure risks, however, is still a very challenging area of research, with very little philosophy that is universally accepted or applied. Below, we utilise the risk measurement framework used in Chapter 3 to study the risk profile of the methods applied in our simulation study.

Synthesis Model		Analysis Model					
		$\tau = 0.05$		$\tau = 0.50$		$\tau = 0.95$	
		β_0	β_1	β_0	β_1	β_0	β_1
ORIG	Est	-2.85	2.00	1.00	2.00	4.84	2.00
	SE	0.52	0.08	0.06	0.01	0.51	0.08
Norm	Est	-15.79	2.00	0.97	1.99	17.77	2.00
	SE	0.95	0.15	0.55	0.09	0.96	0.15
	Len	1.87	1.93	9.20	9.02	1.92	1.95
	Cov	0.00	0.97	0.99	0.99	0.00	0.96
Trans	Est	-10.69	2.00	0.96	2.00	12.69	2.01
	SE	0.67	0.10	0.39	0.06	0.67	0.11
	Len	1.32	1.35	6.41	6.38	1.34	1.35
	Cov	0.00	0.74	1.00	0.99	0.00	0.72
Quant	Est	-2.89	2.00	1.00	2.00	4.87	2.00
	SE	0.57	0.08	0.06	0.01	0.57	0.08
	Len	1.12	1.08	1.06	1.05	1.13	1.06
	Cov	1.00	1.00	1.00	1.00	1.00	1.00
GH	Est	-27.83	1.98	0.95	1.99	29.83	1.98
	SE	3.55	0.55	0.56	0.09	3.58	0.54
	Len	6.96	7.07	9.26	9.13	7.16	6.95
	Cov	0.00	1.00	0.70	0.99	0.00	1.00
GL	Est	-12.38	2.00	1.17	1.99	13.71	1.99
	SE	1.32	0.21	0.31	0.05	1.20	0.18
	Len	2.58	2.66	5.14	5.06	2.40	2.34
	Cov	0.01	0.86	0.47	0.84	0.02	0.84
PP	Est	-10.74	2.00	1.47	1.99	11.84	1.99
	SE	2.38	0.29	0.53	0.08	1.54	0.19
	Len	4.68	3.75	8.82	8.80	3.07	2.41
	Cov	0.44	0.69	0.19	0.81	0.42	0.66

Table 4.4: Properties of $\hat{\beta}_0$ and $\hat{\beta}_1$ over 500 datasets, for quantile regressions with $\tau = \{0.05, 0.50, 0.95\}$, Case 3. Sample size is 1000.

Synthesis Model		Analysis Model					
		$\tau = 0.05$		$\tau = 0.50$		$\tau = 0.95$	
		β_0	β_1	β_0	β_1	β_0	β_1
ORIG	Est	-9.78	2.00	1.02	1.99	11.79	2.00
	SE	0.22	0.03	0.49	0.08	0.22	0.03
Norm	Est	-10.36	1.99	1.03	1.99	12.40	1.99
	SE	0.65	0.10	0.38	0.06	0.65	0.10
	Len	2.97	2.99	0.78	0.78	3.07	3.06
	Cov	0.99	1.00	0.99	0.99	0.98	1.00
Trans	Est	-10.15	1.96	1.03	1.99	12.50	1.99
	SE	0.63	0.10	0.37	0.06	0.68	0.11
	Len	2.88	2.95	0.77	0.78	3.22	3.21
	Cov	1.00	0.98	0.99	0.99	0.97	1.00
Quant	Est	-9.76	2.00	1.03	1.99	11.77	2.00
	SE	0.24	0.04	0.52	0.08	0.23	0.04
	Len	1.09	1.05	1.07	1.06	1.09	1.06
	Cov	1.00	1.00	1.00	1.00	1.00	1.00
GH	Est	-6.53	1.99	1.02	1.99	8.59	1.99
	SE	0.11	0.02	0.38	0.06	0.11	0.02
	Len	0.49	0.47	0.77	0.77	0.52	0.49
	Cov	0.00	0.46	1.00	0.99	0.00	0.51
GL	Est	-3.53	2.00	3.18	2.00	11.94	2.00
	SE	0.63	0.10	0.17	0.03	0.97	0.15
	Len	2.86	2.81	0.36	0.36	4.57	4.41
	Cov	0.03	0.95	0.00	0.43	0.61	0.99
PP	Est	-9.84	1.99	1.01	2.00	11.90	1.99
	SE	0.19	0.03	0.50	0.08	0.19	0.03
	Len	0.86	0.82	1.03	1.03	0.89	0.86
	Cov	0.86	0.76	1.00	1.00	0.83	0.79

Table 4.5: Properties of $\hat{\beta}_0$ and $\hat{\beta}_1$ over 500 datasets, for quantile regressions with $\tau = \{0.05, 0.50, 0.95\}$, Case 4. Sample size is 1000.

Disclosure risks

We now discuss the disclosure risks for the various synthesis models in our simulation study. Table 4.6 documents the values for expected, perceived, number of true matches and false rate risk measures for each of the four data generation mechanisms. For Case 1, the four risks measures have similar values across all synthesis models. In Case 2, we observe that the expected risk is noticeably higher for the Quant approach than any of the other models. Following behind are the PP, GL and Trans approaches. The correct number of matches uniquely made for the GL approach are 8, as compared to 13 for the Quant and 5 for the Norm and Trans approaches. Given the data utility results, not surprisingly, the risks for the GH approach remain low. In Case 3, the expected, perceived and true risks for the Quant approach remain considerably higher than the rest of the models, all of which perform similarly. Case 4 results look very close to those for Case 1 with risks for all models not far from each other, although the expected risk for the GH approach is slightly higher than the rest and about 7 true matches can be made under the Trans approach, as compared to less than 6 for all the other approaches.

The risk measures in Table 4.6 have been calculated over all the observations in a single data set. Given our focus on outliers, we also check how many of the true matches made uniquely belong to either the top or bottom 5% of the observations for Y. The third column in 4.6 shows these numbers in brackets. We observe that whenever any true matches are made, some of these belong to the observations in the tails of Y. In Case 1, overall more than half the true matches made belong to the tails. However, the numbers in Case 2 do not indicate that the more extreme observations in Y are at more risk than any other observations. However, in Case 3, we again observe that most of the true matches belong to the tails of Y, although the PP approach results in no true matches at all. Similar results are observed for Case 4. It is no surprise that the tails of the distribution of Y happen to result in some true matches. The important point to note is that for the data keeper, if the intruder is able to identify even one target correctly, sensitive information for the target may be revealed. As such, it is vital to limit the number of true matches. We find that the number of true matches generally goes hand in hand with the utility of the synthesis models. Nevertheless, the relationship between the two is not perfect. For instance, we did not find the utility of the Norm approach in Case 2 of exceptional quality, yet the true match risks for Norm remain as much as they were in Case 1. Looking at the data utility results, using the Quant approach for synthesis may be an attractive option but as we see from the results for disclosure risks, the data keeper may find it too risky to use.

Synthesis model	Exp	Per	True	False
Norm	2.307	0	4 (2)	0.995
Trans	2.523	0	5 (4)	0.993
Quant	2.438	1	2 (1)	0.997
GH	2.302	1	2 (2)	0.997
GL	2.572	1	1 (1)	0.999
PP	2.332	0	1 (0)	0.999
Case 2				
Norm	2.850	0	5 (0)	0.994
Trans	3.373	3	5 (1)	0.994
Quant	8.372	1	13 (5)	0.986
GH	1.035	0	1 (1)	0.999
GL	4.765	0	8 (3)	0.990
PP	5.075	2	3 (1)	0.996
Case 3				
Norm	1.102	0	1 (1)	0.999
Trans	2.091	0	1 (1)	0.999
Quant	9.470	6	12 (5)	0.987
GH	1.005	0	1 (1)	0.999
GL	1.359	0	2 (2)	0.998
PP	0.881	0	0 (0)	1.000
Case 4				
Norm	2.552	1	6 (5)	0.993
Trans	2.619	0	7 (4)	0.992
Quant	2.490	0	4 (3)	0.996
GH	3.178	0	3 (3)	0.997
GL	2.365	0	2 (1)	0.998
PP	2.819	1	1 (1)	0.999

Table 4.6: Disclosure risks for Y_{syn} for the Norm, Trans, Quant, GH, GL and PP synthesis strategies, Cases 1-4.

4.3.2 Hierarchical study

For the hierarchical data study, we generate 500 balanced datasets with $I = 500$ clusters or individuals with $J = 3$ repeated measures or time points each. The data are generated using a random intercepts (RE) model. As before, we generate data using four different mechanisms with varying error term distributions:

1. Case 1: $y_{ij} = \alpha_i + \beta_0 + x_{ij}\beta_1 + \varepsilon_{ij}$, where, $x_{ij} \sim N(4, 5), \beta_0 = 1, \beta_1 = 2$ and $\varepsilon_{ij}, \alpha_i \sim N(0, 7)$;
2. Case 2: $y_{ij} = \alpha_i + \beta_0 + x_{ij}\beta_1 + \varepsilon_{ij}$, where, $x_{ij} \sim N(4, 5), \beta_0 = 1, \beta_1 = 2$ and $\varepsilon_{ij}, \alpha_i \sim LN(0, 1.39)$;
3. Case 3: $y_{ij} = \alpha_i + \beta_0 + x_{ij}\beta_1 + \varepsilon_{ij}$, where, $x_{ij} \sim N(4, 5), \beta_0 = 1, \beta_1 = 2$ and $\varepsilon_{ij}, \alpha_i \sim t_{1.46}$;
4. Case 4: $y_{ij} = \alpha_i + \beta_0 + x_{ij}\beta_1 + \varepsilon_{ij}$, where, $x_{ij} \sim N(4, 5), \beta_0 = 1, \beta_1 = 2$ and $\varepsilon_{ij}, \alpha_i \sim U(-12, 12)$.

In this study, we do not differentiate between the consequences of misspecification of the within or between error terms. In each of the cases above, we contaminate both at the same time using the same kind of mechanism. We then synthesise Y using the following methods:

1. Norm: $Y_{syn} = \boldsymbol{X}\hat{\beta} + \alpha_i + v_{ij}$, where $v_{ij} \sim N(0, \hat{\sigma_e}^2)$, $\alpha_i \sim N(0, \hat{\sigma_b}^2)$, $\hat{\beta}$, $\hat{\sigma_b}^2$ and $\hat{\sigma_e}^2$ are obtained by maximising the likelihood of a random intercepts model;
2. Trans: same as Norm, with the transformed response variable, $Y^T_{syn} = \boldsymbol{X}\hat{\beta}+\alpha_i+v_{ij}$, where Y^T is obtained by transforming Y using a Box-Cox transformation, Y^T_{syn} is transformed back to obtain Y_{syn};
3. Quant: Y_{syn} is obtained by sampling from the various random intercepts quantile regressions, as described in Section 4.2.2;
4. GH: $Y_{syn} = \boldsymbol{X}\hat{\beta}+V$, where V are obtained by sampling from a multivariate GH as described in Section 4.2.3, $\hat{\beta}$ is obtained by maximising the likelihood of a random intercepts model;
5. GL: $Y_{syn} = \boldsymbol{X}\hat{\beta}+V$, where V are obtained by sampling from a multivariate GL as described in Section 4.2.3, $\hat{\beta}$ is obtained by maximising the likelihood of a random intercepts model;
6. PP: $Y_{syn} = \boldsymbol{X}\hat{\beta}+V$, where V are obtained by sampling from a multivariate PP as described in Section 4.2.3, $\hat{\beta}$ is obtained by maximising the likelihood of a random intercepts model.

For each method, we create $M = 10$ copies for Y_{syn}, and combine analyses over the 10 copies using the partial synthetic data rules in Section 1.1.1. The estimands of interest are:

1. the mean and variance, and sample quantiles of Y at levels $\{0.25, 0.50, 0.75\}$;
2. the estimate, standard errors, length and coverage of confidence intervals, CI, for the coefficients of random intercepts quantile regressions of Y on $\boldsymbol{X}$, $\{\beta_0, \beta_1\}$ at levels $\tau = \{0.25, 0.50, 0.75\}$ (The levels for τ were chosen differently from the non-hierarchical simulation study in light of computation times);
3. the within, σ_e^2, and between, σ_b^2, error variance estimates for each fitted quantile RE model.

Again, we utilise the bootstrap method to obtain the standard errors for the various coefficient estimates for the quantile regressions, and use these in combination with MI combining rules for partially synthetic data to obtain the final results. As before, analysis results from the original data are treated as the benchmark.

We continue to use R (R Core Team, 2016) for the simulations. Additional packages used for the hierarchical study include *lqmm* (Geraci, 2014) for quantile regressions for mixed linear models, *rootSolve* to solve for intermediate correlations for the GH distribution fitting and *cubature* (Johnson and Narasimhan, 2013) for intergration when fitting the GL distribution.

Data utility

Table 4.7 shows the summary statistics of the original and synthetic data for all cases. In Case 1, most of the synthesis models preserve the sample estimates for mean, variance and the different quantiles; the only notable discrepancy is in the performance of the Quant approach. The Quant approach results in quantile estimates that are slightly different from the original ones, and the variability in Y is also lower (~ 137 instead of ~ 160). In Case 2, the Norm approach performs generally well, with very slight discrepancy in the quantile estimates. The Trans approach improves the top tail behaviour of the Norm only, but results in underestimation of the variance of Y. The Quant approach also underestimates the variability in Y. Amongst the three parametric approaches, the GL approach reproduces the original data statistics most closely. The various quantile estimates for the GH approach are all biased, and a few extreme values inflate the variance estimate to unacceptable levels. The variance for the PP approach is also slightly higher than what observed for the original data, but the quantile estimates are generally well-preserved.

For Case 3, the various approaches perform quite differently. Firstly, the Norm approach preserves the mean, median and variance of Y quite well but the tail quantiles are biased. As in Case 2, the Trans approach slightly improves on the bias of the tail quantile

estimates, but the resulting variance estimate for Y is quite low. For the Quant approach, we observe similar results. The shape of the distribution is best preserved using the Quant approach, but the variance estimate is surprisingly low. The GH approach performs similarly to the Norm approach in terms of the quantile estimates, but results in very high variance estimates due to the generation of a few very large values. Again, the GL approach performs generally well, and so does PP but with added variability in Y. In Case 4 most models perform well. There is, however, slight discrepancy for the Quant approach in the tail behaviour, and the final variance estimate is also lower than that of the original data, as in Case 1.

Synthesis Model	Q.25	Q.50	Mean	Q.75	Var
			Case 1		
ORIG	0.55	9.01	9.00	17.47	157.66
Norm	0.53	9.01	9.00	17.46	157.63
Trans	0.54	9.00	9.00	17.46	157.51
Quant	1.04	9.01	8.98	16.95	136.88
GH	0.55	9.01	9.00	17.47	160.66
GL	0.47	9.04	9.00	17.55	157.66
PP	0.53	9.01	9.00	17.48	157.73
			Case 2		
ORIG	5.10	12.40	13.27	20.07	193.52
Norm	4.17	13.26	13.25	22.35	193.19
Trans	4.05	11.88	13.26	20.96	165.30
Quant	5.86	13.28	13.68	20.96	141.42
GH	-0.52	8.47	13.23	19.35	1812.19
GL	4.99	12.95	13.27	21.04	192.85
PP	4.60	12.11	13.26	20.13	236.39
			Case 3		
ORIG	0.70	8.02	8.02	15.31	1127.92
Norm	-5.61	8.06	8.03	21.70	1120.86
Trans	-3.07	8.07	8.06	19.19	498.60
Quant	0.59	8.09	8.42	15.67	185.72
GH	-4.99	8.20	8.06	21.30	27770.64
GL	-1.44	7.96	8.02	17.41	1269.73
PP	-1.78	7.67	8.03	17.43	2569.39
			Case 4		
ORIG	-1.62	7.96	7.98	17.58	195.72
Norm	-1.46	7.97	7.97	17.40	195.89
Trans	-1.49	7.93	7.98	17.39	195.45
Quant	-0.74	8.11	7.96	16.78	162.76
GH	-1.37	7.98	7.97	17.32	195.37
GL	-1.74	8.02	7.98	17.72	195.89
PP	-1.64	7.99	7.98	17.60	195.87

Table 4.7: Summary of Y compared to Y_{syn} over 500 datasets, for Cases 1-4 for the Norm, Trans, Quant, GH, GL and PP synthesis strategies for hierarchical data. Sample size is 1500, 500 clusters, 3 observations per cluster.

We now discuss the results for the quantile regressions of interest. Tables 4.8 - 4.11 show the intercept and slope estimates for our three quantile regressions for the four different cases under study. Throughout the four cases, we find that the estimates, standard errors and coverages for β_1 remain consistently preserved across all three quantiles. There are, however, biases in the 25^{th} and 75^{th} quantile β_1 estimates for the Trans approach, as observed in the non-hierarchical case study.

For β_0, the results are more variable. In Case 1, β_0, alongwith its standard error and CI is generally well preserved by all the synthesis models, although some seem slightly more efficient than others. In Case 2, estimates for β_0 and its standard error generally remain biased, and low coverages are observed throughout Table 4.9. Although the estimates from the Quant approach are still closest to the original data estimates, the Quant approach is not universally better than the other approaches. In estimating β_0 for the 75^{th} quantile regression, PP and GL both perform better than Quant. As observed from the summary statistics, Trans improves upon the point estimates from Norm with slightly smaller standard errors, but these are not always large enough to provide nominal coverage of the true value of β_0. In Case 3, all β_0 related inferences remain biased, except for those for the median regression, and although Quant performs better than the other approaches, coverages of the true value remain lower than 30%. In Case 4, all models perform very well, and estimates, standard errors, and coverages for β_0 observed are as good as those for β_1.

When using mixed effects models for quantile regression, the scale parameter equivalents for within and between error variances as in an RE model may also be calculated (Geraci and Bottai, 2014). These may be used to calculate the intraclass correlation coefficient, ICC, which is normally of interest to analysts. We calculate the estimates for the two variances for our analysis models and compare them across all the synthesis procedures. Table 4.12 shows the results. In Case 1, all the within and between error estimates are generally preserved, with the exception of results from the Quant procedure. With the Quant approach, all within error estimates are slightly lower than those observed for the original data, and the between error estimates are considerably underestimated.

In Case 2, we find that the Norm and Trans approaches consistently overestimate the error variances across all regressions, more noticeably so for the between component. The Quant approach has better results but some of the between components are biased downwards. Estimates from the GH and GL approaches are mostly biased upwards but less severely so for the GL approach. Finally, with the exception of the between component for the median regression, all other PP estimates are estimated close to the original values.

In Case 3, we observe that the Norm, Trans, GH and GL approaches have heavily biased between error variance estimates. The within error variances are less biased but still higher than the estimates from the original data. Results from the Quant approach

Synthesis Model		Analysis Model					
		$\tau = 0.25$		$\tau = 0.50$		$\tau = 0.75$	
		β_0	β_1	β_0	β_1	β_0	β_1
ORIG	Est	-3.65	2.00	1.01	2.00	5.67	2.00
	SE	0.32	0.05	0.30	0.04	0.32	0.05
Norm	Est	-3.66	2.00	1.00	2.00	5.67	2.00
	SE	0.34	0.05	0.31	0.05	0.34	0.05
	Len	1.02	1.02	1.03	1.02	1.04	1.03
	Cov	1.00	1.00	1.00	1.00	1.00	1.00
Trans	Est	-3.64	2.00	1.01	2.00	5.67	2.00
	SE	0.34	0.05	0.31	0.05	0.34	0.05
	Len	1.02	1.02	1.03	1.02	1.03	1.03
	Cov	1.00	1.00	1.00	1.00	1.00	1.00
Quant	Est	-3.32	2.00	1.02	2.00	5.42	1.99
	SE	0.29	0.04	0.29	0.04	0.30	0.04
	Len	0.88	0.87	0.94	0.98	0.92	0.83
	Cov	0.97	1.00	1.00	1.00	0.98	1.00
GH	Est	-3.63	2.00	1.01	2.00	5.65	2.00
	SE	0.34	0.05	0.31	0.04	0.34	0.05
	Len	1.01	1.01	1.03	1.01	1.03	1.01
	Cov	1.00	0.99	1.00	1.00	0.99	1.00
GL	Est	-3.90	2.00	1.12	2.00	6.02	2.00
	SE	0.37	0.05	0.35	0.05	0.37	0.05
	Len	1.12	1.04	1.15	1.11	1.12	0.96
	Cov	0.99	1.00	1.00	1.00	0.91	1.00
PP	Est	-3.70	2.00	1.00	2.00	5.71	2.00
	SE	0.34	0.05	0.32	0.05	0.34	0.05
	Len	1.03	1.02	1.04	1.03	1.04	1.03
	Cov	1.00	1.00	1.00	1.00	1.00	1.00

Table 4.8: Properties of $\hat{\beta}_0$ and $\hat{\beta}_1$ over 500 datasets, for quantile regressions with $\tau = \{0.25, 0.50, 0.75\}$ for hierarchical data, Case 1. Sample size is 1500, 500 clusters, 3 observations per cluster.

are close to the original data estimates; nevertheless, the between error variance is underestimated at the 75^{th} quantile regression. The between variance estimates are also underestimated for the PP approach, while the within ones are overestimated. In Case 4, most models perform generally well, replicating the estimates from the original data, but the between error variance is underestimated for the Quant approach.

Synthesis Model		Analysis Model					
		$\tau = 0.25$		$\tau = 0.50$		$\tau = 0.75$	
		β_0	β_1	β_0	β_1	β_0	β_1
ORIG	Est	1.41	2.00	3.72	2.00	6.92	2.00
	SE	0.09	0.01	0.41	0.01	0.60	0.02
Norm	Est	0.77	2.00	5.27	2.00	9.76	1.99
	SE	0.45	0.05	0.42	0.05	0.45	0.05
	Len	4.68	4.81	1.01	3.22	0.73	2.42
	Cov	0.74	0.98	0.06	0.99	0.01	1.00
Trans	Est	1.57	1.95	4.77	2.06	8.22	2.17
	SE	0.29	0.04	0.28	0.04	0.32	0.05
	Len	3.00	4.26	0.67	2.95	0.52	2.25
	Cov	0.91	0.88	0.11	0.82	0.10	0.05
Quant	Est	1.47	2.00	4.12	2.00	8.33	2.00
	SE	0.08	0.01	0.25	0.03	0.32	0.04
	Len	0.83	1.19	0.59	1.96	0.53	1.90
	Cov	0.97	1.00	0.57	1.00	0.06	1.00
GH	Est	-5.52	2.00	-0.14	2.00	9.36	2.01
	SE	0.42	0.04	0.76	0.05	1.40	0.07
	Len	4.40	3.60	1.82	3.40	2.29	3.61
	Cov	0.01	0.91	0.11	0.99	0.51	1.00
GL	Est	2.14	2.00	4.74	2.00	7.98	2.00
	SE	0.32	0.03	0.27	0.02	0.42	0.03
	Len	3.34	2.81	0.64	1.73	0.68	1.55
	Cov	0.34	0.86	0.10	0.81	0.33	0.91
PP	Est	0.82	2.00	2.68	2.00	7.64	2.00
	SE	0.07	0.01	0.25	0.02	0.79	0.03
	Len	0.73	0.94	0.59	1.52	1.30	1.31
	Cov	0.19	0.37	0.15	0.80	0.90	0.90

Table 4.9: Properties of $\hat{\beta}_0$ and $\hat{\beta}_1$ over 500 datasets, for quantile regressions with $\tau = \{0.25, 0.50, 0.75\}$ for hierarchical data, Case 2. Sample size is 1500, 500 clusters, 3 observations per cluster.

Synthesis Model		Analysis Model					
		$\tau = 0.25$		$\tau = 0.50$		$\tau = 0.75$	
		β_0	β_1	β_0	β_1	β_0	β_1
ORIG	Est	-1.64	2.00	0.00	2.00	1.65	2.00
	SE	0.20	0.02	0.12	0.01	0.20	0.02
Norm	Est	-8.35	2.00	0.03	1.99	8.50	1.98
	SE	0.84	0.09	0.83	0.08	0.85	0.09
	Len	4.04	4.94	6.94	5.82	4.08	4.83
	Cov	0.01	0.99	0.99	0.98	0.00	0.99
Trans	Est	-6.05	2.01	0.05	2.00	6.14	2.00
	SE	0.63	0.07	0.62	0.06	0.64	0.07
	Len	3.03	3.71	5.19	4.34	3.06	3.63
	Cov	0.00	0.94	0.99	0.99	0.00	0.92
Quant	Est	-2.35	2.00	0.00	2.00	2.34	2.00
	SE	0.21	0.03	0.13	0.02	0.21	0.03
	Len	1.01	1.55	1.07	1.35	1.01	1.52
	Cov	0.15	1.00	1.00	1.00	0.15	1.00
GH	Est	-8.98	1.99	0.05	1.99	9.28	1.99
	SE	1.66	0.10	1.31	0.08	1.63	0.10
	Len	7.98	5.86	10.93	5.54	7.85	5.67
	Cov	0.02	0.98	0.79	0.95	0.01	0.98
GL	Est	-4.48	1.99	0.02	1.99	4.59	1.99
	SE	0.59	0.04	0.50	0.04	0.64	0.04
	Len	2.82	2.49	4.14	2.64	3.06	2.47
	Cov	0.08	0.87	0.79	0.81	0.09	0.86
PP	Est	-5.18	1.99	-0.28	1.99	4.70	2.00
	SE	0.51	0.04	0.30	0.04	0.60	0.05
	Len	2.45	2.48	2.54	2.67	2.89	2.71
	Cov	0.31	0.79	0.22	0.66	0.29	0.79

Table 4.10: Properties of $\hat{\beta}_0$ and $\hat{\beta}_1$ over 500 datasets, for quantile regressions with $\tau = \{0.25, 0.50, 0.75\}$ for hierarchical data, Case 3. Sample size is 1500, 500 clusters, 3 observations per cluster.

Synthesis Model		Analysis Model					
		$\tau = 0.25$		$\tau = 0.50$		$\tau = 0.75$	
		β_0	β_1	β_0	β_1	β_0	β_1
ORIG	Est	-5.20	2.01	-0.02	2.00	5.15	1.99
	SE	0.47	0.05	0.45	0.05	0.47	0.05
Norm	Est	-4.97	2.01	-0.04	2.00	4.89	2.00
	SE	0.49	0.06	0.46	0.05	0.49	0.06
	Len	1.02	1.00	1.00	0.95	1.02	1.00
	Cov	0.99	1.00	1.00	1.00	0.99	1.00
Trans	Est	-4.92	2.00	-0.03	2.00	4.89	2.00
	SE	0.49	0.06	0.46	0.05	0.49	0.06
	Len	1.00	1.00	0.99	0.95	1.02	1.00
	Cov	0.99	1.00	1.00	1.00	0.99	1.00
Quant	Est	-4.90	2.01	0.02	2.00	5.27	1.99
	SE	0.40	0.05	0.38	0.05	0.51	0.04
	Len	0.83	0.90	0.82	0.92	1.06	0.80
	Cov	0.97	1.00	1.00	1.00	1.00	0.99
GH	Est	-4.81	2.01	-0.03	2.00	4.75	1.99
	SE	0.48	0.05	0.45	0.05	0.48	0.05
	Len	1.00	0.96	0.98	0.92	1.01	0.97
	Cov	0.92	1.00	1.00	1.00	0.94	0.99
GL	Est	-4.84	2.01	0.08	2.00	4.98	1.99
	SE	0.53	0.05	0.53	0.05	0.57	0.05
	Len	1.08	0.93	1.15	0.96	1.18	0.96
	Cov	0.96	1.00	1.00	1.00	0.99	1.00
PP	Est	-5.08	2.01	-0.03	2.00	5.02	1.99
	SE	0.51	0.06	0.48	0.05	0.51	0.06
	Len	1.05	1.01	1.04	0.99	1.06	1.01
	Cov	1.00	1.00	1.00	1.00	1.00	1.00

Table 4.11: Properties of $\hat{\beta}_0$ and $\hat{\beta}_1$ over 500 datasets, for quantile regressions with $\tau = \{0.25, 0.50, 0.75\}$ for hierarchical data, Case 4. Sample size is 1500, 500 clusters, 3 observations per cluster.

Synthesis Model	Analysis Model					
	$\tau = 0.25$		$\tau = 0.50$		$\tau = 0.75$	
	σ_e^2	σ_b^2	σ_e^2	σ_b^2	σ_e^2	σ_b^2
Case 1						
ORIG	2.00	15.82	2.64	12.44	1.99	15.74
Norm	2.00	15.74	2.64	12.49	2.00	15.74
Trans	1.99	15.80	2.63	12.51	2.00	15.80
Quant	1.68	9.74	2.24	10.36	1.70	7.52
GH	2.00	15.56	2.62	12.53	2.00	15.51
GL	2.00	18.06	2.63	17.17	2.00	14.78
PP	2.00	15.93	2.64	12.97	2.00	15.93
Case 2						
ORIG	1.09	0.00	1.48	5.49	1.24	29.49
Norm	1.66	31.37	2.19	29.92	1.66	31.53
Trans	1.50	19.20	2.03	20.10	1.57	25.06
Quant	1.18	0.00	1.82	2.92	1.44	15.28
GH	3.13	0.00	4.72	10.26	4.27	106.62
GL	1.38	14.50	1.83	11.69	1.47	20.69
PP	1.16	0.00	1.98	0.31	1.61	27.29
Case 3						
ORIG	1.19	4.71	1.65	0.00	1.22	4.78
Norm	1.94	40.20	2.55	35.60	1.93	39.88
Trans	1.72	33.85	2.28	31.47	1.72	33.89
Quant	1.41	3.48	1.87	0.12	1.54	2.39
GH	5.09	37.70	6.13	11.74	5.12	37.71
GL	1.71	17.18	2.14	12.15	1.70	16.02
PP	1.99	0.71	2.33	0.00	1.98	1.35
Case 4						
ORIG	2.01	50.51	2.69	50.90	2.01	50.40
Norm	1.98	50.13	2.62	47.84	1.98	50.15
Trans	1.98	49.99	2.62	47.94	1.99	50.43
Quant	1.89	28.28	2.56	27.19	1.98	18.78
GH	1.94	48.41	2.55	46.27	1.93	48.25
GL	1.95	52.31	2.59	54.38	1.94	51.46
PP	1.97	52.48	2.62	52.87	1.97	52.44

Table 4.12: Summary of $\hat{\sigma}_e^2$ and $\hat{\sigma}_b^2$ over 500 datasets, for Cases 1-4 for the Norm, Trans, Quant, GH, GL and PP synthesis strategies for hierarchical data. Sample size is 1500, 500 clusters, 3 observations per cluster.

Disclosure risks

We now evaluate the relative disclosure risks for all our synthesis methods. Table 4.13 displays the results. We observe that across all four cases, the expected risk is noticeably higher for the Quant approach, although this is also true for the PP approach in Cases 2 and 3. Probabilities of unique matches that exceed 0.2 are counted in the perceived risk measure. Again, the Quant approach leads in the perceived risks, while these remain quite low for the Norm, Trans, GH and GL synthesis procedures. In Cases 1 and 4, the perceived and true risks for Norm are also higher than most of the other competing methods. The number of unique correct matches is also higher for Quant in all the cases. In Cases 2 and 3, PP also appears more risky than other models, followed by the GL approach. False match rates remain consistently high, implying that most of the unique matches made by an intruder are incorrect.

In Table 4.13, we also show the number of true matches that belong to the top and bottom 5% observations of the response variable. As we are dealing with hierarchical data, we use the average of Y across the three time points to determine which units belong to the tails of Y. Our results are shown in brackets within the third column of the table. We find that unlike the non-hierarchical study, the majority of the true matches do not belong to the top or bottom 5% of the units in the dataset. This might also be a consequence of the fact that besides the unit intercepts, we did not design the study such that units have similar values of the predictor variable across the time points.

Discussion

Overall, the hierarchical study resulted in some trends that were also observed for the non-hierarchical study. The slopes were generally better preserved than the intercepts, which is also a consequence of the design of the study. Secondly, the Trans approach can result in biases in the slope estimates at the more extreme quantile regressions. The GH approach often produces a few extreme predicted values that inflate the variance of the synthetic variable, and the standard errors of the coefficients involved. The performance of the GH approach can be improved by regulating the synthesis procedure, discarding the 'bad' copies and generating new ones that are more appropriate. Both the PP and GL approaches generally perform well in preserving the shape of the data. However, synthetic data from the GL better matches the variability in the original data in some cases, while the PP approach sometimes performs better in terms of the intercept estimates and standard errors.

We also observed some differences from the non-hierarchical study. Most notably, although the Quant approach performs well in general, the Norm approach produces better results when the error distributions are normal. Even when the error distributions are not Normal, some of the other approaches perform better than the Quant approach for certain quantile estimates and consequently, quantile regressions. We observed that the

Synthesis model	Exp	Per	True	False
		Case 1		
Norm	4.546	11	8 (2)	0.982
Trans	4.450	5	5 (0)	0.989
Quant	6.764	16	9 (1)	0.981
GH	4.404	5	8 (1)	0.982
GL	4.145	5	5 (1)	0.989
PP	4.655	6	5 (1)	0.989
		Case 2		
Norm	3.473	1	3 (0)	0.993
Trans	5.796	4	11 (0)	0.977
Quant	11.700	15	22 (0)	0.955
GH	1.971	1	2 (0)	0.995
GL	7.190	4	14 (0)	0.970
PP	11.556	14	16 (0)	0.967
		Case 3		
Norm	1.354	0	1 (1)	0.996
Trans	1.874	0	6 (2)	0.985
Quant	13.824	22	19 (0)	0.960
GH	3.803	1	7 (2)	0.981
GL	5.392	3	14 (1)	0.967
PP	6.272	4	19 (1)	0.958
		Case 4		
Norm	3.139	11	5 (1)	0.984
Trans	3.121	4	3 (0)	0.991
Quant	4.479	16	6 (2)	0.984
GH	2.628	4	4 (2)	0.987
GL	2.800	8	2 (0)	0.993
PP	2.589	2	3 (0)	0.991

Table 4.13: Disclosure risks for Y_{syn} for the Norm, Trans, Quant, GH, GL and PP synthesis strategies for hierarchical data, Cases 1-4.

Quant approach often resulted in reduced variability in the synthetic data, and some of the within and between error variance estimates were also underestimated. Nevertheless, Quant is still the most risky of all the synthesis models considered as our disclosure risk results showed. However, we did not find evidence that more unusual units in the dataset were at any more risk than other units in the dataset when considering the number of correct matches made by an intruder.

4.4 Real data application

We now illustrate the proposed methodology on a real data application of interest. The Wealth and Assets Survey (WAS, ONS (2016)) is a longitudinal survey that started running in 2006, and currently has four complete waves of data. Every two years, a sample of private households in Great Britain is interviewed, and it is attempted to

follow the same households for the subsequent waves of data collection. Over 30,000 households are approached in each wave. The WAS is valuable source of information on the economic well-being of households in Britain. Data on household wealth are sparse, and income is usually used as a proxy for analyses. However, families now own properties more than ever before, make investments in stocks, and accumulate wealth in retirement schemes. WAS covers extensive detail on household wealth through levels of assets, savings, pensions, distribution of wealth, and factors that affect financial planning. Potential analysts may wish to link this information to other personal information collected by the survey, such as the sex, ethnicity or other characteristics of the families comprising the household. The WAS has, therefore, a number of variables which are skewed to the right, and contain sensitive information. The WAS has two access licenses, a Special License (SL) and an End User License (EUL). The EUL data do not contain information on geography and certain variables are banded or capped to suppress outliers. Both datasets do not contain direct identifiers, such as National Insurance numbers. However, case numbers are provided to be able to link datasets across waves. Currently, the WAS EUL data may be accessed online through registration on the UK Data Service website (https://www.ukdataservice.ac.uk) subject to agreement to certain terms and conditions. Requests to access the SL version can be made through a special application. In each wave, WAS consists of two surveys, one for the questions on the household as a whole, and another personal survey, completed by all adults in the household individually.

As of date, data from WAS have been mainly utilised by the Office for National Statistics (ONS) to produce various reports, and there are only a handful of research articles from academics from other institutions. With substantial information on various aspects of wealth in Great Britain, we expect that the data will become more popular amongst researchers. In our illustration, we attempt to replicate one of the many analyses carried out by the ONS. We focus on an infographic published by the ONS in October 2013 (ONS, 2013) regarding the amount and distribution of inheritances in Great Britain. The analysis we attempt to replicate is a logistic regression studying the characteristics that make an individual more likely to receive inheritance. Similar analyses have been carried out in the past using the British Household Panel Survey (Ross et al., 2008). For the WAS analysis, the predictors in the regression have been extracted from both household level and personal level data. These include, total household wealth, household type, the socioeconomic classification of the household, and the individual's age, sex, ethnicity, qualifications, father's education, number of siblings and economic activity. One of the findings of the analysis is that individuals belonging to wealthier households are most likely to receive inheritance. While this result may be expected, it has important implications for public policy, which also include tax laws and incentives. We assume a hypothetical situation where the total wealth variable must be synthesised. Although none of our simulation studies deal with synthetic data within predictors, for a non

linear model, we expect that the analysis model should be congenial to our synthesis procedures as long as all the covariates involved are also part of the synthesis process.

The ONS analysis uses waves 1-2 from WAS. The analysis itself in unusual because it is treated as a non-hierarchical analysis on a single set of households, but information on individuals have been gathered from both waves of the data. Variables such as total wealth and inheritance are from wave 2, while household type and tenure are extracted from wave 1. This means that only households that are present in both waves of the data are included. Secondly, no distinction is made between households and individuals; we assume that the household level and personal level data have been linked in each wave and only one person from each household is included in the analysis. Given these details, we focus our illustration on households that participated in both the waves, and only the household reference person (HRP), the person responsible for the accommodation. As such, we suppress the clustering within each wave for this illustration and ignore other members of the household. Synthesising longitudinal household survey data is a notoriously difficult problem, we refer readers to the article by Hu et al. (2014a) for latest research on the topic.

The analysis converts total wealth into deciles to be used as a predictor; this will be affected by the shape of the wealth variable and it is important to preserve it as well as possible to replicate the analysis on the original data. We select all the variables used in the analysis for our synthesis procedures from both the waves. Nevertheless, the variables used for the final analysis are not necessarily good predictors for total household wealth. In the survey data, total wealth is a linear combination of the property, physical, financial and private pension wealth of individuals. While these variables are available in the EUL data, we prefer dealing with a more realistic setting, where such direct predictors are not available to the data keepers. Therefore, we avoid using any of the variables that fall under this category, instead utilising other characteristic variables to predict household wealth. The additional variables we use include number of bedrooms in the house and output area code assigned to the household. Some predictors such as ethnicity were not included in the EUL we obtained, and while they may be good predictors, we were not able to use them in our procedures.

Our final data contain 12900 households, i.e. 25800 observations for data arranged in the long format. We synthesise total wealth of the households from both waves, treating the data as longitudinal, using a random intercepts model. We repeat our procedures as in Section 4.3.2 for all the synthesis models. To aid the synthesis process, we also divide the data into four groups of wealth quantiles and synthesise each group independently from another. We create 10 copies for the synthetic data for each method and run the logistic analysis explained above for each procedure. Using the results on the original data as a benchmark, we then evaluate the utility of our synthesis procedures.

Figure 4.3 displays box plots of the original total wealth variable along with those for a single copy of the synthetic wealth from each of our procedures. We observe that the real data have a long right tail. Synthesising such a variable using the Norm approach fails to capture this tail and also has a left tail that is not observed for the real data. The box plot from the Trans approach indicate better shape with a long right tail but the tail is not long enough to match the shape of the original variable. The best shape is observed for the Quant approach, and although the tail is slightly short as compared to the real data, the shape is replicated well. We also see that the GH and GL approaches somewhat replicate the right tail in the data, but also have some extreme values at the left tail. As expected, the GH approach extends the right tail and has some extreme values not found in the real data. Finally, the PP approach captures the right tail behaviour quite well, but there are some observations on the left that do not exist in the real data. Different copies of the synthetic data show slight variations from the displayed set but the trends are the same.

Next, we consider the estimate and CI for the nine levels of wealth coefficient modelled using the logistic regression. In Figure 4.4 we plot the point estimates and CI for the nine covariates for analysis repeated on the original and synthetic data. In the original analysis (black), we find that each category of wealth has a significant impact (p-value < 0.1 as indicated by asterisks) on the probability of inheritance and the magnitude of the impact increases with increasing wealth. The synthetic data results show that both the Norm and GH approaches fail to capture this effect at all levels of wealth. The Trans and Quant approaches perform best, and their CI overlap the CI from the original analysis. Point estimates obtained using the GL and PP approach, although better than those for the Norm and GH approaches, are still underestimated.

In Figure 4.5, we plot results for some of the other covariates including categories of age and sex, that have not been synthesised. We find that these coefficients are mostly unaffected, although models that did not perform well for the wealth covariates also have slight discrepancies for these coefficients.

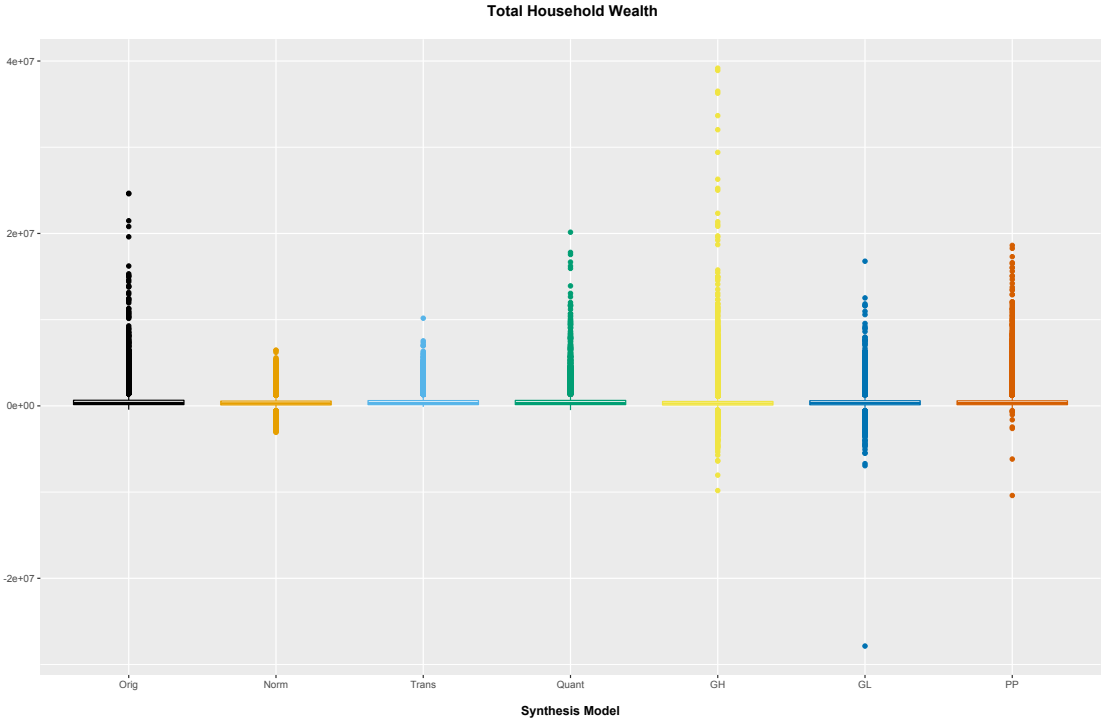

Figure 4.3: Distribution of the original total wealth against synthetic wealth.

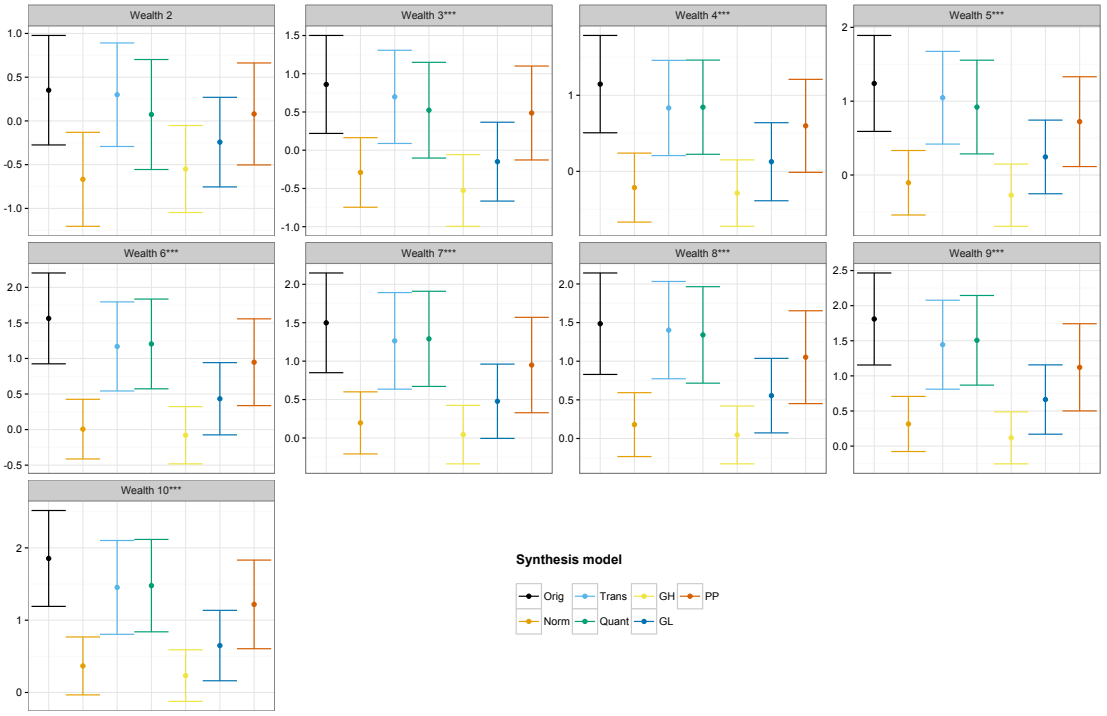

Figure 4.4: Point estimates and confidence intervals for the wealth categories using original and synthetic data.

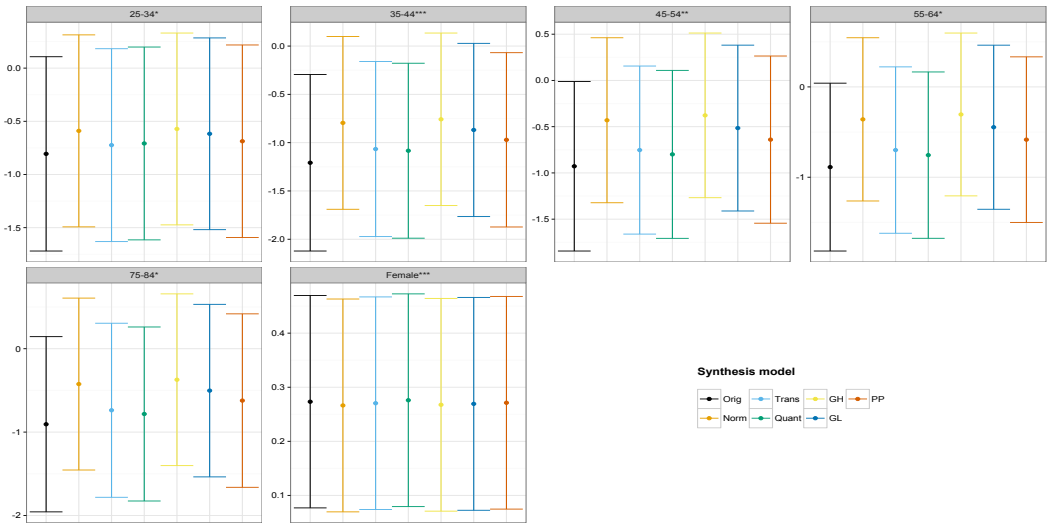

Figure 4.5: Point estimates and confidence intervals for the age categories and sex using original and synthetic data.

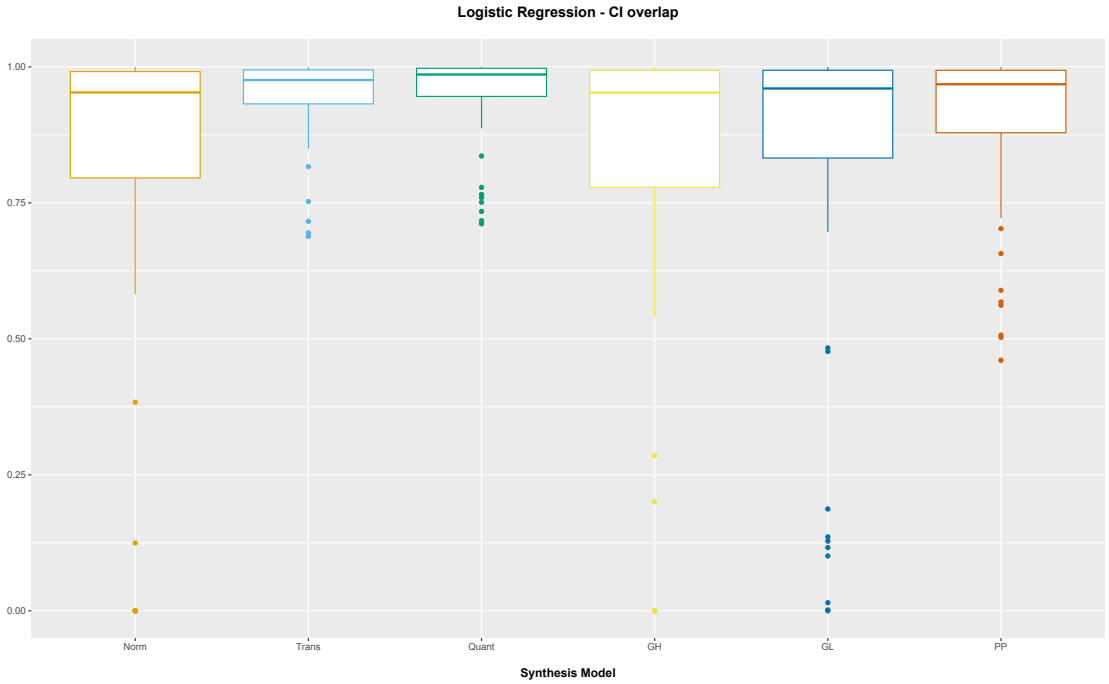

Figure 4.6: Distribution of confidence interval overlap for all covariates in the logistic regression.

Figure 4.6 shows box plots for the CI overlap measure as defined in Section 3.4.1 for all the covariates in the logistic regression. We observe that the Trans and Quant approaches have the highest average of CI overlaps, with none of the overlaps falling below 0.60. Following behind is the PP approach, which has a few coefficients with CI overlap below 0.50. As we know from Figure 4.4, the coefficients with the lowest overlaps belong the wealth deciles. Finally, CI from GH, GL, and Norm have more variability in their CI overlap and have a number of coefficients with CI that do not overlap with the original data CI very much.

In Table 4.14, we document the disclosure risks calculated for our selected data. We help the intruder's process by matching cases based on age categories, assuming these are released unaltered and the intruder knows the age category of the target he is interested in. The results confirm the observations made in the simulations studies. We find that the expected, perceived and true risk measures are the highest for the Quant approach. Although the GL approach did not perform as well as some of the other methods, it results in relatively high risks of identification. Risks from the PP and Trans approach are similar, although the probability of making true matches is higher for PP. This is expected, as we know that PP preserved the shape of the total wealth better than Trans. Given the results for data utility, it is no surprise that the Norm and GH approaches appear less risky as compared to the rest. Furthermore, we display the number of true matches made in brackets for individuals that belonged to the top and bottom tails of the wealth distribution (here, 10%). We find that while a number of true matches belong to the individuals with very high or very low wages, they do not form the majority of the true matches. The highest proportion is about a third, as observed for the Quant approach.

Synthesis model	Exp	Per	True	False
Norm	50.091	41	50 (11)	0.994
Trans	51.813	62	61 (7)	0.993
Quant	67.423	101	78 (25)	0.992
GH	49.507	49	51 (16)	0.994
GL	53.022	92	75 (18)	0.992
PP	58.000	104	64 (13)	0.993

Table 4.14: Disclosure risks for the real data application.

Finally, we present the risk-utility map for our models, plotting average CI overlaps against the true rate of identification. Figure 4.7 shows that for our application using the WAS data, the Trans approach provides the best balance between utility and risk. While the Quant approach provides better data utility, the data keeper must also consider the added disclosure risk associated with its use.

As for any illustration, we caution against the use of these results as proof that some approaches are always better than others. Our results are specific to the analysis model of interest and a consequence of the various choices we have made to synthesise the

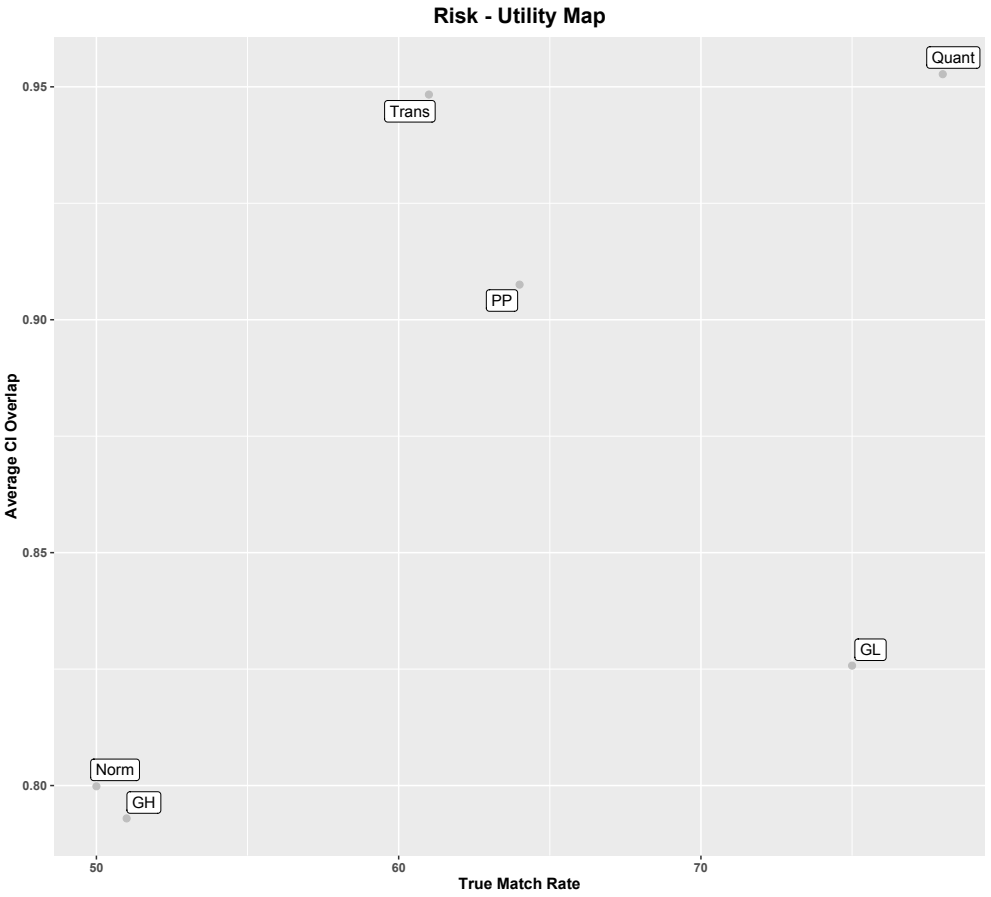

Figure 4.7: Plot of data utility (CI overlap) against disclosure risks (True match rate) for various synthesis models.

dataset. These include factors such as the methodology we have proposed and the choice of variables. Nevertheless, we find that as in the simulation studies in Section 4.3, the Quant approach results in higher utility and disclosure risks. Performance from the three flexible distributions is more varied, but the GH approach does not result in improvements over Norm. In our application, the Trans approach, arguably the most popular in the literature, appears to improve results for Norm as well as provide better data utility than most of the approaches considered.

4.5 Conclusions

In this Chapter, we have taken advantage of various methodological advances in the field of Statistics to propose ideas for generating synthetic data with non normal error terms. We hope that this research will encourage agencies to look beyond simple linear regression with normal errors, to alternatives that may better suit datasets with non normal distributions.

Overall, we conclude that the quantile approach to synthesis can result in excellent data replication for a variety of distributions, providing high utility, but also posing high disclosure risks. This is the first time that quantile regression has been used to partially synthesise data. The method performed well for both non-hierarchical and hierarchical data settings. On some occasions, the quantile regression approach may result in reduced variability in the released data; in the future, this may be remedied by changing the number of copies of synthetic data released or utilising a different approach to fitting quantile regressions to data noted earlier.

Our study of transformations showed significant improvements over the use of normal distribution to synthesise non normal residuals. For the real data illustration, the Box-Cox transformations resulted in excellent data utility. However, our studies illustrated that transformations may still not preserve the shape of the data as well as other flexible distributions. Most importantly, in our simulation studies, we discovered that the use of transformations resulted in biased slope estimates for certain quantile regressions. If the quantile slopes are the key estimands of interest, a preferred approach may be to separate the synthesis of residuals from the estimation of the fitted model. This is the approach we used for all other strategies of parametric form, where the slopes were generally well-preserved. We recommend that whenever the transformation approach is used for data synthesis, the data keeper must also check that the regression estimates are not biased because the parameters are preserved on a transformed scale.

We also investigated the use of flexible distributions to synthesise non normal residuals. In our experience, the GH approach generally performed poorly, and may be improved by adding interventions during the synthesis process. The GL and PP approaches provided varied results over the different scenarios. Overall, the PP approach appeared to be more promising, as the GL approach provided biased results where error distributions were generated through either the normal or uniform distribution. In our setup, we assumed that the error distribution is independent of the distribution of other variables in the model. As this may not hold true in practice, additional strategies may need to be employed to satisfy our assumption.

As there are several competing fitting methods for the GH, GL and PP approaches, we note that our results may not have fully utilised the potential of these methods and further improvements can be made. Nevertheless, we have shown that it is possible to model data using flexible parametric forms of distributions and preserve its shape in the synthetic copies. We also note that each of the three methods proposed have particular bounds on the skewness-kurtosis plane, as well as varying support for the distributions. This implies that when used in scenarios where they are most appropriate, they may have potential for improved results. In our hierarchical data simulations, some of these approaches resulted in better data utility when quantile regression did not perform as well.

We also note the limitations implied by the choice of the data generation mechanisms. Firstly, in some data, the covariates may explain the variation in the data more strongly than in our simulation studies, and this would reduce the affect of misspecifying the shape of the error distribution. Secondly, the results may also change with varying data sizes, studies related to very small or very large data sizes may result in different conclusions. Moreover, non-normal error distributions with more or less extreme tails and various shapes may also result in different consequences.

This is also the first time that congeniality to quantile regressions have been studied for MI for synthetic data. We appreciate that such fine level of preservation may not only be difficult to achieve through the synthesis models, but also pose high risks. Where risks for units are high, it may actually be desirable to utilise certain parametric methods to control the risks for unusual observations. Nevertheless, as our studies show, to preserve estimates for quantile regressions, data keepers may use quantile regressions to synthesise confidential data.

In this Chapter, we have also proposed the use of multivariate distributions to synthesise errors correlated within repeated measures or time points. We conclude that the method generally performs well where the choice of distribution is suited to the data, as the simulations for hierarchical data verified. In the future, the method may be used as an alternative to random effects models for synthesis of the error terms.

Chapter 5

Conclusions and Discussion

In this thesis, we explored models for multiple imputation (MI) for hierarchical data. We aimed to find imputation/synthesis models that may suit a variety of analyses. Although MI was designed to separate the imputer from the analyst, it is not often utilised as such, as the analysts usually impute data for their own specific purposes. For synthesis of confidential data, this feature is not an option, but in fact a necessity. The data keeper has to be mindful of the fact that any possible analysis procedure may be applied to the synthetic data. As such, congeniality (Meng, 1994) is central to the synthesis process, as the analysts are completely dependent on the modelling strategies of the agency.

We investigated two different ways in which synthesis models may be misspecified, either through omitted variables or the shape of the distribution of the error term. In our limited investigations, we discovered that certain models may always perform as well as, or better than others. However, data synthesis is a challenging area, the demands for which can vary considerably from one data set to another. Therefore, it is difficult to dispense general advice for data keepers about specific modelling strategies. Our conclusions, at best, identify areas of caution and methods to avoid certain types of misspecification through alternative modelling techniques.

5.1 Summary of conclusions

In the first part of the thesis, we studied biases in the linear mean specification of the regression model. We discovered that it is possible to impute or synthesise data such that both FE and RE types of analyses may be conducted after MI. The HYB model provided this balance. Our simulation studies showed that HYB preserved estimates for both types of analyses, resulted in appropriate estimates for various variance parameters, and did not allow omitted variable bias to be propagated through the imputation/synthesis procedure. Models such as the WIDE approach and DIFF approaches performed

equally well. Although we suspect that the performance of WIDE will deteriorate with increasing number of variables. Our proposed approach DIFF performed well in a variety of scenarios and is an interesting model to explore, especially to control disclosure risks for the release of sensitive data.

In the second part of the thesis, we focused on the synthesis of the residuals. We set up simulation studies with a focus on finer analyses such as quantile regressions at the extreme quantiles. Preserving these analyses through a synthesis procedure was expected to be a challenging task. We found that by using quantile regressions, it is possible to observe high data utility for such analyses but with high disclosure risks. The use of flexible parametric approaches was also tested and we did not find any of our chosen families of distributions or transformations universally better performing than others. However, we note that different fitting methods may result in improved results. The Box-Cox transformations also resulted in comparable results, and the use of normal distribution in our setup was not suitable.

5.2 Future work

Below, we identify some possible extensions to our research:

- Throughout this thesis, we mainly used models of linear form. The FE versus RE comparison may result in different observations for nonlinear models. For such models, we expect that performance for the HYB and WIDE approaches will also vary. An extension of the DIFF approach to non-linear forms will also be required for comparison.

- Our real data application involved a few interaction terms that we did not model during data synthesis. This is usual in practice. A more detailed study on modelling and preservation of inferences on interactions is required. Realistically, not all interactions may be modelled during imputation. In such a situation, the modelling of error terms may also play a significant role in preservation of certain inferences.

- In Chapters 3 and 4, we introduced new modelling techniques to the synthetic data literature, namely DIFF2 and the Quant approach. Both of these methods excelled in comparison to other models in the studies for data utility. However, as expected, they also resulted in high disclosure risks. We maintain that in most real data applications, it is unlikely that only one variable is synthesised. A natural extension to our methods is to test their performance while synthesising multivariate data using methods such as Sequential Regression Multiple Imputation (SRMI, Raghunathan et al. (2001)). This may reduce disclosure risks associated with the

two models and also assist development of methods for high quality synthetic data for more realistic applications.

- Finally, striking the balance between disclosure risks and data utility means that not all types of analyses can be preserved when generating synthetic data. It would be useful to develop a utility measure that represents the level of utility preserved by the data keepers. For example, it may be possible to provide analysts an indication of which quantiles of data have been covered by a quantile regression synthesis procedure. As data in the extreme quantiles is considered more sensitive, they may have been synthesised using a different synthesis model.

Throughout this thesis, we approached synthetic data generation with the aim of preserving a variety of analyses. An important discussion is whether this is a realistic and practical aim for release of synthetic data. In consideration for disclosure risks, it is not necessary that the synthetic data may be used to perfectly replicate all levels of analyses for all users. For analysts who treat synthetic data as test data, before carrying out analyses on the real data, this may not be a problem. However, SDC methods have been historically popularised as an alternative for researchers who are unable to access the real data. We believe, future research in the field must address the expectations of users of synthetic data, and contribute towards providing a common aim for generation of synthetic data across the globe.

Appendix A

Additional results for missing data simulation study

Imputation Model		Analysis Model									
		ICC = 0.5					ICC = 0.06				
		FE	IGN	IGN2	RE	HYB	FE	IGN	IGN2	RE	HYB
ORIG	Bias	0.03	0.00	0.03	0.02	0.03	0.03	0.01	0.03	0.01	0.03
	VR	0.96	0.50	1.92	0.98	0.96	0.96	0.90	1.02	1.00	0.96
	Len	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Cov	94.48	83.04	99.28	94.72	94.48	94.48	93.84	95.16	95.04	94.48
FE	Bias	0.07	0.02	0.07	0.06	0.07	0.06	0.03	0.06	0.05	0.06
	VR	0.95	0.86	1.27	0.95	0.95	0.95	1.24	1.14	0.97	0.95
	Len	2.90	1.81	2.27	2.85	2.90	2.89	2.29	2.98	3.47	2.89
	Cov	94.84	93.92	97.52	95.08	94.84	95.32	97.28	96.88	95.32	95.32
IGN	Bias	0.02	0.01	0.02	0.01	0.02	0.03	0.02	0.03	0.02	0.03
	VR	1.50	0.73	1.53	0.76	1.50	1.54	0.93	1.55	0.93	1.54
	Len	2.12	1.71	1.51	1.73	2.12	1.56	1.72	1.52	1.64	1.56
	Cov	98.44	91.68	98.44	92.04	98.44	98.60	95.20	98.60	95.28	98.60
IGN2	Bias	0.01	0.01	0.01	0.01	0.01	0.06	0.01	0.06	0.01	0.06
	VR	1.04	0.74	1.05	0.80	1.04	0.91	0.91	0.91	0.91	0.91
	Len	2.68	1.71	1.90	1.74	2.68	1.96	1.71	1.90	1.64	1.96
	Cov	95.64	92.60	95.72	93.48	95.64	94.48	94.00	94.48	94.00	94.48
RE	Bias	0.01	0.01	0.01	0.01	0.01	0.03	0.02	0.03	0.02	0.03
	VR	1.15	0.62	1.54	0.90	1.15	1.49	0.90	1.55	0.96	1.49
	Len	1.84	1.49	1.47	1.75	1.84	1.54	1.70	1.53	1.67	1.54
	Cov	96.12	89.16	98.48	94.00	96.12	98.28	94.88	98.56	95.64	98.28
BG	Bias	0.04	0.01	0.04	0.03	0.04	0.03	0.02	0.03	0.02	0.03
	VR	0.78	0.63	0.98	0.80	0.78	0.72	0.88	0.74	0.90	0.72
	Len	2.18	1.50	1.68	1.96	2.18	1.82	1.69	1.79	1.68	1.82
	Cov	92.36	89.32	94.72	92.16	92.36	91.84	94.44	92.64	94.40	91.84
HYB	Bias	0.03	0.00	0.03	0.02	0.03	0.03	0.01	0.03	0.02	0.03
	VR	0.90	0.64	1.10	0.89	0.90	0.90	0.89	0.93	0.94	0.90
	Len	2.29	1.49	1.75	2.03	2.29	1.94	1.69	1.91	1.70	1.94
	Cov	94.44	89.88	96.44	94.04	94.44	94.00	94.12	94.36	94.56	94.00
WIDE	Bias	0.12	0.10	0.12	0.11	0.12	0.07	0.09	0.07	0.09	0.07
	VR	0.96	0.92	0.99	0.95	0.96	0.97	0.97	0.97	0.97	0.97
	Len	5.79	5.97	4.13	6.00	5.79	5.11	6.86	4.96	6.55	5.11
	Cov	94.28	94.00	94.56	94.56	94.28	95.48	95.52	95.48	95.56	95.48
CC	Bias	0.06	0.00	0.06	0.01	0.06	0.05	0.01	0.05	0.02	0.05
	VR	1.02	0.76	2.03	0.96	1.02	1.01	0.96	1.07	0.96	1.06
	Len	2.55	1.59	2.55	1.63	2.55	2.55	1.59	2.55	1.53	2.61
	Cov	95.48	92.00	99.36	94.52	95.48	94.80	94.20	95.64	94.40	95.48

Table A.1: Properties of $\hat{\beta}$ over 2500 datasets, Case 1, Large and Small ICC, 70% data missing; sample size is 5000, 1000 clusters, 5 observations per cluster.

Imputation Model		Analysis Model									
		ICC = 0.5					ICC = 0.06				
		FE	IGN	IGN2	RE	HYB	FE	IGN	IGN2	RE	HYB
ORIG	σ_e^2	4.00	7.99	7.99	4.00	4.00	4.00	4.25	4.25	4.00	4.00
	σ_b^2	-	-	-	3.99	3.99	-	-	-	0.25	0.25
	ICC	-	-	-	0.50	0.50	-	-	-	0.06	0.06
FE	σ_e^2	4.01	12.99	12.98	4.01	4.01	4.01	9.25	9.24	4.01	4.01
	σ_b^2	-	-	-	8.99	8.98	-	-	-	5.25	5.23
	ICC	-	-	-	0.69	0.69	-	-	-	0.57	0.57
IGN	σ_e^2	7.70	8.00	8.00	7.72	7.72	4.24	4.25	4.25	4.25	4.25
	σ_b^2	-	-	-	0.27	0.27	-	-	-	0.00	0.00
	ICC	-	-	-	0.03	0.03	-	-	-	0.00	0.00
IGN2	σ_e^2	7.71	8.01	8.00	7.73	7.73	4.24	4.26	4.25	4.26	4.25
	σ_b^2	-	-	-	0.28	0.27	-	-	-	0.00	0.00
	ICC	-	-	-	0.03	0.03	-	-	-	0.00	0.00
RE	σ_e^2	4.01	8.00	8.00	4.01	4.01	3.86	4.28	4.28	3.86	3.86
	σ_b^2	-	-	-	3.99	3.99	-	-	-	0.41	0.41
	ICC	-	-	-	0.50	0.50	-	-	-	0.10	0.10
HYB	σ_e^2	4.02	8.01	8.00	4.02	4.02	3.86	4.28	4.28	3.87	3.86
	σ_b^2	-	-	-	4.00	3.99	-	-	-	0.42	0.41
	ICC	-	-	-	0.50	0.50	-	-	-	0.10	0.10
BG	σ_e^2	4.02	8.01	8.00	4.02	4.02	3.86	4.28	4.28	3.87	3.86
	σ_b^2	-	-	-	3.99	3.98	-	-	-	0.42	0.41
	ICC	-	-	-	0.50	0.50	-	-	-	0.10	0.10
WIDE	σ_e^2	5.99	9.74	9.72	5.99	5.99	5.53	5.60	5.60	5.45	5.44
	σ_b^2	-	-	-	3.76	3.74	-	-	-	0.16	0.15
	ICC	-	-	-	0.39	0.39	-	-	-	0.03	0.03
CC	σ_e^2	4.00	7.99	8.00	4.00	4.00	4.00	4.25	4.25	4.19	4.19
	σ_b^2	-	-	-	4.00	4.00	-	-	-	0.06	0.06
	ICC	-	-	-	0.50	0.50	-	-	-	0.01	0.01

Table A.2: Average of $\hat{\sigma}_e^2$ and $\hat{\sigma}_b^2$ over 2500 datasets and resulting ICC, Case 1, Large and Small ICC, 70% data missing; sample size is 5000, 1000 clusters, 5 observations per cluster.

Imputation Model		Analysis Model									
		ICC = 0.5					ICC = 0.06				
		FE	IGN	IGN2	RE	HYB	FE	IGN	IGN2	RE	HYB
ORIG	Bias	0.03	93.30	0.03	5.19	0.03	0.03	93.30	0.03	5.83	0.03
	VR	0.96	0.38	3.32	0.89	0.96	0.96	0.39	2.43	0.87	0.96
	Len	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Cov	94.48	0.00	99.92	0.24	94.48	94.48	0.00	99.72	0.04	94.48
FE	Bias	0.08	93.31	0.08	4.63	0.08	0.03	93.28	0.03	5.06	0.03
	VR	0.90	0.52	1.44	0.90	0.90	0.93	0.56	1.33	0.92	0.93
	Len	2.80	1.31	1.79	2.87	2.80	2.78	1.36	2.00	2.87	2.78
	Cov	94.60	0.00	98.36	64.68	94.60	94.60	0.00	97.56	60.32	94.60
IGN	Bias	64.79	93.35	64.79	90.43	64.79	64.78	93.31	64.78	90.72	64.78
	VR	1.48	0.64	1.50	0.74	1.48	1.41	0.60	1.43	0.71	1.41
	Len	3.68	1.74	2.00	2.89	3.69	3.36	1.72	2.13	2.66	3.37
	Cov	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IGN2	Bias	0.02	93.31	0.02	24.39	0.02	0.04	93.30	0.04	17.95	0.04
	VR	1.19	0.57	1.21	0.96	1.19	1.13	0.55	1.15	0.97	1.13
	Len	3.47	1.44	1.88	4.15	3.47	2.98	1.39	1.88	3.41	2.98
	Cov	96.16	0.00	96.24	0.04	96.16	96.84	0.00	96.96	0.08	96.84
RE	Bias	34.70	96.85	34.70	41.79	34.70	42.07	96.98	42.07	51.41	42.07
	VR	0.33	0.44	0.47	0.29	0.33	0.33	0.50	0.42	0.27	0.33
	Len	3.05	1.26	1.87	3.31	3.05	3.15	1.34	2.17	3.49	3.15
	Cov	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
HYB	Bias	0.45	93.36	0.45	5.65	0.45	0.13	93.29	0.13	5.97	0.13
	VR	0.56	0.44	0.94	0.56	0.56	0.55	0.47	0.81	0.54	0.55
	Len	2.08	1.19	1.38	2.16	2.08	1.96	1.21	1.44	2.05	1.96
	Cov	87.48	0.00	94.56	32.64	87.48	87.36	0.00	92.80	26.00	87.36
BG	Bias	0.05	93.30	0.05	5.24	0.05	0.04	93.28	0.04	5.89	0.04
	VR	0.89	0.45	1.34	0.87	0.89	0.89	0.47	1.19	0.87	0.89
	Len	2.45	1.20	1.54	2.54	2.45	2.33	1.22	1.64	2.45	2.33
	Cov	94.04	0.00	97.80	47.76	94.04	94.24	0.00	96.44	36.32	94.24
WIDE	Bias	0.00	93.34	0.00	7.21	0.00	0.05	93.35	0.05	8.13	0.05
	VR	0.95	0.81	1.06	0.95	0.95	0.99	0.86	1.07	1.00	0.99
	Len	5.12	3.07	2.84	5.36	5.12	5.04	3.26	3.24	5.35	5.04
	Cov	95.28	0.00	96.04	76.52	95.28	95.36	0.00	95.92	69.52	95.36
CC	Bias	0.08	93.34	0.08	41.73	0.08	0.04	93.33	0.04	51.27	0.04
	VR	0.97	0.64	4.92	0.22	0.97	0.98	0.64	4.07	0.18	0.99
	Len	2.48	1.60	3.01	2.31	2.49	2.48	1.60	3.18	2.27	2.50
	Cov	94.56	0.00	99.96	0.00	94.64	94.40	0.00	100.00	0.00	94.64

Table A.3: Properties of $\hat{\beta}$, Case 2, Large and Small ICC, 70% data missing; sample size is 5000, 1000 clusters, 5 observations per cluster.

Imputation Model		Analysis Model									
		ICC = 0.5					ICC = 0.06				
		FE	IGN	IGN2	RE	HYB	FE	IGN	IGN2	RE	HYB
ORIG	σ_e^2	4.00	24.31	13.85	4.02	4.00	4.00	20.56	10.11	4.03	4.00
	σ_b^2	-	-	-	34.30	9.86	-	-	-	34.30	9.86
	ICC	-	-	-	0.90	0.71	-	-	-	0.89	0.71
FE	σ_e^2	4.01	29.14	18.69	4.03	4.01	4.01	25.41	14.95	4.04	4.01
	σ_b^2	-	-	-	39.33	14.70	-	-	-	39.33	14.70
	ICC	-	-	-	0.91	0.79	-	-	-	0.91	0.79
IGN	σ_e^2	22.44	24.33	23.35	23.03	22.48	18.98	20.57	19.59	19.60	19.07
	σ_b^2	-	-	-	1.31	0.87	-	-	-	1.31	0.87
	ICC	-	-	-	0.05	0.04	-	-	-	0.06	0.04
IGN2	σ_e^2	13.08	24.34	13.86	13.62	13.08	9.63	20.58	10.12	9.92	9.63
	σ_b^2	-	-	-	19.35	0.78	-	-	-	19.35	0.78
	ICC	-	-	-	0.59	0.06	-	-	-	0.66	0.07
RE	σ_e^2	5.77	22.89	18.22	5.81	5.77	6.63	19.23	15.57	6.71	6.63
	σ_b^2	-	-	-	22.60	12.47	-	-	-	22.60	12.47
	ICC	-	-	-	0.80	0.68	-	-	-	0.77	0.65
HYB	σ_e^2	4.01	24.23	13.85	4.04	4.01	4.01	20.54	10.11	4.05	4.01
	σ_b^2	-	-	-	34.10	9.85	-	-	-	34.10	9.85
	ICC	-	-	-	0.89	0.71	-	-	-	0.89	0.71
BG	σ_e^2	4.01	24.31	13.85	4.03	4.01	4.02	20.56	10.11	4.05	4.02
	σ_b^2	-	-	-	34.29	9.85	-	-	-	34.29	9.85
	ICC	-	-	-	0.89	0.71	-	-	-	0.89	0.71
WIDE	σ_e^2	5.55	25.71	15.22	5.60	5.55	5.51	21.91	11.43	5.57	5.51
	σ_b^2	-	-	-	33.54	9.68	-	-	-	33.54	9.68
	ICC	-	-	-	0.86	0.64	-	-	-	0.86	00.64
CC	σ_e^2	4.00	24.30	20.32	5.78	4.03	4.00	20.55	16.56	6.66	4.05
	σ_b^2	-	-	-	22.62	16.87	-	-	-	22.62	16.87
	ICC	-	-	-	0.80	0.81	-	-	-	0.77	0.81

Table A.4: Average of $\hat{\sigma}_e^2$ and $\hat{\sigma}_b^2$ over 2500 datasets and resulting ICC, Case 2, Large and Small ICC, 70% data missing; sample size is 5000, 1000 clusters, 5 observations per cluster.

Appendix B

Additional results for synthetic data application

Variable	Synthesis selection	Article selection
Log(wages/full-time equivalent)	7.752(0.432)	7.722(0.457)
Works council (yes = 1)	0.615(-)	0.591(-)
Existence of an opening clause (yes = 1)	0.320(-)	0.313(-)
Application of an opening clause (yes = 1)	0.153(-)	0.152(-)
Existence of an opening clause and works council (yes = 1)	0.240(-)	0.229(-)
Application of an opening clause and works council (yes = 1)	0.118(-)	0.111(-)
Proportion of qualified employees	0.742(0.234)	0.734(0.239)
Proportion of employees with fixed-term contracts	0.063(0.124)	0.063(0.129)
Proportion of casual workers	0.019(0.053)	0.034(0.274)
Proportion of part-time employees	0.211(0.234)	0.212(0.235)
Proportion of trainees	0.049(0.074)	0.050(0.075)
Churning rate	0.045(0.130)	0.050(0.167)
Establishment not part of larger enterprise (Single = 1)	0.603(-)	0.600(-)
Technical state of the establishment (1 = very good, ..., 5 = bad)	2.151(0.723)	2.157(0.743)
Invested in physical capital within the previous year (Invest = 1)	0.792(-)	0.774(-)
Establishment is under foreign ownership (Foreign = 1)	0.074(-)	0.076(-)
5 - 9 employees	0.116(-)	0.117(-)
10 - 19 employees	0.114(-)	0.117(-)
20 - 49 employees	0.167(-)	0.174(-)
50 - 99 employees	0.150(-)	0.142(-)
100 - 199 employees	0.141(-)	0.134(-)
200 - 499 employees	0.166(-)	0.163(-)
500 - 999 employees	0.083(-)	0.078(-)
1000 - 4999 employees	0.059(-)	0.060(-)
5000 or more employees	0.004(-)	0.005(-)
Number of observations	5195	8288

Table B.1: Sample description - comparison of mean and standard deviation of key variables in the synthesis selection and article selection of data. Results from article selection extracted from the article, Table 3, page 103

Model	WOCO	SE	OC	SE	OCxWOCO	SE	OC2	SE	OC2xWOCO	SE	N	R^2
OLS1	0.203***	0.014	0.066***	0.017	-0.012	0.020	-	-	-	-	8288	0.447
	0.192***	0.018	0.077***	0.022	-0.022	0.025	-	-	-	-	5195	0.461
OLS2	0.205***	0.014	0.108***	0.023	-0.048*	0.025	-0.086***	0.030	0.073**	0.033	8288	0.447
	0.193***	0.018	0.104***	0.028	-0.043	0.032	-0.059	0.038	0.048	0.041	5195	0.461
OLS2a	0.165***	0.019	0.109***	0.029	-0.084**	0.039	-0.115***	0.040	0.121**	0.050	2560	0.382
	0.146***	0.024	0.108***	0.036	-0.068	0.047	-0.076*	0.046	0.086	0.057	1612	0.411
OLS2b	0.211***	0.015	0.107***	0.023	-0.052**	0.026	-0.090***	0.030	0.068**	0.033	7098	0.473
	0.205***	0.019	0.113***	0.028	-0.057*	0.032	-0.079**	0.038	0.064	0.040	4516	0.486
OLS2c	0.196***	0.017	0.098***	0.030	-0.038	0.033	-0.072*	0.041	0.053	0.044	4427	0.449
	0.183***	0.020	0.071**	0.034	-0.018	0.040	-0.019	0.047	0.013	0.053	2612	0.455
OLS2d	0.213***	0.019	0.118***	0.032	-0.059*	0.036	-0.100**	0.042	0.094**	0.045	3861	0.451
	0.208***	0.021	0.131***	0.032	-0.065*	0.038	-0.081*	0.044	0.071	0.051	2583	0.472
OLS2e	0.262***	0.026	0.145***	0.047	-0.052	0.052	-0.085	0.060	0.055	0.066	2399	0.458
	0.280***	0.033	0.141**	0.063	-0.052	0.068	-0.064	0.084	0.059	0.088	1507	0.489
OLS2f	0.170***	0.016	0.079***	0.025	-0.033	0.028	-0.081**	0.035	0.064*	0.037	5889	0.452
	0.143***	0.020	0.077**	0.031	-0.030	0.035	-0.061	0.042	0.038	0.045	3688	0.469
FE	-0.011	0.026	-0.003	0.023	-0.033	0.028	-0.003	0.031	0.037	0.036	4410	0.086
	-0.007	0.025	-0.001	0.022	-0.036	0.027	-0.010	0.031	0.042	0.035	4518	-
RE (Bal.)	0.159***	0.017	0.061***	0.021	-0.048*	0.025	-0.046	0.029	0.060*	0.033	4410	0.435
	0.158***	0.016	0.064***	0.021	-0.049**	0.025	-0.047	0.029	0.060*	0.033	4518	-
RE (Unbal.)	0.191***	0.013	0.064***	0.017	-0.044**	0.021	-0.057**	0.023	0.068**	0.027	8288	0.443
	0.176***	0.015	0.053***	0.020	-0.037	0.024	-0.037	0.027	0.051	0.031	5195	-

Table B.2: Key model coefficient and standard error estimates for all the analysis models. Original article results (black) and results from our selection (blue) are presented along with indication for the significance levels of 1% (***), 5%(**) and 10%(*). Results from article selection extracted from the article, Table 4, page 104 and Table A2, page 113.

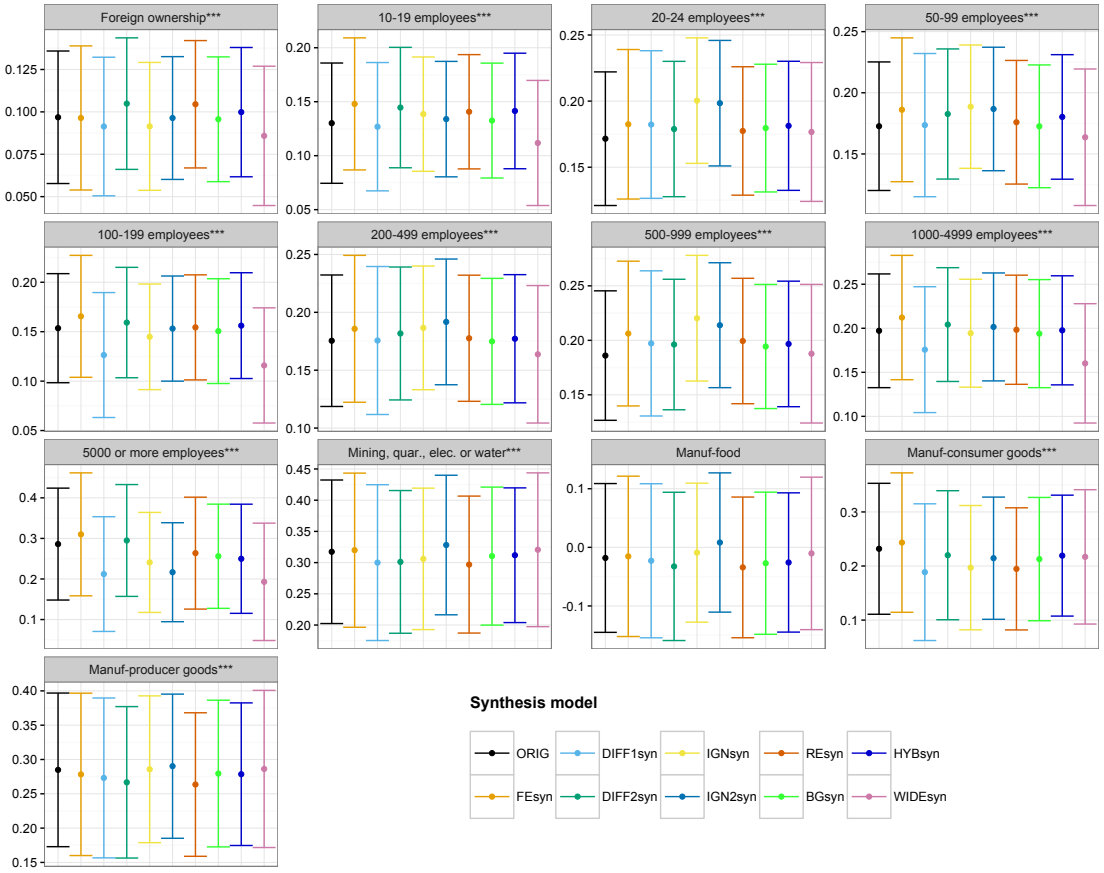

Figure B.1: Set 2 of the OLS1 coefficient estimates and confidence intervals for both original and synthetic data.

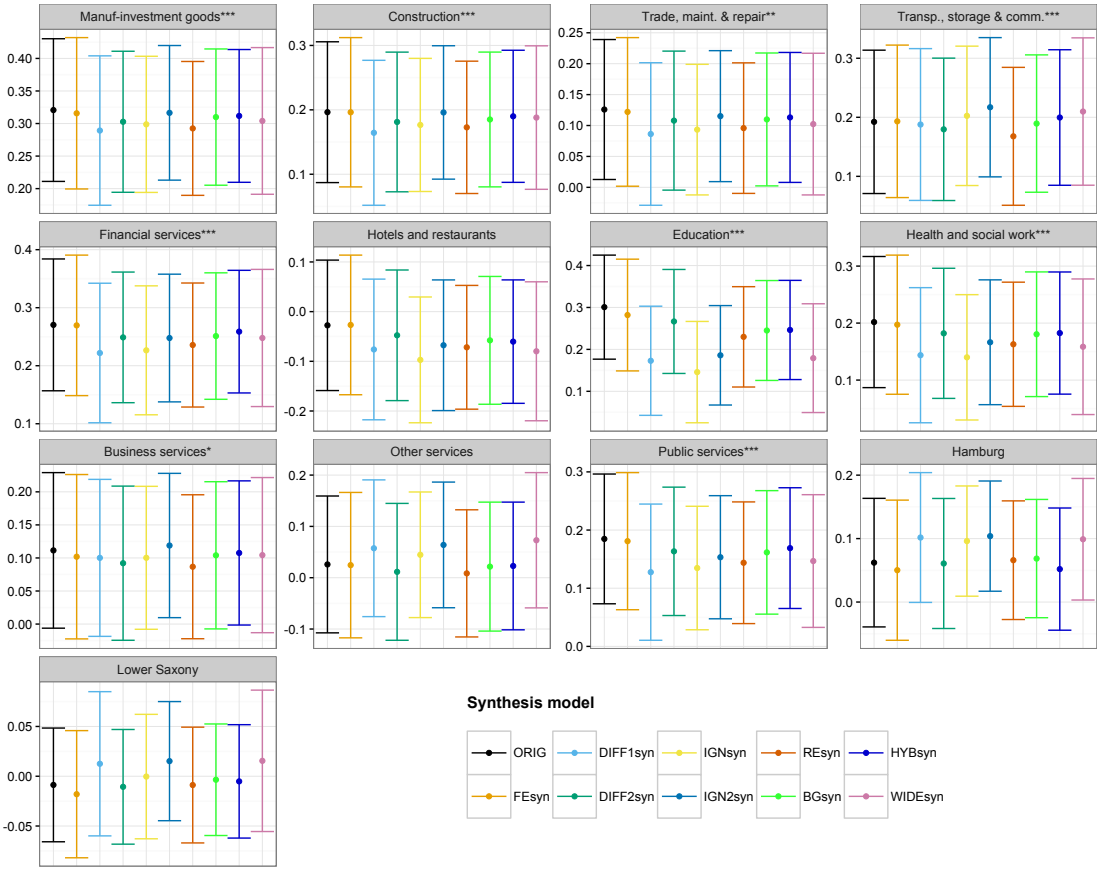

Figure B.2: Set 3 of the OLS1 coefficient estimates and confidence intervals for both original and synthetic data.

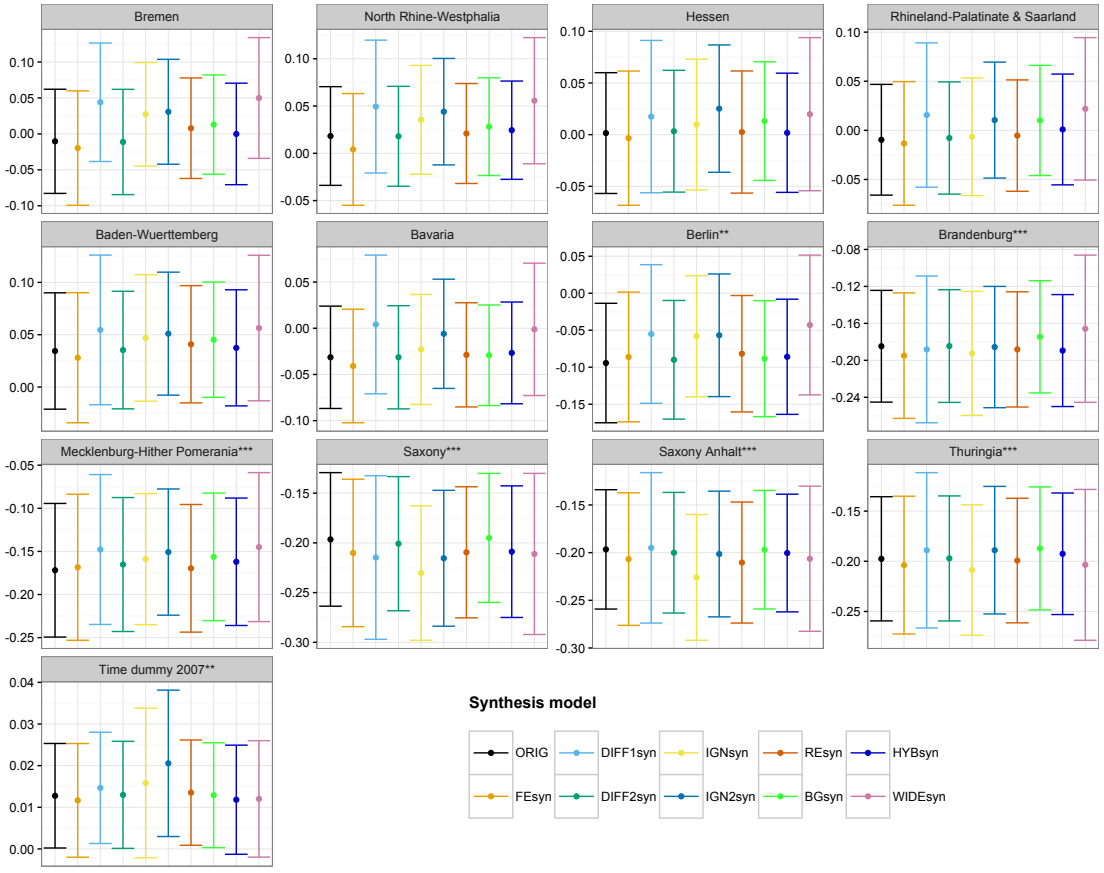

Figure B.3: Set 4 of the OLS1 coefficient estimates and confidence intervals for both original and synthetic data.

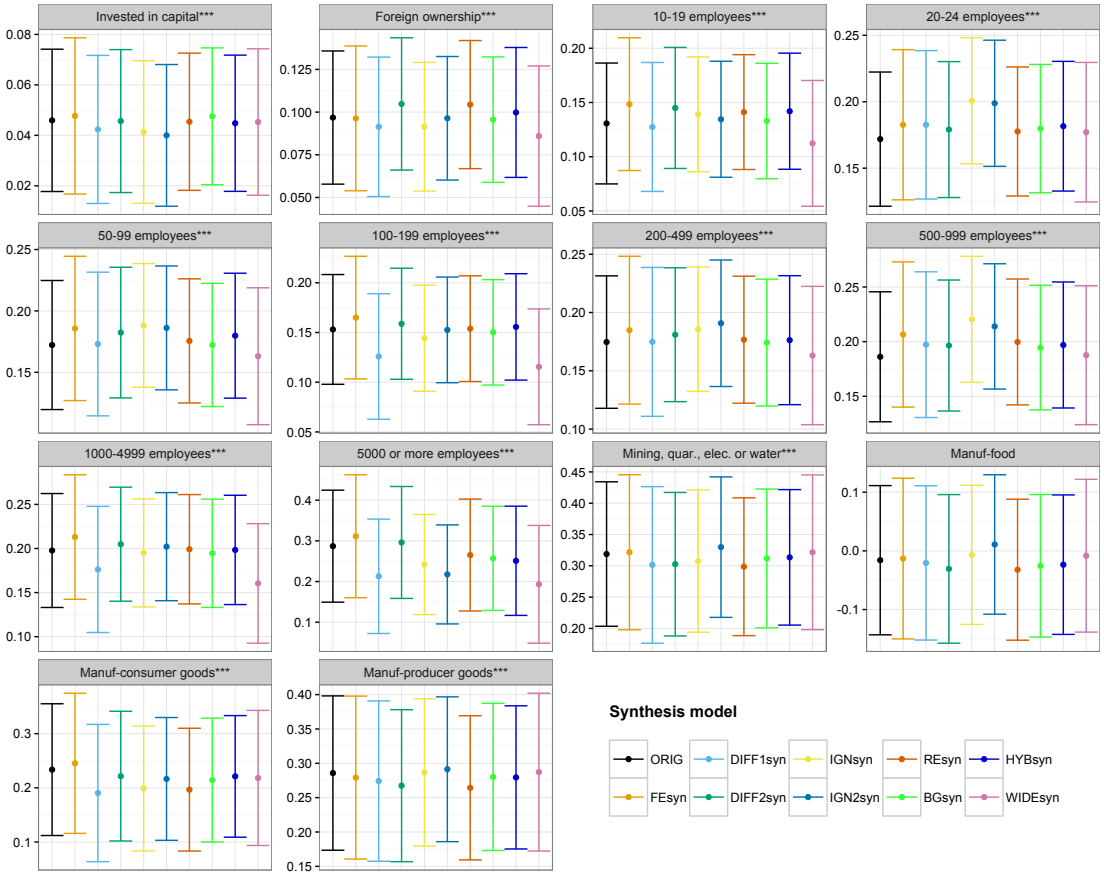

Figure B.4: Set 2 of the OLS2 coefficient estimates and confidence intervals for both original and synthetic data.

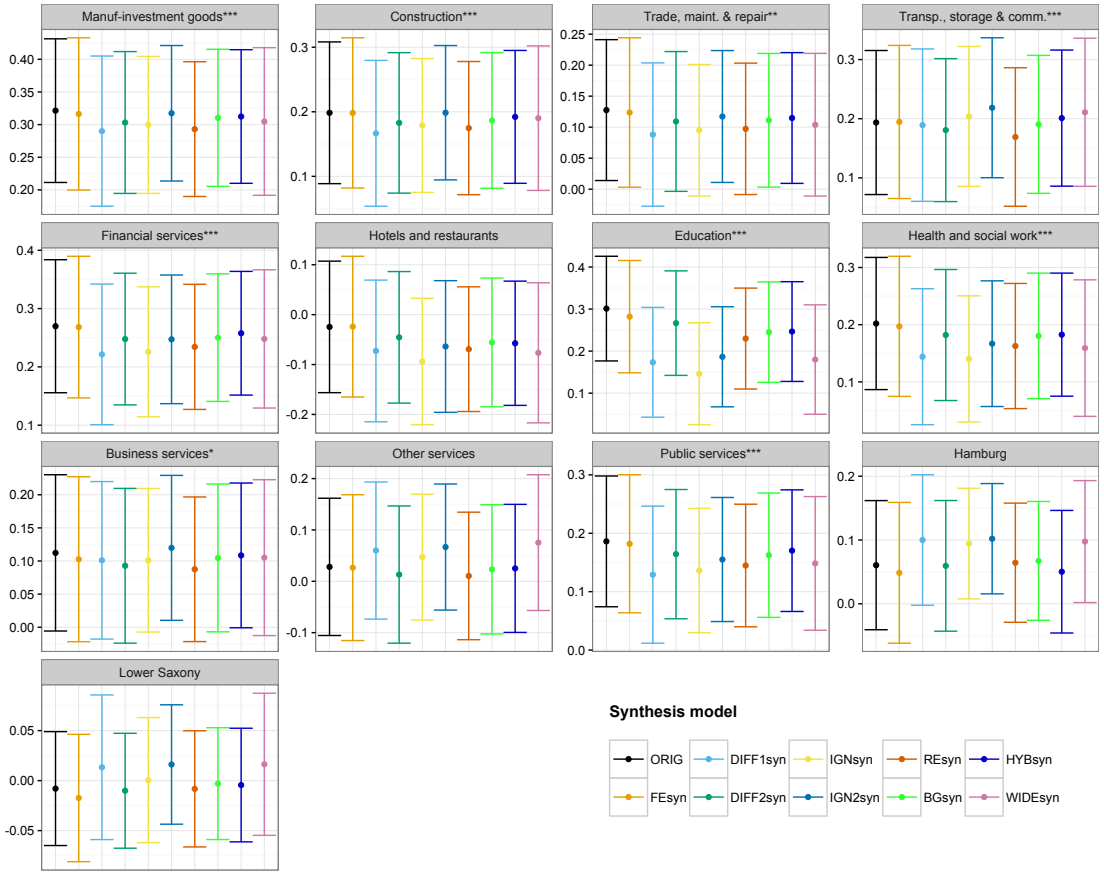

Figure B.5: Set 3 of the OLS2 coefficient estimates and confidence intervals for both original and synthetic data.

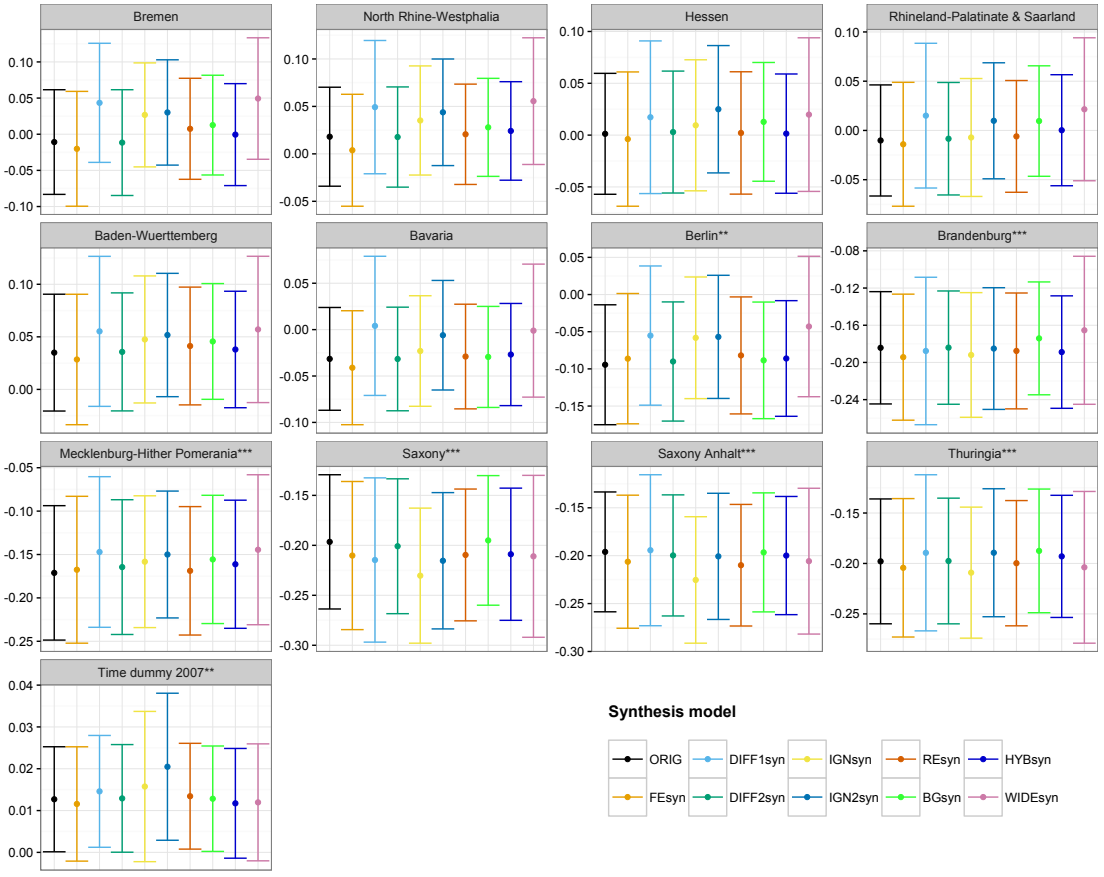

Figure B.6: Set 4 of the OLS2 coefficient estimates and confidence intervals for both original and synthetic data.

References

Abowd, J. M., Stinson, M., and Benedetto, G. (2006). Final report to the Social Security Administration on the SIPP/SSA/IRS public use file project. *Suitland, MD: Census Bureau, Longitudinal Employer-Household Dynamics Program.*

Abowd, J. M. and Vilhuber, L. (2008). *How Protective Are Synthetic Data?*, pages 239–246. Springer Berlin Heidelberg, Berlin, Heidelberg.

Agresti, A., Caffo, B., and Ohman-Strickland, P. (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics & Data Analysis*, 47(3):639–653.

Allison, P. D. (2001). *Missing data*, volume 136. Sage publications.

Allison, P. D. (2009). *Fixed Effects Regression Models*. SAGE, United States of America.

Andridge, R. R. (2011). Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biometrical Journal*, 53:57–74.

Arellano-Valle, R., Bolfarine, H., and Lachos, V. (2005). Skew-normal linear mixed models. *Journal of Data Science*, 3(4):415–438.

Bafumi, J. and Gelman, A. E. (2006). Fitting multilevel models when predictors and group effects correlate. *Annual Meeting of the Midwest Political Science Association, Chicago, Il.*

Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., and Talwar, K. (2007). Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 273–282. ACM.

Barnard, J. and Rubin, D. B. (1999). Miscellanea. small-sample degrees of freedom with multiple imputation. *Biometrika*, 86(4):948–955.

Bartlett, J. (2016). Combining bootstrapping with multiple imputation. [Online; accessed 28-June-2016].

Bartlett, J. W., Carpenter, J. R., Tilling, K., and Vansteelandt, S. (2014). Improving upon the efficiency of complete case analysis when covariates are mnar. *Biostatistics*, 15(4):719–730.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Beck, N. and Katz, J. N. (2001). Throwing out the baby with the bath water: A comment on Green, Kim, and Yoon. *International Organization*, 55(02):487–495.

Bethlehem, J. G., Keller, W. J., and Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, 85(409):38–45.

Blien, U., Wirth, H., and Muller, M. (1992). Disclosure risk for microdata stemming from official statistics. *Statistica Neerlandica*, 46(1):69–82.

Bottai, M. and Zhen, H. (2013). Multiple imputation based on conditional quantile estimation. *Epidemiology, Biostatistics and Public Health*, 10(1).

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252.

Box, G. E. and Tidwell, P. W. (1962). Transformation of the independent variables. *Technometrics*, 4(4):531–550.

Browne, W. (2009). *MCMC Estimation in MLwiN v2.1.* Centre for Multilevel Modelling, University of Bristol.

Butler, S. M. and Louis, T. A. (1992). Random effects models with non-parametric priors. *Statistics in Medicine*, 11(14-15):1981–2000.

Caffo, B., An, M.-W., and Rohde, C. (2007). Flexible random intercept models for binary outcomes using mixtures of normals. *Computational Statistics & Data Analysis*, 51(11):5220–5235.

Cario, M. C. and Nelson, B. L. (1997). Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Technical report, Department of Industrial Engineering and Management Sciences, Northwestern University.

Charest, A.-S. (2010). How can we analyze differentially-private synthetic datasets? *Journal of Privacy and Confidentiality*, 2(2):21–33.

Chen, G. and Keller-McNulty, S. (1998). Estimation of identification disclosure risk in microdata. *Journal of Official Statistics-Stockholm*, 14:79–95.

Chen, S. (2014). *Imputation of missing values using quantile regression.* PhD thesis, Iowa State University.

Chinchilli, V. M., Esinhart, J. D., and Miller, W. G. (1995). Partial likelihood analysis of within-unit variances in repeated measurement experiments. *Biometrics*, pages 205–216.

Clark, T. S. and Linzer, D. A. (2015). Should I use fixed or random effects? *Political Science Research and Methods*, 3(02):399–408.

Clarke, K. A. (2005). The phantom menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science*, 22(4):341–352.

Cleveland, L., McCaa, R., Ruggles, S., and Sobek, M. (2012). When excessive perturbation goes wrong and why IPUMS-International relies instead on sampling, suppression, swapping, and other minimally harmful methods to protect privacy of census microdata. In *Privacy in Statistical Databases*, pages 179–187. Springer.

Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B., and Weidman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using bayesian logistic regression. *Journal of the American Statistical Association*, 86(413):68–78.

Coffey, C., Carlin, J., Lynskey, M., Ning, L., and Patton, G. (2003). Adolescent precursors of cannabis dependence: findings from the victorian adolescent health cohort study. *The British Journal of Psychiatry*, 182:330 – 336.

Cole, T. (1988). Fitting smoothed centile curves to reference data. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 385–418.

Cole, T. J. and Green, P. J. (1992). Smoothing reference centile curves: the lms method and penalized likelihood. *Statistics in Medicine*, 11(10):1305–1319.

Cook, R. D. and Weisberg, S. (1994). Transforming a response variable for linearity. *Biometrika*, 81(4):731–737.

Cox, L., Fagan, J., Greenberg, B., and Hemmig, R. (1986a). Research at the census bureau into disclosure avoidance techniques for tabular data. *Proceedings of the Section on Survey Research Methods*, pages 388–393.

Cox, L., McDonald, S.-K., and Nelson, D. (1986b). Confidentiality issues at the united states bureau of the census. *Journal of Official Statistics*, 2:135–160.

Dale, A. and Elliot, M. (2001). Proposals for 2001 samples of anonymized records: an assessment of disclosure risk. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164(3):427–447.

Davidian, M. and Giltinan, D. M. (1993). Some simple methods for estimating intraindividual variability in nonlinear mixed effects models. *Biometrics*, pages 59–73.

de Jong, R., van Buuren, S., and Spiess, M. (2014). Multiple imputation of predictor variables using generalized additive models. *Communications in Statistics-Simulation and Computation*, 0:1–18.

Dean, B. (2013). *Improved Estimation and Regression Techniques with the Generalised Lambda Distribution*. PhD thesis, University of Newcastle.

Delanius, T. (1977). Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, pages 429–444.

Delanius, T. (1984). Access to information through censuses and surveys. Technical report, Report based on Research Project Supported by Supported by Stiftung Volkswagenwerk.

Delanius, T. and Reiss, S. (1982). Data swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6:73–85.

Demirtas, H. (2009). Multiple imputation under the generalized lambda distribution. *Journal of Biopharmaceutical Statistics*, 19(1):77–89.

Demirtas, H., Freels, S. A., and Yucel, R. M. (2008). Plausibility of multivariate normality assumption when multiply imputing non-gaussian continuous outcomes: a simulation assessment. *Journal of Statistical Computation and Simulation*, 78(1):69–84.

Demirtas, H. and Hedeker, D. (2008). Multiple imputation under power polynomials. *Communications in Statistics - Simulation and Computation*, 37(8):1682–1695.

Demirtas, H., Hedeker, D., and Mermelstein, R. J. (2012). Simulation of massive public health data by power polynomials. *Statistics in Medicine*, 31(27):3337–3346.

Domingo-Ferrer, J. (2002). *Inference control in statistical databases*. Springer.

Domingo-Ferrer, J. and Torra, V. (2003). Disclosure risk assessment in statistical microdata protection via advanced record linkage. *Statistics and Computing*, 13(4):343–354.

Drechsler, J. (2011a). Multiple imputation in practicea case study using a complex German establishment survey. *AStA Advances in Statistical Analysis*, 95(1):1–26.

Drechsler, J. (2011b). *Synthetic datasets for statistical disclosure control: theory and implementation*, volume 201. Springer.

Drechsler, J. (2012). New data dissemination approaches in old europe–synthetic datasets for a German establishment survey. *Journal of Applied Statistics*, 39(2):243–265.

Drechsler, J. (2015). Multiple imputation of multilevel missing datarigor versus simplicity. *Journal of Educational and Behavioral Statistics*, 40(1):69–95.

Drechsler, J., Bender, S., and Rässler, S. (2008a). Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel. *Transactions on Data Privacy*, 1(3):105–130.

Drechsler, J., Dundler, A., Bender, S., Rässler, S., and Zwick, T. (2008b). A new approach for disclosure control in the iab establishment panelmultiple imputation for a better data access. *AStA Advances in Statistical Analysis*, 92(4):439–458.

Drechsler, J. and Raghunathan, T. E. (2008). Evaluating different approaches for multiple imputation under linear constraints. *Invited paper for the Work Session on Statistical Data Editing 2008, United Nations Statistical Commission and Economic Commission for Europe. Vienna.*

Drechsler, J. and Reiter, J. P. (2008). Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In *Privacy in Statistical Databases*, pages 227–238. Springer.

Drechsler, J. and Reiter, J. P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, 105(492):1347–1357.

Drechsler, J. and Reiter, J. P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis*, 55(12):3232–3243.

Drechsler, J. and Reiter, J. P. (2012). Combining synthetic data with subsampling to create public use microdata files for large scale surveys. *Survey Methodology*, 38:73 – 79.

Duncan, G. and Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Business & Economic Statistics*, 7(2):207–217.

Duncan, G. T. and Lambert, D. (1986). Disclosure-limited data dissemination. *Journal of the American Statistical Association*, 81(393):10–18.

Dwork, C. (2006). Differential privacy. *The 33rd International Colloquium on Automata, Languages and Programming*, pages 1–12.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pages 265–284. Springer.

Elamir, E. A. and Skinner, C. (2006). Record level measures of disclosure risk for survey microdata. *Journal of Official Statistics*, 22(3):525–539.

Ellguth, P., Gerner, H.-D., and Stegmaier, J. (2014). Wage effects of works councils and opening clauses: The German case. *Economic and Industrial Democracy*, 35(1):95–113.

Elliot, M. J. (2005). Assessment of disclosure risk for hierarchical microdata. Technical report, Office for National Statistics Microdata Release Panel.

Fabrizi, E., Giusti, C., Salvati, N., and Tzavidis, N. (2014a). Mapping average equivalized income using robust small area methods. *Papers in Regional Science*, 93(3):685–701.

Fabrizi, E., Salvati, N., Pratesi, M., and Tzavidis, N. (2014b). Outlier robust model-assisted small area estimation. *Biometrical Journal*, 56(1):157–175.

Faraway, J. J. (2002). Practical Regression and ANOVA using R.

Fienberg, S. E. (1994). A radical proposal for the provision of micro-data samples and the provision of confidentiality. Technical report, Department of Statistics, Carnegie Mellon University.

Fienberg, S. E. and Makov, U. E. (1998). Confidentiality, uniqueness and disclosure limitation for categorical data. *Journal of Official Statistics-Stockholm*, 14:385–398.

Fienberg, S. E., Makov, U. E., and Sanil, A. P. (1997). A bayesian approach to data disclosure: Optimal intruder behavior for continuous data. *Journal of Official Statistics-Stockholm*, 13:75–79.

Fienberg, S. E., Makov, U. E., and Steel, R. (1998). Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics-Stockholm*, 14:485–502.

Fischer, G., Janik, F., Müller, D., Schmucker, A., et al. (2008). The IAB establishment panel, from sample to survey to projection. *FDZ Methodenreport*, 1.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43(4):521–532.

Foulley, J., San Cristobal, M., Gianola, D., and Im, S. (1992). Marginal likelihood and bayesian approaches to the analysis of heterogeneous residual variances in mixed linear gaussian models. *Computational Statistics & Data Analysis*, 13(3):291–305.

Fung, B. C. M., Wang, K., Chen, R., and Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.*, 42(4):14:1–14:53.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian data analysis*. Chapman & Hall/CRC.

Genest, C. and MacKay, J. (1986). The joy of copulas: bivariate distributions with uniform marginals. *The American Statistician*, 40(4):280–283.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T. (2016). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0-5.

Geraci, M. (2014). Linear quantile mixed models: The lqmm package for laplace quantile regression. *Journal of Statistical Software*, 57(13):1–29.

Geraci, M. (2016). Estimation of regression quantiles in complex surveys with data missing at random: an application to birthweight determinants. *Statistical Methods in Medical Research*, 25(4):1393–1421.

Geraci, M. and Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric laplace distribution. *Biostatistics*, 8(1):140–154.

Geraci, M. and Bottai, M. (2014). Linear quantile mixed models. *Statistics and Computing*, 24(3):461–479.

Geraci, M. and Jones, M. (2015). Improved transformation-based quantile regression. *Canadian Journal of Statistics*, 43(1):118–132.

Ghidey, W., Lesaffre, E., and Eilers, P. (2004). Smooth random effects distribution in a linear mixed model. *Biometrics*, 60(4):945–953.

Ghidey, W., Lesaffre, E., and Verbeke, G. (2010). A comparison of methods for estimating the random effects distribution of a linear mixed model. *Statistical Methods in Medical Research*, 19(6):575–600.

Gilchrist, W. (2000). *Statistical modelling with quantile functions*. CRC Press.

Greenberg, B. and Zayatz, L. V. (1992). Strategies for measuring risk in public use microdata files. *Statistica Neerlandica*, 46(1):33–48.

Gurka, M. J., Edwards, L. J., Muller, K. E., and Kupper, L. L. (2006). Extending the box–cox transformation to the linear mixed model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(2):273–288.

Hasselman, B. (2016). *nleqslv: Solve Systems of Nonlinear Equations*. R package version 3.0.1.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica: Journal of the Econometric Society*, pages 1251–1271.

Hawala, S. (2008). Producing partially synthetic data to avoid disclosure. *Proceedings of the Joint Statistical Meetings*.

He, X. (1997). Quantile curves without crossing. *The American Statistician*, 51(2):186–192.

He, Y. (2005). *Multiple Imputation for Continuous Nonnormal Missing Data*. PhD thesis, University of Michigan.

He, Y. and Raghunathan, T. E. (2006). Tukey's gh distribution for multiple imputation. *The American Statistician*, 60(3).

He, Y. and Raghunathan, T. E. (2009). On the performance of sequential regression multiple imputation methods with non normal error distributions. *Communications in Statistics-Simulation and computation*, 38(4):856–883.

He, Y. and Raghunathan, T. E. (2012). Multiple imputation using multivariate gh transformations. *Journal of Applied Statistics*, 39(10):2177–2198.

Headrick, T. C. (2010). *Power Method Polynomials and Other Transformations*. Chapman & Hall/CRC Press.

Headrick, T. C. and Kowalchuk, R. K. (2007). The power method transformation: Its probability density function, distribution function, and its further use for fitting data. *Journal of Statistical Computation and Simulation*, 77(3):229–249.

Headrick, T. C. and Mugdadi, A. (2006). On simulating multivariate non-normal distributions from the generalized lambda distribution. *Computational Statistics & Data Analysis*, 50(11):3343–3353.

Headrick, T. C. and Sawilowsky, S. S. (1999). Simulating correlated multivariate nonnormal distributions: Extending the fleishman power method. *Psychometrika*, 64(1):25–35.

Headrick, T. C. and Sawilowsky, S. S. (2000). Weighted simplex procedures for determining boundary points and constants for the univariate and multivariate power methods. *Journal of Educational and Behavioral Statistics*, 25(4):417–436.

Heagerty, P. J. and Pepe, M. S. (1999). Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in us children. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(4):533–551.

Heagerty, P. J., Zeger, S. L., et al. (2000). Marginalized multilevel models and likelihood inference (with comments and a rejoinder by the authors). *Statistical Science*, 15(1):1–26.

Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distributions. In Hoaglin, D. C., Mosteller, F., and Turkey, J. W., editors, *Exploring data tables, trends, and shapes*, pages 461–513. New York, NY: John Wiley & Sons, Inc.

Hu, J., Reiter, J. P., and Wang, Q. (2014a). Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data. *arXiv preprint arXiv:1412.2282*.

Hu, Y., Zhu, Q., and Tian, M. (2014b). An effective technique of multiple imputation in nonparametric quantile regression. *Journal of Mathematics and Statistics*, 10(1):30.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., and de Wolf, P.-P. (2012). *Statistical Disclosure Control*. Wiley.com.

Jacobusse, G. W. (2005). WinMICE user's manual for WinMICE prototype version 0.1. *The Netherlands: TNO Quality of Life*.

Jewett, R. (1993). Disclosure analysis for the 1992 economic census. Technical report, US Bureau of the Census, Washington DC.

Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*, 36(1-2):149–176.

Johnson, S. G. and Narasimhan, B. (2013). *cubature: Adaptive multivariate integration over hypercubes*. R package version 1.1-2.

Jorge, M. and Boris, I. (1984). Some properties of the tukey g and h family of distributions. *Communications in Statistics-Theory and Methods*, 13(3):353–369.

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., and Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, 60(3):224–232.

Kennickell, A. B. (1997). Multiple imputation and disclosure protection: The case of the 1995 survey of consumer finances. *Record Linkage Techniques*, pages 248–267.

Kim, J. (1986). A method for limiting disclosure in microdata based on random noise and transformation. In *Proceedings of the Section on Survey Research Methods*, pages 370–374. American Statistical Association, Alexandria, VA.

Kinney, S. K. and Reiter, J. P. (2010). Tests of multivariate hypotheses when using multiple imputation for missing data and partial synthesis. *Journal of Official Statistics*, 26:301 – 315.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., and Abowd, J. M. (2011). Towards unrestricted public use business microdata: The synthetic longitudinal business database. *International Statistical Review*, 79(3):362–384.

Koenker, R. (1984). A note on l-estimates for linear models. *Statistics & Probability Letters*, 2(6):323–325.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press, New York.

Koenker, R. (2016). *quantreg: Quantile Regression*. R package version 5.24.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50.

Koenker, R. and Geling, O. (2001). Reappraising medfly longevity: a quantile regression survival analysis. *Journal of the American Statistical Association*, 96(454):458–468.

Koenker, R. and Hallock, K. (2001). Quantile regression: An introduction. *Journal of Economic Perspectives*, 15(4):43–56.

Koenker, R. and Machado, J. A. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94(448):1296–1310.

Koenker, R., Ng, P., and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, 81(4):673–680.

Kordas, G. (2006). Smoothed binary regression quantiles. *Journal of Applied Econometrics*, 21(3):387–407.

Kottas, A. and Krnjajić, M. (2009). Bayesian semiparametric modelling in quantile regression. *Scandinavian Journal of Statistics*, 36(2):297–319.

Kuo, T. C. and Headrick, T. C. (2014). Simulating univariate and multivariate tukey g-and-h distributions based on the method of percentiles. *ISRN Probability and Statistics*.

Lambert, D. (1993). Measures of disclosure risk and harm. *Journal of Official Statistics-Stockholm*, 9:313–313.

Lazzeroni, L., Schenker, N., and Taylor, J. M. (2011). A comparison of multiple-imputation procedures under model misspecification. *Department of Statistics, UCLA*.

Lee, D. and Neocleous, T. (2010). Bayesian quantile regression for count data with application to environmental epidemiology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(5):905–920.

Lee, Y., Nelder, J. A., et al. (2004). Conditional and marginal models: another view. *Statistical Science*, 19(2):219–238.

Lesaffre, E. and Molenberghs, G. (1991). Multivariate probit analysis: a neglected procedure in medical statistics. *Statistics in Medicine*, 10(9):1391–1403.

Lin, X., Raz, J., and Harlow, S. D. (1997). Linear mixed models with heterogeneous within-cluster variances. *Biometrics*, pages 910–923.

Lipsitz, S. R., Fitzmaurice, G. M., Molenberghs, G., and Zhao, L. P. (1997). Quantile regression methods for longitudinal data with drop-outs: Application to cd4 cell counts of patients infected with the human immunodeficiency virus. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(4):463–476.

Little, R. J. (1988). Missing-data adjustments in large surveys: Reply. *Journal of Business & Economic Statistics*, 6(3):300–301.

Little, R. J. (1993). Statistical analysis of masked data. *Journal of Official Statistics - Stockholm*, 9:407–407.

Liu, J., Gelman, A., Hill, J., Su, Y.-S., and Kropko, J. (2013). On the stationary distribution of iterative imputations. *Biometrika*, 101(1):155–173.

Liu, M., Taylor, J. M., and Belin, T. R. (1995). Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies. *Proceedings of the American Statistical Association, Biometrics Section*, pages 142–147.

Lo, A. Y. (1987). A large sample study of the bayesian bootstrap. *The Annals of Statistics*, 15(1):pp. 360–375.

Machado, J. A. F. and Silva, J. S. (2005). Quantiles for counts. *Journal of the American Statistical Association*, 100(472):1226–1237.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008). Privacy: Theory meets practice on the map. *IEEE 24th International Conference on Data Engineering*, pages 277–286.

Magder, L. S. and Zeger, S. L. (1996). A smooth nonparametric estimate of a mixing distribution using mixtures of gaussians. *Journal of the American Statistical Association*, 91(435):1141–1151.

McCulloch, C. E. and Neuhaus, J. M. (2011). Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Statistical science*, pages 388–402.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9(4):pp. 538–558.

Mitra, R. and Reiter, J. P. (2006). Adjusting survey weights when altering identifying design variables via synthetic data. In *International Conference on Privacy in Statistical Databases*, pages 177–188. Springer.

Moore, E. H. (1920). On the reciprocal of the general algebraic matrix. *Bulletin of the American Mathematical Society*, 26(9):394–395.

Mumtaz, G., Tamim, H., Kanaan, M., Khawaja, M., Khogali, M., Wakim, G., and Yunis, K. (2007). Effect of consanguinity on birth weight for gestational age in a developing country. *American Journal of Epidemiology*, 165:742–752.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica: Journal of the Econometric Society*, pages 69–85.

Muralidhar, K., Sarathy, R., and Dandekar, R. (2006). Why swap when you can shuffle? a comparison of the proximity swap and data shuffle for numeric data. In *Privacy in Statistical Databases*, pages 164–176. Springer.

Nagahara, Y. (2004). A method of simulating multivariate nonnormal distributions by the pearson distribution system and estimation. *Computational Statistics & Data Analysis*, 47(1):1–29.

Neuhaus, J. M., Hauck, W. W., and Kalbfleisch, J. D. (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika*, 79(4):755–762.

Neuhaus, J. M., Kalbfleisch, J. D., and Hauck, W. W. (1994). Conditions for consistent estimation in mixed-effects models for binary matched-pairs data. *Canadian Journal of Statistics*, 22(1):139–148.

ONS (2013). Inheritance in Great Britain, 2008/10. *Office for National Statistics Infographic*. [Online; accessed 30-June-2016].

ONS (2016). Wealth and assets survey, waves 1-4, 2006-2014 [data collection]. *Office for National Statistics Social Survey Division UK Data Service*.

Paass, G. (1988). Disclosure risk and disclosure avoidance for microdata. *Journal of Business & Economic Statistics*, 6(4):487–500.

Palley, M. A. and Simonoff, J. S. (1987). The use of regression methodology for the compromise of confidential information in statistical databases. *ACM Transactions on Database Systems (TODS)*, 12(4):593–608.

Pannekoek, J. (1999). Statistical methods for some simple disclosure limitation rules. *Statistica Neerlandica*, 53(1):55–67.

Pearson, K. (1907). *Mathematical contributions to the theory of evolution XVI. On further methods for determining correlation*. Cambridge University Press, Cambridge.

Peng, L. and Huang, Y. (2008). Survival analysis with quantile regression models. *Journal of the American Statistical Association*, 103(482):637–649.

Piepho, H.-P. and McCulloch, C. E. (2004). Transformations in mixed models: Application to risk analysis for a multienvironment trial. *Journal of Agricultural, Biological, and Environmental Statistics*, 9(2):123–137.

Plümper, T. and Troeger, V. E. (2007). Efficient estimation of time-invariant and rarely changing variables in finite sample panel analyses with unit fixed effects. *Political Analysis*, 15(2):124–139.

Purdam, K. and Elliot, M. (2007). A case study of the impact of statistical disclosure control on data quality in the individual uk samples of anonymised records. *Environment and Planning A*, 39(5):1101.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27(1):85–96.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics-Stockholm*, 19(1):1–16.

Ramberg, J. S., Dudewicz, E. J., Tadikamalla, P. R., and Mykytka, E. F. (1979). A probability distribution and its uses in fitting data. *Technometrics*, 21(2):201–214.

Ramberg, J. S. and Schmeiser, B. W. (1972). An approximate method for generating symmetric random variables. *Communications of the ACM*, 15(11):987–990.

Ramberg, J. S. and Schmeiser, B. W. (1974). An approximate method for generating asymmetric random variables. *Communications of the ACM*, 17(2):78–82.

Rana, S., John, A. H., and Midi, H. (2012). Robust regression imputation for analyzing missing data. In *Statistics in Science, Business, and Engineering (ICSSBE), 2012 International Conference on*, pages 1–4. IEEE.

Reich, B. J. and Smith, L. B. (2013). Bayesian quantile regression for censored data. *Biometrics*, 69(3):651–660.

Reiter, J. (2005a). Significance tests for multi-component estimands from multiply imputed, synthetic microdata. *Journal of Statistical Planning and Inference*, 131(2):365–377.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, 18(4):531–544.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29(2):181–188.

Reiter, J. P. (2004a). Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(1):185–205.

Reiter, J. P. (2004b). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology*, 30(235):l242.

Reiter, J. P. (2005b). Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association*, 100(472):1103–1112.

Reiter, J. P. (2005c). Using cart to generate partially synthetic public use microdata. *Journal of Official Statistics-Stockholm*, 21(3):441.

Reiter, J. P. (2012). Statistical approaches to protecting confidentiality for microdata and their effects on the quality of statistical inferences. *Public Opinion Quarterly*, 76(1):163–181.

Reiter, J. P. and Drechsler, J. (2010). Two stage multiple imputation to protect confidentiality. *Statistica Sinica*, 20:405 – 422.

Reiter, J. P. and Kinney, S. K. (2012). Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. *Journal of Official Statistics*, 28(4):583–590.

Reiter, J. P. and Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality*, 1(1):99–110.

Reiter, J. P., Raghunathan, T. E., and Kinney, S. K. (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology*, 32(2):143.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554.

Rodríguez, R. and Creecy, R. (2007). Synthetic data disclosure control for american community survey group quarters. In *Proceedings of the Joint Statistical Meetings*, pages 1439–1450.

Rogers, W. et al. (1994). Regression standard errors in clustered samples. *Stata Technical Bulletin*, 3(13).

Ross, A., Lloyd, J., and Weinhardt, J. (2008). *The Age of Inheritance*. International Longevity Centre - UK.

Rubin, D. B. (1980). Handling nonresponse in sample surveys by multiple imputation. *US Bureau of Census Monograph, Washington, D.C.*

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley Online Library.

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of Official Statistics*, 9(2):461–468.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489.

Sakshaug, J. W. and Raghunathan, T. E. (2011). Synthetic data for small area estimation. In *Privacy in Statistical Databases*, pages 162–173. Springer.

Sarathy, R. and Muralidhar, K. (2011). Some additional insights on applying differential privacy for numeric data. In *Privacy in Statistical Databases*, pages 210–219. Springer.

Schafer, J. (1997a). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schafer, J. L. (1997b). Imputation of missing covariates under a multivariate linear mixed model. Technical report, Technical report 97-04, Dept. of Statistics, The Pennsylvania State University, http://www. stat. psu. edu/reports/1997/tr97 4. pdf.

Schafer, J. L. and Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, 7(2):147.

Schafer, J. L. and Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, 11(2):437–457.

Schenker, N. and Taylor, J. M. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis*, 22(4):425–446.

Schenker, N., Treiman, D. J., and Weidman, L. (1988). Multiple imputation of industry and occupation codes for public-use files. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pages 85–92.

Schenker, N., Treiman, D. J., and Weidman, L. (1993). Analyses of public use decennial census data with multiply imputed industry and occupation codes. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 42(3):pp. 545–556.

Schenker, N. and Welsh, A. (1988). Asymptotic results for multiple imputation. *The Annals of Statistics*, pages 1550–1566.

Schomaker, M. and Heumann, C. (2016). Bootstrap inference when using multiple imputation. *arXiv preprint arXiv:1602.07933*.

Schunck, R. (2013). Within and between estimates in random-effects models: Advantages and drawbacks of correlated random effects and hybrid models. *The Stata Journal*, 13(1):65–76.

Shen, W. and Louis, T. A. (1999). Empirical bayes estimation via the smoothing by roughening approach. *Journal of Computational and Graphical Statistics*, 8(4):800–823.

Shi, Y. and Demirtas, H. (2015). *PoisNonNor: Simultaneous Generation of Count and Continuous Data*. R package version 1.0.

Si, Y. and Reiter, J. P. (2013). Nonparametric bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioural Statistics*, 38:499 – 521.

Skinner, C. (2008). Assessing disclosure risk for record linkage. In *Privacy in Statistical Databases*, pages 166–176. Springer.

Skinner, C. and Shlomo, N. (2008). Assessing identification risk in survey microdata using log-linear models. *Journal of the American Statistical Association*, 103(483):989–1001.

Skinner, C. J. (1992). On identification disclosure and prediction disclosure for microdata. *Statistica Neerlandica*, 46(1):21–32.

Skinner, C. J., Marsh, C., Openshaw, S., and Wymer, C. (1994). Disclosure control for census microdata. *Journal of Official Statistics*, 10(1):31–51.

SM, S. (1998). A bayesian, species-sampling-inspired approach to the uniqueness problem in microdata disclosure risk assessment. *Journal of Official Statistics*, 14:373–383.

Spruill, N. (1982). Measures of confidentiality. *Statistics of Income and Related Administrative Record Research*, pages 260–265.

Spruill, N. (1983). The confidentiality and analytic usefulness of masked business microdata. *Proceedings of the Section on Survey Research*, pages 602–607.

Spruill, N. (1984). *Protecting confidentiality of business microdata by masking.* The Public Research Institute, Alexandria, VA.

Sramka, M. (2012). Breaching privacy using data mining: Removing noise from perturbed data. In *Computational Intelligence for Privacy and Security*, pages 135–157. Springer.

Stasinopoulos, D. M. and Rigby, R. A. (2007). Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, 23(7):1–46.

Su, Y.-S., Yajima, M., Gelman, A. E., and Hill, J. (2011). Multiple imputation with diagnostics (mi) in r: Opening windows into the black box. *Journal of Statistical Software*, 45(2):1–31.

Sullivan, G. and Fuller, W. (1989). The use of measurement error to avoid disclosure. In *Proceedings of the Survey Research Methods Section*, pages 802–807. American Statistical Association.

Thompson, P., Cai, Y., Moyeed, R., Reeve, D., and Stander, J. (2010). Bayesian nonparametric quantile regression using splines. *Computational Statistics & Data Analysis*, 54(4):1138–1150.

Tokdar, S. T., Kadane, J. B., et al. (2012). Simultaneous linear quantile regression: a semiparametric bayesian approach. *Bayesian Analysis*, 7(1):51–72.

Tukey, J. W. (1997). Modern techniques in data analysis. In *NSF-sponsored regional research conference at Southeastern Massachusetts University, North Dartmouth, MA*.

Tzavidis, N., Salvati, N., Schmid, T., Flouri, E., and Midouhas, E. (2016). Longitudinal analysis of the strengths and difficulties questionnaire scores of the millennium cohort study children in england using m-quantile random-effects regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(2):427–452.

Vale, C. D. and Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, 48(3):465–471.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16(3):219–242.

van Buuren, S. (2010). Multiple imputation of multilevel data. *The Handbook of Advanced Multilevel Analysis*, pages 173–196.

van Buuren, S., Boshuizen, H. C., Knook, D. L., et al. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18(6):681–694.

van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C., and Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12):1049–1064.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91(433):217–221.

Verbeke, G. and Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis*, 23(4):541–556.

von Hippel, P. T. (2013). Should a normal imputation model be modified to impute skewed variables? *Sociological Methods & Research*, 42(1):105–138.

Vonesh, E. F. (1992). Non-linear models for the analysis of longitudinal data. *Statistics in Medicine*, 11(14-15):1929–1954.

Wang, H. J. and Feng, X. (2012). Multiple imputation for m-regression with censored covariates. *Journal of the American Statistical Association*, 107(497):194–204.

Wei, Y., Ma, Y., and Carroll, R. J. (2012). Multiple imputation in quantile regression. *Biometrika*, 99:423–438.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, pages 817–838.

White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in medicine*, 30(4):377–399.

Willenborg, L. and De Waal, T. (2012). *Elements of statistical disclosure control*, volume 155. Springer Science & Business Media.

Winkler, W. E. (2007). Examples of easy-to-implement, widely used methods of masking for which analytic properties are not justified. *Technical Report, Statistical Research Division, U.S. Bureau of the Census, Washington, DC.*

Wright, E. and P., R. (1996). Age-specific reference intervals (normal ranges). *Stata Technical Bulletin*, pages 24–34.

Wu, D.-M. (1973). Alternative tests of independence between stochastic regressors and disturbances. *Econometrica: Journal of the Econometric Society*, pages 733–750.

Yancey, W. E., Winkler, W. E., and Creecy, R. H. (2002). Disclosure risk assessment in perturbative microdata protection. In *Inference Control in Statistical Databases*, pages 135–152. Springer.

Yang, Y., He, X., et al. (2012). Bayesian empirical likelihood for quantile regression. *The Annals of Statistics*, 40(2):1102–1131.

Yang, Y., Wang, H. J., and He, X. (2015). Posterior inference in bayesian quantile regression with asymmetric laplace likelihood. *International Statistical Review*.

Yeo, I.-K. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959.

Yu, K. and Jones, M. (1998). Local linear quantile regression. *Journal of the American Statistical Association*, 93(441):228–237.

Yu, K., Lu, Z., and Stander, J. (2003). Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):331–350.

Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447.

Yuan, Y. and Yin, G. (2010). Bayesian quantile regression for longitudinal studies with nonignorable missing data. *Biometrics*, 66(1):105–114.

Yucel, R. M. (2008). Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 366(1874):2389–2403.

Yucel, R. M. (2011a). Random covariances and mixed-effects models for imputing multivariate multilevel continuous data. *Statistical modelling*, 11(4):351–370.

Yucel, R. M. (2011b). State of the multiple imputation software. *Journal of Statistical Software*, 45(1):1–7.

Yucel, R. M. and Demirtas, H. (2010). Impact of non-normal random effects on inference by multiple imputation: A simulation assessment. *Computational Statistics & Data Analysis*, 54(3):790–801.

Yucel, R. M., Ding, H., Uludag, A. K., and Tomaskovic-Devey, D. (2008). Multiple imputation in multiple classification and multiple-membership structures. *Joint Statistical Meetings*.

Yucel, R. M., Schenker, N., and Raghunathan, T. E. (2006). Multiple imputation for incomplete multilevel data with shrimp. In *Annual Conference on New Methods for the Analysis of Family and Dyadic Processes, Amherst, MA*.

Zhang, D. and Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, 57(3):795–802.

Zhang, P., Song, P. X.-K., Qu, A., and Greene, T. (2008). Efficient estimation for patient-specific rates of disease progression using nonnormal linear mixed models. *Biometrics*, 64(1):29–38.

Zhao, E. and Yucel, R. M. (2009). Performance of sequential imputation method in multilevel applications. In *American Statistical Association Proceedings of the Survey Research Methods Section*, pages 2800–2810.

Zhao, J. H. and Schafer, J. L. (2013). *pan: Multiple imputation for multivariate panel or clustered data*. R package version 0.9.