UNECE

CONFERENCE OF EUROPEAN STATISTICIANS

**Expert Meeting on Statistical Data Confidentiality**

1-3 December 2021, Poland

# Accounting for longitudinal data structures when disseminating synthetic data to the public.

Joerg Drechsler (Institute for Employment Research)
*joerg.drechsler@iab.de*

## *Abstract*

In this talk we evaluate if the concept of differential privacy can be used to disseminate detailed geocoding information without compromising the confidentiality of the individuals included in the database. To enable the release of detailed geographical information we propose a differentially private procedure based on a micro-aggregation algorithm with a fixed minimal cluster size.

We evaluate whether meaningful results can be obtained with this approach using administrative data gathered by the German Federal Employment Agency.  Detailed geocoding information has been added to this database recently and plans call for making this valuable source of information available to the scientific community. We generate differentially private microdata using different levels of geographical detail to identify the most detailed level that still provides acceptable analytical validity while offering strong differential privacy guarantees.

# Accounting for longitudinal data structures when disseminating synthetic data to the public

Sana Rashid[*], Jörg Drechsler[**], Robin Mitra[***]

[*] University of Southampton Highfield Campus, Mathematics Building 54, Southampton SO17 1BJ, sanarashidmahmood@gmail.com

[**] Institute for Employment Research and University of Maryland, Regensburger Str. 104, 90478 Nuremberg, Germany, joerg.drechsler@iab.de

[***] School of Mathematics, Cardiff University, Cardiff, CF24 4AG, mitrar5@cardiff.ac.uk

**Abstract**. When generating synthetic data for public release, careful attention must be given to the selection of appropriate synthesis models. If the dataset has a longitudinal structure it is not obvious which synthesis model should be used to account for the design. Using multiple imputation for missing data, it has been shown previously that employing fixed effects at the imputation stage may adversely affect inferences obtained by an analyst wishing to use random effects to account for the clustering of observations within units and vice versa. Since it is generally unknown which model users of the data will prefer, a synthesis model should be preferred that suits both analysis models. We evaluate several strategies for generating longitudinal synthetic datasets using extensive simulation studies. In our evaluations, we consider both, the analytical validity and the risk of disclosure resulting from the different synthesis strategies. We find that synthesis models should be preferred that cannot be classified as pure random or fixed effects models. We illustrate our findings using data from the German IAB Establishment Panel.

## 1 Introduction

The synthetic data approach for disclosure protection has gained much attention in recent years. With this approach, sensitive and/or identifying information in the data is replaced by random values generated from models fitted to the original data (Rubin, 1993). Over the years, many different modeling strategies have been proposed to flexibly generate synthetic data for various data types (see, for example, Reiter (2005); Drechsler (2011); Kim et al. (2021)). However, to our knowledge all methodological research has only focused on cross-sectional data so far. Longitudinal data poses the additional challenge that the longitudinal structure of the data needs to be preserved if the synthetic data should provide some analytical validity.

Two longitudinal synthetic data products have been released in recent years: the synLBD (Kinney et al., 2011) and the SIPP synthetic beta (Abowd et al., 2006). Both were generated using the “wide approach” for synthesis. With the wide approach, repeated measures of the same variable are stored as separate variables. However, the wide approach has several disadvantages: First, the number of variables in the data increases substantially. Furthermore, the standard advice for synthesis to condition on all other variables in the dataset to preserve the relationships between the variables and to improve the efficiency of the imputation (Little and Raghunathan, 1997) is often no longer feasible with the wide approach. Finally, the underlying imputation model will not be congenial to commonly used analysis models for longitudinal datasets. The synthesis model and analysis model are said to be congenial, if they are based on exactly the same modeling assumptions (Meng, 1994). However, in longitudinal data, random effects models or models based on fixed effects using the within transformation described below are typically used to account for the panel structure. These models are not nested within the wide modeling approach so uncongeniality becomes an issue.

To better understand the implications of the uncongeniality resulting from using different strategies to account for the longitudinal data structure, this paper evaluates the impact of various synthesis/analysis model combinations. Beyond the wide approach and the random and fixed effects models discussed above, we consider two additional synthesis strategies: a standard linear regression model which ignores the longitudinal structure and the so-called hybrid model as discussed in Allison (2009).

The remainder of this paper is organized as follows: In Section 2 we discuss the two modeling strategies which are typically used for analyzing longitudinal data. Section 3 presents alternative modeling strategies which could be used when synthesizing longitudinal data. In Section 4 we conduct an extensive simulation study which evaluates the impacts of different synthesis/analysis model combinations. In Section 5 we present a real data application based on a research paper by Ellguth et al. (2014) which evaluates wage effects of works councils using the IAB Establishment Panel, a large scale establishment survey in Germany. The paper concludes with some final remarks.

## 2 Analyzing longitudinal data

Depending on the context several strategies can be used for analyzing longitudinal data. In this paper, we assume that the analyst is interested in the relationship between an outcome variable $\mathbf{Y}$ and a set of time constant and time varying predictors $\mathbf{X}$ and $\mathbf{Z}$ and uses a parametric regression model for studying this relationship. In general, this regression model can be specified as:

$$Y_{it} = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_{it}\boldsymbol{\gamma} + \delta_i + \varepsilon_{it} \qquad \varepsilon_{it} \sim N(0, \sigma^2), \tag{1}$$

where $i = 1, \ldots, n$ is the index for individuals, $t = 1, \ldots, T$ is the time index, $\mathbf{X}$ and $\mathbf{Z}$ contain the time constant and time varying variables respectively, and $\delta_i$ is an individual specific effect. The two approaches for analyzing longitudinal data differ in how they model $\delta_i$. Below, we present the two modeling strategies in more detail.

### 2.1 Fixed Effects Models

With fixed effects models all parameters are assumed to be fixed in the population. This also holds for the individual effect $\delta_i$. The simplest strategy for estimation is to include a dummy indicator for each individual in the regression model (least squares dummy variable (LSDV) model). However, with this modeling strategy the parameters of the time constant variables are no longer identified and need to be dropped from the model. The LSDV model can be written as:

$$Y_{it} = \mathbf{Z}_{it}\boldsymbol{\gamma} + \mathbf{I}_i^{ind}\boldsymbol{\delta}^f + \varepsilon_{it}, \qquad \varepsilon_{it} \sim N(0, \sigma^2),$$

where $\mathbf{I}^{ind} = \{\mathbf{I}_1^{ind}, \ldots, \mathbf{I}_n^{ind}\}$ is an $nT \times n$ dimensional matrix of indicator variables identifying the $n$ individuals in the dataset. The $n \times 1$ vector $\boldsymbol{\delta}^f$ contains the fixed individual effects. A downside of the LSDV approach is that a large number of parameters needs to be estimated and only limited information is available to estimate the individual effects since typically $T \ll n$.

As the individual effects are typically nuisance parameters, a popular strategy to reduce the number of parameters is the *within transformation.* With this approach, the unit specific mean is subtracted from each variable, and Equation (1) turns into

$$Y_{it} - \bar{Y}_i = (\mathbf{Z}_{it} - \bar{\mathbf{Z}}_i)\boldsymbol{\gamma} + \varepsilon_{it} - \bar{\varepsilon}_i. \tag{2}$$

Note that all time constant variables cancel out, which offers the advantage that time constant omitted variables will not introduce any bias. One disadvantage of this strategy is that similar to the LSDV approach the effects of the time constant variables are no longer directly identifiable. Furthermore, individual records are no longer independent and thus simple ordinary least square solutions are no longer valid (see for example Wooldridge (2010, Chap. 10.5) for further details).

### 2.2 Random Effects Models

Random effects models, also known as multilevel or mixed effects models, assume that there is a random component in addition to the fixed effect for some of the variables. In longitudinal analysis, the standard assumption is that this randomness is limited to the individual effect $\delta_i$, that is, a classical random intercept model is most commonly used when analyzing longitudinal data. The regression model looks exactly similar to Equation (1) with the additional assumption that $\delta_i$ is a random variable following a normal distribution with zero mean and variance $\sigma_\delta^2$:

$$Y_{it} = \mathbf{X}_i\beta + \mathbf{Z}_{it}\gamma + \delta_i + \varepsilon_{it} \qquad \delta_i \sim N(0, \sigma_\delta); \quad \varepsilon_{it} \sim N(0, \sigma_\varepsilon^2).$$

Note that the time constant effects are still identified with this model due to the randomness in the individual effects.

## 3 Synthesizing Longitudinal Data

To evaluate the implications of modeling decisions at the synthesis stage on longitudinal analyses, we consider six synthesis strategies in our simulation study and real data application: a simple OLS regression to illustrate that ignoring the longitudinal structure will generally lead to biased inferences in downstream analyses, the LSDV approach, two variants of the fixed effects approach based on the within transformation, a random effects model, a so-called hybrid model, which will be further discussed below and a synthesis model based on the wide approach. Throughout this paper, we assume that the sequential regression approach is used for synthesis, i.e. variables are synthesized sequentially based on univariate regression models.

To our knowledge the within transformation has never been used for imputation, neither for missing data nor data confidentiality. We present steps for turning this model into a synthesizer below.

### 3.1 The Within Transformation for Synthesis

We note that obtaining parameter estimates for synthesis is straightforward since simple OLS regression using the transformed versions of $Y$ and $\mathbf{Z}$ will provide unbiased estimates of the regression coefficients $\gamma$. To obtain an unbiased estimate of the residual variance, the degrees of freedom need to be adjusted relative to the OLS estimate (see Wooldridge (2010, Chap. 10.5.2)). However, the residuals that need to be added to the predicted values from the model are correlated and cannot be drawn independently as in standard OLS synthesis. Defining $\varepsilon^*_{it} = \varepsilon_{it} - \bar{\varepsilon}_i$ it holds that $Var(\varepsilon^*_{it}) = (1 - 1/T)\sigma^2$, $Cov(\varepsilon^*_{it}, \varepsilon^*_{it'}) = -\sigma^2/T$ for $t \neq t'$, and $Cov(\varepsilon^*_{it}, \varepsilon^*_{jt}) = 0$ for $i \neq j$. Thus, residuals are correlated within individuals but not between. Since residuals are independent between units, we can draw a vector of $T$ residuals $\varepsilon^*_i$ independently for each unit using the covariance structure outlined aboveand add these to the predicted values.

This results in synthetic values for $Y_{it} - \bar{Y}_i$, i.e. synthetic time specific deviations from the unit specific mean. To obtain synthetic versions of $Y_{it}$, we can either 1) add the individual specific means from the original data, or 2) generate synthetic means with a standard OLS synthesizer using the average of the time varying variables $\bar{\mathbf{Z}}$ (and potentially other time constant variables $\mathbf{X}$) as predictors and add these synthetic means to the values obtained in the previous step.

### 3.2 The Hybrid Model

Some authors have proposed a modeling strategy that tries to combine the advantages of the random and the fixed effects approach. Following Allison (2009) we call this approach the hybrid model. The hybrid model has the following general form:

$$Y_{it} = \mathbf{X}_i\boldsymbol{\beta} + (\mathbf{Z}_{it} - \bar{\mathbf{Z}}_i)\boldsymbol{\gamma} + \bar{\mathbf{Z}}_i\boldsymbol{\eta} + \delta_i + \varepsilon_{it} \quad \delta_i \sim N(0, \sigma_\delta); \quad \varepsilon_{it} \sim N(0, \sigma_\varepsilon^2).$$

Similar to the within transformation the model takes care of omitted variable bias caused by time constant variables by including $\bar{\mathbf{Z}}_i$. At the same time the effect of the time constant variables $\mathbf{X}$ can still be estimated directly. However, the efficiency gains of the random effects model are lost with this approach. Still, the approach seems attractive for synthesis since it combines ideas and modeling strategies from both models and thus might provide valid results based on the synthetic data irrespective of which model is selected for analyzing the data.

# 4 Simulation Study

To evaluate the impacts of the different synthesis/analysis model combinations we conduct extensive simulation studies. Specifically, we consider two analysis models (the random effects model and the fixed effects model) and seven synthesis models: simple OLS regression ignoring the longitudinal data structure (IGN), the least square dummy variable approach (LSDV), the random effects model (RE), the hybrid model (HYB), the within transformation approach using synthetic means (FE1) and the true sample means (FE2), and the wide approach (WIDE). Here, we only summarize the main findings of the simulation study for brevity. A longer version of the paper, which presents the simulation study in more details, can be obtained from the authors upon request.

## 4.1 Simulation Design

We consider two cases. In the first case the data are generated using a simple random intercept model with only one time varying predictor. In the second case, we include an additional time constant variable in the data generating process that is not available at the synthesis or analysis stage. Note that this omitted variable will introduce bias in all random effects analysis models and also in several of the synthesis models. We evaluate both settings for two different levels of intra class correlation (ICC={0.06; 0.5}). In all cases the sample size is $n = 1,000$ and the number of time points is set to $T = 5$. We assume that only the dependent variable will be synthesized. The whole process of generating the data, synthesizing the dependent variable $m = 10$ times with the various synthesis strategies, and analyzing the synthetic data using the two different analysis models is repeated 1,000 times for each simulation setup.

## 4.2 Analytical Validity of the Regression Coefficient

If all variables are available for synthesis and analysis, we find that none of the methods introduces any bias for the regression coefficient. However, for the fixed effects analysis model the variances of the regression coefficient are overestimated by up to a factor of two if the IGN or RE model is used for synthesis leading to substantial overcoverage. The overestimation increases with decreasing ICC. On the other hand, the variance is considerably underestimated in the random effects analysis model, if IGN is used for synthesis and overestimated (at least for small ICC) if the LSDV is used. All other methods provide approximately valid results.

If we add a variable for the data generating process that is not available at the synthesis or analysis stage, results will always be biased and coverage rates will be (close to) zero if the analyst uses a random effects model. However, results will also be biased if the analyst uses a fixed effects model and the IGN or the RE model are used at the synthesis stage. Thus, neither the IGN nor the RE model can be recommended as a general strategy for longitudinal data synthesis since both models prevent the analyst from being able to control for omitted variable bias. Otherwise, the findings are similar to the findings in the absence of ommitted variables and we can summarize that either FE1, FE2, HYB, or WIDE should be selected if the analyst is interested in the regression coefficient $\gamma$.

### 4.3 Analytical Validity of the Residual Variance(s)

We also evaluated how the different synthesis/analysis combinations affect the estimated residual variances of the analysis model.We find that if the fixed effects analysis model is used, the residual variance is estimated (almost) unbiasedly for all synthesis strategies except if the IGN model is used for synthesis. We also note a slight overestimation of $\sigma_\varepsilon^2$ for the WIDE approach.

The findings are similar for most synthesizers if the random effects model is used as the analysis model. Both variance components of the random effects model are estimated unbiasedly except if the IGN model or the LSDV are used for synthesis. Finally, we again observe a slight overestimation –for both variance components in this case– if the WIDE approach is used for synthesis.

The findings for the omitted variable case are similar with the exception that the estimated variance of $\delta_i$ in the RE analysis model would always be biased even if the model was run on the original data.

In summary, only the IGN and the LSDV model fail as synthesis models if interest lies in the variance components. We only note a slight overestimation for both variance components if the WIDE approach is used.

### 4.4 Results Regarding the Risk of Disclosure

To evaluate re-identification risks, we propose computing probabilities of identification using methods developed in Reiter and Mitra (2009). A detailed description of the methodology is contained in the full paper, which can be obtained from the authors upon request. Here, we only summarize the main ideas. Suppose the intruder has information on some target records which she will use in a record linkage attack

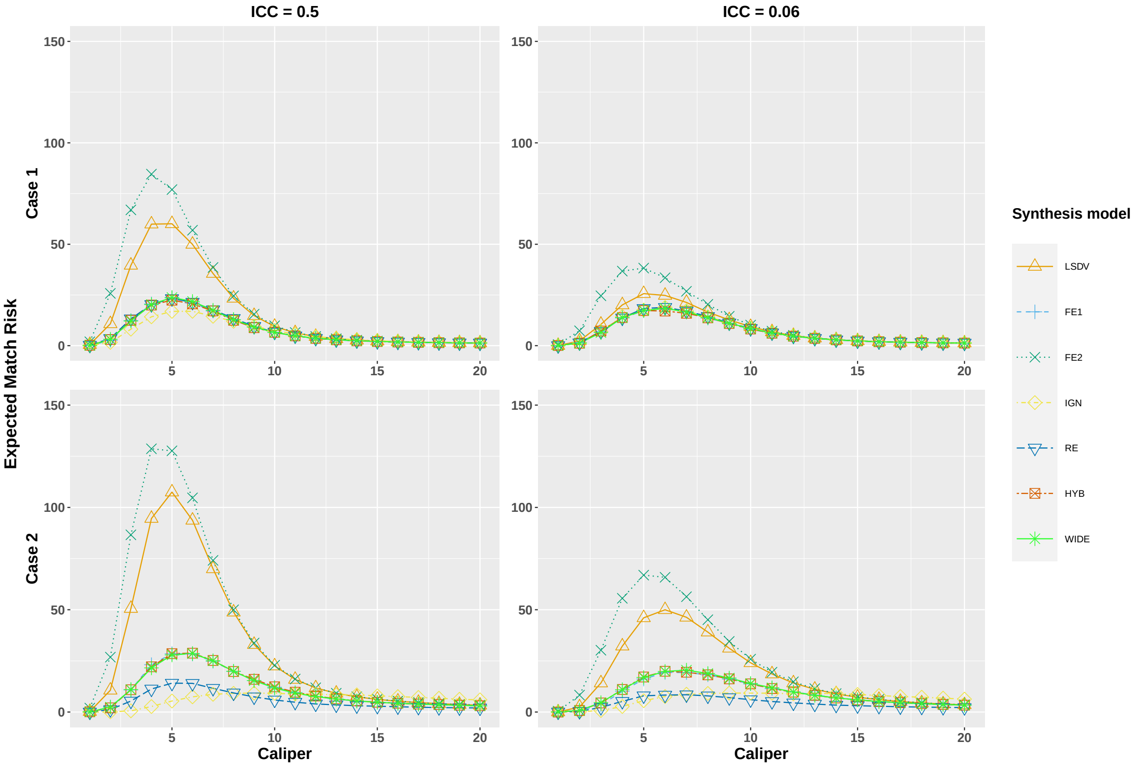

Figure 1: Expected match risk for the synthesis models for two intra-class correlation (ICC) settings. Case 1: Synthesis and analysis models correctly specified. Case 2: Time constant variable not available for synthesizer and analyst.

to identify the targets in the released data. Similar to the concept of probabilistic record linkage, the idea is to estimate the probability of a match between each target record and each record in the released file. The record with highest average matching probability across the synthetic datasets is the declared match. In the evaluation, we use two risk measures that are summaries of these matching probabilities: the expected match risk, which computes the expected number of correctly declared matches, and the true match rate, which computes the number of correct single matches among the target records.

For our risk evaluations we assume that the intruder knows the true values for $Y$ for all records and uses this information to try to identify units in the database. To identify potential matches, the intruder uses a caliper matching approach. For any original record $i$, $i = 1, \dots, n$ he or she computes the Euclidean distance considering all time points for all synthetic records $l$, $l = 1, \dots, n$. All synthetic records with an Euclidean distance less than a pre-specified threshold $\kappa$ ($\kappa = \{1, 2, \dots, 20\}$) are considered potential matches for record $i$.

Figure 1 presents the expected match risks for the various scenarios (results for the true match rates showed similar patterns and are omitted for brevity). We observe a similar pattern in all settings: The FE2 synthesis model, which uses the individual means from the original data, always shows the highest risk followed by the LSDV approach. Most of the other methods show a similar risk profile in Case 1 with the exception of the IGN model which has the lowest risk if the ICC is large. In

Case 2 the risks are further reduced for the RE and the IGN since these two models introduce bias which typically reduces risks. For the LSDV and FE2 approaches, the risks increase in Case 2. This can be explained if we note that the additional time constant variable will increase the variability of $Y$ between units in the original data. Thus, the risks increase for the original data since the different units are easier to distinguish. However, this increased variability between units is only preserved for the LSDV and FE2 synthesizers, hence these are the only models for which the increased risks can also be observed in the synthetic data.

Regarding the different ICC levels, we observe that decreasing the ICC reduces the risk for the FE2 and LSDV approach. These models are only affected by the residual variance, as they treat the variability between individuals as fixed. Th residual variance is larger in the small ICC scenario, hence the risks decrease. For all other models, the risks are approximately similar as the total amount of unexplained variance is similar for both ICC settings in our simulation setup.

Summarizing the results, we find that from a disclosure risk perspective FE2 and LSDV models should be avoided. Also taking into account the results regarding the analytical validity from the previous section, the preferred methods are HYB, FE2, and WIDE.

## 5 Real Data Application

In this section we evaluate to what extend real data analyses could be replicated using the different synthesis strategies. We aim at replicating results from a paper by Ellguth et al. (2014) that uses the IAB Establishment Panel (Fischer et al., 2008) to explore the effects of works councils and opening clauses in collective bargaining agreements on wages. We again only summarize the main findings. A more detailed analysis is included in the full paper available from the authors.

The data used in Ellguth et al. (2014) are extracted from the years 2005 and 2007 of the IAB Establishment Panel survey. The final model of interest is given as:

$$\begin{aligned} ln(wage) &= \beta_0 + \beta_1 WOCO + \beta_2 OC + \beta_3 OC \times WOCO + \beta_4 OC^{app} \\ &\quad + \beta_5 OC^{app} \times WOCO + \mathbf{X}\gamma + \varepsilon, \end{aligned} \tag{3}$$

where $ln(wage)$ is the logarithm of the total monthly wage bill per full-time equivalent employee, $OC$ and $WOCO$ are indicator variables for the existence of opening clauses and works councils, respectively, $OC^{app}$ is an indicator whether the opening clause was actually applied, and $\mathbf{X}$ contains additional control variables.

This article is ideal for our study, as the authors employed a number of sensitivity checks to verify their conclusions. Most importantly, they fit various forms of the model; these correspond directly to our IGN, FE, and RE models.

### 5.1 Data Synthesis

For the synthesis we only use data from the waves 2005 and 2007 and only keep those cases that are observed in both waves. We assume that only the wage variable should be synthesized. We select all variables that are used in the model considered by Ellguth et al. (2014) as predictors, however we do not tailor our synthesis model to the analysis model of the authors, since this information would typically not be available to the data providing agency. Instead we adopt the common synthesis strategy to include as many predictors as possible to preserve the relationships with all these variables but only include main effects without any interactions or nonlinear terms. We also always include all variables on their original scale even if the variables are transformed or re-categorized in the final analysis model (heavily skewed variables such as number of fixed-term, part-time or casual employees are log-transformed in the synthesis model).

We generate $m = 10$ synthetic copies of the wage variable for each of the synthesis models and take a number of steps to improve the quality of the generated data. These steps are described in the full version of this paper.

### 5.2 Results Regarding Analytical Validity

As discussed above, Ellguth et al. (2014) use three different versions of the model described in (3) to evaluate the robustness of the results. In the full paper we evaluate how well the different synthesizers preserve the inferences regarding the five parameters of interest in each of the three models. Here we only summarize the key findings for brevity.

For the OLS analysis model, we observe that the regression coefficients for $WOCO$ and $OC$ are positively biased if FE1, IGN, or WIDE are used for synthesis, although the bias arguably is small. For the fixed effects analysis model, only the IGN synthesis model and the RE synthesis model introduce bias in the coefficients for $WOCO$ and $OC$. For $WOCO$ the bias is so large that the coefficient changes from being insignificant to being highly significant for both models. For the random effects analysis model, we observe a small upward bias for the same two coefficients for the HYB and WIDE synthesis model and a substantial bias for the IGN synthesis model.

Summarizing the results from the different models, we find – comparable to the results from the simulation study – that neither the RE nor the IGN model can be recommended for synthesis. Additionally, we find small biases for some of the estimates for FE1, HYB, and WIDE. These biases are most likely due to model mis-specifications.

For disclosure risk evaluations we use the same measures as presented in Section 4.4. We assume that the intruder knows the federal state and total wage bill for each establishment in the released data and uses this information for re-identification purposes.

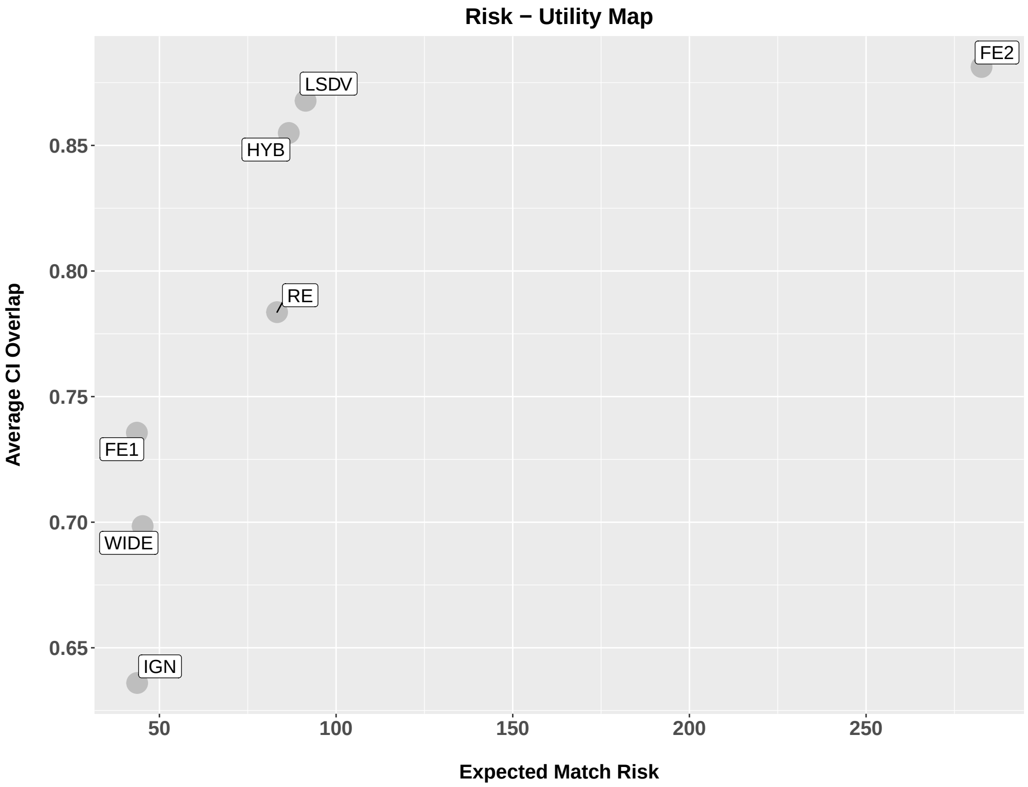

Figure 2: Risk-Utility-Map: Expected match risk against average confidence interval overlap for different synthesis methods.

To visualize the trade-offs between utility and disclosure risk we provide a risk-utility map where we plot the expected match risk against a commonly used utility measure, the confidence interval overlap as proposed by Karr et al. (2006). The measure is computed by calculating the confidence interval for any estimand of interest from the original data and from the synthetic data and looking at the overlap between the two intervals to quantify the data utility.

Figure 2 presents the results for the different synthesizers. The reported values for the confidence interval overlap are based on the average of the five parameters of interest for all three analysis models. The FE2 model stands out because of its exceptionally large risk. However, the LSDV offers an almost similar level of analytical validity as measured by the confidence interval overlap with substantially reduced risk levels. The hybrid models performs similar well with slightly decreased utility and slightly decreased risk compared to the LSDV model. Since risks are comparable to the RE model but utility for the RE model is lower, the hybrid model dominates the RE model. Among the three models with lowest disclosure risk, the FE1 model performs best in terms of utility.

## 6 Conclusions

Research on best strategies for synthesis of longitudinal data is currently still limited. In this paper we tried to fill this gap by proposing a congenial synthesis model based on the within transformation popular in longitudinal analysis and evaluating its performance relative to various other synthesis strategies for longitudinal data. Based on our extensive simulations and real data applications, we find that the wide approach preserves the analytical validity irrespective of the analysis model of the user. Other synthesis models that perform similarly well are models based on the within transformation and the hybrid modeling approach. However, the synthesis model which uses the individual means from the original data leads to substantially increased risks and our real data applications illustrate that these risks could be unacceptably high in practice.

Interestingly, we found that the LSDV approach performed best (if we exclude the unacceptably risky FE2 approach) in terms of preserving the analytical validity in our real data application and the analytical validity of the wide approach and the within transformation based on synthetic means was considerably lower than for the hybrid model. The good performance of the LSDV approach can partly be explained by the fact that we did not account for any biases in the variance estimates in our real data application. We also note that the applicability of the LSDV approach is often limited in practice, since individual effects need to be estimated for each unit. These estimates can become unstable if many units and only few time points are available. But the results also illustrate that once we deal with real data, model mis-specification might often be a bigger problem than the theoretical properties of the different models. In practice, any bias from mis-specification will have bigger impacts on the analytical validity.

Based on our findings from the real data application, the wide approach and the within approach using synthetic means seem to be more susceptible to model mis-specification. For the wide approach separate models need to be fitted for each time point and the within transformation approach needs a separate synthesis model for the mean to sufficiently protect the data. Thus, we would recommend the hybrid model based on our findings. Compared to the wide approach the hybrid approach offers the additional advantage that the modeling task does not get more complicated as the number of time points increase.

**Acknowledgements**
This work was partially supported by The Economic and Social Research Council's Doctoral Training Centre based at the University of Southampton and the DFG grant DR 831/2-2.

## References

Abowd, J. M., M. Stinson, and G. Benedetto (2006). Final report to the Social Security Administration on the SIPP/SSA/IRS public use file project. Technical report, Longitudinal Employer–Household Dynamics Program, U.S. Bureau of the Census, Washington, DC.

Allison, P. D. (2009). *Fixed effects regression models*, Volume 160. SAGE.

Drechsler, J. (2011). Using support vector machines for generating synthetic datasets. In *Privacy in statistical databases*, pp. 148–161. Springer.

Ellguth, P., H.-D. Gerner, and J. Stegmaier (2014). Wage effects of works councils and opening clauses: The German case. *Economic and Industrial Democracy 35*(1), 95–113.

Fischer, G., F. Janik, D. Müller, and A. Schmucker (2008). The IAB Establishment Panel – from sample to survey to projection. Technical report, FDZ-Methodenreport, No. 1, Institute for Employment Research, Nuremberg.

Karr, A. F., C. N. Kohnen, A. Oganian, J. P. Reiter, and A. P. Sanil (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician 60*, 224–232.

Kim, H. J., J. Drechsler, and K. J. Thompson (2021). Synthetic microdata for establishment surveys under informative sampling. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 184*(1), 255–281.

Kinney, S. K., J. P. Reiter, A. P. Reznek, J. Miranda, R. S. Jarmin, and J. M. Abowd (2011). Towards unrestricted public use business microdata: The Synthetic Longitudinal Business Database. *International Statistical Review 79*(3), 362–384.

Little, R. J. A. and T. E. Raghunathan (1997). Should imputation of missing data condition on all observed variables? In *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 617–622. Alexandria, VA: American Statistical Association.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (disc: P558-573). *Statistical Science 9*, 538–558.

Reiter, J. and R. Mitra (2009). Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality 1*, 99–110.

Reiter, J. P. (2005). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics 21*, 441–462.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics 9*, 462–468.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data.* MIT press.