Institute for Employment Research
The Research Institute of the Federal Employment Agency

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

Expert Meeting on Statistical
Data Confidentiality
02. December 2021, Poznań, Poland

Sana Rashid
(University of Southampton)

Jörg Drechsler
(Institute for Employment Research)

Robin Mitra
(Cardiff University)

Background

- Most research on synthetic data only focused on cross sectional data
- More and more longitudinal data available
- Providing access to these data is challenging
  - Longitudinal structure of the data needs to be preserved
  - Repeated measures for the same units tend to increase the risk of disclosure
- Some longitudinal synthetic datasets have been released (SIPP, synLBD)
- Followed standard recommendation to use “wide approach” for synthesis
- Problem: synthesis model will not be congenial to analysis model

Analyzing Longitudinal Datasets (From an Economist's Perspective)

- Only focus on parametric regression modeling
- Two types of models are usually employed
  - Random effects models
  - Fixed effects models based on "within-transformation"
- Both account for hierarchical data by decomposing the variance
- Major difference: treatment of the individual specific effects
- Disadvantage of the random effects approach
  - Parametric model for the random effects might be misspecified
  - Omitted variables will introduce bias
- Disadvantage of the fixed effects approach
  - Effect of time constant variables cannot be estimated directly
  - Less efficient
  - Observations no longer independent after within-transformation

Implications for Data Synthesis

- We do not know which models the users want to fit
- Ideally we want to make everybody happy
- Several open questions
  - Implications of RE synthesis model/FE analysis model and vice versa
  - Implications of the WIDE approach
  - Using the within transformation for synthesis
  - Alternative models

Synthesis Models

- We consider seven models
  - OLS ignoring the hierarchical structure (IGN)
  - Fixed effects model (LSDV)
  - Random effects model (RE)
  - Hybrid model (HYB)
  - Synthesis based on within transformation (FE1 and FE2)
  - Wide format model (WIDE)

- Hybrid model (Allison, 2009) is a mixture of FE and RE model

$$y_{it} = \boldsymbol{X_i\beta} + (\boldsymbol{Z_{it}} - \overline{\boldsymbol{Z}}_{\boldsymbol{i}})\boldsymbol{\gamma} + \overline{\boldsymbol{Z}}_{\boldsymbol{i}}\eta + \delta_i + \varepsilon_{it}, \quad i = 1, \dots, n, \quad t = 1, \dots, T$$

$$\delta_i \sim N(0, \tau^2)$$

$$\varepsilon_{it} \sim N(0, \sigma^2)$$

Synthesis Models

- Ignoring model is standard linear model
- LSDV model is standard linear model with dummies for each individual
- RE model more involved
- Use Bayesian version of RE model based on Gibbs sampler
- Hybrid model is just extended RE model
- Wide approach again set of standard linear models

- Only FE model requires minor adjustments for the synthesizer

The Synthesizer for the DIFF Model

$$y_{it} - \bar{y}_i = (\boldsymbol{Z_{it}} - \bar{\boldsymbol{Z}}_{\boldsymbol{i}})\boldsymbol{\gamma} + \varepsilon^*_{it}, \qquad \varepsilon^*_{it} = \varepsilon_{it} - \bar{\boldsymbol{\varepsilon}}_i$$

- Synthesize $y_{it} - \bar{y}_i$ instead of y_{it}
- Problem: observations are no longer *iid*
- We know that residuals are only correlated within individuals

$Var(\varepsilon^*_{it}) = \left(1 - \frac{1}{T}\right)\sigma^2$, $Cov(\varepsilon^*_{it}, \varepsilon^*_{it'})$=-$\sigma^2$/T for t≠t', $Cov(\varepsilon^*_{it}, \varepsilon^*_{jt})$=0

- We can draw a vector of *T* residuals ε^*_{it} independently for each individual

- Two options to obtain synthetic y_{it} based on synthetic $y_{it} - \bar{y}_i$
  - Add synthetic means based on the model $\bar{y}_i = \bar{\boldsymbol{Z}}_{\boldsymbol{i}}\boldsymbol{\xi} + \boldsymbol{\tau_i}$ (FE1)
  - Add the observed unit specific means $\bar{y}_i$ (FE2)

Simulation Study

- Data generating process

Case 1: $y_{it} = \delta_i + z_{it}\beta + \varepsilon_{it}, \quad i = 1, \ldots, 1000; \;\; t = 1, \ldots, 5$

Case 2: $y_{it} = \delta_i + z_{it}\beta + w_i\gamma + \varepsilon_{it}$

$\delta_i \sim N(12.5, \tau^2)$

$\varepsilon_{it} \sim N(0, \sigma^2)$

- $\mathrm{ICC} = \frac{\tau^2}{\tau^2+\sigma^2} = \{0.06; 0.5\}$ $(\tau^2 = \{4,0.5\}; \sigma^2 = \{4,7.5\})$
- w_i not available for synthesis or analysis
- 7 synthesis models for y_{it}: IGN, LSDV, RE, HYB, FE1, FE2, WIDE
- 2 analysis models: FE and RE
- 1,000 simulation runs

Results for β Case 1 (No Risk of OVB)

- All point estimates unbiased

Synthesis Model		Analysis Model ICC=0.5 FE	ICC=0.5 RE	ICC=0.06 FE	ICC=0.06 RE
ORIG	var.ratio	1.05	1.01	1.05	0.99
	CI.length	1.00	1.00	1.00	1.00
	Cov	96.2	95.1	96.2	94.5
IGN	var.ratio	1.22	0.53	1.95	0.88
	CI.length	1.49	1.06	1.08	1.00
	Cov	97.5	82.8	99.3	93.0
LSDV	var.ratio	1.06	1.06	1.04	1.16
	CI.length	1.05	1.06	1.05	1.18
	Cov	95.4	96.0	95.7	96.4
FE1	var.ratio	1.01	0.96	1.02	0.98
	CI.length	1.05	1.05	1.05	1.05
	Cov	94.8	95.1	95.8	95.4
FE2	var.ratio	1.01	0.99	1.02	0.98
	CI.length	1.05	1.04	1.05	1.02
	Cov	94.8	94.9	95.8	94.4
RE	var.ratio	1.21	0.98	1.93	0.99
	CI.length	1.05	1.05	1.05	1.05
	Cov	96.4	94.4	99.3	95.0
HYB	var.ratio	1.04	1.00	1.06	0.99
	CI.length	1.05	1.05	1.05	1.05
	Cov	95.7	94.2	95.3	95.0
WIDE	var.ratio	1.03	1.00	1.02	0.97
	CI.length	1.06	1.06	1.06	1.05
	Cov	96.4	95.4	95.6	94.1

Results for Case 2 (Omitted Variable)

- RE analysis model is always biased
- RE and IGN synthesis models will cause bias in FE analysis model
- Otherwise results similar to case 1

Results for the Residual Variance(s)

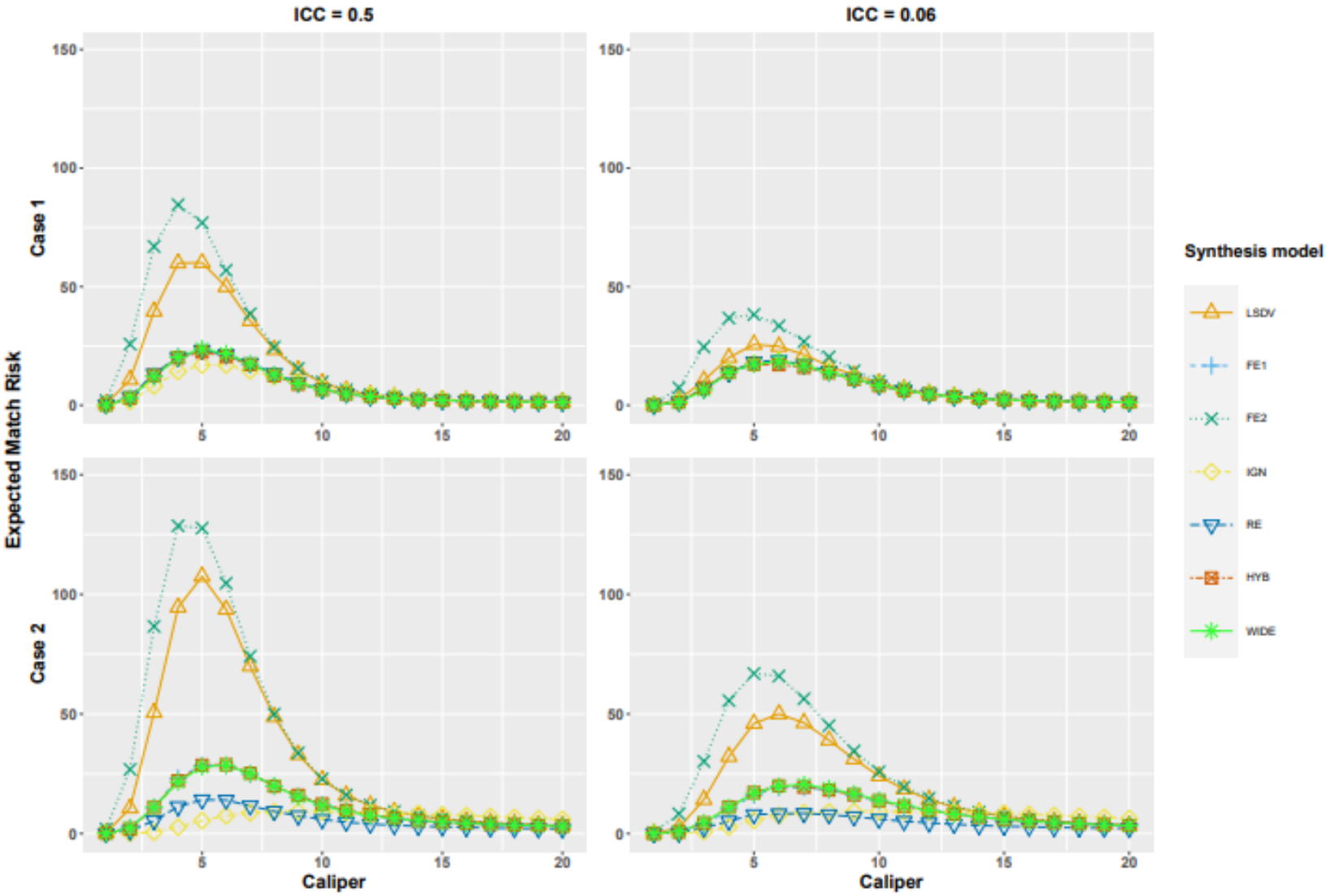
Synthesis Model		ICC=0.5 FE	ICC=0.5 RE	ICC=0.06 FE	ICC=0.06 RE
ORIG	σ^2_ϵ	4.00	4.00	7.49	7.49
	σ^2_δ	-	3.99	-	0.5
IGN	σ^2_ϵ	7.99	7.96	8.00	7.96
	σ^2_δ	-	0.03	-	0.03
LSDV	σ^2_ϵ	3.99	3.99	7.49	7.49
	σ^2_δ	-	4.81	-	2
RE	σ^2_ϵ	3.99	3.99	7.49	7.49
	σ^2_δ	-	4.00	-	0.5
FE1	σ^2_ϵ	4.00	4.00	7.49	7.49
	σ^2_δ	-	4.00	-	0.5
FE2	σ^2_ϵ	4.00	4.00	7.49	7.49
	σ^2_δ	-	4.01	-	0.5
HYB	σ^2_ϵ	3.99	3.99	7.49	7.49
	σ^2_δ	-	4.00	-	0.5
WIDE	σ^2_ϵ	4.04	4.04	7.57	7.57
	σ^2_δ	-	4.04	-	0.5

- Recommendation so far: use FE, HYB or WIDE model

Disclosure Risk Evaluations

- Use risk measures suggested by Reiter and Mitra (2009)
- Assume that intruder has some background knowledge on some target variables
- Tries to find these targets in the released data to learn sensitive information
- Intruder computes matching probabilities for each record in the released file
- Declares the record with the highest matching probability to be the match
- Risk measures evaluate how often this strategy is successful

Expected Match Risk



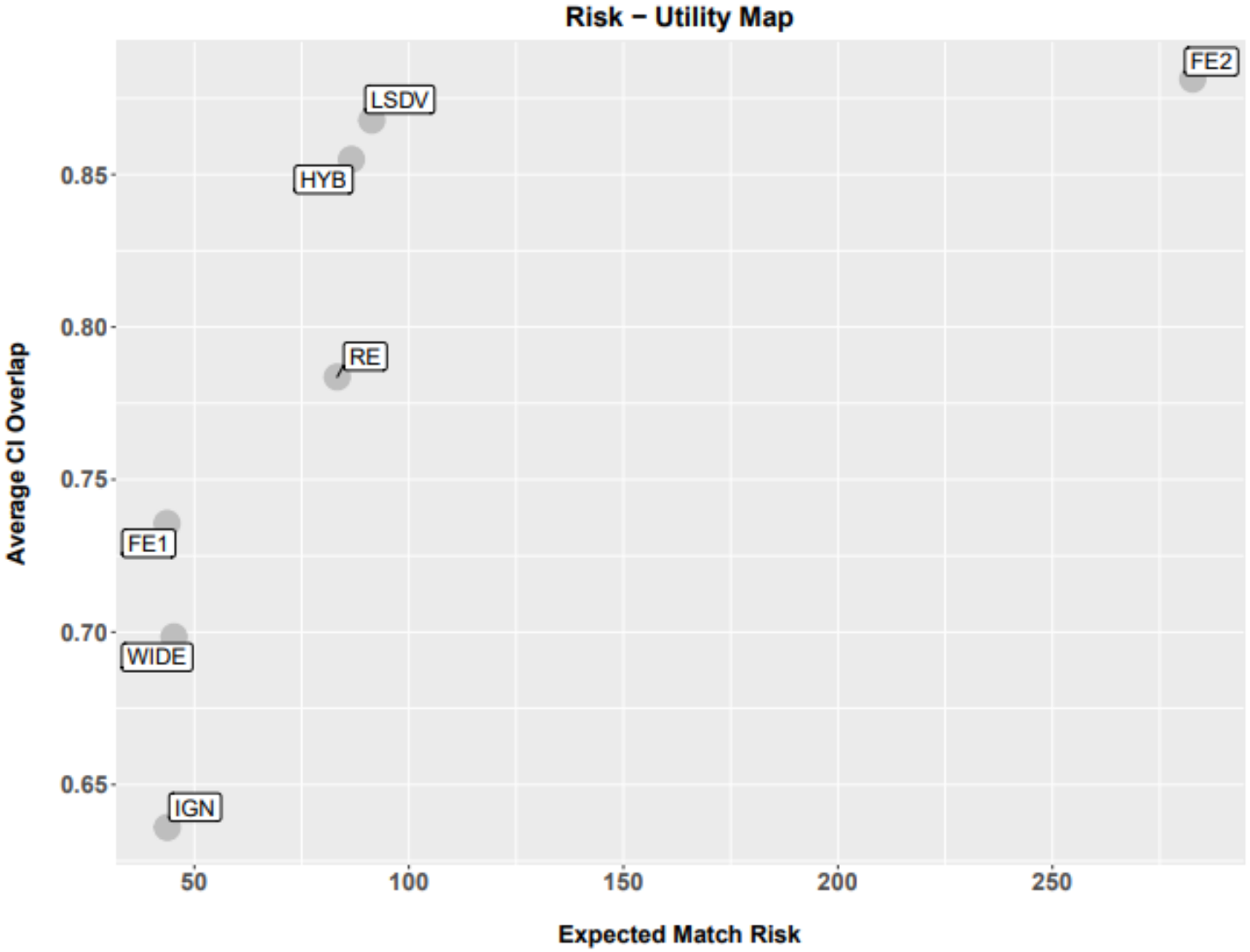
Real Data Application

- Replicate results from a paper by Ellguth et a. (2014)
- Paper explores the effects of works councils and opening clauses in collective bargaining agreements on wages
- Uses two waves of the IAB Establishment Panel
- Authors run sensitivity checks using FE, RE, and IGN models
- Model of interest (notation for the IGN model)

$$\ln(Y) = \beta_0 + \beta_1 WOCO + \beta_2 OC + \beta_3 OC \times WOCO + \beta_4 OC^{app} + \beta_5 OC^{app} \times WOCO + \boldsymbol{x}'\gamma + \boldsymbol{\varepsilon}$$

ln(Y)	total monthly wage bill per full-time equivalent employee
OC	opening clause (y/n)
WOCO	works council (y/n)
OC^{app}	application of opening clause

Risk-Utility Map



Conclusions

- As a general strategy it does not seem advisable to use RE or LSDV models for synthesis
- FE, HYB, and WIDE offer high analytical validity
- Not synthesizing the mean for the FE approach will substantially increase the risks of disclosure
- Standard approach based on WIDE model seems to be ok
- Still need to evaluate the effect of increasing the number of variables and/or the number of waves
- Once we get to real data, bias in model specification probably the bigger problem

Thank you for your attention

joerg.drechsler@iab.de