Integration of survey and administrative data for statistics production - a new framework

Li-Chun Zhang

University of Southampton (L.Zhang@soton.ac.uk)

& Statistics Norway (lcz@ssb.no)

Outline

- Combining survey and administrative (NB. register) data

  beyond survey sampling paradigm

- Case-in-point: use of **VAT** data for **STS**

- **Error sources** for integrated statistical micro data

- **Progressive** nature of **register data**

- A **prediction framework** for progressive data

Register data for use of sampling frame

- Business Register (**BR**) incl. classification, mea-of-size, etc.
- Central Population Register (**CPR**) for household surveys
- **Address** file including post code, etc. for area sampling
- Future: **Geo-referenced** Immobility Register **(IR)**
  - immobility: property, land, natural resource, etc.
  - **multistage sampling**: **cost** vs. **efficiency**
  - traditionally: fixed clusters & random sub-sampling
  - use **Geo-IR** for **dynamic** clustering in order to **maximize within-cluster variance**

Register data as auxiliary information

- Reducing **sampling error**

- household: **post-stratification** and **calibration**
- business: **ratio** and **regression** estimation
- indirect **small area estimation** at detailed levels

- Reducing **nonresponse error**

- **response enhancement** during data collection
- **statistical adjustment** after data collection
- **bias exploration** throughout the statistical processes

Register data for use beyond survey sampling paradigm

- Register = **auxiliary** data in survey sampling paradigm

- Extended roles under **data integration paradigm**

- register as **target** data (e.g. Wallgren & Wallgren, 2006)
- register as **proxy** data
  - lack inherent relevance due to nature-of-source
  - NB. **proxy** $\neq$ **auxiliary**
  - integration (incl. survey, census) to **satisfy quality requirements** incl. **statistical relevance**

Illustrating statistical vs. definitional relevance

- **ILO**-employment status in Labour Force Survey (**LFS**) $\neq$ register-based (**R**) employment status by **definition**

- Individual equality **not** expected in general even with perfect register data = **lack** of **definitional relevance**

- Process register-based employment status such that

  - register-employment total = LFS-employment total
  - **achieve statistical relevance in this respect**
  - at detailed levels, R-employment total has **smaller MSE** than LFS-estimates (Fosen & Zhang, 2011)

Use of VAT data for STS

- Alternative approaches by (data source, coverage):

1. **BR** + **MBS** Monthly Business Survey (all units)

2. **BR** + **VAT/MBS** (super units) + **MBS** (rest units)

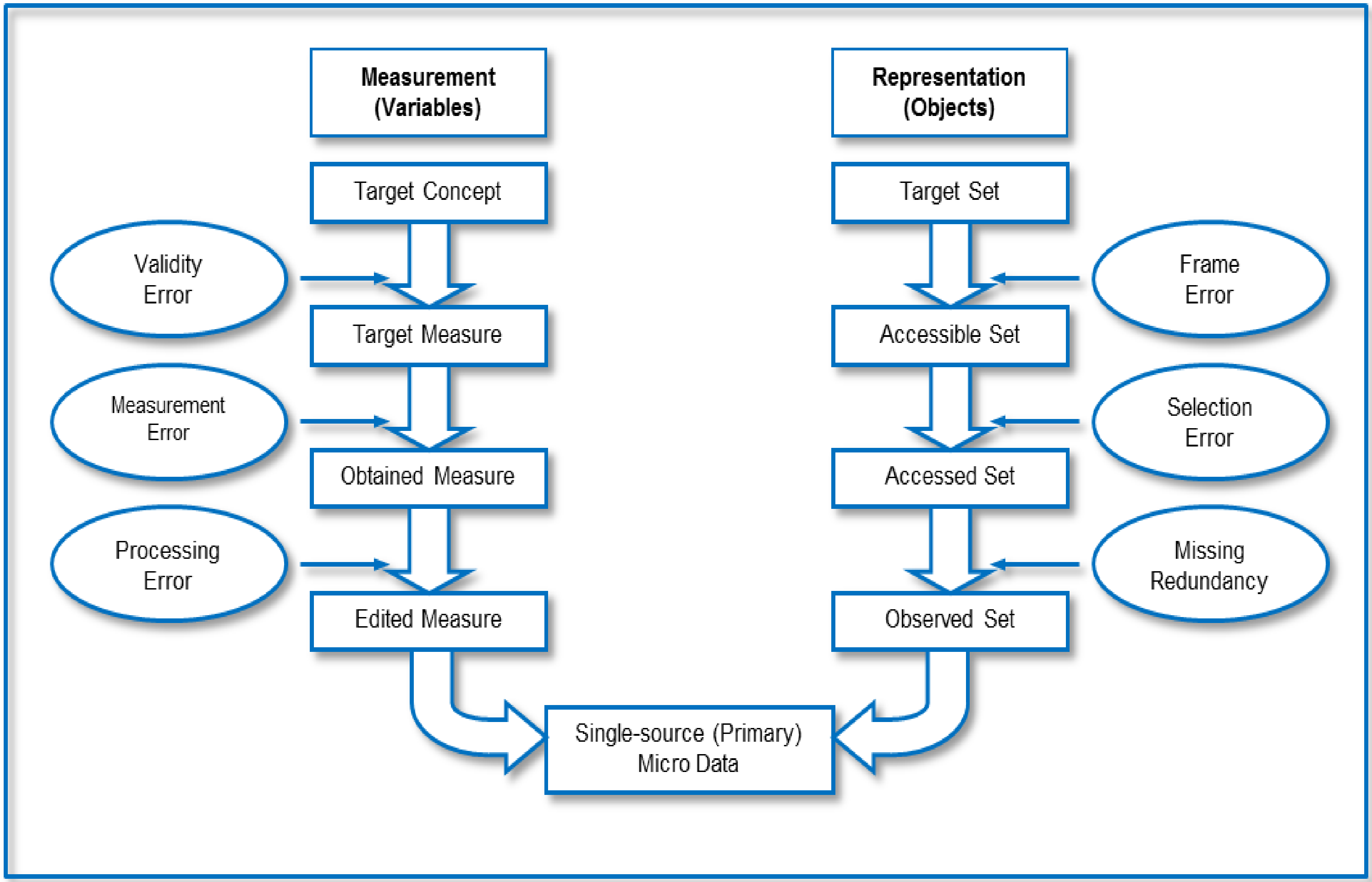
3. **BR** + **MBS** (largest units) + **VAT** (rest units)

4. **BR** + **VAT** (all units; in **retrospect**)

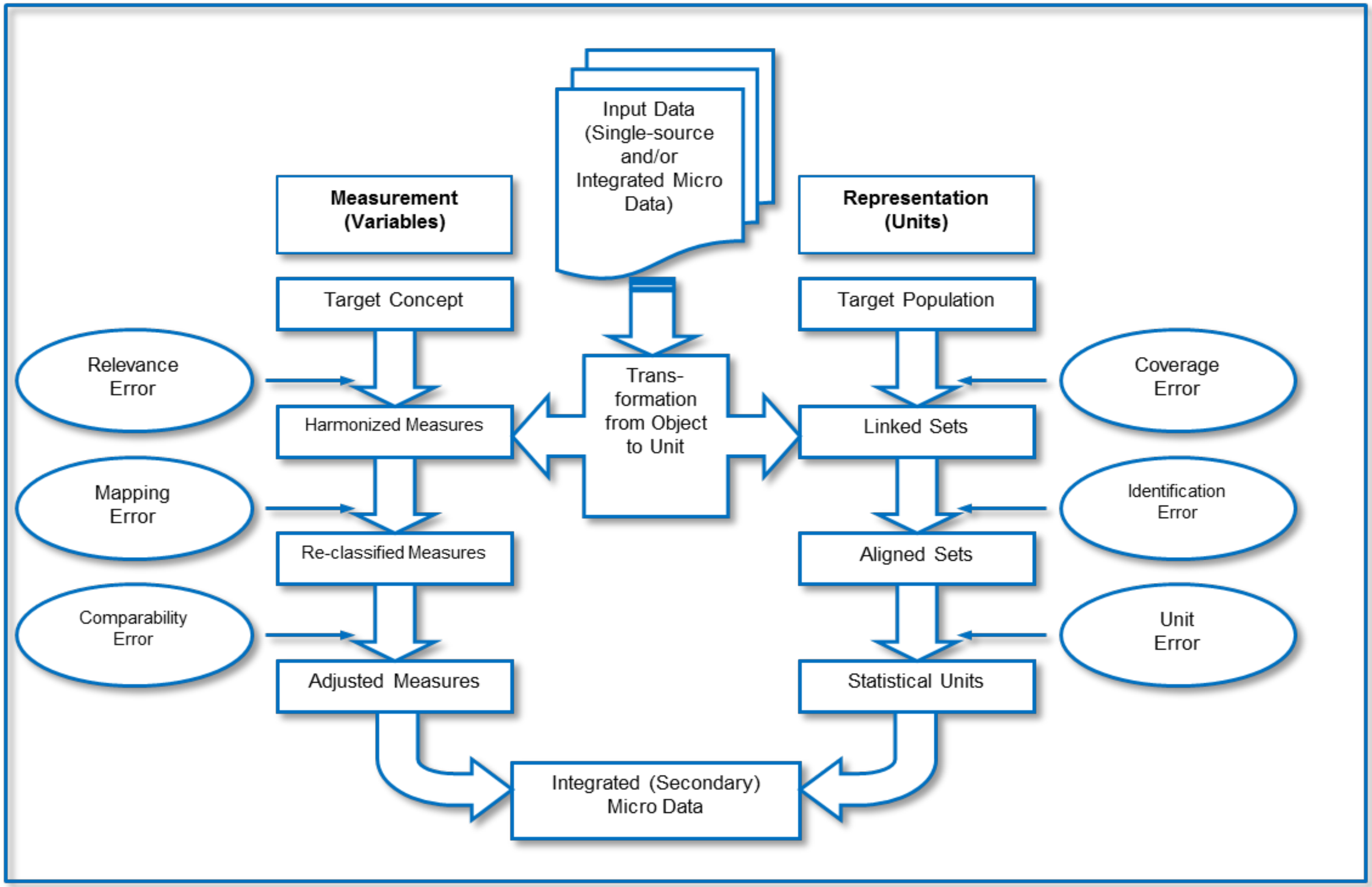
- Key issues

- **target population** & **classification** on combined sources
- **relevance** & **compatibility** between survey and registers
- **timeliness** vs. **burden/resource**

A two-phase life-cycle model for error sources (I): primary-source data



A two-phase life-cycle model for error sources (II): secondary integrated data



Units and measurement

- **Representation**: **units** in **multiple sources**

- **alignment** of business, statistical & VAT units
- ref. **aligned sets** & **identification error**
- a **unit-error theory** (Zhang, 2011)

- **Measurement**:

- ref. **reclassified measures** & **mapping error**
- **apportion** btw. units & **calendarization** over periods
- depends on **alignment** under **representation**

Longitudinal progressive nature of administrative data

- **Longitudinal** data for different time points of interest

- **Progressive measurement**: available value for a given time point of interest may **evolve** over time

  - administrative data typically event-triggered
  - **delay, error & change** of registration
  - distinct feature compared to sample survey & census

- Inference for register-based statistics requires **modelling** (Zhang and Pritchard, 2013)

Illustration: progressive Business Register (Hedlin et al. 2006)

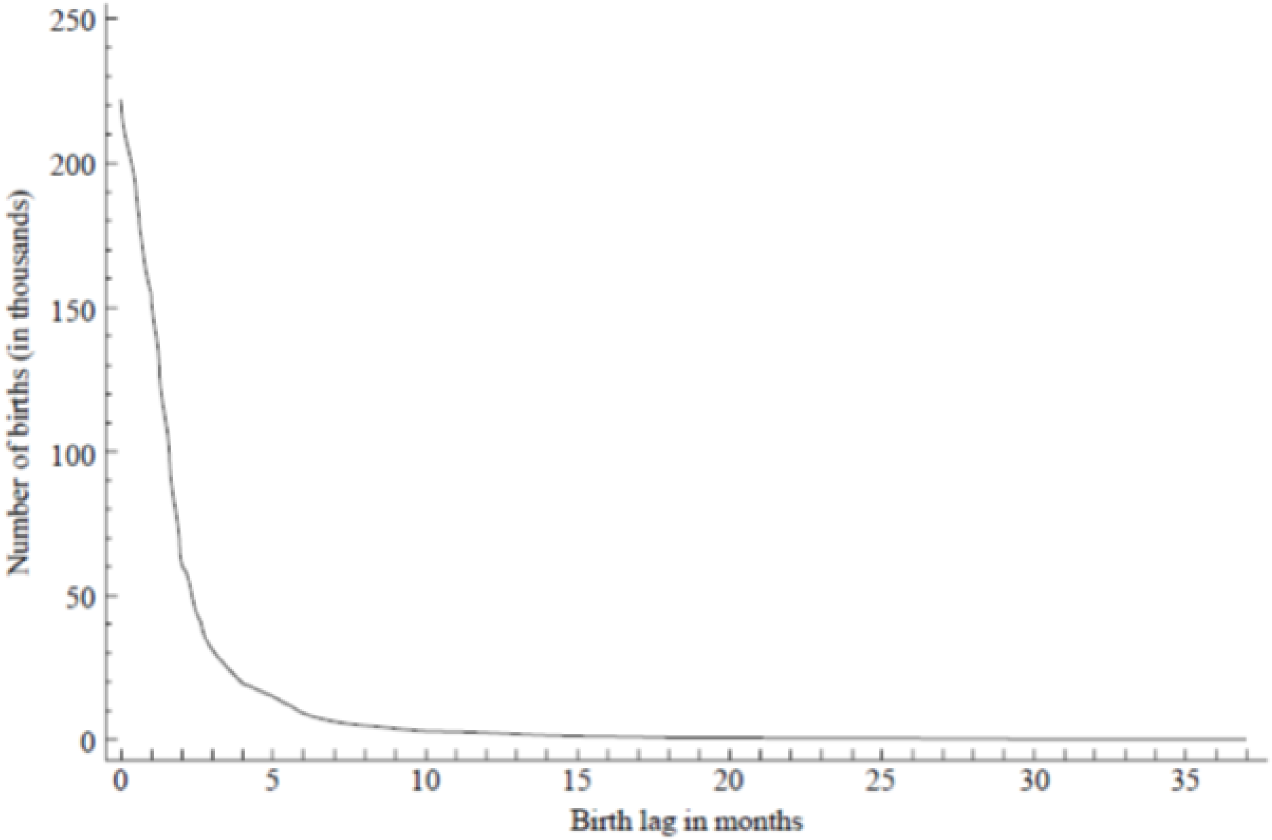


Fig. 1. Number of observed births (in thousands) against birth lag (months)

Illustration: progressive Employment Register

(Zhang & Fosen, 2012)

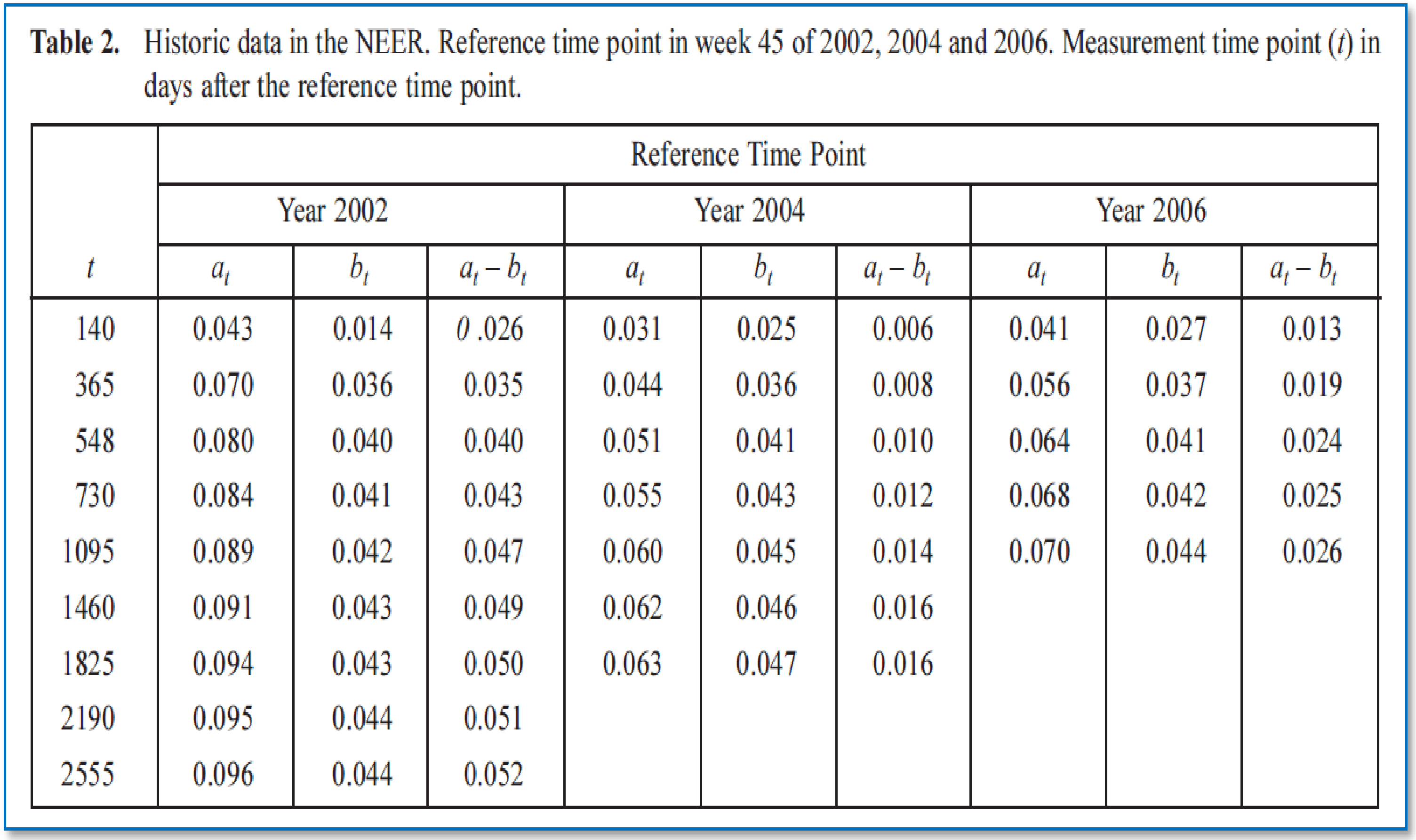


Table 2. Historic data in the NEER. Reference time point in week 45 of 2002, 2004 and 2006. Measurement time point (t) in days after the reference time point.

t	Reference Time Point								
	Year 2002			Year 2004			Year 2006		
	a_t	b_t	$a_t - b_t$	a_t	b_t	$a_t - b_t$	a_t	b_t	$a_t - b_t$
140	0.043	0.014	0 .026	0.031	0.025	0.006	0.041	0.027	0.013
365	0.070	0.036	0.035	0.044	0.036	0.008	0.056	0.037	0.019
548	0.080	0.040	0.040	0.051	0.041	0.010	0.064	0.041	0.024
730	0.084	0.041	0.043	0.055	0.043	0.012	0.068	0.042	0.025
1095	0.089	0.042	0.047	0.060	0.045	0.014	0.070	0.044	0.026
1460	0.091	0.043	0.049	0.062	0.046	0.016			
1825	0.094	0.043	0.050	0.063	0.047	0.016			
2190	0.095	0.044	0.051						
2555	0.096	0.044	0.052						

Note: (a_t, b_t) = (increase, decrease) in employment rate due to updating by time t

Problems regarding target population

- **Target population**: **VAT-active** units in period t

- **Target total** turnover over units in target population

- **VAT turnover y(t;s)** available at time s for $s \geq t$

  - $y(t;s) = \text{NA}$ if **no value** reported for t by s

  - assume **negligible error**, i.e. $y(t;\infty) = y(t;s)$ if $y(t;s) \neq 0$

- Population-t measured s: units with $y(t;s) \neq \text{NA}$

  - **activity delays**: $y(t;s) = \text{NA}$ and $y(t;\infty) \neq 0$

  - **inactivity delays**: $y(t;s) = \text{NA}$ and $y(t;\infty) = 0$

  - **birth delays**: units 'non-existent' at s but $y(t;s) \neq 0$

Methodology regarding target population

- **Birth delays**
  - **typically ignored**; may have limited impact
  - more appropriate: **prediction** at **aggregated level**
- **Activity & inactivity delays**
  - **typically**, categorical classification by **sign-of-life**
  - *ad hoc* division between activity and inactivity delays
  - more appropriate: **prediction** for **existent units**

A prediction framework: existent units

Contribution of existent unit i to target total

$$y_i(t)I_i(t)$$

where $I_i(t) = 1$ if unit is active, and 0 if inactive

- **sign-of-life approach**: set $I_i(t) = 1$ if so-and-so, and 0 otherwise $\Leftrightarrow$ *ad hoc* method in nature

  over-estimation of active delay total more likely?

- **prediction approach**: **joint modelling** of $(y_i(t), I_i(t))$ for all existent units by time s

A prediction framework: target total

A decomposition of target total for t predicted at s

- **reported total**: directly observed

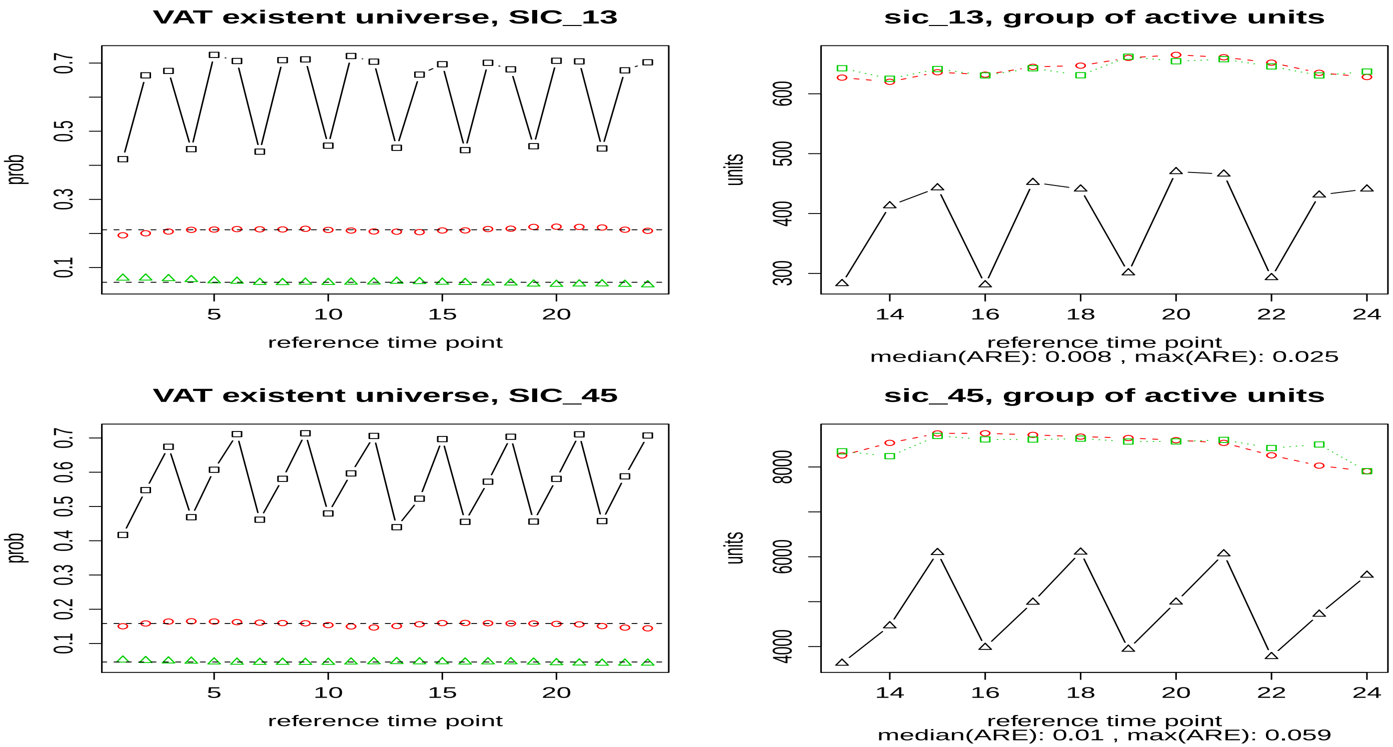
  NB. under the assumption of negligible reporting errors

- **birth delay total**

- **report delay total** of non-reporting existent units

  - avoid *ad hoc* treatment of $I_i(t)$ in practice
  - reporting existent units form a sample of all existent units
  - report sample **not selected by design**
  - extending informative sampling/nonresponse theory

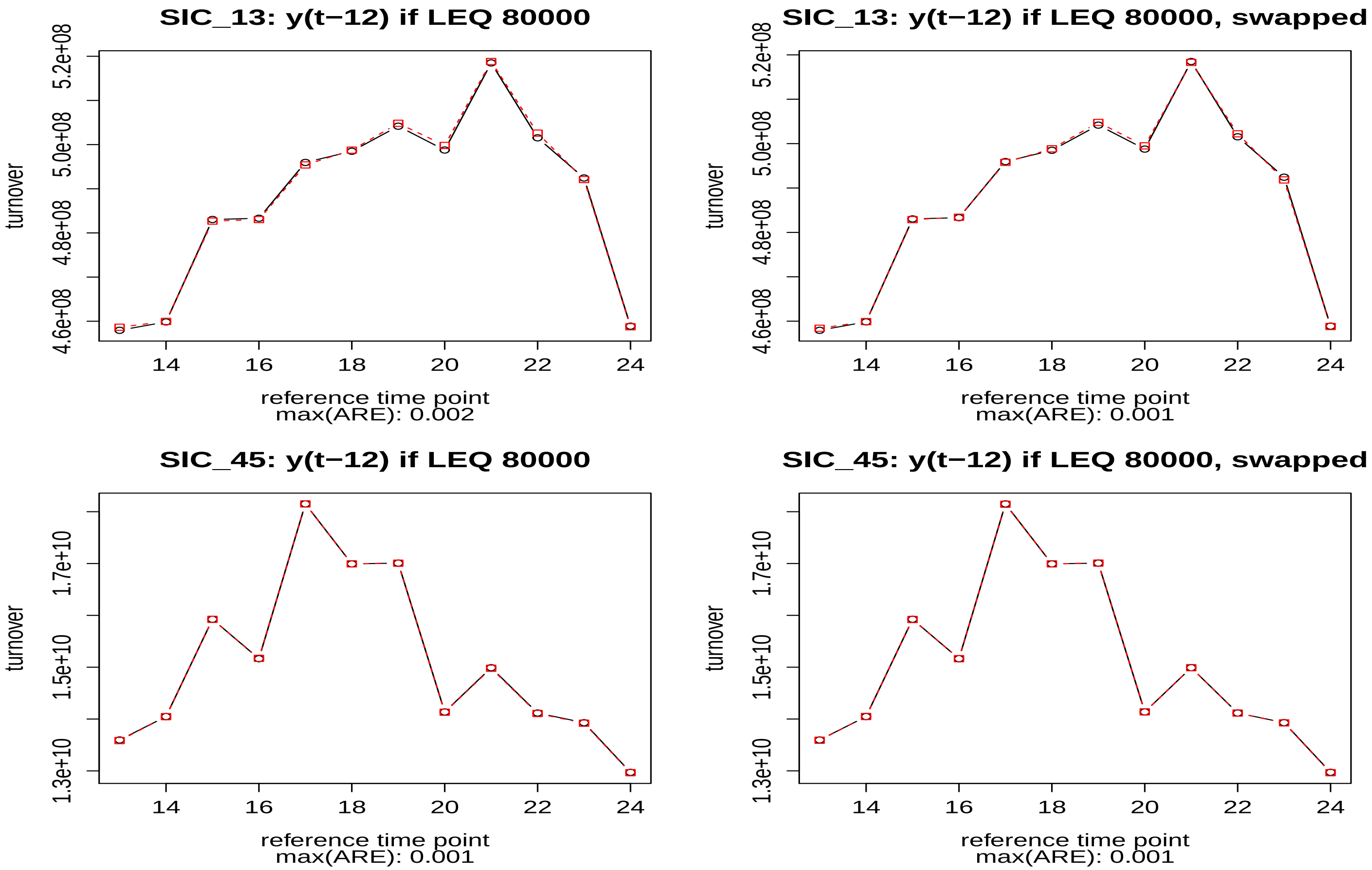
Illustration based on VAT register in UK



Left: reporting rate at $t + 3$ (square), proportion of units with turnover $\geq$ 80000 (circle), proportion of inactive units (triangle); reference time points in 2010 and 2011.

Right: number of reporting existent units at $t + 3$ (triangle), population size (circle) and predicted size (square) of units with turnover $\geq$80000; reference time points in 2011.

A substitution exercise based on UK data



Turnover total of existent units (circle), band-wise substitution of $t-12$ values for non-large units (square) for SIC-13 (80% units) and SIC-45 (85% units), with and without swapping

Developing an approach to minimize survey compliance

- Monthly self-representing sample of the largest units

NB. **cut-in** threshold; emerging in-scope units; outdated units

- **Register-based prediction for the rest**

- **birth delay total** by **projection/forecast**
- **cut-off existent units** (80%+) by **prediction**
- **in-between** existent units (btw cut-off & cut-in thresholds)

  NB. possible supplementary sample due to trade-off between timeliness and accuracy, difference across NACE, etc.?

Final remarks

- **Better uses** of **VAT register** for

- construction of the **target population**
- selection and maintenance of **certainty sample**
- **survey exemption** of majority of units
- **minimizing sample** of remaining units

- For long-term development: **improving integration of VAT register & BR**, in terms of **representation and measurement**, for uses across business statistics

References

[1] Fosen, J. and Zhang, L.-C. (2011). The approach to quality evaluation of the microintegrated employment statistics. In *WP4 Report: Case Studies, pp. 25 - 38. ESSnet on Data Integration.*

[2] Hedlin, D., Fenton, T., McDonald, J.W., Pont, M. and Wang, S. (2006). Estimating the under-coverage of a sampling frame due to reporting delays. *Journal of Official Statistics*, vol. 22, pp. 53 - 70.

[3] Wallgren, A. and Wallgren, B. (2006). *Register-based Statistics - Administrative Data for Statistical Purposes*. John Wiley & Sons, Ltd.

[4] Zhang, L.-C. and Pritchard, A. (2013). Short-term turnover statistics based on VAT and Monthly Business Survey data sources. Paper presented at *the 3rd European Establishment Statistics Workshop*, Nuremberg.

[5] Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, vol. **66**, pp. 41-63.

[6] Zhang, L.-C. and Fosen, J. (2012). A modelling approach for uncertainty assessment of register-based small area statistics. *Journal of the Indian Society of Agricultural Statistics*, vol. **66**, pp. 91 - 104.

[7] Zhang, L.-C. (2011). A unit-error theory for register-based household statistics. *Journal of Official Statistics*, vol. **27**, pp. 415-432.