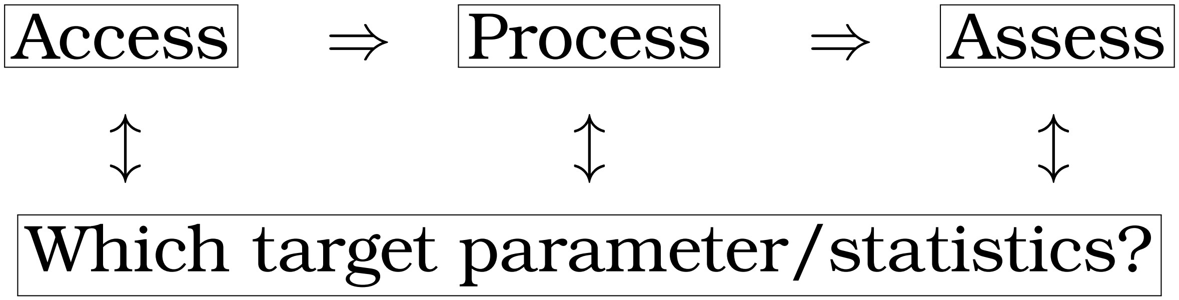
Introduction to
Session 3: Methodologies for big data

Li-Chun Zhang

Non-survey big data sources (Zhang & Haraldsen, 2022)

Type of source	Example of data
Register	vital event, diagnosis
	wage, income tax, VAT, welfare payment
Transaction	scanner data price, point-of-sales receipt
	bankcard or giro payment
	B2B or B2P invoice
	property sales contract
Remote sensing, fixed	smart meter reading
	weather station reading
	traffic loop signal
Remote sensing, mobile	satellite image, drone image
	airborne laser scanning
	maritime AIS, lorry tracking signal
	mobile phone signal
Internet	web page
	social media post

Some key challenges



Access ⇒ Process ⇒ Assess

↕ ↕ ↕

Which target parameter/statistics?

Presentations today more related to **Process**

- Measurement: from organic data (text/image/...)
- Representation: over-/under-coverage, selectivity...
- Integration of multiple sources

Perspective: **Total error** of Process pipeline

- Zhang (2012), Reid et al. (2017), Rocci, et al. (2022)
- Administrative registers... non-survey big data
- Oscillation: generic/standard vs. stovepipe process

Access: confidentiality and data minimisation

> "Survey respondents are usually provided with an assurance that their responses will be treated confidentially. These assurances may relate to the way their responses will be handled within the agency conducting the survey or they may relate to the nature of the statistical outputs of the survey…" (Skinner, 2009).
> … personal data must be "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed" (GDPR)

NSO/NSI is not state-sponsored Facebook
Renewed urgency in the present context

- UNECE: Input Privacy-Preservation Project
- Secure big data collection and processing: Framework, means and opportunities (Zhang & Haraldsen, 2022): to be implemented for Transaction (receipt, debit card)

Assess: Audit sampling inference

Wherever the goal of survey sampling is to produce a point estimate of some target parameter of a given finite population, auditing aims not to estimate the target parameter itself but some chosen error measure of any given estimator of the target parameter, which may be biased due to failure of the underlying model assumptions or other favourable conditions that are necessary.

The framework of inference is design-based given a finite population, from which the random sample is taken under a probability design, but the outcomes of interest and other values known separately from sampling are treated as fixed. Design-based auditing inference is valid regardless the models or algorithms underlying the estimator being assessed. (Zhang, 2021).

Assess: Audit sampling inference

Inference/ Property	Motivation of Estimator $\hat{\theta}$	
	Design-based	Model-based
Design-based	Survey sampling	Audit sampling
Model-based	e.g. “Weighting is inefficient”	Prediction

E.g. model estimator can have smaller design MSE

~~Existence of true model or infallible learning~~

Some applications

- Scanner-data proxy expenditure weights for CPI much better than using Expenditure Survey (Zhang, 2021)
- Dutch CCI and SMI, Corr = 0.88 (Daas et al., 2015) Patone and Zhang (2021): a test for $\mathcal{H}_0 : \ \theta_t - \xi_t = \mu$ where $\theta_t = E(\hat{\theta}_t)$ for audit CCI $\hat{\theta}_t$, assume SMI $\xi_t = E(\hat{\xi}_t)$
- A statistical framework for register-based population size estimation (Bernardini, et al., 2022)

References

[1] Bernardini, A., Cibella, N. and Solari, F. (2022). A statistical framework for register based population size estimation, Technical Report, Istat Advisory Committee on Statistical Methods 2022 Meeting, Rome, Italy.

[2] Daas, P.J., Puts, M.J., Buelens, B., and van den Hurk, P.A. (2015). Big data as a source for official statistics. *Journal of Official Statistics*, 31:249-262.

[3] Patone, M. and Zhang, L.-C. (2021). On two existing approaches to statistical analysis of social media data. *International Statistical review*, 89:54-71. `DOI: 10.1111/insr.12404`

[4] Reid, G., Zabala, F. and Holmberg, A. (2017). Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ. *Journal of Official Statistics*, 33:477-511. `10.1515/jos-2017-0023`

[5] Rocci, F., Varriale, R. and Luzi, O. (2022). Total Process Error: An Approach for Assessing and Monitoring the Quality of Multisource Processes. *Journal of Official Statistics*, 38:533-556. `10.2478/jos-2022-0025`

[6] Skinner, C.J. (2009). *Statistical Disclosure Control for Survey Data*. In: Pfeffermann, D and Rao, C.R. eds. Handbook of Statistics Vol. 29A: Sample Surveys: Design, Methods and Applications. pp. 381-396.

[7] Zhang, L.-C. (2021). Proxy expenditure weights for Consumer Price Index: Audit sampling inference for big-data statistics. *Journal of the Royal Statistical Society, Series A*, 184:571-588. `DOI:10.1111/rssa.12632`

[8] Zhang, L.-C. and Haraldsen, G. (2022). Secure big data collection and processing: Framework, means and opportunities. *Journal of the Royal Statistical Society, Series A*. `DOI:10.1111/rssa.12836`