MNO-MINDS WP3: Methods for combining MNO and non-MNO data

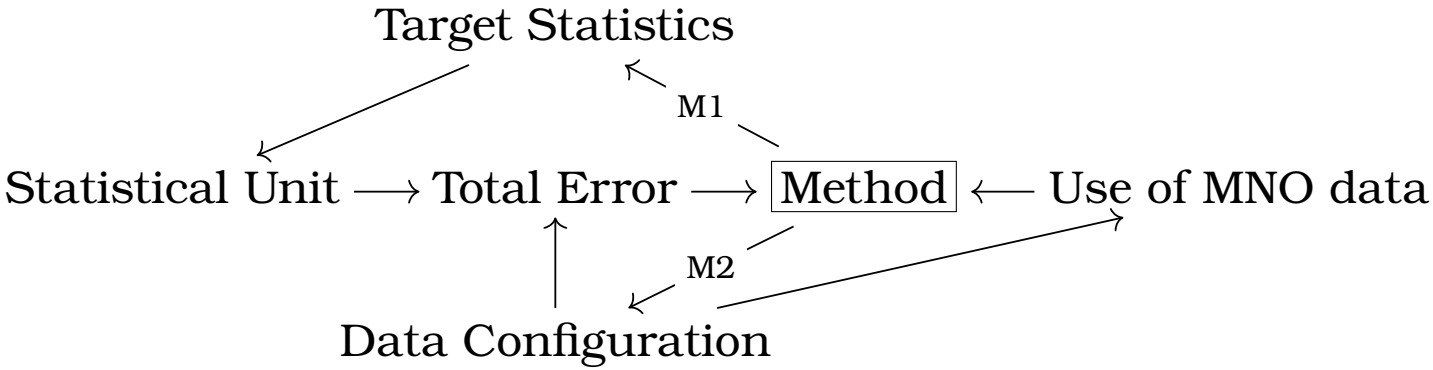
Li-Chun Zhang

Statistisk sentralbyrå (lcz@ssb.no)
University of Southampton

This work was co-funded by the European Commission Project "MNO-MINDS" - 101132744 — 2022-IT-TSS-METH-TOO.

Reference frame of methods

Population, mobility, tourism, environment



Which target statistics, e.g. present population?
Statistical unit? Measurement unit?
Available MNO and non-MNO data, how are they defined?

NB. nano, micro, or macro MNO data

Use of MNO data: target, auxiliary or proxy
Total error: the most important ones in given situation
M1 method: applicable to available data
M2 method: changes available data (& M1 method)

Three broad approaches

Depending on how the associated uncertainty is defined

- **Randomisation** requires a specialised survey to convert the MNO data into the target statistical outputs, the uncertainty of which is considered to be dominated by the survey sampling error.
- Although MNO data are not observed by known probabilities, one may introduce a model of the underlying mechanism *as if* they were, and assess uncertainty accordingly. **Quasi-randomisation** approach is applicable together with suitable non-MNO population data, which can potentially remove the need of specialised surveys.
- It is often possible to build **super-population** models for specific variables using data from MNO and non-MNO sources. Different models are needed for different statistics generally, unlike building quasi-randomisation model that is applicable to all the different variables associated with the same mobile devices.

Inference basis a known sampling design or an assumed model, target-agnostic observation or specific outcome?

Example: long-term *de facto* residents

Target statistics under topic Population, unit **= resident**
Target total Y_i for municipality $i = 1, ..., n$, $\sum_i Y_i = N$
de facto present MNO **device** counts m_i (longitudinal)

- Estimate $w_i = m_i/Y_i$ by surveying sample of persons (out of Y_i)
  $w_i = \xi_i \eta_i$ where ξ_i is #devices (in m_i) per user and η_i = #users/Y_i
  Issues: how to identify devices in m_i, how to cover all users?
- Quasi-randomisation $\hat{Y}_i = m_i N/m$, where $m = \sum_{i=1}^n m_i$ and
  $N = \sum_{i=1}^n N_i = \sum_{i=1}^n Y_i$ given known *de jure* population sizes N_i
  Issue: completely random selection m_i from Y_i plausible?
  Acceptable QR selection model otherwise?
- Let y_i be design-based estimator of Y_i by survey sampling above
  Obtain $\hat{\mu}(m_i, \cdot)$ by super-population model $E(y_i) = E(Y_i) = \mu(m_i, \cdot)$
  Issue: bias due to model misspecification?
  Shrinkage estimator $\tilde{\mu}_i = \gamma_i y_i + (1 - \gamma_i)\hat{\mu}_i$?
  Other models given $\{y_{ij}\}$, sampled from N_i observed among Y_j?

Main topics for method development

Randomisation

- Transfer learning
- User ambiguity

Quasi-randomisation modelling

- Basic selection model, general purpose
- Potential complications

Super-population modelling: statistical calibration

- Similar to scientific calibration of measurement
- Spatial, network, compositional data

Super-population origin-destination flow modelling

- Origin-destination models
- Network flow models, mathematical & statistical

Specific use-cases or applications

Thank you for your attention

MNO data

Event data:

(d, t, j) = (device, time, cell-ID)

NB. unknown physical location (i), known cell-ID j
conditional support of i given j: range of antenna at j

Device data:

$$\{(t_d, \tilde{i}_{d,t}) : t_d = t,\ \exists(d, t, j)\}, \quad \forall d$$

NB. MNO deterministic mapping $j_{d,t} \to \tilde{i}_{d,t}$
support of $\tilde{i}_{d,t}$ depends on output (statistical purpose)

Output (aggregated device) data:

$$m_{ik}^{t_0 t_1} = \sum_d \mathbb{I}\big(z_d^{t_0 t_1} \neq \emptyset\big)\, \mathbb{I}\big(g(z_d^{t_0 t_1}; i, k) = 1\big)$$

$$z_d^{t_0 t_1} = \{(t_d, \tilde{i}_{d,t}) : t_0 \leq t_d \leq t_1\}$$

e.g. $g(z_d^{t_0 t_1}; i, k) = \mathbb{I}(\tilde{i}_{d,t_0} = i)\mathbb{I}(\tilde{i}_{d,t_1} = k) = 1$

define g according to statistical purpose

NB. only macro output data are accessible
NB. however, micro data integration potentially possible by confidential multiparty computing methods