Innovative data for official statistics: Methodological challenges

— Discussion

Li-Chun Zhang

Smart surveys, more troubles

Diversification of survey mode with smart survey

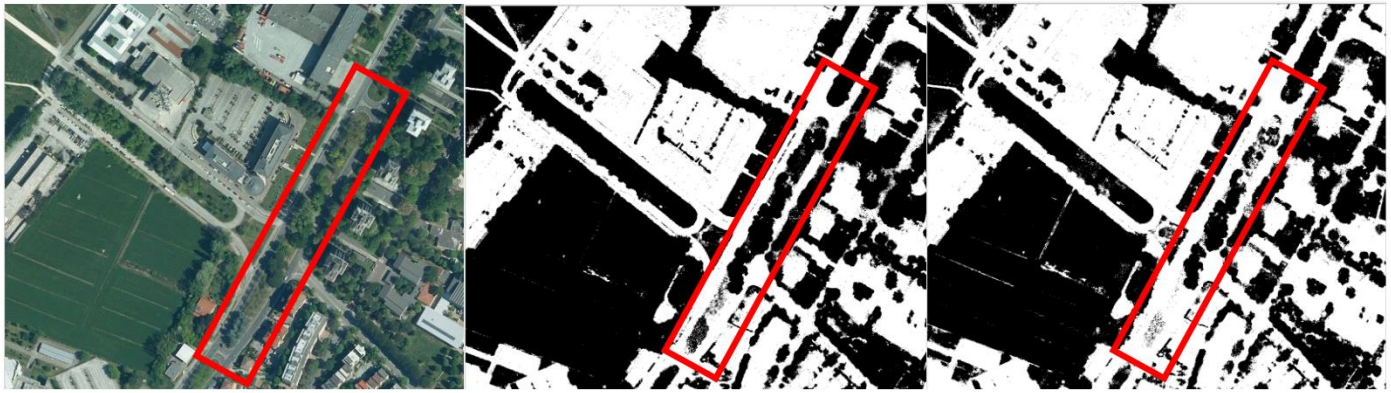
- face-to-face interview (not smart)
- 'stationary': paper, phone, web, ... (not smart, but...)
- mobile: smart phone, tablet, ... (smart)
  - online questionnaire (not really smart)
  - NSI app/sensor/...
    active (self-reporting) vs. passive (cloud-sync)
  - 3rd-party app/sensor/transaction/...
    active (self-reporting) vs. passive (NSI-on-behalf)

Burden of active participation: smart vs. non-smart?
Informed consent of sharing instead of participation?
If yes, then ensuing issues:

selection, mode†, legality, ability, capacity, ...

†: ML-transformed instead of respondent-digested

Valid uncertainty assessment



Target: $Y = \sum_{i \in U} y_i$ of all the pixels $U = \{1, ..., N\}$
Probability sample $s \subset U$ and observe $\{y_i : i \in s\}$
Denote by μ any given ML model or algorithm
Sanguiao-Sande and Zhang (2021): designed-unbiased

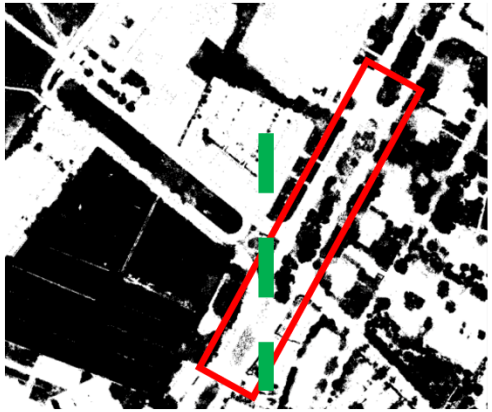
$$E_p\left\{\hat{Y}(s;\mu)\right\} = Y$$

Zhang, Sanguiao-Sande and Lee (2023): design-unbiased predictive inference of $\hat{Y}^{RB} = \sum_{i \in s} y_i + \sum_{i \in U \setminus s} \bar{\mu}_i(s)$, e.g.

$$E_p(\hat{Y}^{RB} - Y), \quad E_p\{(\hat{Y}^{RB} - Y)^2\} \quad \text{or} \quad E_p\{\sum_{i \in U \setminus s} (\bar{\mu}_i - y_i)^2\}$$

Explainability, of what?

For the given pipeline-S2 algorithm for satellite images, there is a question, 'what is the algorithm doing?'
E.g. local interpretable model-agnostic explanation (LIME)



However, the bigger QUESTION of valid inference is not 'which model?' but 'how is a model being used?"
By design-based predictive or auditing inference, one can obtain, say, an estimated MSE of green area total in Rome by S2-algorithm, and the MSE-estimator is *unbiased over repeated sampling* of s and $\{y_i : i \in s\}$ used for estimation.