# Using New Forms of Data in Small Area Estimation

Nikos Tzavidis* & Angela Luna†
Southampton Statistical Sciences Research Institute
University of Southampton

Jessica Steele & Kristine Nilsen
WorldPop
University of Southampton

ITACOSM
Florence, June 5-7 2019

**The research is supported by the MAKSWELL grant - EU-Horizon 2020**

https://www.makswell.eu

*Presenting author
†Presenting author

# Introduction

- Recent interest in the use of new data forms in SAE
  - Remotely sensed data
  - **Mobile phone (CDR) data**
  - Web-scraped data
  - Bank transaction data
- Used as possible source of
  - Response data
  - Auxiliary information

# Motivating the use of new data forms

- Potentially useful in low resource settings
- A typical data scenario in such settings
  - Survey data on demographic & income/wealth available
  - Census data unavailable or infrequently updated
  - Administrative data unavailable

# New forms of data - Some pros and cons

Pros

- Dynamically updated covariates (compare this to Census data)
- Reduced cost of data collection
- More flexible definition of geography

Cons

- Not clear how to extend definition of geography to domains
- Errors in data more difficult to quantify and account for
- Coverage / representativity
- Limited to what can be measured from new data sources
- No obvious reason why covariates correlate with the outcome

# Literature

- Growing body of literature using new data forms
- Targets of estimation at very fine spatial scales
- In most cases mainstream small area literature is ignored with potential consequences (see later in the presentation)
- Some exceptions (see Marchetti et al., 2015; Schmid et al., 2017; Münnich et al., 2019)

# First observations

- Various, mainly supervised methods (Hastie et al., 2008)
- Methodology usually combines survey data with new forms of data and fits a model then used to predict target parameters
- Validation uses correlation plots with estimates derived via alternative data sources
- Commonly used models are not well described
- Challenging for research reproducibility

# Presentation aims

- No new methodology is introduced

Part 1

  - Use SAE model-based methods with new forms of data as covariates
  - Derive SAE point & MSE estimates

Part 2

  - Attempt to decipher a typical model used outside the SAE literature
  - Discuss possible issues with these models
  - Present possible solutions
  - Assess the impact of possible model misspecification on SAE

# Data & Models

- New forms of data processed at very fine spatial levels
- Tempting to specify models at the level the data is available
- Likely to lead to synthetic estimates (implications not discussed in the non-SAE literature)
- Here we use area-level models by aggregating data at the target geography

# Area level models: The Fay-Herriot model

Sampling model

$$\hat{\theta}_i^{direct} = \theta_i + e_i$$

- $\hat{\theta}_i^{direct}$ is a direct design-unbiased estimator
- $e_i$ is the sampling error of the direct estimator

Linking model

$$\hat{\theta}_i^{direct} = \boldsymbol{x}_i\boldsymbol{\beta} + u_i + e_i, \quad i = 1, \ldots, m,$$

where $u_i \sim N(0, \sigma_u^2)$ and $e_i \sim N(0, \sigma_{e_i}^2)$, with $\sigma_{e_i}^2$ assumed known

# Area level models: The Fay-Herriot estimator

The EBLUP under the **Fay-Herriot** (FH) model is

$$\begin{aligned}\hat{\theta}_i^{FH} &= \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}} + \hat{u}_i \\ &= \gamma_i \hat{\theta}_i^{direct} + (1-\gamma_i)\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}},\end{aligned}$$

- Analytic MSE estimator of $\hat{\theta}_i^{FH}$ - Prasad & Rao (1990)
- Alternatively use bootstrap (parametric under the FH model)

# Case study

Poverty estimation in Bangladesh using Wealth Index (WI) as proxy

Aim: Estimate average WI by Upazila (Level 3)

## Survey Data Sources - DHS 2014

- $n = 17$K households
- Stratified 2-stage cluster design
- At least one cluster selected in 365/508 (72%) Upazilas
- Response: WI computed via PCA
- Average Upazila sample-size $\bar{n}_i = 34$

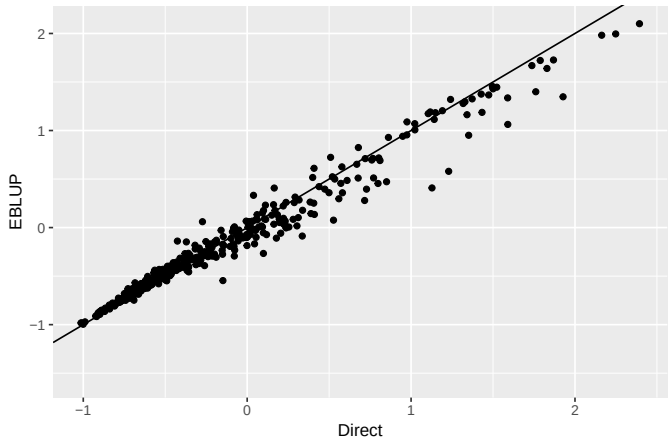
# Case study

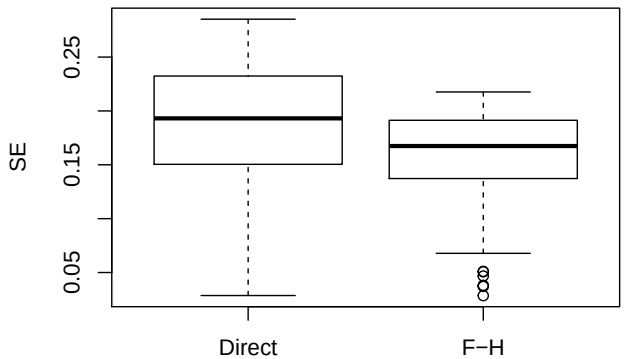
## Auxiliary data sources

- Remote sensing covariates
  - Processed at 1km spatial resolution
  - Aggregated at Upazila level
  - Enhanced vegetation index (EVE)
  - Elevation (ELEV)
  - Accessibility to areas with more than 50K people (ACC)
  - Night time lights (NL)
- CDR data: Ongoing

# Case study - Direct and FH estimation

- Survey weighted direct estimates of $\overline{WI}_i$ at upazila level.
- Estimated variances of the direct:
  - Ultimate cluster variance (UCV) estimator
  - DEFT (One cluster in some Upazilas $\rightarrow$ UCV not applicable)
  - Smoothing via $GVF(WI_i^{1,1/2,1/3}, n_i^{1,1/2})$
  - Ignoring PCA variability
- EBLUPs & Prasad-Rao MSEs under a FH model with RS covariates.



# Case study - Non-SAE literature

- Not easy to decipher the models used - Black box approach
- We use code from one of these papers for comparison reasons
- A Linear Latent Gaussian Model at Upazila level is used
- R-INLA (approx Bayesian inference) used for estimation
- The model also allows for spatial correlation
- We turn this off for direct comparison with FH estimates

# Case study - Non-SAE literature - Model Choices

- In all cases, $u_i$ are iid with variance $\sigma_u^2$
- $e_i$ are independent with variance $\sigma_{e_i}^2$
- Normality of both is assumed

  INLA 1 $\sigma_{e_i}^2 = \sigma_e^2$ assumed unknown

  INLA 2 $\sigma_{e_i}^2 = s_i\sigma_e^2$. $s_i = \frac{v_i}{\bar{v}_i}$ fixed but $\sigma_e^2$ unknown.

  INLA 3 INLA 2 with a highly informative prior for $\tau = 1/\sigma_e^2$.
  $\tau \sim Gamma(shape = 25^2/0.1, rate = 25/0.1)$,
  therefore $E(\tau) = 25 = 1/\bar{v}_i, \quad V(\tau) = 0.1$
- Literature on HB framework for FH models (see You & Chapman, 2006 ; Poletini, 2017)
- Not tested in this presentation

# Case study - Non-SAE literature - Possible pitfalls

- INLA 1 is assumed by the paper
- One observation per Upazila
- Why expect the model to be identifiable ?
- INLA 2 introduces a heteroscedastic structure
- However, $\sigma_e^2$ unknown $\rightarrow$ identifiability?
- INLA 3 uses a highly informative prior on $\sigma_e^2$

# Case study - Results fixed effects

- Same spec as is the non-SAE paper
- Small differences in the estimates of the fixed effects
- FH & INLA 3 almost identical

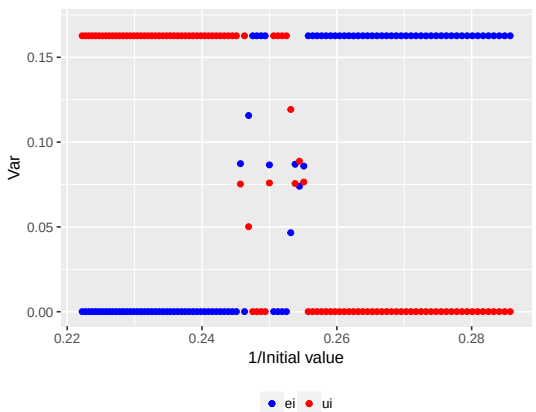
| Variable | Fay-Herriot | | INLA 1 | | INLA 2 | | INLA 3 | |
|---|---|---|---|---|---|---|---|---|
| | Estimate | S.E. | Estimate | S.E. | Estimate | S.E. | Estimate | S.E. |
| (Intercept) | 0.922 | 0.139 | 0.941 | 0.148 | 0.861 | 0.135 | 0.921 | 0.138 |
| evi | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| elev | -0.152 | 0.030 | -0.142 | 0.032 | -0.166 | 0.028 | -0.152 | 0.030 |
| nl | 0.385 | 0.030 | 0.372 | 0.031 | 0.409 | 0.030 | 0.386 | 0.030 |
| acc | -0.084 | 0.034 | -0.090 | 0.036 | -0.068 | 0.033 | -0.084 | 0.034 |

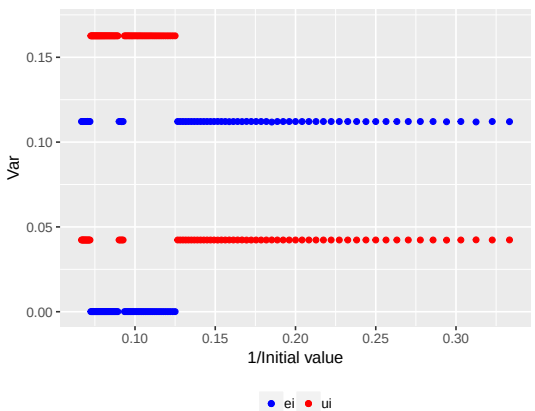
# Case study - Results variance components

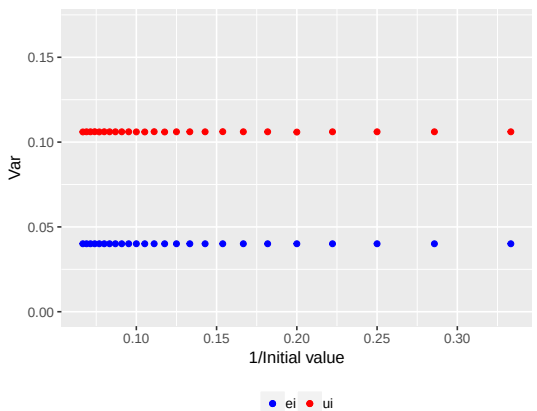
- Large differences in the estimated variance components

| Var.comp | Fay-Herriot | INLA 1 | INLA 2 | INLA 3 |
|---|---|---|---|---|
| $\hat{\sigma}_e^2$ | 0.0391 | 0.0865 | 0.1121 | 0.0401 |
| $\hat{\sigma}_u^2$ | 0.1091 | 0.0759 | 0.0423 | 0.1059 |

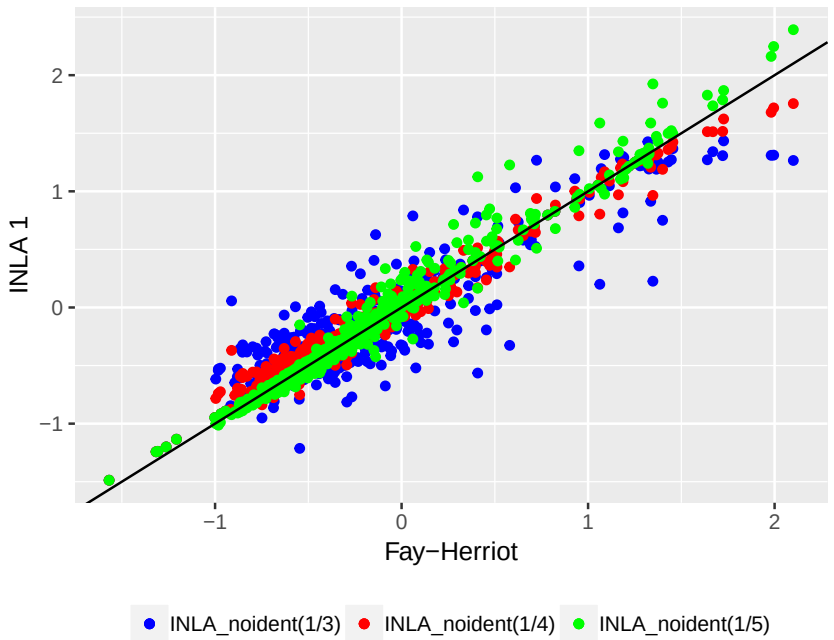
- FH & INLA 3 almost identical
- INLA 1 & 2 differ
- Sensitivity analyses, change the starting values of $\sigma_e^2$
- Other starting values set to default INLA ones

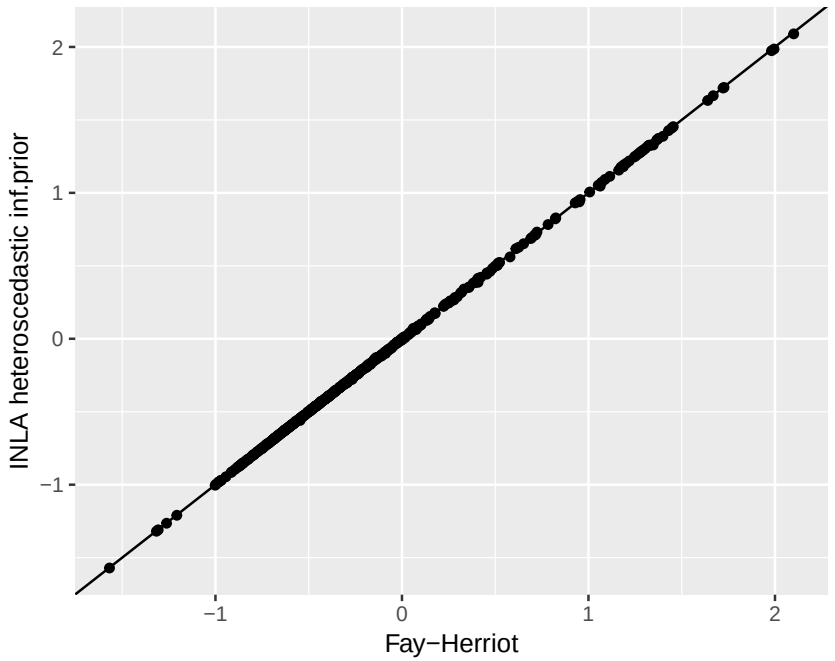


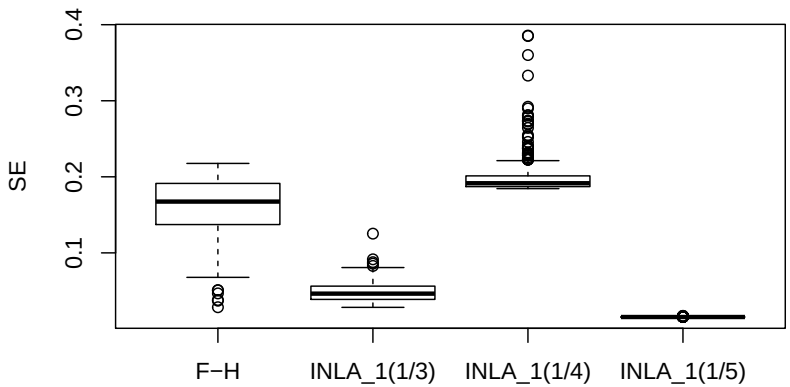
# Case study - Impact on SAE point estimation

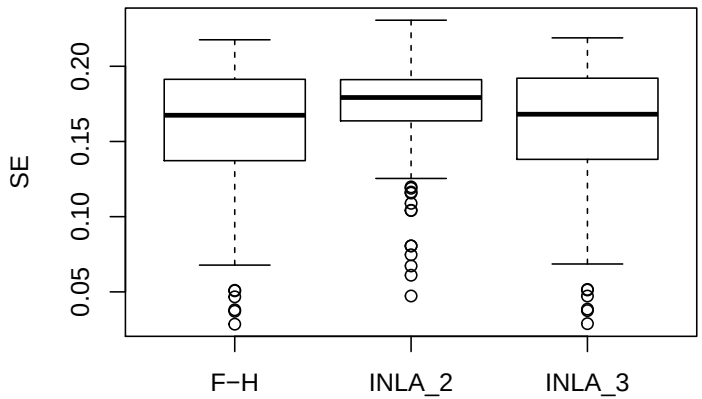



- Positive correlation with FH estimates for INLA 1
- Observe correlation of FH estimates with INLA 3 (highly informative prior)

# Case study - Impact on MSE estimation




- Clear impact on MSE estimates
- Observe distribution of FH analytic MSE estimates with INLA 3 (highly informative prior) MSE estimates (posterior distribution)

# Concluding remarks

- New forms of data offer significant potential for SAE
  - Dynamic updating of estimates
  - Possibly reduced costs
  - Flexible definition of geography
- Risks from black box use of powerful algorithmic tools
- Lack of sensitivity analyses $\rightarrow$ misleading results
- Tempting to produce estimates at very low geographies

Next steps

- Work with CDR data
- Challenges with the definition of geography
- Consider other models as alternatives to FH (see Poletini, 2017)

Thank you for your attention.

n.tzavidis@soton.ac.uk
a.luna.hernandez@soton.ac.uk