Small Area Estimation: A Novel Approach on Estimation of Mean Squared Prediction Error of Small-Area Predictors

Mahmoud Torabi

University of Manitoba, Canada

(Joint work with Jiming Jiang, U of California, Davis, USA)
Outline

- Motivation
Outline

- Motivation
- Model Framework
  - Small Area Predictors
  - Measures of Uncertainty
Outline

- Motivation
- Model Framework
  - Small Area Predictors
  - Measures of Uncertainty
- Some Examples
Outline

- Motivation
- Model Framework
  - Small Area Predictors
  - Measures of Uncertainty
- Some Examples
- Simulation Studies
Outline

- Motivation
- Model Framework
  - Small Area Predictors
  - Measures of Uncertainty
- Some Examples
- Simulation Studies
- Applications
Motivation

Model Framework
• Small Area Predictors
• Measures of Uncertainty

Some Examples

Simulation Studies

Applications

Conclusions

Mahmoud.Torabi@umanitoba.ca
Motivation

- Mixed models are widely used for analyzing correlated data which cover cross-sectional, spatial, and so on.
Motivation

- Mixed models are widely used for analyzing correlated data which cover cross-sectional, spatial, and so on.

- An important case of mixed models is the generalized linear mixed model (GLMM) which has extensively been used in the context of small area estimation (SAE; Rao and Molina, 2015) to cover normal and non-normal responses.
Motivation

- Mixed models are widely used for analyzing correlated data which cover cross-sectional, spatial, and so on.

- An important case of mixed models is the generalized linear mixed model (GLMM) which has extensively been used in the context of small area estimation (SAE; Rao and Molina, 2015) to cover normal and non-normal responses.

- Normal response: state (small area) estimates of median income of four-person families

- Non-normal response: proportions of persons without health insurance for different minority groups (small areas)
Motivation

- However, we do not have enough sample to provide direct estimate (e.g., age-sex-race domains); we use the mixed models to borrow strength from other resources to get reliable estimates.
Motivation

- However, we do not have enough sample to provide direct estimate (e.g., age-sex-race domains); we use the mixed models to borrow strength from other resources to get reliable estimates.

- A major topic in SAE is estimation of mean squared prediction errors (MSPEs) for predictors of various characteristics of interest associated with the small areas.
Motivation

- However, we do not have enough sample to provide direct estimate (e.g., age-sex-race domains); we use the mixed models to borrow strength from other resources to get reliable estimates.

- A major topic in SAE is estimation of mean squared prediction errors (MSPEs) for predictors of various characteristics of interest associated with the small areas.

- A “gold standard” for the MSPE estimation is to produce a second-order unbiased MSPE estimator, that is, the order of bias of the MSPE estimator is $o(m^{-1})$, where $m$ is the number of small areas from which data are collected.
Currently, there are two approaches for producing a second-order unbiased MSPE estimator.
Currently, there are two approaches for producing a second-order unbiased MSPE estimator.

- **Prasad-Rao linearization method** (Prasad and Rao, 1990) which uses Taylor series expansion (tedious derivations; complicated expressions; does not apply to non-differentiable operations such as model selection and shrinkage estimation)
Motivation

Resampling Techniques:

- **Jackknife methods:**
  - Jackknife method (JLW: Jiang et al., 2002) does not apply to non-normal random effects nor to a predictor that is obtained post model selection (PMS).
  - Jiang et al. (2018) proposed a Monte-Carlo jackknife method (McJack) which leads to a second-order unbiased MSPE estimator in situations like PMS; however, it is computationally intensive.

- **Double bootstrap methods (Hall and Maiti, 2006):**
  - Although double bootstrap (DB) is capable of producing a second-order unbiased MSPE estimator, it is, perhaps, computationally even more intensive than the McJack.
Objective

- Propose a second-order unbiased MSPE estimation of small area predictors:
  - which is a hybrid of the linearization method and resampling method, by combining the best part of each method; it is also less computationally intensive compared to the McJack and DB.
Generalized Linear Mixed Model (GLMM)

- Exponential family probability density or mass function:

\[ f(y_i|\theta_i, \phi_i) = \exp\left\{ \frac{y_i\theta_i - a(\theta_i)}{\phi_i} + b(y_i, \phi_i) \right\}, \]

\( (i = 1, \ldots, m), \)

- \( y_i \): variable of interest for the \( i \)-th small area,
- \( \theta_i \): canonical parameters,
- \( \phi_i \): known scale parameters,
- \( a(\cdot) \) and \( b(\cdot) \): known functions,
- \( m \): number of small areas
Generalized Linear Mixed Model (GLMM)

- Natural parameters $\theta_i$:

\[ g(\theta_i) = x_i'\beta + v_i, \]

$g(\cdot)$ : link function,
$v_i \sim N(0, A)$,
$\psi = (\beta, A)$ : model parameters
Mean Squared Prediction Error (MSPE)

Let $\hat{\theta}$ (suppress $i$ for notation simplicity) be a predictor of $\theta$, that is, a function of the observed data, $y$:

- $\hat{\theta}$ can be: EBLUP, or EBP; PMS EBLUP or PMS EBP, to which the standard methods such as Prasad-Rao and JLW do not apply to obtain a second-order unbiased MSPE estimator.
Mean Squared Prediction Error (MSPE)

- Let $\hat{\theta}$ (suppress $i$ for notation simplicity) be a predictor of $\theta$, that is, a function of the observed data, $y$:
  - $\hat{\theta}$ can be: EBLUP, or EBP; PMS EBLUP or PMS EBP, to which the standard methods such as Prasad-Rao and JLW do not apply to obtain a second-order unbiased MSPE estimator

- The MSPE of $\hat{\theta}$ can be expressed as:
  \[
  \text{MSPE} = E(\hat{\theta} - \theta)^2 = E \left[ E \left\{ (\hat{\theta} - \theta)^2 | y \right\} \right] \quad (1)
  \]
Mean Squared Prediction Error (MSPE)

From (1), we can write:

\[ a(y, \psi) = \mathbb{E}\{(\hat{\theta} - \theta)^2|y\} \]
\[ = \hat{\theta}^2 - 2\hat{\theta}\mathbb{E}(\theta|y) + \mathbb{E}(\theta^2|y) \]
\[ = \hat{\theta}^2 - 2\hat{\theta}h_1(y, \psi) + h_2(y, \psi), \]

where \( h_j(y, \psi) = \mathbb{E}(\theta^j|y), j = 1, 2. \)
Mean Squared Prediction Error (MSPE)

- From (1), we can write:

\[
a(y, \psi) = \mathbb{E}\{(\hat{\theta} - \theta)^2 | y\} \\
= \hat{\theta}^2 - 2\hat{\theta}\mathbb{E}(\theta | y) + \mathbb{E}(\theta^2 | y) \\
= \hat{\theta}^2 - 2\hat{\theta}h_1(y, \psi) + h_2(y, \psi),
\]

where \( h_j(y, \psi) = \mathbb{E}(\theta^j | y), j = 1, 2. \)

- Replace \( \psi \) by its consistent estimator \( \hat{\psi} \):

\[
\mathbb{E}\{a(y, \hat{\psi}) - a(y, \psi)\} = O(m^{-1})
\]
Mean Squared Prediction Error (MSPE)

- In other words:

\[ d(\psi) = b(\psi) - c(\psi) = O(m^{-1}), \]

where

\[ b(\psi) = \text{MSPE} = E\{a(y, \psi)\}, \]
\[ c(\psi) = E\{a(y, \hat{\psi})\} \]
Mean Squared Prediction Error (MSPE)

- In other words:

\[ d(\psi) = b(\psi) - c(\psi) = O(m^{-1}), \]

where

\[ b(\psi) = \text{MSPE} = \mathbb{E}\{a(y, \psi)\}, \]
\[ c(\psi) = \mathbb{E}\{a(y, \hat{\psi})\} \]

- One can then show that:

\[ d(\hat{\psi}) - d(\psi) = o_p(m^{-1}) \]
Under some regularity conditions:

$$\hat{\text{MSPE}} = a(y, \hat{\psi}) + b(\hat{\psi}) - c(\hat{\psi}),$$

(2)

in the sense that $E(\hat{\text{MSPE}}) = \text{MSPE} + o(m^{-1})$. 
Under some regularity conditions:

\[ \hat{\text{MSPE}} = a(y, \hat{\psi}) + b(\hat{\psi}) - c(\hat{\psi}), \]  

(2) 

in the sense that \( E(\hat{\text{MSPE}}) = \text{MSPE} + o(m^{-1}). \)

We can calculate \( b(\hat{\psi}) \) and \( c(\hat{\psi}) \) by Monte Carlo methods.
Let $y_{[k]}$ denote $y$ generated under the $k$th Monte-Carlo sample, $k = 1, \ldots, K$. Then, we have:

$$b(\psi) - c(\psi) \approx \frac{1}{K} \sum_{k=1}^{K} \left\{ a(y_{[k]}, \psi) - a(y_{[k]}, \hat{\psi}_{[k]}) \right\},$$

where $\hat{\psi}_{[k]}$ denotes $\hat{\psi}$ based on $y_{[k]}$. 

A Monte-Carlo assisted second-order unbiased MSPE estimator (called Sumca: simple, unified, Monte-Carlo assisted) is given by

$$\hat{\text{MSPE}}_{K} = a(y, \hat{\psi}) + \frac{1}{K} \sum_{k=1}^{K} \left\{ a(y_{[k]}, \hat{\psi}) - a(y_{[k]}, \hat{\psi}_{[k]}) \right\}.$$
Let $y[k]$ denote $y$ generated under the $k$th Monte-Carlo sample, $k = 1, \ldots, K$. Then, we have:

$$b(\psi) - c(\psi) \approx \frac{1}{K} \sum_{k=1}^{K} \left\{ a(y[k], \psi) - a(y[k], \hat{\psi}[k]) \right\},$$

where $\hat{\psi}[k]$ denotes $\hat{\psi}$ based on $y[k]$.

A Monte-Carlo assisted second-order unbiased MSPE estimator (called Sumca: simple, unified, Monte-Carlo assisted) is given by

$$\hat{\text{MSPE}}_K = a(y, \hat{\psi}) + d_K(\hat{\psi})$$

$$= a(y, \hat{\psi}) + \frac{1}{K} \sum_{k=1}^{K} \left\{ a(y[k], \hat{\psi}) - a(y[k], \hat{\psi}[k]) \right\} \quad (3)$$
Remark: A special form of the leading term, \( a(y, \hat{\psi}) \), deserves attention. We can write

\[
a(y, \psi) = \{\hat{\theta} - E(\theta|y)\}^2 + \text{var}(\theta|y)
\]

- the first term on the right side of the above equation vanishes when \( \psi \) is replaced by \( \hat{\psi} \).
- under the general linear mixed model, we then have \( a(y, \hat{\psi}) = \text{var}(\theta|y) = V(\gamma) \), \( \gamma \) is a vector of dispersion parameters, which shows stability of the leading term of the proposed MSPE estimator.
Additional remarks:

- The leading term, $a(y, \hat{\psi})$, in the proposed MSPE estimator is guaranteed positive.

- The Sumca estimator is computationally much less intensive than McJack; McJack requires $m^2/K$ goes to 0, while the Sumca does not have such a restriction (it is recommended that $K = m$ in standard situations). For example, if $m = 100$, the computational cost for Sumca is about 0.001% to 0.01% of that of McJack.
Fay-Herriot Model

- Fay-Herriot (FH) model (1979):

  \[ y_i = x_i' \beta + v_i + e_i = \theta_i + e_i, \quad i = 1, \ldots, m, \]

  where \( v_i \sim N(0, A) \), \( e_i \sim N(0, D_i) \)
Fay-Herriot Model

▶ Fay-Herriot (FH) model (1979):

\[ y_i = x'_i \beta + v_i + e_i = \theta_i + e_i, \quad i = 1, \ldots, m, \]

where \( v_i \sim N(0, A), e_i \sim N(0, D_i) \)

▶ MSPE estimation of \( \hat{\theta}_i \) proposed by Prasad-Rao (1990):

\[
\hat{\text{MSPE}}_{i, \text{PR}} = \frac{\hat{A}D_i}{\hat{A} + D_i} + \left( \frac{D_i}{\hat{A} + D_i} \right)^2 x'_i \left( \sum_{j=1}^{m} \frac{x_jx'_j}{\hat{A} + D_j} \right)^{-1} x_i + \frac{4D_i^2}{(\hat{A} + D_i)^3 m^2} \sum_{j=1}^{m} (\hat{A} + D_j)^2
\]
Fay-Herriot Model

- **Sumca estimator of $\hat{\theta}_i$:**

\[
\text{MSPE}_i = \frac{\hat{AD}_i}{\hat{A} + D_i} + \frac{1}{K} \sum_{k=1}^{K} \left\{ a_i(y_{[k]}, \hat{\psi}) - a_i(y_{[k]}, \hat{\psi}_{[k]}) \right\},
\]

where

\[
a_i(y, \psi) = \frac{AD_i}{A + D_i} + \left( \hat{\theta}_i - \frac{A}{A + D_i} y_i - \frac{D_i}{A + D_i} x_i^T \beta \right)^2
\]
Area-level Model with Model Selection

To test $H_0 : A = 0$ in the FH model (DHM: Datta, Hall, Mandal, 2011):

$$\hat{\theta}_i = \begin{cases} 
\frac{\hat{A}}{\hat{A} + D_i} y_i + \frac{D_i}{\hat{A} + D_i} x_i' \hat{\beta}, & \text{if } T > \chi^2_{m-p}(1 - \alpha) \\
X_i' \beta, & \text{if } T \leq \chi^2_{m-p}(1 - \alpha)
\end{cases} \quad (4)$$

where $T = \sum_{i=1}^m D_i^{-1} (y_i - x_i' \hat{\beta})^2$, $\tilde{\beta} = (X' D^{-1} X)^{-1} X' D^{-1} y$
Area-level Model with Model Selection

To test $H_0: A = 0$ in the FH model (DHM: Datta, Hall, Mandal, 2011):

$$\hat{\theta}_i = \begin{cases} \frac{\hat{A}}{\hat{A} + D_i} y_i + \frac{D_i}{\hat{A} + D_i} x'_i \hat{\beta}, & \text{if } T > \chi^2_{m-p}(1 - \alpha) \\ x'_i \tilde{\beta}, & \text{if } T \leq \chi^2_{m-p}(1 - \alpha) \end{cases} \quad (4)$$

where $T = \sum_{i=1}^{m} D_i^{-1} (y_i - x'_i \hat{\beta})^2$, $\tilde{\beta} = (X'D^{-1}X)^{-1}X'D^{-1}y$

MSPE estimation proposed by DHM:

$$\widehat{\text{MSPE}}_{i, \text{DHM}} = \begin{cases} \widehat{\text{MSPE}}_{i, \text{PR}} & \text{if } T > \chi^2_{m-p}(1 - \alpha) \\ x'_i(X'D^{-1}X)^{-1}x_i & \text{if } T \leq \chi^2_{m-p}(1 - \alpha) \end{cases}$$
Area-level Model with Model Selection

- **Sumca** applies to any kind of predictor including DHM (4):

\[
\hat{\text{MSPE}}_i = \frac{\hat{A}D_i}{\hat{A} + D_i} + \left( \hat{\theta}_i - \frac{\hat{A}}{\hat{A} + D_i} y_i - \frac{D_i}{\hat{A} + D_i} x_i' \hat{\beta} \right)^2
\]

\[+ \frac{1}{K} \sum_{k=1}^{K} \left\{ a_i(y_{[k]}, \hat{\psi}) - a_i(y_{[k]}, \hat{\psi}_[k]) \right\}, \]

where

\[
a_i(y_{[k]}, \hat{\psi}) - a_i(y_{[k]}, \hat{\psi}_[k])
\]

\[= D_i \left( \frac{\hat{A}}{\hat{A} + D_i} - \frac{\hat{A}_{[k]}}{\hat{A}_{[k]} + D_i} \right) + \left( \hat{\theta}_i - \frac{\hat{A}}{\hat{A} + D_i} y_{[k],i} - \frac{D_i}{\hat{A} + D_i} x_i' \hat{\beta} \right)^2
\]

\[- \left( \hat{\theta}_i - \frac{\hat{A}_{[k]}}{\hat{A}_{[k]} + D_i} y_{[k],i} - \frac{D_i}{\hat{A}_{[k]} + D_i} x_i' \hat{\beta}_{[k]} \right)^2 \]
Mixed Logistic Model

Let

\[ P(y_{ij} = 1 | v_i) = p_{ij}, \logit(p_{ij}) = x_{ij}' \beta + v_i, \quad (i = 1, \ldots, m; j = 1, \ldots, n_i) \]

where \( v_i \sim N(0, A) \)
Mixed Logistic Model

- Let

\[ P(y_{ij} = 1|v_i) = p_{ij}, \quad \logit(p_{ij}) = x'_{ij}\beta + v_i, \quad (i = 1, \ldots, m; j = 1, \ldots, n_i) \]

where \( v_i \sim N(0, A) \)

- Assume \( x_{ij} = x_i \), then \( \theta_i = g(x'_i\beta + v_i) \) where

\[ g(u) = \frac{e^u}{1 + e^u}; \quad y_i = \sum_{j=1}^{n_i} y_{ij}, \quad \text{so} \]

\[ f_\beta(y_i|v_i) = \exp \left\{ y_i(x'_i\beta + v_i) - n_i \log(1 + e^{x'_i\beta + v_i}) \right\}, \]

\[ h_{i,s}(y, \psi) = \mathbb{E}(\theta_i^s | y) = \frac{\int \{g(x'_i\beta + v_i)\}^s f_\beta(y_i|v_i)f_A(v_i)dv_i}{\int f_\beta(y_i|v_i)f_A(v_i)dv_i}, \quad s = 1, 2 \]
Mixed Logistic Model

- MSPE estimation of $\hat{\theta}_i$ proposed by Jiang, Lahiri, Wu (JLW) in 2002:

$$\overline{\text{MSPE}}_{i,\text{JLW}} = B_i(\hat{\psi}) - \frac{m - 1}{m} \sum_{i' = 1}^{m} \{B_i(\hat{\psi}_{-i'}) - B_i(\hat{\psi})\}$$

$$+ \frac{m - 1}{m} \sum_{i' = 1}^{m} \{\hat{\theta}_{i, -i'} - \hat{\theta}_i\}^2,$$

where

$$B_i(\psi) = \sum_{k=1}^{n_i} \binom{n_i}{k} \left[ h_{i, 2}(k, \psi) - \{ h_{i, 1}(k, \psi) \}^2 \right] \text{E}\{\theta_i^k (1 - \theta_i)^{n_i - k}\},$$

which is evaluated via numerical integration.
Mixed Logistic Model

- **Sumca estimator of $\theta_i$:**

\[
\widehat{\text{MSPE}}_i = a_i(y, \hat{\psi}) + \frac{1}{K} \sum_{k=1}^{K} \left\{ a_i(y_k, \hat{\psi}) - a_i(y_k, \hat{\psi}_k) \right\},
\]

where

\[
a_i(y, \hat{\psi}) = h_{i,2}(y, \hat{\psi}) - \{ h_{i,1}(y, \hat{\psi}) \}^2
\]

\[
= h_{i,2}(y, \hat{\psi}) - \hat{\theta}_i^2
\]
Fay-Herriot Model

Fay-Herriot (FH) model:

\[ y_i = \beta_0 + \beta_1 x_i + v_i + e_i = \theta_i + e_i, \quad i = 1, \ldots, m, \]

where \( v_i \sim N(0, A), \; e_i \sim N(0, D_i) \)
Fay-Herriot Model

- Fay-Herriot (FH) model:

\[ y_i = \beta_0 + \beta_1 x_i + v_i + e_i \]
\[ = \theta_i + e_i, \quad i = 1, \ldots, m, \]

where \( v_i \sim N(0, A), e_i \sim N(0, D_i) \)

- Simulation set-up:

- \( m = 2m_1; D_i = D_{i1} (1 \leq i \leq m_1); D_i = D_{i2} (m_1 + 1 \leq i \leq m); \)
- \( \beta_0 = \beta_1 = 1; A = 10; D_{i1} \sim U(3.5, 4.5); D_{i2} \sim U(0.5, 1.5); \)
- \( x_i \sim U(0, 1) \) (\( x_i, D_{i1}, \) and \( D_{i2} \) are fixed during simulation);
- \( R = 1000: \) number of simulations
Fay-Herriot Model

- Empirical MSPE (EMSPE):

\[
\text{EMSPE}_i = \frac{1}{R} \sum_{r=1}^{R} \left( \hat{\theta}_i^{(r)} - \theta_i^{(r)} \right)^2, \quad (i = 1, \ldots, m; \ r = 1, \ldots, R)
\]
Fay-Herriot Model

- Empirical MSPE (EMSPE):

\[
\text{EMSPE}_i = \frac{1}{R} \sum_{r=1}^{R} \left\{ \hat{\theta}_i^{(r)} - \theta_i^{(r)} \right\}^2, \quad (i = 1, \ldots, m; \quad r = 1, \ldots, R)
\]

- % RB of a MSPE estimator (MSPE):

\[
\% \text{ RB} = 100 \times \left\{ \frac{\text{E(EMSPE)} - \text{EMSPE}}{\text{EMSPE}} \right\}
\]
Fay-Herriot Model

- **Figure 1**: Boxplots of % RB of MSPE estimates using Prasad-Rao, bootstrap, and Sumca methods: (a) $m = 20$, (b) $m = 50$, and (c) $m = 200$. 

(a) (b) (c)
Fay-Herriot Model

▶ **Figure 2**: Boxplots of MSPE estimates using Prasad-Rao, bootstrap, and Sumca methods: (a) $m = 20$, (b) $m = 50$, and (c) $m = 200$. 

(a)  

(b)  

(c)
Area-level Model with Model Selection

Simulation set-up:

\[ y_i = \beta_01 + \beta_1 x_i + v_i + e_i, \quad i = 1, \ldots, m_1, \]
\[ y_i = \beta_02 + \beta_1 x_i + v_i + e_i, \quad i = m_1 + 1, \ldots, m, \]

\[ m = 2m_1; \quad m = 20; \quad D_i = D_{i1}(1 \leq i \leq m_1); \]
\[ D_i = D_{i2} \quad (m_1 + 1 \leq i \leq m); \quad \beta_01 = 1; \quad \beta_02 = 4; \quad \beta_1 = 1; \]
\[ A(= 0, 0.5, 1); \quad D_{i1} \sim U(0.5, 1.5); \quad D_{i2} \sim U(15.5, 16.5); \]
\[ x_i \sim U(0, 1) \quad (x_i, D_{i1}, \text{and} \ D_{i2} \text{are fixed during simulation}); \]
\[ R = 1000: \text{number of simulations} \]
Area-level Model with Model Selection

Figure 3: Boxplots of % RB for Sumca and DHM methods under different values of $A$ at $\alpha = 0.20$. In each plot, Left: DHM; Right: Sumca

$A = 0$

$A = 0.5$

$A = 1$
Health Insurance of Minority Sub-populations Data

- To consider small domain estimation of the proportion of persons without health insurance for different minority groups in the Asian population in the USA
Health Insurance of Minority Sub-populations Data

- To consider small domain estimation of the proportion of persons without health insurance for different minority groups in the Asian population in the USA.

- Data provided by National Health Interview Survey (NHIS) for the year 2000, which report the individual level binary responses on whether a person has health insurance, along with his or her individual level covariates.
To consider small domain estimation of the proportion of persons without health insurance for different minority groups in the Asian population in the USA.

Data provided by National Health Interview Survey (NHIS) for the year 2000, which report the individual level binary responses on whether a person has health insurance, along with his or her individual level covariates.

The total number of domains \( m \) is 96\((= 3 \times 2 \times 4 \times 4)\) based on age × sex × race × region.
We consider the following model:

\[
\text{logit}(p_{ij}) = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3} + \nu_i, (i = 1, ..., 96; j = 1, ..., n_i),
\]

where

\[
\nu_i \overset{\text{i.i.d.}}{\sim} N(0, A); \quad p_{ij} = P(y_{ij} = 1|\nu_i);
\]

\(y_{ij} = 1\) or 0: whether or not the \(j\)th individual in the \(i\)th small domain does not have health insurance;

\(x_{ij1}, x_{ij2}, x_{ij3}\): family size, educational level, and total family income of the \(j\)th unit in the \(i\)th small domain, respectively.
Health Insurance of Minority Sub-populations Data

Figure 4: Boxplots of square roots of MSPE estimates for health insurance data using JLW, bootstrap, and Sumca methods
COVID-19 Pandemic in Manitoba, Canada

- The goal is to have a better understanding of the COVID-19 pandemic in Manitoba and in particular for some areas which are more vulnerable compared to the rest of Manitoba population.
COVID-19 Pandemic in Manitoba, Canada

- The goal is to have a better understanding of the COVID-19 pandemic in Manitoba and in particular for some areas which are more vulnerable compared to the rest of Manitoba population.

- We provide some summary statistics about the COVID-19 pandemic in Manitoba including daily number of infected cases, number of deaths, number of recovered, and also number of infected cases based on age-sex, and health regions in Manitoba from March 12 to July 24, 2020.
COVID-19 Pandemic in Manitoba, Canada

The goal is to have a better understanding of the COVID-19 pandemic in Manitoba and in particular for some areas which are more vulnerable compared to the rest of Manitoba population.

We provide some summary statistics about the COVID-19 pandemic in Manitoba including daily number of infected cases, number of deaths, number of recovered, and also number of infected cases based on age-sex, and health regions in Manitoba from March 12 to July 24, 2020.

The total number of domains \((m)\) for this analysis is \(80(= 8 \times 2 \times 5)\) based on age \(\times\) sex \(\times\) health region.
COVID-19 Pandemic in Manitoba, Canada

Figure 5: Plots of COVID-19 daily cases, total number of infected cases, total number of deaths, and total number of recovered cases in Manitoba from March 12 to July 24, 2020.
COVID-19 Pandemic in Manitoba, Canada

Figure 6: Histogram of COVID-19 cases stratified by age and sex in Manitoba from March 12 to July 24, 2020
COVID-19 Pandemic in Manitoba, Canada

Figure 7: Maps of COVID-19 cases by sex and health regions in Manitoba from March 12 to July 24, 2020
We consider the following Poisson model:

\[
\log(\theta_i) = \log(E_i) + \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + v_i, (i = 1, \ldots, 80),
\]

where

\[v_i \sim \text{i.i.d. } N(0, A); \theta_i = E(y_i|v_i);\]

\(y_i\): number of infected people in the \(i\)th small domain;

\(E_i\): expected number of infected people in the \(i\)th small domain adjusted by the population size;

\(x_{i1}, x_{i2}\): proportion of immigrants and Indigenous people in the \(i\)th small domain, respectively.
COVID-19 Pandemic in Manitoba, Canada

Figure 8: Prediction of average rate of infected people for 80 domains (age-sex-health region) in Manitoba from March 12 to July 24, 2020
COVID-19 Pandemic in Manitoba, Canada

Figure 9: Boxplots of square roots of MSPE estimates of prediction of average rate of infected people using JLW, bootstrap, and Sumca methods
There is a significant computational advantage of Sumca estimator compared to existing resampling methods in SAE:

- Double bootstrap approach is computationally very intensive.
- McJack method requires the Monte-Carlo sample size, $K$, to satisfy $m^2/K \to 0$. 
Conclusions

- Although, for the second-order unbiasedness of the Sumca estimator, there is no restriction on $K$ (in theory), a larger $K$ would help to reduce the variation of the MSPE estimator.
Conclusions

- Although, for the second-order unbiasedness of the Sumca estimator, there is no restriction on $K$ (in theory), a larger $K$ would help to reduce the variation of the MSPE estimator.

- A practical recommendation is to choose a conveniently large $K$ as long as computational burden is not a concern. In all of our simulation studies and real data analyses, we have chosen $K = m$ to ensure that the results are accurate.
Questions?