On the links between spatial micro-simulation and statistical small area estimation methods

Angela Luna
Joint work with Li-Chun Zhang (UoS) and Paul Williamson and Xin Gu (Liverpool)

Social Statistics and Demography
University of Southampton

July 2018

This project received support by grant ES/N011619/1 - Innovations in Small Area Estimation Methodologies from ESRC via NCRM
<table>
<thead>
<tr>
<th>SAE</th>
<th>Spatial Microsimulation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Aim:</strong> The production of parameter estimates for ‘small’ domains</td>
<td>The creation, analysis and modelling of individual level data allocated to geographic zones¹</td>
</tr>
<tr>
<td><strong>Output:</strong> Set of estimates and their MSEs - Maps</td>
<td>Synthetic individual level data for modelling purposes - Aggregates</td>
</tr>
<tr>
<td><strong>Data:</strong> Survey, census &amp; admin.</td>
<td>Survey &amp; spatial, pop. constraints</td>
</tr>
<tr>
<td><strong>Methods:</strong> Estimators motivated by a statistical model</td>
<td>IPF, Reweighting, Combinatorial Optimisation</td>
</tr>
<tr>
<td><strong>Evaluation:</strong> MSE, external</td>
<td>Diagnostics, MSE and TAE for constraints</td>
</tr>
</tbody>
</table>

Reweighting of a sample from an out-of-area or larger-than-area geography to satisfy a set of local benchmarks X.

- Use of calibration tools (survey sampling) to produce sets of area-specific weights. Area by area calibration. GREGWT algorithm (SAS-R)
- Key difference: Most (or all) survey units do not belong to the area of interest. Worst possible scenario: full suppression of spatial detail on survey data
- Good properties of direct calibration estimators are not directly extensible to this scenario
- Statistical properties of ISC estimates? Potential improvements to this methodology?
1 Statistical properties of ISC
   • Theoretical results & Model-based simulation
2 Calibrated-EBLUP weights
   • Exploration
Set up

- Set of small areas $U_k$ for $k = 1, \ldots, m$; $|U_k| = N_k$
- $y_i$ is an outcome variable for element $i$
- $x_i$ is a vector of covariates for element $i$
- Area-specific benchmark totals $X_k$ known
- Sample $s$ selected from larger-than-area population $U$
- Aim: Provide an estimate for

$$\theta_k = \sum_{i \in U_k} l_i y_i$$

- $l_i = 1 \rightarrow \theta_k = Y_k$. $l_i = 1/N_k \rightarrow \theta_k = \bar{Y}_k$. 

A Luna (UoS)  
Spatial microsimulation and SAE  
July 2018 5 / 24
Find the set of weights $w_i$ that minimise

$$\sum_{i \in s} \frac{(w_i - a_i)^2}{c_i a_i}$$

subject to the constraint

$$\sum_s w_i x_i = \tilde{X}_k = X_k$$

where $c_i$ are fixed constants and $a_i$ are initial weights (arbitrary).

**Notice:**
- Chi-squared distance calibration, e.g. GREGWT
- Non-integer weights (possible $< 1$) are allowed
- No range restrictions (RR) are considered
Theoretical results

Result 1: Unbiased prediction under M1

The ISC estimator is unbiased under the model

\[ y_i = x_i^T \beta + \epsilon_i \quad (M1) \]

\( i = 1, \ldots, N; \ E[\epsilon_i] = 0; \ \text{Cov}(\epsilon_i, \epsilon_j) = \sigma_{ij}, \) given that the calibration constraints ensure unbiased prediction.

Notice that this does not imply unbiasedness for any fixed population.
The theoretical results

Result 2: ISC estimator

The ISC estimator for $\theta_k$ can be written as:

$$\tilde{\theta}_k = X_k^T b + (\hat{Y} - \hat{X}^T b)$$

(1)

where $\hat{Y} = \sum_s a_i y_i$; $\hat{X} = \sum_s a_i x_i$; $b = \hat{A}^{-1}(\sum_s a_i c_i x_i y_i)$ and $\hat{A} = \sum_s a_i c_i x_i x_i^T$.

- Calibration of all areas can be performed in one step.
- $\tilde{\theta}_k$ reduces to the synthetic estimator $X_k^T b$ if there is a constant vector $q$ such that $c_i q^T x_i \equiv 1$ for all $i$, e.g.,
  - model without intercept and $c_i \propto \frac{1}{x_i}$ (all $x_i$ continuous, $\epsilon_i$ heteroscedastic)
  - model with intercept and $c_i = 1$. (all $x_i$ categorical)
Theoretical results

Result 3: Design based Variance

Assuming $a_i = d_i$ (design weights), as $m \to \infty$ and $n_k = O(1)$,

$$V(\hat{\theta}_k) \approx V(\sum_{i \in s} a_i g_{0i} e_i)$$

for $g_{0i} = E(g_i)$; $g_i$ such that $w_i = a_i g_i$ and $e_i = y_i - x_i \mathbf{B}$. This motivates the estimator:

$$\hat{V}_D = \hat{V}(\sum_{i \in s} a_i g_i \hat{e}_i),$$

for $\hat{e}_i = y_i - x_i^T \mathbf{b}$. Furthermore, as $V(\mathbf{b}|s)$ is an approximate design-based variance of $\mathbf{b}$, another possible estimator is given by:

$$\hat{V}_{M1} = \mathbf{X}_k^T \hat{V}(\mathbf{b}|s) \mathbf{X}_k$$
Theoretical results

Result 4: Model-based prediction MSE

Assuming \( a_i = K \), if \( N_k \rightarrow \infty \) as \( m \rightarrow \infty \), \( n_k = O(1) \) and \( \sqrt{n}/N_k \) is small,

\[
V(\tilde{\theta}_k - \theta_k | s) \approx \mathbf{X}_k^T V(b | s) \mathbf{X}_k + V(\epsilon_k | s)
\]

hence, possible estimators are:

- \( \hat{V}_{M1} = \mathbf{X}_k^T \hat{V}(b | s) \mathbf{X}_k \) if \( N_k \) is sufficiently large
- \( \hat{V}_{M2} = \hat{V}_{M1} + \hat{V}(\epsilon_k | s) \) otherwise

Finally, assuming \( y_{ik} = \mathbf{x}_{ik}^T \beta_k + \epsilon_{ik} \), with \( E(\beta_k) = \beta \) and \( V(\beta_k) = \Gamma_{\beta} \), a possible estimator for the prediction MSE of \( \tilde{\theta}_k \) is:

- \( \hat{V}_{M3} = \hat{V}_{M2} + \mathbf{X}_k^T \hat{\Gamma}_\beta \mathbf{X}_k \)
Model-based simulation

Aims:
- Explore $B(\tilde{\theta}_k)$ and $MSE(\tilde{\theta}_k)$
- Explore the properties of $\hat{V}_D$, $\hat{V}_{M1}$, $\hat{V}_{M2}$ and $\hat{V}_{M3}$

Set-up:
- Synthetic population $(300 \times 1000)$
- Auxiliary variables $X_r \sim Multinomial(1, \pi_r); \ p = 1, 2.$
- Response generated under the scenarios:
  - SC1 $y_{ik} = x_{ik}\beta + \epsilon_{ik}; \ \beta = \{5, 3, 1, 4, 2, 8\}$
  - SC2 $y_{ik} = x_{ik}\beta_k + \epsilon_{ik}; \ \beta_k = \beta \times unif(0.85, 1.15)$
  - iid normal errors such that $CV(y) \approx 0.18.$
- Fixed $s_1$ of size 60. Selection of a SRSWOR sample in each domain with size 100. Total sample size 6.000.
- FP-simulation: 5000 samples generated from a fixed population
- Unconditional-simulation: 5000 populations + 1 sample
RAB and RMSE of $\tilde{\theta}_k$ (%)

<table>
<thead>
<tr>
<th></th>
<th>ARB(%)</th>
<th>RMSE(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SC1</td>
<td>SC2</td>
</tr>
<tr>
<td>In sample</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Out of sample</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mod In sample</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Out of sample</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| FP In sample | 0.327 | 4.915 | 0.396 | 4.940 |
| FP Out of sample | 0.363 | 4.591 | 0.430 | 4.618 |
| FP All | 0.356 | 4.656 | 0.424 | 4.682 |
| Mod In sample | 0.005 | 4.687 | 0.511 | 4.760 |
| Mod Out of sample | 0.005 | 4.514 | 0.518 | 4.595 |
| Mod All | 0.005 | 4.549 | 0.517 | 4.628 |
Model-based simulation

Relative Bias Variance estimators (%)

• \( \hat{V}_D = \hat{V}(\sum_{i \in s} a_i g_i \hat{e}_i) \)
• \( \hat{V}_{M1} = \mathbf{x}_k^T \hat{V}(\mathbf{b}|s) \mathbf{x}_k \) if \( N_k \) is sufficiently large
• \( \hat{V}_{M2} = \hat{V}_{M1} + \hat{V}(\epsilon_k|s) \) otherwise
• \( \hat{V}_{M3} = \hat{V}_{M2} + \mathbf{x}_k^T \hat{\Gamma}_\beta \mathbf{x}_k \)

<table>
<thead>
<tr>
<th>Est.</th>
<th>SC 1</th>
<th>SC 2</th>
<th>AMSE</th>
<th>AMSE</th>
<th>AMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( V(\tilde{\theta}_k) )</td>
<td>( V(\tilde{\theta}_k) )</td>
<td>( AMSE(\tilde{\theta}_k) )</td>
<td>( AMSE(\tilde{\theta}_k) )</td>
<td>( AMSE(\tilde{\theta}_k) )</td>
</tr>
<tr>
<td>( \hat{V}_D )</td>
<td>5.868</td>
<td>-</td>
<td>282.642</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>( \hat{V}_{M1} )</td>
<td>10.472</td>
<td>-</td>
<td>11.676</td>
<td>-</td>
<td>-85.58</td>
</tr>
<tr>
<td>( \hat{V}_{M2} )</td>
<td>-</td>
<td>13.853</td>
<td>-</td>
<td>-96.267</td>
<td>0.434</td>
</tr>
<tr>
<td>( \hat{V}_{M3} )</td>
<td>-</td>
<td>72.974</td>
<td>-</td>
<td>10.677</td>
<td>64.206</td>
</tr>
</tbody>
</table>
Summary

- $\tilde{\theta}_k$ is unbiased under model M1. Not unbiased for any finite population.
- Given the expression (1), $\tilde{\theta}_k$ can be calculated in one step.
- In some cases, $\tilde{\theta}_k$ reduces to the synthetic estimator $X_k^T b$. A particular case is when all $x_i$ are categorical and $c_i = 1$.
- FP uncertainty estimation. All proposed variance estimators are biased. For the variance of $\tilde{\theta}_k$, $\hat{V}_D$ seems to perform better if the model holds and $\hat{V}_{M1}$ if it doesn’t. $\hat{V}_{M2}$ and $\hat{V}_{M3}$ seems closer to the average MSE, but this needs to be studied in more detail.
- Unconditional uncertainty estimation. Estimation of area-specific MSE $\tilde{\theta}_k$ does not seem possible with any of the proposed estimators. Under the model, $\hat{V}_{M2}$ shows good performance on estimating the average MSE of $\tilde{\theta}_k$. Although biased the additional term in $\hat{V}_{M3}$ seems to capture some of the additional uncertainty due to model misspecification.
1. Statistical properties of ISC
   - Theoretical results & Model-based simulation
2. Calibrated-EBLUP weights
   - Exploration
Calibrated-EBLUP weights

Consider the nested regression model

\[ y_{ik} = x_{ik}^T \beta + u_i + \epsilon_{ik}, \]

with \( u_i \overset{iid}{\sim} (0, \sigma_u^2) \) and \( \epsilon_{ik} \overset{iid}{\sim} (0, \sigma^2_{\epsilon}). \) An EBLUP of \( \bar{Y}_i \) is given by:

\[ \bar{Y}_i^E = \bar{X}_i^T \hat{\beta} + \hat{\gamma}_i \left( \bar{y}_i - \bar{x}_i \hat{\beta} \right). \]

(2)

As \( \hat{\beta} = \left( X^T \hat{V}^{-1} X \right)^{-1} X^T \hat{V}^{-1} Y = HY \), (2) can be rewritten as:

\[ \bar{Y}_i^E = \left[ \bar{X}_i^T H + \hat{\gamma}_i (\delta_i - \bar{x}_i H) \right] Y = W_i^E Y = \sum_{j=1}^n w_{ij} y_j, \]

(3)

with \( \hat{\gamma}_i = \hat{\sigma}_u^2 / (\hat{\sigma}_u^2 + \hat{\sigma}_{\epsilon}^2 / n_i); \) \( \delta_{ik} = 1 / n_i \) if \( k \in s_i \) and zero otherwise and \( \hat{V} = \text{bdiag} \left( \text{diag} \left( \hat{\sigma}_{\epsilon}^2 \right) + \hat{\sigma}_u^2 1_{n_i} 1_{n_i}^T \right). \)
Considering all domains simultaneously,

$$\bar{Y}^E = \left[ \tilde{X}^T H + \hat{\gamma} (\delta - \bar{x}H) \right] Y = W^E Y.$$ 

$W^E$ is a matrix of dimension $m \times n$, containing in the rows 'optimal' domain-specific weights for $Y$.

- In which situations could the weights in $W^E$ be used to obtain adequate estimates for another variable $Z$?
- Can the weights in $W^E$ be used as a starting point for ISC?
  - In the context presented before, ISC corresponds to the synthetic estimator $\bar{X}_i^T \hat{\beta}$. EBLUP weights can motivate an initial trade-off between bias and variance.
  - The risk of losing optimality for $Y$ can be eliminated by adding $\bar{Y}^E$ to the set of calibration constraints.
• Synthetic population \((100 \times 300)\) generated using a real sample of 10k observations. \(X_1(5), X_2(5), X_3(7)\) and \(Y(6)\).

• Response variables:
  • \(Y_1\) and \(Y_2\) obtained directly from the data.
  • \(Y_3\) has been contaminated to reduce the correlation with \(Y_1\)
  • \(Y_4 = [X_2, X_3] \beta + \zeta; \zeta_{ik} \overset{iid}{\sim} N(0, \sigma^2_{\zeta})\)
  • \(Y_5 = [X_2, X_3] \beta_i + \xi; \xi_{ik} \overset{iid}{\sim} N(0, \sigma^2_{\xi}); \beta_i = \beta + \nu_i; \nu_i \overset{iid}{\sim} MN(0, 0.05 \times \text{diag}(\beta))\)

• Fixed \(s_1\) of size 50. Selection of 1000 independent samples with fixed domain size 25. Total sample size 1.250.
• Estimators:
  1. $\bar{Y}_i^{E_1}$: uses the EBLUP weights calculated for $\bar{Y}_1|\bar{X}_1$
  2. $\bar{Y}_i^{E_1 C_{2,3}}$: uses the weights obtained after applying ISC with starting point the EBLUP weights above, for each domain. Constraints: $X_2, X_3, \bar{Y}_i^{E_1}$.
  3. $\bar{Y}_i^{E_{1,2,3}}$: is an EBLUP for $\bar{Y}_i|\bar{X}_1, \bar{X}_2, \bar{X}_3$
  4. $\bar{Y}_i^{C_{1,2,3}}$: is the ISC obtained using initial weights = 1 and constraints $X_1, X_2, X_3$

• Potential negative weights from $\bar{Y}_i^{E_1}$. In those cases, $W_i^{E*} = W_i^{E} + c$. Around 10% observed, always for $k \notin s_i$
### Results in-sample areas

<table>
<thead>
<tr>
<th>$Y_i$</th>
<th>$E_{1,2,3}$</th>
<th>$E_1$</th>
<th>$E_1 C_{2,3}$</th>
<th>$C_{1,2,3}$</th>
<th>$E_{1,2,3}$</th>
<th>$E_1$</th>
<th>$E_1 C_{2,3}$</th>
<th>$C_{1,2,3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y_1$</td>
<td>7.29</td>
<td>7.46</td>
<td>7.46</td>
<td>21.15</td>
<td>15.75</td>
<td>16.00</td>
<td>16.00</td>
<td>21.60</td>
</tr>
<tr>
<td>$Y_2$</td>
<td>20.31</td>
<td>13.92</td>
<td>18.40</td>
<td>37.51</td>
<td>34.49</td>
<td>38.37</td>
<td>36.16</td>
<td>38.83</td>
</tr>
<tr>
<td>$Y_3$</td>
<td>15.39</td>
<td>8.54</td>
<td>11.49</td>
<td>24.11</td>
<td>22.97</td>
<td>26.28</td>
<td>24.68</td>
<td>25.20</td>
</tr>
<tr>
<td>$Y_4$</td>
<td>0.69</td>
<td>0.29</td>
<td>0.40</td>
<td>0.75</td>
<td>1.03</td>
<td>2.46</td>
<td>1.94</td>
<td>0.97</td>
</tr>
<tr>
<td>$Y_5$</td>
<td>1.27</td>
<td>1.86</td>
<td>2.50</td>
<td>5.22</td>
<td>2.97</td>
<td>3.23</td>
<td>3.31</td>
<td>5.27</td>
</tr>
</tbody>
</table>

- MSE of $E_1$ comparable to that of $C_{1,2,3}$ for other variables, even if the correlation is low.
  \[
  \text{Corr}(Y_1, Y_i) = (-0.363, -0.056, 0.046, 0.029), \ i = 2, \ldots, 5.
  \]
- However, $E_1$ seems substantially more robust to bias.
### Results in-sample areas

<table>
<thead>
<tr>
<th>$Y_i$</th>
<th>$E_{1,2,3}$</th>
<th>$E_1$</th>
<th>$E_1 C_{2,3}$</th>
<th>$C_{1,2,3}$</th>
<th>$E_{1,2,3}$</th>
<th>$E_1$</th>
<th>$E_1 C_{2,3}$</th>
<th>$C_{1,2,3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y_1$</td>
<td>7.29</td>
<td>7.46</td>
<td>7.46</td>
<td>21.15</td>
<td>15.75</td>
<td>16.00</td>
<td>16.00</td>
<td>21.60</td>
</tr>
<tr>
<td>$Y_2$</td>
<td>20.31</td>
<td>13.92</td>
<td>18.40</td>
<td>37.51</td>
<td>34.49</td>
<td>38.37</td>
<td>36.16</td>
<td>38.83</td>
</tr>
<tr>
<td>$Y_3$</td>
<td>15.39</td>
<td>8.54</td>
<td>11.49</td>
<td>24.11</td>
<td>22.97</td>
<td>26.28</td>
<td>24.68</td>
<td>25.20</td>
</tr>
<tr>
<td>$Y_4$</td>
<td>0.69</td>
<td>0.29</td>
<td>0.40</td>
<td>0.75</td>
<td>1.03</td>
<td>2.46</td>
<td>1.94</td>
<td>0.97</td>
</tr>
<tr>
<td>$Y_5$</td>
<td>1.27</td>
<td>1.86</td>
<td>2.50</td>
<td>5.22</td>
<td>2.97</td>
<td>3.23</td>
<td>3.31</td>
<td>5.27</td>
</tr>
</tbody>
</table>

- MSE of $E_1$ comparable to that of $C_{1,2,3}$ for other variables, even if the correlation is low.
  
  \[ \text{Corr}(Y_1, Y_i) = (-0.363, -0.056, 0.046, 0.029), \ i = 2, \ldots, 5. \]

- However, $E_1$ seems substantially more robust to bias.
Results in-sample areas

<table>
<thead>
<tr>
<th>$Y_i$</th>
<th>$E_{1,2,3}$</th>
<th>$E_1$</th>
<th>$E_1 C_{2,3}$</th>
<th>$C_{1,2,3}$</th>
<th>$E_{1,2,3}$</th>
<th>$E_1$</th>
<th>$E_1 C_{2,3}$</th>
<th>$C_{1,2,3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y_1$</td>
<td>7.29</td>
<td>7.46</td>
<td>7.46</td>
<td>21.15</td>
<td>15.75</td>
<td>16.00</td>
<td>16.00</td>
<td>21.60</td>
</tr>
<tr>
<td>$Y_2$</td>
<td>20.31</td>
<td>13.92</td>
<td>18.40</td>
<td>37.51</td>
<td>34.49</td>
<td>38.37</td>
<td>36.16</td>
<td>38.83</td>
</tr>
<tr>
<td>$Y_3$</td>
<td>15.39</td>
<td>8.54</td>
<td>11.49</td>
<td>24.11</td>
<td>22.97</td>
<td>26.28</td>
<td>24.68</td>
<td>25.20</td>
</tr>
<tr>
<td>$Y_4$</td>
<td>0.69</td>
<td>0.29</td>
<td>0.40</td>
<td>0.75</td>
<td>1.03</td>
<td>2.46</td>
<td>1.94</td>
<td>0.97</td>
</tr>
<tr>
<td>$Y_5$</td>
<td>1.27</td>
<td>1.86</td>
<td>2.50</td>
<td>5.22</td>
<td>2.97</td>
<td>3.23</td>
<td>3.31</td>
<td>5.27</td>
</tr>
</tbody>
</table>

- $E_1 C_{2,3}$ performs marginally better than $E_1$. Calibrating would reduce the variance compared to $E_1 C_{2,3}$ as long as $X_2, X_3$ are correlated with $Y_i$. Increase on the bias but still gains respect to ISC and comparable with $E_{1,2,3}$.

- Calibrated alternatives seem to perform particularly poorly for $Y_5$ when compared to $Y_4$. Small population sizes?
### Results out-of-sample areas

<table>
<thead>
<tr>
<th>$Y_i$</th>
<th>$E_{1,2,3}$</th>
<th>$E_1$</th>
<th>$E_1 C_{2,3}$</th>
<th>$C_{1,2,3}$</th>
<th>$E_{1,2,3}$</th>
<th>$E_1$</th>
<th>$E_1 C_{2,3}$</th>
<th>$C_{1,2,3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y_1$</td>
<td>18.02</td>
<td>18.81</td>
<td>18.81</td>
<td>17.84</td>
<td>18.57</td>
<td>19.29</td>
<td>19.29</td>
<td>18.39</td>
</tr>
<tr>
<td>$Y_2$</td>
<td>35.44</td>
<td>36.74</td>
<td>35.53</td>
<td>35.33</td>
<td>36.83</td>
<td>38.16</td>
<td>36.85</td>
<td>36.71</td>
</tr>
<tr>
<td>$Y_3$</td>
<td>24.31</td>
<td>24.52</td>
<td>24.32</td>
<td>24.31</td>
<td>25.48</td>
<td>25.67</td>
<td>25.44</td>
<td>25.48</td>
</tr>
<tr>
<td>$Y_4$</td>
<td>0.79</td>
<td>0.82</td>
<td>0.79</td>
<td>0.79</td>
<td>0.99</td>
<td>1.04</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>$Y_5$</td>
<td>5.76</td>
<td>5.86</td>
<td>5.76</td>
<td>5.75</td>
<td>5.81</td>
<td>5.91</td>
<td>5.81</td>
<td>5.80</td>
</tr>
</tbody>
</table>

- For out-of-sample areas, all estimators are synthetic and perform similarly.
Future work

- Theoretical formulation
- Extension to the possibility of using more than one EBLUP to determine initial weights
  - The key to the bias reduction of $E_1 C_{2,3}$ respect to $C_{1,2,3}$ seem to be the possibility of allocating different initial weights to $k \in s_i$ and $k \not\in s_i$. EBLUP suggest a way to decide on the trade-off bias vs variance.
  - Potential combination of initial weights + EBLUPs as constraints?
- Are negative EBLUP weights an issue?
- MSE estimation