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

SAE	Spatial Microsimulation
Aim: The production of parameter estimates for 'small' domains	The creation, analysis and modelling of individual level data allocated to geographic zones[1]
Output: Set of estimates and their MSEs - Maps	Synthetic individual level data for modelling purposes - Aggregates
Data: Survey, census & admin.	Survey & spatial, pop. constraints
Methods: Estimators motivated by a statistical model	IPF, Reweighting, Combinatorial Optimisation
Evaluation: MSE, external	Diagnostics, MSE and TAE for constraints

[1] Lovelace, R., Dumont, M., 2016. Spatial microsimulation with R. CRC Press.

ISC

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?

Outline

1. Statistical properties of ISC
   - Theoretical results & Model-based simulation
2. Calibrated-EBLUP weights
   - Exploration

Set up

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

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

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

Set up

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

$$\sum_{s} w_i \mathbf{x_i} = \tilde{\mathbf{X}}_\mathbf{k} = \mathbf{X_k}$$

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

The ISC estimator is unbiased under the model

$$y_i = \mathbf{x_i}^T \boldsymbol{\beta} + \epsilon_i \tag{M1}$$

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

Notice that this <u>does not imply</u> unbiasedness for any fixed population.

The ISC estimator for θ_k can be written as:

$$\boxed{\tilde{\theta}_k = \mathbf{X_k}^T\mathbf{b} + (\hat{Y} - \hat{\mathbf{X}}^T\mathbf{b})} \tag{1}$$

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

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

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

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

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

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

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

$$\hat{V}_{M1} = \mathbf{X_k}^T \hat{V}(\mathbf{b}|s) \mathbf{X_k}$$

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

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

hence, possible estimators are:

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

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

- $\hat{V}_{M3} = \hat{V}_{M2} + \mathbf{X_k}^T \hat{\boldsymbol{\Gamma}}_{\beta} \mathbf{X_k}$

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 x 1000)
- Auxiliary variables $X_r \sim Multinomial(1, \boldsymbol{\pi_r})$; $p = 1, 2$.
- Response generated under the scenarios:
  - SC1 $y_{ik} = \mathbf{x_{ik}}\boldsymbol{\beta} + \epsilon_{ik}$; $\boldsymbol{\beta} = \{5, 3, 1, 4, 2, 8\}$
  - SC2 $y_{ik} = \mathbf{x_{ik}}\boldsymbol{\beta}_k + \epsilon_{ik}$; $\boldsymbol{\beta}_k = \boldsymbol{\beta} \times unif(0.85, 1.15)$
  - 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 $\tilde{\theta}_k$ (%)

		ARB(%)		RMSE(%)	
		SC1	SC2	SC1	SC2
FP	In sample	0.327	4.915	0.396	4.940
	Out of sample	0.363	4.591	0.430	4.618
	All	0.356	4.656	0.424	4.682
Mod	In sample	0.005	4.687	0.511	4.760
	Out of sample	0.005	4.514	0.518	4.595
	All	0.005	4.549	0.517	4.628

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

	FP				Unconditional	
	SC 1		SC 2		AMSE	
Est.	$V(\tilde{\theta}_k)$	$AMSE(\tilde{\theta}_k)$	$V(\tilde{\theta}_k)$	$AMSE(\tilde{\theta}_k)$	SC1	SC2
$\hat{V}_D$	5.868	-	282.642	-	-	-
$\hat{V}_{M1}$	10.472	-	11.676	-	-85.58	-99.446
$\hat{V}_{M2}$	-	13.853	-	-96.267	0.434	-96.143
$\hat{V}_{M3}$	-	72.974	-	10.677	64.206	8.054

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 $\mathbf{X_k}^T\mathbf{b}$. 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.

Outline

1. Statistical properties of ISC
   - Theoretical results & Model-based simulation
2. Calibrated-EBLUP weights
   - Exploration

Consider the the nested regression model

$$y_{ik} = \mathbf{x}_{ik}^T \boldsymbol{\beta} + u_i + \epsilon_{ik},$$

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

$$\bar{Y}_i^E = \bar{\mathbf{X}}_i^T \hat{\boldsymbol{\beta}} + \hat{\gamma}_i \left(\bar{y}_i - \bar{\mathbf{x}}_i \hat{\boldsymbol{\beta}} \right). \quad (2)$$

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

$$\bar{Y}_i^E = \left[\bar{\mathbf{X}}_i^T \mathbf{H} + \hat{\gamma}_i \left(\boldsymbol{\delta}_i - \bar{\mathbf{x}}_i \mathbf{H} \right) \right] \mathbf{Y} = \mathbf{W}_i^E \mathbf{Y} = \sum_{j=1}^{n} w_{ij} y_j, \quad (3)$$

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 $\hat{\mathbf{V}} = \text{bdiag}(\text{diag}(\hat{\sigma}_\epsilon^2) + \hat{\sigma}_u^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T)$.

Considering all domains simultaneously,

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

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

- In which situations could the weights in $\mathbf{W}^E$ be used to obtain adequate estimates for another variable $\mathbf{Z}$?
- Can the weights in $\mathbf{W}^E$ be used as a starting point for ISC?
  - In the context presented before, ISC corresponds to the synthetic estimator $\bar{\mathbf{X}}_i^T\hat{\beta}$. EBLUP weights can motivate an initial trade-off between bias and variance.
  - The risk of losing optimality for $\mathbf{Y}$ can be eliminated by adding $\bar{\mathbf{Y}}^E$ 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
  - $Y_4 = [\mathbf{X}_2, \mathbf{X}_3]\,\boldsymbol{\beta} + \boldsymbol{\zeta}$; $\zeta_{ik} \overset{iid}{\sim} N(0, \sigma^2_\zeta)$
  - $Y_5 = [\mathbf{X}_2, \mathbf{X}_3]\,\boldsymbol{\beta}_i + \boldsymbol{\xi}$; $\xi_{ik} \overset{iid}{N} \sim (0, \sigma^2_\xi)$; $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\nu}_i$; $\boldsymbol{\nu}_i \overset{iid}{\sim} MN(\mathbf{0}, 0.05 \times \text{diag}(\boldsymbol{\beta}))$
- 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 $\bar{\mathbf{Y}}_1|\bar{\mathbf{X}}_1$
  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: $\mathbf{X}_2, \mathbf{X}_3$, $Y_i^{E_1}$.
  3. $\bar{Y}_i^{E_{1,2,3}}$: is an EBLUP for $\bar{\mathbf{Y}}_i|\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2, \bar{\mathbf{X}}_3$
  4. $\bar{Y}_i^{C_{1,2,3}}$ is the ISC obtained using initial weights $= 1$ and constraints $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$
- Potential negative weights from $\bar{Y}_i^{E_1}$. In those cases, $\mathbf{W}_i^{E*} = \mathbf{W}_i^{E} + c$. Around 10% observed, always for $k \notin s_i$

Y_i	RAB (%) $E_{1,2,3}$	RAB (%) E_1	RAB (%) $E_1 C_{2,3}$	RAB (%) $C_{1,2,3}$	RMSE (%) $E_{1,2,3}$	RMSE (%) E_1	RMSE (%) $E_1 C_{2,3}$	RMSE (%) $C_{1,2,3}$
Y_1	7.29	7.46	7.46	21.15	15.75	16.00	16.00	21.60
Y_2	20.31	13.92	18.40	37.51	34.49	38.37	36.16	38.83
Y_3	15.39	8.54	11.49	24.11	22.97	26.28	24.68	25.20
Y_4	0.69	0.29	0.40	0.75	1.03	2.46	1.94	0.97
Y_5	1.27	1.86	2.50	5.22	2.97	3.23	3.31	5.27

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

Y_i	RAB (%) $E_{1,2,3}$	E_1	$E_1 C_{2,3}$	$C_{1,2,3}$	RMSE (%) $E_{1,2,3}$	E_1	$E_1 C_{2,3}$	$C_{1,2,3}$
Y_1	7.29	7.46	7.46	21.15	15.75	16.00	16.00	21.60
Y_2	20.31	13.92	18.40	37.51	34.49	38.37	36.16	38.83
Y_3	15.39	8.54	11.49	24.11	22.97	26.28	24.68	25.20
Y_4	0.69	0.29	0.40	0.75	1.03	2.46	1.94	0.97
Y_5	1.27	1.86	2.50	5.22	2.97	3.23	3.31	5.27

- 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.

Y_i	RAB (%)				RMSE (%)			
	$E_{1,2,3}$	E_1	$E_1 C_{2,3}$	$C_{1,2,3}$	$E_{1,2,3}$	E_1	$E_1 C_{2,3}$	$C_{1,2,3}$
Y_1	7.29	7.46	7.46	21.15	15.75	16.00	16.00	21.60
Y_2	20.31	13.92	18.40	37.51	34.49	38.37	36.16	38.83
Y_3	15.39	8.54	11.49	24.11	22.97	26.28	24.68	25.20
Y_4	0.69	0.29	0.40	0.75	1.03	2.46	1.94	0.97
Y_5	1.27	1.86	2.50	5.22	2.97	3.23	3.31	5.27

- $E_1 C_{2,3}$ 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

	RAB (%)				RMSE (%)			
Y_i	$E_{1,2,3}$	E_1	$E_1 C_{2,3}$	$C_{1,2,3}$	$E_{1,2,3}$	E_1	$E_1 C_{2,3}$	$C_{1,2,3}$
Y_1	18.02	18.81	18.81	17.84	18.57	19.29	19.29	18.39
Y_2	35.44	36.74	35.53	35.33	36.83	38.16	36.85	36.71
Y_3	24.31	24.52	24.32	24.31	25.48	25.67	25.44	25.48
Y_4	0.79	0.82	0.79	0.79	0.99	1.04	0.99	0.99
Y_5	5.76	5.86	5.76	5.75	5.81	5.91	5.81	5.80

- 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 $k \notin s_i$. 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