Spatial Approaches in Small Area Estimation

Dr George Sofronov

Centre for Statistical and Survey Methodology,
School of Mathematics and Applied Statistics, University of Wollongong
Introduction

Estimation of population characteristics for sub-national domains (or smaller regions) is an important objective for statistical surveys. In particular, geographically defined domains, e.g. regions, states, counties, wards and metropolitan areas can be of interest.

The problems of small area estimation (SAE) are as follows.

1. How to produce reliable estimates of characteristics of interest (means, totals, quantiles, etc).
2. How to assess the estimation error.
**Design-based approach**

Estimates for these domains based on the usual *design-based approach* to survey sampling inference are called *direct estimates*.

However, sample sizes are typically small (or even zero) within the domains/areas of interest, leading to large sampling variability for these direct estimators.
Model-based approach

An alternative approach that is now widely used in small area estimation is the so-called indirect or *model-based approach*. This uses auxiliary information for the small areas of interest and can be characterized as "borrowing strength" from the relationship between the target variables and the auxiliary information.

Model-based methods of SAE are often based on assuming a *linear mixed model*, with area-specific random effects to account for between area variation beyond that explained by auxiliary variables included in the fixed part of the model.
Linear Mixed Models

\[ Y = X\beta + Zu + e \]

- \( Y \) is a vector of observations,
- \( X \) is a matrix of known covariates,
- \( \beta \) is a vector of unknown regression coefficients (fixed effects),
- \( Z \) is a known matrix,
- \( u \) is a vector of random effects,
- \( e \) is a vector of errors.

The \( u \) and \( e \) are mutually independent normally distributed with zero means and respective variances.
Random area effects

It is customary to assume that the random area effects are *independent*, in practice most small area boundaries are arbitrary and there appears to be no good reason why population units just one side of such a boundary should not generally be correlated with population units just on the other side.

In particular, it is often reasonable to assume that the effects of neighbouring areas (defined, for example, by a contiguity criterion) are *correlated*, with the correlation decaying to zero as the distance between these areas increases.
Models with spatially independent effects

\[ Y = X \beta + Z u + e, \]

where

\[ Z = \begin{pmatrix} 1_{N_1} & 0 & \ldots & 0 \\ 0 & 1_{N_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1_{N_m} \end{pmatrix}, \quad 1_{N_i} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \]

\[ u \sim N(0, \sigma_u^2 I_m), \]

\[ e \sim N(0, \sigma_e^2 I_N). \]
Since different areas are independent, the covariance matrix of $Y$ has block diagonal structure given by

$$V = \begin{pmatrix}
V_1 & 0 & \ldots & 0 \\
0 & V_2 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & V_m
\end{pmatrix},$$

where

$$V_i = \sigma_e^2 I_{N_i} + \sigma_u^2 1_{N_i} 1_{N_i}^t, \quad 1 \leq i \leq m,$$

$I_{N_i}$ is the identity matrix of order $N_i$. 

**Covariance matrix**
Models with spatial dependence

- **Simultaneously Autoregressive models (SAR)**
  The vector of random area effects $v$ satisfies
  \[ v = \rho W v + u, \]
  where $\rho$ is a *spatial autoregressive coefficient*, $W$ is a *proximity matrix*.

- **Conditional Autoregressive models (CAR)**
  The probability of observing a particular value at a given site is a conditional probability, i.e. it depends on the value of $Y$ in the neighbourhood of the site:
  \[ v_i \mid \{v_j \in D_i\} \sim \mathcal{N}(\rho \sum_{j \in D_i} w_{ij} v_j, \sigma_u^2). \]
SAR model

The model can be expressed as

\[ Y = X\beta + Zv + \varepsilon, \]

where

\[ v = (I_m - \rho W)^{-1}u, \quad u \sim \mathcal{N}(0, \sigma_u^2 I_m), \quad \mathbb{E}(v) = 0, \]

\[ \text{Var}(v) = \sigma_u^2 ((I_m - \rho W)(I_m - \rho W^t))^{-1} = G. \]

The \( W \) matrix describes how random effects from neighbouring areas are related, whereas \( \rho \) defines the strength of this spatial relationship.
Proximity matrix

Choices for $W$ (with $w_{ii} = 0$)

- A contiguity matrix, that is, the elements of $W$ take non-zero values only for those pairs of areas that are adjacent.

- Some function of the length of shared border between neighbouring areas.

- A function of the distance between certain locations in each area.

$W$ is typically symmetric, but need not be.

$\tilde{W}$: standardize row $i$ by $w_{i+} = \sum_j w_{ij}$ (so matrix is row stochastic, but probably no longer symmetric). Then $\rho \in (-1, 1)$ is called a spatial autocorrelation parameter.
Covariance matrix

The SAR model can be written as

\[ y = X\beta + Z(I - \rho W)^{-1}u + \varepsilon. \]

It follows that the covariance matrix of \( Y \) is

\[ \text{Var}(Y) = V = \sigma^2_e I_N + ZGZ^t, \]

where

\[ G = \sigma^2_u ((I_m - \rho W)(I_m - \rho W^t))^{-1}. \]
The Australian Bureau of Agricultural and Resource Economics (ABARE) has been conducting farm surveys annually since 1978.

The survey data are highly confidential and farm level records are never disclosed or published in a way that would identify individual landholders.

Variables collected:
- farm area, quantity produced, livestock numbers, etc
- spatial data
- financial variables like receipts, assets, debt, etc
- some social items like age, education, etc
Data set

- A sample of $n = 218$ farms in wheat-sheep zone that first came in the survey in 1994
- Auxiliary variables are farm area and industry
  - specialist croppers
  - mixed livestock croppers
  - sheep specialists
  - beef specialists
  - mixed sheep beef farms
- Variable of interest is cash receipts
- $m = 12$ regions (small areas) in wheat-sheep zone
- *Synthetic coordinates* (longitude, latitude) for each farm
Regions
## Small areas

<table>
<thead>
<tr>
<th>Small area, $i$</th>
<th>Region</th>
<th>$n_i$</th>
<th>Small area, $i$</th>
<th>Region</th>
<th>$n_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>121</td>
<td>57</td>
<td>7</td>
<td>321</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>122</td>
<td>14</td>
<td>8</td>
<td>322</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>123</td>
<td>25</td>
<td>9</td>
<td>421</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>221</td>
<td>14</td>
<td>10</td>
<td>422</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>222</td>
<td>7</td>
<td>11</td>
<td>521</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>223</td>
<td>50</td>
<td>12</td>
<td>522</td>
<td>13</td>
</tr>
</tbody>
</table>
Proximity matrices

Let $d_{i,j}^{(1)}$ be longitudinal distance, $d_{i,j}^{(2)}$ latitudinal distance, $d_{i,j}^{(3)}$ distance between areas $i$ and $j$.

1. $w_{ij} = 1$ if $i, j$ share a common boundary
2. $w_{ij} = 1/d_{i,j}^{(1)}$
3. $w_{ij} = 1/d_{i,j}^{(2)}$
4. $w_{ij} = 1/d_{i,j}^{(3)}$
5. $w_{ij} = \exp\{-\xi d_{i,j}^{(1)}\}, \xi > 0$
6. $w_{ij} = \exp\{-\xi d_{i,j}^{(2)}\}, \xi > 0$
7. $w_{ij} = \exp\{-\xi d_{i,j}^{(3)}\}, \xi > 0$
Proximity matrix $W^{(1)}$

\[
\begin{pmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0
\end{pmatrix}
\]
Measure of spatial association

A standard statistic that is used to measure strength of spatial association among areas is *Moran’s $I$*:

$$I = \frac{m \sum_i \sum_j w_{ij} (Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_{i \neq j} w_{ij} \sum_i (Y_i - \bar{Y})^2}.$$

It is asymptotically normal if $Y_i$ are i.i.d.
Moran’s $I$

Moran’s $I$ for matrices $W^{(1)}$, $W^{(2)}$, $W^{(3)}$, $W^{(4)}$.

<table>
<thead>
<tr>
<th>Matrix</th>
<th>$W^{(1)}$</th>
<th>$W^{(2)}$</th>
<th>$W^{(3)}$</th>
<th>$W^{(4)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I$</td>
<td>-0.1092</td>
<td>0.0067</td>
<td>-0.5421</td>
<td>0.1269</td>
</tr>
</tbody>
</table>

Moran’s $I$ for matrices $W^{(5)}$, $W^{(6)}$, $W^{(7)}$ with decay parameter $\xi$.

<table>
<thead>
<tr>
<th>$\xi$</th>
<th>$W^{(5)}$</th>
<th>$W^{(6)}$</th>
<th>$W^{(7)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>-0.0149</td>
<td>-0.0449</td>
<td>-0.0007</td>
</tr>
<tr>
<td>0.01</td>
<td>-0.0848</td>
<td>-0.3022</td>
<td>0.3472</td>
</tr>
<tr>
<td>0.1</td>
<td>-0.0303</td>
<td>-0.7175</td>
<td>-0.0377</td>
</tr>
</tbody>
</table>
No location effect

<table>
<thead>
<tr>
<th>Variable</th>
<th>lg(Land)</th>
<th>Reg</th>
<th>lg(Land) &amp; Reg</th>
<th>Ind</th>
<th>Lat</th>
<th>Long</th>
<th>Lat &amp; Long</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value</td>
<td>&lt;0.01</td>
<td>0.03</td>
<td>0.18</td>
<td>&lt;0.01</td>
<td>0.80</td>
<td>0.66</td>
<td>0.93</td>
</tr>
</tbody>
</table>
Location by region interaction

<table>
<thead>
<tr>
<th>Variable</th>
<th>Lat</th>
<th>Long</th>
<th>Lat &amp; Long</th>
<th>Reg</th>
<th>Lat &amp; Reg</th>
<th>Long &amp; Reg</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value</td>
<td>0.087</td>
<td>0.678</td>
<td>0.038</td>
<td>0.012</td>
<td>0.006</td>
<td>0.024</td>
</tr>
</tbody>
</table>
Simulation results

Average Relative Bias (ARB), Average Relative Root Mean Squared Error (ARRMSE)

<table>
<thead>
<tr>
<th>Matrix</th>
<th>ARB, %</th>
<th>ARRMSE, %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EBLUP</td>
<td>SEBLUP</td>
</tr>
<tr>
<td>$W^{(1)}$</td>
<td>-3.88</td>
<td>-4.02</td>
</tr>
<tr>
<td>$W^{(2)}$</td>
<td>-4.55</td>
<td>-4.47</td>
</tr>
<tr>
<td>$W^{(3)}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$W^{(4)}$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# ARB and ARRMSE

<table>
<thead>
<tr>
<th>Matrix</th>
<th>Criteria</th>
<th>ξ = 0.001</th>
<th>ξ = 0.01</th>
<th>ξ = 0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W^{(5)}$</td>
<td>ARB</td>
<td>-4.58</td>
<td>-3.06</td>
<td>-5.07</td>
</tr>
<tr>
<td></td>
<td>ARRMSE</td>
<td>123.94</td>
<td>225.05</td>
<td>43.44</td>
</tr>
<tr>
<td>$W^{(6)}$</td>
<td>ARB</td>
<td>-2.72</td>
<td>-3.11</td>
<td>-4.72</td>
</tr>
<tr>
<td></td>
<td>ARRMSE</td>
<td>214.45</td>
<td>114.73</td>
<td>65.77</td>
</tr>
<tr>
<td>$W^{(7)}$</td>
<td>ARB</td>
<td>-1.88</td>
<td>-5.52</td>
<td>-4.61</td>
</tr>
<tr>
<td></td>
<td>ARRMSE</td>
<td>31.03</td>
<td>27.64</td>
<td>60.13</td>
</tr>
</tbody>
</table>
## Relative Bias and Relative Root Mean Squared Error

<table>
<thead>
<tr>
<th>(i)</th>
<th>Relative Bias, %</th>
<th>Relative Root Mean Squared Error, %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EBLUP</td>
<td>(W^{(1)})</td>
</tr>
<tr>
<td>1</td>
<td>-5.43</td>
<td>-3.40</td>
</tr>
<tr>
<td>2</td>
<td>-4.44</td>
<td>-2.20</td>
</tr>
<tr>
<td>3</td>
<td>-10.32</td>
<td>-4.74</td>
</tr>
<tr>
<td>4</td>
<td>-3.98</td>
<td>2.15</td>
</tr>
<tr>
<td>5</td>
<td>-9.90</td>
<td>-5.02</td>
</tr>
<tr>
<td>6</td>
<td>-10.23</td>
<td>-4.13</td>
</tr>
<tr>
<td>7</td>
<td>-6.37</td>
<td>-4.43</td>
</tr>
<tr>
<td>8</td>
<td>6.56</td>
<td>-2.04</td>
</tr>
<tr>
<td>9</td>
<td>-0.87</td>
<td>-7.34</td>
</tr>
<tr>
<td>10</td>
<td>-8.09</td>
<td>-8.20</td>
</tr>
<tr>
<td>11</td>
<td>-2.60</td>
<td>-3.80</td>
</tr>
<tr>
<td>12</td>
<td>0.97</td>
<td>-3.40</td>
</tr>
</tbody>
</table>
Concluding remarks

- Inclusion of spatial structure in small area estimation may give bad results if there is little evidence of spatial correlation in the data.

- Development of alternative robust models for borrowing strength over space in small area estimation.

- Further research with other data sets.