Spatiality in small area estimation: A new structure with a simulation study

Yadollah Mehrabi,1 Amir Kavousi,2 Ahmad-Reza Baghestani,3,4 Mojtaba Soltani-Kermanshahi5

1Department of Epidemiology, School of Public Health and Safety, Shahid Beheshti University of Medical Sciences; 2Workplace Health Promotion Research Center and Department of Epidemiology, School of Public Health and Safety, Shahid Beheshti University of Medical Sciences; 3Physiotherapy Research Centre, Shahid Beheshti University of Medical Sciences; 4Department of Biostatistics, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran; 5Social Determinants of Health Research Center, School of Medicine, Semnan University of Medical Sciences, Semnan, Iran

Abstract

In numerous practical applications, data from neighbouring small areas present spatial correlation. More recently, an extension of the Fay–Herriott model through the Simultaneously Auto-Regressive (SAR) process has been considered. The Conditional Auto-Regressive (CAR) structure is also a popular choice. The reasons of using these structures are theoretical properties, computational advantages and relative ease of interpretation. However, the assumption of the non-singularity of matrix (I−ρW) is a problem. We introduce here a novel structure of the covariance matrix when approaching spatiality in small area estimation (SAE) comparing that with the commonly used SAR process. As an example, we present synthetic data on grape production with spatial correlation for 274 municipalities in the region of Tuscany as base data simulating data at each area and comparing the results. The SAR process had the smallest Root Average Mean Square Error (RAMSE) for all conditions. The RAMSE also generally decreased with increasing sample size. In addition, the RAMSE values did not show a specific behaviour but only spatially correlated coefficient changes led to a stronger decrease of RAMSE values than the SAR model when our new structure was applied. The new approach presented here is more flexible than the SAR process without severe increasing RAMSE values.

Introduction

Sample surveys have for a long time been employed as a cost-effective remedy for data gathering. Such require proper statistics to effectively assess entire populations but need also to produce reliable outcomes for subpopulations, mostly entitled domains or areas, which are said to be “small” if the domain-specific sample size is not large sufficient to produce a direct estimate with acceptable precision. Because of the increasing demand for reliable estimates for these areas, Small Area Estimation (SAE) techniques are exposed to growing popularity in survey sampling (Rao, 2003; Rao and Wu, 2009; Soltani-Kermanshahi et al., 2017a; 2017b).

A direct estimate of a small area parameter(s) depends entirely on the samples from that area which is usually not reliable. However, much of the popularity and usefulness of SAE techniques can be attributed to model-based techniques, now widely used and with a growing application during the past three decades. These approaches take strength from neighbouring areas and use proper techniques that connect direct estimates from the small areas. When an area-level summary of covariate variables exists, we can usefully utilize these models. A fundamental area-level model was first proposed by Fay and Herriot (1979) and followed up by others (Rao, 2003; Waller and Gotway 2004; Rao and Wu, 2009; Soltani-Kermanshahi et al., 2017a; 2017b).

In numerous practical applications, data from neighbouring small areas present spatial correlation. In these instances, unless enough auxiliary variables are accessible, between-area correlations should in some way be illustrated in the covariance structure of the model. Nevertheless, the establishment of a structure of dependence between small areas requires a serious conceptual difference with regard to the traditional structure of independent small areas, where the entire covariance matrix is block-diagonal (Prasad and Rao, 1990; Molina et al., 2009; Gharde et al., 2012). Cressie (1991) already employed a model with spatially correlated random effects in the context of SAE. More recently, an extension
of the Fay–Herriot model through the simultaneously autoregres-
sive (SAR) process has been considered (Singh et al., 2005; 
Petrucci and Salvati, 2006; Pratesi and Salvati, 2008). The model
has the following form:
\[ Y = Xβ + v + e \]  \hspace{1cm} (Eq. 1.1)
where Y is the vector of design-unbiased direct estimator available
for each of the small areas (from 1 to m); X the vector of area-level
auxiliary covariates; e the vector of independent sampling errors;
and v the result of a SAR process with the unknown autoregression
parameter ρ (with range -1 to 1) and proximity matrix W (Anselin,
1988; Cressie, 1993). Then v can be expressed as:
\[ v = (I_m - ρW)^{-1}u \]  \hspace{1cm} (Eq. 1.2)
where the vector of variance components is denoted by ω.

The Conditional Auto-Regressive (CAR) structure is also a
popular choice and, in some papers, more common than the CAR.
The SAR structure is known as a subset of CAR structure (Arab
et al., 2017). The reasons of using these structures are due to their
theoretical properties, computational advantages and ease of inter-
pretations (Cressie, 1993; Yasui and Lele, 1997; Arab et al., 2017;
Li et al., 2009).

One problem of the SAR model is the estimation of ρ. Since,
even for using SAE techniques is the small sample size, then
we have a low precision in estimation of ρ.

A widely adopted choice for this correlation function is the
Mat’ern function, i.e.
\[ p(d) = \exp(-d) \]  \hspace{1cm} (Eq. 1.4)
when the smoothness parameter v = 0.5, the ‘decay parameter’ \( \eta \) =
1 and d is the distance between two areas (Waller and Gotway,
2004; Li et al., 2009).

Other correlation functions commonly used in this structure
are the exponential, Gaussian, and spherical. The most
commonly used correlation function in this structure is the
exponential. The exponential correlation function is given by
\[ G(\omega) = \sigma_u [1 - \rho \omega^2 + \rho \omega^4] \]  \hspace{1cm} (Eq. 1.5)
where \( \rho \) measures the correlation decay with the distance; \( \sigma_u \) the
smoothness parameter; \( \omega \) the correlation decay parameter.

Statistical approach

Suppose that we have M areas in the population and only m
areas are sampled. The SAR model with spatial dependence in
SAR structure has the following form as described by Molina
et al. (2009):
\[ G(\omega) = \sigma_u [(I_m - \rho W)^{-1} (I_m - \rho W)]^{-1} \]  \hspace{1cm} (Eq. 2.1)
where the vector of variance components is denoted by \( \omega \).
Alternatively, we used a new structure in the covariance matrix of
\( \nu \) as follows:
\[ u = (u_1, \ldots, u_m)^T \sim \text{MVN}(0, \Sigma_v) \]
\[ v_i = \sum_{j=1}^{m} \rho(d_{ij})u_j, i, j = 1, 2, \ldots, m \]  \hspace{1cm} (Eq. 2.2)
where V is a variance-covariance matrix with elements \( v_i \) and
the correlation function \( \rho(.) \), an isotropic correlation function
that decays as the Euclidean distance \( d_{ij} = \|s_i - s_j\| \) between two individ-
uals increases. In the exponential model \( \rho(.) \) has the form
described by Eq. 1.4. In this study, we used Restricted Maximum
Likelihood (REML) and Spatial Empirical Best Linear Unbiased
Prediction (SEBLUP) to estimate \( \sigma_u^2 \) and \( \beta \)'s, respectively.
P<0.05 was considered as significant.

Real data analysis

The data used were collected by the Statistical Center of Iran
(SCI) in 2013. The sample size was optimum at the province level
but not at the district level (SCI, 2013). The Iranian Rural and
Urban Household Expenditures and Income Survey (IRUHEIS)
are carried out annually by SCI. The sampling design was three-stage
with stratification. In the first stage, the census areas were deter-
mined and selected; in the second, the urban blocks and rural settle-
ments were selected and at the third, the sample households were
selected. We analyzed urban data of IRUHEIS 2013, collected
between 21 April 2013 and 20 April 2014.

In 2013, a total of 387 districts were selected by IRUHEIS
out of 429 districts of Iran’s urban areas. Of the 18,876 house-
holds that participated in IRUHEIS 2013, we analysed complete
data consisting of 18,850 households. The highest and lowest
sample sizes belonged to Tehran (998 samples) and Narmashir
(5 samples) districts, respectively. The HFE includes all payments
made to purchase the needed nutritious and essential food items
including meat, dairy products, cereal and bean, bread and flour,
biscuits and cakes, oil and butter, fruits and vegetables, nuts,
sweets and sugar, additives and dressings, as well as cigarettes
and tobacco (SCI, 2013). We also took into account some socio-
economic factors that may potentially affect the HFE. These var-
iables have been measured at the area level rather than the individ-
ual one. The data were extracted from reports disseminated by
the Iran Ministry of Interior and also by SCI.

We considered district-level variables emanating from Iran’s
census of 2011 that included the Average Number of Households
(ANH); the average number of rooms of each household; sex,
nutrition status, smoking status of the household head; and
employment status of the household head. We also included
some area-level variables that have been measured by the
Statistical Center of Iran, such as latitude and longitude.
ratio; the Proportion of Male Households (PMH); the proportion of the active population employed; the proportion of population of the following age groups: >65 years, 25-64 years, 15-24 years and <15 years; and the Proportion of Higher Education (PHE), (SCI, 2011). We also considered the Gross Domestic Product (GDP); the proportion of households that had joined a charity organization; the distance from province capital; the per capita income for municipalities; and the migration rate. These data had been produced by the Ministry of Interior in a project conducted for identification of less developed regions of the country. We also used geographic information such as latitude and longitude of each area capital to calculate exponential correlations (Iranian Noojum, 2018).

In order to reach the highest correlation with dependent variables, an appropriate transformation of independent variables (power for continuous variables and logarithm for proportion variables) was used. We also utilized the Variance Inflation Factors (VIF) to assess the collinearity among independent variables, with VIF>5 or coefficients of multiple determination with respect to other independent variables more than 0.8, which indicates serious multicollinearity for the predictor (Besley et al., 1980). The forward selection method was used to establish the final model. In the forward method, independent variables (ordered by their correlation with dependent variable) are inserted into the model serially where only significant variables remain.

According to VIF values, the variable proportion of population at 25 to 64 years was omitted (Supplementary file S1).

### Simulation with grape production data

To investigate the effects of spatiality on precision of estimation a simulation study was carried out using the new structure proposed in this paper. We used synthetic data based on grape production in the region of Tuscany, Italy with a spatial correlation for 274 municipalities. That dataset included a frame with 274 observations of the following variables: i) the direct estimators of the mean agrarian surface area used for production of grape (in hectares) for each Tuscany municipality (grapehect); ii) the agrarian surface area (in hectares) used for production (area); iii) the average number of working days in the reference year 2000 (workdays); and iv) the sampling variance of the direct estimators for each Tuscany municipality (var). We also added longitudinal and latitude for each area to that data. Population data come from the Italian Agricultural Census for the region of Tuscany in the year 2000 (Pratesi et al., 2013; Molina and Marhuenda, 2015). We generated data for each of the 274 areas by the following approach: i) Multivariate normal distribution with mean and co-variance of the correspondent real data was used, i.e., for the new structure \( \sigma^2 \) was 165 and \( d_i \), calculated from the following equation:

\[
d_{ij} = \sqrt{(\ell_i - \ell_j)^2 + (\ell_j - \ell_i)^2}
\]  

(Eq. 4.1)

where \( \ell_i \) and \( \ell_j \) are the latitude and longitude of areas. This gives a \( \sigma^2 \) of 101 and the autoregression parameter \( \rho \) 0.2 for the SAR structure. ii) The sample size at each area was closely mimicked the real data to make the simulation as realistic as possible; iii) Data generation was repeated 1000 times at each of 274 small areas and mean and variance of every area were computed; iv) To achieve small area estimations, we took advantage of the auxiliary variables included in the continuous variables of area (in hectares) and workdays from Tuscany in the year 2000, as they were found to be significant in real data analysis. v) We fitted the SAR and exponential structure spatial model to each of 1000 generated datasets and saved the estimates for each area. vi) To assess the precision of SAEs, we employed Root Average Mean Square Errors (RMSSEs) for small area estimates:

\[
\text{RMSE} = \left( \frac{\sum_{i=1}^{274} \text{MSE}_i}{274} \right)^{1/2}
\]  

(Eq. 4.2)

### Table 1. Results of spatial (exponential) model on households’ food expenditures.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Beta</th>
<th>SE</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANH</td>
<td>4.58x10^3</td>
<td>1.98x10^3</td>
<td>0.021</td>
</tr>
<tr>
<td>PMH</td>
<td>5.45x10^3</td>
<td>9.36x10^3</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>PHE</td>
<td>6.20x10^3</td>
<td>1.84x10^3</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>( \sigma^2 )</td>
<td>1.31x10^4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \text{AIC} )</td>
<td>15619</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ANH: Average Number of Household; PMH: proportion of Male Households (headed by a male); PHE: Proportion of Higher Education; SE=Standard Error; AIC=Akaike Information Criterion.

### Table 2. Descriptive statistics of variables in the Tuscany data.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grapehect</td>
<td>0.25</td>
<td>343.00</td>
<td>69.51</td>
<td>49.45</td>
</tr>
<tr>
<td>Area</td>
<td>35.81</td>
<td>2890</td>
<td>639.20</td>
<td>571.89</td>
</tr>
<tr>
<td>Workdays</td>
<td>22.15</td>
<td>497.70</td>
<td>149.60</td>
<td>67.49</td>
</tr>
</tbody>
</table>

### Table 3. Root Average Mean Square Error of small area estimates with area sample size and spatiality correlation coefficient changes in two types of spatiality structures by simulation.

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Method</th>
<th>Spatiality correlation coefficient</th>
<th>20% reduction</th>
<th>20% increase</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SAR</td>
<td>Exponential</td>
<td>218.79</td>
<td>218.87</td>
</tr>
<tr>
<td></td>
<td>Exponential</td>
<td>SAR</td>
<td>218.11</td>
<td>218.76</td>
</tr>
<tr>
<td>20% reduction</td>
<td></td>
<td>Real</td>
<td>219.42</td>
<td>219.19</td>
</tr>
<tr>
<td></td>
<td>Exponential</td>
<td>SAR</td>
<td>222.19</td>
<td>222.32</td>
</tr>
<tr>
<td></td>
<td>Exponential</td>
<td>Real</td>
<td>222.11</td>
<td>222.15</td>
</tr>
</tbody>
</table>

SAR: Simultaneously Auto-Regressive process.
where has the following form:

$$\text{MSE}_i = \frac{1}{1000} \sum_{q=1}^{1000} (\hat{g}_{(i,q)} - \hat{g}_{(i,M)})^2$$  \hspace{1cm} (Eq. 4.3)

where $\hat{g}_{(i,M)}$ is the estimate of $g_i$ based on real data with the two models (M) and $\hat{g}_{(i,q)}$ shows the estimated $g_i$ at run q (Soltani-Kermanshahi et al., 2017a; 2017b). For better assessment, we used sample size and correlation sensitive analysis using the real sample size and correlation as well as 20% above and below of it. All calculations used utilized the R package SAE (Molina and Marhuenda, 2015) and the codes given in Supplementary files S2 and S3).

Simulation results

The Tuscany data had 274 areas and the descriptive statistics of the three variables used are shown in Table 2. The results of RAMSE with sample size and spatial correlation coefficient sensitivity are shown in Table 3 and Figures 2 and 3. Accordingly, SAR process approximately had smallest RAMSE for all conditions. Also, with increasing sample size, RAMSE roughly decreased. The increase in accuracy powerfully related to areas sample size. In addition, with increasing spatial correlation coefficient, RAMSEs have not specific behavior but in most times, they decreased, approximately.
Discussion and conclusions

This paper proposes a new structure in the covariance matrix of $\nu$ (see Eq. 1.1) by using an exponential model in Mat’ern function (Cressie, 1993; Li et al., 2009) after considering a Fay–Herriot model with correlated random area effects according to the SAR process. Our study shows three significant variables, ANH, PMH and PHE, had a significant effect on HFE (Table 1), the zoning of which showed that the border districts, especially those on the western border of Iran, had a higher HFE than in other parts of the country (Figure 1). Although we had expected the highest HFE in the districts of the capital, our findings showed the opposite. One reason for this could be the way Iran has developed economically.

From a purely statistical point of view, the position of the small area is related to modelling of its parameters and further improvement in the EBLUP estimator can be achieved by including probable spatial interaction among random area effects as discussed previously (Petrucci and Salvati, 2004; Pratesi and Salvati, 2009). Indeed, the inclusion of ancillary variables to obtain the spatial effects may be beneficial even when the strength of the spatial link is weak (Petrucci et al., 2005).

The simulation study with sensitive analysis used to compare the RAMSE values of the SAR and the proposed new structure showed that that sample size changes have a higher impact on RAMSE than spatial correlation coefficient changes. One potential problem with SAE is small subdomain samples. The expected RAMSE than spatiality correlation coefficient changes lead to a stronger decrease of the RAMSE values of the SAR and the proposed new structure is weak (Petrucci et al., 2005).

The simulation study with sensitive analysis used to compare the RAMSE values of the SAR and the proposed new structure showed that the expected RAMSE values with bigger samples. In our simulations, there were no areas without sample data. For such areas, for which the values of the covariates at the area-level are available from any other data source, possible estimators are $\gamma^*=X^*\beta^*$ (Molina et al., 2009). In addition, our results showed that only spatially correlation coefficient changes lead to a smaller decrease of RAMSE values than SAR model when our new structure was applied.

The SAR process had the smallest RAMSE values for all conditions in relation to the new process. Even if SAR has appealing theoretical properties, computational advantages and provides ease of interpretation (Cressie, 1993; Yau and Lele, 1997; Li et al., 2009), it assumes non-singularity of matrix $(I_n-pW)$, which can be avoided by the new process. In general, the RAMSE values were increased about 1.5% without the assumption of non-singularity of matrix $(I_n-pW)$ by the new process. The new process is almost more flexible than the SAR process without severe increasing in RAMSEs. With high spatial correlations in the study variables, spatial EBLUP with correlated random area effects following a SAR process, works better (Petrucci et al., 2005; Pratesi and Salvati, 2009). However, in our work the RAMSE values increased in certain situations with increasing spatially correlation coefficients.

The biggest difference of our study related to similar studies was the insufficient number of households in some districts, but the degree of weakness decreased by using the SAE approach. At least we could find the admissible model to predict the HFE in urban areas of Iran. Furthermore, one of the basic limitations in SAE methodology is the inclusion of highly correlated variables with dependent variables. In this study, we tried to find those variables and gain access to them.

References


Singh BB, Shukla GK, Kundu D, 2005. Spatio-temporal models in...