12-2021

msae: An R Package of Multivariate Fay-Herriot Models for Small Area Estimation

Novia Permatasari

Azka Ubaidillah

msae: An R Package of Multivariate Fay-Herriot Models for Small Area Estimation

by Novia Permatasari and Azka Ubaidillah

Abstract The paper introduces an R Package of multivariate Fay-Herriot models for small area estimation named **msae**. This package implements four types of Fay-Herriot models, including univariate Fay-Herriot model (model 0), multivariate Fay-Herriot model (model 1), autoregressive multivariate Fay-Herriot model (model 2), and heteroskedastic autoregressive multivariate Fay-Herriot model (model 3). It also contains some datasets generated based on multivariate Fay-Herriot models. We describe and implement functions through various practical examples. Multivariate Fay-Herriot models produce a more efficient parameter estimation than direct estimation and univariate model.

Introduction

Survey sampling is a data collection method by observing several units of observation to obtain information from the entire population. Compared to other data collection methods, survey sampling has advantages in cost, time, and human resources. Survey sampling is designed for a certain size of the domain, commonly a large area. However, data demand for small areas is increasing and has become high issue (Ghosh and Rao, 1994). The inadequate sample size causes a large standard error of parameter estimates. This problem is overcome by indirect estimation, namely Small Area Estimation (SAE).

Rao and Molina (2015) said that SAE increases the effectiveness of sample size using the strength of neighboring areas and information on other variables that are related to the variable of interest. There are some estimation methods in SAE, namely Best Linear Unbiased Predictors (BLUP), Empirical Best Linear Unbiased Predictors (EBLUP), Hierarchical Bayes (HB), and Empirical Bayes (EB). The most common SAE estimator is the EBLUP (Krieg et al., 2015). EBLUP has advantages over EB and HB methods. It is a development of the BLUP method that minimizes the MSE among other unbiased linear estimators (Ghosh and Rao, 1994). Area level of EBLUP application for the continuous response variable, called Fay-Herriot model, was firstly employed for estimating log per-capita income (PCI) in small places in the US (Rao and Molina, 2015).

The Fay-Herriot model has extended into a multivariate Fay-Herriot model, which is a model with several correlated response variables. Datta et al. (1991) firstly applied the multivariate model to estimate the median income of four-person families in the US states. Benavent and Morales (2016) developed multivariate Fay-Herriot models with the EBLUP method and introduced four estimation models based on the estimated variance matrix structure. Ubaidillah et al. (2019) also implemented the multivariate Fay-Herriot model and indicated that the multivariate Fay-Herriot model produces a more efficient parameter estimation than the univariate model.

On the Comprehensive R Archive Network (CRAN), there are several packages implementing small area estimation. Some of them are included in the Small Area Estimation subsection of The CRAN Task View: Official Statistics & Survey Methodology (Templ, 2014), including **sae** (Molina and Marhuenda, 2018), **rsae** (Schoch, 2014), **nlme** (Pinheiro et al., 2020), **hbsae** (Boonstra, 2012), **JoSAE** (Breidenbach, 2018), and **BayesSAE** (Chengchun Shi Developer, 2018). Other popular SAE packages not included in that subsection are **mme** (Lopez-Vizcaino et al., 2019) and **saery** (Lefler et al., 2014).

In this paper, we introduce our R package of multivariate Fay-Herriot models for small area estimation, named **msae**. This package and its details are available on CRAN at `http://CRAN.R-project.org/package=msae`. Functions in this package implement four Fay-Herriot Models, namely Model 0, Model 1, Model 2, and Model 3, as proposed by Benavent and Morales (2016).

The paper is structured as follows. First, we explain multivariate Fay-Herriot models in Section 2.2. Then, we describe **msae** package and illustrate the use of this package for SAE estimation by employing simulation studies and applying it to a real dataset in the next Sections 2.3 and 2.4. Finally, we provide a conclusion in Section 2.8.

Multivariate Fay-Herriot model

The multivariate Fay-Herriot model is an extension of the Fay-Herriot model, which utilizes some correlated responses. Fay-Herriot is a combination of two model components. The first component is called the sampling model, and the second component is called the linking model.

Suppose we want to estimate characteristics of R variables in D areas, $\mu_d = (\mu_{1d}, \ldots, \mu_{Rd})^T$, with $d = 1, \ldots, D$. Let $y_d = (y_{1d}, \ldots, y_{Rd})^T$, be a direct estimator of μ_d. The first component, i.e., sampling model is

$$y_d = \mu_d + e_d, \quad e_d \sim N(0, V_{e_d}), \quad d = 1, \ldots, D$$

, where e_d is sampling error with a covariance matrix, V_{e_d} , that is assumed to be known. In the second component, we assume that μ_d is linearly related with p_d area-specific auxiliary variables $X_d = (X_1, \ldots, X_{p_d})^T$ as follows:

$$\mu_d = X_d\beta_d + u_d, \quad u_d \sim N(0, V_{u_d}), \quad d = 1, \ldots, D$$

, where u_d is area random effects, and β_d is a vector of regression coefficient corresponding with X_d. This second component is called the linking model. The combination of the two components forms a multivariate linear mixed model as follows:

$$y_d = X_d\beta_d + u_d + e_d, \quad e_d \sim N(0, V_{e_d}), \quad d = 1, \ldots, D$$

, where u and e are independent.

Benavent and Morales (2016) proposed Fay-Herriot models using four different variance matrices, Model 0, Model 1, Model 2, and Model 3. Model 0 is a univariate Fay-Herriot Model, of which the sampling error and the random effect of target variables are independent. Sampling error and random effect variance matrix are written as follows:

$$V_{ud} = diag_{a \le r \le R}(\sigma^2_{ur})$$

$$V_{ed} = diag_{a \le r \le R}(\sigma^2_{edr})$$

,

$$\text{where } d = 1, \ldots, D$$

.

Model 1, Model 2, and Model 3 are multivariate Fay-Herriot models, of which the variance matrices are no longer diagonal matrices. Model 1 is a multivariate form of Model 0, where the random effect variance of Model 1 is still a diagonal matrix. Model 2 is called the autoregressive multivariate Fay-Herriot model (AR(1)), in which the random variance matrix is written as follows :

$$V_{ud} = \sigma^2_{ur}\Omega_d(\rho)$$

$$\Omega_d(\rho) = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & \rho & \cdots & \rho^{R-1} \\ \rho & 1 & & \rho^{R-2} \\ \vdots & & \ddots & \vdots \\ \rho^{R-1} & \rho^{R-2} & \cdots & 1 \end{bmatrix}$$

Model 3 is called heteroskedastic autoregressive multivariate Fay-Herriot model (HAR(1)), which the element of random error is written as follows:

$$u_{dr} = \rho u_{dr-1} + a_{dr}$$

$$u_{d0} \sim N(0, \sigma^2_{u0}) \quad \text{and} \quad a_{dr} \sim N(0, \sigma^2_r)$$

, where $\sigma^2_{u0} = 1, a_{dr}$, and u_{d0} are independent. The element of random variance matrix is written as follow:

$$\sigma_{drii} = \sum_{k=1}^{i} \rho^{2k}\sigma^2_{i-k}$$

$$\sigma_{drij} = \sum_{k=0}^{|i-j|} \rho^{2k+|i-j|}\sigma^2_{|i-j|-k}$$

.

BLUP and EBLUP

The best linear unbiased prediction (BLUP) of μ is

$$\hat{\mu} = X\hat{\beta} + Z^T\hat{V}_u Z\Omega^{-1}(y - X\hat{\beta})$$

, where $\hat{\beta} = (X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1}y$ is the best linear unbiased estimator (BLUE) of β with the covariance matrix $cov(\hat{\beta}) = (X^T\Omega^{-1}X)^{-1}$. BLUP estimator depends on the random effect variance that is usually unknown. Using Restricted Maximum Likelihood (REML), we estimate and substitute random effect variance estimator for obtaining the multivariate EBLUB estimator. The estimation formula with EBLUP is written as follows:

$$\hat{\mu} = X\hat{\beta} + Z^T\hat{V}_u Z\hat{\Omega}^{-1}(y - X\hat{\beta})$$

$$\hat{\Omega} = Z^T\hat{V}_u Z + V_e$$

, where $\hat{\beta} = (X^T\hat{\Omega}^{-1}X)^{-1}X^T\hat{\Omega}^{-1}y$ is the best linear unbiased estimator (BLUE) of β with the covariance matrix $cov(\hat{\beta}) = (X^T\hat{\Omega}^{-1}X)^{-1}$.

MSE

Benavent and Morales (2016) also proposed MSE estimation for the multivariate Fay-Herriot models as follows:

$$mse(\hat{\mu}) = g_{1i}(\hat{\sigma}_u^2) + g_{2i}(\hat{\sigma}_u^2) + g_{3i}(\hat{\sigma}_u^2)$$

, where each component can be described as follows:

$$g_{1i}(\hat{\sigma}_u^2) = \Gamma V_e$$

,

$$g_{2i}(\hat{\sigma}_u^2) = (1 - \Gamma)X(X^T\Omega^{-1}X)^{-1}X^T(I - \Gamma)^T$$

,

$$g_{3i}(\hat{\sigma}_u^3) \approx \Sigma\Sigma cov(\hat{\sigma}_{uk}^2, \hat{\sigma}_{ul}^2)\Gamma_{(k)}\Omega\Gamma_{(k)}^T, k, l = 1, 2, \ldots, q$$

, where $\Gamma = Z^T\hat{V}_u Z$, $\Gamma_{(k)} = \frac{\delta\Gamma}{\delta\sigma_u^2}$, and $cov(\hat{\sigma}_{uk}^2, \hat{\sigma}_{ul}^2)$ is the inverse of the Fisher information matrix in the estimation of REML.

Overview of R package msae

The R Package **msae** implements multivariate Fay-Herriot models for small area estimation. Here are some functions and the descriptions at a glance:

- `eblupUFH`: This function gives the EBLUP and MSE based on the univariate Fay-Herriot model (Model 0).
- `eblupMFH1`: This function gives the EBLUP and MSE based on the multivariate Fay-Herriot model (Model 1).
- `eblupMFH2`: This function gives the EBLUP and MSE based on the autoregressive multivariate Fay-Herriot model (Model 2).
- `eblupMFH3`: This function gives the EBLUP and MSE based on the heteroskedastic autoregressive multivariate Fay-Herriot model (Model 3).

Those functions return a list of five elements:

- `eblup` is a data frame of EBLUPs for each variable.
- `MSE` is a data frame of the estimated MSEs of the EBLUPs.
- `randomEffect` is a data frame of random effect estimators.
- `Rmatrix` is a block diagonal matrix composed of sampling variances.
- `fit` is a list containing the following objects:
  - `method` shows the type of fitting method.
  - `convergence` shows the convergence of the Fisher Scoring algorithm.
  - `estcoef` shows estimated model coefficients and their significance.
  - `refvar` shows the estimated random effect variance.

- `refvarTest` (only for `eblupMFH3`) shows homogeneity of random effect variance test based on Model 3.
- `rho` (only for `eblupMFH2` and `eblupMFH3`) shows the estimated ρ of random effect variance and their parameter test.
- `informationFisher` is a matrix of information Fisher.

This package also provides datasets generated for each multivariate model. The datasets are generated based on Model 1, Model 2, and Model 3 following steps:

1. Generate sampling error e and auxiliary variables « $X1, X2$ ». Set the parameter as follows:
   - For sampling error e in Model 1, we set $e_d \sim N_3(0, V_{ed})$, where $V_{ed} = (\sigma_{dij})_{i,j=1,2,3}$ with $\sigma_{e11} \sim InvGamma(11,1)$, $\sigma_{e22} \sim InvGamma(11,2)$, $\sigma_{e33} \sim InvGamma(11,3)$, and $\rho_e = 0.5$. We generate different MSE of direct estimates for each area.
   - For sampling error e in Model 2 and Model 3, we set $e \sim N_3(0, V_e)$, where $V_e = (\sigma_{ij})_{i,j=1,2,3}$ with $\sigma_{e11} = 0.1$, $\sigma_{e22} = 0.2$, $\sigma_{e33} = 0.3$, and $\rho_e = 0.5$. It is shown that all the areas have the same MSE of direct estimates.
   - For auxiliary variables « $X1$ $X2$ », we set $X1 \sim N(5, 0.1)$ and $X2 \sim N(10, 0.2)$
2. Generate random effect u, where $u \sim N_3(0, V_u)$. For each dataset, parameter for generating random effect u are as follows:
   - For Model 1, $\sigma_{u11} = 0.2$, $\sigma_{u22} = 0.4$, and $\sigma_{u33} = 1.2$
   - For Model 2, $\sigma_u = 0.4$, and $\rho_u = 0.8$
   - For Model 3, $\sigma_{u11} = 0.2$, $\sigma_{u22} = 0.4$, $\sigma_{u33} = 1.2$, and $\rho_u = 0.8$
3. Set $\beta_1 = 5$ and $\beta_2 = 10$ to calculate direct estimation « $Y1$, $Y2$, and $Y3$ », where $Y_i = X\beta + u_i + e_i$.
4. Auxiliary variables « $X1$ and $X2$ », direct estimates « $Y1, Y2$, and $Y3$ », and sampling variance-covariance « $v1, v2, v3, v12, v13$, and $v23$ » are combined into a data frame called `datasae1` for Model 1, `datasae2` for Model 2, and `datasae3` for Model 3.

Example 1. The multivariate Fay-Herriot model (Model 1)

`datasae1`, which is generated based on Model 1, contains 50 observations on the following 11 variables: 3 dependent variables « $Y1, Y2$, and $Y3$ », 2 auxiliary variables « $X1$ and $X2$ », and 6 variance-covariance of direct estimation « $v1, v2, v3, v12, v13$, and $v23$ ». The procedures for generating such datasets are provided in the previous section. The following R commands are run to obtain EBLUPs of the univariate Fay-Herriot model (Model 0) and the multivariate Fay-Herriot model (Model 1), plot the EBLUPs of the univariate and multivarate model, and plot the MSEs of EBLUPs in the univariate and multivarate model:

```
data('datasae1')

# model specifications
Fo <- list(f1=Y1~X1+X2,
           f2=Y2~X1+X2,
           f3=Y3~X1+X2)
vardir <- c("v1", "v2", "v3", "v12", "v13", "v23")

# EBLUP based on Model 0 and Model 1
u <- eblupUFH(Fo, vardir, data=datasae1)  # Model 0
m1 <- eblupMFH1(Fo, vardir, data=datasae1)  #Model 1

# Figure 1: EBLUPs under Model 0 and Model 1
par(mfrow=c(1,3))

plot(u$eblup$Y1, type = "o", col = "blue", pch = 15, xlab = "area", ylab = "Y1",
     cex.axis = 1.5, cex.lab = 1.5, xaxt = "n")
points(m1$eblup$Y1, type = "o", col = "red", pch = 18)
axis(1, at=1:50, labels = 1:50)
legend("topleft",legend=c("Model 0","Model 1"),ncol=2 , col=c("blue","red"),
    pch=c(15,18), inset=c(0.617,-0.1), xpd=TRUE)
```

```
plot(u$eblup$Y2, type = "o", col = "blue", pch = 15, xlab = "area", ylab = "Y2",
     cex.axis = 1.5, cex.lab = 1.5, xaxt = "n")
points(m1$eblup$Y2, type = "o", col = "red", pch = 18)
axis(1, at=1:50, labels = 1:50)
legend("topleft",legend=c("Model 0","Model 1"),ncol=2 , col=c("blue","red"),
    pch=c(15,18), inset=c(0.617,-0.1), xpd=TRUE)
plot(u$eblup$Y3, type = "o", col = "blue", pch = 15, xlab = "area", ylab = "Y3",
     cex.axis = 1.5, cex.lab = 1.5, xaxt = "n")
points(m1$eblup$Y3, type = "o", col = "red", pch = 18)
axis(1, at=1:50, labels = 1:50)
legend("topleft",legend=c("Model 0","Model 1"),ncol=2 , col=c("blue","red"),
    pch=c(15,18), inset=c(0.617,-0.1), xpd=TRUE)


# Figure 2: MSE of Model 0 and Model 1
par(mfrow=c(1,3))

plot(u$MSE$Y1, type = "o", col = "blue", pch = 15, xlab = "area", ylab = "MSE of Y1",
     cex.axis = 1.5, cex.lab = 1.5, xaxt = "n", ylim=c(0.038,0.12))
points(m1$MSE$Y1, type = "o", col = "red", pch = 18)
axis(1, at=1:50, labels = 1:50)
legend("topleft",legend=c("Model 0","Model 1"),ncol=2 , col=c("blue","red"),
       pch=c(15,18), inset=c(0.617,-0.1), xpd=TRUE)
plot(u$MSE$Y2, type = "o", col = "blue", pch = 15, xlab = "area", ylab = "MSE of Y2",
     cex.axis = 1.5, cex.lab = 1.5, xaxt = "n", ylim=c(0.05,0.28))
points(m1$MSE$Y2, type = "o", col = "red", pch = 18)
axis(1, at=1:50, labels = 1:50)
legend("topleft",legend=c("Model 0","Model 1"),ncol=2 , col=c("blue","red"),
       pch=c(15,18), inset=c(0.617,-0.1), xpd=TRUE)
plot(u$MSE$Y3, type = "o", col = "blue", pch = 15, xlab = "area", ylab = "MSE of Y3",
     cex.axis = 1.5, cex.lab = 1.5, xaxt = "n", ylim=c(0.1,0.42))
points(m1$MSE$Y3, type = "o", col = "red", pch = 18)
axis(1, at=1:50, labels = 1:50)
legend("topleft",legend=c("Model 0","Model 1"),ncol=2 , col=c("blue","red"),
       pch=c(15,18), inset=c(0.617,-0.1), xpd=TRUE)
```

Figure 1 illustrates the EBLUPs based on Model 0 (univariate model) and Model 1 (multivariate model) for all variables of interest. Figure 2 shows the MSEs of Model 1 compared with the MSEs of Model 0. It can be seen that the estimates of both methods show a similar pattern. Meanwhile, EBLUPs based on Model 1 has a lower MSE than Model 0. From this example, we can conclude that the multivariate Fay-Herriot model (Model 1) is more efficient than the univariate Fay-Herriot model (Model 0).

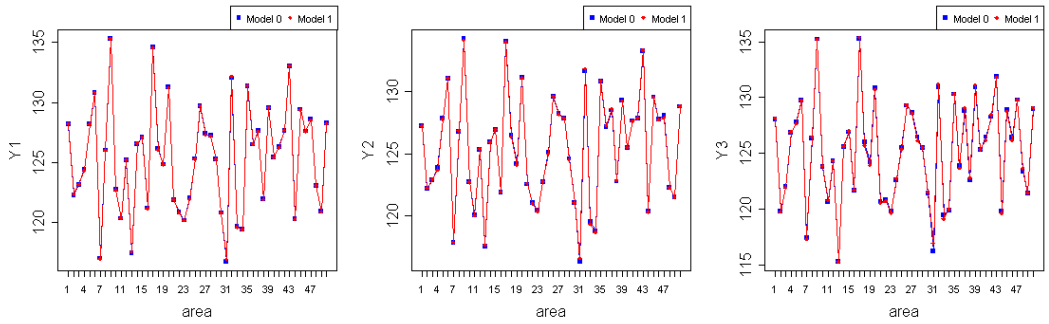

Figure 1: EBLUPs under Model 0 and Model 1.

Example 2. The autoregressive multivariate Fay-Herriot model (Model 2)

datasae2, which was generated based on Model 2, contains 50 observations on the following 11 variables: 3 dependent variables « $Y1, Y2$ and $Y3$ », 2 auxiliary variables « $X1$ and $X2$ », and 6 variance-

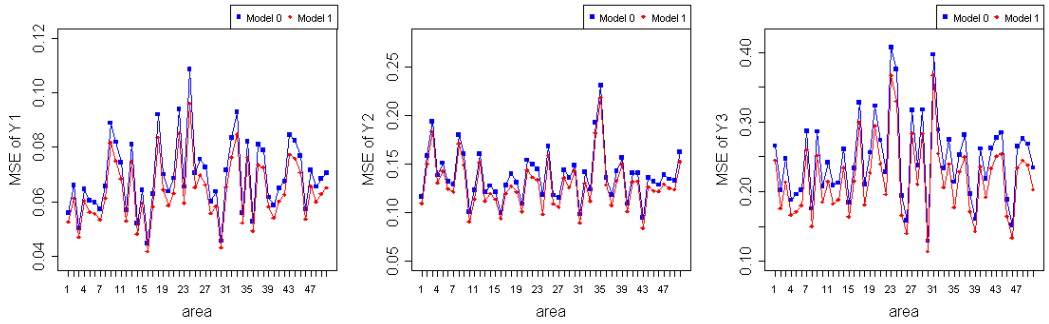

Figure 2: MSE of EBLUPs under Model 0 and Model 1.

covariance of direct estimation « $v1, v2, v3, v12, v13$, and $v23$ ». We compare the effectiveness of the univariate model (Model 0) and autoregressive multivariate Fay-Herriot Model (Model 2) by their MSE. We use `eblupMFH2()` to estimate parameters based on Model 2. Then, we plot the EBLUP and MSE of these methods to compare them.

```
data('datasae2')

# model specifications
Fo <- list(f1=Y1~X1+X2,
           f2=Y2~X1+X2,
           f3=Y3~X1+X2)
vardir <- c("v1", "v2", "v3", "v12", "v13", "v23")

# EBLUP based on Model 0 and Model 2
u <- eblupUFH(Fo, vardir, data=datasae2)  # Model 0
m2 <- eblupMFH2(Fo, vardir, data=datasae2)  # Model 2
```

The EBLUPs based on Model 0 (univariate model) and Model 1 (autoregressive multivariate model) are shown in Figure 3. As it can be seen, both methods show an almost similar result. However, EBLUPs based on Model 2 has a lower MSE than the EBLUPs based on Model 0, as shown in Figure 4. In this example, the autoregressive multivariate Fay-Herriot model (Model 2) seems to be more efficient than the univariate Fay-Herriot model (Model 0).

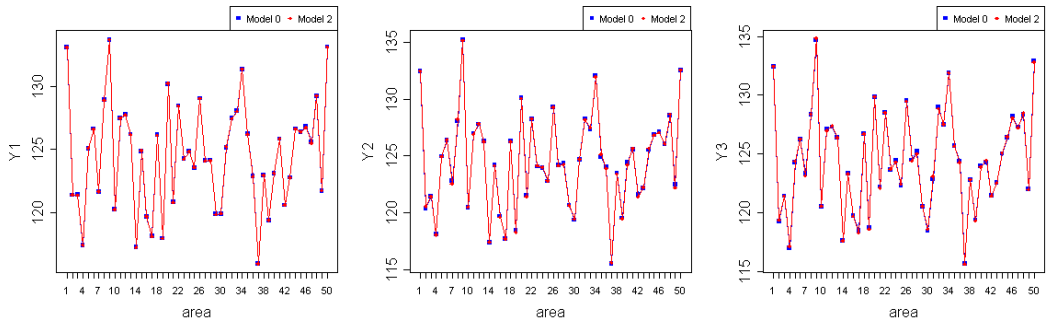

Figure 3: EBLUPs under Model 0 and Model 2.

Example 3. Heteroskedastic autoregressive multivariate Fay-Herriot model (Model 3)

`datasae3`, which was generated based on Model 3, is structured the same as `datasae1` and `datasae2`. We compare the effectiveness of the univariate model (Model 0) and heteroskedastic autoregressive multivariate Fay-Herriot Model (Model 3) by their MSE. We use `eblupMFH3()` to estimate parameters based on Model 3. Then, we plot the EBLUP and MSE of these methods to compare them.

```
data('datasae3')
```

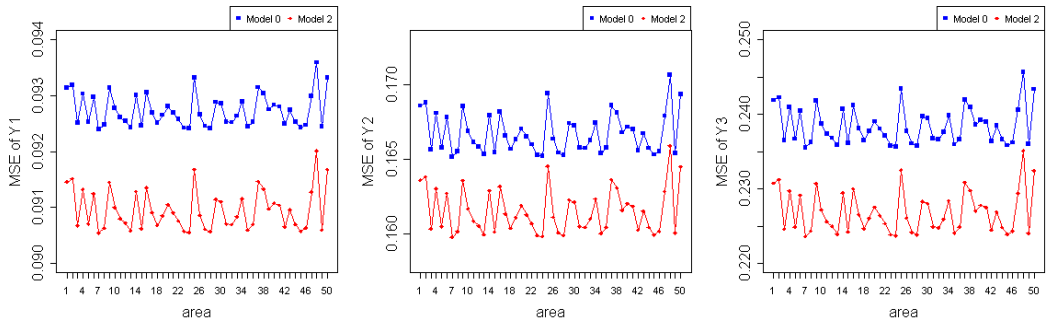

Figure 4: MSE of EBLUPs under Model 0 and Model 2.

```
# model specifications
Fo <- list(f1=Y1~X1+X2,
           f2=Y2~X1+X2,
           f3=Y3~X1+X2)
vardir <- c("v1", "v2", "v3", "v12", "v13", "v23")

# EBLUP based on Model 0 and Model 3
u <- eblupUFH(Fo, vardir, data=datasae3)  # Model 0
m3 <- eblupMFH3(Fo, vardir, data=datasae3)  # Model 3
```

Figure 5 shows EBLUPs based on Model 3 compared to Model 0. The MSEs of both methods are shown in Figure 6. It can be seen that the multivariate EBLUPs will follow the pattern of univariate EBLUPs with smaller MSE values. It can be concluded that the heteroskedastic autoregressive multivariate Fay-Herriot model (Model 3) is more efficient than the univariate model (Model 0).

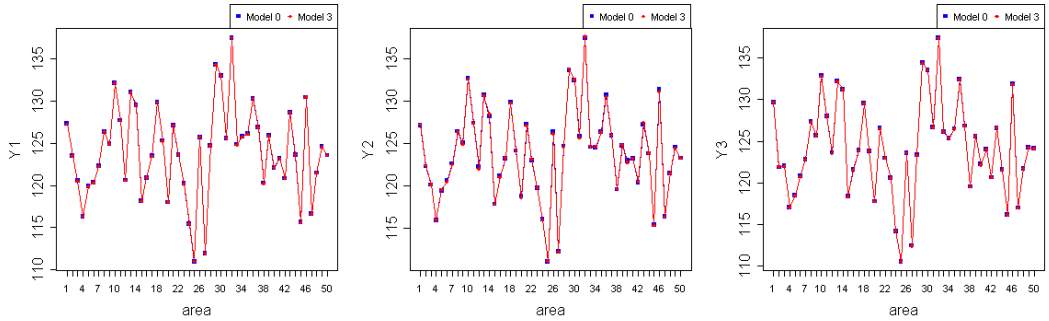

Figure 5: EBLUPs under Model 0 and Model 3.

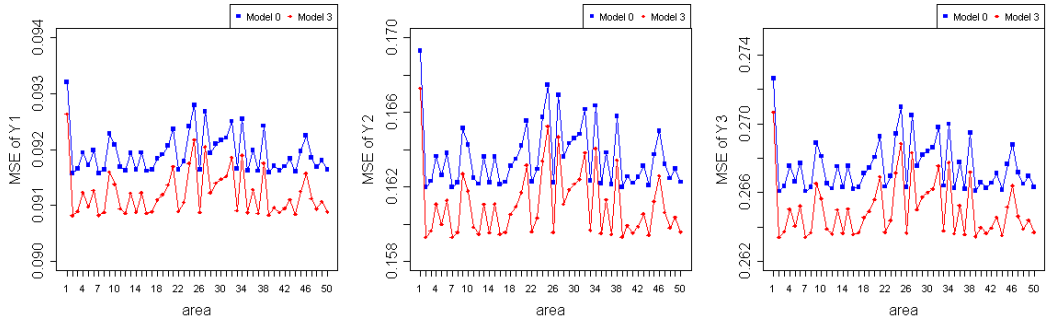

Figure 6: MSE of EBLUPs under Model 0 and Model 3.

Real data example: Poverty index

In this section, we use `incomedata` dataset, which is provided in library **sae**. The dataset contains unit-level data on income and other related variables in Spain. We will demonstrate how to estimate the EBLUP of Foster-Greer-Thorbecke (FGT) poverty index for each province using the multivariate Fay-Herriot Models. The index consists of three indicators, i.e., poverty proportion (p0), poverty gap (p1), and poverty severity (p2).

```
library(sae)
data("incomedata")
```

Firstly, we obtain area-level data by calculating poverty indicators for each unit and aggregating it by province based on Benavent and Morales (2016). We use poverty line z=6557.143 (Molina and Marhuenda, 2015). These following R commands are run to obtain $p0$, $p1$, and $p2$ as variables of interest.

```
library(tidyverse)

pov.line <- rep(6557.143, dim(incomedata)[1]) # poverty line

# calculate unit indictators
incomedata$y <- (pov.line - incomedata$income)/pov.line
incomedata = incomedata %>% mutate(poverty = ifelse(incomedata$y > 0, TRUE, FALSE),
                                   y0 = ifelse(incomedata$y > 0, incomedata$y^0, 0),
                                   y1 = ifelse(incomedata$y > 0, incomedata$y^1, 0),
                                   y2 = ifelse(incomedata$y > 0, incomedata$y^2, 0))

# estimated domain size
est.Nd <- aggregate(incomedata$weight, list(incomedata$prov), sum)[,2]

## estimate P0 P1 dan P2
prov.est = incomedata %>% group_by(prov) %>%
  summarise(p0.prov = sum(weight*y0),
            p1.prov = sum(weight*y1),
            p2.prov = sum(weight*y2)) %>%
  mutate(p0.prov = p0.prov / est.Nd,
         p1.prov = p1.prov / est.Nd,
         p2.prov = p2.prov / est.Nd)
incomedata <- incomedata %>% left_join(prov.est, by = c("prov" = "prov"))
```

We also need variance and covariance of variables of interest to estimate using the multivariate Fay-Herriot model. The following R commands are run to obtain variance and covariance of direct estimation based on Benavent and Morales (2016).

```
# estimate  direct estimation variance-covariance
prov.variance = incomedata %>%
  mutate(v11 = ifelse(incomedata$poverty,
                      (weight*(weight-1))*(y0-p0.prov)*(y0-p0.prov), 0),
         v12 = ifelse(incomedata$poverty,
                      (weight*(weight-1))*(y0-p0.prov)*(y1-p1.prov), 0),
         v13 = ifelse(incomedata$poverty,
                      (weight*(weight-1))*(y0-p0.prov)*(y2-p2.prov), 0),
         v22 = ifelse(incomedata$poverty,
                      (weight*(weight-1))*(y1-p1.prov)*(y1-p1.prov), 0),
         v23 = ifelse(incomedata$poverty,
                      (weight*(weight-1))*(y1-p1.prov)*(y2-p2.prov), 0),
         v33 = ifelse(incomedata$poverty,
                      (weight*(weight-1))*(y2-p2.prov)*(y2-p2.prov), 0)) %>%
  group_by(prov) %>%
  summarise_at(c("v11","v12", "v13", 'v22',"v23","v33"), sum) %>%
  mutate_at(c("v11","v12", "v13", 'v22',"v23","v33"),
            function(x){x/est.Nd^2})
```

We use six explanatory variables selected by stepwise method, i.e., an indicator of age group 50-64 ($age4$), an indicator of age group >=65 ($age5$), an indicator of education level 1 ($educ1$), an indicator

of education level 2 (*educ*2), an indicator of education level 3 (*educ*3), and an indicator of Spanish nationality (*nat*1). The model specifications are written as follow:

```
formula <- list(f1=p0.prov~age4+age5+educ1+educ2+educ3+nat1,
            f2=p1.prov~age4+age5+educ1+educ2+educ3+nat1,
            f3=p2.prov~age4+age5+educ1+educ2+educ3+nat1)
vardir <- c("v11", "v22", "v33", "v12", "v13", "v23")
```

Next, we select the most suitable multivariate Fay-Herriot model using variance homogeneity test and random effect ρ parameter test. In the variance homogeneity test, we test $H0 : \hat{\sigma}^2_{ui} = \hat{\sigma}^2_{uj}$; $i, j = 1, 2, 3$ using model 3. We obtain a non-convergent model, the p-values are 0.98674 and 0.98996. It shows that the difference between variance of random effects is statistically not significant. After that, we test $H0 : \rho = 0$ using model 2. We get the t-statistics value of 19.26 with p-value of 0.00. It shows that there is a correlation between random effects. These results indicate the model that fits the data is Model 2.

The following codes are run to obtain the EBLUPs (under Model 2), to plot the direct estimates and the EBLUPs, and to plot the MSEs of direct estimates and the MSEs of EBLUPs ordered by sample size:

```
# EBLUP based on Model 2
eblup.pov <- eblupMFH2(formula, vardir, data=prov.data)

# Dataframe of Result
result_eblup = data.frame(prov = prov.data$prov,
                          est.Nd = est.Nd,
                          p0 = prov.data$p0.prov, p0.eblup = eblup.pov$eblup$p0.prov,
                          p1 = prov.data$p1.prov, p1.eblup = eblup.pov$eblup$p1.prov,
                          p2 = prov.data$p2.prov, p2.eblup = eblup.pov$eblup$p2.prov,
                          v11 = prov.data$v11, p0.mse = eblup.pov$MSE$p0.prov,
                          v22 = prov.data$v22, p1.mse = eblup.pov$MSE$p1.prov,
                          v33 = prov.data$v33, p2.mse = eblup.pov$MSE$p2.prov)
result_eblup = result_eblup %>% arrange(est.Nd)
result_eblup$prov = as.factor(result_eblup$prov)
result_eblup$id = 1:nrow(result_eblup)

# Figure 7: Direct estimates estimates and EBLUPs (under Model 2) ordered by sample size.
par(mfrow=c(1,3))

plot(result_eblup$id, result_eblup$p0,  type = "o", col = "blue", pch = 15, xlab = "area",
    ylab = "p0", cex.axis = 1.5, cex.lab = 1.5, xaxt = "n")
points(result_eblup$p0.eblup, type = "o", col = "red", pch = 18)
axis(1, at=result_eblup$id, labels = result_eblup$prov)
legend("topright",legend=c("Direct estimates","Model 2"),col=c("blue","red"), pch=c(15,18))
plot(result_eblup$id, result_eblup$p1, type = "o", col = "blue", pch = 15, xlab = "area",
    ylab = "p1", cex.axis = 1.5, cex.lab = 1.5, xaxt = "n")
points(result_eblup$p1.eblup, type = "o", col = "red", pch = 18)
axis(1, at=result_eblup$id, labels = result_eblup$prov)
legend("topright",legend=c("Direct estimates","Model 2"),col=c("blue","red"), pch=c(15,18))
plot(result_eblup$id, result_eblup$p2, type = "o", col = "blue", pch = 15, xlab = "area",
    ylab = "p2", cex.axis = 1.5, cex.lab = 1.5, xaxt = "n")
points(result_eblup$p2.eblup, type = "o", col = "red", pch = 18)
axis(1, at=result_eblup$id, labels = result_eblup$prov)
legend("topright",legend=c("Direct estimates","Model 2"),col=c("blue","red"), pch=c(15,18))

# Figure 8: MSE of Direct estimates and MSE of EBLUPs (under Model 2) ordered by sample size.
par(mfrow=c(1,3))

plot(result_eblup$id, result_eblup$v11, type = "o", col = "blue", pch = 15, xlab = "area",
    ylab = "p0", cex.axis = 1.5, cex.lab = 1.5, xaxt = "n")
points(result_eblup$p0.mse, type = "o", col = "red", pch = 18)
axis(1, at=result_eblup$id, labels = result_eblup$prov)
legend("topright",legend=c("Direct estimates","Model 2"),col=c("blue","red"), pch=c(15,18))
plot(result_eblup$id, result_eblup$v22, type = "o", col = "blue", pch = 15, xlab = "area",
    ylab = "p1", cex.axis = 1.5, cex.lab = 1.5, xaxt = "n")
points(result_eblup$p1.mse, type = "o", col = "red", pch = 18)
```

```
axis(1, at=result_eblup$id, labels = result_eblup$prov)
legend("topright",legend=c("Direct estimates","Model 2"),col=c("blue","red"), pch=c(15,18))
plot(result_eblup$id, result_eblup$v33, type = "o", col = "blue", pch = 15, xlab = "area",
    ylab = "p2", cex.axis = 1.5, cex.lab = 1.5, xaxt = "n")
points(result_eblup$p2.mse, type = "o", col = "red", pch = 18)
axis(1, at=result_eblup$id, labels = result_eblup$prov)
legend("topright",legend=c("Direct estimates","Model 2"),col=c("blue","red"), pch=c(15,18))
```

Variable of Interest	Statistic	Direct Estimation	Model 2
p0	Min	0.05244	0.04601
	Quartil 1	0.17296	0.18947
	Median	0.21850	0.21937
	Mean	0.22659	0.22020
	Quartil 3	0.27753	0.25468
	Max	0.43588	0.35011
	Standard Deviation	0.08173	0.056597
p1	Min	0.01891	0.02551
	Quartil 1	0.05336	0.05969
	Median	0.06810	0.06866
	Mean	0.07567	0.07407
	Quartil 3	0.09208	0.08866
	Max	0.15973	0.14023
	Standard Deviation	0.03227	0.02506
p2	Min	0.005449	0.008481
	Quartil 1	0.025819	0.026867
	Median	0.032344	0.033600
	Mean	0.039042	0.038179
	Quartil 3	0.051379	0.049755
	Max	0.099308	0.087194
	Standard Deviation	0.02083	0.017137

Table 1: Statistics of direct estimation and Model 2

We will compare EBLUPs based on Model 2 with the direct estimates. Both of the estimation results can be seen in Table 1. On the median value, the result of estimated poverty indicators are relatively similar, ranging from 0.218-0.219 for p0, 0.0681-0.0687 for p1, and 0.0323-0.0336 for p2. Model 2 has a lower range and smaller standard deviation than the direct estimation. It means that, in general, the multivariate Fay-Herriot model has lower variability of small area estimates than direct estimation.

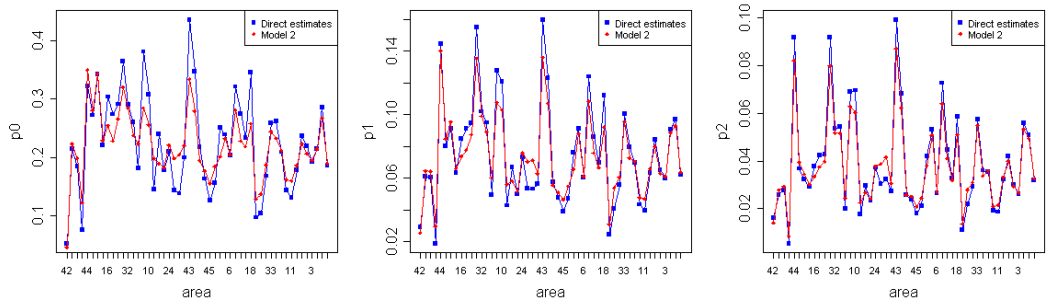

Figure 7: Direct estimates and EBLUPs (under Model 2) of p0, p1, and p2 ordered by sample size.

The results of direct estimates and EBLUPs under Model 2 ordered by sample size are shown in Figure 7. The patterns of small area estimates for both methods are almost the same for all areas. The MSEs of the direct estimates and the EBLUP estimates ordered by sample size are shown in Figure 8. These plots show that the multivariate Fay-Herriot model has lower MSE than the direct estimation. Thus, we can conclude that the multivariate Fay-Herriot model is more efficient than direct estimation.

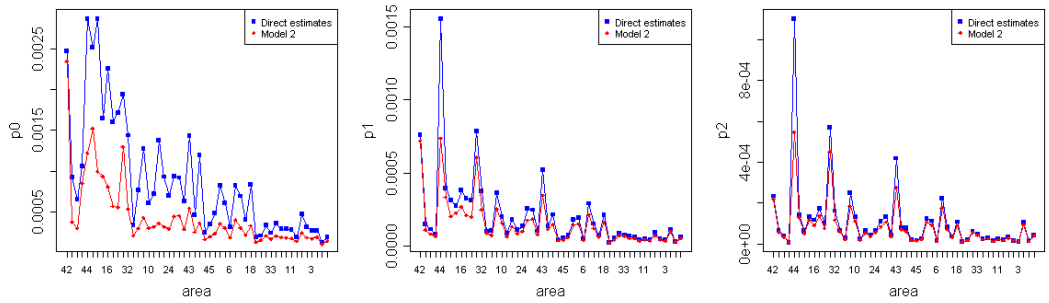

Figure 8: MSE of direct estimates and MSE of EBLUPs (under Model 2) of p0, p1, and p2 ordered by sample size.

Conclusion

This paper introduces the first R package of multivariate Fay-Herriot model for small area estimation named **msae**. The package is available on Comprehensive R Archive Network (CRAN) at http://CRAN.R-project.org/package=msae. This package contains a number of functions for estimating the EBLUP and MSE of EBLUP of each Fay-Herriot Model. This package accommodates the univariate Fay-Herriot model (model 0), multivariate Fay-Herriot model (model 1), autoregressive multivariate Fay-Herriot model (model 2), and heteroskedastic autoregressive multivariate Fay-Herriot model (model 3). The functions are described and implemented using three examples of datasets provided in **msae** package and a real dataset provided in **sae** package, `incomedata`. By these examples, we show that the multivariate Fay-herriot models produce more efficient parameter estimates than direct estimation and univariate model.

Bibliography

R. Benavent and D. Morales. Multivariate fay–herriot models for small area estimation. *Computational Statistics and Data Analysis*, page 372–390, 2016. URL https://doi.org/10.1016/j.csda.2015.07.013. [p111, 112, 113, 118]

H. J. Boonstra. *hbsae: Hierarchical Bayesian Small Area Estimation*, 2012. URL https://CRAN.R-project.org/package=hbsae. R package version 1.0. [p111]

J. Breidenbach. *JoSAE: Unit-Level and Area-Level Small Area Estimation*, 2018. URL https://CRAN.R-project.org/package=JoSAE. R package version 0.3.0. [p111]

Chengchun Shi Developer. *BayesSAE: Bayesian Analysis of Small Area Estimation*, 2018. URL https://CRAN.R-project.org/package=BayesSAE. R package version 1.0-2. [p111]

G. S. Datta, R. E. Fay, and M. Ghosh. Hierarchical and empirical multivariate bayes analysis in small area estimation. In *Proc. of the Bureau of the Census Annual Research Conference*, pages 63–79. Bureau of the Census, Washington D.C., 1991. [p111]

M. Ghosh and J. N. K. Rao. Small area estimation: An appraisal. *Statistical Science*, pages 55–93, 1994. [p111]

S. Krieg, H. J. Boonstra, and M. Smeets. Small area estimation with zero-inflated data - a simulation study. *Statistics Netherland*, pages 1–45, 2015. [p111]

M. D. E. Lefler, D. M. Gonzalez, and A. P. Martin. *saery: Small Area Estimation for Rao and Yu Model*, 2014. URL https://CRAN.R-project.org/package=saery. R package version 1.0. [p111]

E. Lopez-Vizcaino, M. Lombardia, and D. Morales. *mme: Multinomial Mixed Effects Models*, 2019. URL https://CRAN.R-project.org/package=mme. R package version 0.1-6. [p111]

I. Molina and Y. Marhuenda. sae: An r package for small area estimation. *The R Journal*, pages 81–98, 2015. URL https://doi.org/10.32614/RJ-2015-007. [p118]

I. Molina and Y. Marhuenda. *sae: Small Area Estimation*, 2018. URL https://CRAN.R-project.org/package=sae. R package version 1.2. [p111]

J. Pinheiro, D. Bates, and R-core. *nlme: Linear and Nonlinear Mixed Effects Models*, 2020. URL https://CRAN.R-project.org/package=nlme. R package version 3.1-145. [p111]

J. N. K. Rao and I. Molina. *Small Area Estimation 2nd Edition*. John Wiley and Sons Inc., Hoboken, New Jersey, 2015. [p111]

T. Schoch. *rsae: Robust Small Area Estimation*, 2014. URL https://CRAN.R-project.org/package=rsae. R package version 0.1-5. [p111]

M. Templ. *CRAN Task View: Official Statistics & Survey Methodology*, 2014. URL https://CRAN.R-project.org/view=OfficialStatistics. Version 2014-08-18. [p111]

A. Ubaidillah, K. A. Notodiputro, A. Kurnia, and I. W. Mangku. Multivariate fay-herriot models for small area estimation with application to household consumption per capita expenditure in indonesia. *Journal of Applied Statistics*, pages 2845–2861, 2019. URL https://doi.org/10.1080/02664763.2019.1615420. [p111]

Novia Permatasari
Politeknik Statistika STIS
East Jakarta
Indonesia
16.9335@stis.ac.id

Azka Ubaidillah
Politeknik Statistika STIS
East Jakarta
Indonesia
azka@stis.ac.id