THE 1st INTERNATIONAL CONFERENCE ON DATA SCIENCE AND OFFICIAL STATISTICS
ICDSOS
"Strengthening the Contribution of Data Science and Official Statistics to the Society in the Distruption Era"
2021

Estimation of Education Indicators in East Java Using Multivariate Fay-Herriot Model

N Permatasari[1], **A Ubaidillah**[2]

[1] Statistics Indonesia, Jakarta, Indonesia
[2] Polytechnic of Statistics STIS, Jakarta, Indonesia

*Corresponding author's e-mail : novia.permatasari@bps.go.id

Abstract. Education is an important aspect in improving human resources. Data availability of education indicators in a low administrative level is needed as a basis for education planning in that region. The problem of sample size when provide a low administrative level data can be overcome by indirect estimation, namely Small Area Estimation (SAE). SAE is able to increase the effectiveness of the survey sample size by using the strength of neighbouring areas and information from auxiliary variables related to the variables of interest. We obtain simulation study to compare multivariate model to univariate model and implement multivariate model to estimate three education indicators which are obtained from the National Socio-Economic Surveys by Statistics Indonesia. Simulation results are in line with previous studies, where the multivariate Fay-Herriot model with *p* variable has smaller of mean squares error (MSE) than the univariate model. The model implementation to estimate Crude Participation Rate (APK), School Participation Rate (APS), and Pure Participation Rate (APM) also shows that the multivariate model produces smaller RRMSE than the direct estimates. It can be concluded that multivariate model is able to produce more efficient estimates than direct estimation and univariate model.

1. Introduction

Education is an important aspect in improving human resources. The Statistics Indonesia (BPS) produces three education indicators to see the development of education sector in Indonesia, namely Crude Participation Rate (APK), School Participation Rate (APS), and Pure Participation Rate (APM). These three indicators are calculated from the National Socio-Economic Survey (Susenas), which is held every semester. Data availability of these indicators in a low administrative level is needed as a basis for education planning in that region. Unfortunately, based on the survey design, data availability in the second semester is limited to the provincial level. The sample size that are only sufficient for provincial level estimates will result in large relative standard errors and unreliable direct estimates at the district level. This problem can be overcome by indirect estimation, namely Small Area Estimation (SAE).

According to [1], SAE is able to increase the effectiveness of the survey sample size by using the strength of neighboring areas and information on auxiliary variables related to the variable of interest. From the availability of explanatory variables, area-based models are more widely used than unit-based models [2]. Small area estimation using the Empirical Best Linear Unbiased Predictor (EBLUP) method was initiated by [3] to estimate the logarithm of income per capita in the United States, so this model is known as the Fay-Herriot Model.

In general, there are many variables that have strong correlation. By using SAE, these variables can be estimated together using the multivariate SAE method. Research by [4] [5] [6] and [7] showed that the multivariate SAE model has more efficient estimation than the univariate SAE model by utilizing the correlation between the variables of interest. [5] developed Fay-Herriot Model into four different estimation models based on their covariance matrix structure, namely the univariate FH model (Model 0), the multivariate FH model (Model 1), the autoregressive multivariate FH model (Model 2), and heteroscedastics autoregressive multivariate FH model (Model 3).

Until now, small area estimation of education indicators has been studied by [9] and [10] with many models. [9] implemented the EBLUP and EBLUP Benchmarking methods to estimate the Crude Participation Rate at the university level on Kalimantan and [10] applied on Papua. These two studies still used univariate model. Research on the multivariate FH model by [6] and [7] uses the multivariate FH model (Model 1) to estimate the average household expenditure per capita of food and non-food in Central Java, Indonesia. [5] used model 2 to estimate the proportion of poverty in 2005 and 2006, and model 3 to estimate the proportion and depth of poverty at provincial level in Spain.

Various studies on the Fay Herriot model that have been carried out are still limited to one or two variables of interest. In this study, we will conduct simulations and case studies of multivariate models for more than two variables. Simulations were carried out to see the efficiency of the multivariate model compared to the univariate model. Then, we will apply the model to three indicators of education in Indonesia, namely Crude Participation Rate (APK), School Participation Rate (APS), and Pure Participation Rate (APM) at the university level in East Java, Indonesia.

2. Methodology

Data Description

The data used in this study are three education indicators: Crude Participation Rate (APK), School Participation Rate (APS), and Pure Participation Rate (APM) in university level as target variables. Those variables can be calculated by following formula:

$$APK = \frac{number\ of\ university\ student}{number\ of\ people\ at\ 19-24\ years\ old}\ x\ 100\%$$

$$APS = \frac{number\ of\ student\ at\ 19-24\ years\ old}{number\ of\ people\ at\ 19-24\ years\ old}\ x\ 100\%$$

$$APM = \frac{number\ of\ university\ student\ at\ 19-24\ years\ old}{number\ of\ people\ at\ 19-24\ years\ old}\ x\ 100\%$$

The APK, APS, and APM variables are obtained from National Socio-Economic Surveys by Statistics Indonesia which is held every semester. In 2^{nd} semester, this survey was designed for the provincial level, so it was not able to produce reliable estimates for district level.

We use administrative data sourced as auxiliary variables. Those auxiliary variables are: number of family use liquid waste disposal pit, number of family use source of drinking water metered pipe, number of family use source of bathing water washing metered tap, number of private universities, and number of hospital.

Multivariate Fay-Herriot Model

The multivariate Fay-Herriot model is a development of the univariate Fay-Herriot Model by utilizing the correlation between the variables of interest. In matrix, the multivariate Fay Herriot model can be written as follows:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e} = \boldsymbol{X\beta} + \boldsymbol{Z_1 u_1} + \cdots + \boldsymbol{Z_D u_D} + \boldsymbol{e} \quad (1)$$

where $\boldsymbol{y}$ is the direct estimates, $\boldsymbol{X}$ is the matrix of the auxiliary variables, $\boldsymbol{\beta}$ is regression coefficient, $\boldsymbol{Z}$ is the identity matrix, $\boldsymbol{u}$ is area random effect, and $\boldsymbol{e}$ is sampling error. The components $\boldsymbol{e}, \boldsymbol{u_1}, \dots \boldsymbol{u_D}$ are independent with distributions:

$$\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{V_e}), \quad \boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{V_u}), \quad and \quad \boldsymbol{u_d} \sim N(\boldsymbol{0}, \boldsymbol{V_{ud}}) \tag{2}$$

Fay-Herriot model was developed into four models based on the structure of the covariance matrix, namely the univariate FH model (Model 0), the multivariate FH model (Model 1), the autoregressive multivariate FH model (Model 2), and the heteroscedastic autoregressive multivariate FH model (Model 3) [5]. Model 0 is a univariate Fay Herriot model with more than one observation variable. The sampling error covariance matrix in model 0 is a diagonal matrix which shows that there is no correlation between the variables of interest. The sampling error covariance matrix, $\boldsymbol{V_{ed}}$, and random effect covariance matrix, $\boldsymbol{V_{ud}}$, of Model 0 are as follows:

$$\boldsymbol{V_{ed}} = diag_{1 \le r \le R}\left(\boldsymbol{\sigma^2_{edr}}\right), \; d = 1, \dots, D \tag{3}$$

$$\boldsymbol{V_{ud}} = diag_{1 \le r \le R}\left(\boldsymbol{\sigma^2_{udr}}\right), \; d = 1, \dots, D \tag{4}$$

Model 1, Model 2, and Model 3 are multivariate Fay Herriot models where the sampling error covariance matrix is not a diagonal matrix. Model 1 is the simplest multivariate model, which is an extension of Model 0, where the correlation of random effects is ignored.

Model 2 is called the autoregressive multivariate Fay Herriot model (AR(1)), where the value of the random effect covariance matrix is as follows:

$$\boldsymbol{V_{ud}} = \sigma_u^2 \, \boldsymbol{\Omega_d}(\boldsymbol{\rho}) \tag{5}$$

$$\boldsymbol{\Omega_d}(\boldsymbol{\rho}) = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & \rho & \dots & \rho^{R-1} \\ \rho & 1 & & \rho^{R-2} \\ \vdots & & & \vdots \\ \rho^{R-1} & \rho^{R-2} & \dots & 1 \end{pmatrix} \tag{6}$$

Model 3 is called the heteroscedastic autoregressive multivariate Fay Herriot model (HAR(1)), where the elements of the random effect are as follows:

$$u_{dr} = \rho u_{dr-1} + a_{dr} \tag{7}$$

$$u_{d0} \sim N(0, \sigma_0^2) \qquad \sigma_0{}^2 = 1 \qquad a_{d0} \sim N(0, \sigma_r^2) \tag{8}$$

with $a_{dr,} u_{d0}, and\ \sigma_0{}^2$ are independent. The elements of the random effect covariance matrix of Model 3 are as follows:

$$\sigma_{drii} = \textstyle\sum_{k=0}^{i} \rho^{2k} \, \sigma^2{}_{i-k} \tag{9}$$

$$\sigma_{drij} = \textstyle\sum_{k=0}^{|i-j|} \rho^{2k+|i-j|} \, \sigma^2_{|i-j|-k} \;, i \neq j \tag{10}$$

The multivariate Fay-Herriot model was estimated using Empirical Best Linear Unbias Predictor (EBLUP) with the following formula:

$$\boldsymbol{\hat{\mu}_E} = \boldsymbol{X\hat{\beta}_E} + \boldsymbol{Z\widehat{V_u}Z^T\hat{V}^{-1}}\left(\boldsymbol{y} - \boldsymbol{X\hat{\beta}_E}\right) \tag{11}$$

$$\boldsymbol{\hat{V}} = \boldsymbol{Z\widehat{V_u}Z^T} + \boldsymbol{V_e} \tag{12}$$

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

MSE of Multivariate fay-Herriot Model

[5] estimate the MSE value in the multivariate Fay-Herriot model using the method developed by [8] through the following equation:

$$mse(\widehat{\boldsymbol{\mu}}) = \boldsymbol{g}_{1i}(\widehat{\boldsymbol{\theta}}) + \boldsymbol{g}_{2i}(\widehat{\boldsymbol{\theta}}) + 2\boldsymbol{g}_{3i}(\widehat{\boldsymbol{\theta}}) \quad (13)$$

with each component can be described using the following formula:

$$\boldsymbol{g}_{1i}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\Gamma}\boldsymbol{V}_{e} \quad (14)$$

$$\boldsymbol{g}_{2i}(\widehat{\boldsymbol{\theta}}) = (\boldsymbol{I}-\boldsymbol{\Gamma})\boldsymbol{X}\left(\boldsymbol{X}^{T}\widehat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}\,(\mathbf{1}-\boldsymbol{\Gamma})^{T} \quad (15)$$

$$\boldsymbol{g}_{3i}(\widehat{\boldsymbol{\theta}}) \approx \sum\sum cov\left(\hat{\theta}_{k},\hat{\theta}_{l}\right)\boldsymbol{\Gamma}_{(k)}\widehat{\boldsymbol{V}}\boldsymbol{\Gamma}^{T}{}_{(k)},\ k,l = 1,\dots,q \quad (16)$$

where $\boldsymbol{\Gamma} = \mathbf{Z}(\widehat{\boldsymbol{V}}_{\boldsymbol{u}})\mathbf{Z}^{T}\widehat{\boldsymbol{V}}^{-1}$, $\boldsymbol{\Gamma}_{(\boldsymbol{k})} = \frac{\partial\boldsymbol{\Gamma}}{\partial\boldsymbol{\theta}_{\boldsymbol{k}}}$, and $\boldsymbol{cov}\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{k}},\widehat{\boldsymbol{\theta}}_{\boldsymbol{l}}\right) = \left(\boldsymbol{F}_{\boldsymbol{a},\boldsymbol{b}}\right)^{-\mathbf{1}}$.

Relative Root Mean Square Error

In the evaluation, we use Relative Root Mean Square Error **(**RRMSE) to compare the estimates from several estimation methods. It aims to obtain standardized comparison because it has removed the units of the results. RRMSE can be calculated by the following formula:

$$RRMSE = \frac{\sqrt{mse(\widehat{\boldsymbol{\mu}})}}{\widehat{\boldsymbol{\mu}}}$$

where $mse(\widehat{\boldsymbol{\mu}})$ is the MSE of EBLUP of multivariate FH model and $\widehat{\boldsymbol{\mu}}$ *is* the EBLUP.

3. Result and Discussion

3.1. Simulation Study

Simulation of the multivariate Fay Herriot model with 3 observation variables was carried out for domains D = 30, 50, 100, and 200. Simulations were carried out for all possible combinations of models in generating data and their correlation parameters (ρ,ρ_{e}). Simulation data is generated by the following steps:

1. Generate $\left\{e_{dr}^{(b)}, u_{dr}^{(b)},\ x_{dr}^{(b)}\right\}$; $d = 1,\dots,D$; $r = 1,\dots,R,$; $b = 1,\dots,B$.
2. Create the formulas. Simulations were carried out for multivariate data with three observed variables, namely $y1$, y2, and $y3$. Each dependent variable is influenced by three participating variables, namely $x1$, $x2$, and $x3$.

$$f1 = y1 \sim x1 + x2 + x3$$

$$f2 = y2 \sim x1 + x2 + x3$$

$$f3 = y3 \sim x1 + x2 + x3$$

3. Calculate the value of the target parameter, $\mu_{dr}^{(b)} = x_{dr}^{T}\beta + u_{dr}^{(b)}$.
4. Calculate the value of EBLUP, $\hat{\mu}_{dr}^{(b)}$.
5. Calculate the MSE of each model.
6. Repeat steps 1 to 5 for $B = 100$ repetitions.

In this study, we set number of domain $D = 30, 50, 100,$ and 200 observations and number of variabel of interest $R = 3$ variables. We take $\sigma_{U11} = 2,\ \sigma_{U22} = 3,$ and $\sigma_{U33} = 4$ for area random effects; $\sigma_{e11} = 1,\ \sigma_{e22} = 2,$ and $\sigma_{e33} = 3$ for sampling errors, and $\beta_{1} = \beta_{2} = 1$ for the regression coefficients. The values of ρ_{e} and ρ are defined as 0 or 0.5 for each simulation. For the auxiliary variables, X_{1}, X_{2} and X_{3}, are generated with the following distributions: $X_{1} \sim N(10,1)$, $X_{2} \sim N(10,2)$, and $X_{3} \sim UNIF(9,12)$. These initial parameter consider to the simulation on [5] and [6].

Table 1. MSE of Multivariate Fay-Herriot Model Simulation

Variable	Simulation	$\rho(e)$	ρ	k	a	Number of Domains			
						30	50	100	200
Y1	1	-	-	0	0	0.754	0.696	0.685	0.668
					3	0.755	0.707	0.688	0.668
	2	0.5	-	1	0	0.723	0.714	0.678	0.677
					1	0.658	0.642	0.603	0.599
	3	0.5	0	2	0	0.729	0.678	0.69	0.68
					2	0.644	0.595	0.582	0.573
	4	0	0.5	2	0	0.756	0.775	0.746	0.732
					2	0.727	0.744	0.715	0.698
	5	0.5	0.5	2	0	0.909	0.774	0.818	0.802
					2	0.868	0.738	0.775	0.766
	6	0.5	0	3	0	0.733	0.741	0.683	0.669
					3	0.685	0.67	0.612	0.591
	7	0	0.5	3	0	0.764	0.722	0.722	0.699
					3	0.753	0.705	0.693	0.671
	8	0.5	0.5	3	0	0.75	0.712	0.707	0.692
					3	0.737	0.708	0.696	0.675
Y2	1	-	-	0	0	1.386	1.315	1.257	1.198
					3	1.412	1.331	1.262	1.202
	2	0.5	-	1	0	1.405	1.305	1.243	1.218
					1	1.305	1.188	1.114	1.103
	3	0.5	0	2	0	1.19	1.152	1.059	1.02
					2	1.041	1.025	0.932	0.92
	4	0	0.5	2	0	1.298	1.229	1.199	1.178
					2	1.126	1.085	1.038	1.028
	5	0.5	0.5	2	0	1.45	1.416	1.374	1.346
					2	1.408	1.392	1.345	1.328
	6	0.5	0	3	0	1.355	1.251	1.224	1.197
					3	1.296	1.158	1.123	1.089
	7	0	0.5	3	0	1.428	1.369	1.337	1.31
					3	1.405	1.313	1.276	1.245
	8	0.5	0.5	3	0	1.407	1.325	1.304	1.312
					3	1.412	1.308	1.296	1.295
Y3	1	-	-	0	0	2.022	1.852	1.77	1.755
					3	2.048	1.862	1.777	1.756
	2	0.5	-	1	0	1.981	1.879	1.821	1.725
					1	1.829	1.705	1.643	1.568
	3	0.5	0	2	0	1.538	1.42	1.277	1.261
					2	1.368	1.286	1.174	1.159
	4	0	0.5	2	0	1.68	1.594	1.523	1.467
					2	1.517	1.458	1.397	1.347
	5	0.5	0.5	2	0	2.007	1.821	1.781	1.752
					2	1.961	1.78	1.76	1.735
	6	0.5	0	3	0	2.03	1.862	1.779	1.743
					3	1.935	1.725	1.63	1.598
	7	0	0.5	3	0	2.09	1.986	1.943	1.903
					3	2.091	1.974	1.898	1.854
	8	0.5	0.5	3	0	2.145	1.898	1.939	1.884
					3	2.148	1.88	1.925	1.861

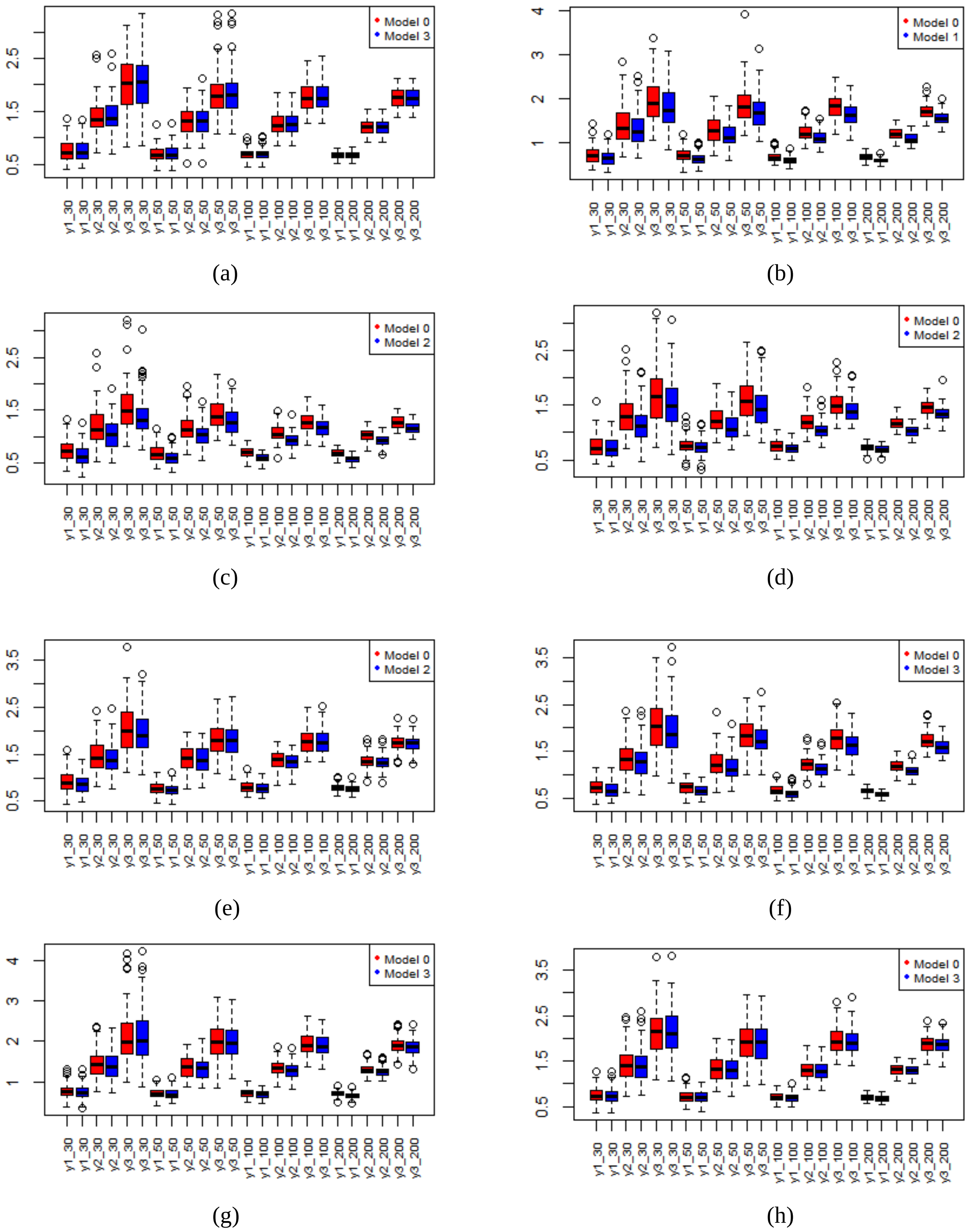


Figure 1. Boxplots of MSE values from Y1, Y2, and Y3 for both models in the simulation study. (a) Boxplots for simulation 1, (b) Boxplots for simulation 2, (c) Boxplots for simulation 3, (d) Boxplots for simulation 4, (e) Boxplots for simulation 5, (f) Boxplots for simulation 6, (g) Boxplots for simulation 7, and (h) Boxplots for simulation 8.

For multivariate FH model (Model 1), autoregressive multivariate FH model (Model 2), and heteroscedastics autoregressive multivariate FH model (Model 3) the value of MSE of EBLUP will be compared with the MSE of model 0, which is a univariate model. While for Model 0, the MSE will be compared with value MSE EBLUP of Model 3.

Table 1 shows that if the data is generated using univariate model, model 0 will produce MSE smaller than multivariate model. The simulation results from 2 to 6 show that the multivariate EBLUP will have a lower MSE than the univariate model (Model 0). In simulations 7 and 8, it can be seen that the MSE in one of the variables in Model 0 is smaller than Model 3 for small domain (domain = 30). However, this MSE will decrease as the number of domains increases. The MSE of Model 3 will be smaller than Model 0 when the number of domains is more than or equal to 50. It indicates that for small domain sizes, a simple model will be more useful than a complex model.

Figure 1 also shows that the MSEs of EBLUP based on Multivariate FH model are smaller that the Univariate model. It means that multivariate model is able to produce more efficient estimates than the univariate model. Simulations that carried out with four different domains showed an inverse relationship between the number of domains and the MSE. The simulation results show that the larger of domains will produce the smaller and more consistent of MSE estimation.

3.2. Application to Real Data

The multivariate Fay Herriot model will be implemented to estimate three education indicators in East Java from National Socio-Economic Surveys in September 2018 at the district level. The three indicators include Crude Participation Rate (APK), School Participation Rate (APS), and Pure Participation Rate (APM) at the university level. The correlation between the three variables that is greater than 0.8 indicates a strong correlation of the variables, so it is appropriate to be estimated using a multivariate model. The auxiliary variables used are administrative data sourced (Podes) in 2018.

The selection of the auxiliary variables is carried out on each variable of interest. The auxiliary variables used were shown in table 2.

Table 2. Variable of Interest and the Auxiliary Variables

Variables of Interest	Variables
Crude Participation Rate	Number of family use liquid waste disposal pit
	Number of family use source of drinking water metered pipe
	Number of family use source of bathing water washing metered tap
	Number of private universities
	Number of hospital health facilities
Pure Participation Rate	Number of family use source of drinking water metered pipe
	Number of private universities
School Participation Rates	Number of family use source of drinking water metered pipe
	Number of hospital

Selection of the best multivariate model is aimed to determine the model that fits the data. This process is carried out by testing the homogeneity of variance followed by ρ parameter testing. In the homogeneity of variance test, we used Model 3 and obtained a p-value more than 5% significance level. This shows that there is no significant difference between the random effect variance on the data. In the ρ parameter test, we conducted modelling based on model 2 and obtained a p-value of 7.1507e-13. This shows that there is a correlation between random effects. These results indicate that the best model that fits the data is Model 2 or autoregressive multivariate FH model. The random effect variance using Model 2 is 0.0013169. It shows that there are random effects on the estimation results. Therefore, modelling with SAE is feasible.

In this case study, a comparison was provided between the results of the direct estimates and the multivariate model. The estimation summary of the two methods shown in table 3. From the median, we know that the estimation of the education indicators in the two methods are relatively the same. Meanwhile, based on the size of the variability, the estimation using the multivariate Fay Herriot model have a lower range than the direct estimates. In line with the range, standard deviation of the multivariate Fay Herriot model is also lower than the direct estimates. In general, the multivariate Fay Herriot model has less variability than the direct estimates.

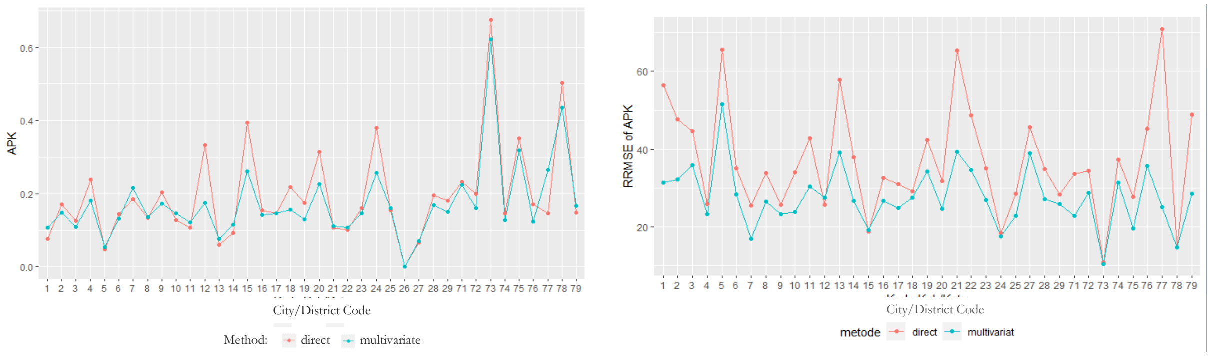
Table 3. Descriptive statistics of the estimation of APK, APM, and APS

Variable	Statistics	Direct Estimates	Model 2
APK	Minimum	0.000	0.000
	Quartile 1	0.126	0.122
	Median	0.157	0.147
	Mean	0.194	0.173
	Quartile 3	0.215	0.180
	Maximum	0675	0621
	standard deviation	0131	0107
APM	Minimum	0.000	0.000
	Quartile 1	0.076	0.085
	Median	0.124	0.122
	Mean	0141	0133
	Quartile 3	0.180	0.147
	Maksmum	0518	0469
	standard deviation	0103	0082
APS	Minimum	0.039	0.065
	Quartile 1	0144	0138
	Median	0.187	0.180
	Mean	0218	0194
	Quartile 3	0266	0214
	Max	0.555	0.506
	Standard deviation	0.108	0.087

Table 4. The results of an autoregressive multivariate model (Model 2)

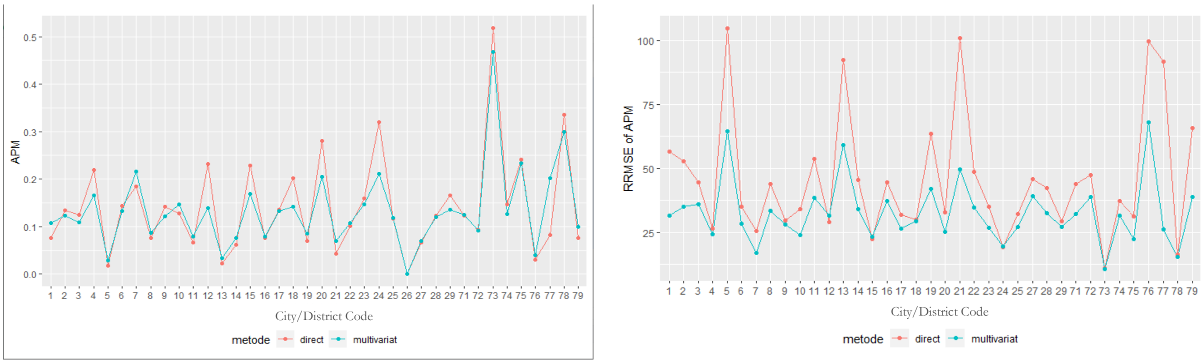
Variables of Interest	Auxiliary Variables	Beta	Standard Error	t-statistics	p-value
APK	(Intercept)	0.089	0.029	3.113	0.002
	Number of family use liquid waste disposal pit	-0.079	0.035	-2.234	0.025
	Number of family use source of drinking water metered pipe	0.314	0.118	2.651	0.008
	Number of family use source of bathing water washing metered tap	-0.067	0.084	-0.790	0.429
	Number of private universities	0.007	0.002	4.022	0.000
	Number of hospital health facilities	-0.001	0.001	-0.409	0.683
APM	(Intercept)	0.041	0.017	2.440	0.015
	Number of family use source of drinking water metered pipe	0.247	0.085	2.913	0.004
	Number of private universities	0.005	0.001	4.088	0.000
APS	(Intercept)	0.090	0.018	5.090	0.000
	Number of family use source of drinking water metered pipe	0.210	0.091	2.302	0.021
	Number of hospital	0.006	0.001	4.766	0.000

Table 4 shows the results of modelling using the autoregressive multivariate Fay-Herriot model (Model 2). It can be seen that almost all auxiliary variables have p-values less than the significance level (5%), which indicates that the variable has a significant effect on the corresponding response variable. Variables that are significant to the APK include number of family use liquid waste disposal pit, number of family use source of bathing water washing metered tap, and number of private universities. Variables that are significant to the APM include: Number of family use source of drinking water metered tap and the number of private universities. While the variables that are significant to the APS include: number of family use source of drinking water metered tap and the number of hospitals.



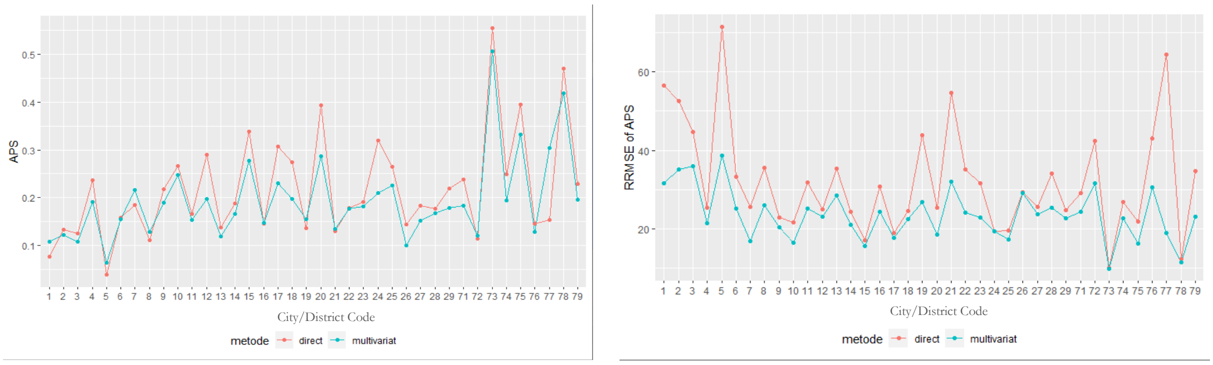
(a) (b)

Figure 2. Comparison of direct estimation and Model 2 for APK variables. (a) Comparison of direct estimates and the EBLUPs. (b) Comparison of RRMSE of the direct estimates and EBLUPs.



(a) (b)

Figure 3. Comparison of direct estimation and Model 2 for APM variables. (a) Comparison of direct estimates and the EBLUPs. (b) Comparison of RRMSE of the direct estimates and EBLUPs.



(a) (b)

Figure 4. Comparison of direct estimation and Model 2 for APS variables. (a) Comparison of direct estimates and the EBLUPs. (b) Comparison of RRMSE of the direct estimates and EBLUPs.

Based on figure 2 to figure 4, the results of the two methods gave almost the same value for each district in East Java. Districts with high estimates in one method will also have high estimates in other models. For APK and APM variables, the district with the highest estimates is Malang city and the lowest is Bangkalan district. For the APS variable, the highest estimates is Malang city and the lowest is Blitar district.

Table 5. Average RRMSE of direct estimates and EBLUP using Model 2.

Variable	RRMSE	
	direct estimates	EBLUP using Model 2
(1)	(2)	(3)
APK	0.3713	0.2772
APM	0.4582	0.3289
APS	0.3225	0.23603

In comparing the estimation results of the two methods, researchers used RRMSE. The average RRMSE from the direct estimates and EBLUPs using Model 2 for each variable can be seen in table 5. Based on table 5, it can be seen that the average value RRMSE of EBLUPs Model 2 is smaller than RRMSE of direct estimates for all variables. In APK variable, EBLUP using Model 2 was able to reduce the RRMSE from 37% to 27%. In the APM variable, EBLUP using Model 2 is able to reduce the RRMSE from 45% to 32%. While the APM variable, EBLUP using Model 2 is able to reduce the RRMSE from 32% to 23%.

For each domain in Figures 2 to 4, the RRMSE of EBLUP using Model 2 will be smaller than the RRMSE of direct estimates in almost all domains. In the APK variable, EBLUP using Model 2 is not able to decrease the RRMSE in Situbondo district and Sidoarjo district. In the APM variable, EBLUP of Model 2 was not able to decrease the RRMSE in three districts, which are Situbondo, Sidoarjo, and Lamongan. While on the APS, EBLUP using Model 2 has been able to decrease the RRMSE in all domains. The differences in RRMSE in the previously mentioned domains are not more than 3%, which is much smaller than the average of RRMSE decrease in other domains.

In general, the estimation results using Model 2 have a lower RRMSE than the direct estimation for the three variables. This shows that the multivariate Fay-Herriot model is able to produce more efficient estimation than direct estimation.

4. Conclusion and Recommendation

Simulation results shows the multivariate models, which are multivariate FH model, autoregressive multivariate FH model, and heteroscedastics autoregressive multivariate FH model, is able to produce more efficient estimates than the univariate model. It is can be seen from the smaller MSE of multivariate model than the univariate model. It is also shown that the larger of domain size, the smaller and more consistent MSE resulted. The implementation of multivariate Fay-Herriot model to estimate the APK, APM, and APS values also shows that the multivariate FH model is able to produce more efficient estimation than the direct estimates. These results are in line with previous studies ablout multivariate Fay Herrot Model by [5], [6], and [7].

References

[1] Rao J and Molina I 2015 *Small Area Estimation 2nd Edition* (New Jersey: John Wiley and Sons Inc.)

[2] Dedianto D and Wulansari I Y 2019 Aplikasi Small Area Estimation (SAE) metode pseudo-EBLUP pada official statistik di Indonesia *Jurnal Aplikasi Statistika & Komputasi Statistik* **10** 33-38

[3] Fay R E and Herriot R A 1979 Estimates of income for small places: An application of James-Sten procedures to cencus data *Journal of the American Statistical Association* **74** 268-77

[4] Datta G, Fay R E, and Ghosh M 1991 Hierarchical and Empirical Multivariate Bayes Analysis

in Small Area Estimation *Bureau of the Census Annual Research Conference* 63-79 (Washington DC: Bureau of the Census, Washington DC)

[5] Benavent R and Morales D 2016 Multivariate Fay – Herriot models for small area estimation *Computational Statistics and Data Analysis* **94** 372-390

[6] Ubaidillah A, Notodiputro K A, Kurnia A, and Mangku I W 2019 Multivariate Fay-Herriot models for small area estimation with application to household consumption per capita expenditure in Indonesia *Journal of Applied Statistics* **46** 2845-61.

[7] Nurizza W A 2018 *Application of the Multivariate Fay-Herriot Model* [Thesis]. (Jakarta: Polytechnic of Statistics STIS)

[8] Prasad N G N and Rao J N K 1990 The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association* **85** 163–171.

[9] Ikhsan E, Ratu Y R, and Nurizza W A 2019 Small Area Estimation of Gross Enrollment Rates in universities in Kalimantan Island in 2018. *Proceedings of the National Seminar on Mathematics and Statistics* **1** 82-88.

[10] Subandriyo B, Ikhsan E, and Muchlishoh S 2019 Estimating the gross participation rate of Papua province universities through small area estimation. *National Seminar on Official Statistics* **2019** 104-109.