Forum Statistika dan Komputasi : *Indonesian Journal of Statistics*
ISSN : 0853-8115
Vol. 20 No.7, Oktober 2015, p: 34-38
available online at:
journal.ipb.ac.id/index.php/statistika

EBLUP METHOD OF TIME SERIES AND CROSS-SECTION DATA FOR ESTIMATING EDUCATION INDEX IN DISTRICT PURWAKARTA

Febriyani Eka Supriatin[1], Budi Susetyo[2], Kusman Sadik[2]

[1,2] Department of Statistics, Bogor Agricultural University, Indonesia
E-mail : febriyaniekas@yahool.co.id[1]

ABSTRACT

Since decentralisation was implemented in Indonesia, more detailed information about the condition of an area becomes very necessary to know as an evaluation of development that the government has done. the success development of a region can be seen through the Human Development Index (HDI). HDI consists of three basic dimensions, knowledge as one of that three basic measured by the index of education. This index is measured by the Adult Literacy Rate and Mean Years of Schooling. Education is one of the important factors in improving human development. The enhancement of education index results in increasing the HDI of an area. Purwakarta has a vision that is made as a district that excels in education in West Java, but until now Purwakarta's education index is still below the West Java province. One step that can be done is to seek information on the education index each district in Purwakarta, with the aim to provide the right policy in each region. Direct estimation of the components forming the HDI for districts is not feasible because these estimates will generate a great value of variance, This is due to the size of the sample used is too small. This study proposes a statistical method by performing the estimation using small area estimation. These estimates using information from surrounding areas that can improve the effectiveness of the sample size and the lower the standard error. Some surveys are conducted regularly every year, in conducting indirect estimation in the survey such as this, efficiency of estimating education index for district level can be improved by including the random effect of the area as well as the random effect of time (Sadik and Notodipuro, 2006). So in this study will be used Empirical Best Linear Unbiased Prediction (EBLUP) by combining the time series and cross-section data for estimating the education index at the level of districts in Purwakarta. The direct estimation of education index produce a larger variance than our methode, it shown by comparing mean square error (MSE) of direct method and indirect method, direct method have the largest MSE.

Key words : Indirect Estimation, Small Area Estimator, EBLUP, Time Series and Cross-Section, HDI, Education Index.

INTRODUCTION

Background

Since the implementation of decentralisation in Indonesia, local governments are given the freedom to manage their respective regions. The need for data or more detailed information about the condition of an area become indispensable. That information is used to determine as well as an evaluation of development that local government has been done. Successful development of a region can be seen through the Human Development Index (HDI). HDI is based on three basic dimensions: long and healthy life, knowledge and a decent life dimension (BPS 2007). Each dimension is measured by an index there are life expectancy index, education index and the purchasing power index.Education is an important and effective factor for human development. According to Todaro and Smith (2006) in Ilhami (2014), Education plays an important role in create human capabilities of a country to absorb and create modern technology and to develop the capacity to create sustainable growth and development. Increased of education index will result in increasing a person's capacity as a result of education, by increasing a person's capacity is expected to get a better job and getting a higher income. Higher income levels will

increase purchasing power Index, on the other hand someone with a high level of income can get better health care so that the index life expectancy increased too.

Purwakarta is one of the districts in West Java who pay big attention to education, with the vision is makes Purwakarta as districts that excel in education in West Java, but until now Purwakarta's education index was still below West Java Province. Based on data from BPS Purwakarta's education index in 2013 amounted to 0.82, while the province of West Java 0.83. Therefore, to realize this vision, the government purwakarta should give special attention in the field of education.One step that can be done is to look for the amount of information on the education index each district in Purwakarta. This was done with the aim to provide the right policy for each sub-district. Education index calculation is done using data from the National Socioeconomic Survey (Susenas). Direct estimation of Education index for districts is not feasible because these estimates will generate a large standard error due to small samples size (Datta and Lahiri, 2000). To solve these problems, statistical methods by performing the small area estimation (SAE) can be used. This estimation are based on models (implicit or explicit) which provide a link to related small area through supplementary data such as administratif record and recent census counts (rao-yu 1994). the whole of supplementary data should have relevance or corelation to the parameters observed (Rao, 2003).

Models for small area estimation is linear mixed models. One of technique completion linear mixed model is Empirical best linear prediction Unbiased (EBLUP) that was developed by Harville (1977). In the survey conducted periodically as SUSENAS, efficiency of education index estimation for the district level can be increased by including random effect of the area as well as the random effect of time (Sadik and Notodipuro, 2006). objective of this study is estimating the education index at the district level in Purwakarta using Empirical Best Linear Unbiased Prediction (EBLUP) by combining the time series data and cross-section.

LITERATURE REVIEW

Education index in HDI

dimension in HDI is measured by education index, the index is composed by two basic components, namely literacy rate and mean years school. The letter literacy rate is the achievement of basic education to the population, in this way the population is expected to apply in everyday life, so as to develop the social and economic condition. The literacy rate is the percentage of the population aged 15 years and above who can read and write Latin letters or other characters. Mean years school indicating higher school education achieved by the people in an area. The higher the mean years school means higher levels of education were undertaken. The mean years school describes the number of years used by the population aged 15 years and above in undergoing formal education.

EBLUP Rao-Yu (*Time Series and Cross-Sectional Model*)

Some surveys are often conducted repeatedly by replacing some elements of its samples which is SUSENAS. Rao and Yu (1992, 1994) extend Fay-Heriot models to overcome time series and cross-section data. Assume that

$$y_{it} = \theta_{it} + e_{it},$$
$$i = 1,...,m \text{ and } t = 1,...,T \qquad (1)$$

Where

$$\theta_{it} = \underset{\sim}{x}_{it}^{'}\underset{\sim}{\beta} + v_i + u_{it} \qquad (2)$$

Combining (1) and (2) may be written as

$$y_{it} = \underset{\sim}{x}_{it}^{'}\underset{\sim}{\beta} + v_i + u_{it} + e_{it} \qquad (3)$$

u_{it} follow AR (1) process for each *i*,

$$u_{it} = \rho u_{i,t-1} + \varepsilon_{it}, \ |\rho| < 1$$

y_{it} is direct estimator from survey for *i*th small area at time t , $\theta_{it} = g(\bar{Y}_{it})$ is means function of small area, while e_{it} is sampling error, it's assumed normally distributed with zero mean and block diagonal covariance matrix $\underset{\sim}{\sum} = blockdiag(\underset{\sim}{\sum}_1,...,\underset{\sim}{\sum}_m)$. x_{it} is vector of fixed covariate for *i*th small area at time t. Here $v_i \square N(0,\sigma_v^2)$, $\varepsilon_{it} \square N(0,\sigma^2)$,and v_i , ε_{it} are assumed independent. Based on (2) it show that θ_{it}

depend on random small area effect (v_i) and random area by time spesific effect (u_{it}).

Arranging data as $\underset{\sim}{y} = (y_{11},...,y_{1T};y_{21},...,y_{2T};y_{m1},...,y_{mT})' = (\underset{\sim}{y}'_1,...,\underset{\sim}{y}'_m)'$, equation persamaan (3) can rewrite as matrix form

$$\mathbf{y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{Z}\underset{\sim}{v} + \underset{\sim}{u} + \underset{\sim}{e} \qquad (4)$$

With

$\underset{\sim}{X} = (\underset{\sim}{X}'_1,...,\underset{\sim}{X}'_m)'; \underset{\sim}{X}'_i = (\underset{\sim}{x}_{i1},...,\underset{\sim}{x}_{iT})$

$\underset{\sim}{Z} = \underset{\sim}{I}_m \otimes \underset{\sim}{\mathbf{1}}$

$\underset{\sim}{v} = (\underset{\sim}{v}'_1,...,\underset{\sim}{v}'_m)'; \underset{\sim}{u} = (\underset{\sim}{u}'_1,...,\underset{\sim}{u}'_m)'$

$\underset{\sim}{e} = (\underset{\sim}{e}'_1,...,\underset{\sim}{e}'_m)'$

$\underset{\sim}{u}_i' = (u_{i1},...,u_{iT})$, $\underset{\sim}{e}_i' = (e_{i1},...,e_{iT})$

Where $\underset{\sim}{\mathbf{1}}_t$ is t-vektor of 1 and $\underset{\sim}{I}_m$ is identity matrix with order m. further

$E(\underset{\sim}{v}) = 0$, $\text{cov}(\underset{\sim}{v}) = \sigma_v^2 \underset{\sim}{I}_m$

$E(\underset{\sim}{u}) = 0$, $\text{cov}(\underset{\sim}{u}) = \sigma_u^2 \underset{\sim}{I}_m \otimes \underset{\sim}{\Gamma} = \sigma_u^2 \underset{\sim}{R}$

$E(\underset{\sim}{e}) = 0$,

$\text{cov}(\underset{\sim}{e}) = \underset{\sim}{\Sigma} = blockdiag(\underset{\sim}{\Sigma}_1,...,\underset{\sim}{\Sigma}_m)$

$\underset{\sim}{\Gamma}$ is a TxT matrix with elements $\rho^{|i-j|}/(1-\rho^2)$ and

$\text{cov}(\underset{\sim}{y}) = V = \underset{\sim}{\Sigma} + \sigma_u^2 R + \sigma_v^2 ZZ' = blockdiag_i(\underset{\sim}{\Sigma}_i + \sigma_u^2 \underset{\sim}{\Gamma} + \sigma_v^2 \underset{\sim}{J}_T)$

, with $\underset{\sim}{J}_T = \underset{\sim}{\mathbf{1}}_T \underset{\sim}{\mathbf{1}}'_T$.

Rao-Yu (1994) analogize model (2.3) as a General Linear Mixed Model (GLMM) due to a combination of fixed and random effects. In GLMM performed estimate of the linear combination of parameters τ , while equation (2) is special case of linier combination $\tau = \underset{\sim}{l}'\underset{\sim}{\beta} + \underset{\sim}{l}_1\underset{\sim}{v} + \underset{\sim}{l}_2\underset{\sim}{u}$ with $\underset{\sim}{l} = x_{it}$, $\underset{\sim}{l}_1$ is m-vector with 1 in the *i*th position and 0 elsewhere, and $\underset{\sim}{l}_2$ is (mT) vector with 1 in the *it*th and 0 elsewhere.

There are various methods of parameter estimation in General Linear Mixed Models, which is the Best Linear Unbiased Prediction (BLUP) and EBLUP. BLUP used to estimate random and mixture effects of a small area models and it has the smallest Mean Square Error (MSE) of all the unbiased and linear estimators (Rahman 2008). Assuming that $\sigma_u^2, \sigma_v^2, \rho$ are known, the BLUP of τ is

$$\tilde{\tau} = \underset{\sim}{l}'\tilde{\underset{\sim}{\beta}} + (\sigma_v^2 \underset{\sim}{l}_1'\underset{\sim}{Z} + \sigma_u^2 \underset{\sim}{l}_2'\underset{\sim}{R})V^{-1}(\underset{\sim}{y} - \underset{\sim}{X}\tilde{\underset{\sim}{\beta}})$$

And can rewrite as form

$$\tilde{\theta}_{it} = t(\sigma_v^2, \sigma_u^2, \rho, y_{it})$$
$$= \underset{\sim}{x}'_{it}\tilde{\underset{\sim}{\beta}} + (\sigma_v^2 \underset{\sim}{1}'_T + \sigma_u^2 \underset{\sim}{\gamma}_t)'(\underset{\sim}{\Sigma}_i + \sigma_u^2 \underset{\sim}{\Gamma} + \sigma_v^2 \underset{\sim}{J}_T)^{-1}(\underset{\sim}{y}_i - \underset{\sim}{X}_i\tilde{\underset{\sim}{\beta}})$$

where $\tilde{\underset{\sim}{\beta}} = (\underset{\sim}{X}'\underset{\sim}{V}^{-1}\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{V}^{-1}\underset{\sim}{y}$ is *Generalize Least Square* estimator from $\underset{\sim}{\beta}$, $\underset{\sim}{\gamma}_t$ is *i*th row of $\underset{\sim}{\Gamma}$. In practice parameters $\sigma_u^2, \sigma_v^2, \rho$ are unknown, therefore *empirical* BLUP is obtained by replacing those parameters with their consistent estimator. The parameters estimator ($\hat{\sigma}_u^2, \hat{\sigma}_v^2, \hat{\rho}$) are obtained by using *Maximum Likelihood* (ML) and *Restricted Maximum Likelihood* (REML) method (see Diallo 2014),this study using ML method to estimate the parameters and MSE.

METHODOLOGY

Data exploration

Variable respone (y_{it}) of this study is education index which measured by SUSENAS's data from 2009 until 2013 and auxiliary variables are obtained from cencus PODES 2008 and 2011. There are 9 variables :

x_1 : Total Population
x_2 : Number of Farming Family
x_3 : Number of farm workers Family
x_4 : Number of non Electricity Family
x_5 : Male/Female
x_6 : Number of elementer school
x_7 : Number of Junior High school
x_8 : Number of Senior High school
x_9 : Number of Letter of proverty

Data exploration conducted on all auxiliary variables and education indices in each sub-district in the district of Purwakarta. Exploration results data can be seen in Table 1 which shows that overall auxiliary variables have a different values of range it is necessary to standardize the data. Standardization of data is done into normal standard form

$$Z = \frac{x_i - \bar{x}}{\sigma}$$

Where

x_i : *i* th Data
$\bar{x}$: Mean Data

σ : Standar Deviation

Table 1. Result of Data Eksploration

	N	Minimum	Maximum	Mean	Std. Deviation
x1	80	13315.00	164853.00	50981.48	31252.11
x2	80	60.00	13038.00	3614.113	3309.54
x3	80	1252.00	6761.00	3019.23	1397.33
x4	80	.00	4842.00	446.25	946.45
x5	80	.86	1.100	1.023	.05
x6	80	10.00	93.00	29.76	17.47
x7	80	2.00	21.00	6.48	4.02
x8	80	.00	21.00	3.54	4.38
x9	80	76.00	2196.00	758.45	548.08

Once the data are standardized, then performed a correlation analysis to find variables that had a significant correlation to the response variable (education index). Correlation analysis results can be seen in Table 2.

Table 2. Correlation analysis results

	Pearson Correlation	Sig. (2-tailed)	N
Indeks Pendidikan	1		80
Zscore (x1)	**.309****	0.005	80
Zscore (x2)	-0.007	0.951	80
Zscore (x3)	**-.231***	0.039	80
Zscore (x4)	-0.012	0.918	80
Zscore (x5)	0.031	0.787	80
Zscore (x6)	**.314****	0.005	80
Zscore (x7)	0.231	0.051	80
Zscore (x8)	**.296****	0.008	80
Zscore (x9)	0.21	0.061	80

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

From the table it show that after standardization, there are four variables that correlated significantly with Education Index, which Zscore variables x1 (zx1), Zscore x3 (ZX3), Zscore x6 (ZX6) and Zscore x8 (zx8)

RESULT AND DISCUSSION

Purwakarta district consists of 17 sub-districts. There is a district that does not use as sample, the average number of population in each sub-district from years 2009-2013 is 50. 981 inhabitants (Podes, 2008 and 2011) and just 0.228% are used as sample, the number of this sample very small for each district. This small size of samples make the direct estimators generated much greater MSE than proposed small area model. It can be seen in Figure 1.

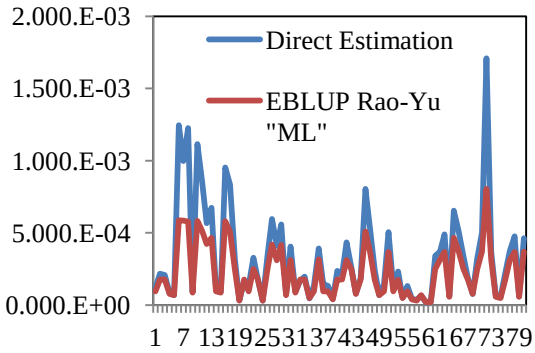

Figure 1. MSE of Education Index

The resulting equation model from EBLUP Rao- Yu is

$$Education\_Index = 0{,}870 - 0{,}000000384x_1 - 0{,}00000501x_3 + 0{,}000973x_6 + 0.00228x_8$$

With $\hat{\sigma}_\upsilon^2 = 0.00099$, $\hat{\sigma}_u^2 = 0.000495$

dan $\hat{\rho} = 0.0516$

Education index estimation results for each sub-district using the EBLUP Rao- Yu method shown in table 3.

Table 3 Education Index estimation result

	Year				
Sub- Distric	2009	2010	2011	2012	2013
10	0.84	0.82	0.83	0.80	0.85
11	0.75	0.77	0.76	0.79	0.78
20	0.76	0.76	0.77	0.80	0.80
30	0.70	0.73	0.73	0.76	0.76
40	0.80	0.75	0.78	0.76	0.75
50	0.73	0.79	0.78	0.80	0.80
60	0.80	0.80	0.79	0.78	0.83
70	0.76	0.79	0.78	0.80	0.77
80	0.76	0.78	0.80	0.79	0.80
81	0.78	0.76	0.79	0.81	0.80
90	0.81	0.86	0.81	0.81	0.84
100	0.88	0.88	0.76	0.87	0.89
101	0.79	0.85	0.85	0.78	0.80
110	0.74	0.84	0.81	0.78	0.79
111	0.82	0.78	0.81	0.84	0.85
112	0.85	0.79	0.86	0.88	0.84

Based on the results estimation by using a model-Yu Rao, the average of education index during the 2009-2013 each sub-district can be seen in figure 2.

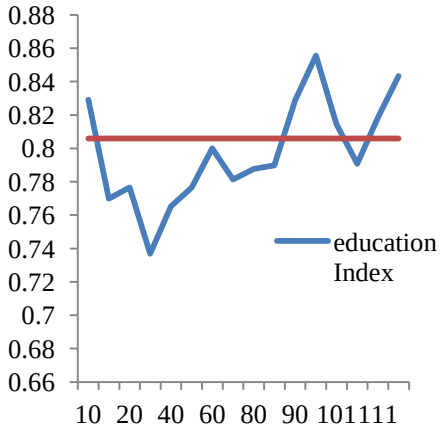

Figure 2. Average of Education Index

From Figure 2 shows that there are three sub-districts that have an average education index is above average education index districts purwakarta namely the districts 10,100 and 112 (Jatiluhur, Purwakarta and Bungursari) districts 100 (districts Purwakarta) is a district that has the most high average education index during the 2009-2013 period. It is very reasonable because Purwakarta is the city center so very easy access to educational facilities in sub-distric Purwakarta. Sub-districts 30 (districts Tegal Waru) has the lowest average education index, this could have been due to the location of the district that is quite far from the district purwakarta. It makes inadequate educational facilities .

CONCLUSION

Estimation of Education Index in Purwakarta district using a model of small area Rao -Yu is much better than the direct estimation, it shown by the MSE, direct estimation produces a much larger MSE than model of small area Rao –Yu. The small area estimation model its estimation results would be better if it included a random effect of area and time (model-Yu Rao).

REFERENCES

[BPS] Badan Pusat Statistik. 2007. *Publikasi Indeks Pembangunan Manusia 2006-2007*. Jakarta (ID):Badan Pusat Statistik.

.

Diallo MS. 2014. *Small Area Estimation Under Skew-Normal Nested Error Models* [disertasi]. [Canada]: Carleton University

.

Ilhami S. 2014. *Analisis Pengaruh Anggaran Pendidikan terhadap Indeks Pembangunan Manusia di Indonesia* [Tesis]. ID[Bogor]: Institut Pertanian Bogor.

Rahman A. 2008. *A Review of Small Area Estimation Problems and Methodological Developments*. Australia: University of Canberra.

Rao JNK. 2003. *Small Area Estimation*. New Jersey (US): John Wiley & Sons, Inc.

_____.2003. Some New Developments in Small Area Estimation. *Journal of Iranian Statistical Society*. 2(2) : 145-169.

Rao JNK dan Yu M. 1994. Small Area Estimation by Combining Time Series And Cross-Sectional Data. *Proceeding of the Section on survey Research Method. American Statistical Association*.

_____. 1994. Small Area Estimation By Combining Time Series and Cross-Sectional Data. *The Canadian journal of statistics*. 22(4): 511-528.

Sadik K dan Notodipuro KA. 2006. Model E-BLUP dalam Small Area Estimation untuk Model yang Mengandung Random Walk. *Jurnal Forum Statistika dan Komputasi*.11(2): 37 – 41.