IJCSN - International Journal of Computer Science and Network, Volume 6, Issue 6, December 2017
ISSN (Online) : 2277-5420
www.IJCSN.org
Impact Factor: 1.5

Zero Inflated Beta Model in Small Area Estimation to Estimate Poverty Rates on Village Level in Langsa Municipality

[1] Meita Jumiartanti; [2] Indahwati; [3] Anang Kurnia

[1] Department of Statistics, Bogor Agricultural University, IPB
Bogor, 16680, Indonesia

[2] Department of Statistics, Bogor Agricultural University, IPB
Bogor, 16680, Indonesia

[3] Department of Statistics, Bogor Agricultural University, IPB
Bogor, 16680, Indonesia

Abstract - Village level poverty rates are needed as a consideration for allocating village funds. The national socio economic survey samples are designed to estimate poverty rates in province and distric level. Direct estimate for calculating estimates of village level poverty rates does not have a good precision due to small sample sizes. Small Area Estimation (SAE) technique is used to produce a good precision with small sample sizes. The estimates of poverty rates should also be produced for non sampled area and when no poor are included in the sample. We propose zero inflated beta model because poverty rates takes value in the intervals [0,1). Clustering technique is used to acommodate random effect area for non sampled area. The purpose of this research is to estimate poverty rates on village level in Langsa Municipality. The result showed that estimates poverty rates on village level with zero inflated beta model is better than direct estimates.

Keywords - *Clustering, Poverty Rates, Small Area Estimation, Zero Inflated Beta Model*

1. Introduction

Village level poverty rates are needed as consideration for allocating village funds. Statistics Indonesia (BPS) is an institution that officially releases poverty rates in Indonesia. BPS releases annual poverty rates sourced from the National Socio-Economic Survey (Susenas). Susenas was first implemented in 1963, by collecting data on household consumption expenditure. Furthermore, Susenas has several times development. Since 2015 the implementation of Susenas is conducted twice a year, i.e. March and September. Enumeration in March resulted in data for the estimation of district-level poverty rates. While the enumeration in September produced data for the estimation of province and national level poverty rates.

However, the estimation of the poverty rates at village level can not be done using Susenas data in March 2016, because the sample of Susenas in March 2016 is only sufficient to estimate poverty rates at districts level. In this case, the village area is a small area. If we predict the poverty rates in a small area directly, direct estimator have no good precision (Kurnia and Notodiputro 2006). In order to produce good precision with small sample sizes, one way is to apply the Small Area Estimation (SAE) technique.

Small area estimation uses additional information from areas around the small area i.e. from the census and administrative records (Rao 2015). The model to be used in the small area estimation is the beta regression model. This is based on the assumption that the response variable is beta distributed has a value in intervals (0.1) and it can be related to a set of predictor variables linearly. According to Swearingen *et al*. (2011) the beta regression model is a more accurate and more efficient model to estimate parameter than the ordinary least squares method if the response variable is asymmetric or when there is heteroscedasticity.

The response variable used in this study is the poverty rates who have value intervals between 0 and 1. The problem is due to the small sample sizes, the response variable at the small area level (village) many are worth 0.

This research uses the zero inflated beta model that can accommodate the value 0 to overcome this problem. The zero inflated beta model is expected to model poverty rates of village level well.

Another problem that appear in this research is when to predict poverty rates the villages that are not taken as a sample (non sampled area). Model estimation results obtained from the sample area for non sampled areas will ignore the random effects of the area. One of the approaches used in this research is by clustering technique. This technique is based on the assumption that an area has a proximity relationship with other areas based on specific characteristics.

This paper aimed to estimate poverty rates using zero inflated beta regression model on village level in Langsa municipality for sampled area and non sampled area.

2. Literature Reviews

2.1 Beta Regression Model

Beta regression model can be used for modelling response variables with intervals (0,1). It is usefull for modelling proportion data without zeros and ones in data. The probability density function (pdf) of beta distribution with parameter μ and ϕ can be written as

$$f(y;\mu,\phi)=\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1},\ 0<y<1 \quad (1)$$

where $0<\mu<1$, $\phi>0$, $\Gamma(.)$is the gamma function, y is response variable, $E(y)=\mu$ is the mean of the response variable, $var(y)=\frac{\mu(1-\mu)}{1+\phi}$. ϕ is interpreted as a precision parameter, for fixed μ the larger value of ϕ then the larger homogenity as well.

2.2 Zero Inflated Beta Regression Model in Small Area Estimation

We use zero inflated beta regression for modelling proportional data when there is a value of zero in the data. The probability density function (pdf) of response variable y can be written as:

$$bi_0(y;\alpha,\mu,\phi)=\begin{cases}\alpha, & \text{jika } y=0\\ (1-\alpha)f(y;\mu,\phi), & \text{jika } y\in(0,1)\end{cases} \quad (2)$$

where $f(y;\mu,\phi)$ is probability density function of beta distribution as shown in (1) and α is probability when y=0. Equation (2) is called zero inflated beta distribution.

The mean and the variance from equation (2) can be written as:

$$E(y)=(1-\alpha)\mu \quad (3)$$

$$Var(y)=(1-\alpha)\frac{\mu(1-\mu)}{1+\phi}+\alpha(1-\alpha)(\mu)^2 \quad (4)$$

Parameters μ and α for poverty rates can be modeled as follows:

$$g_1(\mu_i)=\text{logit}(\mu_i)=\boldsymbol{x_{1i}^T\beta_1}+z_{1i}v_{1i} \quad (5)$$

$$g_2(\alpha_i)=\text{logit}(\alpha_i)=\boldsymbol{x_{2i}^T\beta_2}+z_{2i}v_{2i} \quad (6)$$

where $\boldsymbol{\beta_1}$ is the vector of the fixed effects regression coefficients of the beta distribution mean, μ_i. $\boldsymbol{\beta_2}$ is the regression coefficient for α_i. $\boldsymbol{x_{1i}^T}$ and $\boldsymbol{x_{2i}^T}$ are design matrices of auxiliary variables corresponding to the vectors of fixed effects. v_i is the design matrix for random effects area, where:

$$v_{2i},v_{2i}\sim N(\mathbf{0},\boldsymbol{\Sigma}) \text{ ,where } \boldsymbol{\Sigma}=\begin{pmatrix}\sigma_{v1}^2 & \sigma_{v,12}\\ \sigma_{v,12} & \sigma_{v2}^2\end{pmatrix} \quad (7)$$

Based on Eq. (5) and (6), the estimator for $\mu_i$and $\alpha_i$can written as :

$$\hat{\mu}_i=\frac{\exp(\boldsymbol{x_{1,i}^T\hat{\beta}_1}+z_{1,i}\hat{v}_{1,i})}{1+\exp(\boldsymbol{x_{1,i}^T\hat{\beta}_1}+z_{1,i}\hat{v}_{1,i})} \quad (8)$$

$$\hat{\alpha}_i=\frac{\exp(\boldsymbol{x_{2,i}^T\hat{\beta}_2}+z_{2,i}\hat{v}_{2,i})}{1+\exp(\boldsymbol{x_{2,i}^T\hat{\beta}_2}+z_{2,i}\hat{v}_{2,i})} \quad (9)$$

Indirect estimator for small area estimation using zero inflated beta regression model can be written as :

$$\hat{Y}_i=E(y_i|v_{1,i},v_{2,i})=(1-\hat{\alpha}_i)\hat{\mu}_i$$

$$=\frac{\exp(\boldsymbol{x_{1,i}^T\hat{\beta}_1}+z_{1,i}\hat{v}_{1,i})}{(1+\exp(\boldsymbol{x_{1,i}^T\hat{\beta}_1}+z_{1,i}\hat{v}_{1,i}))(1+\exp(\boldsymbol{x_{2,i}^T\hat{\beta}_2}+z_{2,i}\hat{v}_{2,i}))} \quad (10)$$

3. Methods

3.1 Data

The data used in this study were obtained from National Socio economic Survey (Susenas) March in 2016, Village Potential (Podes) Database in 2014 and publication Sub district in Figures 2016 which contain all information about villages in Langsa Municipality. The calculation of

poverty rates conducted in all villages in Langsa Municipality, which is 66 villages, consisting of 34 villages that become Susenas samples in March 2016 and 32 non-sampled villages. The sample sizes of Susenas in March 2016 in Langsa Municipality is 52 census blocks and each census block consists of 10 households, so the total sample sizes is 520 households in Langsa Municipality.

The response variable used in this study is the poverty rates, calculated based on the average monthly spending data per capita Susenas March 2016.

The auxiliary variables contain information about all villages in Langsa Municipality. The auxiliary variables used in this study are obtained from Podes 2014 data. The following is the auxiliary variables selected from Podes 2014.

Table 1. Auxiliary variables

Variables	Information
X1	Proportion of agricultural families
X2	Proportion of population receiving public health insurance for the poor
X3	Proportion population receiving the Poor Certificate
X4	Closest distance to junior high school
X5	Closest distance to high school
X6	Closest distance to vocational high school
X7	Closest distance to academy/university
X8	Percentage of families using the phone
X9	Proportion of families using electricity
X10	Number of minimarkets (unit)
X11	Ratio of doctor's clinic per 1000 population

The variables used for clustering were obtained from the Village Potensial 2014 (Podes) and publication Sub district in Figures 2016 (KCDA). In this study, the variables used for clustering represent five areas i.e. population, education, health, economics and telecommunications.

3.2 Method of Analysis

The steps used in analyzing the data are:

a. Estimates poverty rates at the village level using direct estimation methods, using the Foster, Greer and Thornbecke equations as follows:

$$\hat{P}_{0i}=\frac{1}{n_i}\sum_{i_0=1}^{n_i}\left(\frac{GK-y_{i_0}}{GK}\right)^0 I(y_{i_0}<GK) \quad (11)$$

where:
$\hat{P}_{0i}$ = poverty rates
n_i = population in village i-th
y_{i_0} = average monthly per capita expenditure of the poor(i_0=1,2,…..,n_{0i})
GK = the poverty line of Langsa Municipality

b. Choosing the variables to be used as the auxiliary variables in the small area estimation model
c. Modeling poverty rates in village level with a zero inflated beta regression model. Poverty rates (response variable) in intervals value (0.1) using Eq. (5), while for poverty rates with zero value using Eq. (6)
d. Estimates poverty rates for sampled area in Langsa Municipality by indirect estimator method in a small area using zero inflated beta regression model using Eq. (10)
e. Estimates poverty rates for non sampled area in Langsa Municipality, with the following stages :
   1. Clustering villages based on variables chosen from Podes 2014 and KCDA 2016
   2. Identify non sampled area (villages) to the cluster which has been made.
   3. Calculate the mean of the random effects area which is located on one cluster in each cluster, with the following equation:
   $\bar{\hat{v}}_{(k)} = \frac{1}{m_k}\sum_{i=1}^{m_k}\hat{v}_i$
   where m_k is the number of sample areas in the k-th cluster.
   4. Estimates poverty rates in village level by modifying the zero inflated beta model for small area estimation by adding the mean of the random effect area predictor ($\bar{\hat{v}}_{(k)}$) in each cluster in the poverty rates estimation model of the non sampled area. Poverty rates for non sampled area is estimated using the following equation:

$$\hat{Y}_i^* = \frac{\exp(\boldsymbol{x}_{1,i}^{T*}\hat{\boldsymbol{\beta}}_1+z_{1,i}\,\bar{\hat{v}}_{1(k)})}{(1+\exp(\boldsymbol{x}_{1,i}^{T*}\hat{\boldsymbol{\beta}}_1+z_{1,i}\,\bar{\hat{v}}_{1(k)}))(1+\exp(\boldsymbol{x}_{2,i}^{T*}\hat{\boldsymbol{\beta}}_2+z_{2,i}\,\bar{\hat{v}}_{2(k)}))} \quad (12)$$

4. Result

4.1 Direct Estimation of Poverty Rates

The direct estimation of poverty rates in Langsa Municipality was conducted using the average monthly per capita income data using Eq. (11), in 34 villages selected for sample. The statistical value of direct estimator of the poverty rates at the village level in Langsa Municipality can be seen in Table 4. Based on the

statistical value of direct estimator it can be seen that the minimum and median value of direct estimator is 0, so it can be said that the results of the direct estimator does not reflect the population because there are not villages where the whole population is not poor. This is due to the small sample sizes, so direct estimators do not have a good precision to estimate the poverty rates in village level.

4.2 Variables Selection

Poverty rates in village level contain zeros more than 50 percent, so we fitted the data using zero inflated beta regression. In this case, the response models are mixture of two sub-models i.e. a "beta" model that models the expected proportion of poverty rates in interval (0,1) and a "zero" model that models the expected proportion of poverty rates contain value zeros. The random effect for the data in this study was village (area).

The selection of auxiliary variables is very important in the small area estimation, since the indirect estimator will yield good results when the auxiliary variables are used is appropriate. We using R package GAMLSS using function stepGAICAll.B() that fitted zero inflated beta distribution to select variables in Table 1. The final model auxiliary variables chosen using the lowest AIC (Akaike Information Criterion) are shown in Table 2.

Table 2. Selection variables for fitting zero inflated beta regression

Tahap	Model	AIC
1	logit (μ_i) = X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11 logit (α_i) = X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11	-79.68
2	logit (μ_i) = X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X11 logit (α_i) =X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X11	-83.31
3	logit (μ_i) = X1 + X2 + X3 + X4 + X6 + X7 + X8 + X9 + X11 logit (α_i) = X1 + X2 + X3 + X4 + X6 + X7 + X8 + X9 + X11	-85.30
4	logit (μ_i) = X1 + X2 + X3 + X4 + X6 + X7 + X8 + X11 logit (α_i) = X1 + X2 + X3 + X4 + X6 + X7 + X8 + X11	-87.18
5	logit (μ_i) = X1 + X2 + X3 + X4 + X6 + X7 + X11 logit (α_i) = X1 + X2 + X3 + X4 + X6 + X7 + X11	-89.01
6	logit (μ_i) = X1 + X2 + X3 + X4 + X6 + X7 logit (α_i) = X1 + X2 + X3 + X4 + X6 + X7	-90.88
7	logit (μ_i) = X1 + X2 + X3 + X4 + X6 + X7 logit (α_i) =X6	-94.98

Using the generalized Akaike information criterion as model selection criterion, the following final model was selected:

$$\text{logit}(\mu_i)=\beta_{1,0} + \beta_{1,1}X1 + \beta_{1,2}X2 + \beta_{1,3}X3 + \beta_{1,4}X4 + \beta_{1,6}X6 + \beta_{1,7}X7 \quad (13)$$

$$\text{logit}(\alpha_i)=\beta_{2,0} + \beta_{2,6}X6 \quad (14)$$

It is important to note that only variable X6 was incorporated to μ and α.

4.3 Model Fitting

Zero inflated beta regression models with random effect were implemented and estimated using R package GAMLSS. Estimation of fixed effect parameters under zero inflated beta model for a sample of 34 villages are presented in Table 3.

The auxiliary variables chosen for modeling is based on the results of the stepwise method. Modeling has a R-squared value of 97 percent, meaning that the auxiliary variables is able to explain the diversity of the response variable by 97 percent.

As can be seen in Table 3, the most influential auxiliary variable to the response variable is variable X3 i.e. Proportion population receiving the Poor Certificate.

Table 3. Table of fixed effects parameters, precision parameter and random effect parameter

Effect	Parameter	mean
Logit(μ_i)		
Intercept	$\beta_{1,0}$	-2.337
X1	$\beta_{1,1}$	2.166
X2	$\beta_{1,2}$	-0.717
X3	$\beta_{1,3}$	7.305
X4	$\beta_{1,4}$	0.036
X6	$\beta_{1,6}$	-0.034
X7	$\beta_{1,7}$	-0.029
Logit(α_i)		
Intercept	$\beta_{2,0}$	2.402
X6	$\beta_{2,6}$	-0.302

4.3 Non Sampled Area Estimation

Clustering is used to estimate the poverty rates of non sampled area (village). Clustering is used to 'borrow' a random effect area from sampled area. Clustering is conducted by hierarchy due to the number of cluster to be formed has not been determined. The method used in clustering analysis is the Ward method. Determination of number of cluster is subjective researcher based on result of dendogram and also information and characteristic of village. Clustering analysis produces 8 cluster. Dendogram is showed in Fig 1

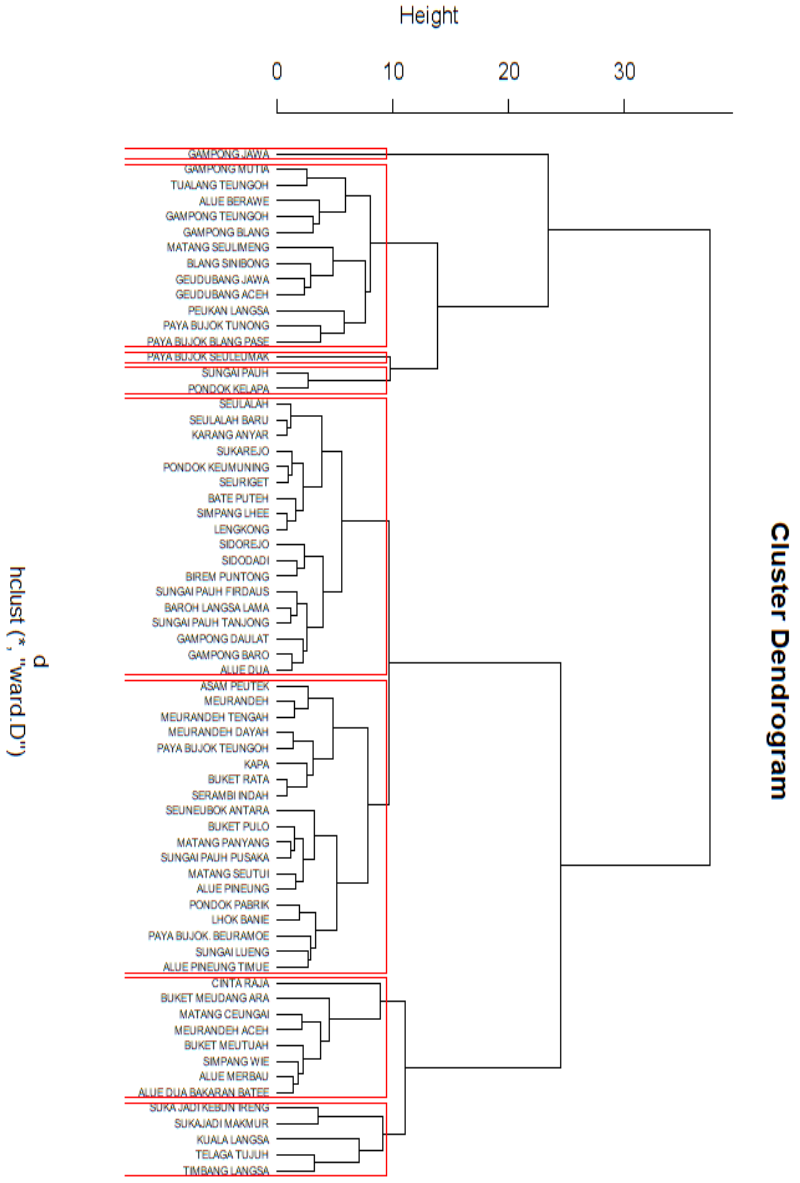

Fig. 1 Dendogram analysis of village cluster in Langsa Municipality

The statistical value of indirect estimator of non sampled villages using zero inflated beta model can be seen in Table 4.

4.4 Estimation of Poverty Rates

According to rao (2015), assuming direct predictor $\hat{Y}_+$ in higher level has good and reliable precision, it is desirable that the estimator of a small area mean when aggregated will be the same as the result of a direct estimator at a higher level, if $\hat{Y}_+$ is regarded as the most correct. This is because direct estimators at higher levels have large sample sizes to produce estimator with good precision and reliable. The poverty rates in Langsa Municipality based on direct estimator is 11.09 percent. The estimator of poverty rates in village level in Langsa Municipality when aggregated is not the same as direct estimators of Langsa Municipality poverty rate.

$$\hat{Y}_+ \neq \sum_{i=1}^{m} W_i \hat{Y}_i \quad (15)$$

It is therefore necessary to modify/calibrate the indirect estimator produced by the zero inflated beta regresion model. According to Rao (2015) a simple adjustment for indirect estimator can be done with the following equation:

$$\hat{Y}_i^{DB} = \hat{Y}_i + \left(\hat{Y}_+ - \sum_{i=1}^{m} W_i \hat{Y}_i\right) \quad (16)$$

After calibration with Eq. (16), the small area estimator of poverty rates at the village level when in aggregated is similar to the direct estimator of the poverty rates of Langsa Municipality, i.e. $11.09 = \hat{Y}_+ = \sum_{i=1}^{m} W_i \hat{Y}_i^{DB}$ with $\hat{Y}_i^{DB}$ is an estimator of poverty rates in village level with a calibrated zero inflated beta model.

In Table 4 can be seen the results of indirect estimator after the calibration. In the calibration model it can be seen that the average of indirect allegations of 8.08 percent with standard deviations of 3.64 percent is smaller than the standard deviation of the direct estimates of the proportion of the poor at the village level.

Table 4 Summary of statistics of poverty rates (%) in village level based on direct estimation method and the zero inflated beta model

Statistics	Estimation Method	
	Direct Estimators (34 villages)	Zero inflated beta with calibration model (66 villages)
Mean	3.67	8.08
Standard deviation	7.89	3.64
Minimum	0.00	4.48
Median	0.00	6.70
Maximum	30.31	18.51

Poverty rates of sampled and non sampled area in Langsa Municipality can be seen in Table 5 and Table 6. From table 5 we can see that the minimum value of poverty rate in sampled area is 4.48 i.e. Gampong Daulat and maximum is 10.68 i.e. Pondok Pabrik.

Table 5. Poverty rates estimate of sampled village in Langsa Municipality

Villages	Estimate
Alue Berawe	5,24
Alue Dua Bakaran Batee	7,58
Alue Merbau	8,49
Alue Pineung	6,36
Alue Pineung Timue	8,83
Asam Peutek	10,03
Baroh Langsa Lama	4,85
Buket Rata	6,21
Gampong Baro	6,12
Gampong Blang	5,07
Gampong Daulat	4,48
Gampong Jawa	4,55
Gampong Mutia	4,73
Gampong Teungoh	4,61
Geudubang Aceh	7,27
Geudubang Jawa	6,45
Karang Anyar	5,21
Lengkong	8,73
Matang Seulimeng	7,40
Meurandeh Dayah	5,31
Paya Bujok Seuleumak	5,11
Paya Bujok Teungoh	4,94
Paya Bujok Tunong	4,61
Paya Bujok. Beuramoe	6,78
Pondok Kelapa	6,04
Pondok Keumuning	8,05
Pondok Pabrik	10,68
Seulalah	6,40
Seulalah Baru	6,39
Seuriget	6,18
Sidodadi	6,73
Sidorejo	6,25
Sungai Pauh	5,90
Timbang Langsa	9,08

From table 6 we can see that the minimum value of poverty rate in non sampled area is 4.55 i.e. Paya Bujok Blang Pase and maximum is 18.51 i.e. Sukajadi Makmur.

Table 6. Poverty rates estimate of non sampled villages in Langsa Municipality

Villages	Estimate
Alue Dua	7,29
Bate Puteh	6,14
Birem Puntong	4,89
Blang Sinibong	6,08
Buket Meudang Ara	9,23
Buket Meutuah	10,54
Buket Pulo	6,51
Cinta Raja	13,76
Kapa	9,54
Kuala Langsa	17,42
Lhok Banie	12,04
Matang Ceungai	12,42
Matang Panyang	7,38
Matang Seutui	16,57
Meurandeh	17,53
Meurandeh Aceh	8,41
Meurandeh Tengah	9,30
Paya Bujok Blang Pase	4,55
Peukan Langsa	4,79
Serambi Indah	6,70
Seuneubok Antara	15,54
Simpang Lhee	6,62
Simpang Wie	11,32
Suka Jadi Kebun Ireng	17,42
Sukajadi Makmur	18,51
Sukarejo	6,70
Sungai Lueng	10,07
Sungai Pauh Firdaus	5,32
Sungai Pauh Pusaka	8,20
Sungai Pauh Tanjong	4,87
Telaga Tujuh	12,12
Tualang Teungoh	4,86

4. Conclusion

The conclusion is zero inflated beta model in the small area estimation produce a better estimator because no value is 0 at village level, whereas in the direct estimator there is a value of 0. Zero inflated beta model in the small area estimation to estimate non sampled area also produce good estimator values while still considering the fixed influence and random diversity of village areas. However, the result of zero inflated beta model in the small area estimation with the weighted average in village level still has a value under the direct estimator of the poverty rates of Langsa Municipality level, so it is necessary to calibrate the model, this is probably due to the value in the response variable contains a lot of value 0.

Acknowledgments

The authors wish to thank BPS-Statistics of Indonesia for allowing author to be part of 2nd Batch of BPS-APBN Master Scholarship Programme

References

[1] Anisa. R, "Kajian Pengaruh Penambahan Informasi Gerombol Terhadap Hasil Prediksi Area Nircontoh (Studi Kasus Pengeluaran per Kapita Kecamatan di Kota dan Kabupaten Bogor)", M.Si. thesis, Department of Statitics, Bogor Agricultural University, Bogor, Indonesia, 2014.

[2] [BPS] Badan Pusat Statistik. Penghitungan dan Analisis Kemiskinan Makro Indonesia Tahun 2016. Jakarta: Katalog BPS, 2016

[3] Ferrari. SLP, and Cribari-Neto. F, "Beta Regression For Modelling Rates and Proportions", Journal of Applied Statistics, Vol. XXXI, No. VII, 2004, pp. 799-815.

[4] Foster. J, Greer. J, and Thorbecke. E, "A class of decomposable poverty measures", Econometrica, 52, 1984, pp.761-766.

[5] Kurnia. A, and Notodiputro. KA, "Penerapan Metode Jackknife dalam Pendugaan Area Kecil", Forum Statistika dan Komputasi, Vol. XI, No. I, 2006, pp. 12-15.

[6] Ospina. R, and Ferrari. SLP, "Inflated beta distributions", Statisticals Papers, 51, 2010, pp. 111–126.

[7] Rao. JNK, Small Area Estimation. New York (US): John Wiley & Sons. 2015

[8] Swearingen CJ, Castro MSM, Bursac Z. "Modeling percentage outcomes: the %beta_regression macro", SAS Global Forum 2011: Statistics and Data Analysis Paper, 335, 2011, pp.1-12.

Authors -

Meita Jumiafrtanti currently pursuing masters degree program in Applied Statistics in Bogor Agricultural University, Indonesia. Attained bachelor degree from STIS Jakarta in 2007.

Indahwati is lecturer at Department of Statistics , Bogor Agricultural University, Indonesia. Her main interest is in Mixed Model and Small Area Estimation.

Anang Kurnia is lecturer at Department of Statistics , Bogor Agricul-tural University, Indonesia. His main interest is in Small Area Estimation.