Quasi Poisson Model for Estimating Under-Five Mortality Rate in Small Area

Nofita Istiana[1], Anang Kurnia[2], Azka Ubaidillah[3]
{nofistiana@gmail.com[1], anangk@apps.ipb.ac.id[2],azka@stis.ac.id[3]}

[1,2]Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB University, Bogor,16111, Indonesia
[2,3]STIS Polytechnic of Statistics, Jakarta, 13330, Indonesia

Abstract. Under-Five Mortality Rate (U5MR) is an important indicator because it reflects the socio-economic conditions and developments in health sector. U5MR is obtained from Demographic and Health Survey (DHS) where the level of estimation is designed for national and provincial level. The decentralization system makes the importance of U5MR for sub-domain of province such as district/municipality level. Small area estimation (SAE) can be used for estimating U5MR in district/municipality level by using a mixed model. The model that is often used is generalized linear mixed model (GLMM). Direct estimation of U5MR produces a large proportion of zero values (excess zero), so the Poisson model is not suitable for modeling the data. Excess zero is the reason for violating the equidispersion in Poisson model. In this study, quasi Poisson modelproduces better predictions than direct estimation. In addition, the U5MR estimation for municipality makes it possible to produce U5MR maps in municipality level.

Keywords: excess zero, GLMM, quasi Poisson, SAE, U5MR

1 Introduction

Mortality and morbidity rate are health indices that can be used to determine the degree of public health. One important type of mortality rate is Under-Five Mortality Rate (U5MR). The mortality rate of children under the age of five is an indicator that is sensitive to socio-economic changes in families and advances in health [1]. This indicator reflects the social, economic and environmental conditions of children where the children lives. U5MR is often used to identify the economic difficulties of the population.

U5MR in Indonesia is quite high but shows a decline, from 71 in 1997 to 32 in 2017. However, U5MR in Indonesia did not reach the MDGs target in 2015 which was 32 per thousand live births. The U5MR target for the SDGs program is 25 per thousand live births in 2030, so that appropriate policies are needed in an effort to reduce U5MR.

Indonesia is one of the developing countries whose population is always increasing every year, causing population density to increase in Indonesia. Based on the results of the population census in 2000 and 2010, Java Island is an island with the highest population density. In addition, with various facilities in Java, U5MR in Java Island is expected to be

ICSA 2019, August 02-03, Bogor, Indonesia

DOI 10.4108/eai.2-8-2019.2290343

below the national U5MR. However, there are two provinces that have U5MR above the national U5MR, namely Banten and East Java.

U5MR is obtained from Demographic and Health Survey (DHS) where the level of estimation is designed for national and provincial level. The decentralization system makes the importance of U5MR for sub-domain of province such as district/municipality, sub-district and village levels. Estimation at sub-domain level are sometimes difficult because it has a relatively small sample size or there are areas that are not surveyed. According to Rao and Molina [2], when there are sub-populations with small sample sizes, direct estimation can produce very large errors. In addition, direct estimation cannot be made for sub-populations that do not have sample. An indirect estimation technique is needed that can increase the effectiveness of sample size so that it can reduce standard errors. One analysis technique that can be used to overcome this problem is Small Area Estimation (SAE).

SAE can be done using modeling, which includes auxiliary variables as fixed effect and area as random effect. The modeling used is Generalized Linear Mixed Model (GLMM). U5MR data based on DHS is a count data, so Poisson regression can be used for modeling the data. However, the data shows a large of zero value. The excess zero problem is usually dealt with Zero Inflated Poisson (ZIP) [3], hurdle models [4], and other mixed Poisson distribution. ZIP and hurdle models are suitable for analyzing data with excessive amounts of zero. However, these models apply only when overdispersion occurs. Some cases have found that excessive amounts of zero are related to underdispersion [5], so ZIP and hurdle models are not suitable. Modeling that can overcome overdispersion and underdispersion is quasi Poisson [6].

This study aims to obtain U5MR for district/municipality level and predict U5MR for nonsampled areas with quasi Poisson model. Estimation with quasi Poisson model will be compared with ZIP and direct estimation. The measurements used for comparison are root mean square error (RMSE) with parametric bootstrap techniques and coefficient of variation (CV).

2 Materials

The data used in this study were 2017 IDHS data and the provincial health profile in 2016. The IDHS data was used to obtain direct estimation of U5MR, while the health profile was needed to obtain the auxiliary variables. The auxiliary variables used were proportion of children under five with pneumonia (X_1), proportion of babies with exclusive breastfeeding (X_2), ratio of children under five with malnutrition per 1000 under five children (X_3), proportion of population with access to clean drinking water (X_4), proportion of children under five with complete basic immunization (X_5).

3Methods

3.1 Direct Estimation of U5MR

Based on the 2017 IDHS report, under-five mortality is defined as the probability of death between birth and before reaching the fifth birthday (0-4) years, while the U5MR is under five mortality per thousand live births.U5MR estimation for district/municipality directly use data on birth date, survival status, and date of child mortality or age at the time of child death obtained from the IDHS 2017 data. Direct estimation of U5MRwas done with the

childhoodmortality package in R program. This package uses synthetic life table approach, combining the chancesof death from each age segment and the death of the actual cohort[7].This method allows for estimating U5MR trends over time [8].

3.2Small Area Estimation

Small Area Estimation is a method for estimating parameters in small area in a survey pilot by utilizing information from outside the area, from within the area itself, and from outside the survey [9]. Estimation of small areas are divided in two, namely design based and model based [2]. Design based uses weighting surveys and survey designs in the process of inferencing. While the model based or indirect estimation uses additional information in its inferential process [9]. Indirect estimation is done by adding information from neighboring areas and auxiliary variables derived from the census or administrative records and has a relationship with the observed variables so as to increase the effectiveness of sample sizes [2]. Based on the availability level of auxiliary variables, the small area estimation model is divided into two, namely, the unit level model and the area level model [2]. This study used indirect estimation with generalized linear mixed model (GLMM) and area level model to estimate U5MR at district/municipality level in Java island.GLMM is an extention of generalized linear model (GLM).

3.2.1 Generalized Linear Model (GLM)

GLM is a development of conventional linear models where population mean values depend on linear predictors through a nonlinear link function. In GLM, the response variable is assumed to follow the distribution of exponential family and is a function of the explanatory variable. According to McCullagh and Nelder [10], the GLM component consists of random components, systematic components, and functions that connect random components and systematic components. In the Poisson regression model, the link function used is log, so the Poisson regression model can be written as follows:

$$\log(\mu_i) = \boldsymbol{x_i'\beta} = \sum_{j=1}^{k} \beta_j x_{ij}\,, i = 1, \dots, n \qquad (1)$$

3.2.2Generalized Linear Mixed Model (GLMM)

GLMM combine the ideas of generalized linear models with the random effects [11]. Let $\boldsymbol{v}$ is vector random effect with distribution function $h(\boldsymbol{v}|\boldsymbol{u})$ and parameter $\boldsymbol{u}$, Poisson GLMM for response variable $y_i$is:

$$\log(\mu_i) = \boldsymbol{x_i'\beta} + \boldsymbol{z_i'v} \qquad (2)$$

with $\boldsymbol{\beta}$ is fixed effect, $\boldsymbol{x_i}$ and $\boldsymbol{z_i}$ are auxiliary variables for fixed effect and random effect, $\boldsymbol{v}$is assumed to be normally distributed ($\boldsymbol{v} \sim N(\boldsymbol{0}, \sigma_v^2 \boldsymbol{I})$)

3.3Overdispersion And Excess Zero

The variance of a Poisson distribution depends on the mean, with the mean equal to the variance. Count data frequently depart from the Poisson distribution due to a larger frequency of extreme observations resulting in spread (variance) greater than the mean in the observed distribution, called overdispersion. Overdispersion can occur due to clustering (grouping) in the population [10], unobserved heterogeneity [12], and great incidence of zero counts [13]. The implication of overdispersion is that Poisson regression is no longer suitable for modeling data. In addition, the formed model will produce biased parameter estimation [13].

The distribution of counts often has a much larger than expected number of observed zeros than assumed by Poisson distribution, called "excess zero" [14]. It turns out that excess zeros can be accommodated by the quasi Poisson model or, in fact, by any Poisson mixture model, for example ZIP.

3.4Zero Inflated Poisson

The ZIP model is a mixture of Poisson and Logistic model.An individual observation will belong to the Always-0 Group or zero state with a probability of p and will belong to group A̰ (the Not Always-0 Group or non-zero state), where the value zero and positive data, both of which are generated by a count distribution, for example Poisson or Negative Binomial with probability of *1-p*. ZIP is defined as follows:

$$f(y_i|v_i) = p_i + (1-p_i)e^{-\mu_i}, \text{for} y_i = 0 \quad (3)$$

$$f(y_i|v_i) = \frac{(1-p_i)e^{-\mu_i}\mu_i{}^{y_i}}{y_i!}, \text{for} y_i > 0 \quad (4)$$

with $E(y_i|v_i) = (1-p_i)\mu_i$and $V(y_i|v_i) = (1-p_i)[\mu_i^2 + \mu_i] - (1-p_i)^2\mu_i^2$. If $p_i = 0$ then (4) become poisson distribution.

3.5Quasi-Poisson

Extended quasi Poisson model is quasi Poisson with random effect. This model defined by Efron in 1986 [6] as follows:

$$f(y_i|v_i) = \phi^{-\frac{1}{2}}\exp\left[-\frac{\mu_i}{\phi}\right]\frac{exp\left[\left(\frac{1}{\phi}-1\right)y_i\right]y_i{}^{y_i}}{y_i!}\left(\frac{\mu_i}{y_i}\right)^{\frac{y_i}{\phi}}, \quad (5)$$

$$\approx \phi^{-1}exp\left[-\frac{\mu_i}{\phi}\right]\frac{\left(\frac{\mu_i}{\phi}\right)^{\frac{y_i}{\phi}}}{\left(\frac{y_i}{\phi}\right)!}$$

with ϕ is dispersion parameter, $E(y_i|v_i) = \mu_i$ and $Var(y_i|v_i) = \phi\mu_i$. If $\phi = 1$ then (5) become poisson distribution. This model allows overdispersion ($\phi > 1$) or underdipersion ($\phi < 1$).

3.6Cluster Analysis

Cluster analysis aims to group objects based on a set of measured variables into several groups so that similar objects will be in the same group. The similarity between objects can be measured byusing the concept of distance. In this study, cluster analysis was used to group districts/municipalities in Java island. According to Wahyudi *et al.* [15] and Sundara *et al.*[16], hierarchical cluster methods withWard shows that the addition of area specific random effects for estimating non-sampled areas will result in better estimation than synthetic models.

4Results and Discussion

Administratively, Java Island consists of six provinces and 119 districts/municipalities. U5MR direct estimation were made on 113 cities/municipalities that were sampled. The direct estimation of U5MR have a large error due to the small sample size in each district/municipality. In addition, the estimation of U5MR for non-sampled

districts/municipalities cannot be obtained. Small Area Estimation was needed by utilizing the auxiliary variables for better estimation results.

U5MR direct estimation shows that the number of districts/municipalities with zero values is 44 or 38.94 percent of all districts/municipalities in Java Island. This shows that there is indication of excess zero in the data. Excess zero is the reason for violating the assumptions of equality in average and variance in Poisson model. Therefore, an approach with ZIP and quasi Poisson model were carried out.

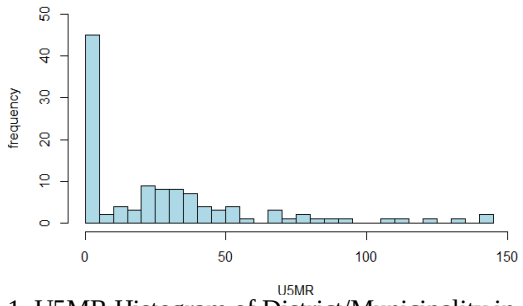
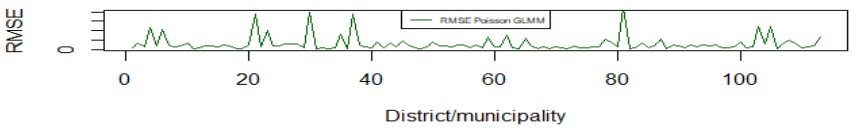


Figure 1. U5MR Histogram of District/Municipality in Java Island

Table 1 shows that the Poisson GLMM produces the closest approximation to the direct estimator and has the largest standard deviation compared to the other models. Quasi Poisson produces estimates with the same averages as direct estimates and lower standard deviations compared to the ZIP model.An evaluation of the value of RMSE with parametric bootstrap was done to see which modeling is better in estimating U5MR at district/municipality level.

Table 1Summary statistics of U5MR estimation

Statistic	Direct estimation	Poisson GLMM	Quasi Poisson GLMM	ZIP GLMM
Min	0.00	0.12	12.12	4.04
Quartil 1	0.00	0.32	22.77	19.54
Median	22.00	21.78	26.25	23.27
Mean	27.86	27.82	27.86	26.85
Quartil 3	39.00	38.76	29.24	29.29
Max	145.00	144.53	98.80	86.77
Standard deviation	33.64	33.41	11.11	15.00



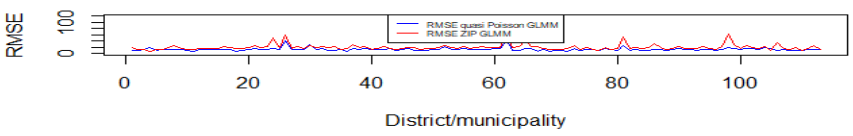


Figure 2. Comparison of RMSE Poisson GLMM, quasi Poisson GLMM, and ZIP GLMM

The best model is one that has the lowest RMSE. Figure 2 is a comparison of RMSE Poisson GLMM, quasi Poisson GLMM, and ZIP GLMM. The figure shows that Poisson GLMM produces a largest RMSE compared to the other two models. Quasi Poisson GLMM produces RMSE with a relatively constant pattern compared to ZIP. Moreover, the graph of quasi Poisson GLMM is under ZIP GLMM, so it can be concluded that the RMSE of quasi Poisson GLMM is lower than ZIP GLMM.

If the quasi Poisson GLMM is compared with the direct estimator, the model has a lower RMSE than the direct estimator. It can be concluded that this model can improve the results of direct estimates and can be used to estimate U5MR at the district/municipality level in Java.

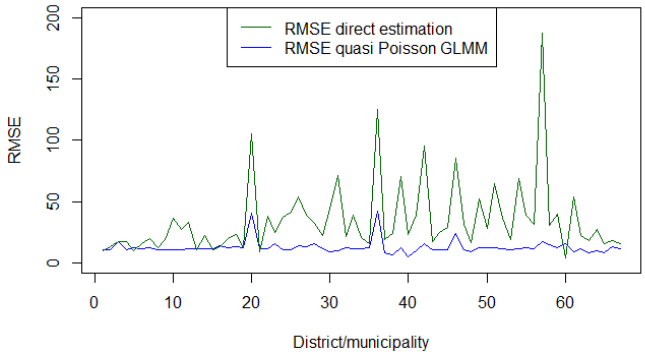


Figure 3. Comparison of RMSE quasi Poisson GLMM anddirect estimation

Further estimation was made of districts/municipalities that were not sampled. Estimationwas carried out with clustering corrections. Random effect for non-sampled districts/municipalities were obtained from the average random effect of sample districts/municipalities in one cluster with the non-sampled districts/municipalities.Districts/municipalities in Java Islandwere clustered through hierarchical clustering techniques. The method used in the cluster was the Ward method. The size of the similarity used was euclidean distance. The dendogram of the cluster analysis was shown in figure 4. The U5MR estimation for districts/municipality made it possible to produce U5MR maps in districts/municipality level. The map was shown in figure 5.

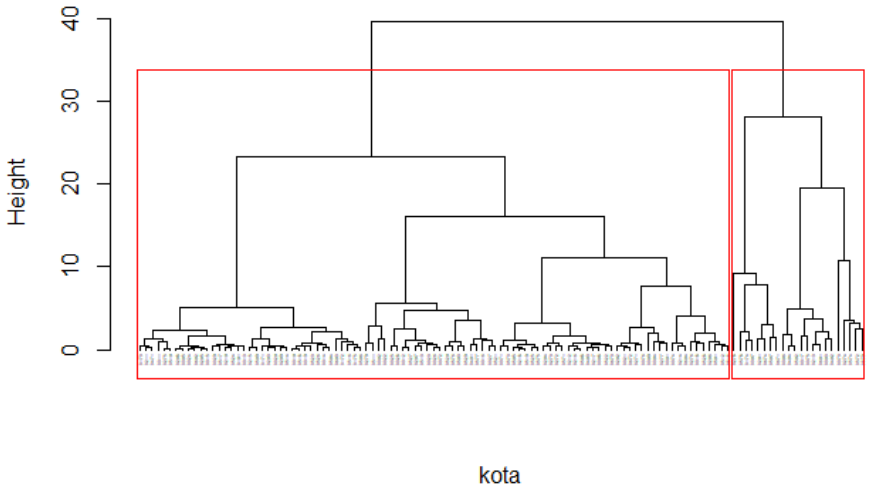


Figure 4. Dendogram of Cluster Analysis

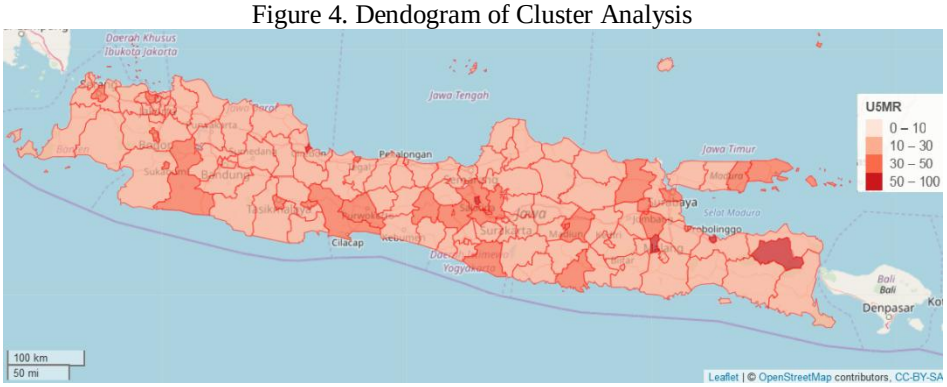


Figure 5. Map of AKBa at district/municipality level on Java Island

Based on the map, it appears that district/municipality with U5MR value of 50 per thousand births or moreare mostly in East Java. West Java has a relatively lower U5MR compared to other regions, except Cirebon city which has U5MR above 50 per thousand births. Table 2 presents the estimated U5MR for districts/municipalities that are not sampled.

Table2Estimation of U5MR not sampled districts/municipalities (per 1000 live births)

Kode	District/municipality	U5MR
3101	Kepulauan Seribu	27
3218	Pangandaran	28
3279	Banjar city	42
3574	Probolinggo city	66
3577	Madiun city	33
3579	Batu city	63

5Conclusion

The current U5MR estimate in Indonesia can only predict the provincial level. The use of SAE with quasi Poisson GLMM can be used to estimate U5MR in district/municipality level. This is because the estimation can improve direct estimates of small samples from the IDHS survey. In addition, SAE can also be used to estimate U5MR in areas that are not sampled by using cluster correction. According to USAID, U5MR calculation based on synthetic cohort allows the production of trends in child mortality. Therefore, the U5MR calculation for the district/municipality level can be used by the local government to reduce U5MR and reaching the SDGs target in 2020.

References

[1] Setyowati T, Budiarto E, Anggraeni D. 2002. Faktor Lingkungan yang Mempengaruhi Kematian Anak Balita. *Jurnal Ekologi Kesehatan* Vol.1, No.1, Februari 2001:1-6

[2] Rao JNK, Molina I. 2015. *Small Area Estimation 2nd ed*. New Jersey: John Wiley & Son.

[3] Lambert D. 1992. Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*, 3, 1–14

[4] Cragg JG. 1971. Some Statistical Models for Limited Dependent Variables with Application to The Demand for Durable Goods. *Econometrica*, 39, 829–844

[5] Tin A. 2008. Modeling Zero-Inflated Count Data with Underdispersion And Overdispersion. SAS Global Forum 2008: Statistics and Data Analysis. Retrieved from https://support.sas.com/resources/papers/proceedings/pdfs/sgf2008/372-2008.pdf

[6] Lee Y, Nelder J, Pawitan Y. 2006. *Generalized Linear Models with Random Effects (Unified Analysis via H-Likelihood)*. New York: Chapman & Hall/CRC.

[7] Breen C. 2018. Package 'childhoodmortality'. Retrieved from https://cran.r-project.org/web/packages/childhoodmortality/childhoodmortality.pdf

[8] [USAID] United States Agency for International Development. 2018. Guide to DHSstatistics.

[9] Kurnia A. 2009. Prediksi Terbaik Empirik Untuk Model Transformasi Logaritma Di Dalam Pendugaan Area Kecil Dengan Penerapan Pada Data Susenas [Disertasi]. Bogor (ID): Institut Pertanian Bogor

[10] McCullagh P, Nelder JA. 1989. *Generalized Linear Models*, 2nd Ed. New York: Chapman and Hall

[11] Faraway JJ. 2016. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models (2nd ed.)*. CRC Pr.

[12] Cameron A, Trivedi P. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press

[13] Ridout M, Demetrio CG, Hinde J. 1998. Models for Count Data With Many Zeros. *International Biometric Conference*, Cape Town, Desember 1998

[14] Winkelmann R. 2008. *Econometric Analysis of Count Data 5th edition*. Berlin: Springer

[15] Wahyudi, Notodiputro KA, Kurnia A, Anisa R. 2016. A study of area clustering using factor analysis in small area estimation (An analysis ofper capita expenditures of subdistricts level in regency and municipality of Bogor). AIP Conference Proceedings 1707, 080017 (2016). doi: 10.1063/1.4940874

[16] Sundara VY,Sadik K, Kurnia A. 2017. Clustering Information of Non-Sampled Area in Small Area Estimation of Poverty Indicators. AIP Conference Proceedings 1827, 020026 (2017). https://doi.org/10.1063/1.4979442