Hierarchical Bayesian Models for Small Area Estimation under Overdispersed Count Data

Ita Wulandari, *Member IAENG*, Khairil Anwar Notodiputro, Anwar Fitrianto, Anang Kurnia

***Abstract*—Bayesian analysis was applied to small area models with overdispersed response variables. The benefits of implementing this strategy by Markov Chain Monte Carlo methods make inference straightforward and computationally feasible. In this paper, we apply the strategy into area-level modeling to predict the under-five mortality rate at the district level in Java Island, the most populated region in Indonesia. The result shows that the zero-inflated negative binomial model yields the reduced relative standard error and relative mean squared error when compared to district estimates, the zero-inflated generalized Poisson and Poisson models.**

***Index Terms*—Count Data, Hierarchical Bayesian, Overdispersion, Zero-inflated, Under-five Mortality Rate**

I. Introduction

In order to meet the growing demand for accurate small-area estimation (SAE) in the public and private sectors, theoretical and practical approaches to small-area estimation have been actively and thoroughly explored [1]. Various data sources, including continuous and discrete (binary and count) data, have been extensively employed in small area estimates. Additionally, this sector has generated a type of data known as zero-inflation, which has a semi-continuous distribution with a mix of positive and negative values disspersed constantly. If we pay attention to these data characteristics, issues with inference can arise due to the zero-inflation on the data and can invalidate the assumptions of the model.

In relation to a data count, the Poisson Regression model is the simplest and most basic model. In order to determine the number of events which occur at a particular time, a Poisson regression model can be used [2]. It is assumed that the mean and variance are the same, a fundamental characteristic of this model. The observed variance is typically higher than expected, a condition is known as overdispersion, hence, this only sometimes holds. If the observed variance is greater than the theoretical variance predicted by the assumed distribution, overdispersion occurs. Overdispersion may occur for several causes. The major cause of overdispersion is the excessive number of zero observations (referred to as zero excess" or "zero-inflation"). Not accounting for the overdispersion may lead to underestimating the standard error and the test statistic has an excessive rate of type I error and poor confidence interval coverage.

The zero-inflated model is also called the two-part model in the generalized linear mixed model (GLMM), and it models nonzero values independently. The existence of zero values contributes greatly to the explanation of excess zeros in the data, see [2], and [3]. According to the SAE model, excess zeros are more significant in a small area than in a large area, particularly when the total sample size is large [3]. The best linear unbiased predictor (EBLUP) was found to be unsuitable for areas with a large number of zeros in other studies [4]. In addition, this method was also implemented by Irlandia et al. [5] by using the K-medoids cluster to estimate parameters. Furthermore, the SAE employing a Bayesian approach in a two-part random effects model was carried out by Chandra and Chambers [6]. The log-transformed linear mixed model (LMM) for nonzero data was developed by Pfeffermann et al. [7] in contrast to the frequentist approach in this study. In subsequent research, Kreig et al. [8] developed an SAE method for data with many zero-inflated values.

Estimating model parameters to calculate response variable data using the Bayesian model approach has begun to be widely used to solve SAE problems. In particular, according to Ghosh *et al*. [9], the hierarchical Bayesian (HB) and empirical Bayesian (EB) approaches have been applied to model the systematic component of the local area. The theory and application of the HB and EB methods used for the SAE have been widely discussed by Ghosh and Maiti [10], [11], Datta *et al*. [12], [13], and Torabi and Rao [14].

The HB model extends Fay-Herriot's model, as well as generalized linear models, by using two prior distributions on the target parameters. The benefits of implementing this strategy include a simple definition and the ability to consider many sources of variation and inference, both of which are obvious and, In most cases, the common Markov Chain Monte Carlo (MCMC) technique is computationally practical. When the target variable is quantified using this method, other model specifications can be examined. The posterior mean of the target parameter is used to estimate it, while the posterior variance is used to establish its precision. The posterior distribution is used to make the inference. The Bayesian model approach to the count data is applied to

Manuscript received December 12, 2022; revised August 14, 2023.

I Wulandari is a doctoral student at IPB University. She is a lecturer at Politeknik Statistika STIS, Indonesia (e-mail: ita.wulandari@stis.ac.id).

K. A. Notodiputro is a professor in Statistics and Data Science at Department of Statistics, IPB University, 16680, Indonesia. (e-mail: khairil@apps.ipb.ac.id).

A. Fitrianto is a lecturer of Statistics and Data Science at Department of Statistics, IPB University, 16680, Indonesia. (e-mail: anwarfitrianto@apps.ipb.ac.id)

A. Kurnia is an associate professor in Statistics and Data Science at Department of Statistics, IPB University, 16680, Indonesia. (e-mail: anangk@apps.ipb.ac.id).

estimate the mortality rate. For example, an alternative model for estimating the mortality rate according to specific age variables in one category of cancer in the United States has also been carried out by Nandram *et al*. [15]. They utilized the Bayesian approach to test four alternative models. Each model assumes that fatalities in a specific area and age group are Poisson distributed. The results showed that their proposed model could capture small area and regional effects well and detect residual spatial correlations, thus facilitating the parameter estimation. Nandram *et al*. [16] also performed a model similar to what has been done previously. However, the model was applied to estimate the mortality rate according to specific age variables in one chronic obstructive pulmonary disease category in the United States. Trevisani and Torelli [17] developed an SAE for the count data using the HB approach.

A likelihood approach for estimating parameters in small area modeling for a child mortality estimate has been carried out [18]. This study applied a model to overcome equidispersion violations in the Poisson Regression model. The models used were Poisson, quasi-Poisson, and ZIP models. The quasi-Poisson model produces better predictions than the other models.

This study investigated how to estimate model parameters and predictions with a high precision and accuracy while considering the overdispersion issue, the Poisson distribution, and the non-normal data distribution. We used three models to address the overdispersion issue brought on by an excess of zeros. The zero-inflated generalized Poisson (ZIGP) model, the zero-inflated negative binomial (ZINB) model, and the Poisson Regression model are compared, and the goal is to improve the fit of the model. We hope that the three models can manage overdispersion and zero inflation. In other words, when the overall dispersion parameter or zero inflation is insufficient, the model we apply can handle the problem. In contrast to Istiana *et al*. [18], we decided that the HB technique was the best option for parameter estimation.

The statistical model would be applied in estimating the under-five mortality rate (U5MR) Using data from the 2017 Indonesia Demographic and Health Survey (IDHS), at the district level on Java Island. We considered the specifications of the relevant HB model. for this case, considering the properties of the quantity to be estimated, i.e., infant mortality rate. In addition, the model that we would show was the best version of the three models. This model was undoubtedly a model that could handle overdispersion in the count data that we used.

The following is how the paper is structured: Section 2 introduces the Poisson mixed model at the area level and details the SAE for zero-inflated data; Section 3 explains the hierarchical Bayesian technique; and Section 4 reports on the real data application; and Section 5 concludes with a brief discussion.

II. The Area-level Poisson Mixed Model

We introduced the Poisson mixed model at the area level in this part. Let's suppose that there is a finite population that contains elements, and each subpopulation or domain is further subdivided into samples that contain elements according to some sampling strategies. The sample size and population size in the *i*-th area are represented by n_i and N_i. The observed value of the response variable for the *i*-th area is denoted by y_i. The explanatory variables are assumed to be available for each area level with area-specific data vectors denoted as $\mathbf{x} = \left(x_i, ..., x_D\right)'$. Let v_d be a set of random effects, in matrix notation, we have $v = \left(v_i, ..., v_D\right)' \sim N_D\left(0, \mathbf{I}_D\right)$ where $\boldsymbol{I}_D$ is the unit $D \times D$ unit matrix. We assume that the target variable's distribution y_i conditionally on the random effect v_i is

$$y_i | v_i \sim Poisson\left(\mu_i\right), i = 1, ..., D \tag{1}$$

Within the framework of Poisson Regression, it is assumed that variance is proportional to the mean in terms of the statistical model, specifically $\text{var}\left(Y\right) = \phi E\left(Y\right) = \phi\mu$. If $\phi = 1$, the variance equals the mean, the Poisson Regression is used. If $\phi > 1$, the model has over-dispersion in comparison to Poisson. If $\phi < 1$, we would have under-dispersion, although this is uncommon.

The empirical data frequently exhibit more zeros than anticipated under either model, a phenomenon known as zero-inflation, which is another frequent issue with count data models, including the Poisson Regression model. This model has two different types of zeroes: random zeroes from one class and structural zeroes from the always-zero class. Frequentist and Bayesian approaches were developed for this sort of data by Chandra and Sud [19] and Pfeffermann *et al*. [7]. Kreig *et al*. [8] considered both approaches in their research. Instead of utilizing simulated and actual data, the EBLUP estimator and the two small area estimators are constructed using models that explicitly account for zero-inflation. The results indicated that when dealing with variables with zero-inflation, all of the SAE estimators outperformed the design-based techniques in terms of accuracy. Nevertheless, the outcomes of the two procedures were nearly identical. The downside of the Bayesian technique is that the computing time is longer, however, the benefit is that information on forecast accuracy follows directly. Thus yet, no formula for the mean squared error has been found for the frequentist method.

III. SAE for Zero-inflated Data

The ZIGP distribution is defined in the same way as the ZIP distribution, with the addition of a zero-inflated parameter. Recently, the ZIGP regression models have proven to be beneficial for assessing count data with a high percentage of zeros [20]. This model has three parameters and will be denoted by $ZIGP\left(\mu, \varphi, \omega\right)$. One of its key advantages is this model's ability to accommodate overdispersion in two ways: a zero-inflation parameter φ and an additional overdispersion parameter ω. This can be demonstrated by reducing it to Poisson Regression when $\omega = 1$ and $\omega = 0$ are present.

We denote the ZIGP regression model with response y_i, auxiliary variables $\mathbf{x} = (x_i, ..., x_D)'$, $\omega = (\omega_i, ..., \omega_D)'$ for overdispersion and $z_i = (1, z_1, ..., z_D)'$ for zero-inflation is defined as

$$P(Y = y_i | x_i, z_i) = \varphi_i + (1-\varphi_i) f(\mu_i, \omega, y_i), y_i = 0$$
$$= (1-\varphi_i) f(\mu_i, \omega, y_i), y_i > 0 \quad (2)$$

where $f(\mu_i, \omega, y_i), y_i = 0,1,2,...$ is the GP model as follows:

$$f(\mu_i, \omega, y_i) = \left(\frac{\mu_i}{1+\omega\mu_i}\right)^{y_i} \frac{(1+\omega y_i)^{y_i-1}}{y_i!} \times \exp\left[\frac{-\mu_i(1+\omega y_i)}{1+\omega\mu_i}\right] \quad (3)$$

and $0 < \varphi_i < 1$. The function $\mu_i = \mu_i(x_i)$ and $\varphi_i = \varphi_i(z_i)$ satisfy $\log(\mu_i) = \sum_{i=1}^{D} x_i\beta_i + v_i$ and $\log it(\varphi_i) = \log(\varphi_i[1-\varphi_i])^{-1} = \sum_{i=1}^{D} z_i\gamma_i + u_i$. The mean and variance are given, respectively by

$E(y_i|x_i) = (1-\varphi_i)\mu_i(x_i)$ and

$V(y_i|x_i) = E(y_i|x_i)\left[(1+\omega y_i)^2 + \varphi_i\mu_i\right]$

The distribution of y_i exhibits overdispersion when $\varphi_i > 0$. The ZINB regression model is derived from the Poisson Gamma mixed distribution. This distribution was chosen because the probability is simple to compute; nevertheless, this simplicity does not ensure a good fit. [21]. It is acceptable to use this model to model count data or discrete data that contains large numbers of zero values in the response variable (zero-inflation) because this model uses a large number of zero values in the response variable, thus reducing the chances of an overdispersion problem [22].

In the ZINB regression, there are two states. The first is a zero-valued state, whereas the second is a negative binomial state. The ZINB regression model's probability function is expressed as follows:

$$P(Y = y_i | x_i, z_i) = \varphi_i + (1-\varphi_i) g(\mu_i, \omega, y_i), y_i = 0$$
$$= (1-\varphi_i) g(\mu_i, \omega, y_i), y_i > 0 \quad (4)$$

where $0 \le \varphi_i \le 1, \mu_i \ge 0, \omega$ is the dispersion parameter and $(.)$ is the gamma function, $g(\mu_i, \omega, y_i)$ is the probability function of the negative binomial distribution

$$g(\mu_i, \omega, y_i) = \frac{\Gamma(y + \omega^{-1})}{\Gamma(\omega^{-1})\Gamma(y_i+1)}\left(\frac{1}{1+\omega\mu_i}\right)^{\omega^{-1}} \times \left(\frac{\omega\mu_i}{1+\omega\mu_i}\right)^{y_i} \quad (5)$$

IV. Hierarchical Bayes Estimator

There are several ways for estimating small area models, one of which is to employ hierarchical HB models. This difficulty is addressed in this work by using HB modles when small area estimates consist of counts. Bayesian specifications are derived form traditional models for the SAE, such as the Fay-Herriot model or, more specifically, a generalized kinear Poisson Regression model.

The Bayesian theorem approach in statistical inference has recently been very developed. In contrast to traditional statistical theory, Bayesian statistics consider all unknown parameters to be random variables. In Bayesian theorem, estimation is done by considering and combining information from both the sample and other available information.

In general, the Fay-Herriot model using the HB method produces the following specifications:

$$\hat{\theta}_i | \theta_i, \sigma_i \sim N(\theta_i, \sigma_i) \quad (6)$$

$$\theta_i | \beta, \tau \sim N(\mathbf{x}'_i\beta, \tau) \quad (7)$$

$$\beta, \tau \sim p(\beta, \tau) \quad (8)$$

For each area i, with $\theta_i, \hat{\theta}_i$ and x_i identifying the features of interest, survey estimate (where available), and possible supplementary data. Linking models have mixed coefficients, which means that they consist of fixed coefficients β, which account for the effects that are applied to the entire population, and random coefficients v_i which account for the effects that are applied to individual areas. It is important to estimate the parameters β and τ as sampling variances, σ_i which are normally considered to be known. Sampling models (6) and linking model (7) remain intact compared to the different hyperprior stages required in a full HB strategy.

In order to perform a full Bayesian analysis, it would be appropriate to use a prior distribution that is sufficiently informative about the hyperparameters. We frequently utilize noninformative priors when we need to get more informed or want the conclusion entirely based on the available data. Priors are typically thought to prevent posterior density from being incorrect and dispersed but precise (otherwise said, less informative). A reasonable conclusion is guaranteed by such a choice, which, nonetheless, requires a careful study, mainly when models are hardly known generally.

The MCMC algorithms used in the HB approach make inference simple and computationally practical. Therefore, more realistic models (i.e. generalized linear models) for the SAE problems [9] are also made much more feasible within the HB approach than the alternative methods.

Our research yields domain mean estimates as MCMC approximations of the posterior means, such as

$$\hat{y}_{i,mcmc} = \frac{1}{r}\sum_{R-1}^{r} y^*_{i.R} \quad (9)$$

The posterior predictive as follows

$$y^*_{i.R} = \frac{1}{N_i}\sum_{i=1}^{N_i} \hat{y}^*_{i.R}\delta^*_{i.R} \quad (10)$$

where $\delta^*_{i.R}$ is the Bernoulli distribution for the The ZIP model and the ZINB model's negative binomial distribution. We define

$$\delta_i = \begin{cases} 1, y_i \neq 0 \\ 0, y_i = 0 \end{cases} \quad (11)$$

The first model in the estimator above is a linear mixed model, describing the distribution of the non-zero target variable, $\hat{y}^*_{i.R}$. Furthermore is the general linear mixed model for the binary zero indicators, $\hat{p}^*_{i.R}$. We define nz as the nonzero portion of the population or sample in the model. The subscript nz is used to denote the nonzero part of the population or sample.

$$\hat{y}^*_{i.R} = x^{'}_{nz.i}\beta_{nz.R} + v_{nz.i.R} + e_{i.R}\delta^*_{i.R} \quad (12)$$

and

$$\hat{p}_{i.R} = \frac{\exp\left(x^{'}_{z.i}\beta_{z.R} + v_{z.R}\right)}{1+\exp\left(x^{'}_{z.i}\beta_{z.R} + v_{z.R}\right)} \quad (13)$$

Both models resulting in the estimates $\hat{\beta}_{nz.R,}\hat{v}_{nz.R}, \hat{\beta}_{z.R,}\hat{v}_{z.R}, \hat{\sigma}_{v.nz.R}, \hat{\sigma}_{e.nz.R}, \hat{\sigma}_{v.z.R}$.

The approximated mean square error of the model, $\hat{y}_{i,mcmc}$ takes the form

$$mse\left(\hat{y}_{i,mcmc}\right) = \frac{1}{r}\sum\nolimits_{R-1}^{r}\left(y^*_{i.R} - \hat{y}_{i,mcmc}\right) \quad (14)$$

In all models, we have taken iteration $R = 1,...,r, r = 10.000$ swith a burning of 600 (default) and thinning by retaining each 10th iteration.

V. Under-five Mortality Rate Analysis

This paper presents the application of the SAE model for Poisson distributed data with overdispersion problems. This model estimates the U5MR in provinces of Java Island (the Special District of Jakarta, West Java, Central Java, the Special District of Yogyakarta, East Java, and Banten) using data from the 2017 IDHS. There were 119 districts, with six districts that were not sampled. The total figure of the under-five in this study was 25,339 children from 49,627 women of childbearing age in 47,963 households.

The U5MR is an indicator that is directly related to the child survival target and reflects the social, economic, and environmental conditions in which children live including their health care. Conceptually, the U5MR is the number of deaths of children aged 0-4 years (0-59 months) in a given year per 1000 children of the same age in the middle of the same year (including infant deaths). This figure is also one of the 100 primary health indicators in the World Health Organization's (WHO) Global Reference List and is the third goal of the Sustainable Development Goals (SDGs). This indicator connects universally-known objectives for children's rights and general development standards. As stated by the WHO, this indicator is significant as it provides a baseline to measure how a nation is doing concerning granting children's rights, particularly those to life, health care, nourishment, water, social security, and protection.

The indicator (U5MR) estimates were obtained from the 2017 IDHS data in the Special Capital District of Jakarta, West Java, Central Java, the Special District of Yogyakarta, East Java, and Banten. Not all of the districts (119) had the same sample size sufficient to estimate the U5MR at the district level. This is because there were districts with a very small sample of children under five, and even six districts were not selected as examples. For this reason, if a direct U5MR estimation was carried out at the district level, it would produce an estimate with a large error.

The calculation of U5MR estimates used seven variables. There was the century month code for the date of birth of the child (b3, the year when the survey was fielded (year), the variable indicating the primary sampling unit (v021), a weighting factor to produce a representative estimate (v005), the age of the child at death in months (b7), century month code for the date on which the interview took place (v008), the relative wealth of the household where the woman lived, divided into quintiles from the poorest to the richest (v190). The data used were obtained from individual women's data (IDIR), the 2017 IDHS.

First of all, we explored data. Of 113 districts with a sample, 44 districts (38.94%) had a U5MR estimate of zero from districts in Java Island. This does not mean there were no under-five deaths in the districts, but it could be due to the small sample size. The percentage of zero value in the direct estimation of the quite large U5MR indicated the presence of an excess of zero which is one of causes of overdispersion in the Poisson Regression model. For this reason, we carried out statistical tests to detect these problems. As a result, we obtained a p-value of 2.22e-16. This means that the SAE model in the Poisson Regression could not handle an excess of zero values (or zero-inflation).

Under this investigation, alternative models ZIGP and ZINB would be used. The ability of this class of regression models to manage overdispersion and zero-inflated is what piqued our interest in them. Here, we permitted regression on the overdispersion and zero-inflation factors in addition to the mean. In situations in which the general dispersion or zero-inflation parameter is insufficient, the goal is to increase the model's fit. For all the three models, we applied a HB method (Poisson, ZIGP, and ZINB).

This section estimates U5MR by district, excluding the six districts not sampled. Three models (Poisson, ZINB and ZIGP) were implemented using several auxiliary variables. These auxiliary variables include the density of health facilities, health centers, and health workers. From the census data, three variables were obtained, the 2018 *Potensi Desa* (PODES). This data was the only village-level database conducted two years prior to the Indonesian population census.

The stationarity of the selected posterior distribution and whether the Markov Chain has reached must be determined using the HB inference. Convergence diagnostics were applied to carry out this process by taking a trace plot or other pertinent statistical measurements. If the distribution of a Markov Chain's points does not alter along with the Markov Chain, the chain is said to be stationary. In this instance, it is visible through the trace plot, which, compared

to the probability density plot and the autocorrelation function plot between cases, is quite constant between mean and variance. This requirement was met in the study.

We used direct estimates of the domain means as responses in area-level models in three different models based on the direct estimates of the domain means. A comparison of the direct and U5MR estimates for the Poisson mixed model, the generalized Poisson mixed model, and the mixed model area-level zero-level negative binomial is shown in Figure 1. We observed that three estimators had almost the same pattern. We needed to find the best model for the U5MR prediction in this condition.

We also compared the RMSE values (Fig. 2). Considering the previous results that the Poisson model has an excess of zero, or in other words, there is an overdispersion problem, we would only compare the two models, namely ZINB and ZIGP. The RMSE values in the ZINB model are below the EMSE values in the ZIGP model. Furthermore, if seen from the average value, the average RMSE of the ZINB model is smaller at 14.65, while the average RMSE of the ZIGP model is 17.38.

Model evaluation is done by looking at the relative standard error (RSE) and root mean squared error (RMSE) values. For simplicity, we used a notation 1-113 for districts. Figure 4 shows a plot of the RSE values of the three models, Poisson (top), ZIGP (center), and ZINB (bottom), against the direct estimator. The Poisson model's RSE value has a larger range than the other models. When comparing the ZINB and ZIGP models, both have narrower ranges than the Poisson model. However, if we look more closely, the ZINB model has a narrower range of 17.62-67.68 and a range of 39.89-144.12 for the ZIGP model. This result was reinforced by the average RSE value of the ZINB model and the ZIGP model, respectively, which were 37.62 and 79.19, in other words, the ZINB model is better. The diversity of the ZINB model also can be considered smaller than the other models.

TABLE I
SUMMARY MODEL ZINB

Parameter	Estimate	Std. Error	z value	Pr (>|z|)	
Conditional model:					
β_0	3.4468	0.1948	19.7150	<2e-16	****
β_1	0.0008	0.0038	0.2170	0.8280	
β_2	0.0101	0.0090	1.1290	0.2590	
β_3	-0.0018	0.0216	-0.0850	0.9320	
Zero-inflation model:					
γ_0	-1.2624	0.5188	-2.4330	0.0150	**
γ_1	0.0141	0.0081	1.5660	0.1174	
γ_2	-0.0876	0.0431	-2.0350	0.0419	**
γ_3	0.0829	0.0437	1.8980	0.0576	*

In our study, the Poisson Regression model could not be utilized to estimate the U5MR. This is because the model had a dispersion problem and zero overload occurred. Therefore, we applied two alternative models to deal with the problem. As we explained earlier, we explored the ZIGP and ZINB models to obtain the best model. Both models actually could deal with the problem of overdispersion. However we selected one model for us to use in estimating unsampled districts. The small RMSE and RSE values were used as the basis for selecting the best model, and the ZINB model was selected.

Table I presents the proposed ZINB model's calculated regression parameters. It also contains the appropriate p-value. We observe this for the zero-inflation ratio model, the health center is negatively related to the U5MR, which means that each additional one percent of health center facilities will reduce the probability of the number of the under-five deaths by 1.0864 assuming other variables are constant. While, health workers have a positive effect on the U5MR in Java.

TABLE II
PREDICTED OF U5MR FOR NIRSAMPLE DISTRICS IN JAVA

Districts/ Municipality	$\hat{y}_{nir.ZINB}$	Lower bounds	Upper bounds	RMSE	RSE
Kepulauan Seribu	44.27	25.86	77.16	13.24	29.91
Pangandaran	42.41	25.09	70.00	11.51	27.14
Banjar	34.32	13.57	71.38	15.20	44.29
Probolinggo	41.00	22.21	60.26	9.34	22.78
Madiun	59.80	33.30	87.68	13.72	22.94
Batu	36.53	21.45	53.51	8.19	22.42

We consider the U5MR predictions using the best model, namely ZINB. Table II presents the U5MR predictions with a 95% confidence interval and the values of RMSE and RSE. Six districts estimated are Kepulauan Seribu, Pangandaran, Banjar, Probolinggo, Madiun, and Batu. The length of the confidence interval indicates the accuracy of our estimation. If the range is narrow, the margin of error is small, which means that the estimate obtained is between reasonable values, or that this estimate is correct. However, if the interval is wide and the margin of error is significant, the final estimate is less accurate. Banjar has a wide confidence interval compared to the other five districts, ranging from 13.5686 to 71.3773. This value means the under-five mortality rate in Banjar ranges from 14 to 72 for every 1,000 live births. In addition to having a wide confidence interval, Banjar has a RSE value of 44.2953, which requires anyone to be careful in using the estimated results obtained. However, the overall RSE value of the six districts, none of them is greater than 50, which means it is still acceptable.

We only show models addressing the overdispersion problem: the ZIGP and ZINB models in the 113 sample districts (See TABLE III). It is important to note that the RMSE and RSE values from the ZINB model are smaller than the direct estimates, as well as the Poisson Regression and the ZIGP models. Furthermore, apart from dealing with the overdispersion problem, the ZINB model can better estimate the U5MR. The RSE value measures the feasibility of using the resulting data. As much as 37.1681% of 113 districts have $RSE < 25\%$, which means that the U5MR estimate is accurate, or in other words, the results can be interpreted very well. Then, 59.2920% of 113 districts have $25\% \le RSE \le 50\%$, which means that a caution is

needed in using the estimated results obtained, the remaining four districts have $RSE > 50\%$, which means that the estimation results obtained are not accurate.

Figure 3 shows that districts in Java with the lowest U5MR (in white) are mostly spread across the top. Most districts in Java have values in the range of 32 to 53.

VI. Conclusion

In this study, we have explored the response variable, the U5MR data. It was shown that the data experienced overdispersion, in particular excess zero, so that the Poisson model could not be applied. Therefore, we apply alternative models that are ZIGP and ZINB to solve the problem. Furthermore, we discovered that our proposed model enhances the accuracy of both direct estimates and the Poisson model. In other words, both models can overcome overdispersion. However, when compared to all other models, the ZINB model performs the best, as evidenced by decreasing RSE and RMSE values.

The auxiliary variable in the present study is assumed to measure without error. However, if the error is not accounted for in the model, the results may be worse than the direct estimator, or the resulting parameter estimator may be biased. For this reason, auxiliary variables measured with errors can be used as materials for further research. As a result, in our next study, we will investigate the structure of an SAE model for data generated while taking into account the overdispersion issue in the response variable, and the auxiliary variable is supposed to be measured with error. We use data sources from multiple surveys (census, surveys, and administrative data), the census of the population, the IDHS, and *Potensi Desa*. Accordingly, the question for further research is on how to obtain the best estimator if it considers measurement errors in both the response and explanatory variables and the multiple surveys used in the study.

References

[1] D. Pfeffermann, “Small area estimation: new developments and directions". International Statistical Review.70. pp.125-143, 2002.

[2] D. Lambert, “Zero-inflated poisson regression, with an application to defects in manufacturing,”. *Technometrics*, vol. 34, no. 1, pp. 1–14, 1992.

[3] D. Fletcher, D.MacKenzie, and E. Villontra. "Modeling skewed data with many zeros: A simple approach combining ordinary and logistic regression." *Journal of Environmental and Ecological Statistics.* 12:45-54, 2005.

[4] H. Chandra, and U. C. Sud, “Small area estimation for zero-inflated data,” *Commun. Stat. Simul. Comput.*, vol. 41, no. 5, pp. 632–643, 2012.

[5] I. Ginanjar, M. Iaeng, S. Wulandary, and T. Toharudin, “Empirical best linear unbiased prediction method with K-medoids cluster for estimate per capita expenditure of sub-district level,” *IAENG Int. J. Appl. Math.*, vol. 52, no. 3, 610-616, 2022.

[6] H. Chandra, and R. Chambers. "Multipurpose weighting for small area estimation." *Journal of Official Statistics.* 25:379-395, 2009.

[7] D. Pfeffermann, B. Terryn, and F. A. S. Mours. "Small area estimation under a two-part random effects model with application to estimation of literacy in developing countries.' S*urvey Methodology.* Vol. 34.No.2. pp. 235-249, 2008.

[8] S. Krieg, H. J. Boonstra, and M. Smeets, “Small-area estimation with zero-inflated data – a simulation study,” *J. Off. Stat.*, vol. 32, no. 4, pp. 963–986, 2016.

[9] M. Ghosh, K. Natarajan, T. W. F. Stround, and B.P. Carlin. "Generalized linear models for small area estimation." *Journal of the Aerican Statistical Association.* 93.273-282, 1998.

[10] M. Ghosh, and T. Maiti. "Adjusted bayes estimators with applications to small area estimation." *Sankhya: The Indian Journal of Statistics.* 61. 71-90, 1999.

[11] M. Ghosh and T. Maiti. " Empirical bayes confidence intervals for means of natural exponential family-quadratic variance function distributions with application to small area estimation." *Scandinavian Journal of Statistics.* 35. 484-495, 2008.

[12] G. S. Datta, P. Lahiri, and K. L. Lu. "Hierarchical bayes estimation of unemployment rates for the states of the U.S." *Journal of the American Statsitical Association, 984.* 1074-1082, 1999.

[13] G. S. Datta, M. Ghosh, and l. A. Waller. "Hierarchical and empirical bayes methods for environmental risk assessment." *Handbook of Statistics. Vpl. 18. Amsterdam: Elsevier Science.* pp. 223-245, 2000.

[14] M. Torabi, and J. N. K. Rao, “Small area estimation under a two-level model." *Survei Methodology, 34.* 11-17, 2008.

[15] B. Nandram, J. Sendranskn, and L. Pickle."Bayesian analysis of mortality rates for U.S. health service areas." *Sankhya: The Indoan Journal of Statistics, 61.* 146-165, 1999.

[16] B. Nandram, J Sendransk, and l. W. Pickle."Bayesian analysis and mapping of mortality rates for chronic abstractive pulmonary disease." *Journal of the American Statistical Association, 95* 1110-1118, 2000.

[17] M. Trevisani and N. Torelli."Hierarchical Bayesian models for small area estimation with count data." *Working paper: Diparimento di Science Economiche e Statistiche.* Universita degli Studi di Trieste, Italy, 2007.

[18] N. Istiana, A. Kurnia, and A. Ubaidillah. Quasi Poisson model for estimating under-five mortality rate in small area. EAI Conference Proceedings of the 1st International Conference on Statistics and Analytics, ICSA 2019, 2-3 August 2019, Bogor, Indonesia, 2020.

[19] H. Chandra, and U Sud. “Small area estimation for zero-inflated data.” Communications in Statistics - Simulation and Computation. 41:5, 632-643, 2012.

[20] F. Famoye, and K. P. Singh. Zero-infated generalized Poisson model with an application to domestic violence data. Journal of Data Science 4 (1), 117-130, 2006.

[21] P. Hougaard, M.-L.T. Lee, G.A. Whitmore, Analysis of overdispersed count data by mixtures of Poisson variables and Poisson processes, Biometrics 53. 1997. 1225–1238, 1997.

[22] A. Garay, M. Hashimoto, E. M. Ortega, and V. H. Lachos, On estimation and influence diagnostics for zero inflated negative binomial regression model. *Computational Statistics and Data Analysis, 55,* 1304-1318, 2011.

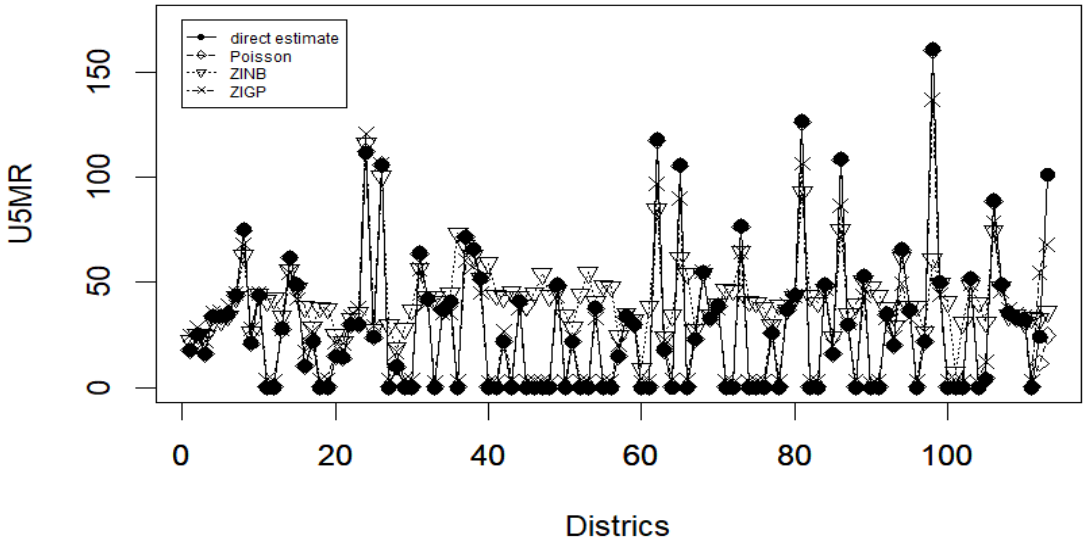

Fig. 1. Direct Estimate $\left(\hat{y}^{dir}\right)$ and $\hat{y}$ for three estimators, Poisson, ZIGP, and ZINB

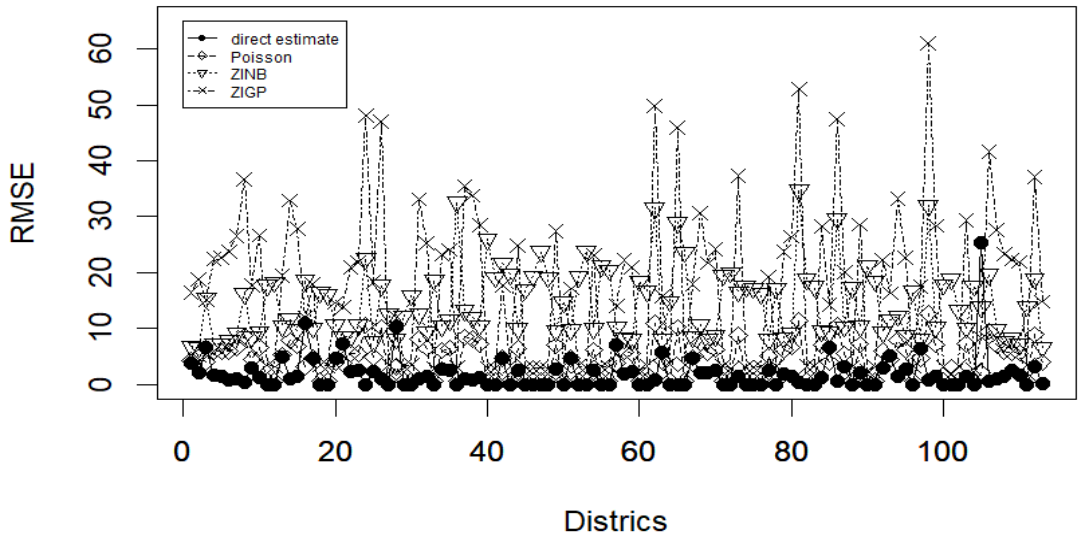

Fig. 2. RMSE plot for direct against three estimators, Poisson, ZIGP, and ZINB

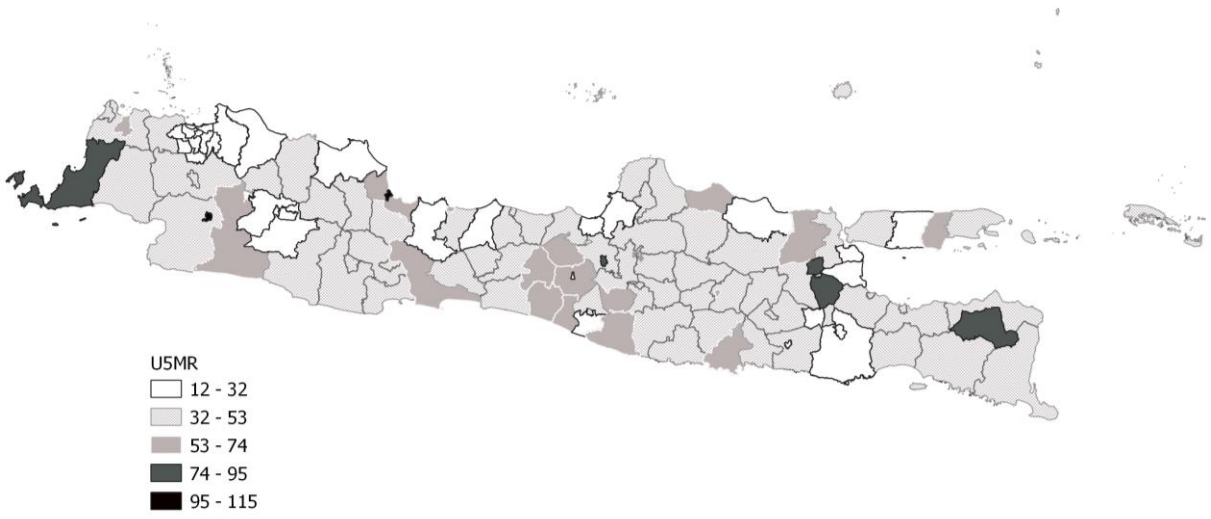

Fig. 4. The under-five mortality rate estimated mapping of districts in Java

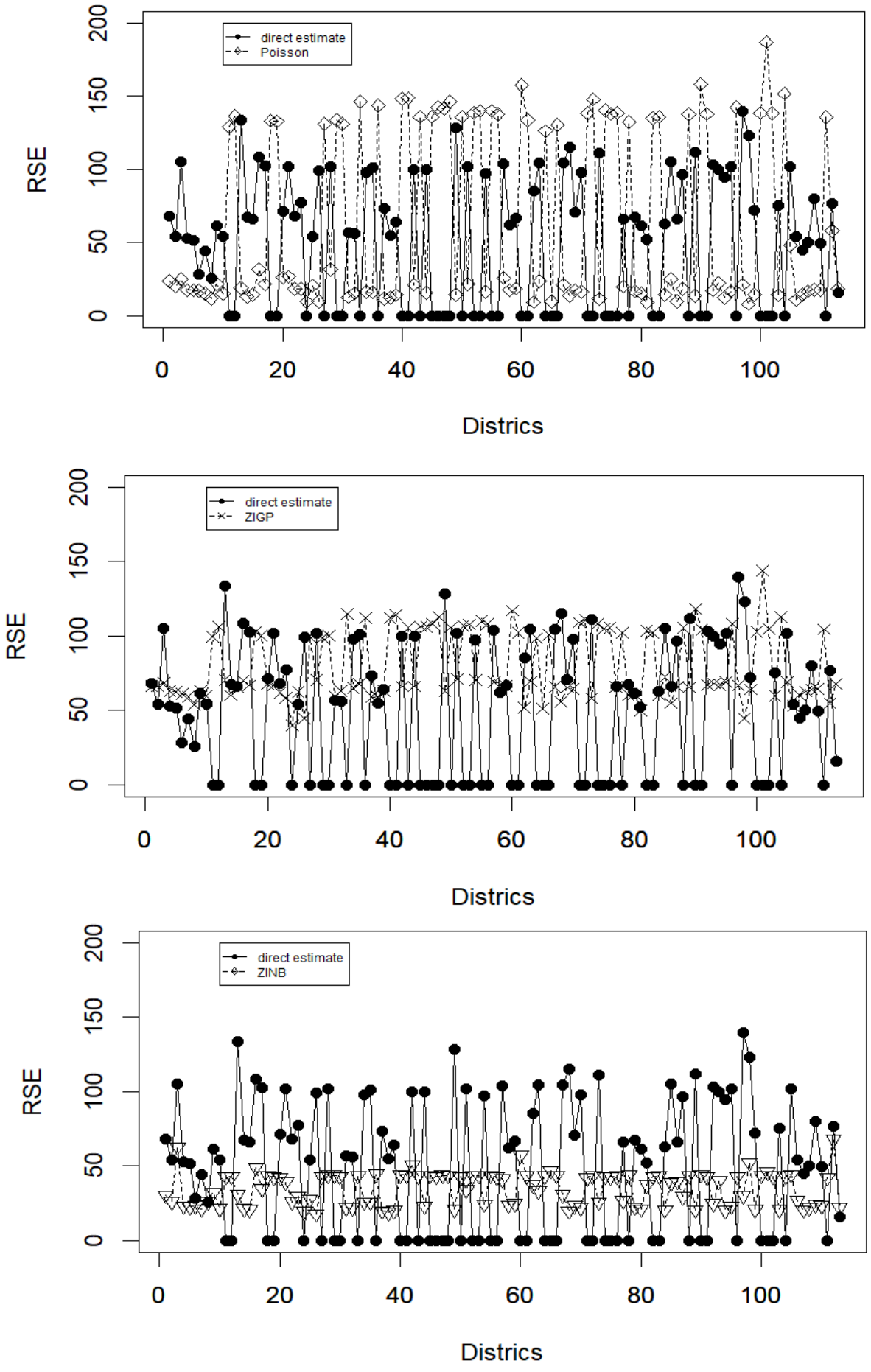

Fig. 3. RSE plot for direct against three estimators, Poisson (top), ZIGP (middle), ZINB (bottom)

TABLE III
PREDICTED OF U5MR WITH CORRESPONDING RSE OBTAINED WITH ZIGP AND ZINB MODEL

Districts	ZIGP			ZINB			Districts	ZIGP			ZINB		
	$\hat{y}$	RSE	RMSE		RSE	RMSE		$\hat{y}$	RSE	RMSE	$\hat{y}$	RSE	RMSE
Jakarta Selatan	24.73	66.19	16.37	22.19	30.22	6.71	Grobogan	2.82	106.53	3.00	39.05	43.48	16.98
Jakarta Timur	28.47	66.02	18.80	25.39	25.96	6.59	Blora	2.84	106.20	3.01	45.06	42.97	19.37
Jakarta Pusat	21.97	68.44	15.04	24.61	62.72	15.43	Rembang	2.78	108.14	3.01	54.09	43.80	23.69
Jakarta Barat	35.59	62.84	22.37	31.35	23.00	7.21	Pati	2.41	112.57	2.71	43.79	43.62	19.10
Jakarta Utara	36.32	62.27	22.61	31.73	22.74	7.22	Kudus	43.49	62.99	27.40	45.99	20.73	9.53
Bogor	38.80	61.16	23.73	33.99	23.14	7.87	Jepara	3.00	105.84	3.18	34.59	42.59	14.73
Sukabumi	44.68	59.27	26.48	42.94	21.07	9.05	Demak	24.10	71.14	17.15	28.52	34.06	9.72
Cianjur	68.25	53.78	36.70	62.88	25.78	16.21	Semarang	3.03	106.59	3.23	44.57	43.25	19.28
Bandung	27.77	63.44	17.62	26.38	32.16	8.48	Temanggung	2.81	107.89	3.03	54.73	43.60	23.86
Garut	44.60	59.76	26.65	44.14	21.37	9.43	Kendal	32.77	70.58	23.13	40.76	24.42	9.96
Tasikmalaya	3.39	99.32	3.37	41.31	42.83	17.69	Batang	2.83	110.27	3.12	48.55	43.51	21.12
Ciamis	2.95	106.1	3.13	42.60	42.86	18.26	Pekalongan	2.91	107.92	3.14	48.09	42.41	20.40
Kuningan	27.71	70.29	19.48	33.83	30.66	10.37	Pemalang	20.35	69.12	14.07	24.49	41.41	10.14
Cirebon	54.77	60.02	32.87	55.79	21.15	11.80	Tegal	33.09	67.14	22.21	35.21	23.53	8.28
Majalengka	42.14	65.85	27.75	47.25	20.49	9.68	Brebes	31.42	66.49	20.89	32.31	24.70	7.98
Sumedang	16.97	69.98	11.88	38.24	48.93	18.71	Magelang	2.19	117.10	2.57	8.83	209.22	18.47
Indramayu	26.75	66.74	17.85	28.85	34.79	10.04	Surakarta	3.17	101.94	3.23	38.32	43.40	16.63
Subang	3.31	102.8	3.40	38.17	43.03	16.42	Salatiga	96.81	51.52	49.88	84.69	37.19	31.50
Purwakarta	3.32	100.3	3.33	37.43	42.60	15.95	Semarang	22.71	69.36	15.75	23.88	33.35	7.96
Karawang	21.46	67.85	14.56	25.08	42.20	10.58	Pekalongan	3.70	98.49	3.64	34.17	43.21	14.77
Bekasi	20.70	66.85	13.84	21.89	39.20	8.58	Tegal	89.94	51.00	45.87	61.90	46.78	28.96
Bandung Barat	33.87	61.89	20.96	32.55	25.81	8.40	Kulon Progo	3.36	100.68	3.38	54.35	43.56	23.67
Bogor	37.55	58.25	21.87	35.72	29.59	10.57	Bantul	27.28	65.46	17.85	27.50	30.56	8.40
Sukabumi	120.6	39.89	48.13	116.0	19.38	22.48	Gunungkidul	55.19	55.55	30.66	54.28	19.58	10.63
Bandung	29.32	62.53	18.33	27.60	27.61	7.62	Sleman	33.14	65.38	21.67	34.53	23.30	8.05
Cirebon	106.1	44.27	46.98	100.5	17.62	17.71	Yogyakarta	37.45	64.38	24.11	39.41	21.91	8.63
Bekasi	3.44	98.99	3.40	29.56	42.78	12.65	Pacitan	2.91	109.19	3.18	46.79	41.83	19.58
Depok	16.98	70.20	11.92	18.51	43.96	8.14	Ponorogo	2.56	110.87	2.84	46.12	43.38	20.01
Cimahi	3.34	99.08	3.31	27.96	43.79	12.24	Trenggalek	64.25	58.06	37.30	64.52	25.65	16.55
Tasikmalaya	3.42	100.2	3.43	36.98	43.02	15.91	Tulungagung	2.75	108.05	2.97	41.03	42.78	17.55
Cilacap	56.27	58.83	33.10	56.78	22.06	12.52	Blitar	2.91	104.62	3.05	40.42	42.58	17.21
Banyumas	40.02	63.18	25.29	43.21	21.87	9.45	Kediri	2.98	105.28	3.14	37.71	43.11	16.26
Purbalingga	2.52	115.0	2.89	43.04	43.47	18.71	Malang	29.21	65.52	19.14	29.68	26.80	7.96
Banjarnegara	35.60	65.14	23.19	40.65	25.51	10.37	Lumajang	3.18	102.25	3.26	39.23	43.76	17.17
Kebumen	35.55	67.65	24.05	44.66	25.52	11.40	Jember	39.40	60.32	23.77	37.01	22.29	8.25
Purworejo	2.71	112.0	3.04	73.60	44.35	32.64	Banyuwangi	43.50	60.78	26.44	43.19	21.02	9.08
Wonosobo	60.18	58.98	35.49	67.34	19.66	13.24	Bondowoso	106.5	49.59	52.80	93.03	37.40	34.80
Magelang	57.25	58.93	33.74	63.10	18.84	11.89	Situbondo	3.14	103.31	3.25	43.66	42.92	18.74
Boyolali	45.02	63.18	28.44	52.14	20.14	10.50	Probolinggo	3.13	102.03	3.19	40.44	43.14	17.45
Klaten	2.57	111.9	2.88	59.12	43.91	25.96	Pasuruan	46.73	60.31	28.18	46.73	20.37	9.52
Sukoharjo	2.44	114.2	2.79	43.84	43.55	19.09	Sidoarjo	19.63	72.17	14.16	24.10	38.71	9.33
Wonogiri	26.90	66.01	17.76	43.07	50.40	21.71	Mojokerto	86.65	54.74	47.44	75.12	39.51	29.68

Karanganyar	3.09	105.50	3.26	45.54	43.24	19.69	Jombang	30.56	65.88	20.14	35.11	29.27	10.28
Sragen	37.49	99.67	24.78	43.20	22.98	9.93	Nganjuk	2.98	105.70	3.15	39.67	43.67	17.33
Madiun	43.69	65.37	28.56	50.73	20.38	10.34	Malang	3.04	104.69	3.18	30.76	43.18	13.28
Magetan	1.95	118.14	2.30	47.84	44.21	21.15	Pasuruan	49.11	59.80	29.37	48.60	20.66	10.04
Ngawi	3.01	104.37	3.15	43.63	42.37	18.49	Mojokerto	2.39	112.89	2.70	40.52	43.41	17.59
Bojonegoro	33.06	67.20	22.22	37.99	24.58	9.34	Surabaya	12.04	69.03	8.31	31.15	44.32	13.81
Tuban	23.97	68.02	16.30	29.09	39.81	11.58	Pandeglang	78.26	53.14	41.59	74.31	26.47	19.67
Lamongan	49.60	66.87	33.17	61.59	19.86	12.23	Lebak	45.93	60.26	27.68	48.08	20.37	9.79
Gresik	32.88	69.05	22.70	38.37	22.90	8.79	Tangerang	37.27	62.72	23.37	34.33	22.09	7.58
Bangkalan	2.99	108.18	3.24	38.60	43.01	16.60	Serang	34.86	64.04	22.32	34.60	23.93	8.28
Sampang	26.07	66.70	17.39	26.61	30.13	8.02	Tangerang	33.92	64.57	21.90	30.89	23.15	7.15
Pamekasan	136.83	44.56	60.97	61.35	52.05	31.93	Cilegon	3.25	104.52	3.39	33.04	42.13	13.92
Sumenep	44.26	64.04	28.35	49.38	20.68	10.21	Serang	54.72	54.72	37.03	32.56	67.68	18.91
Kediri	3.07	102.98	3.16	40.88	43.54	17.80	Tangerang Selatan	67.82	67.82	14.79	36.07	21.81	6.58
Blitar	1.24	144.12	1.79	6.99	268.77	18.80							