Estimating the Unemployment Rate at Sub-District Level in West Java Province in 2024 Using Hierarchical Bayesian Approach with Cluster Information

R D Aditya[1,*], A Y Zukhrufah[1], E Auliya[1], D Widyastuti[1], A K P Lubis[1], A D Nugraha[1], and S Muchlisoh[1]

[1]Politeknik Statistika STIS, DKI Jakarta, Indonesia

*Corresponding author's email: randydaffaa@gmail.com

Abstract. Unemployment is a substantial obstacle to growth in Indonesia, affecting both social and economic stability. The Unemployment Rate is a crucial metric that quantifies the proportion of the labor force actively pursuing work opportunities. The unemployment rate serves as a critical indicator of labor market imbalances, essential for labor policy formulation and assessment. Nonetheless, unemployment data has limitations, particularly at the micro-level, owing to sample constraints. Small Area Estimation (SAE) can address these constraints. This study estimates the unemployment rate at the sub-district level in West Java province for 2024 utilizing the Hierarchical Bayes Beta methodology and clustering techniques. The modeling results indicate that most sub-districts exhibit a low to medium unemployment rate, however 21 locations demonstrate a very high unemployment rate, ranging from 23.00 percent to 48.06 percent.

Keyword: clustering, hierarchical bayes, small area estimation, unemployment.

1. Introduction

Unemployment poses a considerable obstacle to growth in Indonesia, affecting both social and economic stability [1]. The disparity between work possibilities and labor force expansion results in inadequate absorption of human resources and a reduction in community income levels. Unemployment serves as a metric for evaluating the efficacy of regional labor and economic policy [2]. Consequently, a sustainable development strategy is essential that prioritizes the creation of productive and dignified employment opportunities. This initiative corresponds with Indonesia's dedication to the Sustainable Development Goals (SDGs), specifically target 8.5 on decent work, and adheres to the stipulation of Article 27 Paragraph (2) of the 1945 Constitution, which ensures every citizen's right to employment and a respectable level of living.

The open unemployment rate (TPT) is a critical metric that quantifies the proportion of the labor force actively pursuing employment. The unemployment rate serves as a critical indicator of labor market disparities, essential for labor policy formulation and assessment. Indonesia targets an unemployment rate of 5.5 to 6.35 percent within the Sustainable Development Goals [3]. According to data from the Badan Pusat Statistik (BPS), Indonesia's unemployment rate in August exhibits a declining trend since its apex in 2020. In 2024, the unemployment rate attained roughly 4.9 percent. This signifies the resurgence of the labor market and the efficacy of the government's policy in integrating the workforce. Most provinces in Indonesia have favorable labor market conditions, marked by unemployment rates that fall below the national average. Nonetheless, regional disparities persist, as 12provinces exhibit an unemployment rate exceeding the national average. The provinces of West Java, Banten, and Papua exhibit the greatest unemployment rates. West Java warrants particular attention due to its status as the province with the highest unemployment rate in Indonesia, reflecting considerable difficulties in labor absorption within the region (figure 1). This research underscores the necessity of developing labor policies that prioritize regions with elevated unemployment rates and advocates for the dissemination of more granular unemployment data at the regency/municipality and sub-district levels to enhance the precision of unemployment reduction strategies.

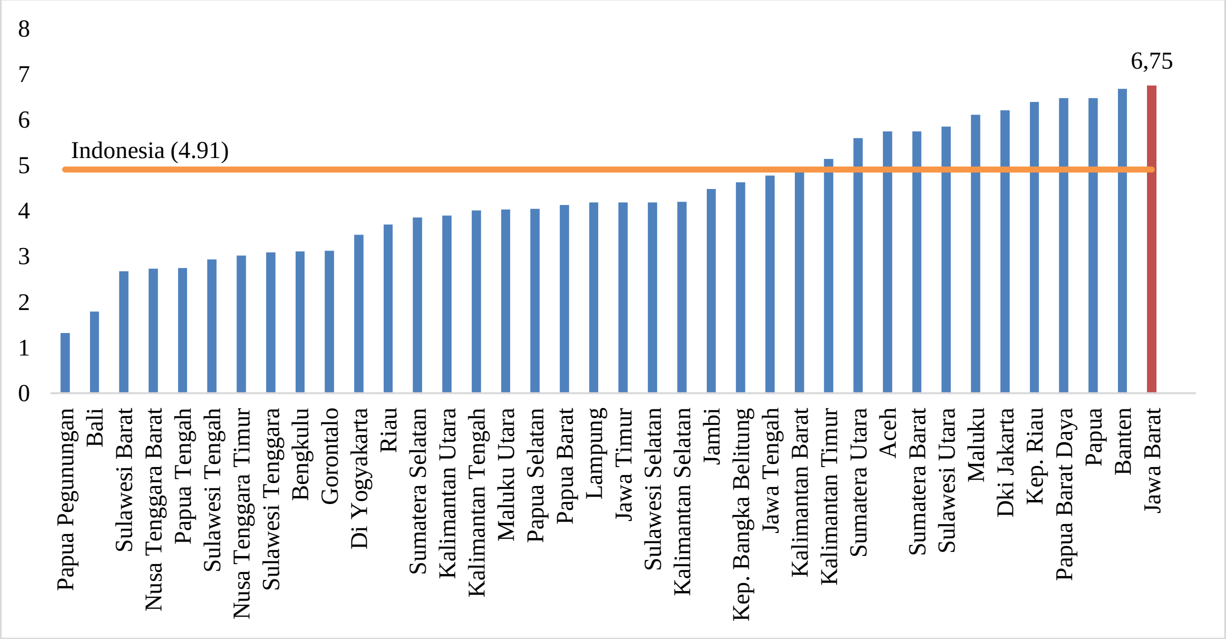


Figure 1. Provincial Unemployment Rates in Indonesia for August 2024

While unemployment data at the sub-district level is crucial, the National Labor Force Survey (Sakernas) does not furnish such data due to the limited sample size, leading to a substantial and statistically inadmissible standard error. Small Area Estimation (SAE) can address these constraints. The benefit of SAE lies in its capacity to generate more accurate estimates with minimal data, hence facilitating more focused labor policy development at the sub-district level [4].

Some research have undertaken research on unemployment rates utilizing the SAE approach [5]–[7]. The study's findings demonstrate that the SAE approach, especially Hierarchical Bayes (HB), yields more accurate estimations than direct estimate methods, as indicated by reduced RSE values. Nevertheless, the majority of research remains confined to the regency/municipality level and has yet to incorporate a clustering-based methodology. Grouping regions according to homogenous traits has been demonstrated to enhance the accuracy of estimations [8], [9].

The selection of the Beta distribution in the HB model is crucial due to TPT being proportional data with a value range of 0 to 1 [10]. This study aims to generate more precise estimates at the sub-district level in West Java Province by integrating the HB Beta methodology with cluster data. This study has three primary objectives: (i) to estimate the sub-district level TPT in West Java using SAE HB Beta with and without cluster information; (ii) to assess the precision of the SAE HB Beta estimation results and comparing them with direct estimation results; (iii) to delineate the sub-district level TPT in West Java Province based on the best model. This research contributes by offering an open unemployment map at the sub-district level, previously unavailable, thereby serving as a foundation for developing evidence-based labor policies locally.

The use of clustering at the sub-district level in West Java for estimating Hierarchical Bayesian (HB) Beta Small Area Estimation (SAE) is based on the high heterogeneity of characteristics between regions in the province, including significant social, economic, and demographic variations between urban and rural areas. Research by [11] shows that West Java faces inequalities in access to infrastructure, housing conditions, and household welfare between urban and rural areas, highlighting the importance of a group-based approach for more representative analysis.

In the context of small area estimation, clustering approaches play a crucial role in improving the reliability of estimates for subpopulations with similar characteristics. [12] asserts that grouping small areas based on Euclidean distance between relevant covariates can result in a smaller Mean Squared Prediction Error (MSPE) in area-level linear mixed models, especially when the between-cluster variance components differ significantly. Thus, clustering helps minimize estimation bias and improve prediction stability. Additionally, [13] also proved that integrating smoothing and clustering in SAE can improve the accuracy of inter-area estimation with similar characteristics and accelerate convergence in the Markov Chain Monte Carlo (MCMC) process. Therefore, applying clustering in the SAE HB Beta at the sub-district level in West Java is a logical and empirical methodological step toward producing more efficient and stable estimates.

2. Research Method

2.1. Small Area Estimation

Small Area Estimation (SAE) is a statistical technique employed to get parameter estimates for small geographic units, such districts, sub-districts, or villages. Small Area Estimation (SAE) yields dependable parameter estimates for diminutive target populations within limited geographical regions [4]. Small Area Estimation (SAE) operates by leveraging strength from several domains using statistical models that link regions using auxiliary variables derived from alternative data sources, like censuses, registrations, or extensive surveys [14]. [4] categorize the SAE approach into two primary types: direct estimation and indirect estimation. Direct estimation entails utilizing solely sample data from the observed region, disregarding external information, which often leads to a significant error rate or renders calculation unfeasible in the absence of samples from that area. Indirect estimate entails employing data from alternative domains through statistical models.

Two prevalent modeling approaches employed for small area estimation are area-level models and unit-level models.

Area-level model

Area-level model are employed when data is exclusively accessible in aggregated form by region (e.g., district, sub-district). The observed target variable is the direct estimate (e.g., from a survey) for each area, with supplementary information provided by auxiliary variables at the area level. The area-level model comprises two components (Fay-Herriot) and is articulated using an equation.

Model Sampling

Let $\hat{\theta}_i$ represent an unbiased direct estimator for parameter θ_i which $\hat{\theta}_i$ encompasses sampling error, leading to the equation:

$$\hat{\theta}_i = \theta_i + e_i \tag{1}$$

Where $\hat{\theta}_i$ denotes the direct estimate for area i, θ_i signifies the true parameter for area i, e_i represents the sampling error, presumed to be $e_i \sim N(0, \sigma_i^2)$ with σ_i^2 indicating the sampling variance.

Model Linking associates the true parameter θ_i with auxiliary variables x_i through the equation:

$$\theta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + b_i v_i \qquad (2)$$

Where x_i denotes the vector of auxiliary variables $(x_{1i}, x_{2i}, \ldots, x_{pi})^T$ that are correlated with the parameter to be estimated θ_i for area i, b_i is a known positive constant, β is the vector of regression coefficients of size $p \times 1$ or $(\beta_1, \beta_2, \ldots, \beta_p)$, v_i represents the random area effect assumed to follow a normal distribution $v_i \sim N(0,\ \sigma_v^2)$, and $i = 1, 2, \ldots, m$ with m indicating the number of areas.

The integration of the two aforementioned models yields a linear mixed model at the area level, referred to as the Fay-Herriot model, represented by the subsequent equation.

$$\hat{\theta}_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + b_i v_i + e_i \qquad (3)$$

The Fay-Herriot model incorporates two error components: sampling error e_i and model error v_i. The amalgamation of these two stochastic elements (sample error and area impact) establishes a framework that facilitates the prediction of tiny area parameters θ_i utilizing regression data and random effects across areas.

Unit-level model

The unit-level model is employed when data is accessible at the micro or individual unit level. This model employs the correlation between unit attributes and the target variable while incorporating inter-area variability through random components. The Battese-Harter-Fuller (BHF) model, established by [15], is one of the most commonly employed unit-level models in the realm of Stochastic Frontier Analysis (SAE). Model BHF posits that the data adheres to a mixed regression model (linear mixed model) incorporating random effects at the small area level. The unit-level model presupposes the availability of data from the unit-level auxiliary variable vector $x_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijp})^T$ for each population unit j inside area i. Let θ_{ij} represent the variable to be estimated, which is related to the unit-level predictor variable xij through the equation:

$$\theta_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + b_{ij} v_{ij} + e_{ij} \qquad (4)$$

Here θ_{ij} denotes the value of the target variable for unit j in area i, β is the vector of regression coefficients of size $p \times 1$ atau $(\beta_1, \beta_2, \ldots, \beta_p)$, v_i represents the area random effect assumed to be $v_i \sim iidN(0,\ \sigma_v^2)$, $e_{ij} = k_{ij} \times e_{ij}$ where k_{ij} is a known constant, and e_{ij} is a random variable that is independent, identically distributed, normally distributed, and independent of v_i. Let $i = 1, 2, \ldots, m$ where m denotes the number of areas, and $j = 1, 2, \ldots, n_i$ where n_i signifies the number of units in the-i small area.

The estimate process in Small Area estimate (SAE) commences with the development of a statistical model that associates the target variable with auxiliary variables. The model's parameters are calculated via statistical techniques, including Restricted Maximum Likelihood (REML). Subsequently, parameter value estimations for minor locations are determined utilizing either a frequentist or Bayesian inference methodology. The frequentist methodology use empirical best linear unbiased prediction (EBLUP), whereas the Bayesian methodology utilizes either Empirical Bayes (EB) or Hierarchical Bayes (HB) [16].

The Bayesian approach is used because in Small Area Estimation (SAE) modeling, there are two Bayesian theorem-based approaches, Empirical Bayes (EB) and Hierarchical Bayes (HB). EB methods estimate the prior distribution using sample data with unknown parameters obtained thru classical methods such as the method of moments, ML/REML, or a combination of both [17]. However, this approach introduces uncertainty in the prior estimation, which impacts the posterior distribution [18]. Additionally, the EB method does not have a precise measure for the standard error estimator [17].

Meanwhile, the HB method overcomes the weaknesses of EB by treating unknown parameters as random components with a specific prior distribution. The HB method has a hierarchical model, which involves many parameters (hyperparameters) [19]. Bayesian theorem is then used to define the posterior distribution of a small area parameter. The HB method has a hierarchical model, which involves many parameters (hyperparameters) [19]. The hierarchical nature of the HB method allows for modeling to be done in several relatively simple and easy-to-understand stages, but it is quite complex overall. Nevertheless, the HB method can produce more accurate and reliable inferences [18].

One application of the Hierarchical Bayesian approach in Small Area Estimation is the HB Beta model. The Hierarchical Bayes (HB) Beta method is an approach used to estimate parameters in data grouped into several small areas, utilizing the Beta prior distribution. This technique allows for modeling data with a hierarchical structure, which can incorporate information from higher levels to improve estimates at lower levels. This approach relies on a gradual updating of beliefs, which is achieved by combining information from prior distributions and observed data using Bayesian theory [20]. In the HB Beta model, the prior distribution for the model parameters is based on the assumption of a Beta distribution. The Beta distribution is a very useful distribution when the applied model relates to proportions or probabilities (interval $0 \leq y \leq 1$).

2.2. Hierarchical Bayes (HB)

Hierarchical Bayes (HB) is a methodology in Bayesian statistics that organizes parameter structures hierarchically, intending to address uncertainty in both the data and the model parameters, as well as in the parameters of those parameters, referred to as hyperparameters. In the conventional Bayesian framework, analysis is conducted by computing the posterior distribution of parameters θ based on the observed data y, utilizing Bayes' Theorem:

$$p\,(y) \propto p\,(\theta).p(\theta) \tag{5}$$

Conversely, in the hierarchical approach, the parameters θ are treated as random variables contingent upon other parameters, referred to as hyperparameters ϕ, resulting in a tiered model structure:
Level 1 (Data level) : Conditional data model on parameters θ, e.g., $y \mid \theta$
Level 2 (Parameter level) : Parameters θ are dependent on ϕ, i.e., $\theta \mid \phi$
Level 3 (Hyperior level): Given a prior for ϕ, i.e., $\phi \sim \pi(\phi)$

The HB Beta model, a Hierarchical Bayesian technique in Small Area Estimation, employs a Beta distribution at the data level (likelihood) to address non-normal sample distributions for proportion parameters. The HB Beta model is utilized when the goal variable represents a proportion or prevalence, specified within the interval [0,1]. The HB Beta model is considered the most appropriate because the target variable, Unemployment Rate (TPT) is a continuous proportion bounded within the interval (0,1) and not directly derived from individual counts. In addition, this distribution is also flexible because its shape can adapt to data patterns through two shape parameters, namely α and β [21]. The HB Beta model is structured in three hierarchical tiers as follows:
Sampling model

$$\hat{\theta}_i|\theta_i \sim Beta\,(a_i, b_i) \tag{6}$$

with Beta parameters defined as:

$$a_i = \theta_i \cdot \phi \text{ and } b_i = (1 - \theta_i) \cdot \phi \qquad (7)$$

Where θ_i represents the unknown proportion parameter for area i, and ϕ denotes the precision parameter, showing the central tendency of the Beta distribution.

Linking model

The value of θ_i associates with auxiliary variables using a logit function:

$$logit(\theta_i)|\beta, \sigma_v^2 \sim N(\boldsymbol{x_i^T \beta}, \sigma_v^2) \qquad (8)$$

Where x_i represents the vector of auxiliary variables and is the regression coefficient to be estimated, $\beta_p \sim N(\mu_{\beta_p}, \sigma_{\beta_p}^2)$ and $\sigma_v^2 \sim IG(c_1, c_2)$.

The estimate phase of the HB Beta method is conducted via a Markov Chain Monte Carlo (MCMC) simulation, enabling the numerical sampling of values from the posterior parameter distribution. In contrast to the frequentist approach, which generates only point estimates, the Bayesian approach provides a comprehensive posterior distribution, facilitating the computation of the posterior mean as the principal predictive value, along with Bayesian credible intervals that indicate the confidence level in the estimate. This methodology yields more reliable estimates for areas with limited sample numbers and can even produce estimates for locations with no respondent data, as it incorporates information from other regions within the model framework. The trustworthiness of the study results is significantly contingent upon the quality and convergence of the MCMC sample utilized [22].

2.3. *Relative Standard Error (RSE)*

Survey activities are subject to errors, encompassing both sampling and nonsampling errors. An inaccuracy arising from the sampling procedure is termed a sampling error. Nonetheless, if the error does not originate from the sample technique, it is classified as a nonsampling error. The Relative Standard Error (RSE) value of an estimate can quantify sampling error. The accuracy of an estimate can be assessed by the Relative Standard Error (RSE) value. This statistic represents the ratio of the standard error to the estimated value of a variable or indicator, articulated as a percentage [23]. The formula for computing RSE is expressed as follows:

$$RSE(\hat{\theta}) = \frac{SE\ (\hat{\theta})}{\hat{\theta}} \times 100\% \qquad (9)$$

Where $\hat{\theta}$ represents the estimate of the variable to be observed, and $SE\ (\hat{\theta})$ denotes the standard error of the variable estimate, or the square root of the variance. The RSE value can be classified into three groups by [23], as follows:

Table 1. RSE categories.

RSE Value	Estimation Result Criteria
RSE < 25%	Accurate
25% < RSE < 50%	Use with caution
RSE > 50%	Very inaccurate

2.4. *Research data and variables*

The analytical unit for this study encompasses all 627 sub-districts within West Java Province. The data utilized comprises the projected TPT value outcomes from the Sakernas August 2024 as the dependent variable, alongside several auxiliary factors derived from the 2024 Village Potential (Podes) data. A compilation of 146 auxiliary variables were obtained from the literature review. Various variables or factors theoretically influence the unemployment rate. The factors are categorized into four primary dimensions: demographic [24]–[26], educational [20] & [21], economic [7] & [22], and structural and policy [6].

3. Result and Discussion

3.1 Overview of direct estimation of TPT at the sub-district level in West Java Province in 2024

The direct estimation of TPT at the sub-district level in West Java encountered difficulties because to the non-sampling of 149 sub-districts, representing roughly 25 percent of the total sub-districts. The average unemployment rate is 7.69 percent, signifying that roughly 7 to 8 individuals are unemployed for every 100 members of the labor in each sub-district. The Parigi sub-district in Pangandaran Regency exhibits the lowest poverty rate at 0.69 percent, whereas the Cidahu sub-district in Kuningan Regency records the greatest poverty rate at 30.57 percent.

Of the 627 sub-districts analyzed, 333 exhibited a relative standard error over 25 percent. Furthermore, 157 of these sub-districts have a relative standard error (RSE) exceeding 50 percent, with the maximum RSE recorded at 120.37 percent in the Taraju sub-district of Tasikmalaya Regency. This study classified locations with an RSE of 0 percent as non-sample sub-districts, as it was presumed that only one observation sample existed in those sub-districts, potentially introducing bias in the estimation. Initially, 48 sub-districts with an RSE of 0 percent were reclassified to non-sample status, increasing the total number of non-sample sub-districts to 197. Under these circumstances, it might be inferred that the direct estimation results are imprecise and unsuitable for subsequent analysis. The Hierarchical Bayes methodology enhances estimations in smaller regions.

3.2 Small Area Estimation using Hierarchical Bayes Beta without cluster information

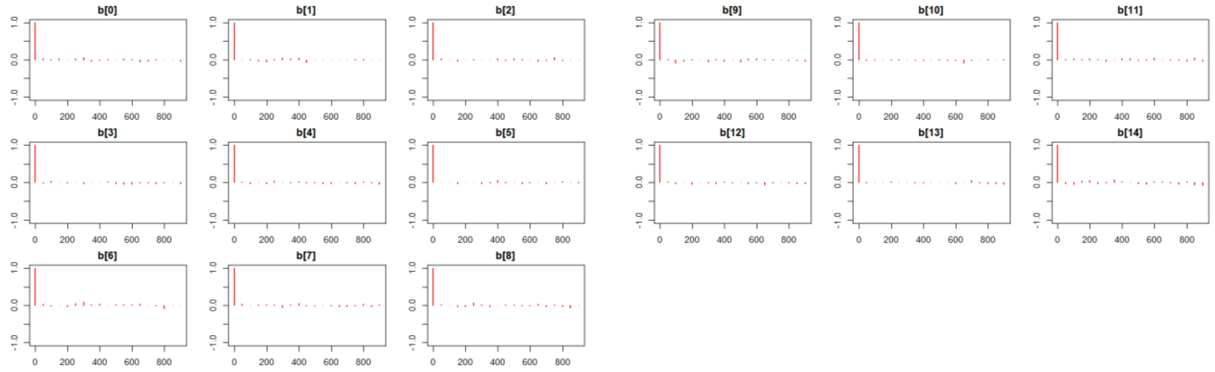
Initially, the SAE approach was implemented collectively for all sub-districts in West Java. Auxiliary variables were chosen by Pearson's correlation significance test between the logit of the unemployment rate and all auxiliary variables, subsequently followed by stepwise regression and multicollinearity assessments. The variable selection yielded 14 candidate variables, which were further modeled using 10 update iterations, 50,000 MCMC cycles, and a thinning of 50. The parameter estimations and credible intervals are presented in the subsequent table 2.

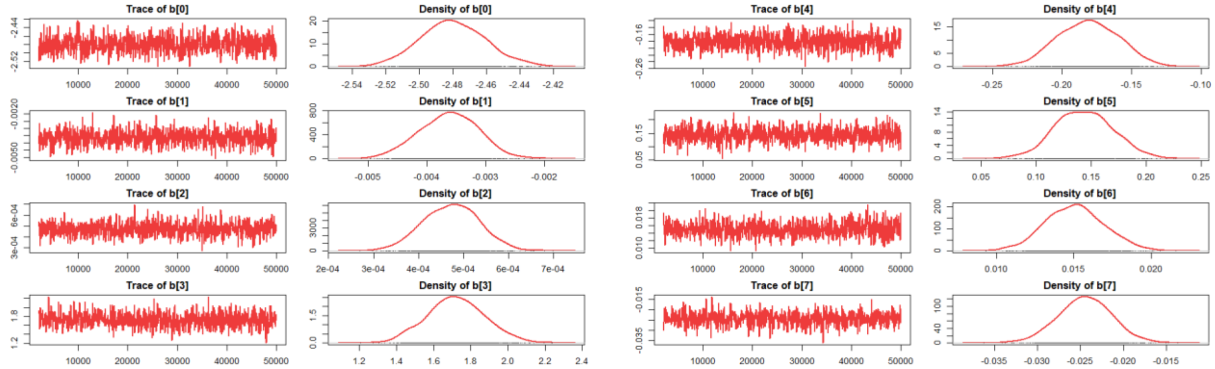
Table 2. Results of parameter estimation for SAE HB Beta without cluster information.

Parameter	Mean	SD	2,5%	97,5%
Intercept	-2.479	0.0191	-2.5164	-2.4396
X_{49}	-0.0036	4.9e-04	-0.0046	-0.0027
X_{2}	0.0005	6.2e-05	0.0003	0.0006
X_{22}	1.7207	0.1549	1.4312	2.0327
X_{56}	-0.1814	0.0216	-0.2223	-0.1402

X_{67}	0.1436	0.0263	0.0918	0.1976
X_{93}	0.0151	0.0019	0.0114	0.0187
X_{114}	-0.245	0.0031	-0.0305	-0.0183
X_{27}	-0.0014	1.7e-04	-0.0017	-0.0010
X_{30}	-0.0057	7.8e-04	-0.0073	-0.0042
X_{32}	-0.0025	3.4e-04	-0.0032	-0.0018
X_{127}	0.2364	0.0333	0.1694	0.2985
X_{128}	-0.0285	0.0042	-0.0363	-0.0198
X_{42}	0.0243	0.0029	0.0189	0.0303
X_{44}	-0.0018	2.3e-04	-0.0022	-0.0013

The estimation results indicate that all variables are significant, as their credible intervals exclude the value zero. A visual assessment of diagnostic plots is conducted to confirm the model's convergence. An MCMC algorithm is deemed to have converged if the autocorrelation plot diminishes after the first lag, the trace plot exhibits no periodic patterns, and the density plot is bell-shaped. The autocorrelation plot is presented in the subsequent figure 2.





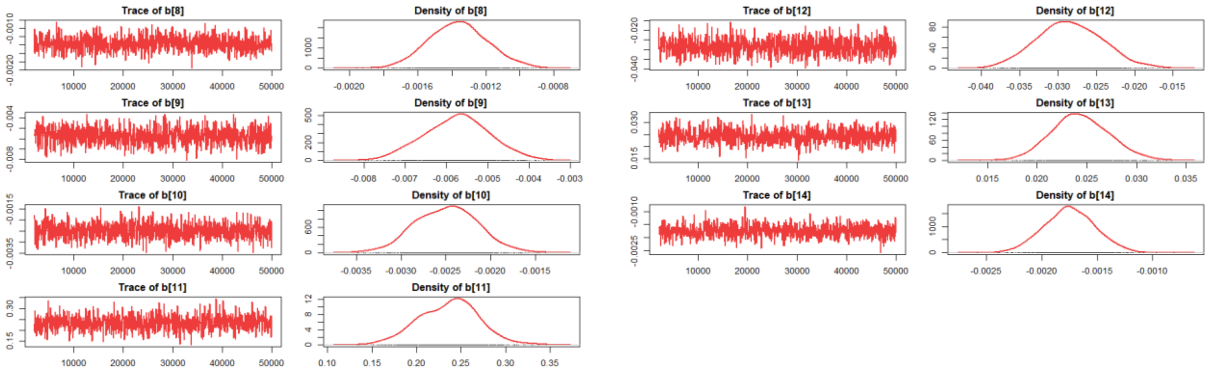
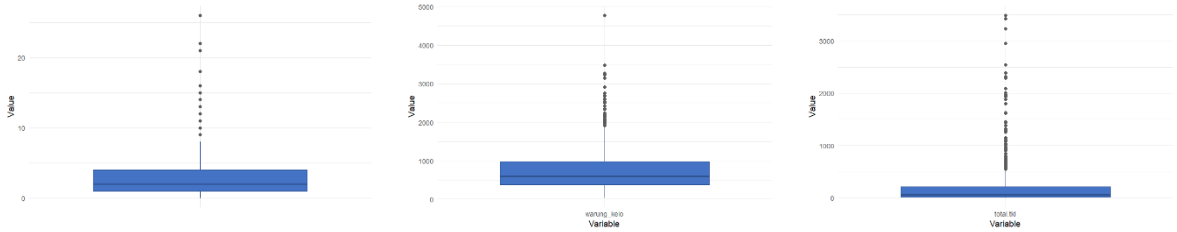


Figure 2. Diagnostic Plot of the Results of the SAE HB Beta Noncluster Estimation.

The results of the SAE HB Beta modelling without clusters have reached a converged state and are proven by figure 2. This approach accurately calculated all sub-districts, including those not included in the sample, as indicated by the image above. This method can enhance precision, as indicated by a reduction in the RSE value relative to direct estimation findings. But, the resultant RSE remains elevated, averaging 30.76 percent, over the set threshold of 25 percent. A clustering method will be conducted to categorize the sub-districts according to their regional features.

3.3 *Clustering results*

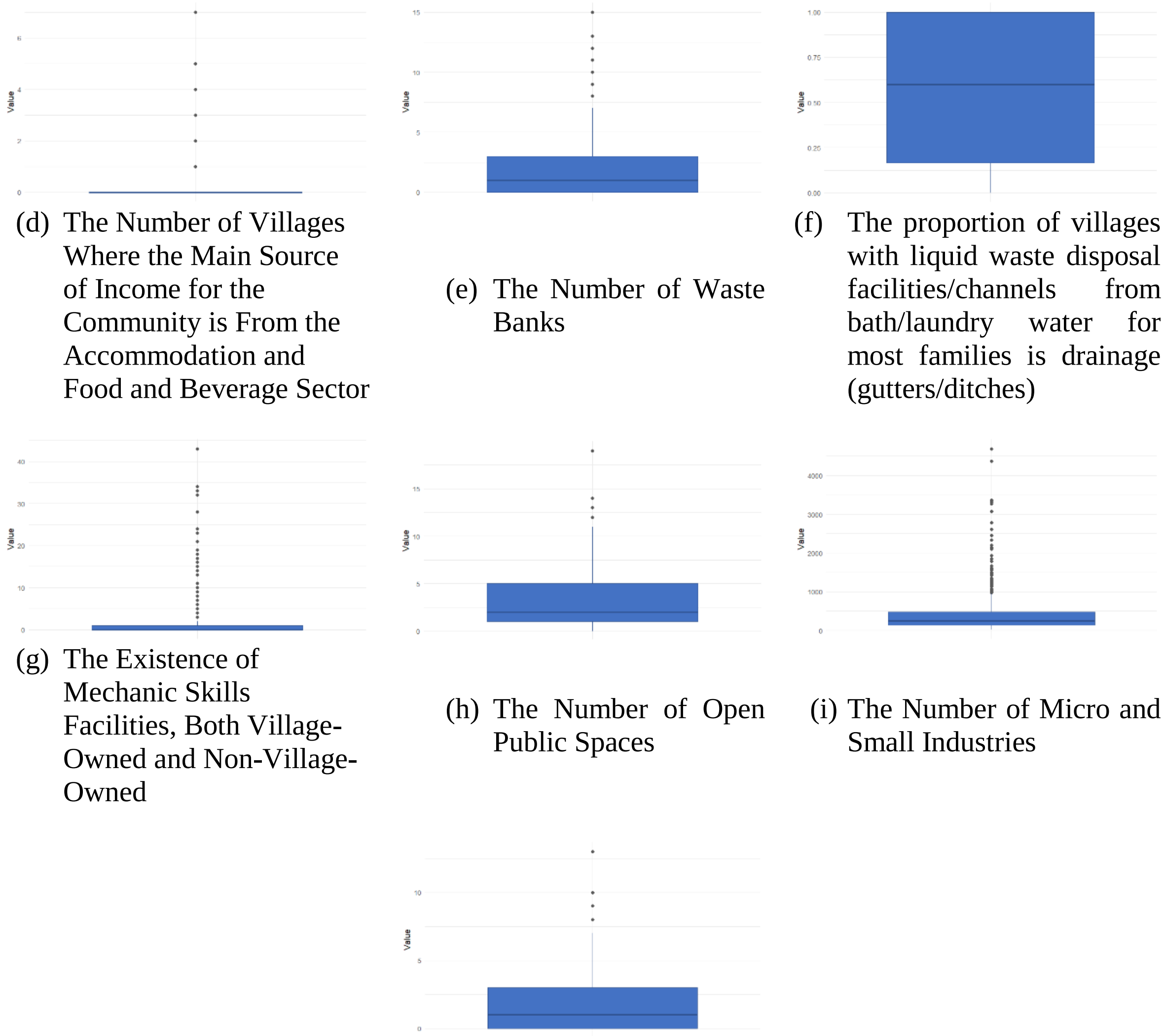
Clustering is conducted to divide the sub-districts into more homogeneous sub-regions or clusters. Each cluster is ideally anticipated to comprise 20 to 60 subdistrict units. Consequently, the suggested quantity of clusters varies between 15 and 18. The variables for cluster analysis were identified by Pearson correlation tests, stepwise regression, and multicollinearity assessments. These variables include the number of senior high schools (X6), the number of grocery stores (X50), the number of male and female Indonesian migrant workers (X3), the number of villages where the main source of income for the community is from the accommodation and food and beverage sector (X22), the number of waste banks (X69), the proportion of villages with liquid waste disposal facilities/channels from bath/laundry water for most families is drainage (gutters/ditches) (X79), the existence of mechanic skills facilities, both village-owned and non-village-owned (X11), the number of open public spaces (X114), the number of micro and small industries (X33), and the number of credit facilities for joint business groups (X128).



(a) The Number of Senior High Schools

(b) The Number of Grocery Stores

(c) The Number of Male and Female Indonesian Migrant Workers



(d) The Number of Villages Where the Main Source of Income for the Community is From the Accommodation and Food and Beverage Sector

(e) The Number of Waste Banks

(f) The proportion of villages with liquid waste disposal facilities/channels from bath/laundry water for most families is drainage (gutters/ditches)

(g) The Existence of Mechanic Skills Facilities, Both Village-Owned and Non-Village-Owned

(h) The Number of Open Public Spaces

(i) The Number of Micro and Small Industries

(j) The Number of Credit Facilities for Joint Business Groups

Figure 3. Boxplot of Variables Used in Clustering

Based on the boxplot results, it can be seen that most sub-districts have relatively low values, but there are some areas with very high values that appear as outliers. The presence of these outliers indicates regional disparities and non-normal (skewed) data distribution. This happens because the Village Potential data is mostly in the form of counts (count data), such as the number of schools, economic facilities, or village infrastructure, so differences in capacity between sub-districts can lead to extreme values.

To handle these data characteristics, K-Medoids is applied as a clustering method because it is more robust to outliers than K-Means [30], [31]. Unlike K-Means, which uses the mean value as the cluster center, K-Medoids uses the medoid (an actual data point) that minimizes the total distance between cluster members, making the clustering results more stable and unaffected by extreme values. Additionally, K-Medoids is also suitable for data with many auxiliary variables and uneven distribution across regions, as seen in the PODES data [32]. After the clustering process, the clustering results were validated using three evaluation indices: the Connectivity Index, the Dunn Index, and the Silhouette Index. These three measures are used to assess the level of compactness and separation between clusters. The results of comparing the values of the three indices are presented in table 3.

Table 3. Validation results to determine the optimum number of clusters.

Criteria	Score	Cluster optimal
Connectivity	569.3710	15
Dunn	0.0711	17
Silhouette	0.1020	15

The analysis indicates that the clustering results categorize 627 sub-districts into 15 distinct clusters. The quantity of subdistricts inside a cluster differs, with cluster 5 comprising the highest number at 83, while cluster 13 contains merely 11. Cluster 10 contains the highest number of non-sample subdistricts at 33, whereas cluster 4 has none.

3.4 Small Area Estimation Hierarchical Bayes Beta utilizing cluster information

Following the clustering process, the SAE HB Beta model was executed for each cluster with varying variables and iteration parameters. The settings are modified until each cluster attains a converged state. This convergence was assessed through diagnostic plots, which consists of the autocorrelation plot, trace plot, and density plot. The MCMC algorithm is considered to have converged if it satisfies the following three conditions: autocorrelation plot for all parameters demonstrates a fading pattern after the first lag, trace plot no longer exhibits a periodic pattern, and density plot has a smooth shape resembling a bell curve. Meanwhile, the significance of the parameters was determined using the credible interval. If a parameter's credible interval crosses zero within the range of 2.5 to 97.5 percent, then the variable is deemed not significant in the model. The modeling composition for each cluster is presented in the subsequent table 4.

Table 4. Modeling composition for each cluster.

Cluster	Number of variables	iter.update	iter.MCMC	thin
1	10	50	300000	20
2	22	20	100000	50
3	12	25	100000	80
4	6	10	50000	100
5	34	20	50000	100
6	6	20	50000	30
7	10	30	70000	50
8	20	15	80000	100
9	13	100	100000	80
10	10	15	70000	50

Cluster	Number of variables	iter.update	iter.MCMC	thin
11	1	20	50000	50
12	4	20	100000	50
13	2	20	100000	15
14	15	50	100000	60
15	3	35	100000	80

Evaluation of modeling results

The objective of the SAE HB Beta model is to generate more accurate estimates than those obtained by direct estimating. A comparison was conducted between the directly determined RSE and the RSE findings derived using SAE. This comparison is illustrated in table 5 and figure 4.

Table 5. Statistical summary of RSE for each estimation method (%).

Descriptive Statistics	Direct Estimation	SAE HB Beta without cluster information	SAE HB Beta with cluster information
Min	0.62	8.06	0.98
Q1	26.15	16.06	2.88
Median	42.39	21.04	5.57
Mean	46.16	30.76	10.41
Q3	63.77	56.19	19.16
Max	120.37	64.86	41.41
NA	197	0	0

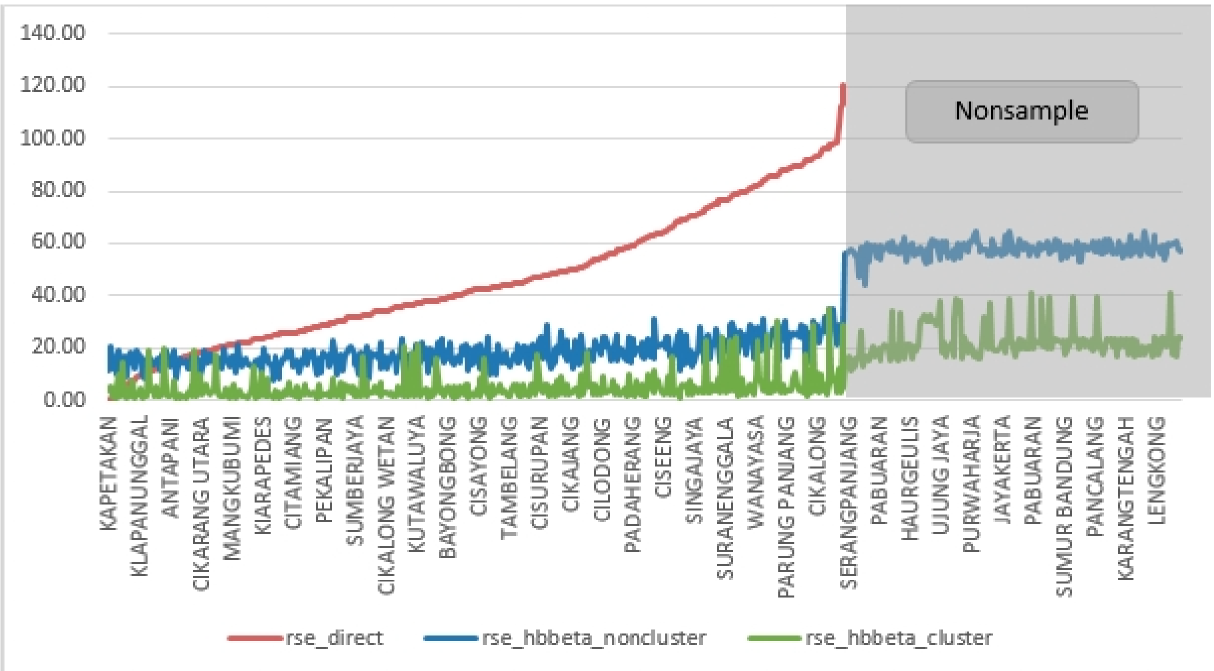


Figure 4. Comparison of the Direct Estimation Method with SAE HB Beta.

The aforementioned table 5 and figure 4 indicate that the SAE HB Beta model incorporating cluster information achieves a relative standard error below 25 percent for most of sub-districts (594 out of 627 subdistrict). Consequently, the approach has enhanced the precision of the sub-district level TPT in West Java Province. The subsequent stage is to perform spatial mapping to assess the unemployment status.

3.5 Mapping of Unemployment Rates at the Sub-district Level in West Java Province for 2024

Mapping was performed using the best model, SAE HB Beta with cluster information. To facilitate interpretation, the proportion figures were multiplied by 100 to obtain percentages. The mapping results are shown in figure 5.

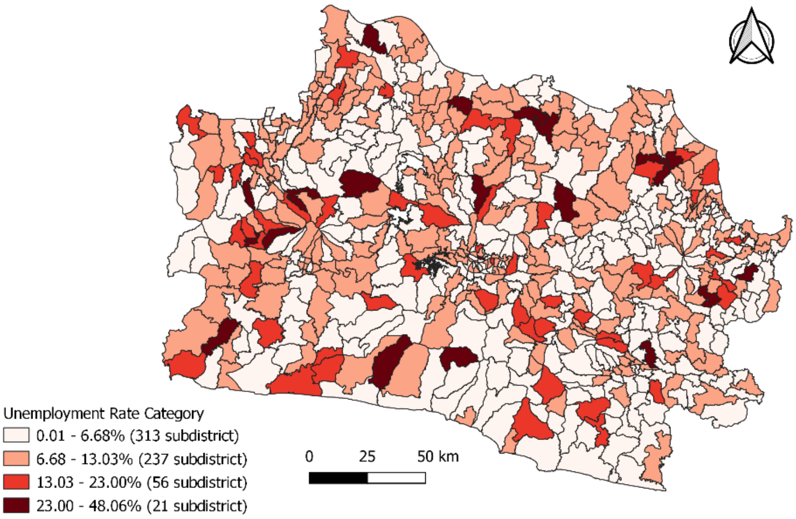

Unemployment Rate Category
0.01 - 6.68% (313 subdistrict)
6.68 - 13.03% (237 subdistrict)
13.03 - 23.00% (56 subdistrict)
23.00 - 48.06% (21 subdistrict)
0
25
50 km

Figure 5. Choropleth Map Depicting TPT Levels by Sub-district in West Java Province for 2024, Constructed Using SAE HB Beta with Clustering Data

The chart above classifies 627 sub-districts into four categories utilizing the natural breaks approach. Natural breaks were used as they reduce volatility within each category and enhance distinctions across categories. The image above indicates that most areas in West Java have low to moderate unemployment rates. There are 313 sub-districts with unemployment rates (TPT) between 0.06% and 6.68%, the lowest being in the Kersamanah sub-district of Garut. Additionally, 237 sub-districts exhibit a TPT range from 6.68% to 13.03%.

There are 56 sub-districts marked in red with elevated TPT, varying from 13.03% to 23.00%, and 21 sub-districts marked in dark red exhibiting the highest TPT, ranging from 23.00% to 48.06%. Cikoneng subdistrict in Ciamis Regency possesses the greatest TPT among sub-districts.

One limitation of this research is that the data used in this study is processed data from BPS (Statistics Indonesia) and not raw data. Consequently, this constraint makes it challenging to perform in-depth checks on data quality at the unit level or to incorporate additional variables that might only be available in the original raw dataset.

The policy recommendations derived from the analysis focus on targeted intervention and the integration of advanced statistical modeling into planning. Firstly, it is recommended that the West Java Provincial Government designate sub-districts with relatively moderate and high TPT as priority areas for employment intervention. These sub-districts, particularly those with high unemployment, should be targeted for local economic stimulus. This stimulus is advised to take practical forms, such as providing capital assistance and mentoring for micro-businesses, implementing productive labor-intensive programs (padat karya produktif), and offering support for business digitalization alongside access to financing through regional banks. Secondly, the policy strongly suggests the utilization of the Small Area Estimation (SAE) Model as a critical basis for both employment planning and evaluation. Specifically, the Hierarchical Bayes (HB) model, which is capable of producing more precise estimates, needs to be integrated into the system of evidence-based planning. This integration is essential for key activities like the preparation of the Regional Manpower Planning (Rencana Tenaga Kerja Daerah or RTKD) and for evaluating the annual achievements of unemployment reduction programs.

4. Conclusion

This study identifies the SAE HB Beta model with 15 clusters as the optimal model for estimating the unemployment rate at the sub-district level in West Java Province for 2024. This model was selected due to its ability to address the data constraints associated with less precise direct estimation outcomes, wherein numerous sub-districts have an RSE value exceeding 25 percent, and 197 of these were not sampled. The application of the SAE HB Beta model markedly enhances estimation quality, as demonstrated by the RSE for most of sub-districts being below 25 percent. Nonetheless, the mapping results reveal an inequitable distribution of job circumstances in West Java. A majority of sub-districts (313 out of 627) exhibit low to moderate unemployment rates (0.01% - 6.68%), whereas 21 sub-districts have relatively high unemployment rates (23.00% - 48.06%).

Acknowledgement

The authors gratefully acknowledge the Tim Pengembangan Statistik Kependudukan dan Ketenagakerjaan BPS RI for their valuable support and contribution to this research.

References

[1] S. F. Azzahra, L. D. Putri, F. Y. Purba, D. Tanjung, A. Rezkitaputri, and R. Z. D. Zulva, “Dampak Pengangguran Terhadap Stabilitas Sosial Dan Perekonomian Indonesia,” *MENAWAN J. Ris. dan Publ. Ilmu Ekon.*, vol. 2, no. 4, pp. 220–233, 2024, doi: 10.61132/menawan.v2i4.719.

[2] R. C. Rambe, P. H. Prihanto, and Hardiani, “Analisis faktor-faktor yang mempengaruhi pengangguran terbuka di Provinsi Jambi,” *e-Jurnal Ekon. Sumberd. dan Lingkung.*, vol. 8, no. 1, pp. 54–67, 2019, doi: 10.22437/pdpd.v7i3.5512.

[3] Kementerian PPN/BAPPENAS, “Laporan Pelaksanaan Pencapaian Tujuan Pembangunan Berkelanjutan Tahun 2023,” 2023.

[4] J. N. . Rao and I. Molina, *Small Area Estimation*, 2nd ed. New Jersey: Wiley, 2015.

[5] S. Muchlisoh, A. Kurnia, K. A. Notodiputro, and I. W. Mangku, “Estimation of unemployment rates using small area estimation model by combining time series and cross-sectional data,” *AIP Conf. Proc.*, vol. 1707, 2016, doi: 10.1063/1.4940870.

[6] Apriliansyah and I. Y. Wulansari, “Application of Spatial Empirical Best Linear Unbiased Prediction (SEBLUP) of Open Unemployment Rate on Sub-District Level Estimation in Banten Province,” *Proc. Int. Conf. Data Sci. Off. Stat.*, vol. 2021, no. 1, pp. 905–913, 2022, doi: 10.34123/icdsos.v2021i1.205.

[7] S. Aprizkiyandari and A. Kurnia, “Small Area Estimation in Estimating Unemployment Rate in Bogor District of Sampled and Non-Sampled Areas Using A Calibration Modeling Approach,” *IJCSN-International J. Comput. Sci. Netw.*, vol. 6, no. 6, pp. 760–765, 2017, [Online]. Available: www.ijcsn.orgimpactfactor:1.5760

[8] J. Torkashvand, K. Godini, A. J. Jafari, A. Esrafili, and M. Farzadkia, “Assessment of littered cigarette butt in urban environment, using of new cigarette butt pollution index (CBPI),” *Sci. Total Environ.*, vol. 769, 2021, doi: 10.1016/j.scitotenv.2020.144864.

[9] A. R. Fusur and A. Ubaidillah, “Pendugaan Area Kecil untuk Persentase Balita Miskin Tingkat Kabupaten/Kota Pendugaan Area Kecil untuk Persentase Balita Miskin Tingkat Kabupaten/Kota di Provinsi Papua Tahun 2023 Menggunakan Pendekatan EBLUP dengan Informasi Klaster (Small Area Estimation,” *Semin. Nas. Off. Stat.*, pp. 927–936, 2024.

[10] B. Liu, “Hierarchical Bayes Estimation and Empirical Best Prediction of Small Area Proportions,” University of Maryland, 2009. [Online]. Available: http://dx.doi.org/10.1016/j.bpj.2015.06.056%0Ahttps://academic.oup.com/bioinformatics/article-abstract/34/13/2201/4852827%0Ainternal-pdf://semisupervised-3254828305/semisupervised.ppt%0Ahttp://dx.doi.org/10.1016/j.str.2013.02.005%0Ahttp://dx.doi.org/10.10

[11] S. Hafsah, Rifda Nida’ul Labibah, Anwar Fitrianto, Erfiani, and L.M. Risman Dwi Jumansyah, “Visualization and Mapping of Household Housing Conditions in West Java Using Multidimensional Scaling,” *J. Stat. dan Apl.*, vol. 8, no. 2, pp. 138–151, 2024, doi: 10.21009/jsa.08201.

[12] E. Torkashvand, M. Jafari Jozani, and M. Torabi, “Clustering in small area estimation with area level linear mixed models,” *J. R. Stat. Soc. Ser. A Stat. Soc.*, vol. 180, no. 4, pp. 1253–1279, 2017, doi: 10.1111/rssa.12308.

[13] R. C. Steorts, “Smoothing, Clustering, and Benchmarking for Small Area Estimation,” *Int. Stat. Rev.*, vol. 88, no. 3, pp. 580–598, 2014, doi: 10.1111/insr.12373.

[14] I. Molina, P. Corral, and M. Nguyen, “Estimation of poverty and inequality in small areas: review and discussion,” *Test*, vol. 31, no. 4, pp. 1143–1166, 2022, doi: 10.1007/s11749-022-00822-1.

[15] G. E. Battese and W. A. Fuller, “An Error Components Model for Prediction of County Crop Areas Using Survey and Satelite Data,” no. 15, 1982.

[16] N. R. Ver Planck, A. O. Finley, J. A. Kershaw, A. R. Weiskittel, and M. C. Kress, “Hierarchical Bayesian models for small area estimation of forest variables using LiDAR,” *Remote Sens. Environ.*, vol. 204, no. April 2017, pp. 287–295, 2018, doi: 10.1016/j.rse.2017.10.024.

[17] M. Ghosh and P. Lahiri, “A Hierarchical Bayes Approach to Small Area Estimation with Auxiliary Information,” pp. 107–125, doi: https://doi.org/10.1007/978-1 4612-2944-5_6.

[18] A. Noviani, “Small Area Estimation dengan Pendekatan Hierarchical Bayesian Neural Network Untuk Kasus Anak Putus Sekolah dari Rumah Tangga Miskin di Provinsi Jawa Timur,” Institut Teknologi Sepuluh November, 2016. [Online]. Available: https://repository.its.ac.id/1470/

[19] J. K. Kruschke, “Tutorial: Doing Bayesian Data Analysis with R and BUGS,” in *Expanding the Space of Cognitive Science - Proceedings of the 33rd Annual Meeting of the Cognitive Science Society, CogSci*, 2011, pp. 56–57.

[20] R. G. Chambers and C. J. Skinner, “Analysis of survey data,” *John Wiley Sons*, 2003, doi: https://doi.org/10.1002/0470867205.

[21] H. E. Catalán Nájera, “Small-area estimates of stunting. Mexico 2010: Based on a hierarchical Bayesian estimator,” *Spat. Spatiotemporal. Epidemiol.*, vol. 29, no. 2019, pp. 1–11, 2019, doi: 10.1016/j.sste.2019.01.001.

[22] G. D. Johnson, “Small area mapping of prostate cancer incidence in New York State (USA) using fully Bayesian hierarchical modelling,” *Int. J. Health Geogr.*, vol. 3, 2004, doi: 10.1186/1476-072X-3-29.

[23] BPS, “Laporan Indeks Khusus Penanganan Stunting 2018 - 2019,” 2020. [Online]. Available: http://dx.doi.org/10.1016/j.bpj.2015.06.056%0Ahttps://academic.oup.com/bioinformatics/article-abstract/34/13/2201/4852827%0Ainternal-pdf://semisupervised-3254828305/semisupervised.ppt%0Ahttp://dx.doi.org/10.1016/j.str.2013.02.005%0Ahttp://dx.doi.org/10.10

[24] A. Juliyanto, “Model Hierarchial Bayes pada Small Area Estimation untuk Pendugaan Proporsi Penganngguran pada Desain Survei Kompleks,” Intitut Teknologi Sepuluh Nopember, 2016.

[25] M. Mauliani, M. Maiyastri, and R. Diana, “Pendugaan Angka Pengangguran Di Kabupaten Padang Pariaman Menggunakan Small Area Estimation Dengan Pendekatan Hierarchical Bayes (Hb) Lognormal,” *J. Mat. UNAND*, vol. 7, no. 4, pp. 15–21, 2019, doi: 10.25077/jmu.7.4.15-21.2018.

[26] R. P. Wicaksono, I. K. G. Sukarsa, and I. P. E. N. Kencana, “Memodelkan Tingkat Pengangguran Di Kota Denpasar Dengan Pendugaan Area Kecil Empirical Bayes,” *E-Jurnal Mat.*, vol. 10, no. 2, p. 81, 2021, doi: 10.24843/mtk.2021.v10.i02.p325.

[27] D. N. F. A. Puspita and A. Ubaidillah, “Pendugaan Area Kecil Tingkat Pengangguran Terbuka Level Kecamatan Di Provinsi Kepulauan Riau Tahun 2022,” *Semin. Nas. Off. Stat.*, vol. 2024, no. 1, pp. 51–62, 2024, doi: 10.34123/semnasoffstat.v2024i1.1982.

[28] R. S. Winanda, Y. I. Khairani, and F. W. Permana, “Assessing Unemployment Rates in Tanah Datar Regency: Insights From Small Area Estimation,” *Barekeng*, vol. 19, no. 2, pp. 1433–1444, 2025, doi: 10.30598/barekengvol19iss2pp1433-1444.

[29] B. Hartono and R. Hapsari, “Kajian Metode Small Area Estimation Untuk Menduga Tingkat Pengangguran Terbuka,” *J. Litbang Sukowati Media Penelit. dan Pengemb.*, vol. 1, no. 2, pp. 95–106, 2018, doi: 10.32630/sukowati.v1i2.27.

[30] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data*. 1990. [Online]. Available: http://www.biblioteca.pucminas.br/teses/Educacao_PereiraAS_1.pdf%0Ahttp://www.anpocs.org.br/portal/publicacoes/rbcs_00_11/rbcs11_01.htm%0Ahttp://repositorio.ipea.gov.br/bitstream/11058/7845/1/td_2306.pdf%0Ahttps://direitoufma2010.files.wordpress.com/2010/

[31] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. 2012. doi: 10.1016/C2009-0-61819-5.

[32] T. A. Munandar, “Penerapan Algoritma Clustering Untuk Pengelompokan Tingkat Kemiskinan Provinsi Banten,” *JSiI (Jurnal Sist. Informasi)*, vol. 9, no. 2, pp. 109–114, 2022, doi: 10.30656/jsii.v9i2.5099.