2020 Asia–Pacific Statistics Week
A decade of action for the 2030 Agenda: Statistics that leaves no one and nowhere behind
15-19 JUNE 2020 | Bangkok, Thailand

# Estimation of SDGs Indicator for Non-Sampled Area Using Small Area Estimation with Partitioning Around Medoids Clustering

Rizky Zulkarnain[1*]; Dwi Jayanti[2]

[1*,2]BPS-Statistics Indonesia
Jl. Dr. Sutomo 6-8 Jakarta 10710, Indonesia
E-mail: [1*]zulqarnaen@bps.go.id; [2]dwijayanti@bps.go.id

**Abstract:**
Indonesia has a strong commitment to achieve Sustainable Development Goals (SDGs) at national and local level, including in the field of family planning. Indicators for family planning targets are generally derived from Indonesia Demographic and Health Survey (IDHS). However, IDHS is designed to produce estimates at national and provincial levels only. This generates gaps for district level policies. Small area estimation (SAE) techniques are usually considered as alternative for overcoming this issue. SAE enables the sampled small areas are estimated reliably by utilizing auxiliary information from other sources. However, there is problem for non-sampled area since it has no observation. Roughly estimating non-sampled area using synthetic model ignoring area random effect merely produces considerable bias. This paper attempts to estimate non-sampled area random effect instead of ignoring it. This is achieved by utilizing similarities among particular areas using clustering techniques. The similarities among areas are measured using Manhattan distance, while clusters are generated using partitioning around medoids (PAM) algorithm. PAM is preferred since it is less sensitive to outliers. Non-sampled area random effect is estimated as average of sampled area random effects within the same cluster. Moreover, cluster information from PAM is also incorporated into standard SAE model to achieve better estimates for both sampled and non-sampled areas. This model is applied to estimate contraception prevalence rate (CPR) at district levels in North Sumatera province. From 33 districts in North Sumatera, there are 6 districts that are not sampled in 2017 IDHS. To ensure that CPR estimates fall between 0 and 1, this study uses logit transformation. There are four auxiliary variables that are used to estimate CPR: number of active acceptors, number of family planning clinics, number of family planning institution, and number of pre-prosperous/1st prosperous family. These variables are obtained from Family Planning Coordinating Board of North Sumatera. The results showed that there are five clusters of districts in North Sumatera province. Small area estimates considering cluster information (SAE-cluster) revise the direct estimates upward or downward. Confidence intervals of SAE-cluster are generally shorter than the direct estimates. This indicates that SAE-cluster produces more precise estimates than that direct method. Finally, estimates of CPR in non-sampled districts could be derived and appropriate district level policies could be undertaken.

**Keywords:** cluster; non-sampled area; partitioning around medoids; SAE; SDGs

## 1. Introduction:

In Indonesia, Sustainable Development Goals (SDGs) targets are not merely adopted into national development plan, but also into local development plan. Since decentralization system, each district in Indonesia is an important policy maker for its development. Indonesia has a strong commitment to achieve SDGs, including in the field of family planning. Target 3.7 of SDGs declares that by 2030 ensure universal access to sexual and reproductive health-care services, including for family planning, information and education, and the integration of reproductive health into national strategies and programmes. In the National Medium Term Development Plan (RPJMN) of Indonesia, contraception prevalence rate for modern method (mCPR) is targeted at 63.41% in 2024, while unmet need for family planning is expected to decline at 7.4% in 2024.

Indicators for family planning targets are generally derived from Indonesia Demographic and Health Survey (IDHS). IDHS is conducted every 5 years with collaboration of National Population and Family Planning Board (BKKBN), Statistics Indonesia (BPS), and the Ministry of Health (Kemenkes), which collects information about fertility, family planning, maternal and child health, etc. IDHS is last updated in 2017. The 2017 IDHS covered about 1,970 census block samples in urban and rural areas

and 47,963 successfully interviewed households. In the interviewed households, there are 49,627 eligible women and 10,009 eligible men were completely interviewed.

The 2017 IDHS was a two-stage stratified survey that was designed to produce estimates at national and provincial levels. Therefore, there are gaps for district level policies. This issue is usually handled by small area estimation (SAE) techniques. SAE enables the small areas are estimated reliably by utilizing auxiliary information from other sources. Empirical Best Linear Unbiased Predictor (EBLUP) is an indirect method to predict small areas parameters. It has been recognized that there is a problem when EBLUP technique is used to estimate the parameters of non-sampled areas. Standard EBLUP uses synthetic model that ignores area random effects for non-sampled areas (Saei and Chambers, 2005). As a result, the resulting estimates will be distorted into a single line of the synthetic model and may cause considerable bias (Anisa *et al.*, 2014).

This paper utilizes similarities among particular areas to estimate area random effects for non-sampled areas. Study is applied to estimate contraception prevalence rate (CPR) at district levels in North Sumatera province. From 33 districts in North Sumatera, there are 6 districts that are not sampled in 2017 IDHS. The idea of considering cluster information into standard EBLUP has been proposed by Anisa *et al.* (2014). This paper uses the model with several modifications: 1) This paper uses partitioning around medoids algorithm to generate clusters; 2) Logit transformation is used to ensure that CPR estimates fall between 0 and 1; and 3) This paper uses area level model rather than unit level model.

## 2. Methodology:

Data for CPR in North Sumatera province is acquired from 2017 IDHS. Auxiliary variables are obtained from Family Planning Coordinating Board of North Sumatera. In order to eliminate the scale effects, these auxiliary variables are transformed into standardized values. Table 1 presents auxiliary variables used in this study.

Table 1. Lists of auxiliary variables

| No. | Variable | Description | Unit of measure |
|---|---|---|---|
| 1 | $Z_1$ | Number of active acceptors | person |
| 2 | $Z_2$ | Number of family planning clinics | unit of clinics |
| 3 | $Z_3$ | Number of family planning institution | unit of institution |
| 4 | $Z_4$ | Number of pre-prosperous and 1st prosperous family | unit of family |

These auxiliary variables are used in generating clusters and in estimating CPR using SAE model. For clustering process, Human Development Index (HDI) is also considered to control that districts are grouped based on their development levels. Partitioning around medoids (PAM) clustering method is preferred since it is less sensitive to outliers. A medoid is simply a representative object from a cluster, whose average distance to all other objects in the same cluster is minimal. Another advantage of PAM is that, since the data does not change, the whole proximity matrix could be computed once and re-use it on each iteration, even across different number of clusters and random repetitions (Sarda-Espinosa, 2019). As in *K*-means clustering, the number of clusters (*K*) in PAM (or sometimes called *K*-medoids) is predetermined. Algorithm of PAM or *K*-medoids clustering is as follows (Hastie *et al.*, 2009):

1. For a given cluster assignment $C$ find the observation in the cluster minimizing total distance to other points in that cluster:
$$i_k^* = \underset{\{i:C(i)=k\}}{\operatorname{argmin}} \sum_{C(i')=k} D(x_i, x_{i'})$$
Then $m_k = x_{i_k^*}, k = 1,2,\dots,K$ are the current estimates of the cluster centers.
2. Given a current set of cluster centers $\{m_1,\dots,m_K\}$, minimize the total error by assigning each observation to the closest (current) cluster center:
$$C(i) = \underset{1\le k\le K}{\operatorname{argmin}} D(x_i, m_k)$$
3. Iterate steps 1 and 2 until the assignments do not change.

In this study, similarities among districts are measured using Manhattan distance. The distance is calculated as $D(x_i, x_{i'}) = \sum_{j=1}^{p} |x_{ij} - x_{i'j}|$, where $p$ is the number of variables. The optimal number of

clusters (*K*) is determined using elbow method. The principle of elbow method is to choose a number of clusters so that adding another cluster does not give much decrease in total within sum of squares.

Cluster information from PAM is further incorporated into standard SAE model to achieve better estimates for both sampled and non-sampled areas. This paper adopts the model developed by Anisa *et al.* (2014), but with several modifications that will be explained later. Consider SAE unit level model:

$$y_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + v_i + e_{ij}$$

where $y_{ij}$ is the sample observation of *j*-th unit in *i*-th area, $\boldsymbol{x}_{ij}$ is the vector of auxiliary variables in *j*-th unit and *i*-th area, whose values are known for all units in population, $\boldsymbol{\beta}$ is the vector of parameters, $v_i$ is area random effects which is distributed $v_i \sim iid\ N(0, \sigma_v^2)$, and $e_{ij}$ is error term which is distributed $e_{ij} \sim iid\ N(0, \sigma_e^2)$.

If it is defined that parameter of interest is the *i*-th small area mean, EBLUP for the sampled area mean can be written as:

$$\bar{Y}_i = \frac{1}{N_i}\left(\sum_{j\in s_i} y_{ij} + \sum_{j^*\in r_i} \hat{y}_{ij^*}\right)$$

where $s_i$ denotes sampled units and $r_i$ denotes non-sampled units in the *i*-th area. Thus, $\hat{y}_{ij^*}$ is estimated value for non-sampled units which calculated with following formula:

$$\begin{aligned}\hat{y}_{ij^*} &= \boldsymbol{x}_{ij^*}^T\hat{\boldsymbol{\beta}} + \hat{\gamma}_i\left(\bar{y}_{is} - \bar{\boldsymbol{x}}_{is}^T\hat{\boldsymbol{\beta}}\right)\\ &= \boldsymbol{x}_{ij^*}^T\hat{\boldsymbol{\beta}} + \hat{v}_i\end{aligned}$$

where $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X^T\Sigma^{-1}X}\right)^{-1}\boldsymbol{X^T\Sigma^{-1}y}$ is generalized least squares estimator of $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$ is covariance matrix of observation $\boldsymbol{y}$, and $\hat{\gamma}_i = \hat{\sigma}_v^2(\hat{\sigma}_v^2 + \hat{\sigma}_e^2/n_i)^{-1}$. EBLUP estimator for non-sampled area mean can be written as follow:

$$\bar{Y}_{i^*} = \frac{1}{N_{i^*}}\left(\sum_{j^*\in r_{i^*}} \hat{y}_{i^*j^*}\right)$$

where $\hat{y}_{i^*j^*}$ is an estimated value that is calculated by the following formula:

$$\hat{y}_{i^*j^*} = \boldsymbol{x}_{i^*j^*}^T\hat{\boldsymbol{\beta}}$$

Anisa *et al.* (2014) proposed a model that modifies standard EBLUP by incorporating the effects of dummy variables from each *k*-th cluster $D_1, D_2, \dots, D_{K-1}$ and the interaction effects of dummy variables and auxiliary variables into basic model. This model also use the average of area random effects within each *k*-th cluster to estimate non-sampled area random effects which can be written as $\bar{\hat{v}}_{(k)} = \frac{1}{m_k}\sum_{i=1}^{m_k}\hat{v}_i$ where $m_k$is the number of sampled area in *k*-th cluster. The model can be written as follow:

- Model for population:

$$y_{ijk} = \beta_0 + \sum_{p=1}^{P}\beta_p x_{ijkp} + \sum_{d=1}^{K-1}\delta_d D_d + \sum_{d=1}^{K-1}\sum_{p=1}^{P}\tau_{dp}D_d x_{ijkp} + v_i + e_{ijk}$$

- Prediction model for sampled area:

$$\hat{y}_{ij^*k} = \hat{\beta}_0 + \sum_{p=1}^{P}\hat{\beta}_p x_{ij^*kp} + \sum_{d=1}^{K-1}\hat{\delta}_d D_d + \sum_{d=1}^{K-1}\sum_{p=1}^{P}\hat{\tau}_{dp}D_d x_{ij^*kp} + \hat{v}_i$$

- Prediction model for non-sampled area:

$$\hat{y}_{i^*j^*k} = \hat{\beta}_0 + \sum_{p=1}^{P}\hat{\beta}_p x_{i^*j^*kp} + \sum_{d=1}^{K-1}\hat{\delta}_d D_d + \sum_{d=1}^{K-1}\sum_{p=1}^{P}\hat{\tau}_{dp}D_d x_{i^*j^*kp} + \bar{\hat{v}}_{(k)}$$

In this study, auxiliary variables are available for area level only. Hence, this paper uses area level model rather than unit level model. These auxiliary variables are transformed into standardized values $(Z)$ to eliminate the scale effects. Moreover, this paper uses CPR as indicator to be estimated, which needs to be non-negative and not larger than one. To ensure that CPR estimates fall between 0 and 1, this paper uses logit transformation. The interaction effects of cluster dummy variables and auxiliary

variables could not be estimated in this study since the limitation of number of areas. Therefore, the model is modified as follow:

- Model for population:

$$ln\left(\frac{p_{ik}}{1-p_{ik}}\right) = \beta_0 + \sum_{p=1}^{P}\beta_p Z_{ikp} + \sum_{d=1}^{K-1}\delta_d D_d + v_i$$

- Prediction model for sampled area:

$$ln\left(\frac{\hat{p}_{ik}}{1-\hat{p}_{ik}}\right) = \hat{\beta}_0 + \sum_{p=1}^{P}\hat{\beta}_p Z_{ikp} + \sum_{d=1}^{K-1}\hat{\delta}_d D_d + \hat{v}_i$$

- Prediction model for non-sampled area:

$$ln\left(\frac{\hat{p}_{i^*k}}{1-\hat{p}_{i^*k}}\right) = \hat{\beta}_0 + \sum_{p=1}^{P}\hat{\beta}_p Z_{i^*kp} + \sum_{d=1}^{K-1}\hat{\delta}_d D_d + \bar{\hat{v}}_{(k)}$$

where $p_{ik}$ is the proportion of married women using contraception (CPR) in *i*-th area and *k*-th cluster, $Z_{ikp}$ is the *p*-th standardized auxiliary variable in *i*-th area and *k*-th cluster, $i^*$ denotes non-sampled area, and $\bar{\hat{v}}_{(k)}$ is the average of sampled area random effects in *k*-th cluster.

## 3. Result:

Figure 1 depicts the resulted total within sum of square as a function of number of cluster. Using elbow method, the optimal number of cluster is chosen when adding another cluster does not give much decrease in total within sum of square. The figure shows that total within sum of square is substantially decreased until five clusters are used. Adding more clusters do not reduce the total within sum of square significantly. Thus, five clusters are considered to be optimal. Figure 2 visualizes the generated clusters using PAM algorithm. It is clearly showed that the five clusters are well separated.

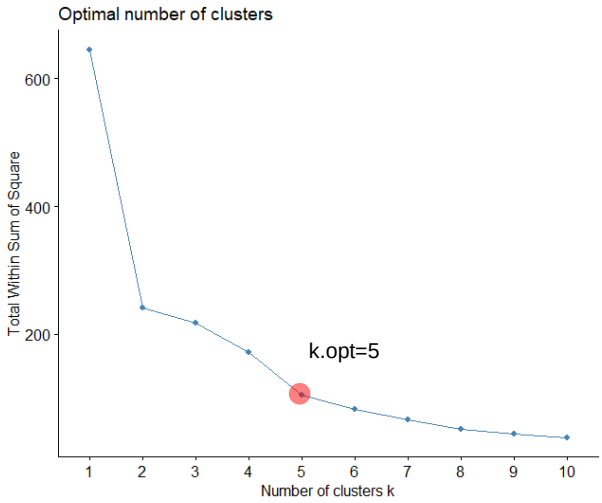


Figure 1. Optimal number of clusters using elbow method

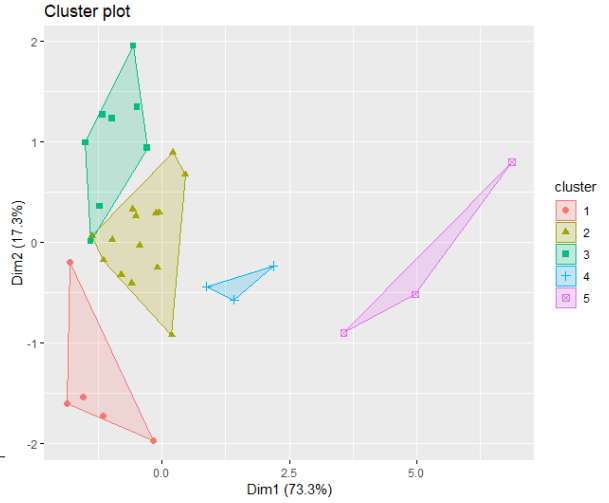


Figure 2. Clustering result using partitioning around medoid (PAM) algorithm

Table 2 presents the members of each cluster. Cluster 1 has five members of districts, including one non-sampled district (Nias Utara). Cluster 1 has the lowest development and family planning status. Cluster 2 consists of fourteen members of districts, including three non-sampled districts: Mandailing Natal, Padang Lawas Utara, and Labuhanbatu Selatan. Cluster 2 has moderate development and family planning status. Cluster 3 consists of eight members of districts, including two non-sampled districts: Samosir and Tanjungbalai. Cluster 3 has lower family planning status. Cluster 4 and cluster 5 each has three members. There is no non-sampled district in both cluster 4 and cluster 5. Cluster 4 and cluster 5 constitute the districts having higher and highest family planning status, respectively.

Table 2. Cluster of Districts in North Sumatera Province

| Cluster | Number of cluster members | Districts |
|---|---|---|
| 1 | 5 | Nias, Nias Selatan, Pakpak Bharat, **Nias Utara***, Nias Barat |
| 2 | 14 | **Mandailing Natal***, Tapanuli Selatan, Tapanuli Tengah, Tapanuli Utara, Labuhan Batu, Dairi, Karo, Humbang Hasundutan, Batu Bara, **Padang Lawas Utara***, Padang Lawas, **Labuhanbatu Selatan***, Labuhanbatu Utara, Gunungsitoli |
| 3 | 8 | Toba Samosir, **Samosir***, Sibolga, **Tanjungbalai***, Pematangsiantar, Tebing Tinggi, Binjai, Padangsidimpuan |
| 4 | 3 | Asahan, Simalungun, Serdang Bedagai |
| 5 | 3 | Deli Serdang, Langkat, Medan |

Note: ***non-sampled area**

The results from above clusters is further utilized to estimate area random effects for non-sampled area $\left(\bar{\hat{v}}_{(k)}\right)$ and to construct area level model incorporating cluster information as follows:

- Prediction model for sampled area:

$$ln\left(\frac{\hat{p}_{ik}}{1-\hat{p}_{ik}}\right) = 1.4600 + 0.1018\,Z_{ik1} - 0.1087\,Z_{ik2} - 0.2425\,Z_{ik3} - 0.1742\,Z_{ik4} - 1.9793\,D_1 - 1.0201\,D_2 - 1.5596\,D_3 - 0.6563\,D_4 + \hat{v}_i$$

- Prediction model for non-sampled area:

$$ln\left(\frac{\hat{p}_{i^*k}}{1-\hat{p}_{i^*k}}\right) = 1.4600 + 0.1018\,Z_{i^*k1} - 0.1087\,Z_{i^*k2} - 0.2425\,Z_{i^*k3} - 0.1742\,Z_{i^*k4} - 1.9793\,D_1 - 1.0201\,D_2 - 1.5596\,D_3 - 0.6563\,D_4 + \bar{\hat{v}}_{(k)}$$

The comparison between direct estimates and small area estimates considering cluster information (SAE-cluster) is presented in 95% confidence interval form in Figure 3. The point in the middle of interval represents point estimate of CPR, while upper and lower limits of interval represent upper and lower bounds of 95% confidence interval for CPR, respectively. SAE-cluster revise the direct estimates upward or downward. Confidence intervals of SAE-cluster are generally shorter than the direct estimates. This indicates that SAE-cluster produces more precise estimates than that direct method.

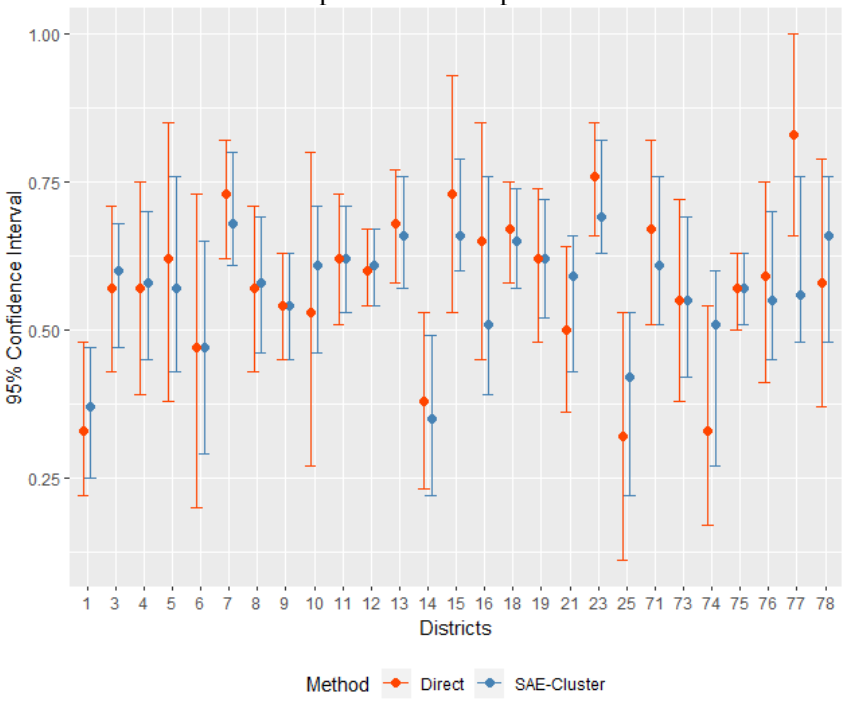


Figure 3. Comparison of 95% confidence interval between direct estimates and SAE-cluster estimates

Spatial distribution of CPR estimates using direct estimation and SAE-cluster estimation are visualized in Figure 4. Missing values in direct estimates occur since the areas are non-sampled in 2017 IDHS. This issue is well handled by SAE-cluster utilizing information from sampled areas within the same cluster. SAE-cluster revises the spatial distribution of CPR estimates in North Sumatera province and change the relative position of several districts.

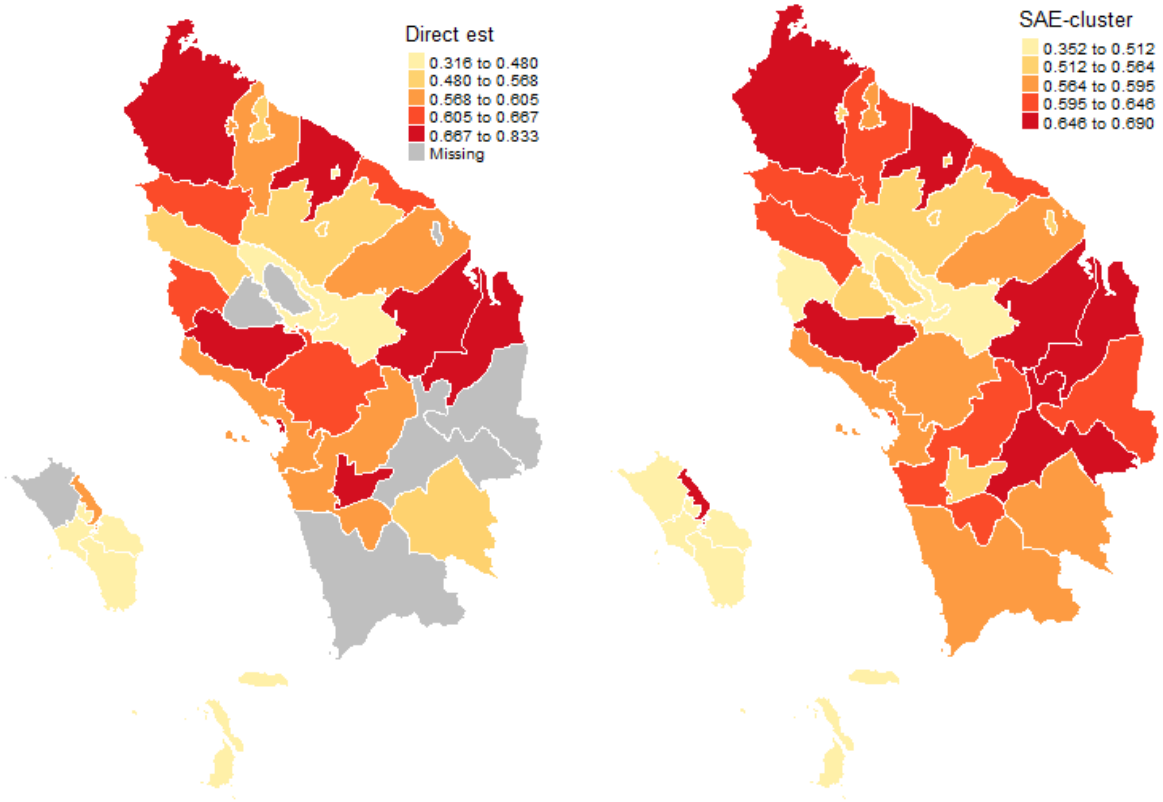


Figure 4. Comparison of spatial distribution between direct estimates and SAE-cluster estimates

**4. Discussion, Conclusion and Recommendations:**

Small area estimation considering cluster information offers two advantages. First, it enables the sampled small areas are estimated reliably. It enhances the precision of estimates substantially. Hence, it overcomes the issue of small sample for district level policies. Second, it serves procedure to estimate area random effects for non-sampled areas, so that better estimates for non-sampled areas could be produced and comprehensive policies could be undertaken.

**References:**

1. Anisa R, Kurnia A, Indahwati. (2014). Cluster Information of Non-Sampled Area in Small Area Estimation. IOSR Journal of Mathematics, 10, 15-19.
2. Anisa R, Notodiputro KA, Kurnia A. (2014). Small Area Estimation for Non-Sampled Area Using Cluster Information and Winsorization with Application to BPS Data. Proc. ICCS-13, 27, 453-462.
3. Hastie T, Tibshirani R, Friedman J. (2009). The Elements of Statistical Learning 2nd Edition. New York: Springer-Verlag.
4. Rao JNK. (2003). Small Area Estimation. New York: John Wiley & Sons.
5. Saei A and Chambers R. (2005). Empirical Best Linear Unbiased Prediction for Out of Sample Area. Working Paper M05/03, Southampton Statistical Sciences Research Institute.
6. Sarda-Espinosa A. (2019). Time-Series Clustering in R Using the dtwclust Package. The R Journal. https://doi.org/10.32614/RJ-2019-023.