Twitter utilization in application of small area estimation to estimate electability of candidate central java governor

IOP Conf. Series: Earth and Environmental Science **299** (2019) 012033 doi:10.1088/1755-1315/299/1/012033

Twitter utilization in application of small area estimation to estimate electability of candidate central java governor

F A Muhyi[1], B Sartono[1*], I D Sulvianti[1] and A Kurnia[1]

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

*Email: bagusco@gmail.com

Abstract. A survey was designed to estimate electability of a governor candidate in the province of Central Java in 2018. The estimate was planned to be done at the province level. However, it is also interesting to do estimation at the district level. The problem is that the sample size at some district is too small or even zero so that the estimate would have inadequate precision. Using a small area estimation method, this research tried to have good estimation by utilizing information from Twitter and other auxiliary data in the modeling. To evaluate the estimation, we calculated the mean square error (MSE) of the estimated electability using resampling method. We noted that both indirect and direct estimations at the province level have equal precision, but the indirect estimation outperformed the direct estimation at the region level. We also compare our result to the election result of Governor Election.

1. Introduction

Towards election of the Candidate Central Java Governor in 2018, electability of a candidate in each region becomes valuable information. A survey about Electability of the Candidates Central Java Governor was designed to estimate electability at the province level. Data gathered from each region is not able to provide a region level estimate with adequate precision because the sample size in some region is small or even zero. Increasing sample size was not very effective in term of time, cost, and source. Small area estimation can provide estimation with adequate precision without increasing sample size [1].

In the context of survey sampling, a subpopulation or area estimate is usually referred to as a “direct estimate” if it is based only in the specific sample data coming from that area [2]. An area estimate is referred to as "indirect estimate" when it is not only based specific sample data coming from that area. Estimators are developed by borrowing strength from auxiliary information gathered from another source of data.

In this research, auxiliary information gathered from a big data source like social media data, especially Twitter. Three characteristics of big data are volume, velocity, and variability. Twitter is micro-blogging service letting people stream information in individual feeds known as tweets [3]. According to [4], tweets can be obtained from a certain location in order to extract auxiliary information from each region. In 2017 according to Beritasatu, Indonesia was on the list of the top five most active users in the world. Then, Katadata in the same year also stated that from the top ten trending topics on Twitter, four of them are about politics. In politics, the benefit of social media for a Leader Candidate is information, service, access to political power, and space [5].

Auxiliary information and survey data are combined by a general linear mixed model to make an estimator for electability. The region is used as a random effect in the model. Each of nirsample region

doesn't have random effect in the model. That condition will cause biased in nirsample area estimation [6]. Some modification in modeling to overcome that problem [6].

Purpose of this research is providing auxiliary information from Twitter for small area estimation, estimate Electability of a Candidate Central Java Governor and compare it with Central Java Governor 2018 election result.

1.1. Small Area Estimation

An area estimate can be done directly or indirectly. A subpopulation or area estimate is usually referred to as a "direct estimate" if it is based only in the specific sample data coming from that area [2]. An area estimate is referred to as "indirect estimate" when it is not only based specific sample data coming from that area. Indirect estimation needs auxiliary information from any source of data. Requirements for auxiliary information is measured without error [1].

There are two types of auxiliary information. The is unit level auxiliary information and area level auxiliary information. Unit level auxiliary information is auxiliary information available for each unit in the population. Area level auxiliary information is auxiliary information available in the form of aggregation in each area [1].

In Indonesia, research about small area estimation usually uses auxiliary information from SUSENAS (*Survei Ekonomi Nasional*) and other administrative data. Small area estimation applied to use auxiliary information from BPS (*Badan Pusat Statistik*) [7].

1.2. Small Area Estimation with Big Data

These days, big data and small area estimation is a rapidly growing topic and will get more attention in the future. Big data have some big potential to use in small area estimation. One of that potential is using big data to provide auxiliary information [8]. Small area estimation was applied to American Community Survey data by using auxiliary information from Google Trends [4]. Small area estimation was also applied by using auxiliary information from big data on an individual's mobility in the region of Tuscany in Italy [8]. Due to technical problems and legal restrictions, it is unfeasible at this stage to have unit level auxiliary data that can be linked with survey data [8]. To overcome this problem, Area level auxiliary information is feasible enough to be linked with survey data.

1.3. Generalized Linear Mixed Binomial Model

Generalized Linear Mixed Binomial Model is used to develop estimator by combining auxiliary information and survey data. Here some explanations about the model :

$$\begin{gathered} Y_i|p_i, n_i \sim Binomial(n_i, p_i) \\ logit(p_i) = \boldsymbol{x'_i}\boldsymbol{\beta} + v_i,\ i = 1,2,3,\dots m \end{gathered} \tag{1}$$

with :

$\boldsymbol{x'_i} = (1 \quad x_{1i} \quad x_{2i} \quad \cdots \quad x_{ki})$

$\boldsymbol{\beta} = (\beta_0 \quad \beta_1 \quad \beta_2 \quad \cdots \quad \beta_k)$

Y_i: binomial random variable at i-th area

n_i: sample size at i-th area

p_i: electability at i-th area

$\boldsymbol{x'_i}$**:** auxiliary information vector at i-th area

$\boldsymbol{\beta}$**:** regression parameter vector

v_i: random effect at the i-th area with $v_i \overset{iid}{\sim} N(0, {\sigma_v}^2)$

m: quantity of observed area

k: quantity of auxiliary information.

Table 1. Auxiliary information.

Notation	Auxiliary Information	Variable's Category
$\boldsymbol{X_1}$	The activity of Candidate A's supporter in Twitter[a]	Big data
$\boldsymbol{X_2}$	The activity of Candidate A in Twitter[b]	Big data
$\boldsymbol{X_3}$	The activity of the Central Java society in Twitter[c]	Big data
$\boldsymbol{X_4}$	Percentage of Candidate A`s oppositional vote of political parties (general election 2014)	KPU
$\boldsymbol{X_5}$	Percentage of Candidate A`s supportive political vote of political parties (general election 2014)	KPU
$\boldsymbol{X_6}$	Percentage of the election attendance (central java governor election 2013)	KPU

[a]$\boldsymbol{X_1}$ is the frequency of emergence of tweets about support to Candidate A, [b]$\boldsymbol{X_2}$ is the frequency of emergence of Candidate A`s activity, [b]$\boldsymbol{X_3}$ is the frequency of emergence of tweets about Central Java society's activity

Some method is compared to estimate parameters from the generalized linear mixed model. They are Pseudo-Likelihood, Laplace integral approximation, and Gauss-Hermit Quadrature integral approximation [9]. Based on some criterion, parameter estimation is best applied on Laplace integral approximation and Gauss-Hermit Quadrature integral approximation when the sample size is small. Pearson Statistics χ^2/df to detect overdispersion in modeling. Overdispersion occurs when variance from data is bigger than the assumed distribution [9].

2. Materials

Data used in this research is survey data about the electability of the candidates of the Central Java Governor. In this survey, Individuals is units, and each region is regarded as a small area. Question asked in the survey is "If Governor election held today who would you choose?" and the options are Candidate A, Candidate B, Candidate C, Candidate D, and refuse to choose. Electability will only be estimated for Candidate A because he is the only interesting candidate in the survey. Auxiliary information used is this research coming from Twitter and KPU(*Komisi Pemilihan Umum*). Data gathered from Twitter are tweets with certain keywords. Tweets were extracted to provide some auxiliary information. Tweets gathered from January 2018 until March 2018. Auxiliary information used can be seen in Table 1.

Information about the sample region and nirsample region can be seen in Table 2. There are 21 sample region and 14 nirsample regions. The sampling method used in the survey was stratified random sampling. The region in Central Java clustered based on political condition, and the clustered region was used as a stratum in stratified random sampling

Table 2. Sample Region and Nirsample Region.

Area	Region
Sample	Banjarnegara, Banyumas, Batang, Blora, Boyolali, Brebes, Grobogan, Jepara, Kebumen, Magelang, Pati, Pemalang, Rembang, Semarang, Temanggung, Sragen, Magelang City, Salatiga City, Semarang City, Tegal City, Pekalongan City.
Nirsample	Cilacap, Demak, Karanganyar, Kendal, Klaten, Kudus, Pekalongan, Purbalingga, Purworejo, Sukoharjo, Tegal, Wonogiri, Wonosobo, Surakarta City.

3. Method

Table 3. Auxiliary information from Twitter

Notation	Auxiliary Information	Keywords
$\boldsymbol{X_1}$	The activity of Candidate A's supporter in Twitter	Hashtag about support for Candidate A
$\boldsymbol{X_2}$	The activity of Candidate A on Twitter	@nameofCandidateA
$\boldsymbol{X_3}$	The activity of the Central Java society in Twitter	Whole keywords used

3.1. Auxiliary Information Extraction

Tweets are obtained from twitter by using the R programming language. One thing needed to gather tweets from each region is geocoded with an approximated radius of the region. Tweets can only be in Obtained in the past seven days, so tweets needed to be obtained routinely in a period of three months. Step needed in the extraction of auxiliary information are :

1. Input keywords, geocode, and radius of the region in an r programming language.
   Keywords used to obtain *tweets* : Pilgub, Pilgub 2018, Pilgub Jateng, Gubernur, Gunernur jateng, Gubernur jawa tengah, Politik, Pilkada, Pilkada jawa tengah , Pilkada 2018 and Hashtag about support for Candidate A.
2. Extract tweets routinely in a period of three month
   Information extracted from Twitter can be seen in Table 3, from all those keywords we will make X1, X2, and, X3. Each auxiliary information contains certain keywords.
3. Compute frequency based on certain keywords to extract auxiliary information in Table 3 by the formula :

$$\hat{f}_i = f_i^{(a)} + \sum_{j=1}^{k_i} f_{ij}^{(b_{j(i)})} / b_{j(i)}\,, i = 1,2,\ldots,35 \;\; j = 1,2,..,k_i \qquad (2)$$

with :

$\hat{f}_i$: approximated frequency of tweets at i-th region
$f_i^{(a)}$: frequency of unsliced tweets at i-th region
$f_{ij}^{(b_{j(i)})}$: frequency of sliced tweets at i-th region which is sliced with j(i)-th neighbour
k_i : quantity of neighbour at i-th region
$b_{j(i)}$: quantity of region in j(i)-th slice

In Figure 1, we can see why equation 2 used. Radius used to obtain tweets are overlapped. So, there is needed some correction to compute frequency.

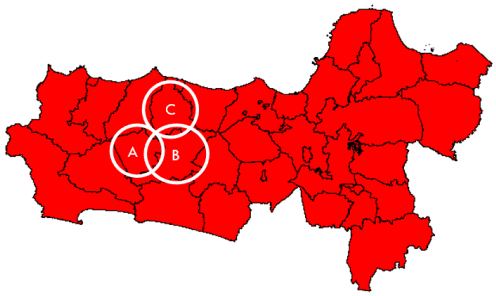


Figure 1. Radius example to obtain tweets.

3.2. Small Area Estimation Method

Here is the procedure to estimate the electability of Candidate A by small area estimation method :

1. Data exploration to get any insight about the relationship between electability and each auxiliary information and also each auxiliary information's characteristic.
2. Modeling to develop estimator

$$\log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i, i = 1,2,\dots,35$$

$$p_i = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}+v_i)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}+v_i)} \quad (3)$$

with :

$\boldsymbol{x}_i' = (1 \;\; x_{1i} \;\; x_{2i} \;\; x_{3i} \;\; x_{4i} \;\; x_{5i} \;\; x_{6i})$

$\boldsymbol{\beta} = (\beta_0 \;\; \beta_1 \;\; \beta_2 \;\; \beta_3 \;\; \beta_4 \;\; \beta_5 \;\; \beta_6)$

p_i : electability of the Candidate A at i–th region

$\boldsymbol{x}_i'$: auxiliary information vector for i-th region

$\boldsymbol{\beta}$: regression parameters vector

v_i : random effect for i-th region

3. Modeling modification for nirsample area. A random effect for nirsample area used random effect from most similar areas. The similarity is determined from an area with the shortest distance in clustering
4. Mse estimation with the bootstrap method
   a. Generate a population with estimated from the indirect estimate
   b. Sampling from the following population
   c. Do step 2 and 3 for data sampled from the generated population
   d. Repeat 4.b and 4.c 1000 times

   MSE estimate [10]:

$$KTG_i{}^B\left(p_i^H\right) = B^{-1}\sum_{b=1}^{B}\left(p_i^{H(b)} - p_i^H\right)^2, i = 1,2,\dots,35,\; b = 1,2,\dots.,B \quad (4)$$

   with :

   $KTG_i{}^B(p_i^H)$: mse estimate for Candidate's A electability at the i-th region

   B: repetition in resampling

   $p_i^{H(b)}$: indirect estimated Candidate's A electability from b-th resampled at the i-th region

   p_i^H: indirect estimated Candidate's A electability at the i-th region

5. Direct estimation at the province level

$$p = \sum_{j=1}^{C}\frac{N_j}{N}p_j, j = 1,2,\dots c \quad (5)$$

with :

$p_j = \frac{y_j}{n_j}$

p: direct estimate Candidate's A electability at the province level

p_j: direct estimate Candidate's A electability at j-th stratum

y_j: Quantity of the respondent who chooses Candidate's A at j-th stratum

n_j: sample size at j-th stratum

Table 4. Estimated parameters in the modelling.

Parameter	Estimated	Stdev	z-score	Pr(>|z|)	VIF
$\boldsymbol{\beta_0}$	-2.573	0.859	-2.875	0.004	
$\boldsymbol{\beta_1}$	0.749	0.213	3.516	0.000	1.097
$\boldsymbol{\beta_3}$	0.803	0.199	4.027	0.000	1.085
$\boldsymbol{\beta_5}$	0.022	0.013	1.697	0.090	1.017
$\boldsymbol{\beta_6}$	0.020	0.008	2.647	0.008	1.016
$\boldsymbol{\sigma^2}$	0.031	0.175			

N_j: Quantity of the voter at j-th stratum
N: Quantity of the voter at the province level
c: number of stratum

6. Indirect estimation at the province level

$$p^H = \sum_{i=1}^{M} \frac{N_i}{N} p_i^H \, , i = 1,2,3, \ldots . 35 \quad (6)$$

with :
p^H: indirect estimate Candidate A's electability at the province level
p_i^H: indirect estimate Candidate A's electability at the i-th region
N_i: Quantity of the voter at the i-th region
N: Quantity of the voter at the province level
M: number of the region in Central Java

Comparing the result with Central Java Governor Election by the correlation coefficient.

4. Results and Discussion

4.1. Data Exploration

Total tweets obtained as long as three months is 273440 tweets. Each of the auxiliary information extracted from tweets has right-skewed distribution. Some region has higher activity on Twitter than any other region. Semarang city is one of the most active Twitter users. Some region near Semarang city also has high activity on Twitter.

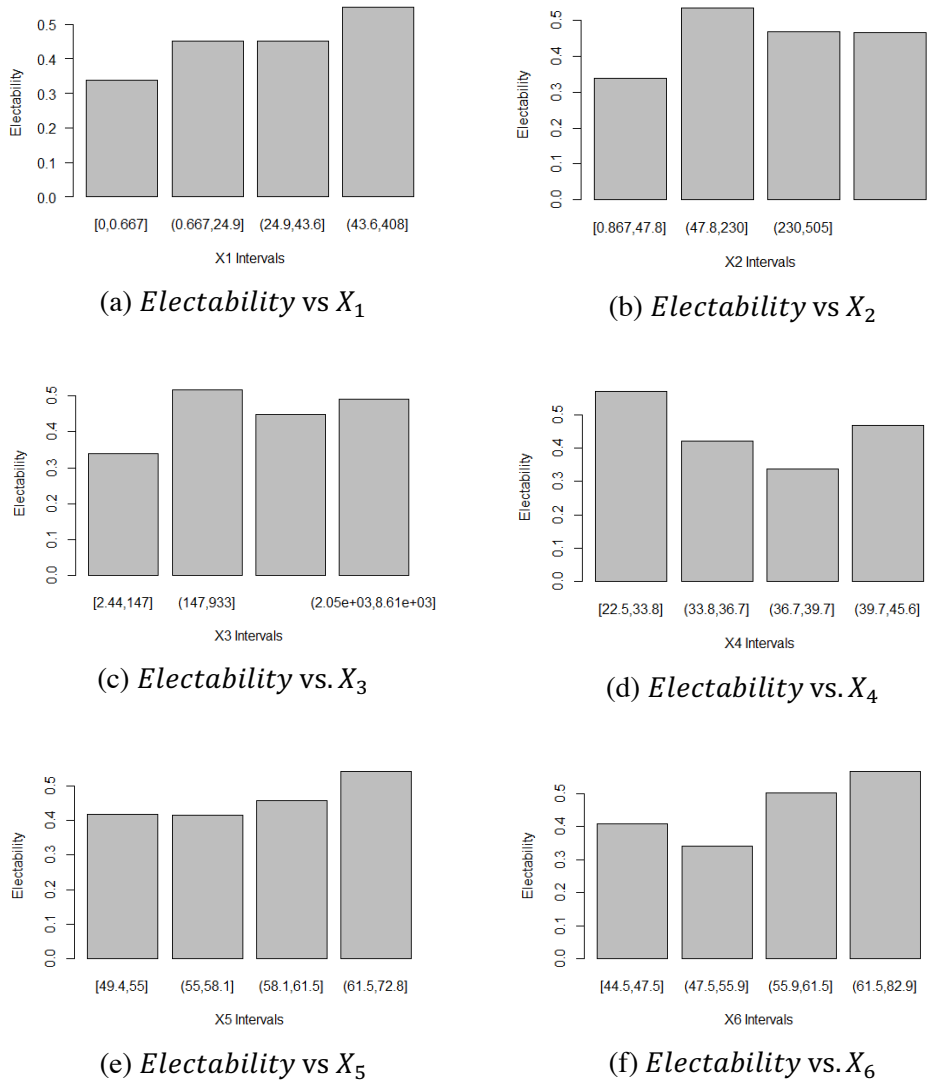
As seen in Figure 2, Relationship pattern between electability and each of the auxiliary information from Twitter seems a positive trend. Relationship between Percentage of Candidate A`s supportive political parties' vote and Percentage of the election attendance with electability is seemed to be positive. The only auxiliary information that has a negative trend is the Percentage of Candidate A`s oppositional political parties'vote. There are needed some treatment in the modeling for this form of a relationship like discretization.

4.2. Modeling to Develop Estimator

Discretization used in the modeling. Purpose of the discretization is to catch relationship pattern Between each auxiliary information extracted from Twitter with electability. Result of the discretization was transformed into a weighing of evidence to make modeling simpler. Table 4 provides result in the, which also include a variable selection in the process.

Two of the three auxiliary variables from twitter have a significant effect on electability at 5% level. They are Activity of Candidate A's supporter in Twitter and Activity of the Central Java society on Twitter. As seen in Figure 2.a and Figure 2.c, both of the auxiliary information has a positive relationship with the electability. Another two auxiliary information included in the model is the Percentage of Candidate A`s supportive political parties' vote (general election 2014) and Percentage of the election

attendance (central java governor election 2013). Both of the auxiliary information has a positive relationship with the electability.



(a) $Electability$ vs X_1 (b) $Electability$ vs X_2

(c) $Electability$ vs. X_3 (d) $Electability$ vs. X_4

(e) $Electability$ vs X_5 (f) $Electability$ vs. X_6

Figure 2. Relationship between variables.

4.3. Estimator used for Candidates A's electability

$$p_i^H = \frac{\exp(-2.573 + 0.749WoE_{1i} + 0.803WoE_{3i} + 0.022x_{5i} + 0.020x_{6i} + \hat{v}_i)}{1 + \exp(-2.573 + 0.749WoE_{1i} + 0.803WoE_{3i} + 0.022x_{5i} + 0.020x_{6i} + \hat{v}_i)}$$

$i = 1,2,\ldots,35$ (7)

with :

p_i^H: indirect estimated of the electability for the i-th region

WoE_{1i}: the weight of evidence transformation for X_1 auxiliary information for the i-th region

WoE_{3i}: the weight of evidence transformation for X, auxiliary information for the i-th region
x_{5i}: Percentage of Candidate A`s supportive political parties' vote (general election 2014) for the i-th region
x_{6i}: percentage of the election attendance (central java governor election 2013) for the i-th region
$\hat{v}_i$: estimated random effect for the i-th region.

Direct estimate electability at province level is 46.83 % while indirect estimate electability is 46.68 %. Both estimates had an almost equal result at the province level. Indirect estimate electability available at the attachment

4.4. Indirect Estimate Electability Comparison with Election's Vote

Indirectly estimated electability would be compared with Governor Election's vote by correlation in several ways. Purpose of this comparison is to evaluate how good is the result. First, here is the scatter plot between Indirectly estimated electability and Governor Election's vote.

From Figure 3, it can be seen that both of indirectly estimated electability and Governor Election's vote have the same direction but there are four regions have suspicious indirectly estimated electability because the values are way too small compared to Governor Election's vote. Correlation between those two is 0.1887802. The correlation is still too small. When the correlation without those four suspicious indirect estimated electability, the correlation is 0.4372723.

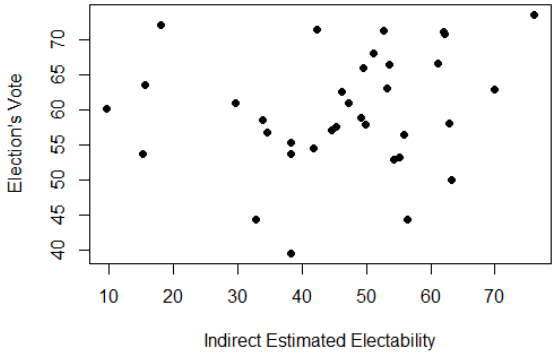


Figure 3. Scatterplot between Indirectly estimated electability and Governor Election's vote.

We were trying to explore more of the advantages of the indirect estimation result. We mark some region that has high activity on Twitter. We separate the region that has a number of tweets of more than 100 tweets. Correlation between Indirectly estimated electability and Governor Election's vote in an area that highly active on Twitter is 0.4184.

5. Conclusion

Estimated electability from direct estimation and indirect estimation at the province level have equal precision but the indirect estimation outperformed the direct estimation at region level because direct estimation can't give any estimations. An indirect estimation has some advantages in a certain condition. There are four regions that have suspicious indirect estimated electability. Indirect estimation performed better when those region excluded from the comparison. Another advantage of the indirect estimation is an indirect estimate performed better in a region that highly active on Twitter.

References

[1] Rao J N K and Molina I 2015 *Small Area Estimation* (New Jersey: John Wiley & Sons, Inc)
[2] Notodiputro K A and Kurnia A 2007 Development Of Small Area Estimation Research In Indonesia *ICoMs*
[3] Kwartler 2017 Text Mining in Practice with R (Chennai: John Wiley & Sons, Inc)
[4] Porter AT, Holan S H, Wikle C K, Cressie N 2013 Spatial Fay-Heriot Model for Small Area Estimation with Functional Covariates *NIASRA* 1208-13
[5] Ardha B 2014 Sosial Media Sebagai Media Kampanye Partai Politik 2014 di Indonesia *Jurnal Visi Komunikasi* 13105-120
[6] Anisa R, Kurnia A and Indahwati 2014 Cluster Information of Non-Sampled Area In Small Area Estimation *IOSR-JM* 1015-19
[7] Sadik K and Notodiputro K A 2008 Small Area Estimation with Time and Area Effect Using Dynamic Linear Model *ICoMs*
[8] Marchetti S, Giusti C, Pratesi M, Salvati N, Giannotti F, Padreschi D, Rinzivillo S, Pappalardo L and Gabrielli L 2015 Small Area Model-Based Estimators Using Big Data Sources *JOS* 31263-281
[9] Sroup W W 2013 *Generalized Linear Mixed Models Modern Concepts, Methods, and Applications* (Boca Raton: CRC Press)
[10] Gonzâlez-Manteiga W, Lombardfa MJ, Molina I, Morales D and Santamaria L 2007 Estimation of The Mean Square Error of Predictors of Small Area Linear Parameters Under a Logistic Mixed Model *Computational Statistics & Data Analysis* 512720-1733

Appendices

Table A1. Electability Estimation Result.

Region	Area	Direct Estimate(%)	Indirect Estimate(%)	Root MSE(%)
Kabupaten Banjarnegara	Sample	-	41.81	3.3
Kabupaten Banyumas	Sample	-	49.26	5.9
Kabupaten Batang	Sample	-	15.55	4.8
Kabupaten Blora	Sample	-	45.35	3.1
Kabupaten Boyolali	Sample	-	52.78	4.5
Kabupaten Brebes	Sample	-	38.21	4.8
Kabupaten Grobogan	Sample	-	61.1	5.1
Kabupaten Jepara	Sample	-	42.35	3.2
Kabupaten Kebumen	Sample	-	32.86	5.9
Kabupaten Magelang	Sample	-	62.9	4.8
Kabupaten Pati	Sample	-	33.9	3
Kabupaten Pemalang	Sample	-	34.53	3.4
Kabupaten Rembang	Sample	-	62.04	11.2
Kabupaten Semarang	Sample	-	53.5	4.9
Kabupaten Sragen	Sample	-	38.28	3.5
Kabupaten Temanggung	Sample	-	53.15	5.8
Kota Magelang	Sample	-	49.5	5.7
Kota Pekalongan	Sample	-	46.17	6.3

Region	Area	Direct Estimate(%)	Indirect Estimate(%)	Root MSE(%)
Kota Salatiga	Sample	-	51.17	5.7
Kota Semarang	Sample	-	62.27	5
Kota Tegal	Sample	-	15.28	5.3
Kabupaten Cilacap	Nirsample	-	38.24	3.7
Kabupaten Demak	Nirsample	-	55.84	3.7
Kabupaten Karanganyar	Nirsample	-	44.58	4.2
Kabupaten Kendal	Nirsample	-	29.74	5.1
Kabupaten Klaten	Nirsample	-	47.22	4.1
Kabupaten Kudus	Nirsample	-	76.1	8.8
Kabupaten Pekalongan	Nirsample	-	9.66	4
Kabupaten Purbalingga	Nirsample	-	63.25	11.1
Kabupaten Purworejo	Nirsample	-	54.32	4.1
Kabupaten Sukoharjo	Nirsample	-	49.89	3.6
Kabupaten Tegal	Nirsample	-	56.34	9.7
Kabupaten Wonogiri	Nirsample	-	69.94	8
Kabupaten Wonosobo	Nirsample	-	55.07	3.6