Bias Not Linear Prediction Method The Best for Small Area on Empirical Suspect Spending Per Capita (Case: South Sulawesi Province District Gowa)

M Nusrang1*, Sudarmin1, S Annas1, Asfar1, Z Rais1
1Department of Statistics Faculty of Mathematic and Natural Science, Mathematics, Universitas Negeri Makassar, Indonesia
*e-mail: Muh.nusrang@unm.ac.id

Abstract. Indirect estimation is the estimation of parameters that use additional information about the same parameters in a small area. The combination of the basic assumption of random effects and fixed effects forms a mixed influence model. In this study, the data used are secondary data from the SUSENAS and Regression methods. The value of Average Root Mean Square Error (ARSME) estimator in the EBLUP estimator is smaller when compared to the direct estimator and simple linear regression. So it can be said that the EBLUP estimator is better than the direct estimator of per household at the District level in Gowa Regency based on SUSENAS data for 2017.

Keyword: Bias not linear prediction, small area, empirical suspect spending

1. Introduction
Small area statistics are now taking the attention of world statisticians very seriously. Much research has been developed both for the improvement of techniques and the development of methods and applications in a variety of real cases and problems encountered. [1] Moreover in Indonesia with the era of regional autonomy, where the constitutional system shifted from a centralized system to a decentralized system. In a decentralized system, local governments have greater authority to regulate themselves, especially at the district / city level. [2] Thus the statistical needs up to the village level become a basic need as a basis for district / city governments to develop a planning, monitoring and evaluation system for regional development or other important policies [3]

The problem arises when going to guess at the village level, which is a small sample size condition because the survey which is the source of information is generally designed to estimate population parameters on a national scale. [4] This condition cannot provide an accurate and precise guessing result. Therefore we need a statistical procedure that can combine data from small and large samples, by taking advantage of the details in the sample survey and census. The right method to provide a solution in this case is the Small Area Estimation [5] [6]

Model which is the basis for estimating small areas assumes that the random effect of the area error is mutually independent. [7] But in some cases, this assumption is often violated. The reason is the diversity of an area influenced by the surrounding area, so that spatial influences can be included in random effects. Spatial influence is a common thing that occurs between one areas with another area, this means that one area affects other areas. [8]

The purpose of this study is to estimate per capita expenditure in Gowa Regency, South Sulawesi Province.
2. Literature Review

2.1 Small area models
Small areas describe a small subpopulation for demographics or groups of people who have socioeconomic (race, sex, age) who are in a wider geographical area [9]. Small areas are often used as a description of a small geographical area, such as districts, sub-districts, or villages. [10]

Types of small area estimation models are divided into two, namely the basic area level models and unit level models [11] [12]. The area level model is used if accompanying data that matches the observed variable data is not available up to the sample unit level, while the unit level model is used if the accompanying data corresponding to the observed variable data is available up to the sample unit level. [13] [14]

2.2 Predictions of the Best Empirical Linear Bias (EBLUP)
If there is a model \( \theta_i = x_i^T \beta + v_i \), where \( \theta_i \) is a parameter of concern and \( y_i \) is the direct estimator value based on the survey design, then: [7]
\[
y_i = \hat{\theta}_i + e_i = x_i^T \hat{\beta} + v_i + e_i
\]
for \( i = 1, \ldots, m \) with \( x_i = (x_{i1}, x_{i2}, \ldots, x_{ip}) \) is a companion variable at the regional level, \( \beta \) is a fixed parameter, \( v_i \) is a small area random effect with \( v_i \sim iid N(0, \sigma_v^2) \), \( e_i \) is an example withdrawal error with \( e_i \sim iid N(0, \psi_i) \), \( e_i \) and \( v_i \) independent. Assume that \( \beta \) and \( \sigma_v^2 \) (variance between small area) unknown, but \( \psi_i \) for \( i = 1, 2, \ldots, m \) known. [15]

Model completion techniques in Equation (1) to obtain BLUP for \( \theta_i = x_i^T \beta + z_i v_i \) has been developed by Henderson [3], with assume \( \sigma_v^2 \) known. Estimation BLUP from \( \theta_i \) based equation (1):
\[
\hat{\theta}_i^{BLUP} = x_i^T \hat{\beta} + y_i (y_i - x_i^T \hat{\beta})
\]
\[
\hat{\theta}_i^{BLUP} = y_i y_i + (1 - y_i) x_i^T \tilde{\beta}
\]
with \( y_i = \sigma_v^2 / (\sigma_v^2 + \psi_i) \). The BLUP method developed by Henderson assumes the knowledge of the various random influence components in a linear mixed model, when in fact these variance components are unknown. As a result, the variety of random effects must be suspected. [16] [17] Conducted a review of several methods for estimating variance components, using the maximum likelihood (ML) method and the restricted maximum likelihood (REML) method. Estimation of \( \sigma_v^2 \) with both the ML and REML methods is carried out using the Fisher scoring algorithm. [18] By replacing \( \sigma_v^2 \) with \( \tilde{\sigma}_v^2 \) we get a new estimator of EBLUP, as follows:
\[
\hat{\theta}_i^{EBLUP} = \tilde{y}_i y_i + (1 - \tilde{y}_i) x_i^T \tilde{\beta}
\]

3. Methodology
The data used in this study are secondary data from the SUSENAS and PODES of Gowa Regency in 2017. The data analysis techniques used in this study are as follows: [9]
1) Preparation Stage
   a. Prepare data as a direct estimator
   b. Prepare data used as a companion variable based on a questionnaire obtained from BPS in collecting PODES data for 2017.
   Prepare variance data from response variables
2) Estimating Phase
   a. Estimating direct multiple linear regression and EBLUP for each district
   b. Estimating MSE for the direct estimation method, multiple linear regression and EBLUP for each district
   c. Evaluate the estimation results by comparing the average root mean square error (ARMSE) estimator.
4. Result and Discussion

The variables observed in this study were expenditure per household per month for sub-districts in the Gowa Regency. Whereas for auxiliary variables (accompanying variables) in expenditure issues can be seen from several proxies (approaches) namely health and income. From the health side, the number of families that received JAMKESMAS / JAMKESDA cards was used and the number of families living in slums was used. From the income side, the number of families in which there are family members becomes farm laborers, the number of families receiving SKTM letters. Thus, additional variables used are the number of families working as farm laborers (x_1), the number of families that receive JAMKESMAS / JAMKESDA cards in a year (x_2), the number of families who receive a SKTM card in a year (x_3), and the number of families living in slums (x_4).

![Figure 1. Shows the distribution of household expenditure increased](image)

Based on Figure 1 above, it can be seen that the average level of sub-district level household expenditure in Gowa Regency approaches the normal distribution. From the picture, information is obtained that there is an average household expenditure that includes extreme data.

After going through the data exploration process, estimation of small areas is done by direct methods, simple linear regression and EBLUP, and the results can be seen in Table 1 below.

**Table 1. Estimating expenditure per household at District level in Gowa Regency based on SUSENAS data for 2017 (Thousand Rupiahs)**

| Regency | Direct | Regresi | EBLUP  |
|---------|--------|---------|--------|
| 1       | 2828.036 | 2228.799 | 2436.944 |
| 2       | 1708.369 | 1054.567 | 1443.140 |
| ...     | ...     | ...     | ...    |
| 17      | 1568.187 | 1592.347 | 1545.597 |
| 18      | 2367.502 | 1825.926 | 2164.218 |

From Table 1 it can be seen that all estimators show that the district with code 1 has the highest average expenditure per household compared to other districts. Meanwhile, the sub-district that has the lowest average expenditure per household is the sub-district with code 7.

**Table 2. Estimator ARMSE (Average Root Mean Square Error)**

| Estimator          | ARMSE   |
|--------------------|---------|
| Direct             | 719.0664|
| Multiple Regression| 607.4732|
| EBLUP              | 371.3421|

To evaluate the best small area estimation method, it can be seen in Table 2, which shows that the Average Root Mean Square Error (ARMSE) estimator value of the EBLUP estimator is smaller when
compared to the direct estimator and simple linear regression. So it can be said that the EBLUP estimator is better than the direct estimator of per household at the District level in Gowa Regency based on SUSENAS data for 2017.

5. Conclusion
Based on the description in the previous section above it can be concluded that the EBLUP estimator is a good estimation in overcoming the small sample size by giving a better ARMSE value compared to the direct and regression estimator.

In this study, we have not considered the use of weighting from BPS and spatial effects which in the results seen in several adjacent areas there is a tendency to have similar expenditure indicators. So there is a need for further studies to pay attention to the weighting of BPS and spatial effects in order to obtain a more comprehensive estimation of small areas. Furthermore, there needs to be a study of the comparison of similar methods both based on conventional models and based on robust methods.

References

[1] H. Folmer and J. Oud, "How to ged rid of W: a latent variables approach to modelling spatially lagged variables," *Environment and Planning*, vol. 40, pp. 25-38, 2008.
[2] B. Singh, K. Shukla and D. Kundu, "Spatial-temporal models in small area estimation," *Survey Methodology*, vol. 31, pp. 183-95, 2005.
[3] A. Kurnia, "Prediksi Terbaik Empirik untuk model transformasi logaritma didalam pendugaan Area Kecil dengan penerapan pada data sensus," Bogor: Program Pascasarjana, Institut Pertanian Bogor, Bogor, 2009.
[4] M. Ghosh and J. Rao, "Small Area Estimation: an appraisal (with discussion)," *Statistical Science*, vol. 9, pp. 55-76, 1994.
[5] C. Elbers, J. Lanjouw and P. Lanjouw, "Micro-Level Estimation of poverty and inequality," *Econometrica*, vol. 71, 2003.
[6] J. Rao, Small Area Estimation, New York: John Wiley and Sons, 2003.
[7] I. Molina, N. Salvati and M. Pratesi, "Bootstrap for estimating the MSE of the Spatial EBLUP Working Paper," *Statistics and Econometrics Series*, vol. 08, 2007.
[8] R. Fay and R. Herriot, "Estimates of income for small places an application of james stein procedures to census data," *Journal American Stat*, vol. 74, pp. 69-77, 1979.
[9] Badan Pusat Statistik, "Pengeluaran Perkapita," 2015. [Online]. Available: www.bps.go.id/subjek/view/id/5.
[10] J. Rao, "Inferential issues in model based small area estimation: some new developments," *Statistics in Transition New Series and Survey Methodology*, vol. 16, pp. 491-510, 2014.
[11] J. Rao, "Some Recent advances in model based small area estimation," *Survey Methodology*, vol. 25, pp. 175-86, 1999.
[12] N. Chand and C. Alexander, Using Administrative records for small area estimation in the American Community survey, United State America: US Bureau of the census, 1995.
[13] H. Chandra, N. Salvati and R. Chambers, "Small are estimation for spatially correlated populations a comparison of direct and indirect model based methods," *Statistics in Transition*, vol. 8, no. 3, pp. 31-50, 2007.
[14] D. Pfeffermann, "Small area estimation, new developments and directions," *International Statistical Review*, vol. 70, pp. 125-43, 2002.

[15] D. Zimmerman and N. Cressie, "Mean Squared prediction error in the spatial linear model with estimated covariance parameters," *Annals of the Institute of Statistical Mathematics*, vol. 44, pp. 27-43, 1992.

[16] R. Pace and R. Barry, "Spare Spatial Autoregressions," *Statist & Probab Letters*, vol. 33, pp. 291-7, 1997.

[17] N. Prasad and J. Rao, "The Estimation of the mean squared error of small area estimators," *Journal American Stat*, vol. 70, pp. 163-71, 2002.

[18] M. Pratesi and N. Salvati, "Small Area estimation the EBLUP estimator based on spatially correlated random effects," *Statistics, Mathematics & Application*, vol. 17, pp. 113-41, 2008.