Small Area Statistic in Indonesia

Khairil A. Notodiputro and **Anang Kurnia**
(khairiln@bima.ipb.ac.id)
Department of Statistics, Institut Pertanian Bogor
Jl. Meranti Wing 22 Level 4
Kampus IPB Darmaga, Bogor 16680
INDONESIA

Abstract

Formerly small area statistic in Indonesia is produced using census data which were broken down into smaller domain such as province or district. These data has been published in Province Statistic or District Statistic. It is uneasy to determine how the small area statistic was first applied in Indonesia. However, the literature showed that small area estimation was first introduced by Smeru Research Institute in collaboration with BPS (Statistics Indonesia) to produce poverty maps at three provinces in Indonesia as a pilot project on 2001 – 2003. Our research in SAE is started from 2006. Two sources of main data are used in our study: (i) National Socio-Economic Survey and (ii) Village Census. These data were collected by BPS.

Keywords: small area estimation, Indonesia case data

1. Introduction

Formerly small area statistic in Indonesia is produced using census data which were broken down into smaller domain such as province or district. These data has been published in Province Statistic (*Propinsi dalam Angka*) or District Statistic (*Kabupaten dalam Angka*).

It is uneasy to determine how the small area statistic was first applied in Indonesia. However, the literature showed that small area estimation was first introduced by Smeru Research Institute in collaboration with BPS (Statistics Indonesia) to produce poverty maps at three provinces in Indonesia (DKI Jakarta, East Java and East Kalimantan) as a pilot project. The ELL (Elbers, Lanjou and Lanjou) method was applied using consumption model to produce poverty maps. In principle, this method combines detailed information collected from a household survey with the complete population coverage of a population census. The pilot project has been done during 2001 – 2003.

Research on small area estimation in Indonesia, especially at IPB, has been started since 2006 and carried out through support from DGHE with title: **Small Area Estimation Models and Its Application for BPS Data**. Two sources of main data are used in our study: (i) SUSENAS (National Socio-Economic

Survey) and (ii) PODES (Village Census). These data were collected by BPS. The household data was obtained from the SUSENAS, while the sub-district (village) data was obtained from the PODES.

SUSENAS, the National Socio-Economic Survey, is a nationally representative household survey, covering all areas of the country. One part of the SUSENAS is conducted every year, collecting information on the characteristics of over 250,000 households and 1,000,000 individuals. This part of the SUSENAS is known as the core of SUSENAS. Another part of the SUSENAS is conducted every three years, specifically collecting information on very detailed consumption expenditure from around 65,000 households. This is known as the Consumption Module of SUSENAS. The sample households are randomly selected as a subset of the 250,000 households in the Core SUSENAS sample of the same year.

PODES, meanwhile, is a complete enumeration of village data throughout Indonesia. The information collected through this village census includes village characteristics such as size of area, population, infrastructure and local industries. The information is obtained from official village documents as well as interviews with village officials. The PODES survey is usually conducted three times in every ten years, usually prior to and as a preparation for an agricultural census (year ending with the digit '3'), an economic census (year ending with the digit '6'), and a population census (year ending with the digit '0').

2. Application SAE in Indonesia Data

There are two main problems faced in applying small area estimation concept in Indonesia case, especially for SUSENAS and PODES data. The first is model pattern of auxiliary variables and response variables which may not be linearly related. Secondly, the ratio between small area variances the sampling error variances.

To overcome the problems, we have proposed the generalized additive mixed models (Kurnia and Notodiputro, 2007); a nonparametric approach (Kurnia, Notodiputro and Ibrahim, 2007) and modified GREG approach (Kurnia, Sartono and Wulandari, 2007).

In the context of design-based method, we arrange optimum weights based on multistage sampling. Furthermore, we also used GREG approach combined with robust regression to estimate the regression parameter. The result showed that the methods produced significant improvement at least when applied to the Indonesia data.

In the context of model-based method, we also proposed generalized additive mixed model (GAMM) for SAE. The GAMM approach has demonstrated significant advantages over its parametric approach to model auxiliary variables, and then we adopted this approach in small area estimation. We have considered an extension of the Fay-Herriot model for the basic area level model

$$y_i = x_i'\beta + \upsilon_i + e_i \text{ , } i = 1, 2, ..., k$$

where β is coefficient regression parameters, υ_i are random effect area, and e_i are sampling errors. We also assume $e_i \sim (0, D_i)$, $\upsilon_i \sim (0, A)$ and that they are independent. D_i is usually assumed to be known, see Rao (2003). We assume that y_i and x_i are related by a smooth function m(.). Let X be the random vector of predictors, thus

$$y_i = m(x_i) + \upsilon_i + e_i \text{ , } i = 1, 2, ..., k$$

where $\upsilon_i|X \sim (0, \upsilon(x_i))$, $e_i \sim (0, D_i)$, and e_i and υ_i are independent. The small area mean functions

$$\theta_i(x_i) = m(x_i) + \upsilon_i$$

is linear combination of mean $m(x_i)$ and the random effects υi.

Some authors have started to adopt nonparametric approaches in SAE. Zheng and Little (2004) proposed a model-based estimator for cluster sampling, in which the regression model combines a spline model with a random effect for the cluster. Opsomer et.al. (2008) showed how the inclusion of a spatial spline can improve the fit relative to a model which only uses a random effect for the small areas, as would be done in traditional small area estimation. We adopted method of the Opsomer (2008) in our research. The result showed slight improvement for our case and it is not satisfactory.

3. The importance of SAE in Indonesia (Case in poverty mapping and hotspot geoinformatics)

Poverty mapping is useful in devising policy strategies for tackling poverty. There are four major classes of poverty mapping measures identified on the basis of data sources, assumptions, and the statistical routines utilized: econometric, social, demographic, and vulnerability-based measures (Davis 2003). Econometric poverty indicators are current consumption expenditures, income, and wealth. Examples of social indicators are nutrition, water, health, and education. Demographic measures may use gender, health, and household age structure indicators. Vulnerability measures are concerned with the level of household exposure to shocks (Henninger 1998).

SAE combines survey and census data to estimate welfare or other indicators for disaggregated geographical units such as sub-districts or villages. With the use of detailed household survey data, the resulting parameter estimates are used to weight the direct and synthetics estimator of a target population in order to determine its expected poverty level. For example, using a set of explanatory variables that are common to both the household survey and the census (for example household size, infrastructure characteristics, and demographic variables), parameters of SAE model are estimated and applied to calculate (we can use GREG, modified GREG, synthetic, or mixed model based) the interested variable characteristics in each small area.

SAE can be used with household unit (unit level) data or community level (area level) data. The first method uses census data and survey data of the same period, while the second method uses average values from communities and small towns. The principles utilized in the two methods for calculating total per-capita consumption or any other poverty measure used are the same. In estimating the per-capita consumption, the parameters required can be categorized into household-level characteristics and geographic-level characteristics.

Other important features of SAE technique is the ability to estimate area parameters although the availability of data in the related area is extremely limited. The estimator can be provided by using synthetic SAE methods which are based on borrowing strength from other areas. Thereby, in general SAE can provide data for every small area of interest. Based on this condition, we can estimate the parameter of interest in each area (villages or sub-districts), followed by the geo-informatics concepts (Patil and Tallie, 2004) that can be used for hotspot detection and other related scan statistic.

Geo-informatics is a method to describe and analyse data expressing any geographic (area in SAE) information. A hotspot means unusual phenomenon, anomalies, aberrations, outbreaks, elevated clusters, or critical areas. Hotspot geo-informatics is a method to detect unusual phenomenon geographically.

References

Davis, Benjamin. 2003. Choosing a method for poverty mapping, March 2006. http://www.povertymap.net/publications/doc/CMPMDAVIS13apr03sec.pdf.

Fay, R.E. and Herriot, R.A. 1979. “Estimates of income for small places: an application of James-Stein procedures to Census data”. Journal of the American Statistical Association, Vol. 74, p:269-277.

Henninger, N. 1998. Mapping and Geographic Analysis of Human Welfare and Poverty: Review and Assessment. Washington, DC: World Resources Institute. http://www .povertymap.net/ publications/doc/henninger

Kurnia, A. and Notodiputro, K.A. 2007. Generalized Additive Mixed Models for Small Area Estimation. Proceeding at the 2nd International Conference on Mathematical Sciences 2007 (ICoMS-2007), 28 - 29 May 2007. Universiti Teknologi Malaysia

Kurnia, A., Notodiputro, K.A. and Ibrahim,N.A. 2007. A Nonparametric Approach in Small Area Estimation. Proceeding at the ICCS-IX, 12 - 14 December 2007. Universiti of Malaya, Shah Alam – Malaysia

Kurnia, A., Sartono, B. and Wulandari, R. 2007. Pengaruh Misspesifikasi Desain Survey pada Pendugaan Area Kecil dengan Pendekatan *Generalized Regression. Prosiding pada* Seminar Nasional Matematika dan Pendidikan Matematika, 24 November 2007. Universitas Negeri Yogyakarta

Patil, G. P. and Taillie, C. 2004. Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics* 11(2): 183-198.

Rao, J.N.K. 2003. Small Area Estimation, New York : John Wiley and Sons.