ECONOMIC JOURNAL OF EMERGING MARKETS April 2009 1(1) 27-35

SPATIAL DISTRIBUTION BASED ON SEMIVARIOGRAM MODEL

Gandhi Pawitan
Fakultas Ilmu Sosial and Ilmu Politik
Universitas Katolik Parahyangan
E-mail: gandhi_p@home.unpar.ac.id

Abstract

This article aims to discuss some aspects in conducting inferential analysis of census data. In this analysis, the assumptions of normality and IID (independently and identically distribution) in the observations are no longer realistic. Hence conventional analyses which are based on these assumptions are invalid and unreliable. Other alternatives can be considered, such as semivariogram analysis. Semivariogram analysis assumes that observations are dependent geographically. The analysis is useful in understanding spatial distribution of characteristics under investigation.

Keywords: *census, aggregation, semivariogram, autocorrelation, spatial distribution*
JEL classification numbers: *C89, J11*

INTRODUCTION

Census is a process in collecting data which involves all observations in a population. Census data represent important information which is crucial in decision making process. They become widely used following the need in decision making process.

Analyses on census data are commonly descriptive in nature. An example for such analyses is the statistical estimation such as the estimation of total unemployment rate in a certain period, population percentage in the labor force, and the difference between male and female in the labor force.

Shen (2006) estimates the urbanization in China based on a census data from 1982-2000. He uses the estimation results to analyze the linkages between the changes in urbanization level and economic development in the region. Pont (2007) illustrates the use of descriptive information in a survey on income in the U.K., which is essential in the policy and decision making.

Descriptive approach aims to describe the characteristics of a population. The description includes statistical or graphical summary about the population. In another situation, the decision making process needs not only the description about a population, but also the linkages among the characteristics of variables under investigation. In such situation, one might use inferential approach which involves analysis of correlation or causality among variables. As an example, in a survey on unemployment, the estimation of the volume and level of unemployment are important in describing policy decision making. In addition, one also needs the information about the linkages between the changes in unemployment level and macroeconomic factors such as inflation and money supply; and demographic factors such as education and other courses levels.

Suryahadi et al. (2005) construct poverty map in Indonesia. They use both census and sample data which cover various level of government level which are pro-

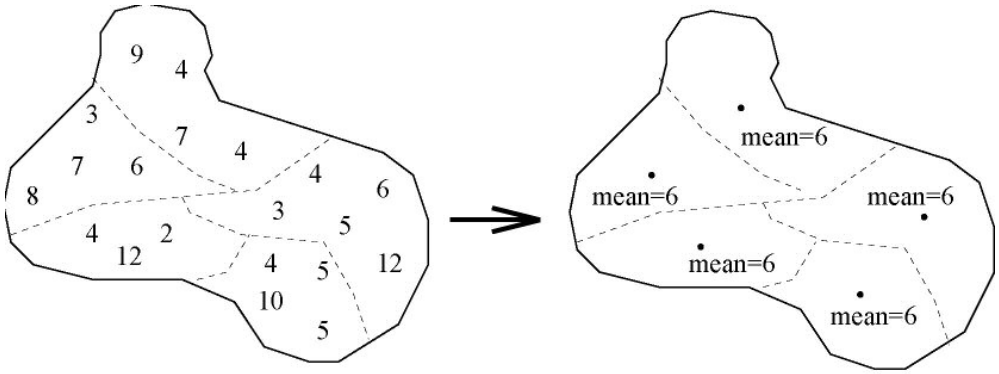
vince, district, and sub-district. The construction involves the description of poverty level. In addition, in defining poverty rate in one level, it considers inferential from several factors that contribute to the poverty rate such as family background, health, education, and economic access.

Ahmad and Lerrut (2000) present and conduct descriptive analysis on energy consumption in province level. Based on the descriptive results, they construct arguments for the need of compensation that can be distributed to as far as all the households. Implicitly, they show the linkages between energy consumption in province level and the need of compensation to the households.

Frees (2006) forecasts participation rate of labor force using data of U.S. Bureau of Labor Statistics to describe the future labor force. In conducting the forecast, he includes some variables of demographic factors, namely age composition, gender, and marital status. He also explains the importance of labor force participation rate projection on the social security financial planning

Hierarchy structure in a population is important in data analysis. There are two issues regarding the hierarchy structure:

i. In reality, it is difficult for the census to meet the identically and independently distribution in the observations. As an example, there exists spatial dependency across observations (Manley et al., 2006).
ii. The loss of variation in observation as the result of aggregation process. Figure 1 illustrates how variation in unit level becomes zero if it is aggregated in the group level. The loss in variation leads to different analysis results in the group level compares to those of individual level (Openshaw, 1984).



(i) Unit level (ii) CD level

Figure 1: Illustration of Loss in Variation Measurement in the Process of Aggregation from Unit Level to the Group Level.

It can be inferred from both issues that there are some important factors in analyzing census data, namely elementary unit which construct the population, hierarchy structure in the population, characteristic definition, and measurement. Elementary unit definition is the foundation in measuring the characteristic under investigation. Hierarchy structure in a population describes the interconnection among variables in the structure. Hierarchy structure also shows geographical position of each elementary unit in the population.

Geographical position shows interactions across elementary units (Pawitan and Steel, 2006). Therefore, the data are not independent because of its geographical position. In social studies, it is often found that social interaction among individual affects other individual which is close to each other.

To overcome such problem, this paper uses semivariogram analysis. It applies to data in the group level, where each group has its own geographical position as the latitude and longitude coordinates. This analysis can explain the distribution of geographical characteristics which enable to interpret the dependence in a certain location.

This paper aims to explain the method of inferential analysis in census data. It focuses on spatial distribution of labor force participation rate. This paper applies semivariogram model to analyze the problem. The next section explains research method, followed by the analysis result.

Spatial distribution in this paper considers only one factor, namely labor force participation rate, assuming that all other factors are constant.

METHODS

This paper explains the spatial distribution based on interdependency in the observations of the social characteristic under investigation. Analyses on census data generally explain a population using the method of estimation and hypothesis testing, and minimizing the possible bias (Brunsdon, 1995). The standard method of statistics assumes that elementary unit meauserements are independent. This assumption simplifies mathematical derivation of estimation and hypothesis testing procedures. However, census data are commonly resulted from a complex process involving stratification or considering hierarchy structure in a population (Skinner et al., 1989). The application of statistical method assumes that the data follows a normally, independently and identically distribution to get valid and reliable results (Haslett, 1992).

Descriptive analysis assumes that a population is fixed and finite which means that population parameters are not random variables. However, inferential analysis assumes that parameters of a population are random variables (Arbia, 1993). Skinner et al. (1989) considers such data analysis as analysis through super population approach.

Hierarchy structure in a population and the assumption of spatial autocorrelation between observations make the data become dependence on each other. Since the assumption of independency is not hold, weaker assumption is needed.

Interactions across observations lead to autocorrelation between the observations. This can be formulated as spatial correlation, which is a function of the distance between elementary units as follows:

$$\rho(d) = \frac{C(d)}{\sigma^2}$$

where d is the distance between elementary units, $C(d)$ is covariogram, and σ^2 is variation.

Variogram can be defined as the difference between two observations from different location (Pawitan and Steel, 2006), namely $Var(y_i - y_j)$. Semivariogram is a statistic shows the average of variation of the

difference between two observations, ½γ. The estimator of semivariogram is formulated as follows: $\hat{\gamma}_{ij} = \frac{1}{2}(y_i - y_j)^2$.

Semivariogram analysis is conducted by modeling the value of semivariogram $\hat{\gamma}_{ij}$ on the distance between pairs of observations (d_{ij}). The distance, d_{ij}, is the Eucledian distance between two observations of y_i and y_j. One of the commonly used models is exponential model, as follows:

$$\hat{\gamma}(d_{ij}) = \hat{n} + (\hat{s} - \hat{n}).\left(1 - \exp\frac{-d_{ij}}{\hat{r}}\right);\ d_{ij} \geq 0$$

where $\hat{n}, \hat{s},$ and $\hat{r}$ are nugget, sill, and range, respectively. Figure 2 shows that the curve will reach a certain distance. The distance shows the limit of the dependency between individual observations. Observations in the range $\hat{r}$ show a certain dependency. Observations outside the range are independent of each other.

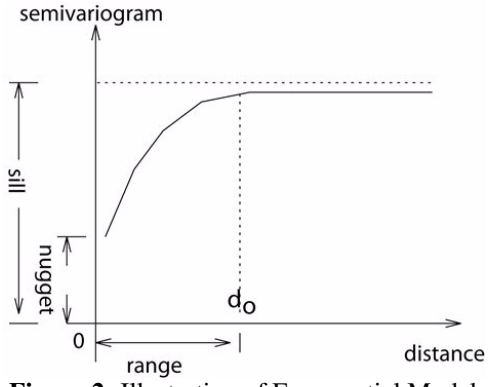


Figure 2: Illustration of Exponential Model in Semivariogram Analysis

Semivariogram model can describe spatial dependency between observations. As an example, in measuring individual preference, the model can provide information whether individual preference influences that of other individual, shown by range and sill.

In conventional statistics methods assuming IID in the data, $\rho(d)=0$. In the presence of spatial autocorrelation, $\rho(d)$ is no longer constant. It varies with the distance between elementary units. Wackernagel (1988) states that data analysis should consider dependency resulted from geographical position.

RESULTS DISCUSSION

Most census data are in aggregate, namely in total or average. Aggregation is a collection of elementary unit in a population based on certain factors such as geography and time. The example of aggregation process based on geography is the collection of data in the level of district, regency, or province. If the characteristic of geographical location is added, what we get are spatial data. The example of aggregation process based on time is a time series data. Aggregation process not only summaries the data which are easy to read, but also maintain the confidentiality of each and every individual, an important aspect in census data.

This paper uses data from census on population and family in Australia in the year 1991. The variable of interest is labor force participation rate in Wollongong Australia (see Figure 3), called Collection District (CD), a collection of about 200 families. This region has 377 CDs. Figure 3 does not present the border of each CD. Labor force participation rate represents the percentage of active labor force plus unemployment over the total population of more than 15 years old (Australian Bureau of Statistics, 2001).

Figure 3: Map of Regional Location, Wollongong, Australia

Semivariogram analysis can be used to explain interdependency among variables from a sampling or census process. Besides economic social characteristic commonly recorded in a survey, researchers also record spatial characteristic of an elementary unit such as geographical location in *longitude* and *latitude coordinates* (Australian Bureau of Statistics, 2001).

Labor force rate distribution can be seen in Figure 4, which shows a distribution similar to normal distribution with the mean and variance of 58,6 and 73,4, respectively. From the spatial analysis based on exponential model, using PROC VARIOGRAM of SAS application, we get the parameter estimations, which are provided in Figure 5. The exponential semivariogram empirical model can be written as follows:

$$\hat{\gamma}(d_{ij}) = 44.5 + (76.0 - 44.5).\left(1 - \exp\frac{-d_{ij}}{9.8}\right);\ d_{ij} \geq 0 \cdot$$

The model shows the estimation of nugget, sill, and range, which are 44.5, 76.0, and 9.8, respectively. Sill provides the variance in labor force participation rate of 44.5. Nugget describes the variation in labor force in zero distance. This can be interpreted as the variation in labor force rate within the CD region. Range gives the distance indication, namely 9.8 km across CDs, showing that labor force participation rates are dependent on each other, and become independent for those of shorter distances. In other words, the values of labor force participation within the CD level are interdependent in the range of 9.8 km.

The spatial dependency analysis using semivariogram model can be compared with spatial autocorrelation coefficient of Moran I, presented in Table 1. Figure 6 shows the harmony between the autocorrelation coefficient of Moran and exponential model. Both Table 1 and Figure 6 suggest that within the range of 1 km, there exist strong interdependency, while outside the range, the interaction tends to weaken. Both statistics show the existence of spatial distribution from the characteristics under investigation.

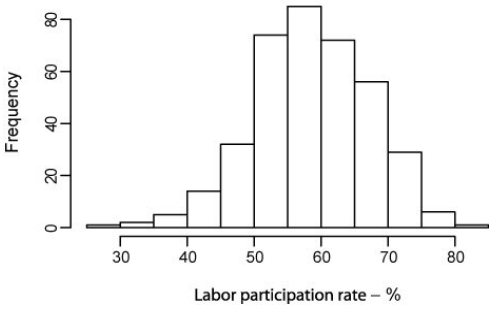


Figure 4: Distribution of Labor Force Par ticipation rates

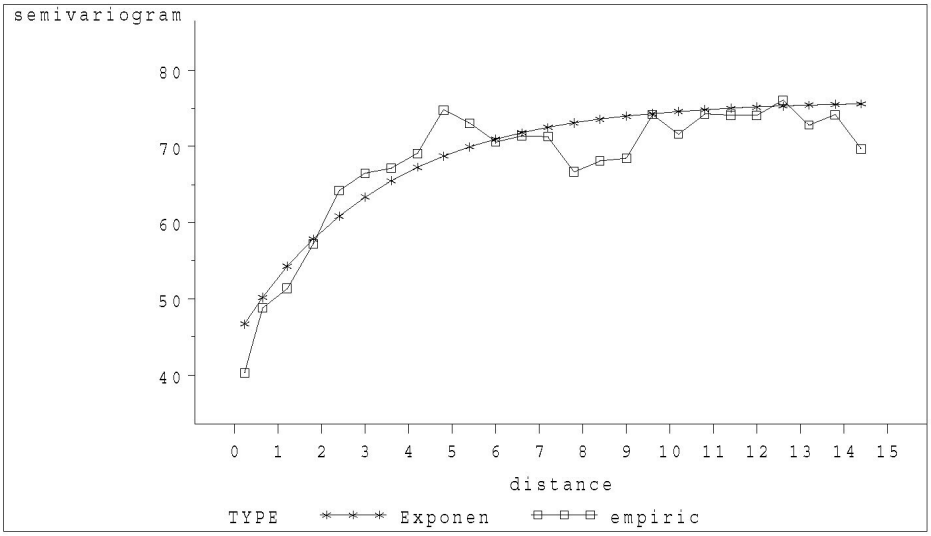


Figure 5: The Estimation of Semivariogram (□) and Its Exponential Model (∗), with Nugget = 44.5, Sill = 76.0, and Range = 9.8

Table 1: Coefficient of Moran and Spatial Autocorrelation Based on Exponential Model

Neighborhood distance (km)	1.0	2.0	3.0	4.0	5.0	7.0	9.0	10.0
Moran coefficient	0.46	0.27	0.20	0.16	0.14	0.11	0.09	0.08
The exponential model	0.31	0.23	0.17	0.12	0.09	0.05	0.03	0.02

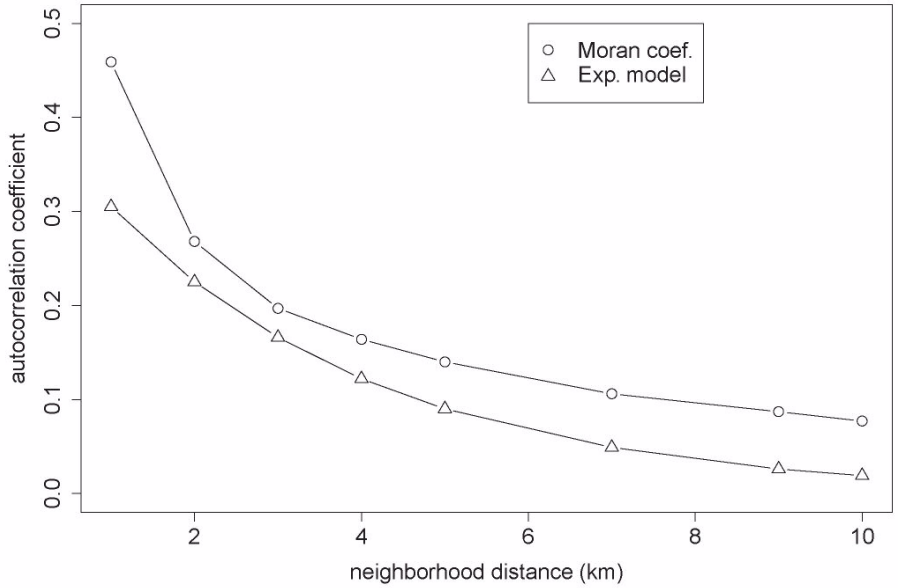


Figure 6: Graphical Comparison between Moran and Spatial Autocorrelation Coefficients Based on Exponential Model

Moran I spatial autocorrelation coefficient and exponential model provide useful information in explaining spatial distribution of labor force participation rates. In general, both measures of spatial autocorrelation tend to decrease as the distance (geographically) from two observation points increases. Exponential models gives more information in spatial dependency through their parameters, namely nugget, sill, and range.

CONCLUSIONS

Census data were collected in the individual level but the publications were in aggregate, namely total and average. The aggregation can be conducted in higher level such as district, regency, or province.

Census data analysis was commonly limited by the assumption that the observations were fixed, so that statistical methods were of little help. The assumption was not relevant since a population was a realization from various populations in the future or in the past. In this sense, a population can be assumed as a collection of observations which vary.

The analysis on census data became relevant if it has been based on the super population approach. In this analysis, the assumption of normally, identically and in-

dependently distribution in the data were weaken, so that conventional statistical methods needed to be carried out carefully.

Another assumption that was no longer relevant in census data analysis was independency. To overcome such problem, semivariogram analysis explained spatial interdependence across units of analysis. It provided important information related to nugget effect and interdependency across unit of analysis in the context of geographical distance.

Spatial autocorrelation coefficient explained spatial distribution in the social characteristics under investigation. The coefficient, represented by exponential model of semivariogram provided results which were in the same spirit with the Moran spatial autocorrelation coefficient. Hence, semivariogram model can be used to explain spatial distribution.

REFERENCES

Ahmad, E. and L. Lerruth (2000), "Indonesia: Implementing National Policies in a Decentralized Context: Special Purpose Programs to Protect the Poor," *IMF Working Paper*, WP/00/102.

Arbia, G. (1993), "The Use of GIS in Spatial Statistical Surveys," *International Statistical Review*, 61(2), 339-359.

Australian Bureau of Statistics (2001), *Census of Population and Housing: Socio-economic Indexes for Area's (SEIF),* Technical Report No. 2039.0.55.001, Australian Bureau of Statistics.

Brunsdon, C. (1995),"Analysis of Univariate Census Data," In Openshaw, S. (Ed.), *Census Users' Handbook,* 213–237, United Kingdom: Geo Information International.

Frees, E.W. (2006), "Forecasting Labor Force Participation Rates," *Journal of Official Statistics*, 22(3), 453-485.

Haslett, J. (1992), "Spatial Data Analysis – Challenges," *The Statistician*, 41, 271–284.

Manley, D., R. Flowerdew and D. Steel (2006), "Scales, Levels and Processes: Studying Spatial Patterns of British Census Variables," *Computers, Environment and Urban Systems*, 30, 143–160.

Openshaw, S. (1984), "Ecological Fallacies and the Analysis of Areal Census Data," *Environment and Planning A*, 6, 17–31.

Pawitan, G. and D.G. Steel (2006), "Exploring a Relationship between Aggregate and Individual Levels Data through Semivariogram Models," *Geographical Analysis*, 38, 310–325.

Pont, M. (2007), "Coverage and Non-response Errors in the Uk New Earnings Survey," *Journal of the Royal Statistical Society: Series A*, 170(3), 713 – 733.

Shen, J. (2006), "Estimating Urbanization Levels in Chinese Provinces in 1982-2000," *International Statistical Review*, 74(1), 89-107.

Skinner, C.J., D. Holt and T.M.F. Smith (1989), *Analysis of Complex Survey,* New York: John Wiley & Sons.

Suryahadi, A., W. Widyanti, R.P. Artha, D. Perwira, D., and S. Sumarto (2005), "*Developing a Poverty Map for Indonesia (A Tool for Better Targeting in Poverty Reduction and Social Protection Programs),"* Unpublished paper, the SMERU Research Institute.

Wackernagel, H. (1988), "Geostatistical Techniques for Interpreting Multivariate Spatial Information," In Chung, C.F., A.G. Fabbri and R. Sinding-Larsen (Eds.), *Quantitative Analysis of Mineral and Energy Resource,* 393– 409, Dordrecht: D. Reidel Publishing Company.