Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

20-09

Exploring the MAUP from a spatial perspective

Gandhi Pawitan, David G Steel

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Exploring the MAUP from a spatial perspective [1]

Gandhi Pawitan,
Faculty of Social and Political Science,
Universitas Katolik Parahyangan, Bandung, Indonesia,
Email : gandhi_p@home.unpar.ac.id,

David G. Steel,
Center for Statistical and Survey Methodology,
School of Mathematics and Applied Statistics,
University of Wollongong, Australia, Email : dsteel@uow.edu.au

Abstract

This paper aims to explore the modifiable areal unit problem (MAUP) from a spatial analysis perspective. Using different scales and zoning on a particular set of spatial data may lead to problems in interpreting the results. Pawitan et Steel (2006) explored the bias associated with aggregate level semivariogram analysis in comparison with the corresponding individual level analysis. Minot et Baulch (2005a) discussed some consequences of using aggregated level data in poverty mapping, which may affect the validity of the output. Tagashira et Okabe (2002) investigated the consequences of the MAUP in developing regression models for spatially aggregated data.

The MAUP will be examined theoretically and empirically based on census data, as well as simulations. Discussion focuses on the expectation of differences between analysis undertaken at two different aggregation levels. We introduce a semivariogram model for aggregated data, and explore relationships using the model at different levels of aggregation.

Keywords : aggregation effect, MAUP, semivariogram model.

1 Introduction

Statistical analysis of social data often uses spatially aggregated data, because such data is readily available and limitations exist on the availability of individual level data. Unit level data is often aggregated into artificial areal units. For example, Australian Census

[1]This research is supported by the Endeavour Research Indonesia Fellowship 2007, DEST, Australia, award ID 0013-2007

data was aggregated according to the collection district, which may contain around 200-300 households.

Some studies have used aggregated data in their analysis and have reported limitations on their results due to issues such as ecological bias and the modifiable areal unit problem (MAUP), for example the study reported by Robinson (1950), Holt, Steel, et Tranmer (1996), Gotway et Young (2002), Young et Gotway (2007).

The MAUP consists of two aspects, referred to as the scale and zoning effect (Openshaw, 1984). The zoning effect refers to how the region is partitioned for a particular number of zones. The scale is determined by how many zones are formed. Holt et al. (1996) showed that the MAUP is caused by the failure to incorporate area or spatial effects into the analysis. They argued that the MAUP can be explained by incorporating the area effects into the model underpinning the analysis.

Recent studies of the MAUP have mainly focused on issues such as the aggregation process, aggregation effect, scale problem, and ecological fallacy, which may be found in Steel, Tranmer, et Holt (2006), Manley, Flowerdew, et Steel (2006), and Pawitan et Steel (2006). Other issues were in regression models and generating a map.

Tagashira et Okabe (2002) developed a regression model using aggregated spatially data. The result showed that the variance of the estimator for the slope coefficient in the aggregated model is larger than that in the disaggregated model. They found the zoning system that has the minimum variance for a fixed number of zones.

Minot et Baulch (2005a) studied a practical aspect of the MAUP when developing a poverty map from aggregated census data. They claimed that the map's precision was reduced as it was created using aggregated census data instead of household-level data

to generate poverty estimates. This situation may impact policy making, as discussed in Minot et Baulch (2005b). They showed that poverty mapping at the district level may reveal that most poor people do not live in the poorest districts, but in the areas where poverty incidence is intermediate. This suggest that estimates of poverty at the district level are not closely correlated with poverty estimates at the household or individual level. A similar applications were found from Minot (2000), Ratcliffe (2005), Wakefield (2007), Baschieri, Falkingham, Hornby, et Hutton (2005), and Hentschel, Lanjouw, Lanjouw, et Poggi (1998).

In this paper, we show how the MAUP is due to the spatial relationships at different geographical levels. The next section will introduce some theoretical background, assumptions and definitions. Some cases of socio-economic data from census and survey data will be presented and discussed to give evidence of the MAUP. Simulations will be used to show how the MAUP can be directly related to spatial relationships as reflected in the semivariogram.

2 Theoretical background

Scale and zoning are special characteristics of spatial data, whose effect cannot be removed but can be controlled in some ways. The M spatial units to which the higher resolution data are aggregated, such as census district, postal code districts, or administrative divisions at various levels, are arbitrarily created by some decision-making processes. These units represent only one of an almost infinite number of ways to partition a region into subregions. Each partitioning will result in different values for the aggregated statistics. Steel et Holt (1996) used the term aggregation effect to cover the effects of

allocating individual units into spatial groups and combining spatial groups at one level into higher level groups.

2.1 Definitions and assumptions

In our case the term *individual* denotes the smallest available level of data. Consider a finite population within a particular region $\mathcal{D}$ of $\mathcal{R}^2$, which contains N individual units. Locations of individuals are denoted by $\mathbf{L} = \{\ell_1, \ell_2, \cdots, \ell_N\} \in \mathcal{D}$. Assume a random process $Y_i|\mathbf{L}$ is defined in region $\mathcal{D}$, where $i \in \mathcal{U} = \{1, \cdots, N\}$. The region is partitioned into M non-overlapping sub-areas, $\mathcal{D}_g$, for $g \in \mathcal{U}_\mathcal{G} = \{1, \cdots, M\}$. Partitioning the region will create groups of individuals, where the gth group contains N_g individuals. Assume that $Y_i|\mathbf{L}$ have a moment structure $E(Y_i|\mathbf{L}) = \mu_i(\mathbf{L})$ and $Cov(Y_i, Y_j|\mathbf{L}) = \Delta_{ij}(\mathbf{L})$. If $i = j$ then $Cov(Y_i, Y_j|\mathbf{L}) = V(Y_i|\mathbf{L}) = \Sigma_i(\mathbf{L})$, which is a variance of the Y_i. A population's mean and variance are defined by

$$\bar{Y} = \frac{1}{N}\sum_{i\in\mathcal{U}} Y_i \quad \text{and} \quad S_{yy} = \sum_{i\in\mathcal{U}} \frac{(Y_i - \bar{Y})^2}{N-1} \tag{1}$$

The population mean has first and second moment, $E(\bar{Y}) = \bar{\mu}$ and $V(\bar{Y}) = \frac{1}{N}(\bar{\Sigma} + (N-1)\bar{\Delta})$, respectively, where $\bar{\mu} = \frac{1}{N}\sum_{i\in\mathcal{U}} \mu_i$, $\bar{\Sigma} = \frac{1}{N}\sum_{i\in\mathcal{U}} \Sigma_i$; and $\bar{\Delta} = \frac{1}{N(N-1)}\sum_{i\neq j\in\mathcal{U}} \Delta_{ij}$. The term $\bar{\Delta}$ can be considered as a summary of the population covariance structure. It may contain important information regarding with interdependency among individual within the population.

Aggregation of individual data leads to a set of M group's means which are available for analysis, denoted by $\bar{Y}_g$ for $g \in \mathcal{U}_\mathcal{G}$ and called aggregated data. It was assumed

that the individuals' locations $\mathbf{L}$ were not available but some spatial characteristics of the groups were, such as area ($\mathcal{A}_g$), perimeter (P_g), and centroid's location (ℓ_g). A centroid is defined as the center of gravity of all boundary points of the region or subregion. Distances between individuals (d_{ij}) or groups' centroids (d_{gh}) are defined by the *Euclidean* distance, $\|\ell_i - \ell_j\|$ and $\|\ell_g - \ell_h\|$ respectively. Given the group level data $\bar{Y}_g$, then $E(\bar{Y}_g) = \bar{\mu}_g$, $V(\bar{Y}_g) = \frac{1}{N_g}\left(\bar{\Sigma}_g + (N_g - 1)\bar{\Delta}_g\right)$, and $Cov(\bar{Y}_g, \bar{Y}_h) = \bar{\Delta}_{gh}$, where $\bar{\Sigma}_g = \sum_{i\in g} \frac{\Sigma_{ii}(\mathbf{L})}{N_g}$; $\bar{\Delta}_g = \sum_{i\neq j\in g} \frac{\Delta_{ij}(\mathbf{L})}{N_g(N_g-1)}$; and $\bar{\Delta}_{gh} = \sum_{\substack{i\in g \\ j\in h}} \frac{\Delta_{ij}(\mathbf{L})}{N_g N_h}$. The weighted group level variance is defined by

$$ {}_N\bar{S}_{yy} = \sum_{g\in\mathcal{U}_\mathcal{G}} \frac{N_g(\bar{Y}_g - \bar{Y})^2}{M-1} \tag{2} $$

2.2 Semivariogram and spatial autocorrelation

Define $\gamma_{ij} = \frac{1}{2}V(Y_i - Y_j)$ as the individual level semivariogram. Intrinsic stationarity of the spatial process, which is $E(Y_i - Y_j) = 0$ and $V(Y_i - Y_j)$ is a function of d_i. It is often assumed and also second order stationarity, which is $E(Y_i) = \mu$, $V(Y_i) = \sigma^2$, and $Cov(Y_i; Y_j) = C(\ell_i - \ell_j)$. These assumptions state that observations have a constant mean and variance over the population and the covariogram, $C(\cdot)$, is a function of relative location of two observations. The semivariogram model is usually developed utilizing the relationship between an empirical semivariogram, $\widehat{\gamma}_{ij} = \frac{1}{2}(Y_i - Y_j)^2$, and the absolute distance between the two observations d_{ij}. The exponential model is one example of the semivariogram model, which is

$$ \gamma(d_{ij}) = \mathsf{n} + (\mathsf{s} - \mathsf{n}) \cdot \left(1 - exp\left[\frac{-3d_{ij}}{\mathsf{r}}\right]\right), \qquad d_{ij} \geq 0 \tag{3} $$

where parameters n, s, and r are the nugget, sill, and range, respectively. For an isotropic process, $\widehat{\gamma}_{ij}$ is an unbiased estimator of $\gamma(d_{ij})$. The relationship between the semivariogram and covariogram is defined by $\gamma(d_{ij}) = \sigma^2 - C(d_{ij})$. The quantity $C(0)$ represents a covariogram value at distance zero, $d_{ij} = 0$. The observations corresponding to $d_{ij} = 0$ in social data may come from one or more different individuals at the same location (e.g within the household). A connection between semivariogram and spatial autocorrelation can be derived as $\gamma(d_{ij}) = \sigma^2(1 - \rho(d_{ij}))$, where $\rho(d_{ij})$ is a spatial autocorrelation function which depends only on the distance of location between observations.

In terms of the empirical semivariogram values $\widehat{\gamma}_{ij}$, the individual and weighted group level variance can be formulated by

$$S_{yy} = \bar{\tilde{\gamma}}; \quad \text{and} \quad {}_N\bar{S}_{yy} = \bar{\tilde{\gamma}} \left(\frac{N-1}{M-1} \right) - \bar{N}\tilde{\tilde{\gamma}}_W \left(\bar{C}_{N\bar{\tilde{\gamma}}} + \frac{M}{M-1} \cdot \frac{\bar{N}-1}{\bar{N}} \right) \tag{4}$$

where $\tilde{\tilde{\gamma}}_W = \sum_{g \in \mathcal{U}_\mathcal{G}} \frac{\bar{\tilde{\gamma}}_g}{M}$, $\bar{C}_{N\bar{\tilde{\gamma}}} = \frac{\bar{S}_{N\bar{\gamma}}}{\bar{N} \cdot \tilde{\tilde{\gamma}}_W}$, and $\bar{S}_{N\bar{\tilde{\gamma}}} = \sum_{g \in \mathcal{U}_\mathcal{G}} \frac{(N_g - \bar{N}) \cdot (\bar{\tilde{\gamma}}_g - \tilde{\tilde{\gamma}}_W)}{M-1}$ (see appendix for a proof). The $\bar{\tilde{\gamma}}$ represents the overall average of $\widehat{\gamma}_{ij}$, $\tilde{\tilde{\gamma}}_W$ is an average of the within group semivariogram, $\bar{C}_{N\bar{\tilde{\gamma}}}$ is a relative covariance between N_g and $\bar{\gamma}_g$, and $\bar{S}_{N\bar{\gamma}}$ represents covariance between N_g and $\bar{\gamma}_g$.

Semivariogram modeling may be attempted using the groups' means and the distances between the groups, which is often taken as the distance between centroids, d_{gh}. In this case a group level semivariogram may be applied and defined by $\Gamma_{gh} = \frac{1}{2} V(\bar{Y}_g - \bar{Y}_h)$ and the associated empirical values is $\widehat{\Gamma}_{gh} = \frac{1}{2}(\bar{Y}_g - \bar{Y}_h)^2$.

2.3 Expectation of the variance

Pawitan et Steel (2006) showed that the expectation of individual level variance and weighted group level variance can be expressed in terms of the individual level semivariogram values, respectively as $E(S_{yy}) = \frac{1}{N(N-1)} \sum_{i \neq j \in \mathcal{U}} \gamma_{ij} = \bar{\gamma}$ and $E({}_N\bar{S}_{yy}) = \frac{N-1}{M-1}\bar{\gamma} - \bar{S}_{N\bar{\gamma}} - \frac{M(\bar{N}-1)}{M-1}\tilde{\gamma}_W$. The $\bar{S}_{N\bar{\gamma}}$ represents the covariance between N_g and $\bar{\gamma}_g$, $\tilde{\gamma}_W$ is an average within group semivariogram, and $\bar{\gamma}_g$ is an average of individual level semivariogram in the group g. They are formulated respectively by, $\bar{S}_{N\bar{\gamma}} = \sum_{g \in \mathcal{U}_\mathcal{G}} \frac{(N_g - \bar{N}) \cdot (\bar{\gamma}_g - \tilde{\gamma}_W)}{M-1}$, $\tilde{\gamma}_W = \sum_{g \in \mathcal{U}_\mathcal{G}} \frac{\bar{\gamma}_g}{M}$, and $\bar{\gamma}_g = \sum_{i \neq j \in g} \frac{\gamma_{ij}}{N_g(N_g-1)}$.

2.4 Aggregation effects

Pawitan et Steel (2006) discussed the effect of aggregation on semivariogram analysis and proposed a method for estimating individual level semivariogram from aggregated data. In this paper we show how the aggregation effect can be directly related to spatial relationships as reflected in the semivariogram. The aggregation effect can be examined in terms of the difference between statistics calculated from a data set at two different scales, for example between the individual level and group level (Steel, Holt, & Tranmer, 1996).

Consider the aggregation effect on variances, which is the difference between the variance calculated from the aggregated data and from individual level, that is ${}_N\bar{S}_{yy} - S_{yy}$. Based on (4) the empirical aggregation effect can be formulated by

$$ {}_N\bar{S}_{yy} - S_{yy} = \left(\frac{N-M}{M-1}\right)\bar{\hat{\gamma}} - \bar{N}\tilde{\hat{\gamma}}_W \left(\bar{C}_{N\bar{\hat{\gamma}}} + \frac{M}{M-1} \cdot \frac{\bar{N}-1}{\bar{N}}\right) \tag{5} $$

Equations (5) express the aggregation effect in terms of the difference between weighted group level variance and individual level variance. The result shows how the aggregation effect can be related to spatial relationship using semivariogram values. The key factors in the model are $\bar{\hat{\gamma}}$ and $\tilde{\hat{\gamma}}_W$. The value of $\bar{\hat{\gamma}}$ is free from any zoning or scaling effect, but $\tilde{\hat{\gamma}}_W$ depends on the zoning or scale. The factor $\bar{N}$ is determined by the scale and is usually known, and is the same for all zoning at a given scale. Consider a case where $\bar{C}_{N\bar{\hat{\gamma}}}$ is small and M large, then we have ${}_N\bar{S}_{yy} - S_{yy} \approx (\bar{N}-1)(\bar{\hat{\gamma}} - \tilde{\hat{\gamma}}_W)$, and so the scale effect is determined by the scale as reflected in $\bar{N}$. The impact of different zoning at the same scale depends on how the different zoning affects $\tilde{\hat{\gamma}}_W$, which is an average of the within group semivariogram values. The expectation of (5) can be derived as

$$E({}_N\bar{S}_{yy} - S_{yy}) = \left(\frac{N-M}{M-1}\right)\bar{\gamma} - \bar{S}_{N\bar{\gamma}} - \left(\frac{M(\bar{N}-1)}{M-1}\tilde{\gamma}_W\right) \tag{6}$$

We can apply equation (5) to explore the $\tilde{\hat{\gamma}}_W$, and how it changes with a different scale and zoning. The advantage of this approach is that we only have to consider the distribution of within group distances. Consider a case with the exponential semivariogram model (3) and d represent a distance between individuals within the group. We have a semivariogram model for the gth group that is

$$\bar{\gamma}_g = \mathsf{s} - (\mathsf{s}-\mathsf{n})exp\left[\frac{-3\bar{d}_g}{\mathsf{r}}\right] \tag{7}$$

where $\bar{d}_g$ is an average distances among all individuals in gth group. Using the Taylor series expansion, we have that $exp\left[\frac{-3\bar{d}_g}{\mathsf{r}}\right] \approx 1 - \frac{3\bar{d}_g}{\mathsf{r}}$. The approximation is applicable

in situations where $\frac{3\bar{d}_g}{\mathsf{r}}$ is small, that is the case when average size of groups are smaller than r. Applying this to (7), we have an approximation for the $\bar{\gamma}_g$, that is $\bar{\gamma}_g \approx \mathsf{s} - (\mathsf{s} - \mathsf{n}) \cdot \left(1 - \frac{3\bar{d}_g}{\mathsf{r}}\right)$.

Matérn (1986) discussed the approximation for the $\bar{d}_g$ when the group's shape is a circle, which is $E(d) \approx k_1\sqrt{\mathcal{A}}$, where $\mathcal{A}$ is the area of the group and the constant $k_1 = 0.511$. Applying this approximation into group g, and substituting into (7), we get $\bar{\gamma}_g \approx \mathsf{n} + (\mathsf{s} - \mathsf{n}) \cdot 3 \cdot \frac{k_1\sqrt{\mathcal{A}_g}}{\mathsf{r}}$, where $\mathcal{A}_g$ is the area of the gth group (see Pawitan & Steel, 2006 for detail). Hence the $\tilde{\gamma}_W$ can be derived and approximated into

$$\tilde{\gamma}_W \approx \mathsf{n} + \frac{\mathsf{s} - \mathsf{n}}{\mathsf{r}} \cdot 3 \cdot k_1 \tilde{\mathcal{A}}^*_W, \quad \text{where} \quad \tilde{\mathcal{A}}^*_W = \frac{1}{M} \sum_{g \in \mathcal{U}_\mathcal{G}} \sqrt{\mathcal{A}_g} \tag{8}$$

This suggests that $\tilde{\mathcal{A}}^*_W$ is a relevant summary of the characteristics of scale and zoning in the aggregation effect. Its value may change as the scale or zoning change. The term $\mathcal{A}_g$ is interchangeable with other factor, such as $\mathcal{A}_g = \frac{N_g}{D_g}$, where D_g is a density of the gth group.

In the semivariogram model, the sill (s) represents the variance of the data (S_{yy}), hence we assume that $S_{yy} = \mathsf{s}$, then from equation (5), the weighted group level variance can be approximated by

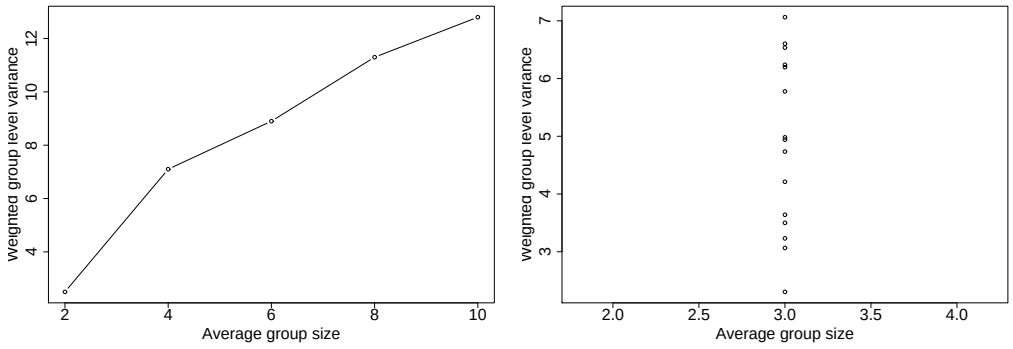
$${}_N\bar{S}_{yy} \approx \frac{M}{M-1}\left(\mathsf{s} + (\bar{N} - 1)(\mathsf{s} - \mathsf{n})\left[1 - 3k_1\frac{\tilde{\mathcal{A}}^*_W}{\mathsf{r}}\right]\right) \tag{9}$$

Equation (9) provides a connection between the non-spatial statistic ${}_N\bar{S}_{yy}$ and the spatial parameters of the population as described by n, s, and r. The important aspect of

the zoning also is shown by the $\tilde{A}^*_W$, which is the group area factor. Equation (9) can be thought of as an another semivariogram model from a spatial structure with the $\tilde{A}^*_W$ factor as an analog of the $\bar{d}$. The impact of the scale is clearly seen through the appearance of the factor $(\bar{N}-1)$. Although this derivation was done for the exponential semivariogram model case, the idea can be applied for other semivariogram models easily.

2.5 The scale effect

Given some particular grouping, the original groups can be formed into larger M_k groups. This can be done several times, each resulting in an average group size $\bar{N}_1, \cdots, \bar{N}_K$; for $k = 1, \cdots, K$, where $\bar{N}_k = N/M_k$. Each realization also gives a different ${}_N\bar{S}_{yy}$ and $\tilde{A}^*_W$, say ${}_N\bar{S}_{yy\,1}, \cdots, {}_N\bar{S}_{yy\,K}$; and ${}_1\tilde{A}^*_W, \cdots, {}_K\tilde{A}^*_W$. This can be drawn in figure (1 a). Based



(a) Scaling effect (b) Zoning effect

Figure 1: The weighted group level variance at different scale

on aggregation effect model (9) we have

$$ {}_N\bar{S}_{yy\,k} = \frac{M_k}{M_k - 1}\left(\mathsf{s} + (\bar{N}_k - 1)(\mathsf{s} - \mathsf{n})\left[1 - 3k_1\frac{{}_k\tilde{A}^*_W}{\mathsf{r}}\right]\right) \qquad (10) $$

Given the values of ${}_N\bar{S}_{yy_k}$, $\bar{N}_k$, k_1, and ${}_k\tilde{\mathcal{A}}^*_W$ then equation (10) has three unknown parameters n, s, and r. In relation to the spatial autocorrelation, define $\rho(0) = \left(1 - \frac{\mathsf{n}}{\mathsf{s}}\right)$, then we may have ${}_N\bar{S}_{yy_k} = \frac{M_k}{M_k - 1} \cdot \mathsf{s}\left(1 + (\bar{N}_k - 1)\rho(0)\left[1 - 3k_1 \frac{{}_k\tilde{\mathcal{A}}^*_W}{\mathsf{r}}\right]\right)$, where the $\rho(0)$ can be interpreted as the intra-household spatial correlation. If the observations are independently and identically distributed (IID) then the $\mathsf{n} = \mathsf{s}$ and $\rho(0) = 0$. This case implies ${}_N\bar{S}_{yy_k}$ is proportional to the s by a factor $\frac{M_k}{M_k - 1}$. If the observations are not IID, then variation in ${}_N\bar{S}_{yy_k}$ depends on the magnitude of the $\rho(0)$ and also the value of r.

2.6 The zoning effect

The zoning effect can be examined by varying the arrangement of the groups at a particular $\bar{N}$. Suppose that we have $t = \{1, \cdots, T\}$ realization of the zoning at a given scale. This situation implies a variation in ${}_N\bar{S}_{yy}$ and $\tilde{\mathcal{A}}^*_W$, that is ${}_N\bar{S}_{yy_1}, \cdots, {}_N\bar{S}_{yy_T}$; and ${}_1\tilde{\mathcal{A}}^*_W, \cdots, {}_T\tilde{\mathcal{A}}^*_W$. This zoning realization can be drawn as in figure (1 b). Based on model (9), we have

$$ {}_N\bar{S}_{yy_t} = \frac{M}{M - 1}\left(\mathsf{s} + (\bar{N} - 1)(\mathsf{s} - \mathsf{n})\left[1 - 3k_1 \frac{{}_t\tilde{\mathcal{A}}^*_W}{\mathsf{r}}\right]\right) \tag{11} $$

Equation (11) which shows that the value of ${}_N\bar{S}_{yy}$ at the tth realization of the zoning is dependent on the $\tilde{\mathcal{A}}^*_W$. The $\tilde{\mathcal{A}}^*_W$ reflects the area of the group, which may change on every realization of the zoning. Given T realization of the zoning scheme, then the ${}_N\bar{S}_{yy}$ and $\tilde{\mathcal{A}}^*_W$ for every realization can be computed.

3 The MAUP from socio-economic census data

The data comes from the 1991 Australian Census of Population and Housing for the Adelaide region. The region is divided into non-overlapping collection districts. There are 1,713 collection district and 767,030 people over 15 years old were counted by the census. There are three groupings that the data can be readily aggregated to (of which CD is the lowest level). The groupings are SSC, DPC, and LGA (ABS & MapInfo, 1993). The SSC refers to a collection district derived suburb. It is composed of one or more collection districts that lie wholly within a suburb. If the CD is split across two or more suburban boundaries, then the CD is allocated to the most appropriate suburb. The DPC derives from Australian Post Postcode boundaries. It may contain one or more collection districts. The LGA is the legal local government area, and it may contain one or more collection districts.

Three characteristics have been considered, these are employment rate, unemployment rate, and labor force participation rate. Some statistics have been tabulated in table (1), and boxplot is used to give some descriptions of characteristic group means for different groupings (Fig. 2).

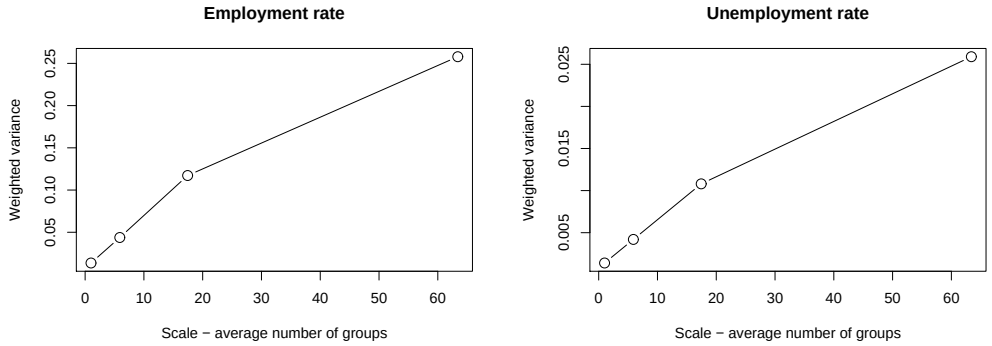
Table (1) shows that mean (unweighted) of the characteristics are not affected much by the grouping, but variances ($_N\bar{S}_{yy}$) are affected by the grouping. It shows that the weighted group level variances increases with the scale. If the data was IID or no spatial relationship, then it would be equal to the individual level variance (Steel & Holt, 1996).

Table 1: Some statistics of the variables at different grouping level, with CD level is the lowest level.

Characteristics	Grouping level			
	CD	SSC	DPC	LGA
Number of zones	1713	313	102	27
$\bar{N}$	1.00	5.90	17.44	63.44
Population density (person/km^2)	1967.8	1591.3	1311	1376.8
Average area (km^2)	0.39	2.23	6.61	24.84
Employment rate				
mean	0.5303	0.5240	0.5387	0.5183
$_N\bar{S}_{yy}$	0.0136	0.0438	0.1172	0.2579
Unemployment rate				
mean	0.0731	0.0719	0.0735	0.0729
$_N\bar{S}_{yy}$	0.0014	0.0042	0.0108	0.0259
Labor participation rate				
mean	0.6034	0.5959	0.6123	0.5912
$_N\bar{S}_{yy}$	0.0115	0.0335	0.0874	0.1939

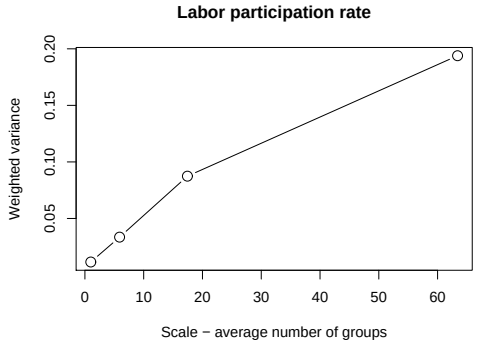
Figure (2) shows increasing values for the weighted group level variance ($_N\bar{S}_{yy}$) when the $\bar{N}$ gets larger. This gives an indication of the scaling issue. The zoning issue is more difficult to show using the census data, since the census only has one realization of the zoning scheme.

At a particular level of grouping, the MAUP can be investigated by looking at the presence of the spatial autocorrelation (Arbia, 1989). Consider the Adelaide data available at the CD level, and a neighborhood distance between centroids of 1.0 km, 2.0 km, and 5.0 km. The neighborhood distance represents all the CDs within a particular distance, i.e. 1, 2, and 5 km. The connectivity matrix can be generated by S+Spatialstats



(a) Employment rate (%)

(b) Unemployment rate (%)



(c) Labor participation rate (%)

Figure 2: Plot of the ${}_N\bar{S}_{yy}$ at different grouping levels, from CD level to LGA level.

software, `find.neighbor()` and `spatial.neigbor()`, then the Moran coefficient can be calculated by `spatial.cor()`. Different neighborhood distances can be considered to be different groupings of the CD. For example, a neighborhood distance of 1.0 km will aggregate the CDs which have distances of less than or equal to 1.0 km between themselves. Table (2) shows that the Moran autocorrelations coefficient decrease as neighborhood distance increases.

Table 2: Moran autocorrelation coefficient of Adelaide CD data at different neighborhood distances

	Distance (km)		
Characteristics	1.0	2.0	5.0
Employment	0.5241	0.3690	0.2132
Unemployment	0.5074	0.3513	0.1816
Labor	0.4711	0.3374	0.1987

4 Simulation

Equation (5) shows how aggregation effects can be decomposed into several factors related to spatial relationships within the population . The impact of these factors are explored through a simulation study. The simulation considered the scaling and zoning effect. The simulation was done using the following steps :

step 0 Specify the dimensions of the rectangular region, number of groups required (M), and number of repetitions of the simulation. The dimension of the rectangular region for example can be 10x10, 15x10, etc, denote this as G. The region is partitioned into a regular grid resulting in equal size rectangles.

step 1 Number the rectangles from 1 to the number of rectangles counted, i.e. G.
step 2 Randomly choose $G - M$ rectangle, which will be merged with other rectangles to create the required number of groups.
step 3 Assign each rectangle chosen for merging to one of the two different merging processes, (a) merge to the left or right of the adjacent grid, and (b) merge to the top or bottom of the adjacent grid. Assignment of the merging process was done randomly between the alternatives.
step 4 List all the members of the defined groups, and calculate the groups level data.
step 5 Compute relevant group level statistics.
step 6 Repeat step 1 to 5 for the required number of repetitions.

Generating inter dependency among observations within the population can be achieved by considering a semivariogram model, i.e. exponential model in (3). Most models for the isotropic semivariogram contain three parameters, the nugget (n), sill (s), and range (r), which are in the interval $[0, \infty)$. The covariance structure of population can be computed by considering a relationship between semivariogram and covariogram, $C(d_{ij}) = \sigma^2 - \gamma(d_{ij})$ (see Pawitan et Steel, 2004 and Arbia, 1989).

4.1 Individual and aggregated level data

The simulation created a population contained individual level observations, which were generated according to an exponential semivariogram model with parameters, $\mathsf{n} = 5$, $\mathsf{s} = 20$, and $\mathsf{r} = 15$. There are 10,000 individual points in the population within a region of 60 by 80 length unit or 4,800 square area units. Each individual is located at one particular point. Figure (3) shows a summary of the population. Twenty different

scales were considered, with M varying from 50 to 2000 as shown in Table (3). Sixty repetitions (zoning) were applied to each scale in the simulation.

```
1 ******************************************
2 * Simulation program to generate individual *
3 * level data with a specified semivariogram *
4 * model, eg. exponential model :            *
5 *       n + (s-n)*(1-exp(-3*d/r))            *
6 *                                            *
7 * written by Gandhi Pawitan                  *
8 *           School of Math. and App. Stat.   *
9 ******************************************
10    Semivariogram model : n + s Exp(r)
11   n,s,r            :    5.000  20.000  15.000
12   Number of points :  10000
13   Region shape     : rectangle
14   X-min,X-max      :   20.000  80.000
15   Y-min,Y-max      :   10.000  90.000
16 ---------------------------------------------
17 Performance of the generated points :
18  min max mean var of x :  20.000  79.992  49.855 298.042
19  min max mean var of y :  10.012  89.990  49.867 539.949
20  mean var of original z :   0.004   1.003
21  min max mean var of z : -17.113  16.818   0.758  18.397
```

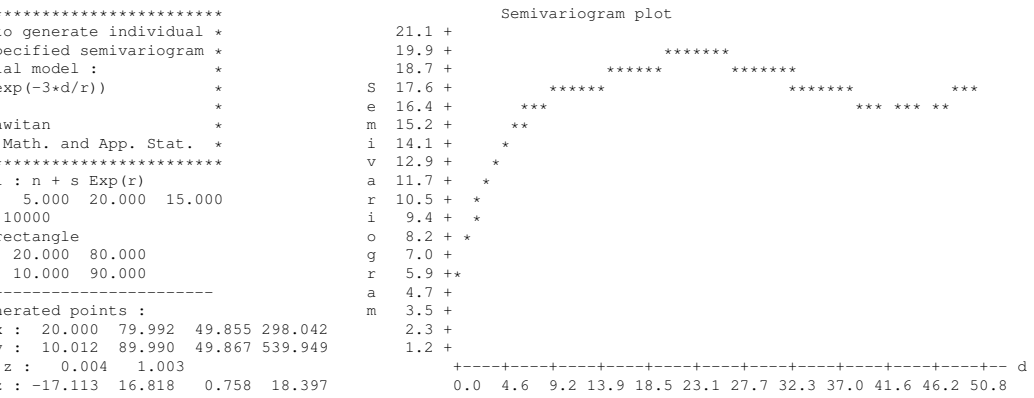


Figure 3: Individual level data and its semivariogram model, with $\mathsf{n} = 5$, $\mathsf{s} = 20$, and $\mathsf{r} = 15$

Table 3: Twenty different number of groups (scales)

	1	2	3	4	5	6	7	8	9	10
grid	50×50	50×50	40×50	40×40	40×40	40×35	40×35	30×35	30×30	30×30
M	2000	1800	1600	1400	1200	1000	900	800	700	600
$\bar{N}$	5.0	5.56	6.25	7.14	8.33	10.0	11.11	12.50	14.29	16.67
	11	12	13	14	15	16	17	18	19	20
grid	25×30	25×25	25×25	25×20	20×20	20×20	20×20	20×15	20×15	20×15
M	500	450	400	350	300	250	200	150	100	50
$\bar{N}$	20.0	22.22	25.0	28.57	33.33	40.0	50.0	66.67	100.0	200.0

4.2 Exploring scaling and zoning effect

The average group size, $\bar{N}$, plays an important role in exploring the MAUP as shown in (5), and $\bar{N}$ is also the easiest factor to be measured. We will also show that the impact

of other factors can be explored mainly based on their relationship with $\bar{N}$. Figure (4) shows a relationship between scale ($\bar{N}$) on the ${}_N\bar{S}_{yy}$. The figure shows the value of ${}_N\bar{S}_{yy}$ increases non-linearly with an increasing scale ($\bar{N}$). The linear fitting comes out to $R^2 = 96.5\%$, and $R^2 = 99.6\%$ for the quadratic. The fitting models suggest a very close relationship between $\bar{N}$ and ${}_N\bar{S}_{yy}$.

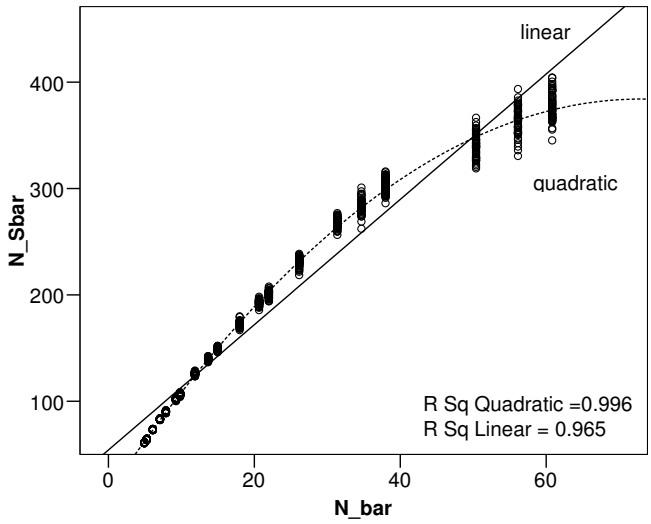
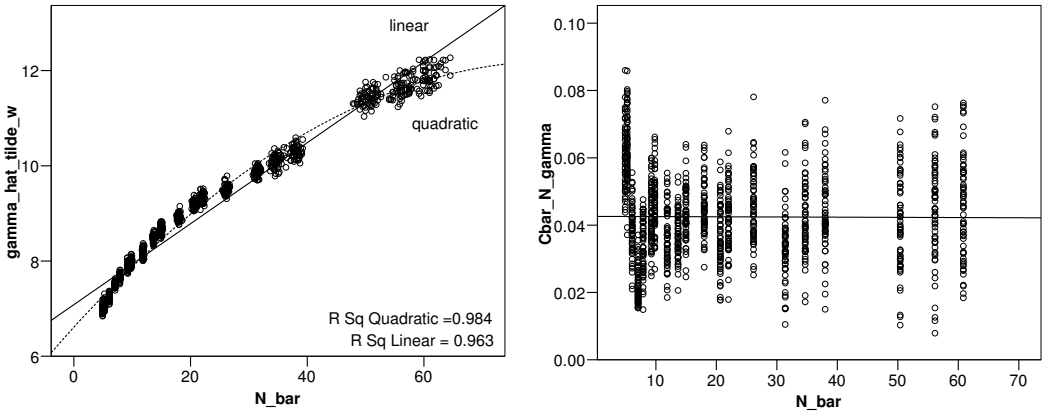


Figure 4: Scatter plot of ${}_N\bar{S}_{yy}$ versus $\bar{N}$.

Equation (5) shows that variation in aggregation effects can be related to variation in the terms $\bar{\hat{\gamma}}$, $\bar{N}$, $\tilde{\hat{\gamma}}_W$, and $\bar{C}_{N\hat{\gamma}}$. The term $\bar{\hat{\gamma}}$ is not affected by scale or zoning. But $\bar{N}$, $\tilde{\hat{\gamma}}_W$, and $\bar{C}_{N\hat{\gamma}}$ are affected by scale and zoning. These three terms will be affected by scale, but the zoning will affect only the $\tilde{\hat{\gamma}}_W$, and $\bar{C}_{N\hat{\gamma}}$. The $\bar{N}$ is usually known, but the others may not be. Since $\bar{N}$ is usually known, we may look at relationship between $\bar{N}$ versus

$\tilde{\hat{\gamma}}_W$ or $\bar{C}_{N\bar{\gamma}}$. We argue that $\tilde{\hat{\gamma}}_W$ is more important in explaining aggregation effects than $\bar{C}_{N\bar{\gamma}}$. This will be investigated through a graphical approach.



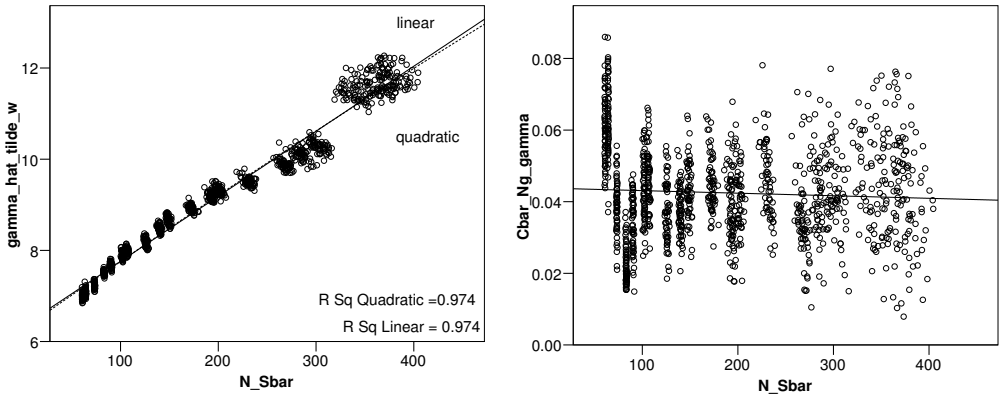
(a) (b)

Figure 5: Scatter plot of (a) $\bar{N}$ versus $\tilde{\hat{\gamma}}_W$, (b) $\bar{N}$ versus $\bar{C}_{N\bar{\gamma}}$

The term $\tilde{\hat{\gamma}}_W$ represents the average of the average within group semivariogram values $\bar{\hat{\gamma}}_g$, thus it will be a measure of average within group variance. Figure (5-a) shows a close relationship between $\tilde{\hat{\gamma}}_W$ and $\bar{N}$. The $\tilde{\hat{\gamma}}_W$ is increasing as $\bar{N}$ gets larger. The linear fitting comes with $R^2 = 96.3\%$ and $R^2 = 98.4\%$ for quadratic. Meanwhile, figure (5-b) shows a pattern for $\bar{C}_{N\bar{\gamma}}$. It shows a stationary pattern for $\bar{C}_{N\bar{\gamma}}$, with values ranging from (< 0.02) to (0.08) and the mean at around the value 0.04. This figure suggests that the $\bar{C}_{N\bar{\gamma}}$ has a little impact on ${}_N\bar{S}_{yy}$.

Figure (6) shows a relationship between ${}_N\bar{S}_{yy}$ versus $\tilde{\hat{\gamma}}_W$. The first plot indicates that variation for $\tilde{\hat{\gamma}}_W$ increases as ${}_N\bar{S}_{yy}$ increases. The linear and quadratic fittings are very

close with the value of $R^2 = 97.4\%$. The second figure shows a relationship between $\bar{C}_{N\bar{\hat{\gamma}}}$ and ${}_N\bar{S}_{yy}$. The figure shows a stationary pattern of the $\bar{C}_{N\bar{\hat{\gamma}}}$, where it goes from 0.02 up to 0.08. These figures suggest that ${}_N\bar{S}_{yy}$ is more closely related to $\tilde{\hat{\gamma}}_W$ than to $\bar{C}_{N\bar{\hat{\gamma}}}$.

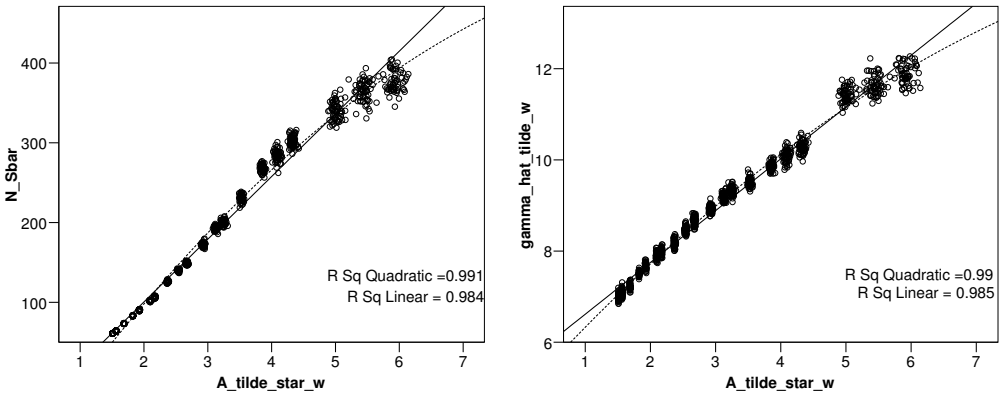


(a) (b)

Figure 6: Scatter plot of (a)$\tilde{\hat{\gamma}}_W$ versus ${}_N\bar{S}_{yy}$, (b) $\bar{C}_{N\bar{\hat{\gamma}}}$ versus ${}_N\bar{S}_{yy}$

Figure (**??**) shows a relationship between the variation of aggregation effect against $\bar{N}$ and $\tilde{\mathcal{A}}^*_W$. The trend shows that variations are very small for a small $\bar{N}$ and get larger as $\bar{N}$ increase. This could be caused by a decreasing number of groups followed by $\bar{N}$ increases, since $M = \frac{N}{\bar{N}}$. This pattern is also the same for the $\tilde{\mathcal{A}}^*_W$. This pattern indicates that a non-linear relationship exists between a scale and variation of the aggregation effect. This gives important information regarding the impact of the aggregation effect at a particular $\bar{N}$ or $\tilde{\mathcal{A}}^*_W$.

Meanwhile, zoning effect may be observed in terms of variation of statistic at the



(a) ${}_N\bar{S}_{yy}$ versus $\tilde{\mathcal{A}}^*_W$ (b) $\tilde{\hat{\gamma}}_W$ versus $\tilde{\mathcal{A}}^*_W$

Figure 7: Relationship between scale ${}_N\bar{S}_{yy}$ and $\tilde{\hat{\gamma}}_W$ versus $(\tilde{\mathcal{A}}^*_W)$

same scale. For example the variation of ${}_N\bar{S}_{yy}$ at the same scale can be illustrated in a boxplot, such as the one in figure (8). The boxplot in every scale $(\bar{N})$ indicates a distribution of ${}_N\bar{S}_{yy}$. The figure shows a boxplot is getting larger when the scale $(\bar{N})$ increases, which indicate increasing variation for ${}_N\bar{S}_{yy}$ when the $\bar{N}$ gets larger.

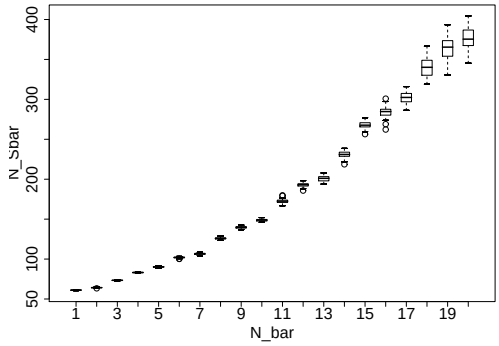
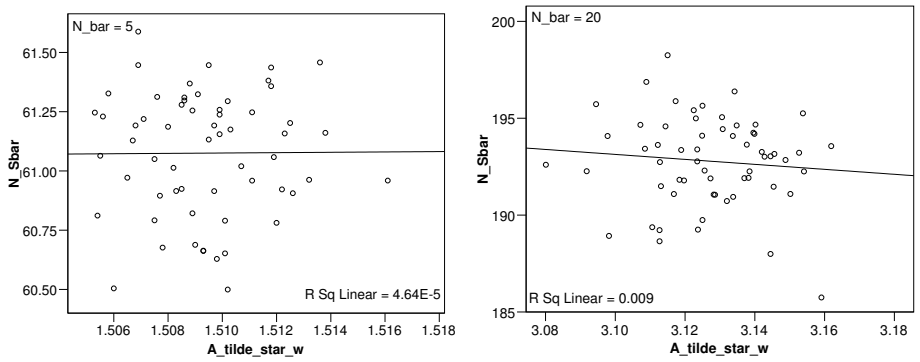


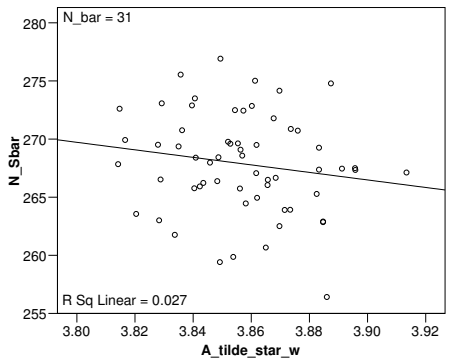
Figure 8: Boxplot of ${}_N\bar{S}_{yy}$ at difference $\bar{N}$

Figure (9) shows a relationship between $\tilde{\mathcal{A}}^*$ versus ${}_N\bar{S}_{yy}$ at a particular $\bar{N}$ ($\bar{N} = 5$,

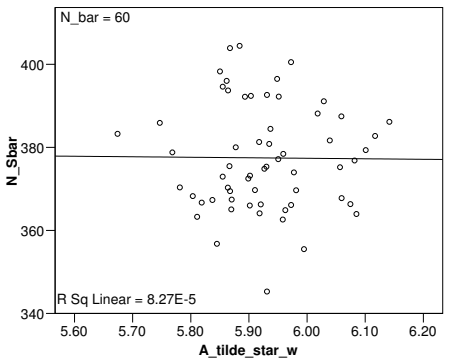


(a) $\bar{N} = 5$

(b) $\bar{N} = 20$



(c) $\bar{N} = 31$



(d) $\bar{N} = 60$

Figure 9: Relationship between ${}_N\bar{S}_{yy}$ versus $\tilde{\mathcal{A}}^*_W$ at different level of $\bar{N}$

$\bar{N} = 20$, $\bar{N} = 31$, and $\bar{N} = 60$). The chosen $\bar{N}$ represent low, medium, and large scale. Random pattern were shown at every particular scale, which indicates no relationship between $\tilde{\mathcal{A}}^*_W$ and ${}_N\bar{S}_{yy}$. These figures suggest that the zoning effect is more difficult to assess. Examining the scales of the figures, we see the relative zoning effect becomes larger as $\bar{N}$ increases.

5 Discussion

The evidence of the MAUP from the Australian Census data mainly shows the scale effect, since it only concerned one zoning scheme. The scale effect was shown by weighted variance of the characteristics (employment, unemployment, and labor participation rate) and was also indicated by spatial autocorrelation (Moran coefficient).

The simulation gave a great opportunity to look at the scale and zoning effect. The simulated data shows a similar pattern with the scale effect as shown by the Australian Census data. The simulated scale is represented by $\bar{N}$ and $\tilde{\mathcal{A}}^*_W$. It was found that in addition to the non-linear relationship between scale and ${}_N\bar{S}_{yy}$, there was also a non-linear relationship between scale and $\tilde{\bar{\gamma}}_W$. Meanwhile, the zoning effect was not apparent as shown by a relationship between $\tilde{\mathcal{A}}^*_W$ and ${}_N\bar{S}_{yy}$, instead there was a random pattern of ${}_N\bar{S}_{yy}$ along the $\tilde{\mathcal{A}}^*_W$ values.

Expression of the aggregation effect, as defined in (5), contains the factor $\bar{C}_{N\bar{\gamma}}$, which

is the coefficient of co-variation between N_g and $\bar{\gamma}_g$. If the N_g or $\bar{\gamma}_g$ is constant, then $\bar{C}_{N\bar{\gamma}}$ is zero. In general, we expect that the factor $\bar{C}_{N\bar{\gamma}}$ to be very small. Therefore the main factor in addition to $\bar{N}$ affecting scale and zoning is $\tilde{\gamma}_W$.

6 Summary

The MAUP can be used as a tool in semivariogram analysis, and can be used to estimate the individual level of semivariogram parameters from the group level data. Different zoning and scaling realizations are needed to estimate the individual level of semivariogram parameters. The estimated parameters can be used to adjust the estimated group level of semivariogram parameters, which is useful for further analysis, such interpolation procedure by kriging method.

Future work may involve exploring equation (9) further. The equation can also be considered as a starting point in using the MAUP to investigate the individual level of spatial parameters. This equation can be sketched as displayed in figure (10).

Using the $\bar{N}$ as the x-axis, the figure shows that variation at a particular $\bar{N}$ is due to the zoning effect and variation among $\bar{N}$ indicates the scale effect. Using (9) we could develop a procedure to estimate n, s, and r. This could then be applied to data, such as that presented in figure (10). The independent variable in (9) is $\tilde{A}^*_W$, and as a result different zoning and scales that create appreciable variation in the variable should be

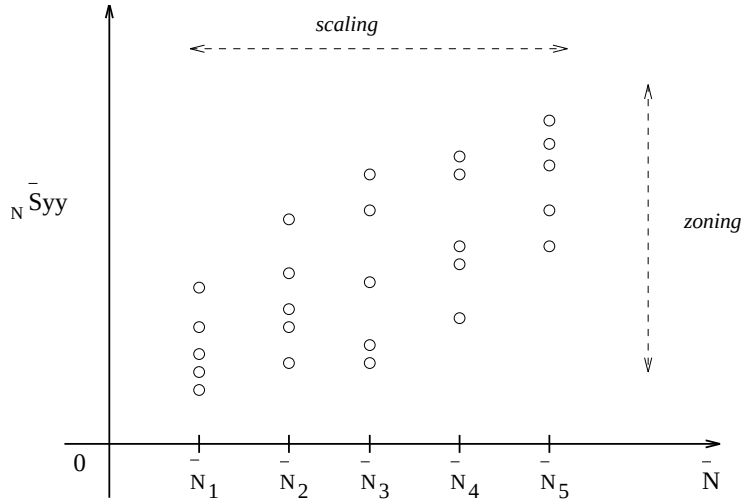


Figure 10: The scale and zoning effect

used. However, (9) was obtained using the assumption that $3\bar{d}_g/\mathsf{r}$ was small, and so the groupings used should be consistent with this assumption. Alternatively higher order terms in the Taylor series expansion of exp() could be used.

Appendix : Proof of equation (4)

We have if that $S_{yy} = \frac{1}{N(N-1)} \sum_{i,j} \frac{1}{2} \left(Y_i - Y_j\right)^2$, and define that $\widehat{\gamma}_{ij} = \frac{1}{2} \left(Y_i - Y_j\right)^2$. Hence we will get that $S_{yy} = \bar{\hat{\gamma}}$

Proof. $S_{yy} = \frac{1}{N(N-1)} \sum_{i,j \in \mathcal{U}} \frac{1}{2}(Y_i - Y_j)^2 = \frac{1}{N(N-1)} \sum_{i,j \in \mathcal{U}} \widehat{\gamma}_{ij} = \bar{\hat{\gamma}}$ □

We have that ${}_N\bar{S}_{yy} = \frac{N-1}{M-1}\bar{\hat{\gamma}} - \bar{N}\tilde{\hat{\gamma}}_W \left(\bar{C}_{N\bar{\hat{\gamma}}} + \frac{M}{M-1} \cdot \frac{\bar{N}-1}{\bar{N}}\right)$.

Proof. Define $\widehat{\Gamma}_{gh} = \frac{1}{2} \left(\bar{Y}_g - \bar{Y}_h\right)^2$. The weighted group level variance can be expressed

in term of $\widehat{\Gamma}_{gh}$, that is

$$ {}_N\bar{S}_{yy} = \frac{1}{N(M-1)} \sum_{g,h \in \mathcal{U}_\mathcal{G}} N_g N_h \widehat{\Gamma}_{gh} \tag{12} $$

Relationship between the group level semivariogram and the individual level semivariogram can empirically be formulated as

$$ \widehat{\Gamma}_{gh} = \bar{\hat{\gamma}}_{gh} - \frac{N_g - 1}{2N_g} \bar{\hat{\gamma}}_g - \frac{N_h - 1}{2N_h} \bar{\hat{\gamma}}_h \tag{13} $$

where $\bar{\hat{\gamma}}_{gh} = \frac{1}{N_g N_h} \sum_{\substack{i \in g \\ j \in h}} \widehat{\gamma}_{ij}$ and $\bar{\hat{\gamma}}_g = \frac{1}{N_g(N_g-1)} \sum_{i \neq j \in g} \widehat{\gamma}_{ij}$. The mean square error within the group is defined $S_{yy}^{<W>} = \frac{1}{N-M} \sum_{g \in \mathcal{U}_\mathcal{G}} \sum_{i \in g} (Y_i - \bar{Y}_g)^2$, and can be expressed in to

$$ S_{yy}^{<W>} = \tilde{\hat{\gamma}}_W \left(1 + \bar{C}_{N\bar{\hat{\gamma}}} \frac{\bar{N}(M-1)}{M(\bar{N}-1)} \right) \tag{14} $$

where $\tilde{\hat{\gamma}}_W = \frac{\sum_{g \in \mathcal{U}_\mathcal{G}} \bar{\hat{\gamma}}_g}{M}$, $\bar{C}_{N\bar{\hat{\gamma}}} = \frac{\bar{S}_{N\bar{\hat{\gamma}}}}{\bar{N} \cdot \tilde{\hat{\gamma}}_W}$, and $\bar{S}_{N\bar{\hat{\gamma}}} = \frac{1}{M-1} \sum_{g \in \mathcal{U}_\mathcal{G}} (N_g - \bar{N})(\bar{\hat{\gamma}}_g - \tilde{\hat{\gamma}}_W)$. Substituting (13) into equation (12) gives

$$ {}_N\bar{S}_{yy} = \frac{1}{N(M-1)} \sum_{g,h \in \mathcal{U}_\mathcal{G}} N_g N_h \left(\bar{\hat{\gamma}}_{gh} - \frac{N_g - 1}{2N_g} \bar{\hat{\gamma}}_g - \frac{N_h - 1}{2N_h} \bar{\hat{\gamma}}_h \right) $$

Modifying this equation becomes

$$
\begin{aligned}
{}_N\bar{S}_{yy} &= \frac{1}{N(M-1)}\left(\sum_{g,h\in\mathcal{U}_\mathcal{G}}\sum_{\substack{i\in g\\ j\in h}}\gamma_{ij} - \sum_{g,h\in\mathcal{U}_\mathcal{G}}\frac{N_h}{2}(N_g-1)\bar{\hat{\gamma}}_g - \sum_{g,h\in\mathcal{U}_\mathcal{G}}\frac{N_g}{2}(N_h-1)\bar{\hat{\gamma}}_h\right)\\
&= \frac{1}{N(M-1)}\left(N(N-1)\bar{\hat{\gamma}} - \sum_{h\in\mathcal{U}_\mathcal{G}}\frac{N_h}{2}\sum_{g\in\mathcal{U}_\mathcal{G}}(N_g-1)\bar{\hat{\gamma}}_g - \sum_{g\in\mathcal{U}_\mathcal{G}}\frac{N_g}{2}\sum_{h\in\mathcal{U}_\mathcal{G}}(N_h-1)\bar{\hat{\gamma}}_h\right)\\
&= \frac{1}{N(M-1)}\left(N(N-1)\bar{\hat{\gamma}} - \sum_{h\in\mathcal{U}_\mathcal{G}}\frac{N_h}{2}\left[(M-1)\bar{S}_{N\bar{\hat{\gamma}}} + M\bar{N}\tilde{\hat{\gamma}}_W - M\tilde{\hat{\gamma}}_W\right]\right.\\
&\quad \left. - \sum_{g\in\mathcal{U}_\mathcal{G}}\frac{N_g}{2}\left[(M-1)\bar{S}_{N\bar{\hat{\gamma}}} + M\bar{N}\tilde{\hat{\gamma}}_W - M\tilde{\hat{\gamma}}_W\right]\right)\\
&= \frac{N-1}{M-1}\bar{\hat{\gamma}} - \bar{S}_{N\bar{\hat{\gamma}}} - \tilde{\hat{\gamma}}_W\frac{M(\bar{N}-1)}{M-1}
\end{aligned}
$$

We have $\bar{S}_{N\bar{\hat{\gamma}}} = \bar{C}_{N\bar{\hat{\gamma}}} \cdot (\bar{N} \cdot \tilde{\hat{\gamma}}_W)$ to complete the proof. □

Références

ABS, & MapInfo. (1993). *CDATA91 with MapInfo User's Manual*. Australia : MapInfo Australia.

Arbia, G. (1989). *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*. Dordrecht : Kluwer.

Baschieri, A., Falkingham, J., Hornby, D., & Hutton, C. (2005). *Creating a poverty map for Azerbaijan* (Working Paper N° 3793). World Bank Policy Research.

Gotway, C. A., & Young, L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association*, *97*(458), 632–648.

Hentschel, J., Lanjouw, J., Lanjouw, P., & Poggi, J. (1998). *Combining census and survey data to study spatial dimensions of poverty* (Policy Research Working Paper N° 1928). The World Bank Development Research Group.

Holt, D., Steel, D. G., & Tranmer, M. (1996). Area Homogeneity and the Modifiable Areal Unit Problem. *Geographical System*, *2*, 83–101.

Manley, D., Flowerdew, R., & Steel, D. (2006). Scales, levels and processes: Studying

spatial patterns of British census variables. *Computers, Environment and Urban Systems*, *30*, 143–160.

Matérn, B. (1986). *Spatial Variation*. Berlin Heidelberg : Springer-Verlag.

Minot, N. (2000). Generating disaggregated poverty maps: An application to vietnam. *World Development*, *28*(2), 319 – 331.

Minot, N., & Baulch, B. (2005a). Poverty mapping with aggregate census data: What is the loss in precision? *Review of Development Economics, 9(1), 524, 2005*, *9*(1), 5–24.

Minot, N., & Baulch, B. (2005b). Spatial patterns of poverty in vietnam and their implications for policy. *Food Policy*, *30*, 461 – 475.

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

Pawitan, G., & Steel, D. (2004). Generating inter-correlated observations under a specified spatial model. *Integral Journal of Mathematics and Natural Science,Department of Mathematics, Parahyangan Catholic University, Indonesia*, *9*(2), 58 – 65.

Pawitan, G., & Steel, D. G. (2006). Exploring a relationship between aggregate and individual levels data through semivariogram models. *Geographical Analysis*, *38*, 310–325.

Ratcliffe, J. H. (2005). Detecting spatial movement of intra-region crime patterns over time. *Journal of Quantitative Criminology*, *21*(1), 103 – 123.

Robinson, W. S. (1950). Ecological Correlations and the Behaviour of Individuals. *American Sociological Review*, *15*, 351–357.

Steel, D. G., & Holt, D. (1996). Rules for Random Aggregation. *Environment and Planning A*, *28*, 957–978.

Steel, D. G., Holt, D., & Tranmer, M. (1996). Making Unit-Level Inferences From Aggregated Data. *Survey Methodology*, *22*(1), 2–15.

Steel, D. G., Tranmer, M., & Holt, D. (2006). Unravelling ecological analysis. *Journal of Applied Mathematics and Decision Sciences*, *2006*, Article ID 38358, 18 pages. (doi:10.1155/JAMDS/2006/38358)

Tagashira, N., & Okabe, A. (2002). The modifiable areal unit problem in a regression model whose independent variabel is a distance from a predetermined point. *Geographical Analysis*, *34*(1), 1–20.

Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics*, *8*(2), 158 – 183.

Young, L. J., & Gotway, C. A. (2007). Linking spatial data from different sources: the effects of change of support. *Stoch Environ Res Risk Assess*, *21*, 589 – 600.