NIASRA
NATIONAL INSTITUTE FOR APPLIED
STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

The University of Wollongong

Working Paper

15-13

Empirical Zoning Distributions for Small Area Health Data

Dr Sandy Burden and Professor David Steel

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email:
anica@uow.edu.au

Empirical Zoning Distributions for Small Area Health Data

Dr Sandy Burden [*1] and Professor David Steel[2]

[1] Research Fellow, National Institute of Applied Statistics Research Australia, School of Mathematics and Applied Statistics, University of Wollongong, Australia.

[2] Director, National Institute of Applied Statistics Research Australia, School of Mathematics and Applied Statistics, University of Wollongong, Australia

Abstract

Many health studies use aggregate data such as the means or totals for areal units or zones when individual level data are not available. An ecological analysis using these data typically produces estimates that differ from those obtained using the corresponding individual level analysis. This is due to the modifiable area unit problem (MAUP) whereby the results of the analysis depend on the scale and zoning effects. In this paper empirical zoning distributions are used to study the effect of zoning on the parameter estimates from ecological analyses. The zoning distribution is defined as the distribution of the parameter estimates obtained from a given ecological analysis which is repeated for different sets of M zones.

For simulated population data, generated using the 2007-2008 Australian National Health Survey and 2006 Australian Census, we create zoning distributions for estimates from ecological regression models at

[*] Email: sburden@uow.edu.au

multiple scales of analysis. These distributions are typically relatively symmetrical and unimodal and appear to be normally distributed. They often have appreciable variation, which should not be ignored. Using the distribution, the “ecological average” or mean of the empirical estimates at each scale of analysis, displays systematic variation with the number of zones. The variance of the zoning distribution is related to the average zone population per zone.

Using empirical zoning distributions, the parameter estimates obtained for a given set of zones at the same or a different scale can be compared and the influence of zones on the results of an ecological analysis.

Keywords: modifiable areal unit problem, ecological bias, scale, zoning, regression analysis, multilevel modelling

1 Introduction

A set of zones is formed when a given study region is partitioned into M geographically contiguous, non-overlapping areal units. Individuals in the population are assigned to the zones using indicators of geographic location, such that all individuals belong to exactly one zone and each zone has $N_g \geq 1$ observations, $g = 1, ..., M$. Health studies ideally use individual level data in a multilevel modelling framework to incorporate the clustered nature of the data in the analysis. The model has a hierarchical structure with individual observations nested within the cluster or group to which they belong, which for geographical data are the zones. For privacy reasons, or to allow data from different sources to be brought together, frequently only aggregate geographic summaries are available and an ecological analysis is used which substitutes area level summaries for the individual level data. However, ecological parameter estimates may be biased and the amount and direction of bias depends on the zones used to analyse the data.

This sensitivity is called the modifiable area unit problem (MAUP) and it occurs because aggregation removes the direct link between individual response and covariate values. It has two aspects: the scale effect, which occurs when the number of areas in the study region is changed and the zoning effect which arises when the areas at a given scale are defined using different boundaries (Flowerdew et al., 2001). Changing either factor may alter the estimates which are obtained. Changes in scale substantially affect the variance of the estimates in a somewhat systematic way (Steel and Holt, 1996), but the effect of moving the zone boundaries is not apparently systematic (Openshaw, 1984; Stafford et al., 2008; Haynes et al., 2008). Although highly homogenous areas will result in an analysis with more power and potentially a smaller bias due to the MAUP (Briant et al., 2010) as in this case more of the variability between data values is between areas.

While the MAUP remains unresolved, the results of area level or ecological analyses can only legitimately be applied to the particular zones used in the study, as estimates may change for alternative sets of zones, even at the same scale. When the particular set of zones is of direct substantive interest, the possibility of a different result for an alternative set of zones is irrelevant. In other cases the effect of the zones on the parameter estimates and variation in the results for an alternative set of zones can assist interpreting the results of the analysis. Its importance has been recognised in many different types of analysis including studies of health (Diez-Roux and Mair, 2010; Parenteau and Sawada, 2011; Swift et al., 2008; Cockings and Martin, 2005; Schuurman et al., 2007; Best et al., 2001). However, the geographical scale of a study is still frequently determined by data availability (Wakefield, 2004), as is the choice of zones.

The zones used to analyse data can either be derived from existing zoning systems or created for the purpose of the analysis. Existing systems typically utilise official or administrative boundaries which provide a convenient way to disseminate data, such as post-codes, output areas, enumeration districts and local authorities. However, it has long been recognised that the use of existing zoning systems has limitations (Openshaw, 1977). For example the UK Census Enumeration Districts display “wide variations in population size, geographical shape, area and social composition” (Cockings and Martin, 2005, pp. 2732–2733). Alternatives to administrative boundaries include zones formed using geometric shapes (such as a rectangular grid), Voronoi tessellations (Swift et al., 2008), local knowledge of the area, or automatic zone design procedures. A recent review of zone design techniques is provided by Duque et al. (2007). Stand alone zone design algorithms which have been used for small area health data include ZDES (Openshaw and Rao, 1995) and AZTool (Cockings et al., 2011; Martin, 2003) which are based on the AZP algorithm (Openshaw, 1977; Openshaw and Rao, 1995). Other zone design packages, including the scale-space clustering method (Mu and Wang, 2008), are available for use with Geographic Information System (GIS) packages.

Several aspects of the zoning effect have previously been studied, including the appropriate zones to use in an analysis (Haynes et al., 2007); the definition of neighbourhoods; and the inclusion of contextual or neighbourhood effects, particularly for the analysis of individual level data (Diez-Roux and Mair, 2010). The effect of scale and the ecological bias associated with using aggregate data to estimate individual level relationships have been widely considered in the fields of spatial epidemiology, geography and the social sciences (see Greenland, 2002; Steel et al., 2003; Richardson et al., 1987; Wakefield, 2004, for example). However, these analyses do not consider the distribution of estimates obtained using multiple sets of zones at each of several given scales, which is the focus of this paper.

In this paper we define the zoning distribution of a parameter estimate as the distribution or density function of the estimate over all possible sets of M zones, given the scale and any constraints used in constructing the zones (such as minimum population thresholds). It can be used to obtain the expected value and variability of estimates at a given scale. The ecological average, defined as the expected value of the zoning distribution, along with the variance of the zoning distribution provide a way to compare and standardise the results for a set of zones at a given scale. Zoning distributions may also be used to make inferences about a parameter for one set of zones or at one scale, given the data for another set of zones at a different scale.

There are presently no established rules or guidelines which can be used to consider the form of the zoning distribution, other than in the case of random aggregation, when the expected value of each estimate is unbiased for the appropriate individual level parameters and their variation can be derived from standard statistical theory (see Steel and Holt, 1996, for example). Since their comprehensive demonstration by Openshaw (1984), zoning distributions have not been widely considered in the literature, with the notable exception of Cockings and Martin (2005) who create 10 sets of zones at several scales to determine the sensitivity of a correlation coefficient to changes in scale and zoning.

In this paper empirical estimates of zoning distributions are created for the parameter estimates from a statistical model for small area health data so that they can be better understood. They are used throughout this paper to describe the variation in estimates for different sets of zones at multiple scales; to identify their impact on statistical analyses; to identify any systematic changes with scale; and to identify appropriate assumptions for zoning distributions so they may be incorporated in future statistical analyses. In section 2, the methodology used to create empirical zoning distributions for aggregate health data is described. The results of the analysis are presented in section 3 and discussed in section 4. In addition to the usual results on ecological bias and scale effects, we focus on the characteristics of the zoning distribution such as its dispersion, its shape, its variance and how it changes with scale and how important it is to account for the zoning distribution in interpreting the results of an ecological analysis.

2 Methodology

Evaluation of zoning distributions requires spatially detailed data which is not available for the Australian population. Instead, these data were simulated for 6,378,163 individuals in the 11879 populated Census Collection Districts (CDs) in New South Wales (NSW), Australia. Unit record data from the 2007-2008 National Health Survey (see Australian Bureau of Statistics, 2008, 2009, for a description of the data) were combined with summary (benchmark) data defining the characteristics of CDs from the 2006 Australian Census (Australian Bureau of Statistics, 2006a) using a spatial microsimulation model (MSM) as described in Burden and Steel (2013). For this approach, for each area an initial set of individual records was randomly selected from the National Health Survey. These records were then swapped with stochastically selected alternatives using a simulated annealing procedure to optimise the fit of the data to the pre-selected benchmarks for the area from the 2006 Australian Census. The resulting simulated population reflects the characteristics of the survey data and also the geographical structure at the CD level.

Health outcomes considered in the study were: measured body mass index (BMI); type 2 diabetes mellitus (diabetes); and angina. The relationship between each outcome and three binary indicators: a sedentary lifestyle (little or no physical activity); dietary fat (consumption of whole milk with $\geq$3% fat);

and current smoking status (for BMI) or obesity ($BMI \geq 25$ - for angina and diabetes) was investigated. Table 1 shows summary statistics for the distribution of each variable across the CD's. The variables age and sex were also included in the model for BMI and used to calculate the expected cases of angina and diabetes in each area, which were included as an offset in their respective models.

[Table 1 about here.]

As health studies often includes a measure of socio-economic status, an area level indicator of socioeconomic status (denoted HSEIA) was created, following the procedure used to define the Australian Bureau of Statistics Socioeconomic Index for Areas, Index of Relative Socio-Economic Advantage and Disadvantage (SEIFA) (Australian Bureau of Statistics, 2006b, p.17–23). HSEIA was created using principle component analysis applied to the CD-level correlation matrix obtained from the simulated data. The index was standardised to have a mean of 1000 and standard deviation of 100. Despite being created using different datasets and some different variables (due to data availability limitations), the distribution of deciles assigned to each area for HSEIA and SEIFA was similar. Overall, HSEIA included 16 variables (see Appendix 5 for details), had an eigenvalue of 7.09 and explained 44% of the variation in the variables used in the index. For comparison, SEIFA included 21 variables, had an eigenvalue of 9.16 and also explained 44% of the variation in the variables.

The final simulated dataset comprised a set of individual health records (from the health survey) for the population of NSW in private dwellings with location information known to the CD level. The simulated data were then rezoned to higher levels of aggregation using the AZTool Software (Cockings et al., 2011; Martin, 2003). AZTool randomly allocates CDs to analysis zones whilst preserving geographic contiguity. It then iteratively swaps CDs between adjacent zones to improve predefined targets and ensure each zone lies within specified upper and lower population limits. At eight scales of analysis, 1000 sets of zones were defined, each using a single run of AZTool with 15 sets of swap iterations. Table 2 shows the population constraints used for each scale and the resulting average population statistics. At each scale, the average population per zone was achieved and the range in population per zone was always narrower than the specified limits. The coefficient of variation (standard deviation divided by the mean) decreased with scale indicating less variability in zone population with increasing scale.

[Table 2 about here.]

Aggregate data summaries were produced for each set of zones and the summaries were analysed to obtain regression parameter estimates, $\hat{\boldsymbol{\beta}}^E$, for the covariates in X.

For a given set of M zones, the average BMI in each zone ($\bar{Y}_g; g = 1, ..., M$) was modelled in terms of the average for age, sex, each indicator variable and HSEIA using the zone level regression model defined in Equation 1.

$$\begin{aligned} \bar{Y}_g | \bar{\boldsymbol{x}}_g &\sim N(\mu_g, \sigma^2) \\ E[\bar{Y}_g | \bar{\boldsymbol{x}}_g] = \mu_g &= \bar{\boldsymbol{x}}_g^T \boldsymbol{\beta}^{\boldsymbol{E}} \end{aligned} \tag{1}$$

where $\bar{x}_g = \frac{1}{N_g} \sum_{i=1}^{N_g} X_i$ is the mean covariate value for individuals $i = 1, ..., N_g$ in area g.

Angina and diabetes were modelled as count variables. Making a rare disease assumption, the ecological model defined in Equation 2 was specified using a Poisson distributed response and log link function. A normally distributed random effect ν_g was modelled in terms of the relative risk of disease for area g,

θ_g. This approach was used because for rare diseases and large group sizes, the count of positive responses for each area can be approximated by an independent Poisson random variable given the covariates and a normally distributed random effect for the zones. An offset, E_g, was included in the model to account for differences in the population at risk in each area according to its age and sex structure. It was calculated as the weighted sum of the expected counts of disease in each age×sex stratum in the population. The expected counts in each stratum were calculated across all areas using the individual level data for the population.

$$\begin{aligned} Y_g|\nu_g^E, \bar{\boldsymbol{x}}_g &\sim Pois(\mu_g) \\ E[Y_g|\nu_g^E, \bar{\boldsymbol{x}}_g] &= \mu_g = E_g\theta_g \\ \theta_g &= exp(\bar{\boldsymbol{x}}_g^T \boldsymbol{\beta}^{\boldsymbol{E}} + \nu_g^E) \\ \nu_g^E &\sim N(0, \sigma_\nu^2) \end{aligned} \tag{2}$$

Model parameters were estimated using either ordinary least squares (for BMI) or second order penalised quasi-likelihood (PQL2) in the MlWiN software (Rasbash et al., 2009). Markov Chain Monte Carlo (MCMC) techniques are also frequently utilised to analyse these models and can provide better estimates. However, these techniques are also time-consuming. Due to the large number of simulations and as the aim of this project was development and demonstration of zoning distributions comparing changes in estimates with scale and zoning, MCMC techniques were not utilised. Using the resulting parameter estimates for each covariate, zoning distributions were defined at each scale using kernel density estimation with a Gaussian kernel in the R Statistical Software (R Development Core Team, 2008).

For the rth set of zones, $r = 1, ..., 1000$, at scale $L_k, k = 1, ..., 8$, the estimate of the regression parameter is denoted $\hat{\boldsymbol{\beta}}^E_{r,k}$. The associated standard variance estimate from the analysis is denoted $\widehat{Var}(\hat{\boldsymbol{\beta}}^E_{r,k})$, from which the estimated standard error can be obtained as $\widehat{SE}(\hat{\boldsymbol{\beta}}^E_{r,k}) = \sqrt{\widehat{Var}(\hat{\boldsymbol{\beta}}^E_{r,k})}$. We can calculated the average of each of these quantities over the $R = 1000$ sets of zones generated at scale L_k. For the zoning distribution this gives: $E_k[\hat{\boldsymbol{\beta}}^E] = \frac{1}{R}\sum_{r=1}^{R} \hat{\boldsymbol{\beta}}^E_{r,k}$, the average of the ecological regression estimates, $E_k[\widehat{Var}(\hat{\boldsymbol{\beta}}^E)] = \frac{1}{R}\sum_{r=1}^{R} \widehat{Var}(\hat{\boldsymbol{\beta}}^E_{r,k})$, the average of the estimated variance of the regression estimates, and $E_k[\widehat{SE}(\hat{\boldsymbol{\beta}}^E)] = \sqrt{E_k[\widehat{Var}(\hat{\boldsymbol{\beta}}^E)]}$, the average estimated standard error of the regression estimates. The estimated variance is not the same as the empirical variance of the zoning distribution, which is given by $Var_k(\hat{\boldsymbol{\beta}}^E) = \frac{1}{R}\sum_{r=1}^{R} \left(\hat{\boldsymbol{\beta}}^E_{r,k} - E_k[\hat{\boldsymbol{\beta}}^E]\right)^2$.

The individual level data were also analysed using appropriate models to provide a reference for comparison with the ecological estimates. A linear multilevel statistical model was used to obtain parameter estimates for the regression coefficients, $\hat{\boldsymbol{\beta}}^I_{r,k}$ for $r = 1, ..., 1000$ at scale L_k, and the associated variance estimate $\widehat{Var}(\hat{\boldsymbol{\beta}}^I_{r,k})$, for each covariate on BMI (including age and sex). The statistical model is defined in Equation 3 for a given set of M zones where ν^I_g is a normally distributed random effect with mean zero and variance σ^2_ν.

$$\begin{aligned} Y_i|\nu_g, \boldsymbol{x}_i &\sim N(\mu_i, \sigma^2) \\ E[Y_i|\nu_g, \boldsymbol{x}_i] = \mu_i &= \boldsymbol{x}_i^T \boldsymbol{\beta}^{\boldsymbol{I}} + \nu^I_g \\ \nu^I_g &\sim N(0, \sigma^2_{\nu^I}) \end{aligned} \tag{3}$$

Using equivalent notation, binary indicators of prevalence of angina and diabetes were modelled using Equation 4 as Bernoulli random variables with a logistic link and random effect ν_g^I.

$$\begin{aligned} Y_i|\nu_g, \boldsymbol{x}_i &\sim Bern(\mu_i) \\ E[Y_i|\nu_g, \boldsymbol{x}_i] = \mu_i &= \frac{exp(\boldsymbol{x}_i^T \boldsymbol{\beta^I} + \nu_g^I)}{1 + exp(\boldsymbol{x}_i^T \boldsymbol{\beta^I} + \nu_g^I)} \\ \nu_g^I &\sim N(0, \sigma^2_{\nu^I}) \end{aligned} \tag{4}$$

The parameters in these models were estimated using each set of zones to define the group level and a sample of data of size approximately equivalent to that obtained from a large population survey. A 0.33% or 1% simple random sample of individual records was selected with an equal probability of selection for BMI and the binary variables respectively. Some logistic models at levels one to three failed to converge using PQL2, due to the large number of small areas, so first order marginal quasi-likelihood (MQL1) was used for estimation.

3 Results

3.1 Empirical Zoning Distributions

Zoning distribution density plots for the ecological regression parameter estimates ($\hat{\beta}^E_{r,k}$) are shown for BMI, angina and diabetes in Figures 1, 2 and 3 respectively. The domain of each distribution represents the range of parameter estimates which may be obtained for the given covariate and scale. The density curve reflects the probability distribution for the zoning distribution of the estimate for the given statistical model.

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

The zoning distributions were generally unimodal, reasonably symmetric and were a similar shape for all response variables. With an increase in scale, that is with smaller M, the ecological average of each parameter estimate generally increased in absolute magnitude in a consistent direction, although the relative size of the change diminished with scale. Due to the large size of the dataset, parameter estimates were statistically significant at the 5% level for most sets of zones at all scales, based on the regression estimate and the estimated variance for a particular set of zones.

The variance of the distributions increased substantially with scale and the proportional increase in variance with scale was similar for all covariates due to the reduced power of the analysis resulting from fewer data values. This demonstrates the implications for the inferences that can be made using ecological parameter estimates. When the limits of the distribution extend to include zero, inferences may not be significant and in some cases the apparent relationship may change sign, i.e. the covariate obesity in the model for diabetes.

The Shapiro Wilk test for normality was performed (using the R statistical software) for each zoning distribution. For BMI and diabetes, the null hypothesis of normality was retained at the 5% level for approximately 90% and 96% of the tests performed. For angina, the results were mixed, with the assumption of normality retained for only 73% of the tests at the 5% level. This may be due to reduced power with angina having fewer positive responses in the data. For all response variables, the skewness of the zoning distributions for all parameter estimates was well below one (with typical values of 0.01 to 0.03) and the excess kurtosis was also low, confirming that the distributions are all relatively symmetric and are consistent with normally distributed data.

A key advantage of the microsimulation approach is that estimates obtained from ecological analysis and multilevel models can be compared using realistic aggregate and individual level data from the same population. Table 3 provides a summary the ecological average, the average estimated standard error and the empirical variance of the zoning distributions for levels 1, 4, and 8 of the ecological analyses. Average multilevel model parameter estimates, calculated for 1000 sets of zones, and the equivalent CD level estimates are also included. The multilevel model estimates are the average estimates using groups at level 4, although very similar results were obtained for groups at all scales. For angina and diabetes the estimates are obtained using a Bernoulli distributed response variable (not Poisson). The estimated variance of the random effects is generally small and decreases with scale in all cases.

[Table 3 about here.]

For an individual level target of inference, the ecological parameter estimates are consistently biased and have generally higher variances than multilevel models. With increasing scale, the change in bias of the average variance of the difference appears to behave systematically. However, predicting the magnitude of the bias for a particular set of zones is much harder. If the zoning distribution were known, the bias of a particular estimate may be standardised by accounting for the variability due to zoning, although the problem still remains that the relationship between the individual level and area level estimates is not immediately predictable.

The scale effect is demonstrated by the change in the ecological average M changes. In some cases this resulted in parameters changing sign. For example in Figure 3, the distribution of the estimated parameter for the HSEIA covariate takes positive and negative values at different scales. The results suggest that a component of the bias is related to the scale of the analysis, particularly as similar results were obtained for the minimum, median and maximum of the zoning distributions (not shown). The relationship was not universal though, as the pattern of covariate values with scale for some covariates was more complex i.e. obesity in the diabetes model.

The potential magnitude of the zoning effect is quantified by variance parameters. These results suggest that the zoning distributions of the variances vary systematically as a function of scale i.e. as a function of $\bar{N}$ and therefore M, reflecting that the variance is a function of the degrees of freedom, $M - 1$. The trend in the variance of the zoning distribution with scale seems consistent with that expected when data are randomly aggregated even though the average parameter estimates do not exhibit the same characteristics.

The implications of the zoning distribution can also be considered in terms of a predictive interval for the parameter estimates obtained for a new set of zones. Assuming approximate normality of the zoning distribution, the width of a 95% prediction interval for a new parameter estimate calculated from a different set of zones is equal to $2 \times 1.96\sqrt{Var(\hat{\beta})}$. Table 4 shows these intervals as a proportion of the average value of the estimate at the relevant scale. The predictive intervals get much wider as the scale increases, in a similar fashion to the variance of the zoning distribution. This suggests that the impact of the zones increases for higher scales, and hence the importance of taking the zoning distribution into consideration also increases.

[Table 4 about here.]

For some parameter estimates, 95% of the estimates from a new set of zones would lie within 10% of the average value of the parameter estimate, i.e. the covariates for angina at the lower scales. In some studies this may represent a reasonable level of variation due to zoning, in which case the zoning distribution will not substantially affect the parameter estimates or inference. However in many cases, even when the estimates are statistically significant, the prediction interval for 95% of estimates from a new set of zones may be greater than 20%

and could lie between 30% and 100%. For these cases the variation due to zoning may have a substantial impact on inference.

3.2 Relationships in the Data

For a linear model, when data are randomly aggregated, the parameter estimates for a fully specified model are theoretically unbiased and the variance is inversely proportional to the degrees of freedom, $M-1$ (Steel and Holt, 1996). As the size of the total population is fixed, the variance is also proportional to $\bar{N}$. Figure 4 plots the ecological average of the parameter estimates versus M for each scale of analysis. It shows that there are clear trends in the variation of the average ecological parameter estimates with scale and that the relationship is not linear. In general the ecological average decreases in magnitude with increasing scale, although the coefficient for dietary fat in the case of diabetes appears inversely related to M. The key finding is that for statistically significant parameter estimates there is a systematic relationship between the ecological average and number of areas used in the analysis.

[Figure 4 about here.]

The relationship between the empirical variance of the zoning distribution for each regression parameter estimate, $(\widehat{Var}_k(\hat{\beta}^E))$, and $\bar{N}$ is shown in Figure 5. For the response BMI, the expected value of the estimated variance of the parameter estimates is proportional to the average population size in each area, $\bar{N}$ (and $1/M$, which is not shown). Despite non-linearity in the models, similar results are obtained for angina and diabetes. These results suggest that scale effects for the regression parameter estimates are related to M with variances linearly related to $\bar{N}$. It is interesting to see that the same result appears to hold for the parameter estimates for a non-linear model. An important finding from these results is that even though the number of areas at the highest scale

(L_8 M=317) is not small, the variance of the zoning distribution is substantial. Moreover, even at smaller scales where M is large, there is appreciable variation in the zoning distribution.

[Figure 5 about here.]

The average estimated variance of the parameter estimates, $(E_k[\widehat{Var}(\hat{\beta}^E)])$, over 1000 zones (not shown), also exhibits the same linear relationship with $\bar{N}$ (and $1/M$) with increasing scale for all of the response variables. $E_k[Var(\hat{\beta}^E)]$ and $Var_k(\hat{\beta}^E)$ are compared directly in Figure 5 for each of the response variables. In all cases the relationship between the variance estimates is reasonably linear, but not exactly the same. The average estimated variance is close to four times greater than the zoning variance.

[Figure 6 about here.]

4 Discussion

These results provide valuable insights into zoning distributions, showing that they exist and can have appreciable dispersion even when there are a large number of zones. For a continuous or a binary response variable, the zoning distributions of the ecological parameter estimates appear normally distributed. The variance of each zoning distribution is a predictable function of $\bar{N}$ and the average of the parameter estimates over the zoning distribution are related to M. The variance of the zoning distribution is often appreciable, although less than the standard variance estimates obtained from an ecological analysis, and it should not be ignored when interpreting the results of statistical analysis based on aggregate data for geographic zones. For example, zoning distributions at two different scales may overlap and in some cases the variation due to zoning may exceed difference due to scale. Quantifying the zoning distribution allows confidence intervals for the expectation of the estimates over the zoning distribution to be obtained.

The results demonstrate several implications of using data aggregated to a given set of zones for obtaining parameter estimates. In general, the aggregate data estimator is biased compared with the individual level estimator and the magnitude and direction of the bias depends on the estimator. The bias of a given estimate cannot yet be predicted, but for statistically significant estimates the average value of the estimator over the zones varies reasonably systematically with scale for both linear and non-linear ecological models. An extension of this result is that zoning distributions may also be used in the definition and analysis of neighbourhood effects. For example, zoning distributions (particularly at multiple scales) can be used to identify the scale above which zones can be assumed to be randomly formed.

One result which is only available when multiple sets of zones are used to analyse the data at each level is that parameter estimates may only be statistically significant for some sets of zones at a given scale. When this occurs, the zones chosen to analyse the data can affect the statistical significance of the parameter estimates in ways that at present are not predictable. In many cases the primary factor affecting the stability of the estimates is the scale of the analysis. In all cases, a greater number of observations (i.e. zones) improves the variance of the estimates and increases the probability of a statistically significant result. However, even when there are over 1000 zones in the analysis the zoning can have an impact on the parameter estimates.

In some cases, the expectation of the zoning distribution at a given scale may itself be a reasonable target of inference. If the zoning distribution can be characterised, we might then be able to draw conclusions about it. A finding in this study is that in most cases, the bias caused by aggregation is more substantial than the variation due to the zones used in the analysis. However there are some cases - particularly at higher scales - when the variation due to zoning is substantial and cannot be ignored. Similarly, compared with the bias associated with the use of aggregate rather than individual level data, the zoning effect was relatively minor at low scales although its impact increased substantially as the scale of aggregation increased and the zoning distributions became much flatter and wider.

Given a set of zones at a particular scale it is worthwhile evaluating the zoning distribution to assess the sensitivity of the results of the analysis to the zones used. If only one set of zones is available at the scale, the trend in scale effects and the empirical variance of the zoning distribution observed in this paper suggest that it is worthwhile evaluating the zoning distribution at a higher scale to give some indication of the possible means and variance of the zoning distribution at the scale of interest.

If the zoning distribution at a particular scale can be estimated, then given results for one set of zones it may be possible to make a judgment regarding the results which may be obtained for another set of zones at that scale. For example, using prediction intervals, a prediction of the parameter estimates obtained with a different set of zones can be made. For this study, relative prediction intervals of ±10% to ±15% were frequently obtained from the zoning distributions, although some parameter estimates were more substantially affected. Consequently the use of a particular set of zones introduces an additional source of error, and knowledge of the zoning distribution allows it to be quantified and compared with the other sources of error.

By undertaking analyses at several scales or using several sets of zones at a given scale, the average value of the zoning distribution may be obtained. A major finding of the analyses conducted here is that the ecological average,

$E_k[\hat{\beta}^E]$, appears to vary systematically with scale allowing the effect of zoning at a given scale to be predicted. Moreover, it is possible to re-aggregate the data to a higher scale and then to predict the variance of the zoning distribution at a lower scale based on the variance at the higher scale. The implications of these results are that given one observation on a zoning distribution at one scale, if the data are aggregated in a number of ways to several different scales, the relationships between the scales can be exploited to help asses the possible mean and variance of the zoning distribution for the scale of interest. Moreover at a given level above the lowest scale, it is possible to make a partial adjustment of the estimator to its average value, to account for the zoning distribution. For example by extrapolating the observed relationship between the ecological average and the scale of analysis down to the lowest scale at which data are available.

To draw conclusions about a different scale requires an understanding of how the expectation of the zoning distribution varies with scale. Obtaining parameter estimates at a scale which is higher than the scale of interest is not difficult, as with sufficient zones the data can be merged in multiple ways to a higher scale. Going down a level is more difficult, but if the zoning distribution can be related in a systematic way to the scale of the analysis, then prediction and estimation at lower levels is possible. An example of this is that the zoning distribution at level 1, say, may help us in assessing the CD level zoning distribution which should be used with the single estimate that we have for the CD level.

In conclusion, the characteristics of the zones used to aggregate the data are an important aspect of the analysis for any type of study using small area health data or when population grouping is involved. In all studies there is a need to carefully consider the zones used in the analyses and the zoning distribution that applies. This paper provides an extensive systematic investigation of the characteristics of zoning distributions for parameter estimates obtained from the analysis of small area health data using an ecological model.

5 Appendix 1

[Table 5 about here.]

References

Australian Bureau of Statistics (2006a). *Census Dictionary.* cat. no. 2901.0 (reissue) ABS, Canberra.

Australian Bureau of Statistics (2006b). *Socio-Economic Indexes for Areas (SEIFA) - Technical Paper.* cat. no. 2039.0.55.001. ABS, Canberra.

Australian Bureau of Statistics (2008). *National Health Survey 2007-08.* Basic CURF, CD-ROM. Findings based on use of ABS CURF data.

Australian Bureau of Statistics (2009). *National Health Survey: Users' guide - Electronic Publication, 2007-08.* Cat. No. 4363.0.55.001. Viewed 21 June 2012 http://www.ausstats.abs.gov.au/ausstats/subscriber.nsf/0/CC0FB5A08570984ECA25762E0017CF2B/$File/4363055001_2007-08.pdf.

Best, N., Cockings, S., Bennett, J., Wakefield, J., and Elliott, P. (2001). Ecological regression analysis of environmental benzene exposure and childhood leukaemia: Sensitivity to data inaccuracies, geographical scale and ecological bias (Pkg: p141-207). *Journal of the Royal Statistical Society, Series A*, 164(1):155–174.

Briant, A., Combes, P.-P., and Lafourcade, M. (2010). Dots to boxes: Do the size and shape of spatial units jeopardize economic geography estimations. *Journal of Urban Economics*, 67:287 – 302.

Burden, S. and Steel, D. (2013). Microsimulation of health data to retain spatial structure for small areas. Working Paper 20-13, University of Wollongong.

Cockings, S., Harfoot, A., Martin, D., and Hornby, D. (2011). Maintaining existing zoning systems using automated zone-design techniques: methods for creating the 2011 census output geographies for england and wales. *Environment and Planning A*, 43:2399 – 2418.

Cockings, S. and Martin, D. (2005). Zone design for environment and health studies using pre-aggregated data. *Social Science and Medicine*, 60:2729–2742.

Diez-Roux, A. and Mair, C. (2010). Neighbourhoods and health. *Annals of the New York Academy of Sciences*, 1186:125 – 145.

Duque, J., Ramos, R., and Suriñach, J. (2007). Supervised regionalisation methods: a survey. *International Regional Science Review*, 30(3):195 – 220.

Flowerdew, R., Geddes, A., and Green, M. (2001). Behaviour of regression models under random aggregation. pages 89–104. John Wiley and Sons, Chichester.

Greenland, S. (2002). A review of multilevel theory for ecologic analyses. *Statistics in Medicine*, 21(1):389–395.

Haynes, R., Daras, K., Reading, R., and Jones, A. (2007). Modifiable neighbourhood units, zone design and residents perceptions. *Health and Place*, 13:812 – 825.

Haynes, R., Jones, A., Reading, R., Daras, K., and Emond, A. (2008). Neighbourhood variations in child accidents and related child and maternal characteristics: does area definition make a difference. *Health and Place*, 14:693 – 701.

Martin, D. (2003). Extending the automated zoning procedure to reconcile incompatible zoning systems. *Int. J. Geographical Information Science*, 17(2):181–196.

Mu, L. and Wang, F. (2008). A scale-space clustering method: mitigating the effect of scale in the analysis of zone-based data. *Annals of the Association of American Geographers*, 98(1):86 – 101.

Openshaw, S. (1977). A geographical solution to scale and aggregation problems in region-building partitioning and spatial modelling. *Transactions of the Institute of British Geographers*, 2(4):459–472.

Openshaw, S. (1984). *The modifiable area unit problem*, volume 38 of *Concepts and techniques in modern geography*. Geobooks, Norwich.

Openshaw, S. and Rao, L. (1995). Algorithms for reengineering 1991 census geography. *Environment and Planning*, 27:425–446.

Parenteau, M.-P. and Sawada, M. (2011). The modifiable areal unit problem (maup) in the relationship between exposure to no_2 and respiratory health. *International Journal of Health Geographics*, 10:58.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rasbash, J., Charlton, C., Browne, W., Healy, M., and Cameron, B. (2009).

MLwiN version 2.1. Technical report, Centre for Multilevel Modelling,, University of Bristol.

Richardson, S., Stücker, I., and Hémon, D. (1987). Comparison of relative risks obtained in ecological and individual studies: some methodological considerations. *International journal of epidemiology*, 16:111–120.

Schuurman, N., Bell, N., Dunn, J., and Oliver, L. (2007). Deprivation indices, population health and geography: an evaluation of the spatial effectiveness of indices at multiple scales. *Journal of Urban Health*, 84:591 – 603.

Stafford, M., Duke-Williams, O., and Shelton, N. (2008). Small area inequalities in health: are we underestimating them? *Social Science and Medicine*, 67:891 – 899.

Steel, D. and Holt, D. (1996). Rules for random aggregation. *Environment and Planning A*, 28:957–978.

Steel, D., Tranmer, M., and Holt, D. (2003). Analysis combining survey and geographically aggregated data. In *Analysis of Survey Data*, chapter 20, pages 323–343. John Wiley and Sons, London.

Swift, A., Liu, L., and Uber, J. (2008). Reducing maup bias of correlation statistics between water quality and gi illness. *Computers, Environment and Urban Systems*, 32:134 – 148.

Wakefield, J. (2004). A critique of statistical aspects of ecological studies in spatial epidemiology. *Environmental and Ecological Statistics*, 11:31–54.

Table 1: CD level health outcome and risk factor summary data for the simulated population of NSW

	Total	Mean	Standard	Coeff.	2.5 %	97.5 %
	('000)		Deviation	Variation	percentile	percentile
BMI		27.1	5.17	0.191	23.4	39.2
Diabetes	199	0.031	0.0015	0.049	0.028	0.034
Angina	121	0.019	0.0019	0.099	0.015	0.023
Smoking	1049	0.164	0.0071	0.043	0.151	0.178
Sedentary	1955	0.306	0.0070	0.023	0.293	0.320
Obesity	763	0.120	0.0033	0.028	0.113	0.126
Dietary Fat	2924	0.458	0.0070	0.015	0.445	0.472

Table 2: Population targets and constraints used to create sets of zones in AZTool and the resulting population statistics averaged for each scale.

Scale	Pop.	Range	No.	Mean	Std	Min	Max	Coeff.
	Target	('000)	Zones		Dev			Variation
CD			11879	537	258	3	2755	0.481
L1	1000	0.5 – 3	6214	1026	210	636	2755	0.205
L2	2000	1–4	3168	2013	237	1373	3180	0.117
L3	4000	2–8	1585	4024	309	2783	5502	0.077
L4	6000	3–12	1056	6040	376	4429	8269	0.062
L5	8000	4 – 16	792	8053	444	5904	10464	0.055
L6	10000	5 – 20	634	10060	506	8004	12613	0.05
L7	15000	7.5 – 30	423	15078	678	11881	18018	0.044
L8	20000	10 – 40	317	20120	834	17037	23781	0.041

Table 3: Comparison of the ecological averages of the parameter estimates ($E_k[\hat{\beta}^E_{r,k}]$) for levels 8, 4 and 1 with the corresponding CD and multilevel estimates (using groups at scale $l_k = 4$) (line 1); and variance estimates $Var_k(\hat{\beta}^E)$ (and $E_k[\hat{SE}(\hat{\beta}^E_{r,k})]$) (line 2) for the statistical models for BMI, angina and diabetes. For the multilevel estimates $Var_k(\hat{\beta}^E) < 1e-04$ so they are not included. Covariates for age and sex were also included in the model for BMI.

	Level 8	Level 4	Level 1	CD Model	ML Model
BMI					
Constant	27.8◇	27.4	27.9	27.9	29.9
	0.731 (1.76)	0.214 (0.841)	0.0262 (0.283)		(0.543)
Sedentary	-4.4◇	-2.36	-1.28	-0.235	0.508
	0.488 (1.35)	0.128 (0.653)	0.0122 (0.234)		(0.101)
Smoking	6.3◇	4.56	2.24	1.68e-04	-0.203
	0.191 (1.01)	0.0667 (0.511)	0.0108 (0.203)		(0.122)
Dietary Fat	-1.78⋆	-1.98	-0.694	0.584	-1.26
	0.395 (1.40)	0.11 (0.661)	0.0138 (0.217)		(0.1)
HSEIA	-0.00242◇	-0.00229	-0.00185	-0.00129	-0.00279
	1.46e-07 (8.18e-04)	4.62e-08 (3.89e-04)	6.86e-09 (1.24e-04)		(4.77e-04)
ANGINA					
Constant	4.48	3.34	1.59	0.8	-2.61
	0.176 (0.702)	0.0633 (0.383)	0.00745 (0.157)		(0.257)
Sedentary	3.74	3.46	2.75	2.31	0.988
	0.049 (0.478)	0.019 (0.272)	0.00323 (0.129)		(0.0605)
Obesity	-6.38	-5.34	-3.6	-2.67	0.528
	0.144 (0.641)	0.0483 (0.375)	0.00687 (0.19)		(0.0734)
Dietary Fat	-5.65	-4.26	-1.83	-0.74	-0.55
	0.102 (0.551)	0.0384 (0.325)	0.0054 (0.149)		(0.0628)
HSEIA	-0.00223	-0.00178	-0.00117	-8.69e-04	-0.00167
	3.56e-08 (3.18e-04)	1.28e-08 (1.75e-04)	1.65e-09 (7.42e-05)		(2.54e-04)
σ^2_ν	0.012	0.0185	0.0404	0.0545	0.151
DIABETES					
Constant	-2.52	-1.74	-0.477	-0.0708	-2.44
	0.0695 (0.475)	0.0221 (0.268)	0.0027 (0.114)		(0.214)
Sedentary	0.803◇	0.595	0.0697⋆	-0.138	0.629
	0.0202 (0.321)	0.00864 (0.189)	0.00143 (0.0938)		(0.0471)
Obesity	0.416⋆	0.333⋆	0.538	0.822	1.33
	0.0704 (0.443)	0.0216 (0.271)	0.00321 (0.141)		(0.0495)
Dietary Fat	3.6	2.84	1.58	1.07	-0.72
	0.048 (0.38)	0.0157 (0.231)	0.00225 (0.108)		(0.0502)
HSEIA	0.000565◇	0.00021⋆	-0.000348	-4.98e-04	-0.00126
	1.37e-08 (2.14e-04)	4.31e-09 (1.21e-04)	5.37e-10 (5.35e-05)		(2.11e-04)
σ^2_ν	0.0047	0.0082	0.0131	0.0174	0.0815

Standard errors in parentheses

◇ indicates that parameter estimate significant for less than 90% of zones

⋆ indicates that parameter estimate significant for less than 50% of zones

Table 4: Relative width of the predictive interval for a parameter estimate obtained for a new set of zones (in the same study area) at the same scale. It is written as a percentage of the average value of the parameter estimate, X, i.e. $E[\hat{\beta}](1 \pm X)$ where $X = 1.96\sqrt{Var_k(\hat{\beta})}/E_k[\hat{\beta^E}]$

Level	1	2	3	4	5	6	7	8
BMI								
Constant	0.0114	0.0194	0.0271	0.0331	0.0383	0.0424	0.0512	0.0603
Sedentary	0.169	0.245	0.298	0.297	0.302	0.301	0.298	0.311
Smoking	0.091	0.0912	0.0894	0.111	0.119	0.123	0.136	0.136
Diet. Fat	0.331	0.336	0.292	0.329	0.376	0.449	0.569	0.691
HSEIA	0.088	0.129	0.153	0.184	0.198	0.227	0.283	0.31
ANGINA								
Constant	0.106	0.13	0.147	0.148	0.153	0.164	0.17	0.184
Sedentary	0.0405	0.0556	0.0728	0.078	0.0852	0.0909	0.101	0.116
Obesity	0.0452	0.0588	0.0713	0.0807	0.0854	0.0925	0.102	0.117
Diet. Fat	0.0786	0.0867	0.0885	0.0903	0.0899	0.0973	0.102	0.111
HSEIA	0.0679	0.0937	0.118	0.124	0.131	0.143	0.152	0.166
DIABETES								
Constant	0.214	0.21	0.179	0.167	0.169	0.177	0.183	0.205
Sedentary	1.06	0.504	0.32	0.306	0.313	0.304	0.326	0.347
Obesity	0.206	0.568	0.92	0.864	0.921	1.06	1.04	1.25
Diet. Fat	0.0588	0.0759	0.0849	0.0864	0.0876	0.0982	0.106	0.119
HSEIA	0.13	0.465	1.91	0.612	0.478	0.441	0.386	0.406

Table 5: Variables from NHS0708 used in the HSEIA index with their loadings.

Weight	NHS0708 Variable Description
-0.334	% Gross weekly equivalent cash income (∼ 2nd - 3rd decile)
-0.294	% Persons over 15 with no post-school qualification
-0.290	% Persons with a government concession card
-0.266	% Families: one parent with dependent offspring only
-0.256	% Households renting from government/community organisation
-0.226	% Persons less than 70 with disability requiring help with core activities
-0.223	% Occupation: labourers
-0.174	% Labour force status: unemployed
-0.169	% Occupation: machinery operators and drivers
0.123	% occupied private dwellings with four or more bedrooms
0.142	% Persons over 15 years at university or other tertiary institution
0.159	% Occupation: Managers
0.264	% Persons over 15 Diploma/adv. diploma only non-school qualification
0.298	Gross weekly equiv. cash income (∼ 9th - 10th decile)
0.313	% Occupation: professionals
0.324	% Persons with private health insurance

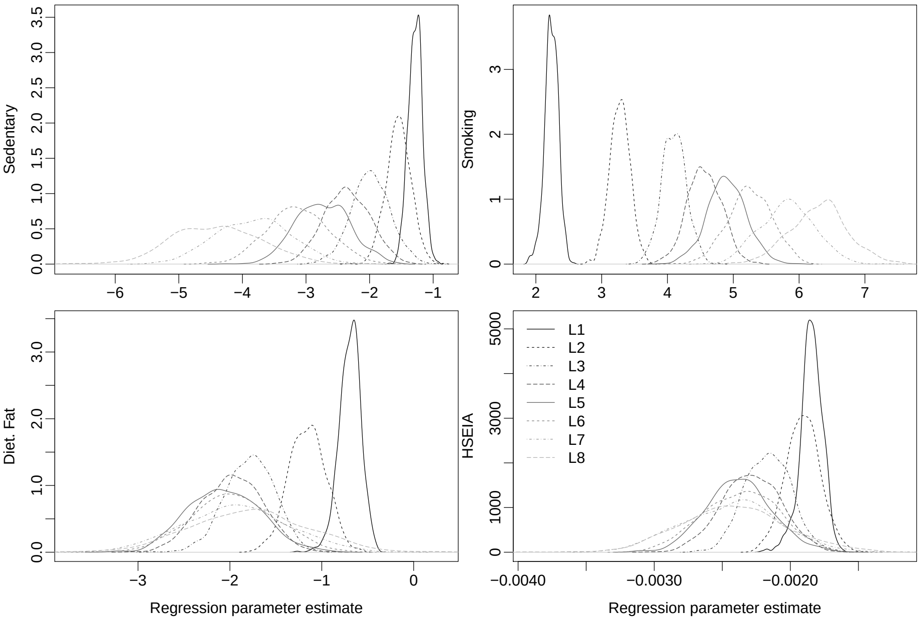

Figure 1: Density plots of the zoning distribution of the ecological regression coefficients for Sedentary, obesity, dietary fat, HSEIA on BMI at eight scales

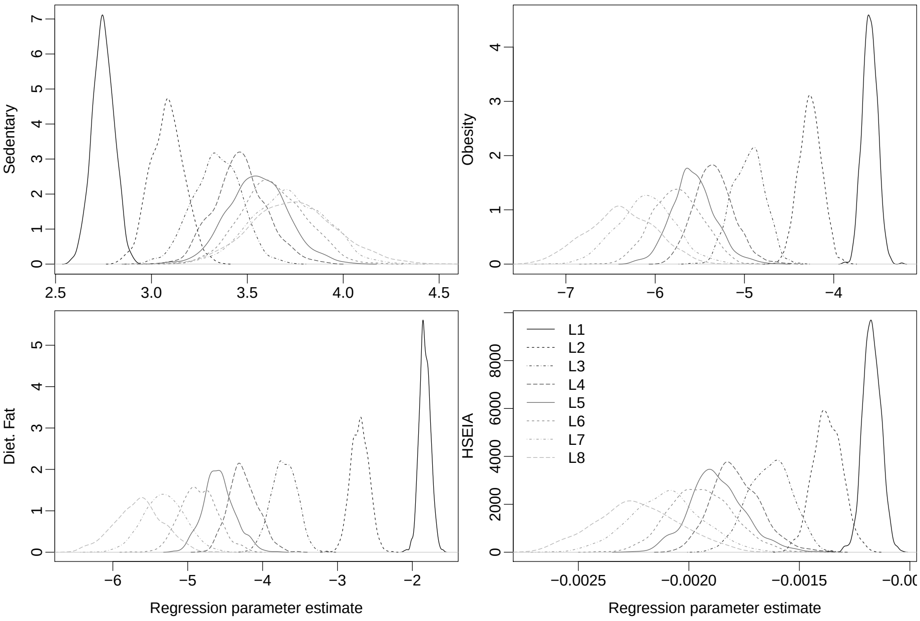

Figure 2: Density plots of the zoning distribution for the ecological regression coefficients for Sedentary, obesity, dietary fat and HSEIA on angina at eight scales

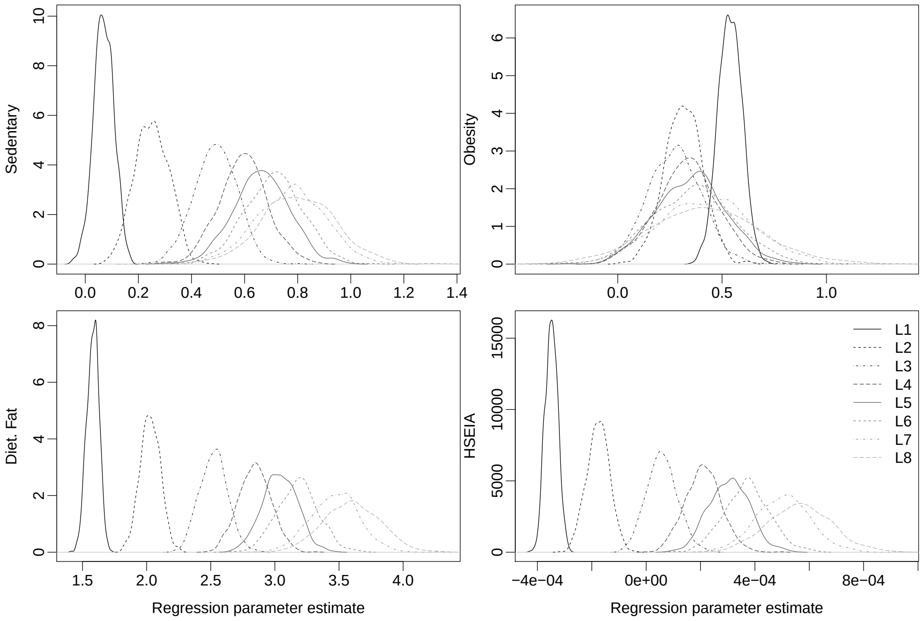

Figure 3: Density plots of the zoning distribution for the ecological regression coefficients for Sedentary, obesity, dietary fat and HSEIA on diabetes at eight scales

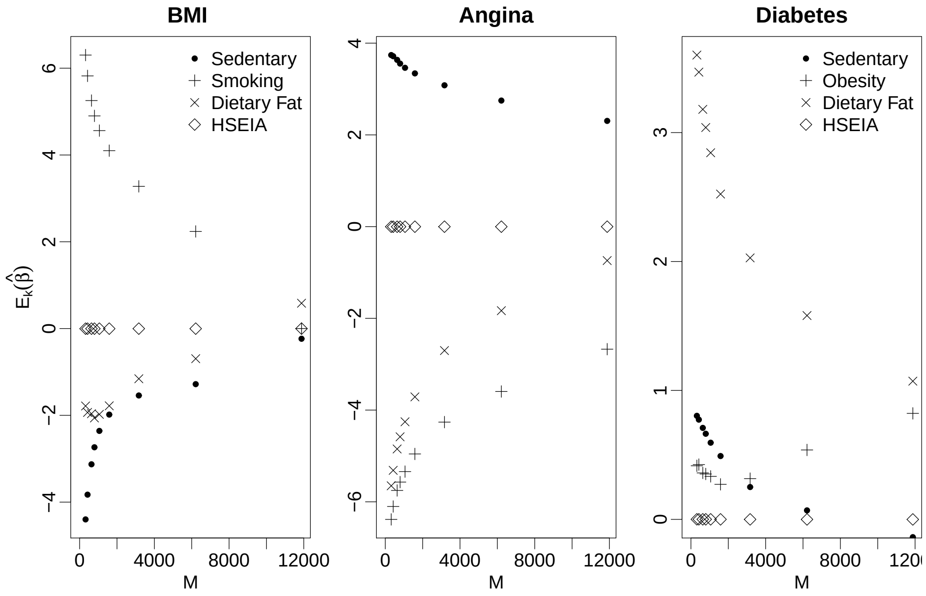

Figure 4: Plot of the ecological average of the parameter estimates $E_k[\hat{\beta}^E]$ versus M for each scale of analysis for BMI, angina and diabetes. The legend for angina is the same as for diabetes

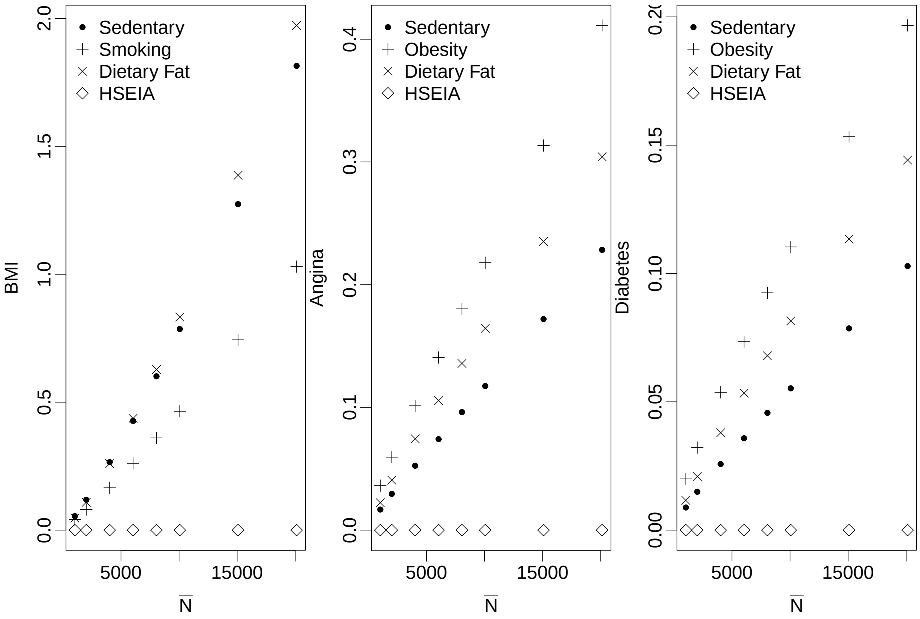

Figure 5: $E_k[\widehat{Var}(\hat{\beta}^E)]$ versus $\bar{N}$ for each scale of analysis for BMI, angina and diabetes

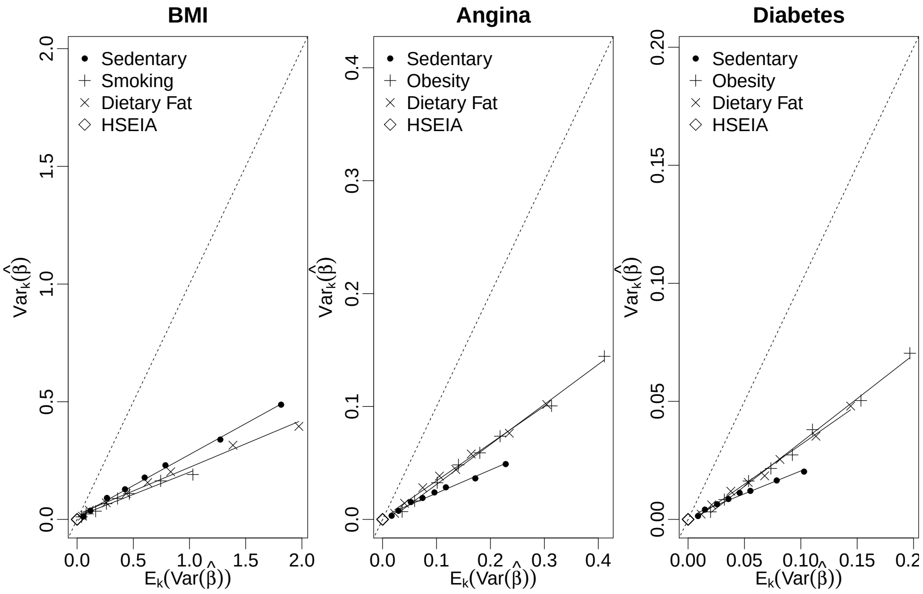

Figure 6: $E_k[\widehat{Var}(\hat{\beta}^E)]$ versus $Var_k(\hat{\beta}^E)$ for each scale of analysis for BMI, angina and diabetes