2013

Characteristics of empirical zoning distributions for small area health data

Sandy Burden
University of Wollongong, sburden@uow.edu.au

David Steel
University of Wollongong, dsteel@uow.edu.au

NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

The University of Wollongong

Working Paper

15-13

Characteristics of Empirical Zoning Distributions for Small Area Health Data

Sandy Burden and 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

Characteristics of Empirical Zoning Distributions for Small Area Health Data

Sandy Burden and David Steel

May 10, 2013

Abstract

Many studies of health utilise a multilevel modelling framework and if individual level data are not available use ecological inference to obtain individual level parameter estimates using area-level data summaries, resulting in biased parameter estimates and increased variance. For these studies, the modifiable area unit problem means that the scale of the analysis and the zones used to aggregate the data affect the amount and direction of the bias and the increase in variance. To investigate the effects of scale and zoning, in this paper the distribution of the parameter estimates for over many sets of zones at the same scale (the zoning distribution) is obtained for parameter estimates from an ecological model at multiple scales of analysis. The distributions are typically symmetrical and unimodal and can be considered to follow a normal distribution. The estimated average parameter estimate (ecological average) displays systematic variation with scale and is related to $\sqrt{M-1}$. The variance of the distribution is related to the average number of observations in the areas. The implications of creating and using a zoning distributions are wide ranging as they allow the estimates for a given set of zones at the same or a different scales to be compared and assessed.

1 Introduction

The modifiable area unit problem (MAUP) is the sensitivity of the results of an analysis to the spatial areas or zones used to aggregate or analyse the data. It has two aspects: the scale problem, which occurs when a smaller or larger number of areas is used to analyse the data and the zoning problem which arises when the areas at a given scale are defined using using different boundaries [Flowerdew et al., 2001]. Changing either of these factors may alter the estimates which are obtained, but to date no apparently systematic trends have emerged [Openshaw, 1984]. The MAUP means that the results of area level analyses can only legitimately be applied to the particular areal units used, as an inference may change when an alternative set of areas is used, even at the same scale. 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].

The MAUP occurs because aggregation removes the link between individual response and covariate values. For example, studies of health may use 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 spatial data may be defined using geographic areas. Frequently, for privacy reasons, only aggregate summaries are available and a common approach is to use ecological inference which substitutes area level summaries for the individual level data. However, the estimates of the parameters for this model may be biased because the within-area variability of the data is not available and the amount and direction of bias depends on the geographic areas used to analyse the data. The parameter estimates are not biased for the individual level target of inference for a properly specified linear model [Tranmer and Steel, 2001], although they will be for a non-linear link function. Changes in scale substantially affect the variance of the estimates [Steel and Holt, 1996], but the effect of moving the zone boundaries

is not apparently systematic [Stafford et al., 2008, Haynes et al., 2008]. Homogenous areas are least affected by the MAUP [Briant et al., 2010] as in this case most of the variability between data values is between areas.

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. 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 aggregate data analysis have been widely considered in the fields of spatial epidemiology, geography and the social sciences [for example, see Greenland, 2002, Steel et al., 2003, Richardson et al., 1987, Wakefield, 2004]. However, these 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, the selection or definition of analysis zones at several scales is used to understand the effect of scale and zoning on regression parameter estimates obtained using aggregate health data. In the next section zoning distributions are introduced. They are used throughout the paper to describe the variation in parameter estimates for different sets of zones. In Section 3 the methodology used to create empirical zoning distributions for aggregate health data is described. The results of the analysis are presented in Section 4 and discussed in Section 5.

2 The Zoning Distribution

A set of zones is formed when a given study region is partitioned into M geographically contiguous, non-overlapping parts. Individuals in the population are assigned to the zones using measures of geographic location, such that all individuals belong to exactly one zone and each zone has $N_g \geq 1$ observations, $g = 1, ..., M$. For a parameter or a statistic θ evaluated for a given set of zones, an estimate, $\hat{\theta}$, is obtained.

Define the zoning distribution of the parameter estimate $f(\hat{\theta})$ as the distribution or density function of the estimate over all possible sets of M zones, given the constraints used in defining the zones. 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, and variance provide a way to compare and standardise the results for a set of zones at a given scale. Zoning distributions can 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 for area level data, other than in the case of purely random grouping, 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 [Steel and Holt, 1996]. 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. To identify appropriate assumptions for zoning distributions, in this paper empirical estimates of zoning distributions are created for the parameter estimates from a statistical model for small area health data.

3 Methodology

In this project, empirical zoning distributions were created for parameter estimates obtained from a regression model for three health outcomes and a set of covariates using the following steps. Firstly, spatially detailed health data was simulated for 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 [Australian Bureau of Statistics, 2008, 2009] were combined with summary data defining the characteristics of CDs from the 2006 Australian Census [Australian Bureau of Statistics, 2006b] using a spatial microsimulation model (MSM) as described in Burden and Steel [2013].

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 each variable. 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.

An area level indicator of socioeconomic status (denoted HSEIA) was also 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

Table 1: Summary of health outcomes and risk factors for the simulated population of NSW

	Total ('000)	Mean	Standard Error	Coeff. Variation	2.5 % percentile	97.5 % 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

Bureau of Statistics, 2006a, p.17–23]. Despite being created using different datasets and some different variables, the distribution of deciles assigned to each area for the two indices was similar. Overall, HSEIA included 16 variables, 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 with spatial location information known to the CD level. The location of individuals within CD's was not defined. The simulated data was then rezoned to higher levels of aggregation using the AZTool Software [Cockings et al., 2011, Martin, 2003]. AZTool randomly allocated CDs to analysis zones whilst preserving geographic contiguity. It then iteratively swapped CDs between adjacent zones to improve the population target and achieve upper and lower population limits for each zone. Table 2 shows the constraints used to define the sets of zones at each scale. At eight scales of analysis, each of 1000 sets of zones was defined using a single run of AZTool with 15 sets of swap iterations. Average population statistics for the zones are summarised for each scale in Table 3. At each scale, the average population per zone was very close to the target 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.

The data were aggregated to each set of analysis zones and the resulting zone level sum-

Table 2: The population target and constraints used in AZTool to create 1000 zones at each of eight scales.

Scale	Population Target	Range ('000)	No. Areas
L1	1000	0.5 – 3	6338
L2	2000	1–4	3169
L3	4000	2–8	1585
L4	6000	3–12	1056
L5	8000	4 – 16	792
L6	10000	5 – 20	634
L7	15000	7.5 – 30	423
L8	20000	10 – 40	317

Table 3: Population statistics for the sets of zones at each scale averaged over the 1000 sets of zones forming the zoning distribution at each given scale.

Level	No. Zones	Mean	Std Dev	Min	Max	Coeff. Variation
CD	11879	537	258	3	2755	0.481
1	6214	1026	210	636	2755	0.205
2	3168	2013	237	1373	3180	0.117
3	1585	4024	309	2783	5502	0.077
4	1056	6040	376	4429	8269	0.062
5	792	8053	444	5904	10464	0.055
6	634	10060	506	8004	12613	0.05
7	423	15078	678	11881	18018	0.044
8	317	20120	834	17037	23781	0.041

maries were modelled to obtain a regression parameter estimate, $\boldsymbol{\beta}^E$, for each covariate. The population weighted average BMI in each area was modelled in terms of the average for age, sex, each indicator variable and HSEIA using an area level regression model. Angina and diabetes were modelled as count variables and making a rare disease assumption, an ecological model was specified using a Poisson distributed response, i.e. for rare diseases and large group sizes, the count of positive responses for each area was approximated by an independent Poisson random variable given the covariates and a normally distributed random effect for the zones. An offset was included in the model to account for differences in the population at risk in each area. It was calculated as the proportional counts of disease based

on the population size and age×sex structure in each area. The models were estimated using either generalised least squares (for BMI) or second order penalised quasi-likelihood (PQL2) in the MlWiN software [Rasbash et al., 2009]. For BMI, 23% and 33% of the models for levels 7 and 8 respectively did not converge, reducing the number of estimates used to define the empirical distributions.

The individual level data were also modelled to provide a reference for comparison with the ecological estimates. A linear (for BMI) or logistic (for angina and diabetes) multilevel statistical model was used to obtain parameter estimates for the regression coefficients, $\boldsymbol{\beta}$, for each covariate. Binary indicators of prevalence of angina and diabetes were modelled as Bernoulli random variables using a generalised linear model with a logistic link. The models were estimated using each set of zones to define the group level. 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 convergence using PQL2, so first order marginal quasi-likelihood (MQL1) was used for estimation.

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].

4 Empirical Zoning Distributions

The average parameter estimate at each scale (ecological average) and its standard error over the sets of zones are shown in Table 4 while the empirical variance and its standard error are given in Table 5. Zoning distribution density plots for the ecological regression parameter estimates ($\hat{\beta}^E$) are shown for BMI in Figure 1, for angina in Figure 2 and for diabetes in Figure 3. The domain of each distribution represents the range of parameter estimates which may be obtained for the given covariate and scale. The density curve defines

the probability of the estimate for the given statistical model. As Table 4 and Figures 1 to 3 show, the zoning distributions were generally unimodal, reasonably symmetric and were similar for all response variables. In general, with an increase in scale, the ecological average of each parameter estimate increased in absolute magnitude in a consistent direction, but the relative size of the change diminished with scale. Similar results were obtained for the quantiles of the zoning distributions. The distribution of the estimated parameter for the HSEIA covariate was more complex, reflecting its lack of statistical significance, taking both positive and negative values at each scale. Its average value did become slightly more negative with scale, but it remained very close to zero.

Table 4: Average parameter estimate $\hat{\beta}^E$ and its average standard error over the sets of zones at each scale for each covariate included in the models for BMI, Angina and Diabetes. Note that covariates age and sex were also included in the model for BMI.

Level	1	2	3	4	5	6	7	8
BMI								
Constant	27.9	27.5	27.5	27.4	27.4	27.4$^\diamond$	27.6$^\diamond$	27.8$^\diamond$
	(0.283)	(0.425)	(0.65)	(0.841)	(1)	(1.15)	(1.48)	(1.76)
Sedentary	-1.28	-1.54	-1.98	-2.36	-2.74	-3.13$^\diamond$	-3.83$^\diamond$	-4.4$^\diamond$
	(0.234)	(0.344)	(0.515)	(0.653)	(0.775)	(0.886)	(1.13)	(1.35)
Smoker	2.24	3.28	4.1	4.56	4.9	5.25$^\diamond$	5.82$^\diamond$	6.3$^\diamond$
	(0.203)	(0.284)	(0.407)	(0.511)	(0.6)	(0.681)	(0.862)	(1.01)
Diet. Fat	-0.694	-1.16	-1.78	-1.98	-2.06$^\diamond$	-1.96$^\diamond$	-1.93$^\diamond$	-1.78$^\diamond$
	(0.217)	(0.33)	(0.51)	(0.661)	(0.792)	(0.912)	(1.18)	(1.4)
HSEIA	-0.00185	-0.00191	-0.00216	-0.00229	-0.00238	-0.00236$^\diamond$	-0.00243$^\diamond$	-0.00242$^\diamond$
	(0.000124)	(0.00019)	(0.000297)	(0.000389)	(0.000466)	(0.000535)	(0.000689)	(0.000818)
ANGINA								
Constant	1.59	2.18	2.92	3.34	3.58	3.77	4.16	4.48
	(0.157)	(0.221)	(0.314)	(0.383)	(0.443)	(0.494)	(0.606)	(0.702)
Sedentary	2.75	3.08	3.34	3.46	3.55	3.64	3.72	3.74
	(0.129)	(0.172)	(0.229)	(0.272)	(0.31)	(0.343)	(0.415)	(0.478)
Obesity	-3.6	-4.26	-4.96	-5.34	-5.57	-5.75	-6.1	-6.38
	(0.19)	(0.244)	(0.318)	(0.375)	(0.425)	(0.467)	(0.56)	(0.641)
Diet. Fat	-1.83	-2.7	-3.71	-4.26	-4.58	-4.85	-5.31	-5.65
	(0.149)	(0.202)	(0.273)	(0.325)	(0.368)	(0.405)	(0.485)	(0.551)
HSEIA	-0.00117	-0.00137	-0.00163	-0.00178	-0.00187	-0.00194	-0.0021	-0.00223
	(7.42e-05)	(0.000103)	(0.000144)	(0.000175)	(0.000201)	(0.000224)	(0.000274)	(0.000318)
DIABETES								
Constant	-0.477	-0.878	-1.41	-1.74	-1.95	-2.1	-2.39	-2.52
	(0.114)	(0.157)	(0.22)	(0.268)	(0.307)	(0.341)	(0.413)	(0.475)
Sedentary	0.0697$^\diamond$	0.251$^\diamond$	0.491	0.595	0.664	0.709	0.773	0.803$^\diamond$
	(0.0938)	(0.122)	(0.16)	(0.189)	(0.214)	(0.235)	(0.28)	(0.321)
Obesity	0.538	0.316$^\diamond$	0.272$^\diamond$	0.333$^\diamond$	0.352$^\diamond$	0.36$^\diamond$	0.425$^\diamond$	0.416$^\diamond$
	(0.141)	(0.179)	(0.232)	(0.271)	(0.304)	(0.332)	(0.391)	(0.443)
Diet. Fat	1.58	2.03	2.52	2.84	3.04	3.18	3.47	3.6
	(0.108)	(0.145)	(0.195)	(0.231)	(0.261)	(0.285)	(0.337)	(0.38)
HSEIA	-0.000348	-0.000177$^\diamond$	5.84e-05$^\diamond$	0.00021$^\diamond$	0.000306$^\diamond$	0.000371$^\diamond$	0.000504$^\diamond$	0.000565$^\diamond$
	(5.35e-05)	(7.26e-05)	(1e-04)	(0.000121)	(0.000139)	(0.000154)	(0.000186)	(0.000214)

Standard errors in parentheses

$\diamond$ indicates that parameter estimate significant for less than 95% of zones

Table 5: Variance of $\hat{\beta}^E$ (Var$[Var(\hat{\beta}^E)]$) over the zones at each scale for BMI, Angina and Diabetes.

Level	1	2	3	4	5	6	7	8
BMI								
Constant	0.0262	0.0742	0.144	0.214	0.288	0.352	0.522	0.731
	(5.86e-06)	(3.51e-05)	(0.000137)	(0.000249)	(0.000436)	(0.000699)	(0.00193)	(0.00337)
Sedentary	0.0122	0.0373	0.0911	0.128	0.178	0.23	0.339	0.488
	(2.03e-06)	(1.07e-05)	(4.7e-05)	(0.000102)	(0.000189)	(0.000337)	(0.000723)	(0.00181)
Smoker	0.0108	0.0232	0.035	0.0667	0.0889	0.11	0.164	0.191
	(1.47e-06)	(9.24e-06)	(3.86e-05)	(7.58e-05)	(0.000137)	(0.000201)	(0.000549)	(0.000845)
Diet. Fat	0.0138	0.0394	0.0703	0.11	0.157	0.202	0.315	0.395
	(2.91e-06)	(1.93e-05)	(9.12e-05)	(0.000206)	(0.000378)	(0.000571)	(0.00158)	(0.00287)
HSEIA	6.86e-09	1.59e-08	2.85e-08	4.62e-08	5.78e-08	7.48e-08	1.24e-07	1.46e-07
	(3.98e-12)	(1.33e-11)	(3.93e-11)	(7.05e-11)	(1.2e-10)	(1.75e-10)	(4.32e-10)	(6.76e-10)
ANGINA								
Constant	0.00745	0.021	0.0477	0.0633	0.0778	0.0989	0.131	0.176
	(2.67e-06)	(8.6e-06)	(2.63e-05)	(4.59e-05)	(8.05e-05)	(0.000117)	(0.00022)	(0.000404)
Sedentary	0.00323	0.00764	0.0154	0.019	0.0239	0.0284	0.0363	0.049
	(9.46e-07)	(2.73e-06)	(9.17e-06)	(2e-05)	(3.46e-05)	(4.9e-05)	(1e-04)	(0.000177)
Obesity	0.00687	0.0164	0.0325	0.0483	0.059	0.0737	0.101	0.144
	(1.36e-06)	(4.97e-06)	(1.78e-05)	(3.34e-05)	(6.14e-05)	(8.67e-05)	(0.000158)	(0.000292)
Diet. Fat	0.0054	0.0143	0.0281	0.0384	0.0442	0.0579	0.0764	0.102
	(1.21e-06)	(5.29e-06)	(1.61e-05)	(2.7e-05)	(4.71e-05)	(6.56e-05)	(0.000121)	(0.000204)
HSEIA	1.65e-09	4.3e-09	9.61e-09	1.28e-08	1.57e-08	2.01e-08	2.65e-08	3.56e-08
	(6.89e-13)	(1.8e-12)	(5.19e-12)	(9.21e-12)	(1.62e-11)	(2.36e-11)	(4.53e-11)	(8.38e-11)
DIABETES								
Constant	0.0027	0.00882	0.0165	0.0221	0.0283	0.0358	0.0498	0.0695
	(1.03e-06)	(3.05e-06)	(1.07e-05)	(2.12e-05)	(3.3e-05)	(4.89e-05)	(9.98e-05)	(0.000142)
Sedentary	0.00143	0.00416	0.00643	0.00864	0.0113	0.0121	0.0165	0.0202
	(3.35e-07)	(1.28e-06)	(4.12e-06)	(8.44e-06)	(1.31e-05)	(1.89e-05)	(3.9e-05)	(5.99e-05)
Obesity	0.00321	0.00837	0.0163	0.0216	0.0273	0.038	0.0503	0.0704
	(6.92e-07)	(2.62e-06)	(8.29e-06)	(1.69e-05)	(2.68e-05)	(3.99e-05)	(7.66e-05)	(0.000106)
Diet. Fat	0.00225	0.00617	0.0119	0.0157	0.0185	0.0254	0.0352	0.048
	(6.09e-07)	(2.1e-06)	(6.9e-06)	(1.28e-05)	(1.94e-05)	(2.78e-05)	(5.5e-05)	(7.2e-05)
HSEIA	5.37e-10	1.76e-09	3.24e-09	4.31e-09	5.57e-09	6.99e-09	9.84e-09	1.37e-08
	(2.52e-13)	(6.35e-13)	(2.15e-12)	(4.3e-12)	(6.6e-12)	(9.91e-12)	(2.03e-11)	(2.91e-11)

Variance of the standard errors in parentheses

Table 5 shows that the variance of the distributions increased substantially with scale and the proportional increase in variance with scale was similar for all covariates. The increased variance with scale is most likely due to the reduced power of the analysis as the mean covariate values for each zone became more similar with scale, reducing the observed variance. 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, or the apparent relationship may change sign, as observed for some covariates in the model for diabetes.

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/or M, reflecting that the variance is a

function of the degrees of freedom, $M - 1$. The trend with scale seems consistent with random aggregation even though the average parameter estimates do not exhibit random aggregation.

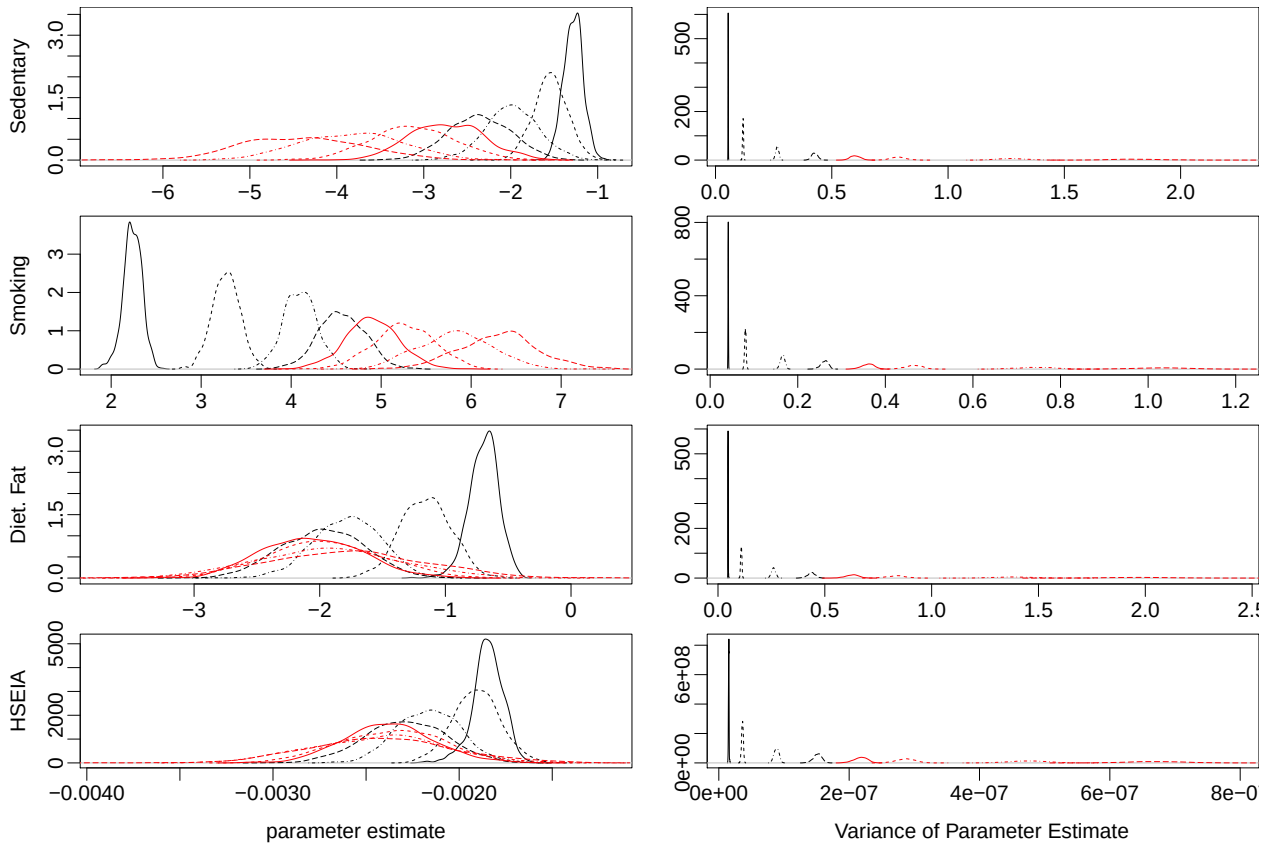

Figure 1: Density plots of the zoning distribution of the ecological regression coefficients Sedentary, obesity, dietary fat, HSEIA and the random effect on BMI at eight scales

For BMI and angina, the parameter estimates obtained for all individual level variables were statistically significant at the 5% level for all sets of zones at all scales but the estimates for HSEIA were not. For diabetes, obesity was not significant at the lowest scale and only significant for half the zones at level 2. Dietary fat was significant in fewer than 10% of the sets of zones as scale increased and sedentary varied between 30% and 90%, becoming more significant at higher scales. In many cases the size of the effects increased with scale, suggesting that a component of the bias is related to the scale of the analysis. However, the trend was not universal. For example, the results for the obesity parameter estimate are more complex and are not always significant.

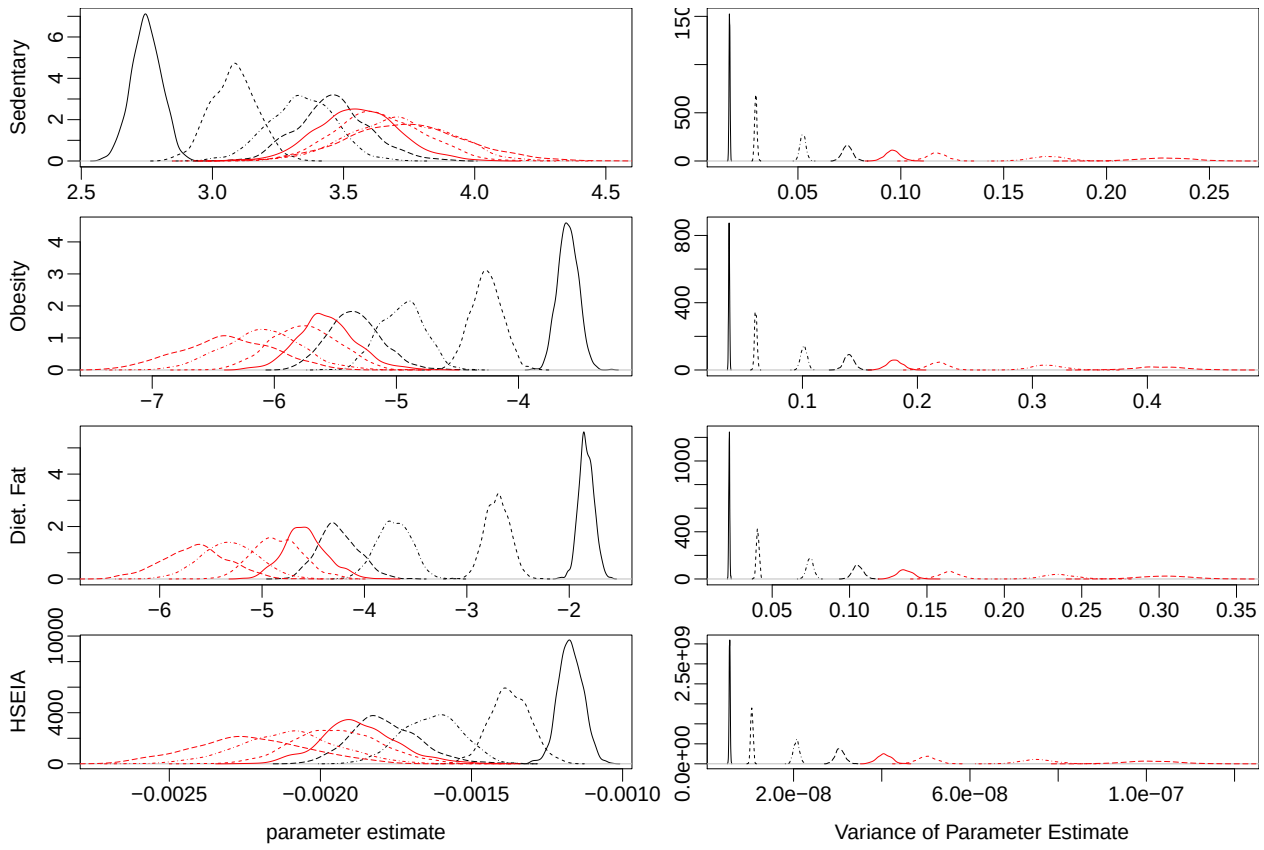

Figure 2: Density plots of the zoning distribution for the ecological regression coefficients Sedentary, obesity, dietary fat and HSEIA on angina at eight scales

The Shapiro Wilk test for normality was performed (using the R statistical software) for each zoning distribution. For BMI and diabetes, the null hypothesis was retained at most scales and normality of the zoning distribution could be assumed for all covariates. For angina, the results were mixed with the assumption of normality rejected for all covariates at level 4, all covariates except sedentary at level 5 and by HSEIA at levels 1 and 8. 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.

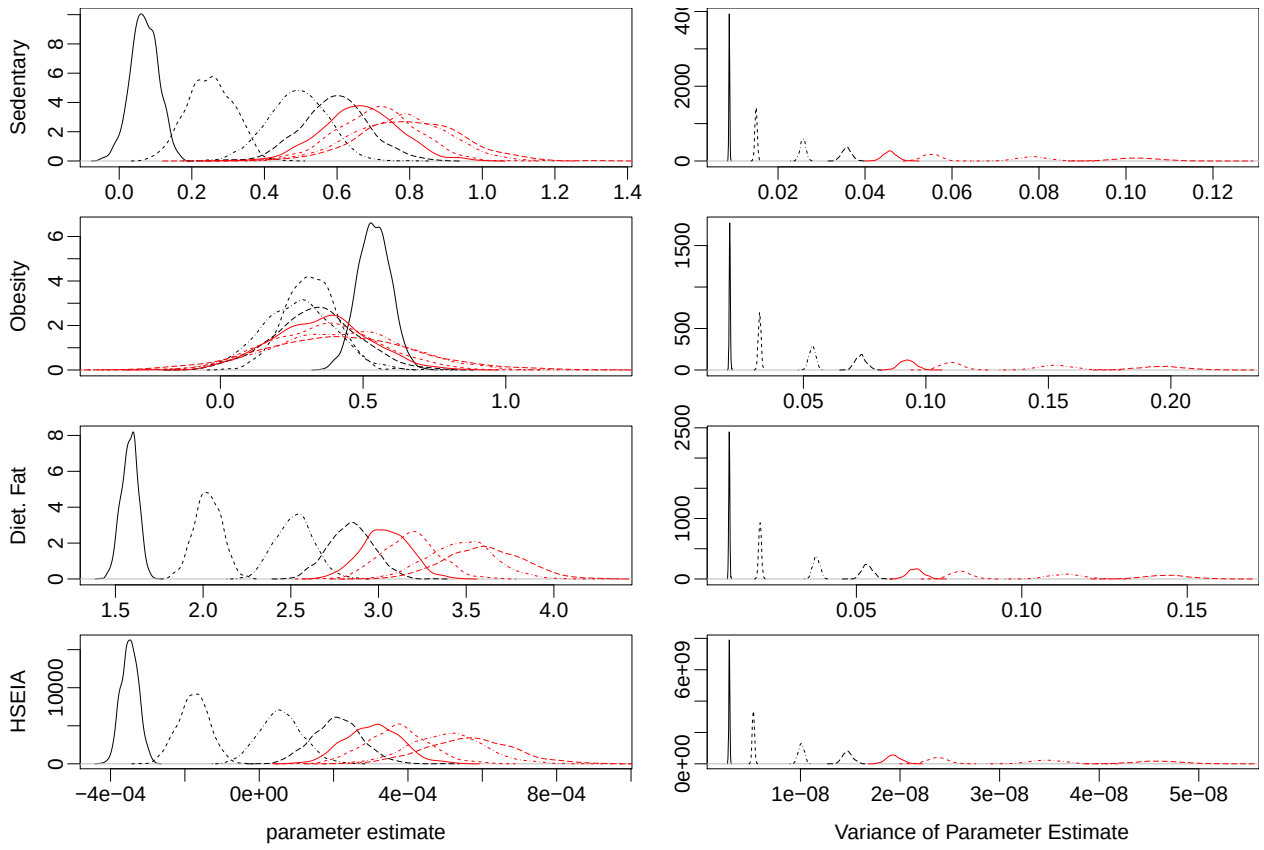

Figure 3: Density plots of the zoning distribution for the ecological regression coefficients Sedentary, obesity, dietary fat and HSEIA on diabetes at eight scales

4.1 Estimation of a Multilevel Target of Inference

A key advantage of the microsimulation approach is that estimates obtained from ecological and multilevel models can be compared using realistic aggregate and individual level data from the same population. Table 6 provides a summary of the average individual, CD level and multilevel model parameter estimates for each response variable. The multilevel model utilises the zones at level 4, although very similar results were obtained for groups at all scales.

Assuming that an individual level inference is required, the ecological parameter estimates are consistently biased and have generally higher variances than multilevel models, except when there are many small zones. In many cases and for most covariates, the difference between the ecological and multilevel parameter estimates was statistically significant due to

Table 6: Parameter estimates for the statistical models for BMI, angina and diabetes.

	Level 8	Level 4	Level 1	CD Model	ML Model
BMI					
Constant	27.8	27.4	27.9	27.9	29.9
Sedentary	-4.4	-2.36	-1.28	-0.235	0.508
Smoking	6.3	4.56	2.24	0.000168	-0.203
Dietary Fat	-1.78	-1.98	-0.694	0.584	-1.26
HSEIA	-0.00242	-0.00229	-0.00185	-0.00129	-0.00279
20-29	-5.2	-5.37	-5.39	-5.32	-1.18
30-39	-2.96	-1.51	-1.54	-2.29	0.311
40-49	0.0204	-0.248	0.432	0.551	0.975
50-59	5.19	4.29	3.16	1.92	1.38
60-69	-0.189	1.2	2.11	1.94	1.78
70-79	5.37	3.08	1.36	0.846	1.41
80+	-8.22	-5.16	-3.02	-3.28	0
Female	5.02	4.59	2.34	1.05	-0.776
ANGINA					
Constant	4.48	3.34	1.59	0.8	-2.61
Sedentary	3.74	3.46	2.75	2.31	0.988
Obesity	-6.38	-5.34	-3.6	-2.67	0.528
Dietary Fat	-5.65	-4.26	-1.83	-0.74	-0.55
HSEIA	-0.00223	-0.00178	-0.00117	-0.000869	-0.00167
DIABETES					
Constant	-2.52	-1.74	-0.477	-0.0708	-2.44
Sedentary	0.803	0.595	0.0697	-0.138	0.629
Obesity	0.416	0.333	0.538	0.822	1.33
Dietary Fat	3.6	2.84	1.58	1.07	-0.72
HSEIA	0.000565	0.00021	-0.000348	-0.000498	-0.00126

the bias associated with the ecological estimates. With increasing scale, the change in the bias of the ecological average and the average variance of the difference appears to behave systematically. However, predicting the magnitude of the bias for any given set of zones is much harder. If the zoning distribution were known, the bias of a particular estimate may be standardised by removing 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 implications of the zoning distribution can also be considered in terms of a predictive confidence 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 is equal to $2 \times 1.96\sqrt{Var(\hat{\beta})(1 + 1/K)}$, where K is the number of observations used to estimate the mean and variance of the zoning distribution. Table 7 shows these intervals as a proportion of the average value of the estimate at the relevant scale. The predictive confidence 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. However, wider relative prediction intervals are also frequently associated with smaller average parameter estimates, which may not be statistically significant.

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.

4.2 Relationships in the Data

In this section, relationships in the data that may affect and/or control the observed biases and increased variance of the parameter estimates are investigated. For a linear model, when data are randomly aggregated and the covariates are approximately normally distributed, the parameter estimates for a conditionally 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 shows that some of the regression parameter estimates for BMI, angina and diabetes appear to be related to $\sqrt{M - 1}$, but not linearly. However, there are parameter estimates for which the relationship does not hold at all and in the case of diabetes one parameter appears inversely

Table 7: Relative width of the predictive confidence 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 proportion of the average value of the parameter estimate, X, i.e. $E[\hat{\beta}](1 \pm X)$ where $X = 1.96 \times \sqrt{Var(\hat{\beta})} \times (1 + 1/1000)$, where K is the number of observations used to estimate the mean and variance of the zoning distribution

Level	1	2	3	4	5	6	7	8
BMI								
Constant	0.0114	0.0194	0.0271	0.0331	0.0384	0.0424	0.0513	0.0603
Sedentary	0.169	0.246	0.299	0.298	0.302	0.301	0.298	0.311
Smoking	0.0911	0.0912	0.0895	0.111	0.119	0.123	0.136	0.136
Diet. Fat	0.332	0.336	0.292	0.329	0.377	0.45	0.569	0.691
HSEIA	0.088	0.13	0.153	0.184	0.198	0.227	0.283	0.31
10-19								
20-29	0.0404	0.0618	0.09	0.11	0.132	0.148	0.198	0.243
30-39	0.213	0.375	0.561	0.641	0.723	0.73	0.684	0.658
40-49	0.641	2.47	1.67	3.14	6.55	4.54	16.8	82.4
50-59	0.0778	0.108	0.149	0.176	0.203	0.223	0.257	0.313
60-69	0.132	0.284	0.539	0.749	1.08	1.68	5.41	11
70-79	0.24	0.279	0.317	0.336	0.346	0.32	0.354	0.371
80+	0.106	0.155	0.185	0.206	0.226	0.247	0.272	0.291
Female	0.0981	0.124	0.141	0.148	0.181	0.199	0.233	0.286
ANGINA								
Constant	0.106	0.13	0.147	0.148	0.153	0.164	0.17	0.184
Sedentary	0.0406	0.0556	0.0728	0.078	0.0853	0.091	0.101	0.116
Obesity	0.0452	0.0588	0.0713	0.0807	0.0855	0.0925	0.102	0.117
Diet. Fat	0.0787	0.0868	0.0886	0.0903	0.09	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	0.207	0.568	0.92	0.864	0.921	1.06	1.04	1.25
Diet. Fat	0.0588	0.0759	0.0849	0.0864	0.0877	0.0983	0.106	0.119
HSEIA	0.131	0.465	1.91	0.612	0.479	0.442	0.386	0.407

related to $\sqrt{M-1}$.

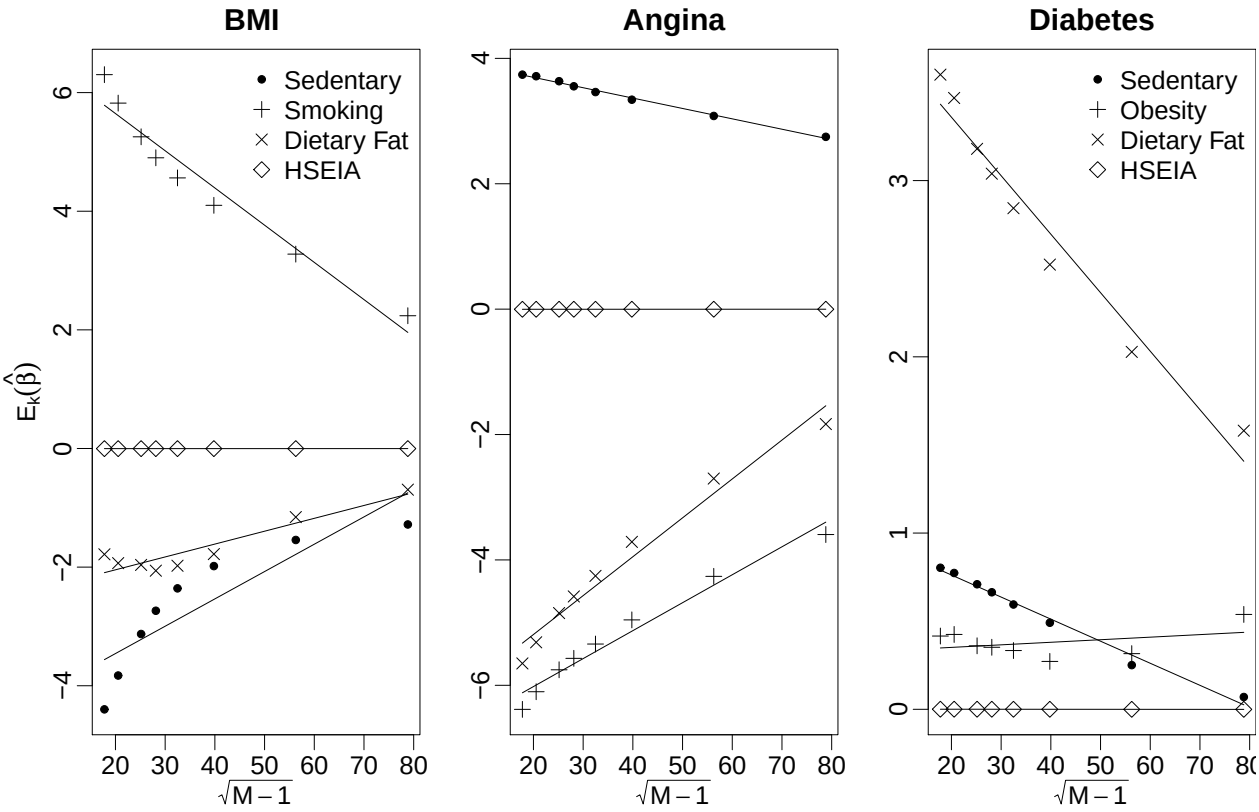

Figure 4: Plot of the ecological average of the parameter estimates $E[\hat{\beta}^E]$ versus $\sqrt{M-1}$ for each scale of analysis for BMI, angina and diabetes. The legend for angina is the same as for diabetes

The average empirical variance of the parameter estimates ($E_k[Var(\hat{\beta}^E)]$), over $k = 1000$ zones, is shown in Figure 5. For the response BMI, the expected value of the variance of the parameter estimates is proportional to the average population size in each area, $\bar{N}$ (and $1/\sqrt{M}$, which is not shown). Despite non-linearity in the models, similar results are obtained for angina and diabetes. In all cases the area level covariate HSEIA has a much lower variance compared with the individual level covariates. 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 a non-linear model. If random aggregation can be identified for both linear and non-linear models, then the scale at which the aggregation becomes random can also be identified. This means that

the ecological average of the parameter estimate can be predicted at lower scales, down to the scale at which the within area homogeneity affects the parameter estimates.

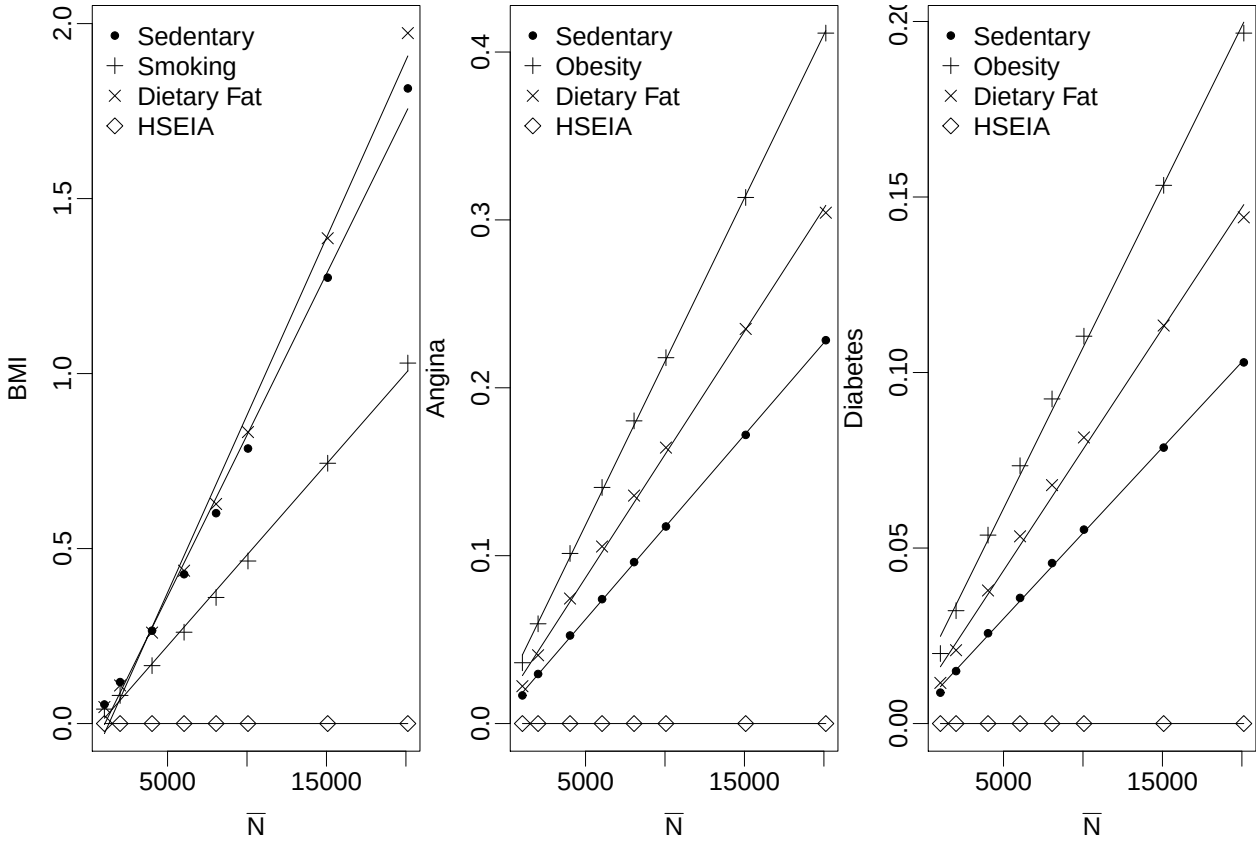

Figure 5: Plot of $E_k[Var(\hat{\beta}^E)]$ versus $\bar{N}$ for each scale of analysis for BMI, angina and diabetes

The relationship between the variance of the zoning distribution for each regression parameter estimate ($Var_k(\hat{\beta}^E)$) and $\bar{N}$ is shown in Figure 6. This estimate of the variance also exhibits the same linear relationship with $\bar{N}$ (and $1/\sqrt{M}$) with increasing scale for all of the response variables.

Comparing the two estimates of variance, the variance denoted by $E_k[Var(\hat{\beta}^E)]$ is the average value of the variance estimate for each regression coefficient obtained using each set of zones. The variance of the regression parameter estimate for the zoning distribution i.e. calculated over the 1000 sets of zones, is denoted $Var_k(\hat{\beta}^E)$. The two estimates are compared directly in Figure 7 for each of the response variables. In all cases the variance estimates are somewhat linearly related, but not exactly the same. The average empirical variance is close

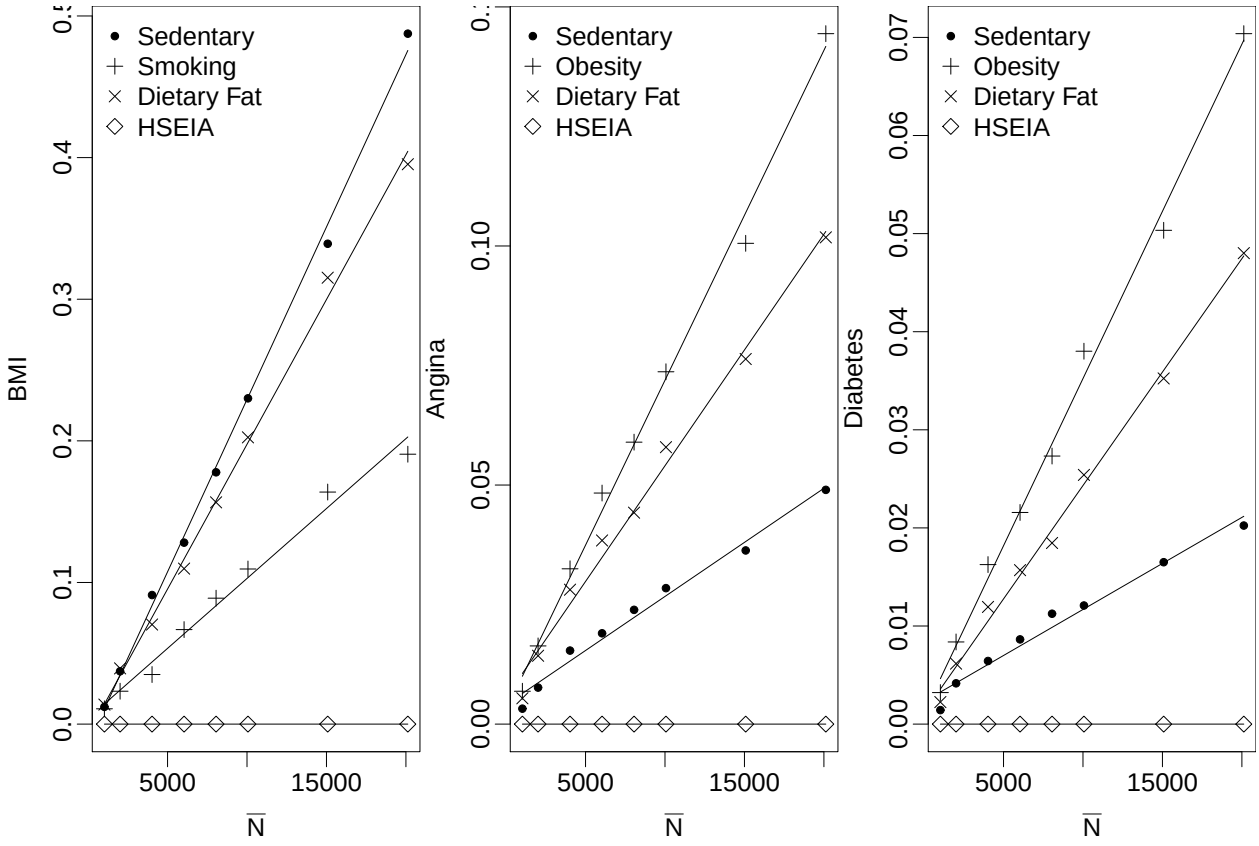

Figure 6: Plot of $Var_k(\hat{\beta}^E)$ versus $\bar{N}$ for each scale of analysis using linear ecological model for BMI, angina and diabetes

to four times greater than the zoning variance.

5 Discussion

These results provide a valuable insight into zoning distributions, because knowledge of the zoning distribution of a parameter estimate allows for its inclusion into a statistical model for area level data, improving the estimates obtained from the model and better characterising the data. It also allows confidence intervals for the expectation of the estimates over the zoning distribution to be obtained. The empirical zoning distributions are all relatively symmetrical and generally unimodal and the regression parameter estimates are approximately normally distributed. They suggest that for a continuous or a binary response variable, the zoning distributions of the parameter estimates obtained using a linear or logistic ecological

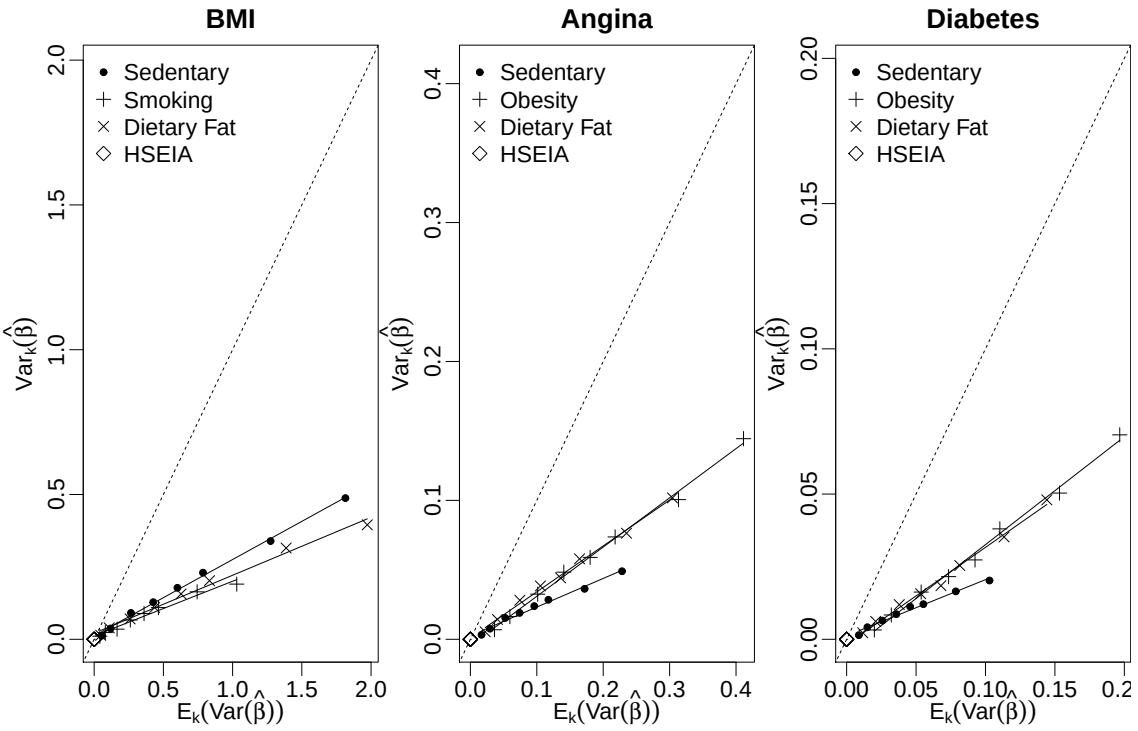

Figure 7: Relationship between $Var_k(\hat{\beta}^E)$ and $E_k[Var(\hat{\beta}^E)]$ for each scale of analysis using ecological models for BMI, angina and diabetes

model have a normal distribution. Moreover, the variance of each zoning distribution is a predictable function of $\bar{N}$ and the parameter estimates are related to $\sqrt{M-1}$. The variance of the zoning distribution is often appreciable, and should not be ignored when interpreting the results of statistical analysis based on aggregate data for geographic zones.

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 aggregation for both a linear and a logistic ecological model. 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. Comparing the parameter estimates for the population model and the multilevel model, in this application both gave similar parameter estimates. Using such a large amount of data, even with groups at the CD level, the effect of within zone homogeneity on the estimates is not appreciable.

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. The results are similar for analyses undertaken using both normally distributed continuous data and count data. However, 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.

Similarly, 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 $\pm 10\%$ to $\pm 15\%$ were frequently obtained from the zoning distributions, although some parameter estimates were more substantially affected. Consequently in many cases 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. 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.

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 $E_k[\hat{\beta}^E]$ appears to vary consistently 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.

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 linear or logistic model.

References

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

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

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

Australian Bureau of Statistics. *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, 2009.

N. Best, S. Cockings, J. Bennett, J. Wakefield, and P. Elliott. 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, 2001.

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

S. Burden and D. Steel. Microsimulation of health data to retain spatial struc-

ture for small areas. Working Paper 16-13, University of Wollongong, 2013. URL http://cssm.uow.edu.au/publications/index.html.

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

S. Cockings, A. Harf, D. Martin, and D. Hornby. 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, 2011.

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

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

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

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

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

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

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

L. Mu and F. Wang. 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, 2008.

S. Openshaw. 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, 1977.

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

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

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

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

J. Rasbash, C. Charlton, W. Browne, M. Healy, and B. Cameron. MLwiN version 2.1. Technical report, Centre for Multilevel Modelling,, University of Bristol., 2009.

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

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

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

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

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

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

M. Tranmer and D. Steel. Using local census data to investigate scale effects. In *Modelling scale in geographical information science*, chapter 6, pages 105 – 122. John Wiley and Sons, ltd, London, 2001.

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