Spatial Variation in Attributable Risks

Peter Congdon, School of Geography, QMUL, p.congdon@qmul.ac.uk

Abstract The attributable risk (AR) measures the contribution of a particular risk factor to a disease, and allows estimation of disease rates specific to that risk. While previous studies consider variability in ARs over demographic categories, this paper considers the extent of spatial variability in ARs estimated from multilevel data with confounders both at individual and geographic levels. A case study considers the AR for diabetes in relation to elevated BMI, and area rates for diabetes attributable to excess weight. Contextual adjustment includes known area variables, and unobserved spatially clustered influences, while spatial heterogeneity (effect modification) is considered in terms of varying effects of elevated BMI by neighbourhood deprivation category. The application is to patient register data in London, with clear evidence of spatial variation in ARs, and in small area diabetes rates attributable to excess weight.

Key words: Attributable risk. Multilevel. Spatial. Diabetes. Obesity. Deprivation. Bayesian.

1. Introduction

The attributable risk (AR) seeks to quantify the proportion of disease due to a particular risk factor, which may be termed the focus risk factor (Uter and Pfahlberg 2001; Benichou, 2001). Other terms include the attributable fraction, population attributable risk, and population attributable fraction. The AR measures impacts of risk factors on disease levels, taking into account both associations (i.e. relative risk) between disease and exposure, and the proportion of subjects exposed. Using attributable risks one may ascertain disease rates and burdens specific to a particular risk factor (Steenland & Armstrong, 2006; Ezzati et al, 2006; Gefeller, 1995). With a risk factor expressed in binary form, and P_E as the proportion of subjects exposed, a point estimator of the attributable risk is

$$[P_E(RR-1)]/ [P_E(RR-1)+1], \qquad (1a)$$

where RR is the relative risk for those exposed as compared to those unexposed. The latter should be adjusted for confounders (Darrow and Steenland, 2011; Benichou, 2001; Steenland and Armstrong, 2006). Another estimator is

$$P_{E|D}(RR-1)/RR= P_{E|D}(1-1/RR), \qquad (1b)$$

where $P_{E|D}$ is the proportion of diseased subjects exposed.

Variations in attributable risks over demographic categories (e.g. age categories, ethnic groups) have been considered in some studies (e.g. Okosun and Boltri, 2006; Oteng-Ntim et al, 2013), and contributions to the Global Burden of Disease study such as Ferrari et al 2014) use the estimator (1a) to derive attributable risks varying both by demographic group and over nations. Variations in attributable risks at subnational geographic scales, down to relatively small area scale, and the underlying methodological issues, have, however, been little explored. One approach (Tanuseputro et al, 2005) assumes confounder adjusted

relative risks based on national epidemiological surveys to be transferable across lower scale geographic settings. However, relative risks may vary across such geographic settings. Estimation of ARs from multilevel data, after adjustment for contextual risks, either measured neighbourhood confounders (e.g. area deprivation), or unmeasured spatially structured risk factors, has not been considered in previous studies. Allowance for spatial heterogeneity (e.g. effects of the focus risk varying by area type) is also not considered in existing studies.

The present paper is particularly concerned with the AR for diabetes in relation to excess weight. The association between elevated bodyweight and diabetes risk has been explored in many studies based on patient level data, either using categorical forms of the bodyweight predictor (Ganz et al, 2014; Field et al, 2001), or with linear regression of diabetes risk on BMI (Wong et al, 2014). However, a few multilevel studies have also considered spatial aspects of rising obesity prevalence or diabetes (e.g. Krokstad et al, 2013; Liu and Núñez, 2014), and a multilevel perspective on obesity is advocated by Huang et al (2009). Taking account of geographic context is important as an increasing number of studies link obesity (and hence diabetes) to environmental influences (Hill and Peters, 1998).

The present paper seeks to assess the potential importance of spatial effects (spatial heterogeneity, spatial clustering) in the estimation of context sensitive attributable risks, and their relevance in estimating area disease rates specific to particular risk factors, specifically small area diabetes rates attributable to excess weight. Spatial variation in the latter is particularly relevant for policy purposes. A subsidiary aim is to demonstrate the utility of a Bayesian approach to estimation using a logistic regression method in which ARs are based on a ratio estimator, while also selecting out significant influences on the disease outcome using Bayesian variable selection.

1.1 Attributable risks in a Multilevel Setting

The application in this paper considers estimation of ARs for diabetes prevalence in relation to excess weight, using multilevel data from health registers or health surveys. Oriented to such data, the paper considers adjustment for both patient confounders (e.g. other diseases, age, ethnicity), and observed and unobserved neighbourhood (contextual) confounders.

Multilevel data presupposes subjects nested within clusters, and observations for subjects within clusters areas may be correlated (Chen and Dey, 2003). As mentioned by Diez-Roux et al (1997) “correlation between individuals within neighborhoods .. may persist even after controlling for [observed] individual level and neighborhood level variables.” Existing multilevel disease risk models generally consider spatial effects in terms of (a) effects of observed area variables, and (b) randomly varying intercepts (typically assumed iid) over areas to represent unobserved area influences. Contextual effects are then assessed in

terms of the relative proportion of variation explained by areas (e.g. Pickett and Pearl, 2001; Merlo et al, 2006). Some analyses go beyond this to allow for spatial clustering in unmeasured neighbourhood influences on disease levels (Dasgupta et al, 2014; Xu, 2014; Chaix et al, 2005).

However, as discussed in Goodchild (2011), spatial effects encompass spatial heterogeneity as well as spatial clustering. There is an extensive literature on spatially varying regression relationships with both Bayesian approaches (Assunçao, 2003), and classical approaches often based on generalized weighted regression (Fotheringham et al 2003). This paper considers a relatively simple form of heterogeneity in regression effects, namely varying impacts of individual risk factors according to area type. In terms of the framework provided by Anselin (2010, p. 6) the form of heterogeneity considered here involves discrete heterogeneity, or spatial regimes.

Such heterogeneity can also be seen as a spatially defined form of effect modification or "hazard heterogeneity" (Ezzati et al, 2006, p. 245), applicable "when the assumption of constant relative risk [is] not appropriate". The potential importance of effect modification in estimating ARs is considered by Flegal et al (2004).

Specifically the analysis below accordingly considers estimation of ARs via multilevel models that admit the potential for (a) spatially correlated but unmeasured risk factors, and (b) neighbourhood group heterogeneity in impacts of bodyweight on diabetes. Regarding the first feature, and as discussed above, multilevel data presupposes subjects nested within clusters, and observations or residuals for subjects within clusters areas may be correlated (Chen and Dey, 2003). When areas constitute the clusters, residuals may show spatial correlation.

There is an extensive literature on modelling spatially correlated residual effects on health outcomes. Such spatial effects often proxy unobserved risk factors (e.g. environmental or cultural), which vary smoothly over space (Best, 1999). As mentioned by Wakefield et al (2000), modelling of spatially correlated errors, denoted v_j (j=1,..,J) for J areas, may proceed by initially specifying either the joint multivariate distribution of the vector $v=(v_1,..,v_J)$, or the univariate density of each areas error, v_j, conditional on errors in other areas. A widely adopted scheme known as the convolution prior, but with potential identification issues, involves an intrinsic autoregressive effect (Besag et al, 1991) combined with an iid (non-spatial) effect. Lee (2011) compares the properties of alternative conditional priors for spatial errors, and recommends instead the method of Leroux et al (1999), on the grounds of including a measure of spatial dependence, and in providing a rational form of conditional variance.

Regarding spatial heterogeneity, a focus here is on the potential interaction between area deprivation category and the effects of overweight, a cross-level interaction in the terminology of multilevel analysis. Possible mechanisms for such interaction are suggested by the large number of studies linking obesity (and diabetes itself) to environmental influences, such as access to healthy food and exercise opportunities (Hill and Peters, 1998; Feng et al, 2012; Salois, 2012). For example, obesity may be related to aspects of food environment (e.g. density of facility types, such as fast food outlets) which adversely influence diet, with less healthy food environments characterised by high consumption of processed food, high in fat and sugar (Lake and Townshend, 2006). Less healthy food environments tend to be in less affluent areas, that is areas with high deprivation (Morland et al, 2002). Exercise has independent effects on diabetes as well as through its effect on obesity (Kriska et al, 2003; De Feo et al, 2006), and exercise access is typically lower in deprived areas (Lamb et al, 2010). Effects on diabetes of diet-related and exercise-related obesity may therefore vary by neighbourhood deprivation category.

A logit regression methodology is adopted (see section 2), with the attributable ratio based on estimating diabetes risks under a reference setting (Traskin et al, 2013; Greenland and Drescher, 1993; Vander Hoorn et al, 2004). Instead of adopting a single binary threshold, the models used here involve a categorisation of bodyweight (Ganz et al, 2014), with three categories for excess weight: overweight (BMI between 25 and 29.99), obese class I (BMI between 30 and 34.99) and grossly obese (BMI over 35), with BMI under 25 as reference category. A Bayesian estimation and inference approach uses the WINBUGS software (Lunn et al, 2009), and Markov chain Monte Carlo (MCMC) estimation, with the main data analysis preceded by an initial analysis of missing data (missingness in BMI itself, and in an ethnicity confounder).

2. Methods

Let D denote presence or not of disease, X denote the focus risk factor (the exposure for which the AR is being developed), and C represent possible confounders in the relationship between X and D. Then for continuous X, one may represent the AR as (Traskin et al, 2013; Greenland and Drescher, 1993)

$$\int \left[1 - \frac{\Pr(D=1|X=x^*,C)}{\Pr(D=1|X=x,C)}\right] dF(x|D=1) \qquad (2)$$

where x^* represents the reference exposure, and $dF(x|D=1)$ is the distribution function of X among diseased subjects. For X in category form, the reference is the category where the exposure is absent.

For nested multilevel data, one may estimate (2) by applying binary regression (e.g. logit regression) with responses D_{ij} (=1 for cases, =0 for non-diseased) for subjects $i=1,..,n_j$ within neighbourhoods $j=1,...,J$, with $N=\sum_j n_j$. Observed predictors are the focus risk X_{ij}, patient confounders C_{ij} and neighbourhood confounders L_j. Let p_{ij} denote the predicted probability

under the observed scenario (X_{ij},C_{ij},L_j), and $T=\sum_{ij} p_{ij}$ denote the predicted total cases. Let p^*_{ij} denote the predicted probability under the reference (counterfactual) scenario (X^*_{ij},C_{ij},L_j), and the corresponding total predicted cases be noted $T^*=\sum_{ij} p^*_{ij}$.

Let ϕ denote the overall attributable risk. Then Greenland and Drescher (1993) propose the ratio estimator

$$\phi=1-T^*/T \qquad (3a)$$

as a confounder adjusted estimator of the AR. Estimators of the AR for subcategories g=1,..,G, such as particular ethnic groups, or area categories, may be obtained by summing only over subjects contained in each subgroup (Deubner et al, 1980). So if g_{ij} denotes the category for an individual subject, the relevant totals for category h using estimator (3a) are $T^*_h=\sum_{ij,g_{ij}=h} p^*_{ij}$, and $T_h=\sum_{ij,g_{ij}=h} p_{ij}$, with the category specific AR estimated as

$$\phi_h=1-T^*_h/T_h \qquad (3b)$$

One has $\phi=\sum_h w_h\phi_h$ with weights $w_h=T_h/T$.

Derivation of the standard errors of (3a)-(3b) in classical approaches may be quite complex, involving delta approximations (Graubard and Fears, 2005; Benichou, 2001), whereas using MCMC sampling the full posterior density of such quantities is readily obtained. Another benefit of a Bayesian approach is in the inclusion of predictor selection in the binary regression (see section 4 for details), leading to a form of model averaging. While choice of confounders and interactions reflects prior substantive knowledge, this does not preclude collinearities that may affect estimates (Fox and Monette, 1992; Hayashi et al, 2013; Ostchega et al, 2012) and reduce precision of structural quantities such as ARs.

2.1 Generic Regression Specification

The envisaged data here are multilevel in the broader sense that subjects i are nested with neighbourhoods j, which in turn may be nested in area typologies or neighbourhood groupings g=1,..,G (e.g. area deprivation quintiles, area socioeconomic classifications) (Joshy et al, 2009; Barnett et al, 2012). This scheme is analogous to that of Langford et al (1999) involving within-area (individual), group and neighbourhood effects.

Multilevel regression to predict the probability p_{ij} of diabetes status can then take account both of observed predictors (X_{ij},C_{ij},L_j), of neighbourhood groups $g=g_{ij}$, and of unmeasured neighbourhood risks v_j. For example, assume $D_{ij} \sim$ Bern(p_{ij}), and a logit link regression. For simplicity, assume BMI is in binary form (obese $X_{ij}=1$ or non-obese $X_{ij}=0$). Then an obesity effect, varying by neighbourhood category g, may be combined with impacts of observed patient and neighbourhood confounders, and unobserved neighbourhood risks v_j, as in

$$\text{logit}(p_{ij})=\beta_0+\gamma_g X_{ij}+\beta C_{ij}+\delta L_j+v_j \quad (4)$$

In practice, it may be necessary to assess whether all these features are required. With regard to the neighbourhood effects v_j, this might involve first fitting a reduced model without such terms, and assessing gain in fit on including such residuals, or assessing correlation patterns in the realized residuals from the reduced model. Additionally, predictor selection should be applied to achieve model parsimony and reduce imprecision caused by any multicollinearity between predictors. Also in practice, X_{ij} is taken as a categorical variable in the analysis below, as overweight (as well as obesity) enhances diabetes risk, and extreme obesity (e.g. BMI over 35) implies additional risk (Ganz et al, 2014; Field et al, 2001).

The term v_j is chosen to represent possible spatial structuring in cluster level residuals, which would be expected substantively when unobserved neighbourhood disease risk factors are spatially clustered. Following the recommendation of Lee (2011), one may assume the v_j follow Leroux et al (1999), since the data then determines the extent of estimated spatial dependence in unobserved neighbourhood risks. The conditional form of this prior is

$$P(v_j|v_{[j]},\kappa^2,\omega) \sim N(\omega \textstyle\sum_{k\sim j} v_k /d_j,\ \kappa^2/d_j), \quad (5a)$$

where $v_{[j]}=(v_1,..,v_{j-1},v_{j+1},..v_J)$, k ~ j denotes neighbourhoods k adjacent to neighbourhood j, ω is a measure of spatial dependence between 0 and 1, and

$$d_j=1-\omega+\omega M_j, \quad (5b)$$

where M_j is the number of areas adjacent to area j. If ω=0 this prior reduces to an iid density with variance κ^2, an advantageous feature mentioned by Lee (2011).

3. Case Study

The goal is to assess spatial variability in attributable risks for the relationship between bodyweight and the presence of type 2 diabetes. The observations are from a primary care register (observations are for the end of 2011) covering two north London boroughs (Havering, Barking and Dagenham), and the analysis focuses on n=82884 middle aged men aged 40-74 with type 2 diabetes diagnosed (D_{ij}=1) or not (D_{ij}=0). There are 6189 subjects with diagnosed diabetes. Individual confounders are hypertension, age group, and ethnicity (4 categories). Ethnicity-BMI interactions are included due to evidence of differential diabetes risk (e.g. due to varying insulin resistance) within normal BMI subjects of different ethnicity (Petersen et al, 2006). Age-weight interactions are indicated because diabetes prevalence and incidence continue to rise at older ages, despite average BMI tending to decline after age 60 (National Obesity Observatory, 2011).

Subjects are nested in J=258 neighbourhoods (lower level Super Output Areas or LSOAs); these can be aggregated to electoral wards, which are small areas (with both administrative and political status) often used in health profiling in the UK (see http://www.localhealth.org.uk). There are 35 wards in the case study area.

An important observed contextual confounder is area deprivation, based on the income domain score from the Indices of Deprivation 2010 (DCLG, 2011); prior evidence is for a positive diabetes gradient with ascending area deprivation (Maier et al, 2014; Cox et al, 2007). To allow for interactions between contextual and individual variables, the present study uses a categorical version of this predictor, namely the income deprivation quintile to which each neighbourhood is assigned. Additional contextual confounders are air quality and environmental greenspace. Potential relevance of air quality and greenspace to diabetes incidence has been demonstrated in recent studies (e.g. Brook et al, 2008; Astell-Burt et al, 2014). The air quality index is based on levels of nitrogen dioxide, particulates, sulphur dioxide and benzene from the UK National Air Quality Archive (Fairburn et al, 2008).

Among questions of interest are the extent of spatial variation in AR estimates, in particular according to neighbourhood deprivation quintiles, though smaller scale variation in ARs can also be considered (e.g. between the 258 neighbourhoods themselves, or an intermediate scale such as electoral wards). Straightforward application of equation (1) suggests an upward gradient in ARs as area deprivation rises. For example, the proportion of diabetics who are obese in the lowest income deprivation quintile (least deprived) neighbourhoods is 40%, compared to 48% of diabetics in the highest quintile.

3.1 Exposure Specific Area Prevalence

As mentioned above, the attributable risk is a measure of impact (Schoenbach and Rosamund, 2000), measuring how much of a disease can be attributed to a risk factor. Using attributable ratios one may ascertain the proportion of the total disease rate or burden attributable to a particular risk factor (Steenland & Armstrong, 2006). Adapting the terminology of Gefeller (1995), using spatial attributable ratios one may estimate exposure-specific prevalence rates by area. Geographic profiling of obesity-related diabetes is of particular relevance to health resource targeting and health promotion initiatives to tackle increasing levels of diabetes.

For concreteness in relation to the case study area hierarchy (and as pursued in the Results section), let ρ_h denote an age standardised prevalence rate (ASPR) for diabetes in electoral ward h (h=1,..,35), and ϕ_h the attributable risk for diabetes in relation to excess weight (overweight and obesity combined). Then the diabetes rate attributable to excess weight is

$$\nu_h = \rho_h \phi_h. \qquad (6)$$

One may also derive attributable numbers of diabetes cases due to excess weight (e.g. Price et al, 2012).

The ASPR may be estimated from another form of model (e.g. using area data only), or by aggregating at each MCMC iteration over the estimated probabilities p_{ij} according to the ward h that LSOA j is located within, and the age group a that subject ij belongs to. Let n_{ah} denote the number of subjects in ward h and quinquennial age group a. Then ward age specific disease rates are obtained as

$$\rho_{ah}=\sum_{ij\epsilon a,j\epsilon h} p_{ij}/\, n_{ah}. \qquad (7)$$

Direct age-adjustment may then be used to combine age specific rates into an overall ward level disease rate ρ_h, with weights w_a being the proportion of each age group in a standard population (Lilienfeld & Stolley, 1994). Here weights are from the European standard population applied to the quinquennial age bands 40-44, 45-49, ...,70-74. If estimation of ρ_h is carried out parallel to that of ϕ_h in this way, then obtaining credible intervals for exposure-specific prevalence rates ν_h is straightforward under MCMC approaches, but would involve complex delta approximations under classical estimation.

4. Regression Framework

The above discussion suggests, from a literature review, that spatial correlation in area residuals may occur for multilevel models applied to geographically nested data. We consider a baseline model (Model 1) without explicit spatial effects, and assess the pattern of realized residuals at area (LSOA) level. If spatial correlation is apparent, we will consider a model including spatial residual effects, namely v_j in (5).

Model 1 includes only observed subject and neighbourhood risk factors. Thus heterogeneity in the effects of overweight, obesity and gross obesity on diabetes risk is included, but not spatial residuals. The effect of BMI is represented using standard cut-points (Ganz et al, 2014). Thus define X_{ij1}=1 for observed BMI between 25 and 29.99, and X_{ij1}=0 otherwise; X_{ij2}=1 for observed BMI between 30 and 34.99 (obesity class I), and X_{ij2}=0 otherwise; and X_{ij3}=1 for BMI over 35 (obesity classes II and III), and X_{ij3}=0 otherwise. Effects of X_{ij1} to X_{ij3} vary between neighbourhood deprivation quintiles g=1,..,5. Individual confounders are diagnosed hypertension H_{ij} (binary), age group A_{ij} (categories: 1=40-49 (reference), 2=50-59, 3=60-69, and 4=70-74), and ethnic category E_{ij} (1=white (reference), 2=black, 3=south Asian, 4=mixed-other). Binary interaction effects are defined by overweight and obesity combined. Thus define $X_{ij}=\sum_{k=1,..,3} X_{ijk}$. Then interaction terms between age and body weight are AB_{ij}={I(A_{ij}=2, X_{ij}=1), I(A_{ij}=3, X_{ij}=1), I(A_{ij}=4, X_{ij}=1)}, where $I(W)=1$ if W is true and $I(W)=0$ if W is false. Interactions EB_{ij} between ethnic group and weight are defined similarly.

The respective parameters are α (intercept); fifteen fixed effects $\{\gamma_{1g}, \gamma_{2g}, \gamma_{3g}\}$ representing impacts on diabetes risk of overweight, obesity class I, and gross obesity, specific to area deprivation quintile g; a hypertension effect β_1; age group effects $\{\beta_2,\beta_3,\beta_4\}$; ethnic group effects $\{\beta_5,\beta_6,\beta_7\}$; parameters $\{\beta_8,\beta_9,\beta_{10}\}$ for age-weight interactions; parameters $\{\beta_{11},\beta_{12},\beta_{13}\}$ for ethnicity-weight interactions; and parameters $\{\delta_1, \delta_2, \delta_3\}$ for effects of known neighbourhood confounders, area income deprivation (L_1), air quality (L_2=1 if a neighbourhood has below average air quality, L_2=0 otherwise), and greenspace (L_3=1 if the percent greenspace in a neighbourhood exceeds 50%, L_3=0 otherwise).

Under this model there are 32 parameters $\{\beta_0,\ldots,\beta_{13}; \gamma_{11},\ldots,\gamma_{35}; \delta_1,\ldots,\delta_3\}$ with six of the β parameters, and the γ parameters, being for interactions. Reference probabilities p_{ij}^* for (3a) and (3b) are obtained by omitting the terms $\{\gamma_{1g}, \gamma_{2g}, \gamma_{3g}\}$ and also parameters $\{\beta_8,\ldots,\beta_{13}\}$ for the interactions (AB_{ij}, EB_{ij}) involving weight.

Predictor selection is included, as in principle it will assist in selecting a parsimonious model with enhanced precision for structural parameters such as ARs, a feature emphasized in some reviews (e.g. Benichou, 2001). Predictor selection is supported by an exploratory classical analysis, using the car package in R, which includes generalized variance inflation factors, applicable to binary regression (Fox and Monette, 1992; O'Brien, 2007). A classical logistic regression, replicating Model 1, showed generalized variance inflation factors exceeding 5 (but under 10) for some predictors. This feature was found to be associated particularly with the inclusion of interaction effects, AB_{ij} and EB_{ij}.

Model parsimony, and control for multicollinearity, is based on predictor selection using a version of the Lasso approach (e.g. Yuan and Lin, 2005; Park and Casella, 2008), though there are other approaches based on various forms of mixture prior for the regression coefficients. Predictor selection is applied to all fixed effect regression parameters in both models below, except the intercept, namely to the 31 regression parameters $\{\beta_1,\ldots,\beta_{13}; \gamma_{11},\ldots,\gamma_{35}; \delta_1,\ldots,\delta_3\}$. Let a particular regression parameter (i.e. a particular β,γ, or δ parameter) be denoted generically as ξ_k. Let $d_k \sim \text{Bern}(\pi_d)$ be binary indicators, with d_k=1 corresponding to inclusion of ξ_k, and d_k=0 corresponding to rejection of ξ_k. Under the case d_k=1, we have $\xi_k \sim \text{DE}(\lambda)$ where DE() denotes the double exponential (or Laplace) density, $p(u|\lambda)=\lambda\exp(-\lambda|u|)/2$ with variance $2/\lambda^2$. This density is expressed in terms of the absolute difference from the mean, and has fatter tails than the normal distribution. Under d_k=0, we have ξ_k=0. Then, following Yuan and Lin (2005), and assuming predictors in standardized form, the mixture prior for retaining or rejecting regression coefficients (and corresponding predictors) is

$p(\xi_k|d_k) = (1-d_k)\delta(0)+d_k DE(\lambda)$. (8)

We assign the priors $\lambda \sim U(0.001,10)$, corresponding to a relatively diffuse prior on the variance for ξ_k when d_k=1, and $\pi_d \sim$ Beta(1,1).

Model 1 has no cluster (area) effects, though for multilevel datasets, observations and residuals for subjects within clusters may be correlated (Chen and Dey, 2003; Diez-Roux et al 1997), and for geographically defined clusters, the correlation may be stronger for nearby areas. To assess the extent of spatial patterning of residuals under Model 1, the averages of the 82884 standardized residuals $(D_{ij}-p_{ij})/[p_{ij}(1-p_{ij})]^{0.5}$ are obtained according to LSOA j=1,..,258. Spatial correlation is assessed in the LSOA average residuals using a Moran's I statistic (Lawson, 2013, p. 91), with spatial interaction defined by adjacency.

If spatial clustering is apparent, an extended model including spatial effects may be considered. Thus model 2 replicates model 1 except in adding a spatial residual, with prior as in (5). Hence model 2 also has 31 fixed effect regression coefficients (β, γ and δ parameters) subject to variable selection, and 258 LSOA spatial random effects v_j, with two hyperparameters $\{\omega, \kappa^2\}$ governing their density. The spatial correlation ω is assigned a $U(0,1)$ prior, and the precision parameter $1/\kappa^2$ assigned a gamma prior with index 1 and scale 0.001.

5. Results

5.1 Estimation, Fit and Model Checks

As mentioned above, a preliminary stage to the main analysis is necessitated by missingness in the variables BMI (18% of values missing) and ethnicity (28% missing). A method allowing for non-random missingness in BMI values (Appendix 1) is used to generate the imputed datasets (Ibrahim et al, 2012). The main analysis then involves MCMC estimation applied to K=5 multiply imputed datasets (Little et al, 2014) containing imputed ethnicity and BMI where these values are missing. Total variances of parameter estimates then take account of within and between imputation variances. Inferences are based on the second halves of two chain runs of 10000 iterations for each of the five multiply imputed datasets, with convergence assessed using Brooks-Gelman-Rubin (BGR) statistics (Brooks and Gelman, 1998) (see Appendix 2 for discussion).

Model fit for the main analysis is assessed using the deviance information criterion (DIC) of Spiegelhalter et al (2002), the WAIC statistic (Watanabe, 2013), and a scoring rule (Gneiting and Raftery, 2007) appropriate to binary outcomes, namely the Brier score. Two posterior predictive checks are applied, based on predictions $D_{new,ij}$ sampled from the posterior predictive density. The first posterior predictive check uses the unweighted residual sum of squares (Hosmer et al, 1997; Copas, 1989), namely $R=\sum_{ij}(D_{ij}-p_{ij})^2$ and $R_{new}=\sum_{ij}(D_{new,ij}-$

$p_{ij})^2$. The posterior predictive p-value is estimated by the proportion of iterations where R_{new} exceeds R. Extreme tail p-values (under 0.05 or over 0.95) indicate model discrepancies (Berkhof et al, 2000). The second check is based on the Hosmer-Lemeshow group statistic H (Hosmer and Lemeshow, 2000). This compares observed and predicted number of diabetes cases (O_g, E_g) within g=1,..,4 patient groups: highly obese (BMI over 35), obese class I (BMI 30-34.99), overweight (BMI 25-29.99), and normal/underweight. Thus $H=\sum_g(O_g-E_g)^2/[N_g\pi_g(1-\pi_g)]$, where N_g and π_g are respectively the total number of subjects and average modelled diabetes probability in each group. As a posterior predictive check, H is compared with the analogous statistic H_{new}, obtained by accumulating predictions $D_{new,ij}$ to obtain totals $O_{g,new}$.

5.2 Model Results

Application of model 1 shows satisfactory predictive checks, in that none of the posterior predictive p-values for the models are in the tail region (Table 1). However, analysis of the standardized residuals $(D_{ij}-p_{ij})/[p_{ij}(1-p_{ij})]^{0.5}$ shows a spatial pattern. For example, for the first imputed dataset, the Moran's I statistic has mean (95% interval) of 0.48 (0.34, 0.67), whereas including spatial effects, as in (5), under model 2, Moran's I now has a 95% interval (-0.18, 0.53) straddling zero.

The advantage of model 2 over model 1 is also apparent for DIC and WAIC estimates, which improve for all imputations. Thus the DIC reductions for model 2 compared to model 1 are respectively -30.4, -24.1, -28.6, -33.2 and -27.8 under the five imputed datasets. It may be noted that DIC and WAIC estimates are very similar. The Brier score also shows improved fit under model 2. Model 2 also has satisfactory predictive checks.

To demonstrate spatial heterogeneity in AR estimates, Table 2 shows overall population-wide attributable ratios under model 2, and ARs for each area deprivation quintile (average posterior means and standard deviations over the K=5 complete datasets). The mean ARs by neighbourhood deprivation quintile are also plotted in Figure 1. Table 2 and Figure 1 show higher ARs in the two most deprived quintiles.

To demonstrate the factors underlying this gradient, Table 3 shows, for the first complete dataset, and by deprivation quintile, the proportion of diabetic subjects who are obese, either obese class 1 or highly obese (i.e. corresponding to a particular form of $P_{E|D}$ in equation 1(b)). Also shown are relative diabetes risks for obese subjects compared to normal weight subjects (with BMI between 18.5 and 24.99). These relative risks are based on comparing p_{ij} under model 2, within patient groups defined both by BMI category and neighbourhood deprivation quintile. The increase in relative risks and in $p_{E|D}$ for higher area deprivation levels is apparent.

Table 4 shows parameter summaries for model 2 in the form of pooled means and standard deviations of parameter estimates (log odds ratios) over the K=5 imputed datasets, as well as the percent relative efficiency of estimation with a finite K imputed datasets rather than an infinite number (Jamshidian, 2004). For regression coefficients, posterior retention probabilities $Pr(d_k=1|D)$, as defined in (8), are also shown. For such coefficients the parameter estimates are for the products $d_k\xi_k$, where ξ_k refers generically to the β, γ and δ coefficients. So on MCMC iterations where $d_k=0$ the regression parameter has value 0.

Table 4 shows posterior retention probabilities exceeding 0.975 for β_1 to β_7, which all have positive 95% intervals. Hence there is significantly elevated diabetes prevalence for hypertensive subjects; for older subjects; and for subjects with black and south Asian ethnicity.

From the viewpoint of establishing spatial effect modification, also highly significant are interactions γ_{2g} and γ_{3g} between area deprivation category g and extreme obesity (BMI over 35), and between area deprivation and obesity class I (BMI 30-34.99). These effects are significant for all deprivation categories but the effects are most pronounced for deprived quintiles 4 and 5. By contrast, effects γ_{1g} of overweight (BMI 25-29.99) in enhancing diabetes risk are only significant for more deprived areas (in quintiles 3, 4 and 5).

While interactions between area deprivation and weight are generally significant, the BMI-age interactions (with parameters β_8 to β_{10}) and the BMI-ethnic interaction (parameters β_{11} to β_{13}) have posterior retention probabilities under 0.5. For contextual variables, the δ coefficients show insignificant effects for air quality and greenspace, and also for the deprivation effect δ_1, with posterior retention probability of 0.45.

There is a considerable amount of model uncertainty. For example, the median probability model (Barbieri and Berger, 2004), defined by retaining only predictors with posterior retention probabilities above 0.5, has a selection probability of only 0.028. Similarly, while all ethnicity-weight interactions have retention probabilities below 0.5, models with all three interaction effects taking the value zero are selected with a probability of only 0.24. Regarding the predictor selection, a sensitivity analysis was undertaken with a gamma prior on λ^2, namely $\lambda^2 \sim Ga(1,0.001)$ (Park and Casella, 2008), and this produced very similar estimates for both λ and for retention rates.

Background analysis (not reproduced in detail here) involved a model replicating model 2, except in omitting interactions between deprivation and excess weight categories; so in this reduced model, the effects of overweight, obesity class I and extreme obesity are summarised in homogeneous coefficients γ_1, γ_2 and γ_3. Unlike the full model 2, this reduced model showed a significant δ_1 coefficient (for area deprivation) with retention probability 1. Hence the effect of deprivation on diabetes risk in model 2 seems to be

mediated very largely by the effect on diabetes of interaction between area deprivation and overweight/obesity status.

The spatial dependence parameter ω for model 2 has a mean of 0.70, showing spatially clustered variability in unobserved neighbourhood risk factors. To indicate where the spatial residual terms v_j have most relevance in representing unmeasured risk factors, Figure 2 plots out the posterior probabilities $Pr(v_j>0|D)$ of elevated spatial effects. There is a spatial pattern apparent in Figure 2. Thus underprediction of high prevalence by measured predictors, with $Pr(v_j>0|D)$ exceeding 0.8 in 37 neighbourhoods, is concentrated in the south west of the region. Overprediction of low prevalence by measured predictors, with $Pr(v_j>0|D)$ under 0.2 in 44 neighbourhoods, is concentrated in extreme NW and SE parts of the region. As discussed above, this patterning may represent unmeasured environmental risk factors relevant to variations in diabetes levels, such as exercise access and aspects of the food environment.

5.3 Small Area Attributable Risk Profiles

The attributable risk is a measure of impact, measuring the proportion of disease cases that can be attributed to a risk factor. As an example of a geographic profile in ARs relevant to health policy prioritisation, Figure 3 presents attributable risks from model 2 evaluated according to the electoral ward of each subject (subsequently denoted as wards). The 258 neighbourhoods are nested within 35 wards, with more deprived wards mostly located in the south west of the region; see also Table 5 which includes deprivation scores, estimated attributable risks, and estimated overweight-specific diabetes rates in each ward (Gefeller, 1995; Steenland & Armstrong, 2006).

The ARs at ward level range widely from 14% to 36%, with higher ARs generally found in more deprived wards, and so it is in such wards that interventions to reduce diabetes by strategies to prevent overweight and obesity would likely have more effect. Applying the ARs to estimated rates for total diabetes prevalence provides estimates of obesity-specific diabetes prevalence, as in (6). These vary more than three-fold, from 0.9% to 3.1%, with clear implications for targetting health promoting interventions.

6. Concluding Remarks

The attributable risk is a measure of impact, taking account both of the association between disease and exposure, and the proportion of subjects exposed. Spatial variation in attributable risks and in risk-specific disease rates reflects the impact of environmental risk factors on health, and is important for health agencies in prioritizing interventions (Narayan et al, 1999; Price et al, 2012). As mentioned by Diez-Roux et al (2008), strategies to prevent chronic disease need to focus not only on changing individual behaviors or treating risk factors, but also on modifying the environments that facilitate the development and maintenance of risk factors.

The present analysis seeks to explore methodological issues in estimating attributable risks under such a broader contextual perspective. The forms of data permitting such an approach are health surveys and disease registers containing geographically configured multilevel data, including patient risk factors, and data on known neighbourhood predictors. Some neighbourhood influences may be unobserved, and the analysis here confirms that spatial structuring of neighbourhood effects in models for multilevel health data should be considered, e.g. by analysing residuals from a baseline model without neighbourhood (cluster) effects (e.g. Chaix et al, 2005).

However, particularly relevant to estimating attributable risks from such data is the finding here of spatial heterogeneity in relative risks according to exposure, and in proportions exposed. Such heterogeneity leads on to spatial variation in attributable risk, and in exposure specific prevalence rates. Spatial heterogeneity is here considered in terms of discrete spatial regimes (Anselin, 2010), but in terms of applications to other multilevel disease datasets, there is scope to investigate continuous heterogeneity, for example via spatially dependent regression effects (Assunçao, 2003). While effect modification for different demographic categories is quite often considered in AR estimation, the present study suggests that spatial effect modification (e.g. according to area category) should be considered more routinely for diseases with a suspected environmental component. By contrast, the review by Uter and Pfahlberg (2001) showed environmental risk factors to be considered relatively infrequently in AR estimation.

One avenue for research is the extent to which estimates of ARs based on individual multilevel data (which may be computationally demanding for large populations) may be approximated by ecological analysis, with areas only. The above analysis has shown significant impacts of both age and ethnic group, and so expected disease cases for each area would ideally be based on the ethnic and age structure of each area's population. With expected events as an offset, one might then consider Poisson regression of area disease counts and an adaptation of the ratio estimator (3) to modelled disease counts by area. This would compare modelled counts based on actually observed neighbourhood risk factors (e.g. area obesity rates, area hypertension rates, area deprivation), with the counts predicted under a counterfactual model with (say) area obesity rates set to zero.

Another benefit of the application here is in demonstrating the utility of a Bayesian estimation strategy. Estimation using MCMC methods leads to straightforward credible intervals (and full posterior densities) for attributable risks, which are based on a ratio estimator when there is regression adjustment for confounders. By contrast, complex variance calculations are needed under classical approaches. The same considerations apply to exposure specific area disease rates (here overweight specific diabetes rates by area) which are a product of area disease rates and attributable risks.

A Bayesian estimation strategy also assists in allowing for model uncertainty via predictor selection, for example, among confounders and exposure-confounder interactions. In the

case study application to diabetes in London, collinearity was associated particularly with interactions between the exposure and confounders (e.g. interactions between excess weight and age group), and considerable model uncertainty was apparent. While classical approaches to variable selection in AR estimation include backward selection as in Stafford et al (2007), variable selection using a Bayesian approach is more flexible with a richer set of inferences, for example in terms of ranking predictors in importance through posterior probabilities of inclusion (Cui et al, 2010).

The analysis here is illustrative and can be extended to estimation of ARs for multiple risk factors, such as joint ARs for diabetes in relation to both body weight and physical exercise (Bruzzi et al, 1985). There is also scope for estimating ARs from regression schemes allowing for both direct and indirect effects of the focus risk factor. The particular application in the paper has been to two London boroughs, and has shown variation in estimated ARs between area deprivation categories. However, the multilevel regression approaches to AR estimation considered are relevant for diseases other than diabetes where environmental influences are increasingly recognized as relevant.

Appendix 1 Missing Data Imputation

BMI values are assumed to be lognormal with regression means based on age (categories 40-49, 50-59, 60-69, 70-74), ethnicity (categories white, black, south Asian, mixed-other), and hypertension diagnosis (binary). Ethnicity is assumed to be multinomial, and based on an additional dataset, the 2011 Census. Thus multinomial sampling is based on proportions in different ethnic groups that are specific to the subject's age band (40-49, 50-59, 60-69, 70-74), and to the middle level super output area (MSOA) level of the subject's residence (there are 53 such areas in the study zone). That is, sampling of the 4 ethnic categories is carried out within 212 strata defined by subject age and MSOA of residence. K=5 data imputations are made after convergence of the missingness model, at intervals 100 iterations apart.

To allow for possible informative missingness in BMI values, the missingness model likelihood includes binary indicators R=1 if BMI is observed and R=0 if BMI is missing, and a logit regression of Pr(R=1) on age group, ethnic category, hypertension status and log(BMI). The analysis in fact shows the probability of being observed is positively related to BMI. This is consistent with the relatively high proportions overweight (all BMI > 25) in the complete data, namely 67%, whereas other sources (http://www.noo.org.uk/LA/obesity_prev/adults) suggest a lower figure of 63-64%.

Suppose MCMC posterior means and variances of the K estimates of a parameter θ are $q_1,..,q_K$ and $V_1,..,V_K$ respectively, then within imputation variance of θ is estimated as

$W = \sum_{k=1}^{K} V_k/K$, between imputation variance as $B = \sum_{k=1}^{K} (q_k - \bar{q})^2/(K-1)$, and total variance of the pooled estimate $\bar{m}$ of θ as $T=B(1+1/K)+W$.

Appendix 2

The models being estimated include multiple effects (fixed, random) and so assessing convergence of MCMC sampling is important. Convergence was satisfactory under both models using two chains, using one chain with default initial parameter settings (e.g. as adopted in the examples in the Winbugs package), and the initial parameter values in the the other chain based on running an exploratory single chain run from the default initial parameter settings. The default settings are zero for fixed effects and ones for precisions.

Convergence is straightforward under all models (as judged by early attainment of BGR statistics approaching and essentially indistinguishable from 1) for fixed effect parameters $\{\alpha, \beta, \delta, \Gamma\}$ and overall regression probabilities p_{ij}. Plots of the evolution of the BGR statistic are available under the "bgr diag" tool in Winbugs, as described in the Inference Menu of the Winbugs User Manual (Lunn et al, 2009). Values of the BGR near 1 indicate convergence, with 1.1 considered acceptable by Gelman and Hill (2007).

To exemplify regression parameters, consider the varying effects $\{\gamma_{1g},\gamma_{2g},\gamma_{3g}\}$ of overweight, obesity class I and extreme obesity under model 2, and with the first imputed dataset. With a burn-in of 100 iterations, chain plots (Figure 4) indicate satisfactory mixing, and BGR statistics tend early to 1. MCMC convergence in hyperparameters for random effects is also sometimes an issue. We consider chain and BGR plots (Figures 5a, 5b) for the standard deviation κ of the spatial random effects v_j, and corresponding plots (Figures 5c, 5d) for the spatial correlation parameter ω. The trace plot for κ show some short term divergences in the sampling paths, but overall mixing is satisfactory, and BGR statistics as judged by plots obtained using the "bgr diag" tool in Winbugs show early convergence in both parameters (well before iteration 5000).

References

Anselin, L. Thirty years of spatial econometrics. Papers in Regional Science, 2010; 89(1): 3–25.

Assunçao, R. Space varying coefficient models for small area data. Environmetrics, 2003; 14(5): 453-473.

Astell-Burt T, Feng X, Kolt G. Is neighborhood green space associated with a lower risk of type 2 diabetes? Diabetes Care 2014; 37(1):197-201.

Barbieri, M, Berger, J. Optimal predictive model selection. Annals of Statistics, 2004; 32(3): 870-897.
Barnett K, Mercer S, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. The Lancet 2012; 380(9836):37-43.
Benichou J. A review of adjusted estimators of attributable risk. Stat Methods Med Res. 2001; 10(3):195-216
Berkhof J, van Mechelen I, Hoijtink H. Posterior predictive checks: principles and discussion. Computational Statistics 2000; 15(3), 337–354.
Besag J, York J, Mollie A. Bayesian image restoration, with two applications in spatial statistics. Annals of the Institute of Statistical Mathematics, 1991; 43:1--21.
Best, N. Bayesian ecological modelling, pp 193-201 in Disease Mapping and Risk Assessment for Public Health, Lawson, A., et al (eds). Wiley: New York, 1999
Brook R, Jerrett M, Brook J, Bard R, Finkelstein M. The relationship between diabetes mellitus and traffic-related air pollution. J Occup Environ Med. 2008; 50(1):32-8.
Brooks S, Gelman A. General methods for monitoring convergence of iterative simulations. J. Computational and Graphical Stat 1998; 7: 434-45
Bruzzi P, Green S, Byar D, Brinton L, Schairer C. Estimating the population attributable risk for multiple risk factors using case-control data. Am J Epidemiol. 1985; 122(5):904-14
Chaix B, Merlo J, Chauvin P (2005) Comparison of a spatial approach with the multilevel approach for investigating place effects on health: the example of healthcare utilisation in France. J Epidemiol Community Health. 59(6):517-26.
Chen, M-H, Dey, D (2003) Variable selection for multivariate logistic regression models. Journal of Statistical Planning and Inference, 2003; 111(1), 37-55.
Copas, J. Unweighted sum of squares test for proportions, Applied Statistics, 1989; 38: 71-80
Cox M, Boyle P, Davey PG, Feng Z, Morris A. Locality deprivation and Type 2 diabetes incidence: a local test of relative inequalities. Soc Sci Med. 2007; 65(9):1953-64.
Cui, G., Wong, M , Zhang, G. Bayesian variable selection for binary response models and direct marketing forecasting. Expert Systems with Applications, 2010; 37(12), 7656-7662.
Darrow L, Steenland N. Confounding and bias in the attributable fraction. Epidemiology 2011; 22(1):53-8.
Dasgupta P, Cramb S, Aitken J, Turrell G, Baade P. Comparing multilevel and Bayesian spatial random effects survival models to assess geographical inequalities in colorectal cancer survival: a case study. Int J Health Geogr. 2014 Oct 4;13:36.
De Feo P, Di Loreto C, Ranchelli A, Fatone C, Gambelunghe G, Lucidi P, Santeusanio F. Exercise and diabetes. Acta Biomed. 2006;77 Suppl 1:14-7.
Department for Communities and Local Government (DCLG) Indices of Deprivation 2010. London: DCLG; 2011
Deubner D, Wilkinson W, Helms M, Tyroler H, Hames C. Logistic model estimation of death attributable to risk factors for cardiovascular disease in Evans County, Georgia. Am J Epidemiol. 1980; 112(1):135-43.

Diez-Roux A, Nieto F, Muntaner C. Neighborhood environments and coronary heart disease: A multilevel analysis. American Journal of Epidemiology. 1997;146:48–63

Diez-Roux, A, Kershaw, K., Lisabeth, L. Neighborhoods and cardiovascular risk: beyond individual-level risk factors. Current Cardiovascular Risk Reports, 2008; 2(3): 175-180

Duncan C, Jones K, Moon G. Context, composition and heterogeneity: using multilevel models in health research. Soc Sci Med 1998; 46(1):97-117

Ezzati, M., Vander Hoorn, S., Lopez, A. D., Danaei, G., Rodgers, A., Mathers, C. et al. (2006) Comparative quantification of mortality and burden of disease attributable to selected risk factors. Chapter 4 in Global burden of Disease and Risk Factors, pp 241-396. Washington: World Bank.

Fairburn J, Butler B, Smith G. Environmental justice in South Yorkshire: locating social deprivation and poor environments using multiple indicators. Local Environment, 2008; 14(2) :139–154.

Feng J, Glass T., Curriero F., Stewart W., & Schwartz B. The built environment and obesity: a systematic review of the epidemiologic evidence. Health & Place 2012; 16 (2): 175–190.

Ferrari A, Norman RE, Freedman G, Baxter A, Pirkis J, Harris M, Page A, Carnahan E, Degenhardt L, Vos T, Whiteford H. The burden attributable to mental and substance use disorders as risk factors for suicide: findings from the Global Burden of Disease Study 2010. PLoS One. 2014 Apr 2;9(4):e91936

Field A, Coakley E, Must A, Spadano J, Laird N, Dietz W, Rimm E, Colditz G. Impact of overweight on the risk of developing common chronic diseases during a 10-year period. Arch Intern Med. 2001; 161(13):1581-6.

Flegal K, Graubard B, Williamson D. Methods of calculating deaths attributable to obesity. Am J Epidemiol. 2004;160(4):331-8.

Fotheringham A, Brunsdon, C, Charlton, M. Geographically Weighted Regression: the Analysis of Spatially Varying Relationships. Wiley: New York, 2003.

Fox J, Monette G. Generalized collinearity diagnostics. J Am Stat Assoc 1992; 87: 178–183

Friedman G. Body mass index and risk of death. Am J Epidemiol. 2014; 180(3):233-4.

Ganz M, Wintfeld N, Li Q, Alas V, Langer J, Hammer M. The association of body mass index with the risk of type 2 diabetes: a case-control study nested in an electronic health records system in the United States. Diabetol Metab Syndr 2014; 6(1):50.

Gefeller O. Definitions of attributable risk-revisited. Public Health Reviews, 1995; 23(4):343-355

Gelman A, Hill J. Data Analysis Using Regression and Multilevel/Hierarchical Models Cambridge: Cambridge University Press; 2007.

Gneiting T, Raftery A. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 2007; 102(477), 359-378.

Goodchild M. Challenges in geographical information science. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science 2011; 467(2133), 2431-2443.

Graubard B, Fears T. Standard errors for attributable risk for simple and complex sample designs. Biometrics, 2005; 61(3):847–855

Greenland S, Drescher K. Maximum likelihood estimation of the attributable fraction from logistic models. Biometrics 1993; 49(3):865-72.
Hayashi T, Boyko E, Sato K, McNeely M, Leonetti D, Kahn S, Fujimoto W. Patterns of insulin concentration during the OGTT predict the risk of type 2 diabetes in Japanese Americans. Diabetes Care. 2013;36(5):1229-35.
Hill J., Peters J. Environmental contributions to the obesity epidemic. Science, 1998; 280 (5368):1371–1374
Hosmer, D, Hosmer T, Le Cessie S, Lemeshow, S. A comparison of goodness-of-fit tests for the logistic regression model. Statistics in Medicine, 1997; 16(9): 965-980.
Hosmer D, Lemeshow, S. Applied Logistic Regression. New York: Wiley. 2000.
Huang T, Drewnosksi A, Kumanyika S, Glass T. A systems-oriented multilevel framework for addressing obesity in the 21st century. Prev Chronic Dis. 2009 6(3):A82, 1-10.
Ibrahim J, Chu H, Chen M. Missing data in clinical studies: issues and methods. J Clin Oncol 2012; 30(26):3297-303.
Jamshidian M. Strategies for Analysis of Incomplete Data. In M Hardy, A Bryman (eds) Handbook of Data Analysis (pp. 113-130), New York: Sage; 2004
Joshy G, Porter T, Le Lievre C, Lane J, Williams M, Lawrenson R. Prevalence of diabetes in New Zealand general practice: the influence of ethnicity and social deprivation. J Epidemiol Community Health 2009; 63(5):386-90.
Kriska A, Saremi A, Hanson R, Bennett P, Kobes S, Williams D, Knowler W. Physical activity, obesity, and the incidence of type 2 diabetes in a high-risk population. Am J Epidemiol. 2003;158:669-75.
Krokstad S, Ernstsen L, Sund E, Bjørngaard J, Langhammer A, Midthjell K, Holmen T, Holmen J, Thoen H, Westin S. Social and spatial patterns of obesity diffusion over three decades in a Norwegian county population: the HUNT Study. BMC Public Health. 2013 Oct 19;13:973.
Lake A, Townshend T. Obesogenic environments: exploring the built and food environments. J R Soc Promot Health 2006;126(6):262-7.
Lamb K, Ferguson N, Wang Y, Ogilvie D, Ellaway A. Distribution of physical activity facilities in Scotland by small area measures of deprivation and urbanicity. Int J Behav Nutr Phys Act. 2010; 18;7:76.
Langford I, Leyland A, Rasbash J, Goldstein H. Multilevel modelling of the geographical distributions of diseases. J R Stat Soc Ser C Appl Stat. 1999;48(2):253-68.
Lawson A. Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology, 2nd edition, 2013 CRC: Boca Raton.
Lee D. A comparison of conditional autoregressive models used in Bayesian disease mapping. Spatial and Spatio-temporal Epidemiology, 2011; 2, 79-89
Leroux B, Lei X, Breslow N. Estimation of disease rates in small areas: a new mixed model for spatial dependence. In: Halloran M, Berry D (eds) Statistical models in epidemiology, the environment and clinical trials (pp. 135–78). New York: Springer-Verlag; 1999
Lilienfeld, D, Stolley, P. Foundations of Epidemiology, 3rd Ed. Oxford University Press, 1994.

Little T, Jorgensen T, Lang K, Moore E. On the joys of missing data. J Pediatr Psychol. 2014; 39(2):151-62

Liu L, Núñez A. Multilevel and urban health modeling of risk factors for diabetes mellitus: a new insight into public health and preventive medicine. Adv Prev Med. 2014;2014:246049

Lunn D, Spiegelhalter D, Thomas A, Best, N. The BUGS project: Evolution, critique and future directions. Statistics in Medicine 2009; 28(25), 3049-3067

Maier W, Scheidt-Nave C, Holle R, Kroll L, Lampert T, Du Y et al. Area level deprivation is an independent determinant of prevalent type 2 diabetes and obesity at the national level in Germany. PLoS One 2014; 9(2):e89661. doi: 10.1371/journal.pone.0089661.

Meng X-L. Posterior predictive p-values. The Annals of Statistics 1994; 22: 1142-1160

Merlo J, Chaix B, Ohlsson H, Beckman A, Johnell K, Hjerpe P, Larsen K. A brief conceptual tutorial of multilevel analysis in social epidemiology: using measures of clustering in multilevel logistic regression to investigate contextual phenomena. Journal of Epidemiology and Community Health, 2006; 60(4): 290-297.

Morland K, Wing S, Diez Roux A, Poole C. Neighborhood characteristics associated with the location of food stores and food service places. Am J Prev Med 2002; 22(1):23–29.

Narayan K, Thompson T, Boyle P, Beckles G, Engelgau M, Vinicor F et al. The use of population attributable risk to estimate the impact of prevention and early detection of type 2 diabetes on population-wide mortality risk in US males. Health Care Manag Sci. 1999; 2(4):223-7.

National Obesity Observatory. Adult Weight. NOO Data Briefing; 2011. http://www.noo.org.uk.

O'Brien, R. A caution regarding rules of thumb for variance inflation factors. Quality & Quantity, 2007; 41(5), 673-690.

Okosun I, Boltri J. Racial/ethnic disparities in hypertension and diabetes ascribed to differences in obesity rate; pp. 53-72 in Racial and Ethnic Disparities in Health and Healthcare, ed E. Metrosa, Nova Publishers: New York, 2006

Ostchega Y, Hughes J, Terry A, Fakhouri T, Miller I. Abdominal obesity, body mass index, and hypertension in US adults: NHANES 2007-2010. Am J Hypertens. 2012; 25(12):1271-8.

Oteng-Ntim E, Kopeika J, Seed P, Wandiembe S, Doyle P (2013) Impact of obesity on pregnancy outcome in different ethnic groups: calculating attributable risks. PLoS One. 2013;8(1):e53749.

Park, T, Casella, G. The Bayesian lasso. Journal of the American Statistical Association, 2008; 103: 681-686.

Petersen K, Dufour S, Feng J, Befroy D, Dziura J, Dalla Man C et al. Increased prevalence of insulin resistance and nonalcoholic fatty liver disease in Asian-Indian men. Proc Natl Acad Sci 2006; 103 (48):18273-7.

Pickett K, Pearl M (2001) Multilevel analyses of neighbourhood socioeconomic context and health outcomes: a critical review. J Epidemiol Community Health. 55(2):111-22.

Price K, Plante C, Goudreau S, Boldo E, Perron S, Smargiassi A. Risk of childhood asthma prevalence attributable to residential proximity to major roads in Montreal, Canada. Can J Public Health. 2012;103(2):113-8.
Rockhill B, Weinberg C, Newman B. Attributable risk estimation for established breast cancer risk factors: considering the issues of high prevalence and unmodifiability. Am J Epidemiol 1998; 147(9):826-33.
Sairenchi T, Iso H, Irie F, Fukasawa N, Ota H, Muto T. Underweight as a predictor of diabetes in older adults: a large cohort study. Diabetes Care 2008; 31(3):583-4.
Salois, M. Obesity and diabetes, the built environment, and the local food economy in the United States. Economics & Human Biology 2012; 10(1): 35-42.
Schoenbach V, Rosamund W. Understanding the fundamentals of epidemiology: an evolving text. http://www.epidemiolog.net/evolving/RelatingRiskFactorstoHealth.pdf.
Spiegelhalter, D, Best, N, Carlin, B, Van Der Linde, A. Bayesian measures of model complexity and fit. J Royal Statistical Society B 2002; 64(4): 583-639.
Stafford R, Schluter P, Kirk M, Wilson A, Unicomb L, Ashbolt R, Gregory J. A multi-centre prospective case-control study of campylobacter infection in persons aged 5 years and older in Australia. Epidemiol Infect. 2007;135(6):978-88
Steenland K, Armstrong B. An overview of methods for calculating the burden of disease due to specific risk factors. Epidemiology 2006; 17(5):512-9.
Tanuseputro P, Manuel D, Schultz S, Johansen H, Mustard C. Improving attributable risk methods: examining smoking-attributable mortality for 87 geographic regions in Canada. Am J Epidemiol 2005; 161(8):787-98.
Traskin M, Wang W, Ten Have T, Small D. Efficient estimation of the attributable fraction when there are monotonicity constraints and interactions. Biostatistics 2013; 14(1):173-88.
Uter W, Pfahlberg A. The application of methods to quantify attributable risk in medical practice. Stat Methods Med Res. 2001; 10(3):231-7.
Vander Hoorn S, Ezzati M, Rodgers A, Lopez A, Murray C. Estimating attributable burden of disease from exposure and hazard data. In Ezzati M, Lopez I, Rodgers A, Murray C (eds) Comparative Quantification of Health Risks. Geneva: World Health Organization. 2004, pp 2129-40,
Wakefield, J, Best, N, Waller L. Bayesian approaches to disease mapping. In: Elliott P, Wakefield J, Best, N, Briggs D (eds), Spatial Epidemiology: Methods and Applications. Oxford: Oxford University Press. 2000; pp 104–27.
Wang W, Small D. A comparative study of parametric and nonparametric estimates of the attributable fraction for a semi-continuous exposure. Int J Biostat 2012; 8(1):32.
Watanabe, S. A widely applicable Bayesian information criterion. Journal of Machine Learning Research 2013; 14(1), 867–897.
Wong R, Chou C, Sinha S, Kamal A, Ahmed A. Ethnic disparities in the association of body mass index with the risk of hypertension and diabetes. J Community Health. 2014; 39(3):437-4

Xu H (2014) Comparing spatial and multilevel regression models for binary outcomes in neighborhood studies. Sociol Methodol, 44(1):229-272.
Yuan, M., Lin, Y. Efficient empirical Bayes variable selection and estimation in linear models. Journal of the American Statistical Association, 2005; 100, 1215:1225

Highlights

Considers issues in estimating attributable ratios for multilevel data for diseases subject to environmental risk

Finds 25% of diabetes is attributable to excess bodyweight, but with the attributable ratio varying almost three fold according to area deprivation level.

Finds considerable model uncertainty using a Bayesian variable selection approach.

Table 1 Model Fit and Check Statistics

	Model 1	Model 2
DIC	38965.9	38937.1
DIC (effective parameters)	29.1	91.6
WAIC	38966.2	38937.6
WAIC (effective parameters)	29.2	91.8
Brier Score (Mean) (x 100)	6.405	6.394
Brier Score 95% LL	6.397	6.383
Brier Score 95% UL	6.413	6.404
Posterior Predictive p-value (%)		
PPPV based on unweighted residual sum of squares	35.2	34.9
PPPV based on Hosmer-Lemeshow statistic	46.7	43.3

Table 2 Attributable Ratios, Diabetes and BMI, by Area Deprivation Quintile

	Mean	St devn	2.5%	97.5%
Overall	0.246	0.016	0.214	0.277
Deprivation Quintile 1	0.161	0.030	0.102	0.219
Deprivation Quintile 2	0.126	0.024	0.078	0.174
Deprivation Quintile 3	0.254	0.024	0.207	0.301
Deprivation Quintile 4	0.319	0.025	0.269	0.369
Deprivation Quintile 5	0.340	0.030	0.281	0.400

Table 3 Variations in Attributable Risk Components by Area Deprivation Quintile

	% of diabetic subjects who are obese	Diabetes relative risk (model 2), (with 95% intervals). Obese compared to normal weight subjects
Quintile 1	40	1.87 (1.86,1.88)
Quintile 2	38	1.77 (1.76,1.78)
Quintile 3	36	1.93 (1.92,1.94)
Quintile 4	47	2.57 (2.56,2.59)
Quintile 5	48	2.73 (2.72,2.75)

Table 4 Parameter Estimates for Model 2 (Pooled Estimates)

Parameter	Description	Mean	St devn	Relative Efficiency	Retention probability
γ_{11}	Overweight effect, deprivation quintile 1	0.030	0.067	0.991	0.462
γ_{12}	Overweight effect, deprivation quintile 2	-0.015	0.056	0.995	0.347
γ_{13}	Overweight effect, deprivation quintile 3	0.332	0.082	0.976	0.998
γ_{14}	Overweight effect, deprivation quintile 4	0.289	0.087	0.982	0.995
γ_{15}	Overweight effect, deprivation quintile 5	0.339	0.098	0.984	0.996
γ_{21}	Obesity class 1 effect, deprivation quintile 1	0.415	0.096	0.960	1
γ_{22}	Obesity class 1 effect, deprivation quintile 2	0.328	0.099	0.954	1
γ_{23}	Obesity class 1 effect, deprivation quintile 3	0.384	0.093	0.981	0.999
γ_{24}	Obesity class 1 effect, deprivation quintile 4	0.577	0.093	0.992	1
γ_{25}	Obesity class 1 effect, deprivation quintile 5	0.639	0.102	0.983	1
γ_{31}	Highly obese effect, deprivation quintile 1	0.709	0.113	0.970	1
γ_{32}	Highly obese effect, deprivation quintile 2	0.591	0.113	0.966	1
γ_{33}	Highly obese effect, deprivation quintile 3	0.872	0.109	0.968	1
γ_{34}	Highly obese effect, deprivation quintile 4	1.115	0.096	0.995	1
γ_{35}	Highly obese effect, deprivation quintile 5	1.156	0.106	0.999	1
β_1	Hypertension	1.204	0.029	0.999	1
β_2	Age 50-59	0.717	0.072	0.987	1
β_3	Age 60-69	1.071	0.068	0.980	1
β_4	Age 70-74	1.197	0.074	0.984	1
β_5	Black Ethnic	0.296	0.092	0.964	0.990
β_6	South Asian Ethnic	1.150	0.071	0.985	1
β_7	Other/Mixed Ethnic	0.234	0.078	0.938	0.979
β_8	Age 50-59 * all overweight (interaction)	0.050	0.078	0.980	0.460
β_9	Age 60-69 * all overweight (interaction)	0.035	0.071	0.963	0.391
β_{10}	Age 70-74 * all overweight (interaction)	0.004	0.071	0.976	0.383
β_{11}	Black *all overweight (interaction)	-0.008	0.083	0.989	0.440
β_{12}	South Asian * all overweight (interaction)	-0.014	0.070	0.981	0.413
β_{13}	Other/Mixed * all overweight (interaction)	0.030	0.074	0.966	0.400
Contextual		Mean	St devn	Relative Efficiency	Retention probability
δ_1	Deprivation	0.021	0.032	0.997	0.451
δ_2	Pollution	0.008	0.029	0.999	0.277
δ_3	GreenSpace	0.000	0.021	0.999	0.227
	Other Parameters				
		Mean	St devn	Relative Efficiency	
ω	Spatial dependence	0.700	0.205	0.998	
κ	St devn of spatial residuals	0.231	0.043	0.998	
λ	Lasso hyperparameter	1.461	0.336	0.995	

Table 5 Area Profile: Deprivation, Obesity-Related Diabetes ARs, Total and Obesity-specific Diabetes Prevalence (%)
Posterior Means and 95% Credible Intervals

Borough	Ward	Income deprivation score	Attributable risk (%)	Diabetes Prevalence (%)	Obesity-specific Diabetes Prevalence (%)
Barking & Dagenham	Abbey	0.267	24.1 (23.4,24.8)	9.57 (9.54,9.6)	2.2 (2.14,2.27)
	Alibon	0.278	34.9 (34.3,35.6)	8.5 (8.47,8.53)	2.84 (2.79,2.89)
	Becontree	0.250	31.7 (31.1,32.5)	9.09 (9.05,9.13)	2.75 (2.7,2.82)
	Chadwell Heath	0.271	28.8 (28,29.5)	8.81 (8.77,8.85)	2.41 (2.35,2.47)
	Eastbrook	0.185	24.9 (24.3,25.7)	7.51 (7.48,7.54)	1.79 (1.74,1.84)
	Eastbury	0.273	33.1 (32.5,33.7)	9.76 (9.72,9.8)	3.09 (3.04,3.14)
	Gascoigne	0.356	31.6 (31,32.3)	8.73 (8.69,8.76)	2.65 (2.59,2.7)
	Goresbrook	0.271	35.3 (34.8,35.9)	8.96 (8.92,8.99)	3.03 (2.97,3.08)
	Heath	0.311	34.3 (33.6,34.9)	8.8 (8.77,8.84)	2.87 (2.82,2.93)
	Longbridge	0.171	23.5 (22.8,24.2)	9.92 (9.88,9.96)	2.23 (2.17,2.3)
	Mayesbrook	0.288	35.8 (35.3,36.4)	9.15 (9.11,9.19)	3.14 (3.09,3.19)
	Parsloes	0.285	36 (35.3,36.7)	8.71 (8.68,8.75)	2.99 (2.93,3.05)
	River	0.246	32.5 (31.9,33)	8.88 (8.84,8.91)	2.75 (2.7,2.81)
	Thames	0.299	33.9 (33.2,34.6)	9.32 (9.28,9.35)	3.03 (2.97,3.09)
	Valence	0.279	34 (33.3,34.6)	9.21 (9.17,9.24)	2.98 (2.92,3.04)
	Village	0.277	33.1 (32.5,33.7)	8.84 (8.81,8.87)	2.79 (2.75,2.85)
	Whalebone	0.189	26.4 (25.7,27.1)	8.48 (8.44,8.52)	2.13 (2.07,2.19)
Havering	Brooklands	0.148	18.7 (18,19.3)	6.61 (6.59,6.64)	1.17 (1.14,1.21)
	Cranham	0.063	17.7 (17,18.4)	6.14 (6.12,6.16)	1.03 (0.99,1.07)
	Elm Park	0.134	18.7 (18.1,19.4)	6.69 (6.67,6.72)	1.19 (1.15,1.24)
	Emerson Park	0.059	14.4 (13.8,15.1)	6.58 (6.56,6.6)	0.9 (0.86,0.94)
	Gooshays	0.266	28.9 (28.3,29.4)	7.02 (7,7.04)	1.93 (1.89,1.97)
	Hacton	0.074	16.7 (16.1,17.4)	6.29 (6.27,6.31)	1 (0.97,1.04)
	Harold Wood	0.125	16.7 (16,17.3)	5.9 (5.89,5.93)	0.93 (0.9,0.97)
	Havering Park	0.172	20.8 (20.2,21.4)	7.06 (7.04,7.09)	1.4 (1.35,1.44)
	Heaton	0.243	27.5 (26.9,28)	7.5 (7.48,7.53)	1.96 (1.92,2)
	Hylands	0.083	17 (16.3,17.6)	6.45 (6.43,6.47)	1.04 (1,1.08)
	Mawneys	0.120	15.6 (14.9,16.3)	6.28 (6.26,6.3)	0.93 (0.89,0.97)
	Pettits	0.079	15.6 (14.9,16.2)	6.27 (6.26,6.29)	0.93 (0.89,0.97)
	Rainham-Wenningt	0.129	20.8 (20.2,21.5)	7.49 (7.46,7.51)	1.48 (1.44,1.53)
	Romford Town	0.122	20.2 (19.7,20.8)	6.45 (6.43,6.47)	1.24 (1.2,1.27)
	South Hornchurch	0.098	17.5 (16.9,18.2)	6.11 (6.09,6.13)	1.02 (0.98,1.05)
	Squirrel's Heath	0.161	22.3 (21.7,22.9)	7.55 (7.52,7.57)	1.6 (1.56,1.64)
	St Andrew's	0.077	15.8 (15.1,16.5)	5.81 (5.79,5.83)	0.87 (0.83,0.91)
	Upminster	0.042	15.9 (15.2,16.5)	5.99 (5.97,6.02)	0.9 (0.86,0.94)

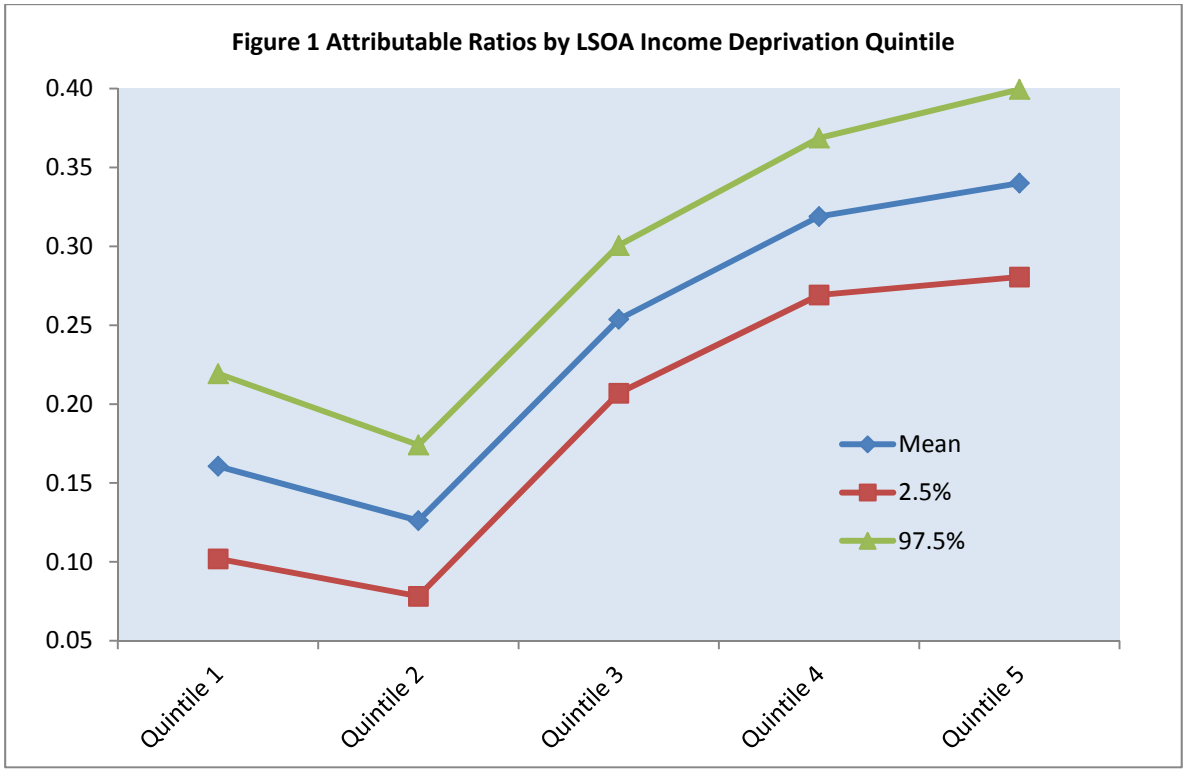
Figure 1 Attributable Ratios by LSOA Income Deprivation Quintile
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
Mean
2.5%
97.5%
Quintile 1
Quintile 2
Quintile 3
Quintile 4
Quintile 5

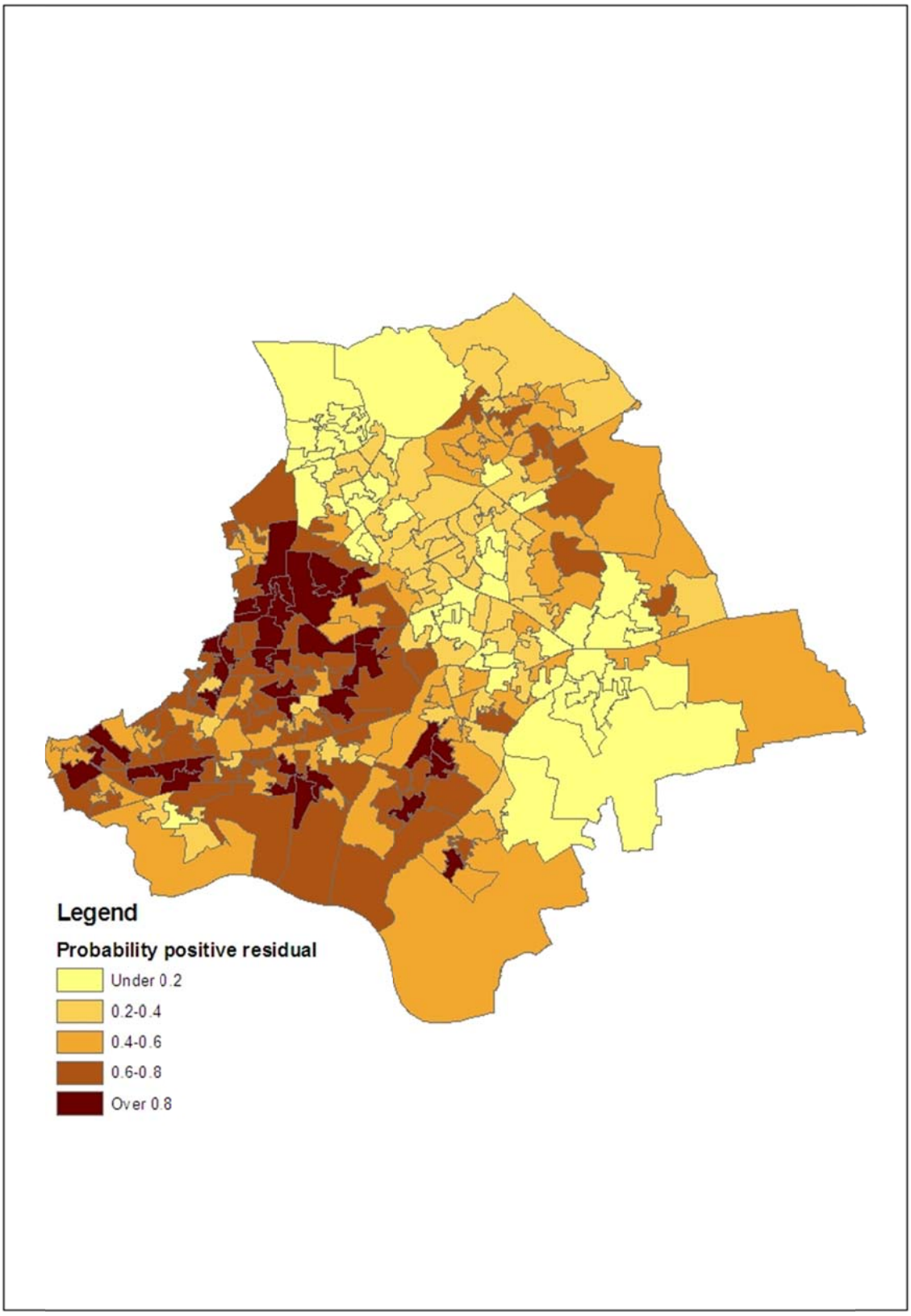

Figure 2 Posterior Probabilities that $v_j > 0$

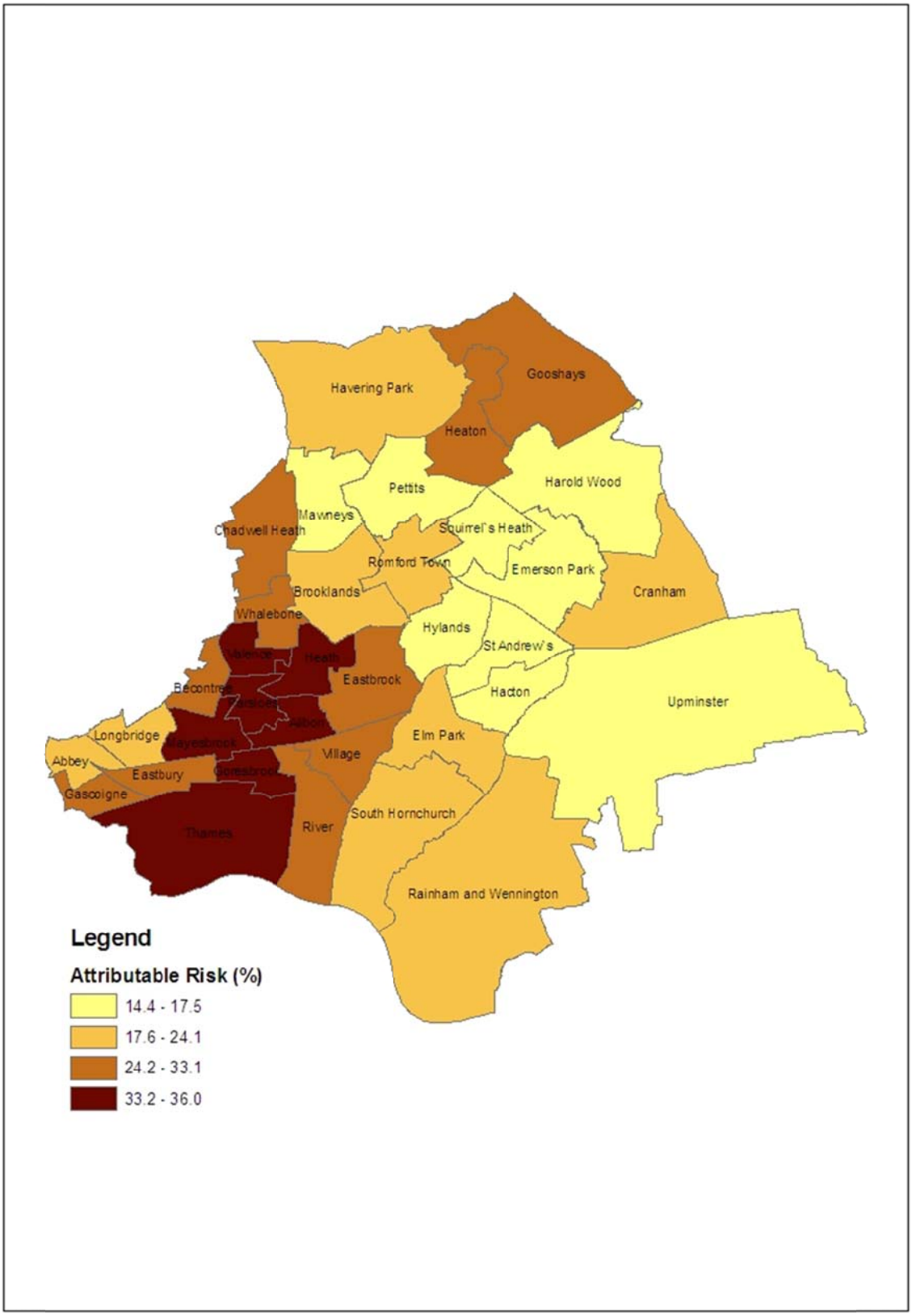

Figure 3 Attributable Risk Estimates for Electoral Wards

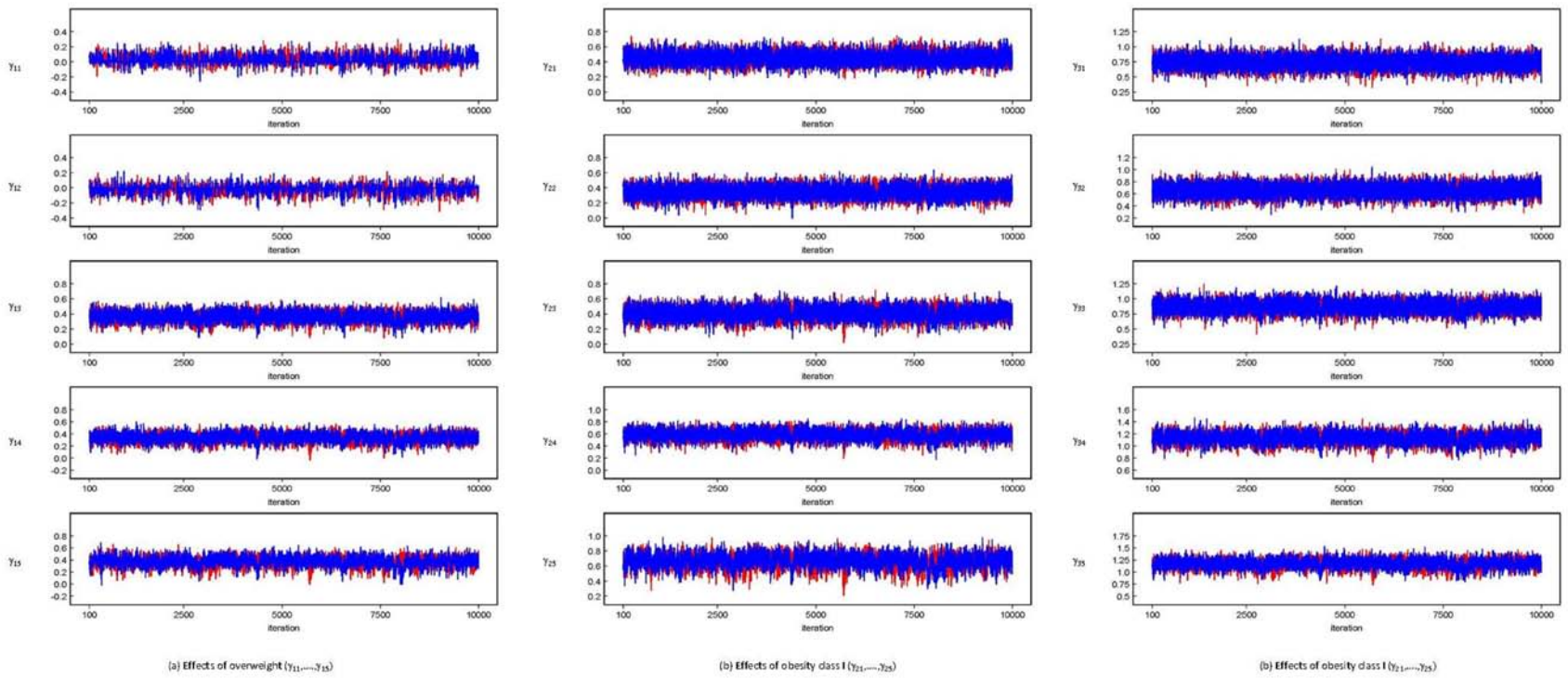

Figure 4 Trace Plots for Effects (γ_{1g},γ_{2g},γ_{3g}) of Overweight, Obesity Class I and Extreme Obesity

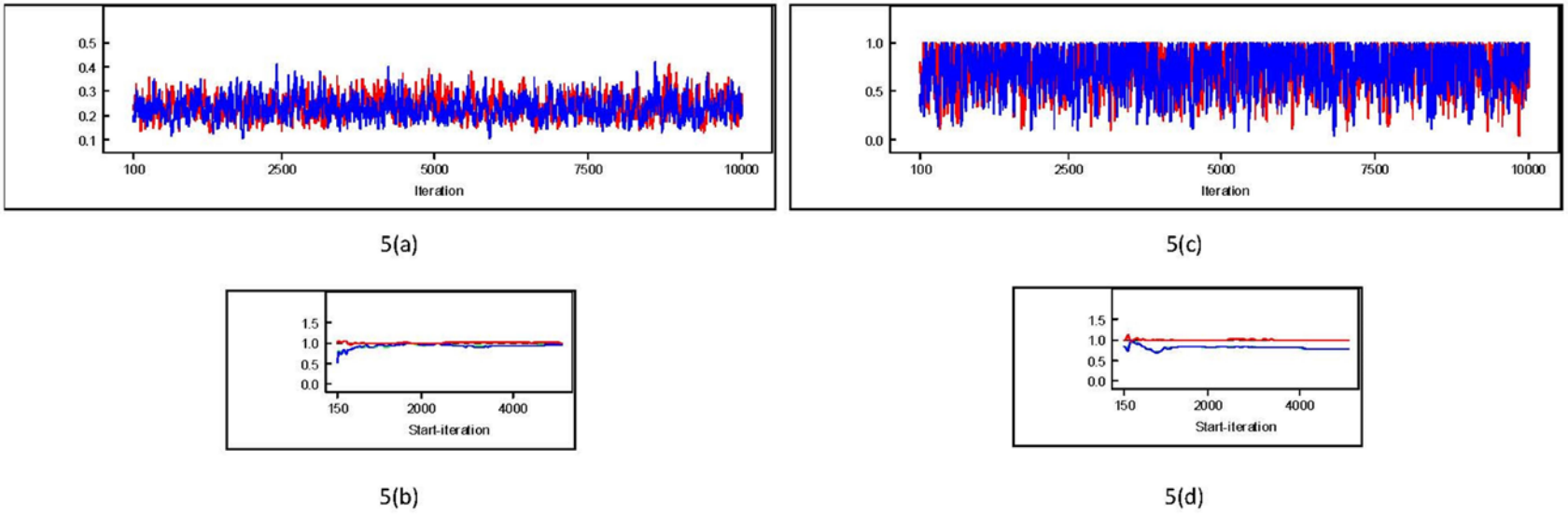

Figure 5 Chain plots, 5(a) and 5(c) for κ and ω respectively; BGR plots, 5(b) and 5(d) for κ and ω respectively