Quantifying the effect of area deprivation on child pedestrian casualties using longitudinal mixed models to adjust for confounding, interference, and spatial dependence

Daniel J. Graham†
Imperial College London, London, UK
Emma J. McCoy
Imperial College London, London, UK
David A. Stephens
McGill University, Montreal, Canada

Summary. This paper investigates the link between area based socioeconomic deprivation and the incidence of child pedestrian casualties (CPCs). The analysis is conducted using data for small spatial zones within major British cities over the period 2001 to 2007. Spatial longitudinal Generalised Linear Mixed Models (GLMMs), estimated using frequentist and Bayesian approaches, are used to address issues of confounding, spatial dependence, and transmission of deprivation effects across zones (i.e. interference). The results show a consistent strong deprivation effect across model specifications. The incidence of CPCs in the most deprived zones is typically greater than 10 times that of the least deprived zones. Modelling interference through a spatially autoregressive covariate uncovers a substantially larger effect.

Keywords: Deprivation, child pedestrian casualties, spatial model, confounding, interference.

1. Introduction

A positive association between between socio-economic deprivation and health has frequently been reported in the literature (e.g. Bajekal, 2005; Lancaster and Green, 2002; Cooper, 2002; Lorant et al., 2001; Jones et al., 2000; Carstairs and Morris, 1991; Townsend et al., 1988). Aside from general health outcomes, studies show effects on the incidence of injuries amongst children. For instance, Hippisley-Cox et al. (2002) find that children from the most deprived socioeconomic groups have a death rate from injury five times that of those from the least deprived social class, while Roberts and Powers (1996) provide evidence suggesting that a socioeconomic gradient is prevalent over the most common mechanisms of child injury.

The road traffic environment presents opportunity for injuries in the form of Child Pedestrian Casualties (CPCs). Over the period of this study, 2001 to 2007, there were 85,536 CPCs reported to the police force in Britain with considerable variation across the country

†*Address for correspondence*: Daniel J. Graham, Department of Civil Engineering, Imperial College London, London, SW7 2AZ, UK. Email: d.j.graham@imperial.ac.uk

in the incidence per unit area or per resident. Statistical evidence shows that children from the most deprived areas tend to have a higher probability of being involved in a pedestrian accident than those from the least deprived areas (for reviews see Christie, 1995; White et al., 2000; Graham and Stephens, 2008; Green et al., 2011). Christie (1995) suggests several reasons for this effect including higher exposure rates for deprived children (as fewer parents own cars), less adult supervision in the traffic environment, and educational disadvantage in understanding issues of road safety. There is also evidence that risk varies with area-based characteristics and that deprived children exhibit different behavioural patterns that increase their susceptibility to road traffic accidents.

Statistical analysis of the link between CPCs and deprivation have generally adopted one of two approaches. They have either used information on the socioeconomic status of victims to test for non-uniform incidence rates by background, or they have taken an area based approach and tested for association between spatial variation in CPC counts and levels of deprivation. While a positive association has generally been found using either type of approach, questions remain over the validity of this result due to difficulties experienced in addressing three key methodological issues: confounding, spatial dependence and interference in exposure to deprivation between zones. For instance, while the victim based studies have produced valuable evidence, they are inherently susceptible to confounding because we might expect deprived people to reside disproportionately within more hazardous traffic environments, for instance in busy inner city areas. The area based studies have attempted to address confounding by disentangling the effect of deprivation from that of other area based characteristics, but have generally assumed that the exposure of interest (i.e. some measure of deprivation) takes a known fixed value for each zone with no interference in exposure between zones. This may be a restrictive assumption since the deprived population travel, thereby giving rise to spillovers between areas. In addition, the area based studies have, to our knowledge, been exclusively cross-sectional in nature and consequently have encountered difficulties in adjusting adequately for spatial dependence and unobserved heterogeneity.

In this paper we present a longitudinal area based spatial analysis of the link between socioeconomic deprivation and CPCs in Britain largest cities. Our objective is to test the robustness of any statistical association to some commonly used modelling assumptions and in particular to address issues of confounding, spatial dependence, and transmission of deprivation effects across zones (i.e. interference). Using Generalised Linear Mixed Models (GLMMs) we introduce spatial covariates and draw on the longitudinal nature of the data to adjust for potential sources of confounding. We then address possible sources of model bias from spatial dependence and interference using spatial GLMMs with spatially autoregressive covariates within a Bayesian Conditional Autoregressive (BCA) framework.

The paper is structured as follows. Section 2 describes the methods and models used and explains the contribution that our analysis makes over existing studies. Section 3 describes the data and the techniques used to construct covariates. Model results are presented in section 4. The paper concludes with a summary of the main finding and a discussion of issues for future research.

2. Data analysis and models

Previous area based statistical work on the link between deprivation and child pedestrian casualties has to our knowledge been exclusively cross-sectional in nature (for reviews of

early work see Christie 1995; White et al. 2000; for recent empirical studies see Graham and Glaister 2003; Noland and Quddus 2004; Graham et al. 2005; Graham and Stephens 2008; Green et al. 2011). These studies have uncovered a positive association between deprivation and CPCs, but typically under fairly restrictive model assumptions. In particular the potential influences of confounding, spatial dependence, and interference in exposure between zones have received little attention. In this paper we adopt a mixed model approach for longitudinal data to address some of the limitation of the existing literature.

A generalised linear mixed model (GLMM) for a longitudinal data structure, comprising N units, $i = 1, ..., N$, each of which has n_i measures made over times t, $t = 1, ..., n_i$, giving a total of $n = \sum_{i=1}^{N} n_i$ sample observations, takes the form

$$h^{-1}\left\{\mathbb{E}(Y_{it}|\boldsymbol{X}_{it}, \boldsymbol{u}_i)\right\} = \boldsymbol{X}_{it}^{\mathsf{T}}\boldsymbol{\beta} + \boldsymbol{Z}_{it}^{\mathsf{T}}\boldsymbol{u}_i \tag{1}$$

where $h^{-1}(\cdot)$ is some known vector valued link function, $\boldsymbol{X}_{it}^{\mathsf{T}}\boldsymbol{\beta}$ is the fixed effects part of the model with design vector $\boldsymbol{X}_{it}$ and parameter vector $\boldsymbol{\beta}$, and $\boldsymbol{Z}_{it}$ is the design vector for the random effects $\boldsymbol{u_i} \sim \mathcal{N}(\boldsymbol{0}, G)$ with G being a valid covariance matrix. For the linear predictor $\eta_{it} = \boldsymbol{x}_{it}^{\mathsf{T}}\boldsymbol{\beta} + \boldsymbol{z}_{it}^{\mathsf{T}}\boldsymbol{u}_i$, the conditional expectation is $\mu_{it} = \mathbb{E}(Y_{it}|\eta_{it})$ and the conditional variance is $\mathrm{Var}(Y_{it}|\eta_{it}) = \phi V(\mu_{ij})$ where ϕ is a dispersion parameter and $V(\mu_{ij})$ is the variance function.

Our GLMM analysis addresses three key issues which are considered in turn below: confounding, spatial dependence, and spatial interference.

2.1. Confounding

The relationship between the incidence of CPCs and area deprivation is likely confounded in the sense that both the response (i.e. CPC count) and the exposure (i.e. deprivation) could depend on a set of pre-exposure characteristics. Under these conditions any observed association between response and exposure could be spurious in the sense that it could be attributed to other factors. While existing cross-sectional studies have sometimes acknowledged this potential problem, covariate adjustment in such models is typically based on a sub-set of the factors confounded with deprivation because some are difficult to measure or are unobserved.

To address confounding we include spatial covariates within design matrix $\boldsymbol{X}_{it}$ in addition to the exposure variable. These covariates are constructed to represent hypothesised sources of confounding which are described in full in the next section. We also specify multilevel binary variables within design matrix $\boldsymbol{X}_{it}$ or $\boldsymbol{Z}_{it}$, depending on whether these should be regarded as fixed or random, which differentiate the data according to different spatial categorisations and provide individual effects for each spatial unit. The inclusion of such effects achieves overdispersion in the model and accommodates the existence of unobserved heterogeneity. The justification for this approach is that some potentially important, but unobserved, characteristics of zones (or some aggregation of zones) will have changed very little, if at all, over the short time period of the data and can therefore be represented to a first-approximation by time-invariant effects.

The GLMM is used as a means of assessing the extent to which confounding is present and whether any observed deprivation effect is evident having adjusted for known confounders. Of course the inclusion of spatial covariates and multilevel effects cannot guarantee that no unmeasured confounders remain. This assumption, which is commonly made in empirical work, is generally regarded as untestable and is the key concern of the vast

literature on causal inference. One simple approach that can be used to test for non-spurious correlation is that of Granger causality (Granger, 1969), which is used widely in the econometrics literature and discussed recently in relation to causal inference by Eichler and Didelez (2010). We implement the Granger model here to provide an additional test of the 'causal' nature of any observed deprivation effect.

The original concept of Granger causality was developed for time series analysis within a vector autoregressive (VAR) framework to test whether past values of a covariate help to predict future values of a response, given the dynamic evolution of the response itself over time (e.g. Sims, 1980). By conditioning on a dynamically specified process the potential for spurious correlation is reduced. There are two key elements underpinning the concept of Granger causality: the existence of time ordering, such that an 'effect' must follow temporally from a 'cause'; and prediction, such that past values of the cause contain unique information which help predict future values of the effect.

For longitudinal data Holtz-Eakin et al. (1988) introduced the linear VAR model

$$Y_{it} = \alpha_i + \sum_{p=1}^{m} \theta_p Y_{i,t-p} + \sum_{p=1}^{m} \psi_p X_{i,t-p} + \eta_t + \varepsilon_{it} \tag{2}$$

where α_i represents unobserved individual time-invariant heterogeneity, $\varepsilon_{it} \sim IID\left(0, \sigma^2\right)$ is an error term, and η_t is a time specific effect that allows for unobserved shocks that are common across individuals. In this model we therefore condition not only on the lagged values of the response but also on any unit level effects which can be either random or fixed (and therefore arbitrarily correlated with the covariates). The variable X is then said to Granger cause Y if the parameters ψ_p are jointly different from zero.

Estimation of (2) is complicated by the fact that the unobserved individual effects (α_i) are correlated with the lagged response variables. To address this problem we use the dynamic Generalized Method of Moments (GMM) instrumental variables estimator for longitudinal data introduced by Arellano and Bond (1991) and extended by Arellano and Bover (1995). This approach specifies equation (2) in both levels and first-differences and uses the time series nature of the data to derive a set of instruments which are assumed correlated with the covariates but orthogonal to the errors. Specifically, lagged first-differences are used as instruments for equations in levels and lag levels as instruments for first-differenced equations. A set of moment conditions can then be defined and solved within a GMM framework to yield consistent estimates of the parameters in equation (2) (for details see Hall, 2005).

The key point about the longitudinal Granger model is that by conditioning on a dynamic process, with individual and temporal effects also included within an instrumental variables estimator, we can be reasonably confident that scope for spurious correlation from confounding has been substantially reduced. Consequently, if we find predictive power for changes in Y from past changes in X, having adjusted for the stochastic evolution of Y itself over time and for individual and temporal heterogeneity, we can take this as an indication that a significant conditional association exists.

2.2. Spatial dependence

The assumption of independence between zones in road traffic accident models is commonly made for convenience, but it may not hold if the covariate vector fails to adequately represent the spatial structure of the response. If spatial dependence is present then we can no longer

assume a typical GLMM structure in which the errors are iid given the covariates and random effects. The existence of spatial dependence can be difficult to detect and ignorance of the nature of the dependency means that it is also a tricky problem to treat. However, if left untreated, spatial autocorrelation can cause an increase in Type I error, give rise to a lack of precision in regression coefficients, and have a large impact on parameter estimates (see Cressie, 1993). Hewson (2005) has demonstrated the importance of accounting for spatial dependency in models for CPCs. Here we use a Bayesian approach to develop spatial GLMMs.

Our choice of spatial models is informed by Beale et al. (2010) who apply several methods for spatial dependency to simulated data sets for linear regression with Gaussian errors. They find that parametric spatially autoregressive models perform poorly as do those that seek to remove autocorrelation, while Bayesian Conditional Autoregressive (BCA) models are found to perform best. For our nonlinear count models the BCA approach offers distinct analytical and computational advantages. While nonlinear spatial models can also be specified within a GLMM framework, dependence between the spatial random effects induces a high-dimensional integral for the marginal density rendering maximum likelihood estimation (MLE) and restricted MLE (REML) in general no longer feasible (see for example Cressie, 1993; Diggle and Ribeiro, 2007). While Gaussian transformations of the response could be considered, the fact that the Bayesian approach allows spatial correlation structures to be specified directly for non-normal responses, and given the results of Beale et al. (2010), we adopt the BCA approach.

The fully hierarchical BCA model for spatial GLMMs with spatially correlated random effects was first proposed by Besag et al. (1991) and is described in full by Diggle et al. (1998) and Best et al. (1999). It has been used extensively in spatial accident research (e.g. Miaou et al., 2003; Song et al., 2006; Aguero-Valverde and Jovanis, 2008; Wang et al., 2011) and is available in the WinBUGS software using GeoBUGS (Thomas et al., 2004). For our application the CPC count response, Y_{it}, is modelled as a Poisson process

$$Y_{it}|\mu_{it} \overset{ind}{\sim} Poiss(\exp(\mu_{it}))$$

where μ_{it} is the log rate of the Poisson process for area i at time t depending on covariates and random effects as follows

$$\mu_{it} = \boldsymbol{x}_{it}^{\mathsf{T}}\boldsymbol{\beta} + s_i + v_i,$$

where $\boldsymbol{x}_{it}$ denotes a vector of spatial covariates, s_i is a spatially correlated random effect for zone i, and v_i is a random term to capture heterogeneity from extra-Poisson variation. Other hierarchical random effects can be added.

As in a frequentist GLMM a zero mean normal prior is typically specified for v_i

$$v_i \overset{ind}{\sim} \mathcal{N}(0, \tau_v^{-1})$$

where τ_v is the precision parameter. Spatial dependence in the residuals is modelled using a conditional autoregressive (CAR) normal prior for the s_i terms according to the Markov random field model

$$s_i|s_{j\neq i} \overset{ind}{\sim} \mathcal{N}\left(\frac{\sum_j s_j w_{ij}}{Q_{ii}}, \frac{\kappa}{Q_{ii}}\right)$$

where κ is a scale parameter, $w_{ij} = Q_{ii}h_{ij}$ is an element of a spatial weight matrix with $h_{ij} = -Q_{ij}/Q_{ii}$ $(i \neq j)$ being weights reflecting spatial dependence between the residual in

zone i and the residual in zone j. Typically, non-informative prior densities are specified for model parameters and hyper-parameters, including normal priors for the parameters of the linear predictor and gamma priors for τ_v and κ. The Bayesian formulation thus provides a fully specified probabilistic spatial model for accident counts. A key advantages of this approach is that it deals directly with sources of error and uncertainty, including those arising from ignorance of true correlation structures, which can propagate through model components.

2.3. Interference

The problem of spatial interference in exposure to deprivation presents a particularly difficult challenge which has not been addressed in the literature to date. Interference is said to occur when the outcome for some unit i does not depend only on the covariate vector for that unit but also on that of units j (Cox, 1958). It is important to stress that interference goes beyond the concept of statistical dependence induced by 'clustering', which may cause units to share similar characteristics that can have a bearing on outcomes but that do not imply inter-unit transmission of covariate effects (e.g. Rosenbaum, 2007).

With regard to exposure to deprivation, the problem we face is that the deprivation characteristics of unit j could affect the outcome of unit i because of travel between zones. Cox (1958) argues that if interference is present, and it is not possible to modify the data generating process, then the overlap between treatment effects should be accepted and incorporated into the analysis. This can be done either through explicit modelling of the interactions between units or by re-defining the units of interest such that they individually internalise any interactions, for instance, through aggregation. Since aggregation is problematic in that it dilutes the accuracy of measures of deprivation and sacrifices precision in parameter estimation, and since there are well established methods of modelling trip generation between zones, we attempt to explicitly model interactions within spatially autoregressive constructions of the deprivation covariate.

First, we simply calculate an exposure variable which is an average value taken across 'proximate' zones from the centroid of zone i. This in effect assumes an autoregressive process corresponding to a Markov random field. To define proximate zones; that is, those that are likely to have significant traffic interactions with zone i; we choose a distance threshold around i. We define proximate zones j, $j = 1, ..., k$, as those captured within catchment area $\mathcal{C}$ extending five kilometres from the centroid of zone i. Denoting the deprived population in unit i at time t as D_{it}, and the total working population by P_{it}, then the quantity

$$\frac{(D_{it}/P_{it}) + \sum_{j=1}^{k \in \mathcal{C}} (D_{jt}/P_{jt})}{(k+1)}.$$

gives a spatially averaged measure of exposure to deprivation which we refer to as the *average inter-zonal deprivation rate*.

Second, drawing on concepts developed by Graham and Megueulle (2006) to model traffic flows, we construct a more general spatially autoregressive model of potential interactions between zones based on an inverse distance weighted 'gravity' approach to trip generation. The gravity model is commonly used to satisfactory effect in modelling interactions between zones for the purposes of transport engineering and planning. This approach characterises the volume of interactions between zones F as proportionate to the 'activity' or 'mass' M at those zones and inversely proportionate to the distance between

them: $F_{ij} = (M_i^{\alpha} \times M_j^{\beta})/d_{ij}^{\gamma}$. There is considerable empirical evidence in support of this approximation, as demonstrated in recent results for the UK by Graham and Melo (2011) and reviewed more generally by El-Geneidy and Levinson (2006). Our gravity approach calculates three components of exposure to deprivation in any zone i.

(a) Exposure density originating in zone i - this is simply measured by the deprived population resident in zone i at time t given the size of the zone.

$$EO_{it} = \frac{D_{it}}{r_{it}}$$

where r_{it} is an approximation to the radius of zone i calculated by assuming that its area is circular. The normalisation of exposure by radius seems intuitively reasonable since it proxies for the density of deprived people active in the traffic environment.

(b) Exposure from trips destinating in zone i - exposure trips destinating in zone i are assumed to depend on the deprived population in proximate zones, the distance between zone i and all other proximate zones, and the relative attractiveness of zone i. Destinating exposure trips in zone i at time t are then proxied by

$$ED_{it} = \sum_{j(i \neq j)}^{k \in \mathcal{C}} \left(\frac{D_{jt}}{d_{ij}} \right) \cdot \left[(E_{it} + P_{it}) / \sum_{j(i=j)}^{k \in \mathcal{C}} (E_{jt} + P_{jt}) \right],$$

where the first term on the RHS is a distance discounted sum of total deprived residents in proximate zones, and the second term represents the attractiveness of zone i to trips measured in terms of the relative proportion of 'activity' (i.e. population and economic activity) contained within zone i.

(c) Exposure trips intersecting zone i - to represent movements that do not originate or destinate in zone i, but which may introduce some exposure in i, we use the following proxy variable

$$EI_{it} = \sum_{j(i \neq j)}^{k \in \mathcal{C}} \left\{ \left(\frac{D_{jt}}{d_{ij}} \right) \cdot \left[\sum_{j(i \neq j)}^{k \in \mathcal{C}} (E_{jt} + P_{jt}) / \sum_{j(i=j)}^{k \in \mathcal{C}} (E_{jt} + P_{jt}) \right] \cdot \frac{2 \arctan (r_i / d_{ij})}{2\pi} \right\}.$$

The first term captures the total deprived population resident in proximate zones j with interactions decaying with distance to zone i, the second term represents the relative attractiveness of destination zones other than i within $\mathcal{C}$, and the third term distributes the proportion of trips from j that go in the direction of i depending on the angle of the cone that delimits the spatial extent of zone i from zone j (see figure 1 below). The underlying assumptions on direction is that trips that originate in proximate zones j, but do not destinate in zone i, are distributed uniformly around j within catchment area $\mathcal{C}$ and so the angle θ_{ij} is used to approximate the proportion of trips from j that could potentially cross zone i.

Since EO_{it}, ED_{it} and EI_{it} are effectively measured in the same units, exposure per unit of distance, they can be summed to form a single exposure variable incorporating spatial interactions which we refer to as the *'gravity' weighted inter-zonal deprivation rate*.

Incorporating these two spatially autoregressive exposure covariates in the GLMMs we seek to test whether any statistical association between CPCs and deprivation is robust to

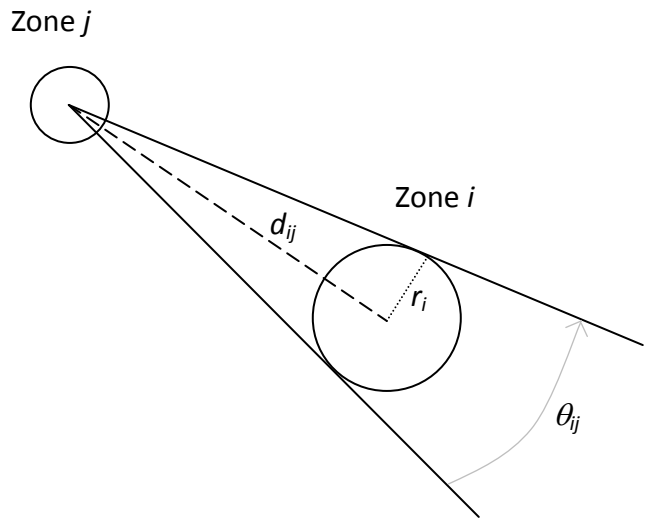

Fig. 1. Geometric allocation of trips from zone j passing through zone i.

assumptions about interference, or lack thereof. We also study any change in observed spatial dependence in the random effects components of the GLMMs to see whether estimates of spatial dependence are influenced by the form of spatial interaction assumed.

3. Data

Our study is based on a longitudinal analysis of CPCs in small zones of Britain's largest urban areas over the period 2001 to 2007. The cities covered include London and the major conurbations of Greater Manchester, Merseyside, South Yorkshire, Tyne and Wear, West Midlands, West Yorkshire and Strathclyde. The data available for analysis and the logic used to construct spatial covariates are described below.

3.0.1. Response variable

The response variable comprises annual counts of CPCs for small spatial zones of Britain, based on ward and intermediate zone classifications, over the period 2001 to 2007. The casualty data are taken from records completed by police officers each time an incident is reported to them. The individual police records are collated and processed by the UK Department for Transport (DfT) as "Road Accident Data - GB", generally known as the STATS 19 data. The data record the age of the victim and we define children as persons under the age of 16.

Using a Geographical Information System (GIS), and coordinate information from STATS 19, we allocated each CPC to one of 10,085 geographic zones of Britain. The spatial framework we use for analysis is the most disaggregate level for which the necessary data exist to construct the spatial covariates and exposure variable. For England and Wales the Census

Area Statistical (CAS) wards are used, of which there are 8,850 in total. CAS wards are also defined for Scotland, however, since the Scottish Government has taken over responsibility for data collection they have preferred to publish data using a different geographic disaggregation. The Scottish geography most similar to that of CAS wards, for which the relevant data are available, is the intermediate zone (izone) classification, which divides Scotland into 1,235 zones, comparable in size to the 1,222 Scottish CAS wards. We then extract data only for London and the conurbations of Britain using an official classification of zones provided by the Department for Transport. This yields 1,820 zones for analysis with 7 years of data giving 12,740 observations.

In total there were 38,497 CPCs over the period of observation with a mean incidence per zone of 3.082 and a a maximum of 36. The total number of CPCs has fallen consecutively each year: 7327 (2001), 6336 (2002), 5842 (2003), 5353 (2004), 4976 (2005), 4447 (2006), 4213 (2007).

3.0.2. Exposure

Our exposure variable is constructed using Department for Work and Pensions Longitudinal Study (DWPLS) data which record the number of working age people claiming some form of benefit from the state. We calculate the exposure variable in rate form (i.e. claimants per working age resident) to represent poverty based *deprivation rate* for each zone. Previous area based studies of health and socio-economic characteristics tend to use official measures of deprivation, such as indices of multiple deprivation, as the exposure variable (e.g. Graham and Stephens, 2008; Lancaster and Green, 2002; Cooper, 2002; Lorant et al., 2001; Carstairs and Morris, 1991; Townsend et al., 1988). A key limitation of these indices is that they are available only at infrequent intervals. As Carstairs (2000) points out, the availability of small-area data on a continuing basis offers alternative sources to represent socio-economic deprivation in the longitudinal setting. Regressing our measure of deprivation on the latest official indices of multiple deprivation and income deprivation we find strong correlations, with R^2 values above 0.9. We are therefore confident that the claimant count based measure provides a good representation of spatial variance in poverty or deprivation.

3.0.3. Constructing spatial covariates to address confounding

In observational data the effect of deprivation may be confounded. Specifically, we hypothesise the following sources of confounding:

i. *Child population* - deprived families may tend to have more children creating a larger supply of victims. To adjust for this effects we include a measure of *zone child population*.

ii. *Traffic generation potential and nature of the urbanised environment* - deprived zones may tend to experience greater volumes of traffic. Traffic flows are not observed at the zone level. Instead, we use a form of 'gravity' trip generation model to represent potential traffic flows. This approach, which was introduced in the previous section, characterises the volume of traffic between zones as proportionate to the 'activity' or 'mass' at those zones and inversely proportionate to the distance between them, and has been shown to provides a good fit to small area traffic data for the UK (e.g. Graham and Melo, 2011). Using employment (E_i) and population (P_i) to represent

the mass at each zone we proxy for traffic generation potential using the following four covariates:

(i) *Zone employment density* - $ZED_i = E_i/r_i$,
(ii) *Zone population density* - $ZPD_i = P_i/r_i$,
(iii) *Proximate employment density* - $PED_i = \sum_j E_j/d_{ij}$,
(iv) *Proximate population density* - $PPD_i = \sum_j P_j/d_{ij}$,

where r_i is the radius of zone i approximated from zone area assuming the zone is circular, and d_{ij} is the distance between the centroids of zone i and j.
These variables proxy for traffic flows, but in a very general way, by simply capturing the volume of 'activity' within and around zones. As such, they also represent variation in the nature of the built environment since 'mass' has been split to measure the relative composition and density of population and economic activity. This is a potentially useful feature since we might expect road safety and zone deprivation to be associated with the nature of land use and the degree of urbanisation. For instance, children living in suburban residential environments may be more affluent with less exposure to traffic risk than those living in dense inner city mixed use locations. Previous research indicates that such covariates can adequately represent relevant characteristics of the built environment (e.g. Graham and Glaister, 2003).

iii. *Scale of the road network* - high capacity networks tend to depress land values which in turn will influence the socio-economic profile of the people that live in close proximity. Using GIS software we generated longitudinal data on *road length* for each zone including a breakdown by road type: A-road, B-road, minor road, and motorway. Network data for the year 2005 are not available.

iv. *Road network density* - with the available GIS data we were also able to represent the *road network density* in each zone using a measure of the number of network nodes per unit of area. A network node is defined as the meeting point of two or more links. Deprived zones may tend to have more extensive dense networks.

v. *Time* - in a longitudinal study, time itself may be a confounder because government policies or other interventions could simultaneously affect deprivation and road accidents. Fuel taxation or motoring policies, for instance, provide relevant examples. We specify time as a factor allowing for 'shocks' that are common across units in each *year*.

Most of the confounding covariates are highly persistent over time. ANOVA results show that the vast majority of the variance in our time-varying covariates (typically over 98%) comprises between unit variance.

In addition to the time-varying confounders we make use of random or fixed intercepts, specified at different spatial levels, to represent unobserved time-invariant effects. The justification for this approach is that some potentially important confounders; for instance network engineering / design, speeds, the availability of open recreational space, and the physical / climatic characteristics of zones; will have changed very little, if at all, over the short time period of the data. We specify individual effects at three levels to give the following additional variables:

vi. *Zone (unit) level individual effects.*

vii. *Area type effects* - we construct binary variables for *area types* according to a five way categorisation of wards in relation to the level of urbanisation: Central London, Inner London, Outer London, Inner Conurbation, and Outer Conurbation .

viii. Region specific effects - we construct binary variables for *regions* of Britain corresponding to the following broad geographical areas: North, North West, Scotland, London, West Midlands, Yorkshire & Humberside.

The unit level effects adjust for the detailed characteristics of each locality, the area effects for unobserved differences between broadly similar area types, and the regional effects for general differences in climate, daylight and other relevant characteristics.

3.0.4. Spatially autoregressive exposure covariates

As outlined in section 2, we attempt to address interference in exposure to deprivation through the construction of two spatially autoregressive covariates that incorporate assumed interactions between zones. One is based on a Markov random field average inter-zonal deprivation rate within distance bands of 5 kilometres from the centroid of each zone. The second measure calculates a 'gravity' weighted inter-zonal deprivation rate to model trip interactions between zones. C++ programs were written to compute these covariates using spatial information derived from GIS analysis.

3.0.5. Spatial weight matrices

Spatial weight matrices are generated to model spatial dependence. Since we cannot observe the true nature of the underlying autoregressive error processes, an element of subjectivity is necessarily involved in the choice of such matrices. For the BCA models we use a contiguity matrix of 'queen' form in which units that share a common border or a common single point are treated as neighbours. The contiguity matrix is created using the `maptools` and `spdep` packages in `R`.

4. Results

The results are organised in three sub-sections: GLMMs and Granger Causality for the analysis of confounding, spatial models to adjust for dependence, and analysis of interference in exposure between zones.

4.1. GLMMs and Granger Causality

Table 1 below shows results from Generalized Linear Models (GLMs) and from a GLMM. For the GLMs, we found that the BIC values supported selection of the Negative Binomial rather than Poisson link function indicating that the data are not equidispersed. A log transformation of the covariates and the exposure variable was also found to improve the BIC relative to absolute values.

Table 1. Negative binomial GLMs and Poisson GLMM.

	Negative Binomial GLM		Negative Binomial GLM		Poisson GLMM	
	Est.	S.E.	Est.	S.E.	Est.	S.E.
(Intercept)	-7.631	0.119	-10.453	0.273	-11.326	0.726
log deprivation rate	0.608	0.015	0.600	0.018	0.583	0.026
log child population	0.941	0.014	0.539	0.042	0.500	0.060
log zone employment density			0.032	0.006	0.034	0.008
log proximate employment density			0.102	0.071	0.130	0.110
log zone population density			0.207	0.050	0.284	0.077
log proximate population density			-0.155	0.089	0.040	0.170
log road length			0.507	0.026	0.487	0.042
log road network density			0.187	0.024	0.128	0.037
year 2002	-0.146	0.022	-0.153	0.021	-0.146	0.017
year 2003	-0.188	0.023	-0.197	0.021	-0.194	0.018
year 2004	-0.266	0.023	-0.277	0.022	-0.277	0.018
year 2006	-0.434	0.024	-0.454	0.023	-0.454	0.019
year 2007	-0.464	0.024	-0.488	0.023	-0.492	0.020
theta	6.685	0.309	8.988	0.497	1.000	
BIC	43256		42393		14429	
n	10920		10920		10920	

The models shown in table 1 adjust in different ways for potential sources of confounding and unobserved heterogeneity. Columns 2 and 3 show results from a 'naive' model in which the CPC count is regressed against zone child population and exposure to deprivation with no covariate adjustment. These results indicate a positive association between deprivation and the incidence of CPCs, with an estimated coefficient of 0.608 (s.e. 0.015). This value can be interpreted as signifying that a 10% increase in exposure to deprivation is associated on average with a 6% increase in the incidence of CPCs. The model predicts a mean incidence of CPCs over all zones of 2.5 and rates for the most and least deprived zones of 5.9 and 0.39 respectively. Thus the model predicts a substantial deprivation gradient with incidence rates over 15 times higher in the most deprived zones compared to the least deprived.

Next, we include a set of covariates in the model to adjust for potential sources of time-varying confounding as discussed in section 3. The results, given in columns 4 and 5, show that our spatial covariates do help to explain variance in the incidence of CPCs. The BIC value for this more complex model indicates a substantial improvement in fit over the previous naive model. As hypothesised, we find that the incidence of CPCs is increasing with urban density (as measured by the population and employment covariates) and also with the density and length of the road network. However, we also again find a positive and significant effect from deprivation on CPCs with an estimated coefficient of 0.600 (s.e. 0.015). On the basis of this estimate, the model predicts a mean incident rate of 2.5 across all zones, 5.6 for the most deprived zones and 0.40 for the least deprived zones. Thus, with the inclusion of spatial covariates we find a slightly less steep but still substantial deprivation gradient (14 rather 15 times the incident rate predicted). The fact that the estimated exposure effect decreases with covariate adjustment is consistent with confounding, but it does not negate the evidence of a positive association between deprivation and CPCs.

We now consider the GLMM results, which in addition to adjustment for spatial covariates, also attempt to account for time-invariant heterogeneity arising at the level of zones or from broader area type and regional classifications. We found that a simple mixed model with adjustment for covariates and an individual specific random intercept for each cross-sectional unit improved the BIC substantially. We then compared this base model to (1) a similar mixed-model but with region and area types as fixed factors, and (2) a multilevel mixed model with random region and area type effects as well as the individual level random intercept. The random specification of regional and area type effects gave the lowest BIC, substantially lower than the BIC achieved using a GLM with time-varying confounders alone (14,429 compared to 42,393).

Columns 6 and 7 of table 1 show results from the Poisson GLMM. The multilevel model results indicate a positive significant effect of deprivation on CPCs. The parameter estimate of 0.583 (s.e. 0.026) is smaller than the Negative Binomial GLM result in column 4. This estimate implies a mean incident rate across all zones of 2.30, but 5.1 for most deprived zones and 0.40 for the least deprived. Thus the inclusion of multilevel random effects again reduces the deprivation gradient, now 13 times as many incidents predicted in the most compared to the least deprived zones, but a clear substantial effect remains. The model therefore provides convincing evidence that higher levels of deprivation are associated with greater numbers of CPCs across heterogeneous spatial units.

It is worth noting that this deprivation effect is largely consistent across cities within the UK. Disaggregating our data by region, and fitting the Poisson GLMM model separately for each city-region, we find evidence of a positive and significant effect from deprivation in all conurbations. The estimated deprivations coefficients are as follows: Tyne & and Wear 0.580 (s.e. 0.131), Manchester & Merseyside 0.622 (s.e. 0.062), Strathclyde 0.723

(s.e. 0.058), London 0.552 (s.e. 0.053), West Midlands 0.584 (s.e. 0.090), and South & West Yorkshire 0.504 (s.e. 0.073).

Another interesting aspect of the results in table 1 is the consecutive fall in accidents year on year over the period of study. In the year 2000, the UK Government, through the Department for the Environment Transport and the Regions (DETR), set out a number of targets for reduction of road traffics which included a 50% fall in children killed or seriously injured and a 10% fall in less serious accidents (DETR, 2000). To achieve these targets the Government sought improvements in vehicle safety and traffic safety management and targeted training and advertising at high risk groups. While the data appear to show that these targets have been met, they do not reveal what the underlying mechanisms are and in fact the decline shown in table 1 can be viewed as part of a long term trend in road accidents which has been evident since the early 1970s. The extent to which this decline will continue is uncertain, but the rate of decline has certainly slowed over the last two decades.

Finally in this section we consider the test for Granger Causality. To implement this model we regress a standardised transformed response derived from the square-root of the accident rate per child on lags of the standardised exposure and lags of the response according to the model specification given in equation (2). We use a two equation version of dynamic GMM which is specified in levels and first-differences, with lagged levels as instruments for the first difference equation and lagged first-differences as instruments for the levels equation. For this approach to produce consistent estimates it is necessary that the instrument matrix be truly exogenous. This in turn requires that there be no second-order serial correlation in the first differenced residuals. For diagnostics we use the Arellano and Bond tests for serial autocorrelation (Arellano and Bond, 1991) and the Sargan/Hansen test of overidentifying restrictions for exogeneity of the instrument matrix (for details see Hall, 2005). The models are estimated using the `plm` packages in `R`.

Several lag permutations, of both covariates and instruments, can be used to estimate the model. We sought to find the most parsimonious model that passes the necessary tests for residual autocorrelation and instrument exogeneity. To successfully eliminate autocorrelation in the differenced residuals it was necessary to use second order lags in the response with a first order lag in the exposure covariate. We then tested different instrument matrix specifications based on lags of the available covariates in relation to the Sargan / Hansen test. We found that an instrument matrix of three lags and deeper satisfies instrument exogeneity.

The result of the Granger causality test indicates that exposure to deprivation helps to predict accident rates, conditional on the dynamic evolution of accident rates themselves and allowing for unit level heterogeneity. We find a positive effect from ward deprivation with an estimated coefficient of 0.330 and associated Wald test that is significant at the 95% level. In other words, according to the Granger criteria, deprivation is causally linked to CPCs in the sense that an increase in CPCs follows from an increase in deprivation.

We therefore find compelling evidence of a deprivation effect on CPCs. While the magnitude of the deprivation gradient reduces having adjusting for confounding, we consistently find a strong positive effect across different model specifications indicating that the relationship is robust to different assumptions on confounding. The Granger causality result supports the hypothesis that deprivation has a prima facie causal, or non-spurious, conditional association with the incidence of CPCs.

4.2. Spatial dependence

In this subsection we implement spatial models to address issues of spatial dependence. As described in section 2, the approach we adopt is based on a Bayesian CAR model incorporating spatially correlated random effect for each zone (s_i) according to a Markov random field specification and with hierarchical random intercepts at the level of zones (v_i), regions (r_i) and area types (a_i)

$$\mu_{it} = \boldsymbol{x}_{it}^{\mathsf{T}} + s_i + v_i + r_i + a_i + \delta_i.$$

Note that zones nest within Regions and Area types. As with the frequentist GLMMs time effects (δ_i) are also included. Following previous literature in the field (e.g. Best et al., 1999; Miaou et al., 2003; Song et al., 2006; Aguero-Valverde and Jovanis, 2008; Wang et al., 2011), we adopt non-informative normal priors for the parameters of the linear predictor, $\beta \sim \mathcal{N}(0, 0.0001)$, and non-informative gamma priors for precision parameters κ, τ_v, τ_r, and τ_a, $Gamma(0.5, 0.0005)$. Initial values are set for all stochastic nodes in the model and we run three chains: one using estimates from a GLMM estimated by Penalised Quasi Likelihood (PQL) (where available) and the other two using PQL GLMM estimates plus and minus four standard errors (e.g. Cowles, 2004). For the spatial models, we do not have frequentist evidence about u_i so instead we use arbitrary values of 0.5, 1.0 and 0.25 in each of the three chains respectively.

Initial runs of the BCA models indicated problems of convergence, which can commonly occur in estimation of Bayesian GLMMs with multilevel random effects (see for example Gilks and Roberts, 1996; Browne et al., 2009). The particular model we are trying to estimate, which includes hierarchical spatial and temporal random effects, is very similar to that estimated by Crainiceanu et al. (2002). They show that simple reparameterisations of the model based around standardised covariates and hierarchical centering can vastly improve MCMC mixing and speed up convergence, and we therefore estimate all the Bayesian models in this way with apparent satisfactory results (see also Gelfand et al., 1995, 1996) .

Convergence of the models was assessed through visual inspection of the trace plots and with reference to the Brooks-Gelman diagnostic. Convergence was achieved for all chains and for each BCA model specification by 4,000 iterations. We discarded the first 5,000 iterations as burn-in and for inference we ran the models for a further 25,000 iterations.

Table 2 below shows the means and 95% credible intervals for model parameters from a non-spatial hierarchically centred GLMM with standardised covariates estimated using Bayesian inference. For comparison the table also includes PQL estimates from a frequentist analysis of the same model.

As we would expect in a model based on a large number of observations with relatively uninformative priors, the frequentist and Bayesian results are very similar. The models indicate a positive statistically significant effect of deprivation on CPCs with a mean posterior value of 0.300 and credible 95% interval (0.273, 0.327). The model results give a gradient predicting 13 times more accidents in the most compared to the least deprived zones (5.1 incidents per annum compared to 0.4, the mean prediction for the sample is 2.3). The results are therefore entirely consistent with those of the Poisson GLMM in the previous section. It is also interesting to note that the posterior means of the precision parameters for zone level random effects and area type random effects correspond closely to the estimated variance components from the PQL model.

Table 3 shows results from the BCA models with spatial random effects.

Compared to the results in table 2 we find little difference having allowed for spatial dependency according to a Markov random field. The only substantial difference relates to

Table 2. PQL GLMM parameter estimates and Bayesian posterior summaries (means and 95% credible intervals) with standardised covariates

	Bayesian GLMM				PQL GLM	
	mean	s.d.	2.5%	97.5%	Est.	S.E.
(Intercept)	1.555	0.604	0.412	2.246	1.142	0.045
log deprivation rate	0.300	0.014	0.273	0.327	0.300	0.014
log child population	0.281	0.033	0.213	0.347	0.281	0.034
log zone employment density	0.040	0.010	0.021	0.059	0.040	0.010
log proximate employment density	0.053	0.062	-0.067	0.179	0.065	0.055
log zone population density	0.156	0.043	0.072	0.243	0.156	0.043
log proximate population density	0.059	0.095	-0.129	0.249	0.017	0.072
log road length	0.296	0.026	0.245	0.347	0.296	0.026
log road network density	0.094	0.027	0.041	0.149	0.097	0.028
year 2002	-0.146	0.017	-0.180	-0.113	-0.146	0.017
year 2003	-0.195	0.018	-0.229	-0.159	-0.194	0.018
year 2004	-0.277	0.018	-0.313	-0.242	-0.277	0.018
year 2006	-0.455	0.019	-0.493	-0.418	-0.454	0.019
year 2007	-0.494	0.020	-0.533	-0.455	-0.492	0.020
BIC					14429	
DIC	40580					
	$\bar{\tau}$	s.d.	$\bar{\tau}^{-1}$			
zone (Intercept)	9.455	0.535	0.106		0.105	0.324
region (Intercept)	131.800	240.900	0.008		0.010	0.098
area (Intercept)	1348.000	1497.000	0.001		0.000	0.013
n	10920				10920	

Table 3. Bayesian posterior summaries (means and 95% credible intervals) for spatial models with standardised covariates

	mean	s.d.	2.5%	97.5%
(Intercept)	-0.126	0.233	-0.613	0.308
log deprivation rate	0.305	0.015	0.275	0.334
log child population	0.220	0.035	0.153	0.289
log zone employment density	0.031	0.010	0.012	0.050
log proximate employment density	0.123	0.068	-0.021	0.252
log zone population density	0.228	0.048	0.134	0.320
log proximate population density	0.065	0.112	-0.145	0.307
log road length	0.309	0.028	0.253	0.366
log road network density	0.049	0.030	-0.009	0.110
year 2002	-0.147	0.017	-0.181	0.113
year 2003	-0.197	0.018	-0.232	0.163
year 2004	-0.283	0.018	-0.319	0.247
year 2006	-0.464	0.020	-0.503	0.426
year 2007	-0.504	0.020	-0.544	0.465
DIC	40545			
τ_v	12.490	1.089	10.560	14.820
τ_r	1063.000	1426.000	18.680	5084.000
τ_a	1555.000	1551.000	91.440	5918.000
κ	21.650	7.443	11.910	41.880
n	10920			

Notes. τ_v, τ_r, and τ_a are estimated precision parameters for the zone, region, and area type effects respectively. κ is the estimated precision parameter of the Gaussian CAR prior.

the proximate employment density covariate for which we find a substantially higher effect in the spatial model. The positive effect of deprivation on CPCs is still clearly indicated and of the same order of magnitude: 0.305 with 95% credible interval (0.275, 0.334). The DIC of the spatial model indicates an improvement in model fit arising from the inclusion of spatial random effects, with the DIC falling in value by 35. Note also that the variances of the zone (v) and regional (r) random effects are lower than in the non-spatial model due to the contribution from spatial dependency.

The results presented in this section indicate that spatial dependency does exist and that we can achieve a better model fit by recognising this source of variation. However, we also find that the spatial and non-spatial models do not differ greatly, certainly with respect to inference on deprivation. We again find a positive association between deprivation and CPCs of similar magnitude to that estimated using non-spatial models. This suggests that the identification of a significant deprivation gradient in our non-spatial models does not suffer from type 1 error due to spatial dependence.

4.3. Interference

Finally in this results section, we present models that acknowledge the potential for interference in exposure between zones. As described in section 2 we use two approaches to construct spatially autoregressive exposure covariates, based on inter-zonal averaging and inter-zonal 'gravity' weighted calculations, which we include in spatial BCA models. The results are shown in table 4 below.

Table 4. Bayesian posterior summaries (means and 95% credible intervals) for spatial models, with spatial interaction exposure covariates. All covariates are standardised.

	mean	s.d.	2.5%	97.5%	mean	s.d.	2.5%	97.5%
(Intercept)	0.1854	0.4123	-0.6550	0.7788	0.2978	0.3183	-0.2252	1.0910
log average inter-zonal deprivation rate (5km radius)	0.2006	0.0240	0.1519	0.2469	-	-	-	-
log gravity weighted inter-zonal deprivation rate (5km radius)	-	-	-	-	0.5762	0.0277	0.5236	0.6312
log child population	0.5175	0.0343	0.4490	0.5828	0.3097	0.0336	0.2444	0.3778
log zone employment density	0.0413	0.0103	0.0210	0.0612	0.0019	0.0099	-0.0172	0.0216
log prox. employment density	0.2978	0.0755	0.1507	0.4502	0.1861	0.0667	0.0526	0.3174
log zone population density	-0.1302	0.0457	-0.2203	-0.0406	-0.1740	0.0421	-0.2559	-0.0911
log prox. population density	-0.2780	0.1370	-0.5354	-0.0325	-0.7429	0.1064	-0.9536	-0.5327
log road network density	0.2458	0.0291	0.1882	0.3036	0.0677	0.0294	0.0118	0.1244
log road length	0.0895	0.0267	0.0352	0.1402	0.2351	0.0269	0.1805	0.2866
year 2002	-0.1416	0.0171	-0.1751	-0.1078	-0.1625	0.0173	-0.1962	-0.1290
year 2003	-0.1850	0.0180	-0.2207	-0.1500	-0.2230	0.0177	-0.2578	-0.1884
year 2004	-0.2666	0.0187	-0.3037	-0.2298	-0.2823	0.0181	-0.3178	-0.2469
year 2006	-0.4289	0.0209	-0.4698	-0.3876	-0.4448	0.0196	-0.4834	-0.4062
year 2007	-0.4644	0.0225	-0.5084	-0.4203	-0.4655	0.0203	-0.5046	-0.4250
DIC	40662				40562			
τ_v	10.1400	0.9962	8.4920	12.4000	12.2500	0.9633	10.5300	14.2900
τ_r	1028.0000	1447.0000	8.6310	5196.0000	924.7000	1379.0000	6.8180	4846.0000
τ_a	224.9000	287.7000	13.4200	922.8000	1704.0000	1687.0000	25.4000	6376.0000
κ	15.1800	6.1160	7.1760	31.1800	22.9900	6.1730	13.7000	37.4000
n	10920				10920			

Notes. τ_v, τ_r, and τ_a are estimated precision parameters for the zone, region, and area type effects respectively.
κ is the estimated precision parameter of the Gaussian CAR prior.

Considering first the model results associated with the average inter-zonal exposure covariate (columns 2 to 5), we see that the mean of the posterior density for this parameter indicates a positive deprivation effect from the spatially averaged exposure variable (0.201). With regard to the other covariates in the model, the results are very similar to those shown in table 3. While the overall indication is that a strong effect of deprivation on CPCs remains having allowed for interference between zones, in fact the predicted gradient is lower than found in other models: 3.7 incident per annum predicted for the most deprived zones compared to 1.1 for the least deprived (mean of 2.3 over the sample). The DIC of the model indicates a significant deterioration in model fit relative to results based on the fixed zone exposure values. Furthermore, it is interesting to note that the estimated variances of the zonal and spatial components (τ_v^{-1} and κ^{-1}) are quite substantially larger, indicating that use of this covariate increases the amount of unobserved heterogeneity. The implication is that modelling interference via the inter-zonal exposure covariate does not improve our ability to account for variance in the incidence of CPCs.

Turning to the results for the model using the inter-zonal 'gravity' weighted exposure covariate (columns 6 to 9), we again find evidence of a positive mean deprivation effect (0.576) with a relatively narrow 95% credible interval (0.524, 0.631). The mean value of the posterior is substantially larger than that reported in previous models. In fact this model predicts a mean accident rate for all zones of 2.26 incidents per annum, but the predicted gap between the most and least exposed zones is substantial, 7.0 incidents and 0.19 respectively. In relation to previous model results, this suggests that both internal levels of deprivation, and proximity to deprivation elsewhere, matter for the incidence of CPCs with the lowest level of incidents being predicted for zones that are not deprived and not near other deprived zones.

There are some other interesting aspects to the results for this model. First, the variance of the spatial random effects is smaller than in the other spatial models estimated. This is likely due to the explicit recognition of zonal interactions within the exposure covariate itself, which accounts for some of the previously unexplained spatial variance in the incidence of CPCs. Second, we find that the DIC of the model using the gravity exposure variable is of the same order of magnitude as that of the previous spatial models in table 3, but shows a substantial improvement over the model using the inter-zonal averaged exposure covariate. The relatively good DIC value and the smaller variance of the spatial random effects indicate that this variable may have been successful to some degree in capturing interference.

In summary, the models with spatially autoregressive exposure covariates show that the existence of a positive association between deprivation and CPCs is robust to some parametric deterministic specifications of interference between zones. There is also evidence that the spatially autoregressive gravity weighted model of interference shows a good fit to the data and that the inclusion of spatial interaction in the linear predictor can help to account for some of the spatial dependency that we observe between zones. Perhaps of most interest, however, is the change in the deprivation gradient that we observe, with a much more substantial effect predicted if we allow for interference of this nature. This implies that the incidence of CPCs depends not only the deprivation status of the zone in question but on that of neighbouring zones as well.

5. Conclusions

This paper has investigated the link between deprivation and the incidence of CPCs in British cities using frequentist and Bayesian approaches to estimate longitudinal spatial GLMMs. To contribute to the existing published cross-sectional evidence on this theme the paper developed models to address confounding, spatial dependence, and interference in exposure between zones.

The results show evidence of confounding and demonstrates that adjustment for known sources of confounding can affect estimates of the deprivation gradient. However, results across various model specifications show consistent evidence of a deprivation effects and analysis according to the Granger model supports the hypothesis that the effect is non-spurious. We also find evidence of spatial dependence, and while it does not appear to induce substantial effects on parameter estimates in our case study, results do show that we can improve model fit by incorporating such dependency. The paper also uncovers evidence of interference in exposure between zones and shows that the assumptions we make about the form of this interference are important and can affect results substantially. In particular, we find that modelling interference through a spatially autoregressive inverse distance weighted covariate provides a reasonably convincing adjustment for interference and indicates a substantially greater deprivation effect.

While the paper has been able to highlight the need to carefully consider the assumptions underpinning area based casualty models, there are several areas for future research remaining. First, we must acknowledge that some of the covariates used, and particularly those representing traffic flows, are essentially modelled proxy variables and this creates uncertainty about the effects being measured and their relations to deprivation. In future work, it would be interesting to directly incorporate this uncertainty into the analysis, for instance through a two-step model with bootstrapping for variance estimation or through specification of a measurement error model. Second, more work is required on the form of autoregressive process assumed for interference and on the potential for use of more flexible forms within semiparametric models. Third, it would be useful to develop a unified approach to confounding, spatial dependence and interference within the same model. Finally, the work presented in this paper could benefit from a more formal framework in which to consider issues of causality, for instance along the lines of the potential outcomes model for causal inference (e.g. Rosenbaum and Rubin, 1983), which has been extended recently to cover continuous treatments and exposures by Imbens (2000); Hirano and Imbens (2004).

Acknowledgements

We are grateful to the Associate Editor and two anonymous referees for helpful comments and suggestions. We also thanks Ordnance Survey for providing the longitudinal road network data used in this study. David Stephens is supported by a Discovery Grant from the Natural Sciences and Engineering Council of Canada

References

Aguero-Valverde, J. and P. P. Jovanis (2008). Analysis of road crash frequency with spatial models. *Transportation Research Record 2061*.

Arellano, M. and S. Bond (1991). Some tests of specification for panel data: Monte carlo

evidence and an application to employment equations. *Review of Economic Studies 58*, 277–297.

Arellano, M. and O. Bover (1995). Another look at the instrumental variable estimation of error component models. *Journal of Econometrics 68*, 29–51.

Bajekal, M. (2005). Healthy life expectancy by area deprivation: magnitude and trends in England, 1994-1999. *Health Statistics Quarterly 25*, 18–27.

Beale, C. M., J. J. Lennon, J. M. Yearsley, M. J. Brewer, and D. A. Elston (2010). Regression analysis of spatial data. *Ecology Letters 13*(2), 246–264.

Besag, J., J. York, and A. Molli (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics 43*, 1–20.

Best, N., R. A. Arnold, A. Thomas, L. A. Waller, and E. M. Conlon (1999). *Bayesian Models for Spatially Correlated Disease and Exposure Data*, Chapter in Jos M. Bernardo, James O. Berger, A. Philip Dawid and Adrian F. M. Smith (eds) Bayesian Statistics 6, pp. 131–156. Oxford: Oxford University Press.

Browne, W. J., F. Steele, M. Golalizadeh, and M. J. Green (2009). The use of simple reparameterizations to improve the efficiency of Markov chain monte carlo estimation for multilevel models with applications to discrete time survival models. *Journal Of The Royal Statistical Society Series A 172*(3), 579–598.

Carstairs, V. (2000). Socio-economic factors at areal level and their relationship with health. In P. Elliot, J. C. Wakefield, N. G. Best, and D. J. Briggs (Eds.), *Spatial epidemiology: methods and applications*, pp. 51–67. Oxford: Oxford University Press.

Carstairs, V. and R. Morris (1991). *Deprivation and health in Scotland*. Aberdeen: Aberdeen University Press.

Christie, N. (1995). Social, economic and environmental factors in child pedestrian accidents: a research overview. Technical Report 116, Transport Research Laboratory, Berkshire.

Cooper, H. (2002). Investigating socio-economic explanations for gender and ethnic inequalities in health. *Social Science & Medicine 54*(5), 693–706.

Cowles, M. K. (2004). Review of winbugs 1.4. *The American Statistician 58*(4), 330–336.

Cox, D. R. (1958). *Planning of experiments*. London: John Wiley & Sons.

Crainiceanu, C. M., D. Ruppert, J. Stedinger, and C. Behr (2002). *Improving MCMC mixing for a GLMM describing pathogen concentrations in water supplies*, Chapter in C. Gatsonis et al (eds) Bayesian Statistics vol VII, pp. 207–221. New York: Springer.

Cressie, N. C. A. (1993). *Statistics for spatial data*. Chichester: John Wiley & sons.

DETR (2000). *Tomorrows Roads Safer for Everyone*. London: DETR.

Diggle, P., R. A. Moyeed, and J. A. Tawn (1998). Model-based geostatistics. *Applied Statistics 47*, 299–350.

Diggle, P. J. and P. J. Ribeiro (2007). *Model-based Geostatistics*. New York: Spinger.

Eichler, M. and V. Didelez (2010). On Granger causality and the effect of interventions in time series. *Lifetime Data Analysis 16*, 3–32.

El-Geneidy, A. and D. Levinson (2006). *Access to destinations: development of accessibility measures*. Mineapolis: Minnesota Department for Transportation.

Gelfand, A. E., S. K. Sahu, and B. P. Carlin (1995). Efficient parametrisations for normal linear mixed models. *Biometrika 82*(3), pp. 479–488.

Gelfand, A. E., S. K. Sahu, and B. P. Carlin (1996). *Efficient Parametrisations for Generalized Linear Models*, Chapter in Jos M. Bernardo, James O. Berger, A. Philip Dawid and Adrian F. M. Smith (eds) Bayesian Statistics 5, pp. 165–180. Oxford: Oxford University Press.

Gilks, W. and G. Roberts (1996). *Strategies for improving MCMC mixing*, Chapter in W.R. Gilks et al (eds) Markov Chain Monte Carlo in practice, pp. 89–110. London: Chapman & Hall.

Graham, D. J. and S. Glaister (2003). Spatial variation in road pedestrian casualties: the role of urban scale, density and land-use mix. *Urban Studies 40*, 1591–1607.

Graham, D. J., S. Glaister, and R. J. Anderson (2005). The effects of area deprivation on the incidence of child and adult pedestrian casualties in England. *Accident Analysis & Prevention 37*, 125–135.

Graham, D. J. and C. Megueulle (2006). A spatial analysis of cycling casualties in England: the effect of exposure. working paper, Imperial College London.

Graham, D. J. and P. C. Melo (2011). Assessment of wider economic impacts of high-speed rail for great britain. *Transportation Research Record 2261*, 15–24.

Graham, D. J. and D. A. Stephens (2008). Decomposing the impact of deprivation on child pedestrian casualties in England. *Accident Analysis & Prevention 40*, 1351–1364.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectoral methods. *Econometrica 37*, 424.

Green, J., H. Muir, and M. Maher (2011). Child pedestrian casualties and deprivation. *Accident Analysis and Prevention 43*(3), 714–723.

Hall, A. R. (2005). *Generalized Method of Moments*. Advanced Texts in Econometrics. Oxford: Oxford University Press.

Hewson, P. J. (2005). Epidemiology of child pedestrian casualty rates: Can we assume spatial independence. *Accident Analysis and Prevention 37*(4), 651–659.

Hippisley-Cox, J., L. Groom, D. Kendrick, C. Coupland, E. Webber, and B. Savelyich (2002). Cross sectional survey of socioeconomic variations in severity and mechanism of childhood injuries in trent 1992-7. *British Medical Journal 324*, 1132–1138.

Hirano, K. and G. W. Imbens (2004). The propensity score with continuous treatments. In A. Gelman and X. Meng (Eds.), *Applied Bayesian modeling and causal inference from incomplete data perspectives*, pp. 73–84. New York: Wiley.

Holtz-Eakin, D., W. Newey, and H. S. Rosen (1988). Estimating vector autoregressions with panel data. *Econometrica 56*, 1371–1395.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika 87*(3), 706–710.

Jones, K., M. Gould, and C. Duncan (2000). Death and deprivation: an exploratory analysis of deaths in the health and lifestyle survey. *Social Science & Medicine 50*(7-8), 1059–1080.

Lancaster, G. and M. Green (2002). Deprivation, ill-health and the ecological fallacy. *Journal of the Royal Statistical Society A 165*(2), 263–278.

Lorant, V., I. Thomas, D. Delige, and R. Tonglet (2001). Deprivation and mortality: the implications of spatial autocorrelation for health resource allocation. *Social Science & Medicine 53*(12), 1711–1720.

Miaou, S. P., P. S. Hu, T. Wright, A. Rathi, and D. SC (2003). Roadway traffic crash mapping: a space-time modeling approach. *Journal of Transportation and Statistics 6*(1), 33–57.

Noland, R. and M. Quddus (2004). A spatially disaggregate analysis of road casualties in England. *Accident Analysis and Prevention 36*, 973984.

Roberts, I. and C. Powers (1996). Does the decline in child injury mortality vary by social class? a comparison of class specific mortality in 1981 and 1991. *British Medical Journal 313*, 784–786.

Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association 102*(477), 191–200.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica 48*, 1–48.

Song, J., M. Ghosh, S. Miaou, and B. Mallick (2006). Bayesian multivariate spatial models for roadway traffic crash mapping. *Journal of Multivariate Analysis 97*(1), 246–273.

Thomas, A., N. Best, D. Lunn, R. Arnold, and D. Spiegelhalter (2004). *GeoBUGS User Manual*.

Townsend, P., P. Phillimore, and A. Beattie (1988). *Health and deprivation: inequality in the North*. London: Croom Helm.

Wang, C., M. Quddus, and S. Ison (2011). A spatio-temporal analysis of the impact of congestion on traffic safety on major roads in the UK. *Transportmetrica*.

White, D., R. Raeside, and D. Barker (2000). Road accidents and children living in disadvantaged areas: A literature review. Technical report, Scottish Executive Central Reseach Unit, Edinburgh.