A segment-based spatial analysis of non-motorised road traffic casualties occurring in non-built up areas of England and Wales, 1999-2008.

Thesis · June 2010

CITATIONS
0

READS
266

1 author:

Marcus Young
University of Southampton
20 PUBLICATIONS **17** CITATIONS

SEE PROFILE

A segment-based spatial analysis of non-motorised road traffic casualties occurring in non built-up areas of England and Wales, 1999-2008.

Marcus A Young B.Sc. (Hons)

Faculty of Life and Health Sciences

University of Ulster, Coleraine

Submitted for the award of Master of Science

Geographic Information Systems

2010

Contents

Acknowledgements

I would like to thank the following people: Dr. Sally Cook, my supervisor for this research project, who provided invaluable advice and suggestions on improvement of this paper; Ianko at ET SpatialTechniques for providing support for ET Geowizards and responding so rapidly to fix bugs that the large datasets used in this study exposed; Dr. Richard Mee for his understanding and support; and Darren Wade without whose endless support and encouragement completion of this project would not have been possible.

Abstract

In the UK road accidents outside of built-up areas accounted for 25% of pedestrian fatalities and 50% of cyclist fatalities in 2008, yet this group of road users has received very little research attention, whilst a large body of research into urban non-motorised road casualties exists. This study adopts a segment-based approach and develops a series of negative binomial regression models to explore the relationship between non-motorised transport casualties that occurred in non built-up areas of England and Wales during 1999-2008 and a range of explanatory factors, including some variables uniquely relevant to this type of road user. Explanatory factors with a significant and positive association with casualty incidence included A-class and B-class roads, the number of intersections, and the presence of National Trails or National Cycle Network routes. A significant negative association with casualties was found for road sinuosity and distance from the nearest built-up area. This research represents the first national-scale segment-based study of road accidents carried out in the UK, and the approach is considered an improvement over several national area-level studies that have been conducted. The methodology developed and outlined in this paper could be applied to similar research within the UK and abroad.

Abbreviations

AADF annual average daily flow

BSU basic spatial unit

CPRE Campaign to Protect Rural England

CSV comma-separated values

CSR complete spatial randomness

DfT Department for Transport

GIS geographical information system

IRR incidence rate ratio

ITN Integrated Transport Network Layer

LISA local indicator of spatial association

NCN National Cycle Network

NGR National Grid Reference

NB negative binomial

NMT non-motorised transport

OLS ordinary least squares

SAR simultaneous autoregressive

ZINB zero-inflated negative bimonial

Note on access to contents

I hereby declare that with effect from the date on which the dissertation is deposited in the Library of the University of Ulster I permit the Librarian of the University to allow the dissertation to be copied in whole or in part without reference to me on the understanding that such authority applies to the provision of single copies made for study purposes or for inclusion within the stock of another library. This restriction does not apply to the copying or publication of the title and abstract of the dissertation. IT IS A CONDITION OF USE OF THIS DISSERTATION THAT ANYONE WHO CONSULTS IT MUST RECOGNISE THAT THE COPYRIGHT RESTS WITH THE AUTHOR AND THAT NO QUOTATION FROM THE DISSERTATION AND NO INFORMATION DERIVED FROM IT MAY BE PUBLISHED UNLESS THE SOURCE IS PROPERLY ACKNOWLEDGED.

List of Figures

List of Tables

1. Introduction

In Britain during 2008 there were 44,885 non-motorised transport (NMT) road users, that is pedestrians, cyclists and horse riders, who were killed or injured in road accidents. Analysis by the Department for Transport (DfT) distinguishes between accidents which occurred on built-up roads and non built-up roads[1]. The vast majority of the NMT casualties in 2008 (42,516) occurred on built-up roads and 1.15% of these were fatal. On non built-up roads 2,285 NMT casualties occurred, and 7.92% of these were fatal, a rate almost seven times that of built-up areas. Non built-up areas accounted for 25% of pedestrian fatalities and 50% of cyclist fatalities in 2008 (Department For Transport, 2009b).

The UK Government is actively promoting walking and cycling due to dual concerns about health of the population and the environment. The UK Government has also recently closed consultation on its Road Safety Strategy Post 2010 document: "A Safer Way: Consultation on Making Britain's Roads the Safest in the World" (Department For Transport, 2009a). This document has been criticised by the Campaign to Protect Rural England (CPRE) for not addressing the issue of country lanes[2]. A research exercise by CPRE found that 65% of respondents felt always or sometimes threatened by traffic when walking, cycling or riding on country lanes, and only 3% felt completely safe from traffic (Campaign To Protect Rural England, 1999). The CPRE is campaigning for a reduction in the national speed limit for C-class and unclassified roads from the current 60mph to 40mph to protect vulnerable road users. The walking charity Ramblers is running a "safe to cross" campaign for the improvement of crossing points where established rights of way have been severed by fast and busy roads passing through rural areas and now pose a danger to horse riders and walkers

[1]Built-up roads exclude motorways and include roads with a speed limit of 40mph or less; non built-up roads exclude motorways and include roads with a speed limit above 40mph.

[2]Country lanes are defined by the CPRE as C-class and unclassified roads in rural areas.

(Ramblers, 2003). In contrast, in evidence submitted to the House of Commons Transport Committee's Eleventh Report of Session 2007–08, the Association of British Drivers called for the 40mph single carriageway national speed limit that applies to heavy goods vehicles to be scrapped, and the Road Haulage Association has indicated it would like the 40mph limit increased on A-class roads (House Of Commons Transport Committee, 2008).

There appears to be no published research that has examined the relationship between road characteristics and the incidence of NMT casualties that specifically occur outside of built-up areas, and no research that has taken into account explanatory factors of unique relevance to NMT road users, such as the presence of footpath crossing points. Research is clearly needed to help inform the debate and assist policy makers in developing an appropriate NMT casualty reduction strategy for non built-up roads. Filling this gap in understanding is the primary aim of the research described in this paper.

This study uses negative binomial (NB) regression models to explore the relationship between aggregated counts of NMT casualties and a range of explanatory factors. As the influence of these factors may vary depending on the type of NMT road user and the injury severity, disaggregated models are developed for these different groups. The spatial unit used for casualty count aggregation is the road segment, and this is believed to be the first national-scale segment-based study of road accidents to have been completed in the UK. In addition to the regression analysis, casualty hot zones are generated for each NMT user type using kernel density estimation techniques. These hot zones are used to investigate the extent to which NMT casualties in non built-up areas are clustered at high-risk locations.

This paper follows the following structure: an initial review of relevant literature which concludes with a statement of the research objectives; a description of the data used and methodological approaches adopted, including how the sta-

tistical models were developed and the issue of spatial autocorrelation was addressed; presentation of the results from the regression models and hot zone analysis along with discussion of the findings and potential limitations of the study; and finally a conclusion where the study is briefly summarised, key findings are drawn out and considered in relation to road safety strategy, and proposals for future study are outlined.

2. Literature review

This review of the literature begins by considering the scope of previous road accident research and how this may be relevant to and inform the specific study of NMT casualties outside of built-up areas. This is followed by individual sections devoted to the different approaches adopted in prior research, including non-spatial, count-based (segment or area-level) and hot zone analysis. The selection of appropriate statistical models for analysing count-based data is then considered, followed by a review of options available for addressing the issue of spatial autocorrelation. Finally, the research objectives of this study, as informed by the literature review, are defined.

2.1. Scope of previous research

There is no standard term to identify the study of road traffic accidents and related casualties that occur outside of urban areas. Common terms used in the literature include rural, non built-up, and non-urban but the meanings of these terms can vary markedly between studies. For example, when Qin and Ivan (2001) examined pedestrian casualties in rural areas of Connecticut, United States, the study sites were within towns with populations of some 10,000 or more located in rural areas. In the UK, the DfT defines rural roads as major and minor roads (excluding motorways) that are outside of urban area polygons for settlements with a population of 10,000 or more (Department For Transport,

2009b). Using this definition many of the "rural roads" will be located in built-up areas in small towns and have more in common with urban roads than roads entirely outside of built-up areas. A report examining the nature of rural road accidents in Cambridgeshire defined rural roads as those with a speed limit greater than 40mph (Hughes, 1994), whilst the same definition is used by the DfT to distinguish built-up roads from non built-up roads (Department For Transport, 2009b). This practice of identifying rural roads by their speed limit is being phased out in the UK as speed limits have been reduced on rural roads in recent years (Lynam, 2007).

This study is concerned with NMT casualties that occur outside of built-up areas, defined by polygons that correspond to settlements with an extent of at least 20 hectares and a resident population of at least 1,500. There has been very little previous research examining road traffic accidents that occur on these non built-up roads in the UK and there appears to have been none that has investigated factors specific to NMT users of these roads - such as public footpaths that are intersected by roads, national walking trails that contain on-road sections, and promoted on-road cycle routes. The research that has been carried includes several DfT sponsored reports by the Transport Research Laboratory. These have included a review of the potential policy options for rural road safety, including some consideration of NMT users (Lynam, 2007) and an investigation into the causes of collisions involving cyclists in urban and rural settings, primarily based on a descriptive analysis of UK national road accident injury data (known as STATS19), but including other data sources and an international literature review (Knowles et al., 2009).

Once the scope is widened to encompass road traffic accidents involving any road user type and in any location a large body of academic research becomes available. These studies can be divided into three broad approaches: non spatial; spatial analysis that models count data to explore contributory factors or develop

prediction models; and spatial analysis to identify and characterise locations with high accident or casualty frequency (often referred to as accident hot spots or zones). In addition there are studies that primarily aim to consider statistical methods or to address specific problems that arise with the statistical modelling of accident data, such as spatial autocorrelation.

2.2. Non spatial studies

Studies of this type analyse reported accident data to gain insight into the characteristics of accidents and casualties, they do not make use of spatial analysis techniques available in geographical information system (GIS) but may have a spatial component, for example disaggregating accidents by region (Edwards, 1996).

Stone and Broughton (2003) examined some 32,000 fatal and serious injury cycling accidents that occurred in Great Britain between 1990 and 1999, based on STATS19 data. They produced a series of univariate tabulations of accident incidence rates and associated fatality rates. They found that while three quarters of the accidents occurred on 30mph roads, the fatality rate showed a marked rise as the speed limit increased. On 30mph roads only 3% of accidents were fatal, rising to 6% at 50mph, 11% at 60mph and 20% on 70mph roads. As the road speed limit increased the proportion of rear impact accidents (not occurring at junctions) rose from 12% on 30mph roads to 56% on 70mph roads. The study also found that 70% of the accidents occurred within 20m of a junction, half of these at T-junctions.

Lynam (2007) carried out an analysis of rural road casualties[3] between 2000 and 2005, again based on STATS19 data. Tabulated data shows that 69% of pedestrian fatalities on rural roads occurred on major roads[4] and of these 83%

[3]In this study "rural roads" were those outside of towns of 10,000 or more population, identified using a GIS-based system.

[4]Major roads are Motorways, Dual Carriageways and A-class roads.

occurred where the speed limit was 50mph or greater. In the case of cyclists, 57% of fatalities on rural roads were on major roads, and of these 86% occurred where the speed limit was 50mph or greater. These figures indicate that a majority of pedestrian and cyclist fatalities in rural areas occur on major roads, and that vulnerable road users are dangerously interacting with motorised vehicles on major rural roads with high posted speed limits.

One of the disadvantages of studies that use this approach to analysing accidents (or casualties) is that they only consider information that is recorded on the official accident reports, they do not consider supplementary information that may be available in other datasets, such as other roadway characteristics or information about land-use classification. It would be possible to append additional attributes to each accident or casualty using a GIS, but a further limitation would remain - this type of analysis can only consider the characteristics of accidents that happened and the location where they happened, it can reveal nothing about the non-accidents - the location and characteristics of locations where no accident occurred. Alternative approaches that use a spatial unit as the basic statistical unit can overcome these limitations (Flahaut, 2004).

2.3. Count-based models

In this approach counts of accidents or casualties are aggregated and then an appropriate statistical model is developed either to aid the understanding of the relationship between explanatory variables and the event, or to be used to predict counts in the future at other locations based on known explanatory variables (Guikema and Coffelt, 2009). The two main methods of count aggregation in the literature are area-level (for example district, ward or census enumeration area in the UK; state or census tract in the United States) or road segment (sometimes called road link). Generally speaking the area studies tend to consider a larger geographic region (for example a national study), whilst the segment studies tend to be smaller in scope, limited to a locality or a specific city. In both cases ex-

planatory variables are restricted to those that can be assigned to the aggregation unit and, in contrast to the non-spatial studies, characteristics of the accidents, vehicles and casualties involved are not considered.

2.3.1. Area-level studies

There have been several national-based area-level studies of road casualties in the UK (e.g.Wang et al., 2009a; Jones et al., 2008; Graham and Stephens, 2005, 2008; Noland and Quddus, 2004).

Graham and Stephens (2005) explored the effects of deprivation on the incidence of child and adult pedestrian casualties. They assigned each pedestrian casualty in England to a census ward based on recorded national grid reference and then developed NB regression models. As well as deprivation measures, these models included other explanatory variables such as population density, network nodes per unit area (used as a proxy for the extent of built development), and length of road by class. Results for both child and adult casualties showed a significant positive relationship between the length of A-road in a ward and the number of casualties, and a significant negative relationship between the length of minor road and the number of casualties. The density of network nodes and population also showed significant positive association with casualties.

In another ward-level study, Wang et al. (2009a) considered the effects of road speed and curvature on traffic casualties in England. Their analysis was disaggregated with separate models for motorised users, non-motorised users (cyclists, horse riders and pedestrians) and vulnerable users (non-motorised users plus motorcyclists) for each injury category (fatal, serious or slight). Results for the non-motorised models showed significant positive coefficients for all injury types for A-road and B-road length, log of population and log of employment. The number of nodes was a significant and positive factor for the serious and slight injuries models. The models for total road casualties showed significant negative coefficients for bend density (curvature measure) for all injury severi-

ties, though for the non-motorised casualties the negative coefficient was significant only in the serious injuries model.

Jones et al. (2008) aggregated casualty counts for 1995-2000 by each local authority district in England and Wales. They explored a large range of explanatory variables relating to traffic exposure, resident population, landcover, elevation and hilliness, climate, road curvature and junction density. Many of the variables within these broad categories were removed prior to the final model as part of the two-stage regression process adopted. Whilst separate models were generated for casualty severity, no distinction was made between the type of road user. The variable "percentage of road length passing through urban areas"[5] was found to be negatively related to the number of fatalities, not significant for the serious casualties model, and positively related to slight casualties.

Noland and Quddus (2004) included dummy variables[6] in their regression model representing the land use classification of the ward - ranging from wholly rural (the reference variable) to wholly urban. They reported that as the level of urbanization increased there were significantly fewer casualties of all types (fatal, serious, and slight). In fact their results suggest that if all wards were to be classified as predominantly rural, slight injuries would increase by 106.5% and serious injuries would increase by 97.5%, whilst if all wards were to be classified as wholly urban, slight injuries would decrease by 6.8% and serious injuries would decrease by 16.49%. The authors do not critically examine these findings despite them appearing to be at odds with what is known from the raw data - in 2001 (Noland and Quddus used 1999 data) there were 1.54 times the number of fatal casualties on rural roads as urban roads, but for serious and slight injuries there were 0.68 and 0.43 times the number of casualties respectively.

[5]This was calculated using a GIS and based on the Institute of Terrestrial Ecology Landcover Map of Great Britain - a raster dataset with 1km grid cells.

[6]Dummy variables are used to introduce a categorical variable into a regression model, with each category converted into a separate binary variable. One of the dummy variables, the reference variable, must be excluded from the model to avoid multicollinearity.

Jones et al. (2008) produced three ward maps for England and Wales that showed "a clear urban rural pattern in the casualty rates, with higher values generally found in the more urban districts". However, in their final model, the percentage of road length passing through urban areas shows a non significant negative relationship with serious casualties, again seemingly at odds both with the raw STATS19 data and their own exploratory data analysis. This raises concerns about drawing inappropriate inferences from area-level studies, at least with respect to certain explanatory variables.

The studies considered here have aggregated casualty counts into areas based on administrative boundaries which have been created for a purpose not related to the study in question. These arbitrary areas, referred to as "modifiable" areas, are subject to a number of issues that can influence the results of statistical analysis, and that are known together as the Modifiable Areal Unit Problem, or MAUP (Openshaw, 1984). One of these issues is the "ecological fallacy" problem, which is concerned with the applicability of analyses that have been carried out on aggregated entities to the original entities themselves. A significant association between a variable and casualty count at ward or district level does not infer a correlation at the level of the individual casualty, and a lack of association at the aggregate level does not mean an association does not exist at the individual level. Another MAUP issue is the scale effect whereby correlations between variables become stronger as areas become larger (Openshaw, 1984), and can also result in the direction of association changing, from positive to negative and vice versa (Flowerdew et al., 2008).

It is possible that MAUP effects are present in the area-level studies considered here and may explain results that appear counter intuitive, another example of which is the relationship between junction density and casualty or accident count. Jones et al. (2008) note in their methodology that "the presence of junctions is a well established risk factor" for road accidents, but the junction density

variable used - the number of junctions per kilometre of road - did not make it into their final regression model. Noland and Quddus (2004) expected junctions to result in more casualties, but found no major association and concluded that the small difference in the number of junctions between the wards has little effect on casualty numbers. Quddus (2008) found the number of junctions to be statistically insignificant in all models and it was omitted from the final list of explanatory variables. However, an analysis of over two million accidents recorded in STATS19 data for 1999-2008 shows that 67% of them occurred at or within 20m of a junction, suggesting that inferring from the area studies above that there is no association between junctions and accidents could be questionable. If the number of junctions per area is very similar across all areas then even if every accident was associated with a junction, the models would not be sensitive to this.

2.3.2. Segment studies

The road section or segment is considered to be the most appropriate aggregation unit for road accident analysis. However, an important factor to consider in this type of study is what the most appropriate segment length for analysis is and how that should be selected (Flahaut et al., 2003). If the segment is too long then characteristics of (or related to) the segment may vary along its length and this could impair statistical analysis and subsequent interpretation, although a large number of segments in a study can minimize these heterogeneity effects (Koorey, 2009). If the segment is too short then a high proportion of the segments may have a zero accident count and the mean count is likely to be very low, presenting difficulties with statistical models (Koorey, 2009). Another issue with short segments is a greater likelihood that the factors contributing to an accident belong to a segment other than the one that the accident has been allocated to. This could occur when the accident location is recorded as the position where the vehicle finally came to rest (Koorey, 2009). Similarly if the segment

size is too short in relation to the accuracy of the accident location data, then the accidents would be assigned to segments of too fine a spatial scale.

Two approaches have been adopted for choosing segment length, either fixed-length or variable-length. In fixed-length studies the road network is divided into equal length segments. In a study of predictive models for accidents on Flemish motorways, Geirt and Nuyts (2006) used a fixed road segment length of 100m, whilst Parida et al. (2006) used 200m segments in a study of crashes on non-urban highways in India. In variable-length studies the road network is divided to derive segments that are homogeneous with respect to specific attributes, such as width or gradient. Berhanu (2004) defined homogeneous segments based on adjacent land use and road characteristics, with segment lengths ranging between 0.4km and 3.2km. The variable-length approach may involve manually collecting and assessing data (e.g Berhanu, 2004) and may be less suited to larger datasets. However, automated approaches are being developed, such as an application to aid crash analysis of the national rural State Highway network in New Zealand which automatically segments the network based on curvature, number of lanes or width and speed limit (Koorey, 2009). In some studies the segment length is variable but only because the segment structure present in the raw road data has been used without adjustment (e.g. Wang et al. 2009b; Grundy et al. 2009).

There does not appear to have been any national segment-based research carried out in the UK, but several studies of smaller scope have been completed. Grundy et al. (2009) examined the impact of the introduction of 20 mph zones on road casualties in London. Using a GIS, STATS19 casualties for 1986-2006 were assigned to London road segments extracted from the Ordnance Survey Integrated Transport Network Layer (ITN) layer and counts were then generated for each road segment for each year prior to statistical analysis. Results showed a 40% reduction in casualties associated with introduction of the zones, and a 50%

reduction in the number of killed or seriously injured children. Other researchers have explored the impact of congestion on road accident occurrence on the M25 motorway and developed an innovative method to assign STATS19 accident data (2004-2006) to the correct carriageway segment. This method utilises the perpendicular distance from the accident to the segment and the angular difference between the vehicle direction prior to the accident (recorded in STATS19) and the direction of the segment - the "correct" segment being the one with a short perpendicular distance and a small angular difference. The study found no evidence that congestion impacted accident frequency (Wang et al., 2009b).

2.4. Hot spot or hot zone analysis

Studies that use this approach are concerned with identifying and analysing concentrations or clusters of traffic accidents. It may then be possible to improve road safety by addressing common causal factors that are present at these locations (Steenberghen et al., 2009).

2.4.1. Identifying hot spots or zones

The two most common techniques for identifying traffic accident clusters are the kernel density method (e.g. Anderson, 2009; Erdogan et al., 2008; Pulugurtha et al., 2007) and the local spatial-autocorrelation method (e.g. Flahaut, 2004; Steenberghen et al., 2004; Geurts et al., 2005).

With the kernel density method the study area is first divided into equal-sized cells (the number predetermined by the chosen cell size) and then a circular search area (the kernel) is constructed around each accident point. A mathematical function is then applied to calculate the kernel value which decreases from 1 at the accident point to 0 at the kernel boundary. Each cell's density value is then calculated by summing the kernel values that overlap it. The radius of the kernel is known as the bandwidth, and the larger the bandwidth the more accident points that will be included within it and the smoother the surface created.

The researcher will decide the threshold density value for a cell to be considered a hot spot, and contiguous cells with high density values can be combined into a hot zone. (Anderson, 2009; Pulugurtha et al., 2007; Steenberghen et al. 2009).

The local spatial-autocorrelation method uses a local indicator of spatial association (LISA), such as Local Moran's I. The accident points are aggregated into spatial units and a count per unit calculated. A spatial weights matrix is then generated to define the neighbourhood relationships, and this matrix is used to calculate a LISA for each spatial unit, which indicates how similar or dissimilar the accident count of the spatial unit is to the count of neighboring spatial units. For example, Flahaut et al. (2003) divided the two-lane N29 road in Belgium into 100m segments and developed a methodology which calculates 10 separate Moran's I values for each segment (based on 2 - 20 contiguous neighbours) and defines a hot zone as the segment under consideration plus the number of neighboring segments which maximise Moran's I (Flahaut et al., 2003).

A potential problem with the kernel method, and the LISA method if not using a contiguity-based matrix, is the use of a Euclidean measure of distance (i.e. the straight-line distance in any direction through space), when it is known that road accidents are constrained to a network. Road accidents may fall within the kernel, or be included as neighbours in a weights matrix, when they are on a non-contiguous section of road that is farther in network distance from the point of interest than the Euclidean distance would suggest. Research indicates that the Euclidean distance may sometimes be a good approximation for shortest-path distance on a network, but the difference is significant when the Euclidean distance is less than 500m (Okabe and Satoh, 2009). Several approaches to overcome these limitations have been suggested, such as the network kernel method within the SANET toolbox (Okabe and Satoh, 2009) and an alternative proposed by Xie and Yan (2008), and a moving segment approach developed by Steenberghen et al. (2009).

2.4.2. Analysing hot spots or zones

A variety of methodologies have been explored to help understand the characteristics of hot spots or zones once they have been identified. Anderson (2009) linked adjacent hot spot cells, assigned accident and environmental attributes to these hot spots, and then used a K-means clustering algorithm to first create clusters[7] of hot spots based on similar attributes, and then to create groups of similar clusters. Geurts et al. (2005) identified hot zones and then divided the accidents into two groups - those that occurred within a hot zone and those that occurred outside a hot zone. A data mining technique was then applied to each group of accidents to identify accident circumstances that frequently occur together (known as frequent item sets). To explore the impact that various road and local environment attributes have on the occurrence of road accident hot zones, Flahaut (2004) used a binary variable indicating whether a road segment belonged to a hot zone (value = 1) or not (value = 0) as the dependent variable in a spatial autologistic regression model. Pulugurtha et al. (2007) evaluated a range of methods for ranking pedestrian accident hot zones.

A potential concern with the hot spot approach is that while it focuses attention on specific locations where a concentration of accidents has occurred, the majority of accidents may well happen outside of these areas. Morency and Cloutier (2006) identified 22 hot spots in Montreal, Canada where there had been at least 8 pedestrian injuries over a 5 year period, and found that accidents at these locations represented only 4% of total pedestrian injuries, and that the 5,082 total injuries had occurred at more than 3,500 different sites. They conclude that focussing on hot spots is a "high risk preventative strategy".

[7]The clustering process is a mechanism to classify or group hot spots based on similarities, clusters do not represent contiguous hot spots.

2.5. Statistical models

There are several reasons why standard ordinary least squares (OLS) regression is not appropriate for the analysis of count data. Firstly OLS regression can predict negative values which is impossible for count data where values must be zero or greater, and secondly two important assumptions of the OLS model - normal distribution and homoscedasticity - are typically violated by count data (Elhai et al., 2008). Count data is intrinsically heteroskedastic with the variance of the residuals increasing as the expected value of the count variable increases, and right skewed with many low values and fewer high values (Hilbe, 2008). These violations mean that if OLS regression is used on count data tests of statistical significance of regression coefficients will be biased, potentially leading to Type I errors (falsely rejecting the null hypothesis that a regression coefficient is zero) (Coxe et al., 2009).

To overcome these limitations a count response regression model is required. Count response models are part of a wider group of models known as discrete response models that are used for modelling data with non-negative integer responses, other examples include binary logistic and probit regression. Poisson regression is the basic count model and other count models are based on it. A limitation of the Poisson model is that it assumes that the mean and variance of the dependent variable are equal, which is often not the case in real data. When the variance is greater than the mean the data are considered to be overdispersed, and the Poisson model may not fit the data well. In these circumstances the NB model can be used (Hilbe, 2008). The use of the NB model is common in the road accident analysis literature (e.g. Noland and Quddus, 2004; Jones et al., 2008; Wang et al., 2009a).

The NB distribution has an expected number of zero counts for any given mean, and as the mean increases the number of expected zero counts decreases. The NB model can be extended to handle situations where the number of zero

counts exceed the theoretical requirements of the model, and this is known as the zero-inflated negative bimonial (ZINB) model (Hilbe, 2008). It is important to note that in a ZINB model the zeros are assumed to come from two different distributions - structural zeros from a binary distribution and sampling zeros from a count distribution (Hilbe, 2008). In the context of road safety modelling, Lord et al. (2007) describe this as a dual-state generating process, and note that for the ZINB model assumptions to hold it must be possible for sites being studied to exist in two states - an inherently safe state (always zero) and a non-zero state (but where zero accidents may occur during a sample period). Lord et al. (2007) criticise the use of ZINB regression simply in order to get a better model fit and argue that inherently safe roads do not exist and that zero-inflated models "should be avoided for modelling motor vehicle crashes on highway entities".

Both the Poisson and NB models can include an exposure variable that is entered into the model to represent the opportunity for the event to occur, such as length of time, population size or geographical area. In statistical software the natural log of the exposure variable is entered as an offset (Hilbe, 2008). In the case of segment-based studies with variable segment lengths, the segment length can be included as the exposure variable, meaning that the accident count of a segment is proportional to its length - the longer a segment the more opportunity there is for an accident to occur (Aguero-Valverde and Jovanis, 2008). The length of road has also been used as an offset in area-level studies (e.g. Wedagama et al. 2006). For studies analysing motorised vehicle accidents a better exposure variable would be one that also accounts for the number of vehicles using the road segment, such as the annual vehicle kilometres travelled for a segment of road (segment length in kilometres multiplied by the vehicle count) (Haynes et al., 2008).

Due to the lack of available data, controlling for NMT road user exposure is typically much more difficult, except for small studies where it is possible to

collect the detailed data required (Graham and Stephens, 2005). Typically, researchers introduce proxy variables to the model which they hope will account for some of the variation in exposure, such as resident population (e.g. Graham and Stephens, 2008), whilst others ignore exposure and note it as a study limitation (e.g. Warsh et al., 2009). Qin and Ivan (2001) note that population density may not correlate well with pedestrian activity, for example in popular tourist areas there can be many more pedestrians than population density would suggest, and in areas with high population density and also high vehicle ownership, pedestrian activity can be less than expected. Qin and Ivan (2001) consider models which use population as a proxy for exposure to be "intrinsically unreliable", and their research found the correlation between population density and exposure to be non-significant.

2.6. Spatial autocorrelation issues

Spatial autocorrelation is said to occur when events or event attributes in geographic space display nonrandomness. When events are more clustered, or nearby events have attributes more similar, than would be expected from complete spatial randomness alone, there is said to be positive spatial autocorrelation. Negative spatial autocorrelation occurs when events are more dispersed, or nearby events have more dissimilar values, than would be expected (Fortin and Dale, 2009).

The presence of positive spatial autocorrelation in count data is problematic as the Poisson and NB regression models, in common with OLS regression, assume that the total count for a study entity (e.g. ward, district or segment) during a specific period is independent of counts in neighboring or nearby entities (Quddus, 2008). If the explanatory variables within a regression model do not account for the spatial clustering, i.e. the regression residuals display positive spatial autocorrelation, then tests of significance for regression coefficients may be biased leading to Type I errors, shifts in coefficient sign leading to mistaken inferences

(Kissling and Carl, 2007), and inflated goodness-of-fit measures (Haining, 2009). The risk of type I errors is a particular problem when coefficients are close to the significance threshold (Haining et al., 2009).

Spatial regression models have been proposed to address the spatial autocorrelation issue, and those that extend the standard OLS model are the most developed and most accessible to researchers. They include the simultaneous autoregressive (SAR) models which assume that the dependent variable at a location, as well as being a function of the explanatory variables is also a function of the dependent variable at neighboring locations. This relationship is incorporated into the model as an additional parameter through the use of a spatial weights matrix which defines the neighbours and, if required, can also weight the neighbours (so that nearer neighbours are considered more important) (Kissling and Carl, 2007). Two commonly used SAR models are known as the spatial lag model and the spatial error model and these are implemented in the software tool GeoDa, developed by Luc Anselin at Arizona State University (Anselin et al., 2006).

The methods discussed above are only suitable for continuous data, they are not appropriate for non-negative random count data (Quddus, 2008). Haining et al. (2009) note that methodology for analysing discrete spatial data "remains undeveloped" and although they have proposed spatial Poisson and NB models that appear to work well, they observe that these are not easy to apply. Other alternative models for count data include Bayesian hierarchical methods that have been explored by several researchers (e.g. Quddus, 2008; Aguero-Valverde and Jovanis, 2008). However, it is perhaps reassuring to note that in a comparison of spatial and non-spatial models of London crash data (an area-level study), Quddus, 2008 found that the non-spatial NB models and the Bayesian hierarchical models "gave quite similar results in many cases".

Probably as a result of the complexities involved in spatial count-based mod-

els, some studies have attempted to take account of spatial effects by including proxy variables within NB models. For example, Wang et al. (2009a) attempted to address unobserved regional variation by included a dummy variable for each region of England, and found all but one region to be statistically significant in all models.

It is also worth noting that although in some studies NB regression residuals are tested for spatial autocorrelation using standard Moran's I (e.g. Quddus, 2008; Gruenewald et al., 2009), this test is only suitable for use with linear regression models (Lin and Zhang, 2007). An extension of the Moran's I test suitable for testing residuals of generalized linear models (which includes the Poisson and NB models) has been proposed by Lin and Zhang (2007).

2.7. Research objectives

The overall aim of this study is to explore the relationship between NMT road casualties in non built-up areas and a range of explanatory factors that can be associated with the road network. The specific research objectives, as informed by the literature review, are as follows: to create an appropriate segment-based dataset for the road network in England and Wales to be used as the basic spatial unit (BSU) in subsequent analysis; to assign each NMT casualty that was reported in England and Wales during 1999-2008 and that occurred outside of a built-up area to a BSU and generate aggregate casualty counts for each BSU; to assign a range of relevant explanatory variables to each BSU; to develop a series of disaggregated NB regression models for each type of NMT road user and each injury severity; to explore options for assessing the impact of spatial autocorrelation on the regression estimation results, given that this is known to be difficult with count-based models; to identify casualty hot zones for each NMT road user type using the kernel density estimation method; to identify the key findings that can inform future road safety strategy for non built-up NMT casualty reduction; and finally to suggest areas for future study.

3. Data and methodology

This study required data from a range of sources to be brought together and manipulated, primarily using the ESRI ArcGIS package and several third-party add-on tools to provide additional functionality. Because of the large datasets and complex spatial queries that were required, where possible all datasets were imported into a single ArcGIS file geodatabase, for ease of storage and management, and to obtain performance and optimization benefits.

In summary, the methodology for this study consisted of the following key steps: reported NMT road casualties were aligned to a polyline feature class representing the road network using a snapping process; the road polyline feature class was divided into nominal 250m sections and additional data attributes representing potential explanatory factors were assigned to each segment; a count of the number of casualties per segment was used as the dependent variable in a series of NB regression models; the kernel density estimation method was used to create a casualty density surface for the study area to enable casualty hot zones to be identified and classified; and road segments with their centroid located within a hot zone were selected and the coincidence of casualties with these hot zone segments was explored. The data required and analysis techniques developed for each of these steps is considered in detail in the sections that follow.

3.1. Boundary data

English administrative counties and Welsh unitary authorities, derived from 2001 Census Boundary Data, were obtained form the EDINA UKBORDERS website in ESRI shapefile format and merged into a single polygon feature class. A shapefile of built-up area polygons was obtained from the Office for National Statistics. This dataset, known as the "Urban area and settlement boundary CD", identifies areas of land of urban character which extend to at least 20 hectares and have a resident population of at least 1,500 - based on Ordnance Survey 1:10,000

scale mapping and population data derived from the 2001 census. Land is considered urban in character if permanent structures are present, and includes roads which have built-up land on one or both sides, airports, motorway service areas and car parks. Playing fields and golf courses are not considered urban unless fully enclosed by built-up land. Areas of urban land that are less than 200m apart are combined together into a single polygon. A polygon layer representing non built-up areas was created by applying the ET Geowizards "Erase" tool to the local authority boundary layer, using the built-up polygons as the erase layer (ET SpatialTechniques, 2009).

3.2. Road data

The road network contained in the Ordnance Survey Meridian 2 vector dataset was chosen for this study, which was obtained in ESRI shapefile format from the EDINA ShareGeo service. The shapefile consists of separate polylines for each road class - motorways, A-roads, B-roads, and minor roads. The road network in Meridian 2 is sourced from the Ordnance Survey Roads Database and is derived from high resolution mapping (1:1,250 in urban areas, 1:2,500 in rural areas, and 1:10,000 in moorland). The road centreline is generalised using a 20m lateral filter, although this does not affect the positional accuracy of retained node points. The road network in Meridian 2 is not complete as it excludes minor roads <200m in length, private roads, and tracks.

Alternative approaches for choosing the segment length for studies of this nature were discussed in Section 2.3.2. Given the size of the study and the amount of road data involved it was not practical to derive homogeneous variable-length segments. Instead, it was decided to aim for a nominal fixed segment length of 250m, on the basis that this would be short enough to limit the extent of heterogeneity within each segment, long enough to avoid the issues associated with segments that are too short, and would ensure that all the explanatory factor measures were meaningful, particularly with respect to sinuosity. Previous

segment-based studies have used a fixed length of 200m (Parida et al., 2006), and although Koorey (2009) used a variable-length method, the resulting average segment length was 250m.

A series of processing steps were carried out to obtain a suitable segment dataset which could then be used as the basis for aggregating the casualty counts and for assigning the explanatory factor attributes. The four road class poylines were merged into a single polyline and the road segments within Scotland were removed. Using the ArcGIS "clip" tool and the non built-up polygon layer as the clip feature, the road segments within built-up areas were removed. In order to maximise the size of the link segments prior to to reducing them to the 250m target length, the ET Geowizards "Clean Pseudo Nodes" tool was used. This tool removes nodes between non-intersecting link segments (known as pseudo nodes) when a specified attribute value (in this case road class) is the same for both segments. Topology is preserved in the process and no regular nodes are removed. Next, the ET Geowizards "split polyline" tool was used with a target length set to 250m and configured to split segments into equal lengths to avoid creating many small segments from the remainders[8] (Figure 1). The frequency histograms in Figure 2 show how segment lengths changed as a result of this process.

Before running the regression models (see Section 2.5), motorway segments and all segments less than 10m in length were removed from the dataset. Prior to clipping the roads polyline with the non built-up polygon layer there were no segment lengths less than 1m, so these very small segments were an artefact of that process and appeared around the edges of the built-up areas and their removal could be justified. Furthermore, as the casualty location is at best accurate to the nearest 10m, including road segments shorter than this in the models could be problematic, as it would not be appropriate to assume that the characteristics of

[8]For example, a segment 800m long would be split into 3 segments of 266.67m .

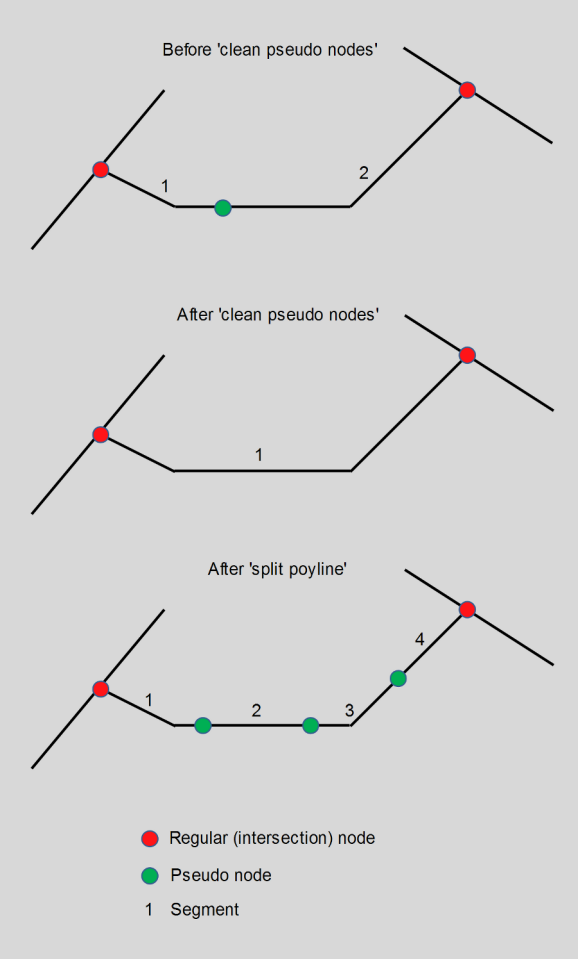

Figure 1: Illustration of the “clean pseudo nodes” and “split polyline” processes.

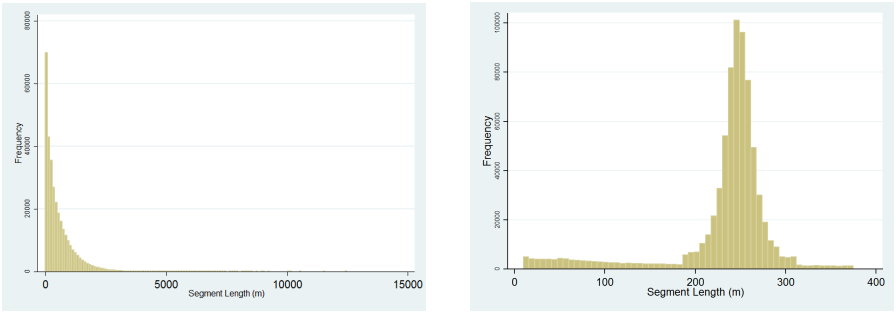

Figure 2: Frequency histogram of segment length of non built-up roads before (left) and after (right) processing for nominal 250m length.

that short segment were representative of the location where the casualty event occurred. Figure 3 summarises all the adjustments made to the road segment dataset.

3.3. Casualties

Police forces in England and Wales follow a standard procedure for recording each road traffic accident that occurs on a public road (including footway) where at least one road vehicle[9] and one casualty is involved and which comes to their attention within 30 days. The data returns are known as STATS19 and include some 50 variables for each qualifying accident, usually compiled by an attending police officer but sometimes reported to the police at a later time (for example when reported by a member of the public at a police station). Accidents which involve no personal injury or which occur on private roads or car parks are excluded. Accidents which involve pedal cycles or ridden horses on a public road and where no motor vehicle or pedestrian is involved are included in the STATS19 data (Department For Transport, 2009b, 2004).

STATS19 data for the UK for each year 1999-2008 (the most recent available) were obtained from the UK Data Archive. For each year the accident data consisted of three tab delimited files, the first with details of the accident, the second with details of each casualty for each accident, and the third with infor-

[9]In this context road vehicle includes pedal cycle and ridden horse.

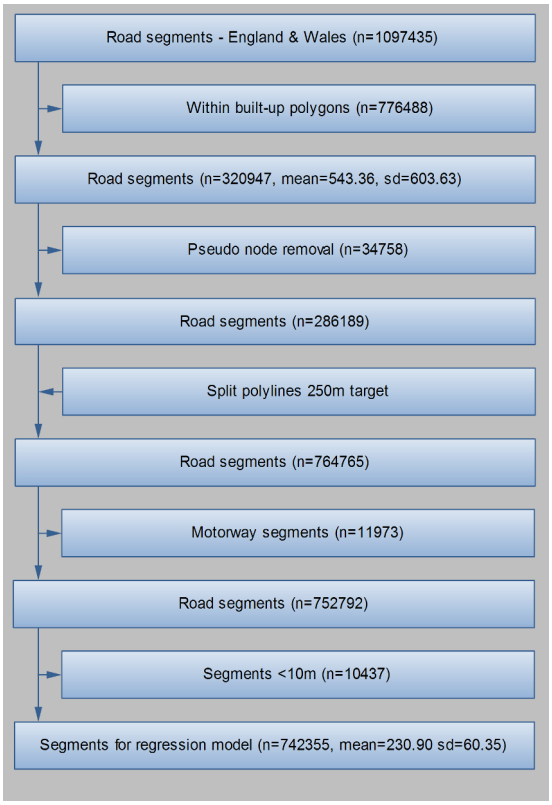

Figure 3: Adjustments made to the road segment dataset (n=total, mean=mean length in metres, sd=length standard deviation).

mation about the vehicle each casualty was either occupying or, in the case of pedestrians, first hit by. All the files were imported into Microsoft Access and combined into a single relational database with separate tables for accidents, vehicles and casualties. The total record count for each database table is shown in Table 1.

Table	Total records
Accidents	2081584
Casualties	2836060
Vehicles	3821564

Table 1: Total STATS19 records 1999-2008

The casualty table was found to have multiple casualty records for 117 accidents where the accident record indicated only one casualty, and these were removed. Queries were then run to create a new table for each NMT casualty type (pedestrian, cyclist and horse rider) containing a separate record for each casu-

alty and drawing relevant fields from the accident, casualty and vehicle tables. Several data cleansing steps were then completed: removing Scottish records; recoding local authority codes for Metropolitan Police districts outside London[10]; and modifying the easting and northing National Grid Reference (NGR) fields to convert them to a suitable format for use in ArcGIS[11]. The tables were then exported as dBase files, plotted in ArcGIS and then saved as point feature classes in the geodatabase. When the casualty data was displayed in ArcGIS it was immediately apparent that a number of points were outside the extent of the England and Wales local authority boundary layer, as shown in Figure 4. Examination of some of the outlying points indicated various causes, including: no NGR provided; ten figure NGR entered but without grid square identifier digits; and for South coast locations zero or blank not placed in the grid square identifier digit of the northing reference and instead 4 location digits and a final zero digit entered. However, the problems were sufficiently diverse and non-standard to make any attempt to correct the issue too time consuming and therefore the decision was made to exclude these points from the study.

In view of the problem with outlying points it was decided not to assume that the NGRs of the points falling within the England and Wales boundary were correct, and a quality control check was introduced. For the period 1999-2008 the local authority codes used in the STATS19 data were consistent and could be directly mapped to the 2001 local authority boundary dataset. By adding the STATS19 local authority codes to the 2001 local authority boundary dataset and running an overlay intersect it was possible to select those casualties that overlaid the correct local authority. Rather than exclude all the non-matching

[10]For example, the Metropolitan Police area covers part of the Epping Forest District Council area and is included as a separate local authority code in the STATS19 accident record - this code was amended to match the main Epping Forest District Council code.

[11]The easting and northing grid reference fields in the STATS19 data each have 5 digits and a 10m resolution, with the first digit in each identifying the grid square. ArcGIS requires each to have six digits with a 1m resolution and this requires a zero suffix to be added, which can be achieved by multiplying by 10.

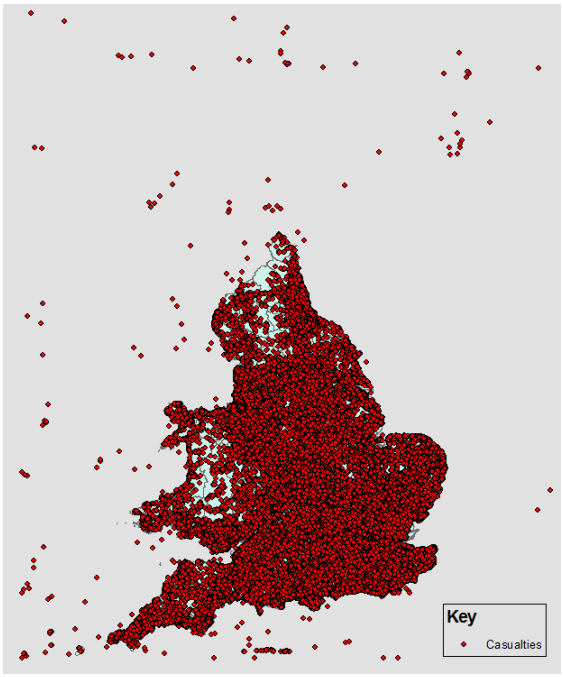

Figure 4: All non-motorised casualties 1999-2008 overlaying England & Wales boundary layer

casualties, some of which might be reasonably close to overlaying the correct local authority[12], a spatial join was used to identify all the local authorities within 500m of each non-matching casualty. If a matching local authority was found within 500m, then the casualty was not excluded from the dataset.

The next stage required the position of each casualty point to be adjusted so that it was coincident with a road polyline. The "Global Snap Points" tool in ET GeoWizards was used for this purpose. To reduce the amount of data to be processed any casualties falling within a built-up area polygon were first removed from the dataset. To improve accuracy of the snapping process the road class variable in the STATS19 data was utilised so that casualties were only snapped to the corresponding class of road[13]. Casualties which were within 300m[14] of

[12]The attending police office may have assigned it to the wrong authority if the accident was fairly close to the boundary, or because of the NGR resolution (nominally +/- 10m) the point may overlay the wrong authority.

[13]For example, a casualty recorded as having occurred on an A-class road would be snapped to the nearest edge of the A-road polyline feature class.

[14]300m was considered a reasonable compromise - it was sufficient to allow for road width on the ground not represented by the polyline, allowed a margin of error for manual or GPS-based NGR readings, and exploratory analysis had indicated that some 98% of the casualties would be

a road polyline of matching class were initially snapped. Any casualties that were not snapped during this first stage were then snapped to a road of any class within 300m[15]. Finally, any casualties that were within a built-up area polygon as a result of the snapping process were removed from the dataset.

Casualty counts for each segment in the segment dataset were generated using spatial joins, with separate counts for total casualties, each casualty type (pedestrian, cyclist, horse rider), and also by injury severity (fatal, serious, slight). One segment with a casualty count of 104 was an obvious outlier, and on investigation the casualties coincident with this segment were found to relate to a variety of road classes and road numbers, suggesting that the NGR had been used as a bucket code. Three casualties were identified as correctly assigned to this segment and the remainder were removed from the casualty dataset and the count adjusted accordingly. It was observed that any casualty point which was coincident with a regular node (intersection) contributed to the casualty count for each segment connecting at that node, resulting in multiple counting of casualties. This raised the casualty count by 1,335, but it was considered valid for the affected casualties to be associated with each of the segments meeting at the intersection. In addition, as mentioned in Section 3.2, motorway segments and segments less than 10m in length were removed from the segment dataset prior to regression analysis, along with their associated casualty counts. Figure 5 summarises all the adjustments made to the casualty data point totals.

3.4. Explanatory factors

3.4.1. Footpath and bridleway crossings

The Ramblers (2003) have produced a report which identifies some 1000 locations where public footpaths or bridleways are severed by fast and busy roads.

snapped at this distance.

[15]Examination of the data showed that some NGRs were clearly correct but the road class did not match between STATS19 and the Meridian road data.

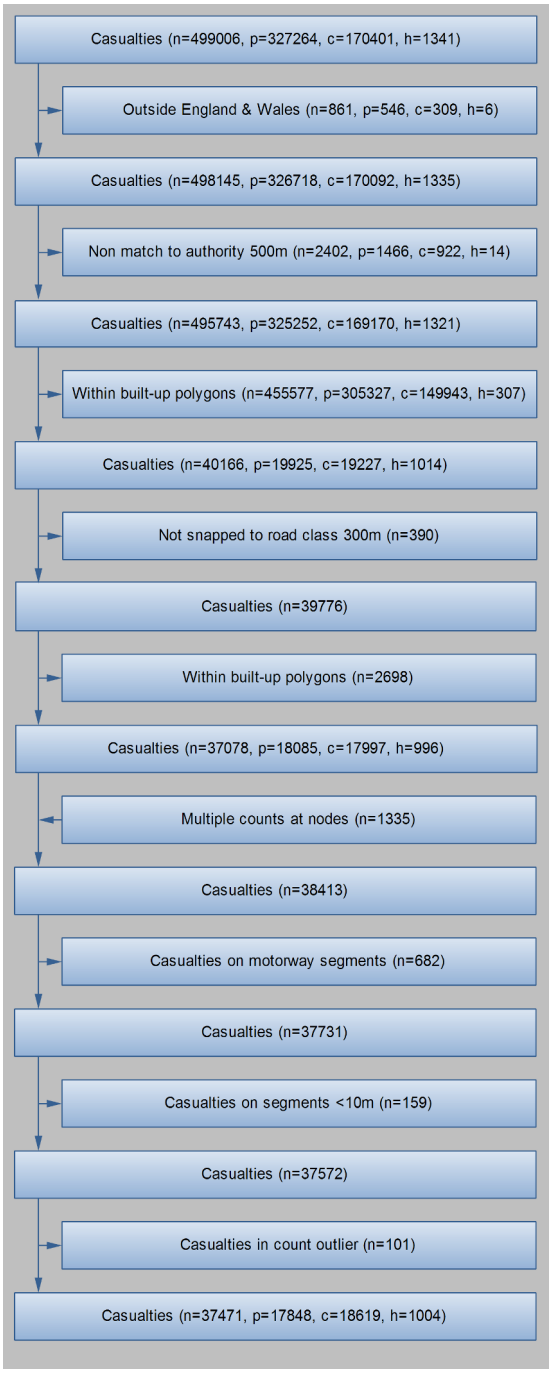

Figure 5: Adjustments to casualty data point totals (n=total, p=pedestrian, c=cyclist, h=horse rider).

They argue that these crossing points pose a risk of serious injury or death to pedestrians or horse riders and are campaigning for improvements to be made. As it was not possible to obtain this data in electronic format, it was manually transcribed from the report into a spreadsheet. The vast majority of the supplied NGRs were six figure (100m resolution), with a few eight figure (10m resolution), but none of them identified the 100km grid square which they related to. Digital mapping software was used to manually identify the correct grid square for each crossing based on the supplied location and road information, and the NGRs were converted into a format suitable for use in ArcGIS. Crossings in the report which did not have NGRs or which referred to paths severed by motorways (so use of the crossing no longer permitted) were excluded, giving a total of 825. The next stage required the position of each crossing point to be adjusted so that it was coincident with a road polyline. The "Global Snap Points" tool in ET GeoWizards was used for this purpose. Crossing which were within 500m of a road polyline of matching class were initially snapped. Any crossings that were not snapped during this first stage were then snapped to a road of any class within 500m. Of the 806 crossings that were snapped to a road polyline, 738 were identified, via a spatial join, to be coincident with a segment in the segment dataset.

3.4.2. National Trails

National Trails are long distance routes primarily for walkers, but with some sections suitable for horse riders and cyclists. The exception is the Pennine Bridleway which is suitable along its entire length for walkers, cyclists and riders. Whilst much of the trail routes are on footpaths, bridleways or byways, all of them will involve road crossings as well as sections of varying length that use the public road in non built-up areas where pavements are unlikely to be provided. The routes of 13 of the 15 National Trails in England, plus the Offa's Dyke path which is shared with Wales, were obtained in ESRI polyline shapefile

format from Natural England. The routes had been digitised from Ordnance Survey 1:25,000 raster maps with a ± 1mm accuracy in 2000 (nine trails) and 2005 (four trails), with no subsequent updates. Datasets for the other Welsh National Trails, Glyndŵrs Way and Pembrokeshire Coastal Path, were not available.

In order to establish which sections of road were used by the trails, the "Global Snap Polylines" tool in ET Geowizards was used, configured to insert vertices into the trail polylines and to snap to the nearest vertex or edge of the road polyline. In this way the snapped part of the trail polyline is entirely coincident with the relevant section of road. A process of trial and error established that a tolerance of 60m produced the most accurate and consistent results (Figure 6).

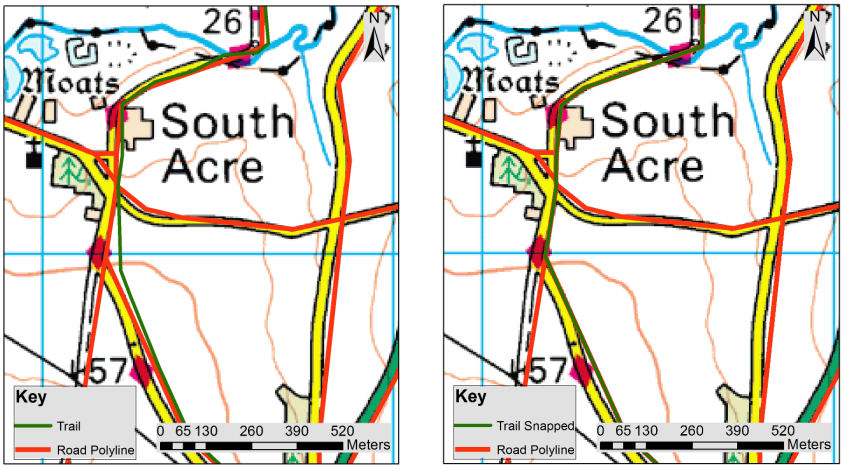

Figure 6: Example of snapping the trail polyline to the road polyline, showing before (left) and after (right), and overlaying Ordnance Survey 1:50,000 raster mapping.

A spatial join was used to identify whether or not a segment in the segment dataset had a coincident National Trail. It should be noted that this process will identify road segments where a trail crosses the road as well as those where a trail follows the line of the road. This is also useful information, as it records locations where pedestrians potentially interact with the road network (albeit that the presence of footbridges, underpasses, or pavements is unknown).

3.4.3. National Cycle Network

The National Cycle Network (NCN), coordinated by Sustrans, consists of more than 12,000 miles of cycle routes in the UK, some traffic free and some on-road. The NCN is divided into national routes, regional routes and local links into the national routes. For this study only the national and regional routes were considered, and these were obtained in ESRI polyline shapefile format direct from Sustrans. The routes had been digitised from Ordnance Survey 1:50,000 raster maps. In the national route dataset a field was present to identify whether line segments were on-road or off-road, allowing the on-road segments to be extracted, and the off-road segments discarded. After routes in Scotland were removed, the "Global Snap Polylines" tool in ET Geowizards was used, configured to insert vertices into the route polylines and to snap to the nearest vertex or edge of the road polyline. In this way the snapped part of the route polyline is entirely coincident with the relevant section of road. A process of trial and error established that a tolerance of 60m produced the most accurate and consistent results. A spatial join was used to identify whether or not a segment in the segment dataset had a coincident cycle route. It should be noted that this process will identify road segments where a cycle route crosses the road as well as those where a cycle route follows the line of the road.

3.4.4. Steepness

It has been suggested that cyclists might be at greater risk when travelling uphill in steep areas due to a greater speed differential between them and motorised vehicles (Sustrans, 2009). In order to explore this aspect, a supplementary dataset to the Ordnance Survey ITN layer was obtained which identifies road links in the ITN which are steep (14-20% gradient) or very steep (≥20% gradient). This data was supplied in comma-separated values (CSV) file format with each record identifying a single point representing the steepest part of a steep ITN road link. The data was imported into ArcGIS and after points in Scotland

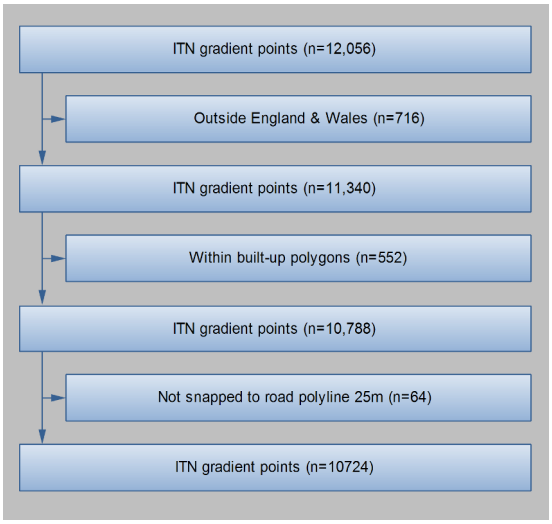

Figure 7: Adjustments to gradient data point totals (n=total).

or within a built-up area polygon had been removed, the "Global Snap Points" tool in ET Geowizards was used to snap the points to the nearest road polyline edge - based on a 25m search tolerance[16]. The adjustments made to the gradient point data are summarised in Figure 7. A spatial join was used to identify whether or not a segment in the segment dataset had a coincident point indicating a steep or very steep gradient.

3.4.5. Intersections

Exploratory analysis of all NMT casualties outside of built-up areas showed that 40% occurred within 20m of a junction, with the proportion increasing to 52% when considering cyclist casualties alone. With other research showing a significant positive relationship between junctions and NMT casualties (Wang et al., 2009a), it was important that junctions were included as an explanatory factor in this study.

Using the ET Geowizards "Export Nodes" tool it was possible to export each

[16]The ITN data has high positional accuracy and contains additional roads (mainly minor) not included in Meridian. The Meridian centreline is generalised to within 20m of real world position, therefore a 25m search tolerance for snapping the ITN gradient points should snap the gradients to the correct road. A larger tolerance would risk snapping the gradient point to the wrong road.

node in the segment dataset to a point feature class with the type of node (pseudo, regular or dangling) identified. The regular nodes occur where three (or more) poylines meet and represent intersections on the road network, although it should be noted that not all of these will be genuine intersections as roads that appear to intersect in the polyline layer may actually pass over or under each other. Using a spatial join the number of regular nodes per segment was counted, with three possible values: zero (segment has no intersections); one (segment has one intersection); or two (segment has two intersections) (Figure 8).

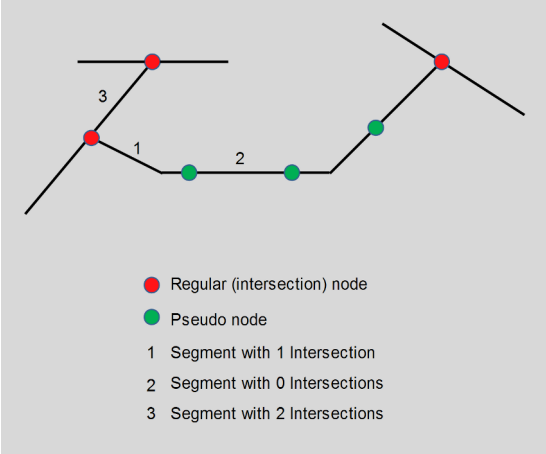

Figure 8: Illustration of segment intersection identification.

3.4.6. Sinuosity

A number of studies have found a negative correlation between NMT casualties and measures of road bendiness, perhaps due to lower traffic speed or more careful driving on these roads compared with long straight sections (e.g. Wang et al., 2009a; Berhanu, 2004). However, these studies do not specifically examine NMT casualties on non built-up roads where the relationship may be different. Slow moving pedestrians or horse riders in the roadway and not separated from traffic could be particularly at risk on bends where visibility is reduced, for example roads lined with tall hedges.

The measure of road bendiness used in this study is the sinuosity index, sometimes referred to as the detour ratio, which is calculated by dividing the

total length (Lt) of the segment (i.e. the road distance between two nodes) by the shortest (aerial) distance between the two nodes (Lsd) (Figure 9). If a road segment is straight then the sinuosity index will equal one, with the index value increasing as segment curvature increases. The sinuosity index was calculated for each segment in the segment dataset using the "Line Metrics" tool in Hawth's Analysis Tools (Beyer, 2004). It was found that in some cases the sinuosity index was given a value of zero, and this occurred when a segment formed a complete loop with the start and finish nodes being coincident. With zero distance between nodes the denominator for the index calculation will be zero and cause a divide by zero error, and to avoid this error the "Line Metrics" tool returns a zero result instead. The highest sinuosity value for the dataset was 60, and it was decided to replace the zero values with 100.

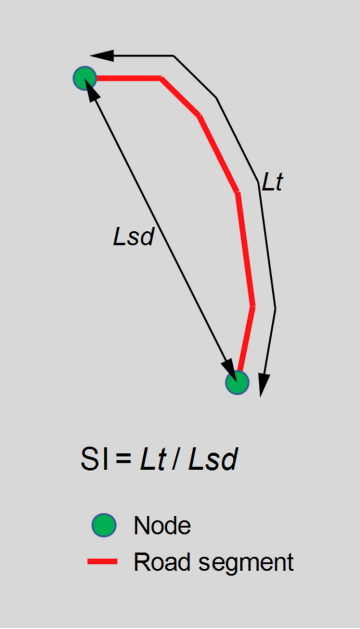

Figure 9: Calculation of Sinuosity Index (SI).

It should be noted that this measure of road bendiness can only take account of changes of direction that occur at vertices. If a change of road angle occurs at a regular (intersection) or pseudo node it will not be identified by this method.

3.4.7. Traffic flow

Different levels of traffic flow may be a contributory factor in road accidents involving NMT users. The opportunity for an accident to occur - the exposure to risk - may be influenced by the amount of motorised traffic using a road. Sufficiently detailed traffic flow data was only available for major roads (A-class and motorways), and this was obtained from the DfT for the years 1999-2008, to match the time period of the STATS19 data. The flow data, which was provided in CSV files by region and then combined into a single database, consisted of 16,202 unique count points, each with a reference number and a ten digit NGR identifying its position on the road network. For each count point for each year (that the count point was in use) an annual average daily flow (AADF) value was provided for all motor vehicles (included pedal cycles) along with disaggregated totals for each vehicle type. The AADF is an estimate of the average number of vehicles which pass a count point each day, and is derived from manual counts taken on every major road link[17] and data from a small number (less than 200) of automatic counters (Road Traffic Statistics Branch, 2007).

There were 12 count points with zero AADF values, and these were removed from the database along with the motorway points, leaving a total of 15,227. A new AADF for all motorised vehicles was calculated (excluding pedal cycles) and the data was then grouped by count point reference number and an average AADF calculated for the ten year period for each count point. This database was then imported into ArcGIS and the count points were snapped to the A-class roads polyline using the ET Geowizards "Global Snap Points" tool with a 300m search tolerance (115 points did not snap at this tolerance and were removed).

As there were 15,112 count points, but over 107,000 A-class segments in the segment dataset, it was necessary to develop a method to derive an AADF value

[17]Links are usually a section of road between consecutive junctions with other major roads. Manual counts are taken between every 1 and 8 years.

for every segment. Wang and Kockelman (2009) note that the standard solution is to assign the AADF value of the nearest count point, but found a spatial interpolation technique using Ordinary Kriging to be more reliable. Ideally these techniques would take into account that roads are constrained to a network and make use of network distances rather than Euclidean distances. However, such techniques are computationally very intensive and not readily available. The SANET toolbox (Okabe and Satoh, 2009) includes a network interpolation tool but unfortunately this was not able to cope with the size of the dataset used in this study.

To select the most appropriate estimation method, the count point dataset was randomly split into two samples - an 80% sample of the count points to be used in the estimation procedure, and the remaining 20% to be used to compare estimated AADF with known AADF. For the spatial interpolation method, the ArcGIS "Geostatistical Wizard" was used to run the Ordinary Kriging procedure with the 80% sample as the dataset to be modelled, and the 20% sample as the validation dataset. The AADF values were log transformed as this gave a better fit to the normal distribution as shown by a Q-Q plot, and different parameter combinations were explored to optimise the model. The selected model gave a root-mean-square error of the predicted values of 12,090. For the nearest count point method, segments from the A-class road polyline with a coincident count point from the 20% sample were selected and exported to a new feature class. Two spatial joins were then run, the first to join the AADF value from the coincident count point[18] (the observed value), and a second to join the closest count point from the 80% sample (the predicted value). The root-mean-square error of the predicted values in this case was 14,375. As the root-mean-square error of the Kriging method was 15.9% lower than the nearest count point method, this

[18]The join method was set to mean in case more than one count point was coincident with a segment.

was chosen as the estimation technique and the same Kriging model was run on the full count point dataset.

To assign an AADF value from the Kriging surface to each A-class road segment in the segment dataset, the ET Geowizards "Polyline Coordinates" tool was used to calculate the centroid of each segment as a new point feature class. A prediction was then run using the Kriging surface with the centroid points as the input data, and a join was then used to associate a predicted AADF value with each segment.

3.4.8. Distance from built-up area

Figure 10 shows fatal NMT casualties in non built-up areas plotted on a map of England and Wales. It indicates that these casualties are clustered around the built-up areas[19]. In order to examine this relationship using the regression analysis, a method to measure segment distance from built-up areas was developed. This variable would also enable the model to account for this component of spatial autocorrelation.

Using the point layer of exported nodes (see Section 3.4.5), the distance from each node to the nearest built-up area polygon was calculated using the "Point Distance" tool in ET Geowizards. A spatial join was then carried out using the segment dataset as the target and the node points layer as the join feature, with the join configured to calculate the mean of the node distance. Each segment has a node at each end (whether regular, pseudo, or dangling), so this procedure obtains the distance each segment node is from the nearest built-up polygon and then takes the mean of the two distances, thus giving an indication of the segment's distance from the built-up area (Figure 11).

[19] A similar pattern is observed for all casualties, but only fatal are shown here for map clarity.

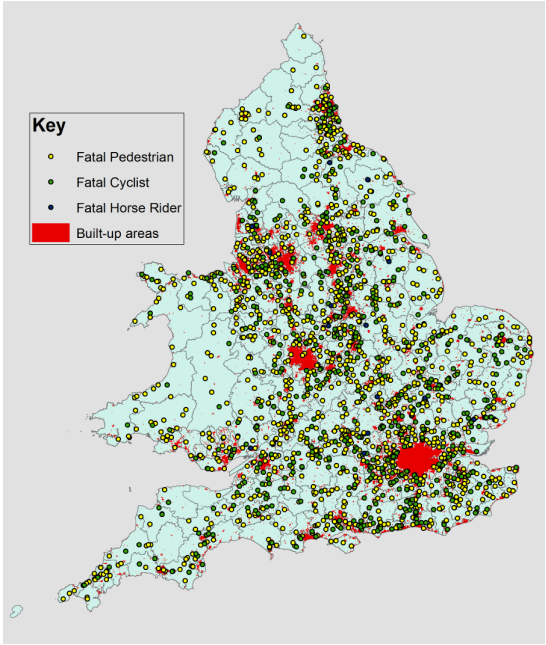

Figure 10: Fatal non-motorised casualties in non built-up areas.

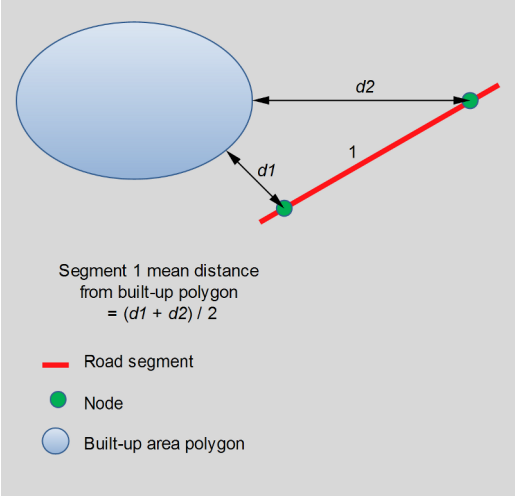

Figure 11: Illustration of segment distance from built-up area calculation.

3.4.9. Population

Local resident population was included in the regression models as a proxy for NMT road user exposure. Population estimates based on the 2001 census at the local authority district or unitary authority level were obtained from the Office for National Statistics, and the data joined to the local authority boundary layer. In order to assign this data to the road segments, the mid point coordinates of each segment (i.e. the centroid) were first calculated using the "Polyline Co-ordinates" tool in ET Geowizards and these were then exported to a new point feature class. Using a spatial join, each segment centroid was assigned a population attribute based on the local authority urban polygon it was within. Then, using a further spatial join, the population attribute from the centroid point layer was joined to the segment dataset. Thus, each segment had assigned to it the population of the local authority that its centroid was within.

3.5. Statistical analysis

Exploratory data analysis and the development of regression models, carried out using the statistical software STATA v11, is discussed below. This is followed by consideration of spatial autocorrelation issues and a proposed methodology to account for this in the regression models.

3.5.1. Regression models

As discussed in Section 2.5, count data is unlikely to have a normal distribution and is usually right skewed. Exploratory analysis of the casualty count for each segment confirms this to be the case (skewness = 8.65, kurtosis = 137.05), as shown in Figure 12 and Table 2. Clearly OLS regression is not appropriate, and a model suitable for a discrete response variable is required, such as Poisson or NB. The mean segment casualty count is 0.05 with a variance of 0.08. As the variance exceeds the mean (1.6x) this indicates the presence of overdispersion and suggests that the NB model is likely to be preferred over the Poisson model.

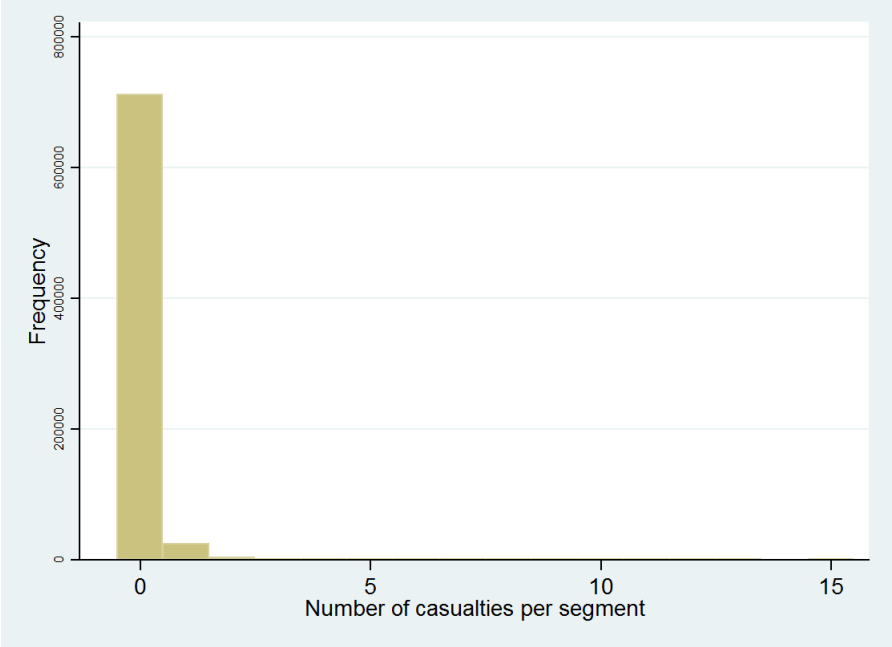

Figure 12: Histogram of segment frequency for each casualty count.

The NB model includes an additional parameter, known as alpha, to represent this overdispersion. An alpha of zero indicates no overdispersion and the NB model reverts to the standard Poisson model. An alpha greater than zero indicates the presence of overdispersion, and higher values of alpha indicate more overdispersion (Coxe et al., 2009). An initial NB model was run to calculate alpha and to perform a likelihood ratio chi-squared test of alpha=0. This gave an alpha of 3.3 and a chi-squared value of 1.1e04 with one degree of freedom ($P<0.0001$), confirming the presence of overdispersion and rejecting the Poisson model.

To provide a visual comparison of the Poisson and NB models, the STATA "prcounts" function (Long and Freese, 2001) was used to calculate and then plot the difference between the observed probability and the mean predicted probability for each count (0-9) and for each model. As Figure 13 illustrates, the Poisson model does not perform well, in particular under predicting zero counts and overpredicting counts of one.

A disaggregated approach to the regression modelling was adopted, with sep-

Casualty count	Frequency	Percent
0	711925	95.901
1	25549	3.442
2	3562	0.480
3	861	0.116
4	268	0.036
5	96	0.013
6	53	0.007
7	17	0.002
8	11	0.001
9	6	0.001
10	1	<0.001
11	2	<0.001
12	2	<0.001
13	1	<0.001
15	1	<0.001
Total	742355	100.000

Table 2: Segment frequency for each casualty count.

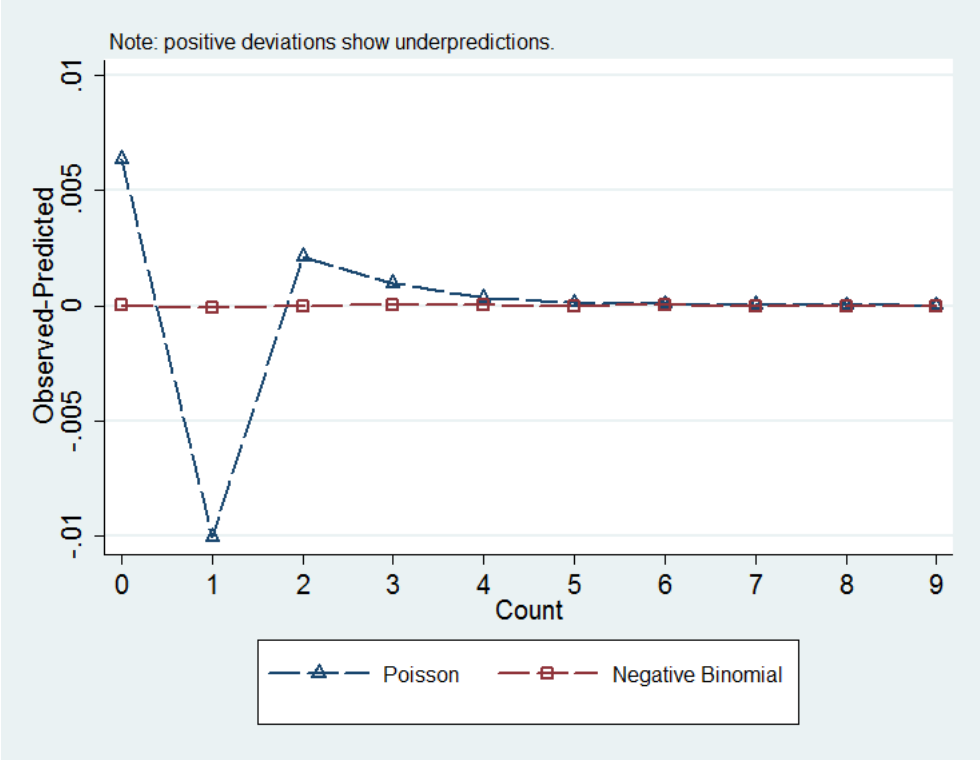

Figure 13: Plot of the difference between observed probability and predicted probability for Poisson and NB models.

Variables	Type	Obs.	Mean	SD	Min.	Max.
Dependent variables						
Total casualties	Discrete	742355	0.05048	0.27782	0	15
Total fatalities	Discrete	742355	0.00266	0.05298	0	4
Total serious injuries	Discrete	742355	0.01179	0.11624	0	7
Total slight injuries	Discrete	742355	0.03603	0.22687	0	11
Pedestrian casualties	Discrete	742355	0.02404	0.18154	0	10
Cyclist casualties	Discrete	742355	0.02508	0.18790	0	12
Horse rider casualties	Discrete	742355	0.00135	0.03913	0	4
Road characteristics						
A-class road	Dummy	742355	0.14434	0.35144	0	1
B-class road	Dummy	742355	0.09283	0.29019	0	1
Minor road (reference)	Dummy	742355	0.76283	0.42535	0	1
Sinuosity	Continuous	742355	1.05075	1.46694	1	100
Steep	Binary	742355	0.01421	0.11836	0	1
Number of intersections	Discrete	742355	0.57893	0.67348	0	2
Non-motorised user interactions						
National Trail present	Binary	742355	0.00667	0.08142	0	1
Sustrans route present	Binary	742355	0.06768	0.25119	0	1
Dangerous crossing present	Binary	742355	0.00099	0.03151	0	1
Sum of interactions	Discrete	742355	0.07535	0.26901	0	3
Demographics characteristics						
Population ('000s)	Continuous	742355	117.6295	69.6194	24.5	976.4
Spatial factors						
Distance from built-up area (m)	Continuous	742355	2276.708	2336.229	0	26831.04

Table 3: Summary statistics of variables used in the all-segment models.

arate models for fatal, serious, and slight injuries, as well as specific models for cyclists, pedestrians, and horse riders. In addition a separate model was generated for all casualties on A-class road segments to enable the motorised vehicle AADF explanatory variable to be be included. Tables 3 and 4 list the variables used in the models along with summary statistics.

Prior to running the models potential correlation between the independent

Variables	Type	Obs.	Mean	SD	Min.	Max.
Dependent variables						
Total casualties	Discrete	107155	0.14765	0.49351	0	15
Road characteristics						
Predicted AADF	Continuous	107155	13152.07	7285.139	1019.22	63096.11
Sinuosity	Continuous	107155	1.00794	0.30642	1	100
Steep	Binary	107155	0.00108	0.03288	0	1
Number of intersections	Discrete	107155	0.63801	0.69763	0	2
Non-motorised user interactions						
Sum of interactions	Discrete	107155	0.03679	0.19357	0	3
Demographics characteristics						
Population ('000s)	Continuous	107155	122.9018	76.16904	24.5	976.4
Spatial factors						
Distance from built-up area (m)	Continuous	107155	1715.673	2101.224	0	26827.86

Table 4: Summary statistics of variables used in the A-class road segment model.

variables was investigated. There were two correlations of note relevant to the A-class model - a negative correlation between the mean distance from built-up polygon and predicted AADF (R^2= 0.194); and a positive correlation between population and predicted AADF (R^2= 0.05). Of relevance to the all segment models, the highest correlation (negative) was between population and mean distance from built-up polygon (R^2= 0.04). None of these correlations was considered strong enough to warrant exclusion from the models.

To take account of the opportunity for a casualty to occur in a segment, the segment length was entered into the NB regression as an exposure variable with its coefficient constrained to one, effectively turning the segment count into a rate.

3.5.2. Spatial autocorrelation

To investigate the presence of spatial clustering or dispersion in the data, two approaches were adopted. For the casualty point dataset Ripley's K function was calculated using the CrimeStat application (Levine, 2004), and for the aggregated segment casualty counts Global Moran's I values were calculated for a range of k-nearest neighbours using ArcGIS.

Ripley's K function is able to indicate the presence of spatial clustering or dispersion at a range of distances. Conceptually it involves drawing a circle of set radius around a point in the dataset and then counting the number of other points that fall within the circle. This is then repeated for every other point in the dataset and the results summed. The circle radius is then incrementally increased and the process repeated. The end result is a K statistic for a range of distances, which can be plotted on a graph and compared with the K statistic that would be expected if the points exhibited complete spatial randomness (CSR). If the K statistic at a certain distance is higher than that expected from CSR then it indicates that points are clustered at that scale, and if the K statistic is lower than that expected from CSR then the points are considered dispersed (Levine, 2005).

There are a number of potential problems with comparing Ripley's K results for the casualty point data in this study with those from a standard CSR simulation. Firstly, the points occur on a road network and research has shown that points randomly distributed on a network are unlikely to appear random when the network structure is removed and the points are viewed as distributed on a continuous plane (Okabe and Satoh, 2009). Secondly, in this study the area under consideration has been deliberately restricted to non-built up areas, so comparing the casualty point distribution with points randomly assigned to the bounding rectangle would be fundamentally flawed. The ideal solution would be to generate random points across the non built-up road network and use this as the CSR for comparison. Whilst tools are available to generate random points on a network, for example within the SANET toolbox (Okabe and Satoh, 2009), these are unable to cope with the size of dataset used in this study, and in addition would not be able to assign random points to a disconnected network that is restricted to roads in non built-up areas. Instead, to provide an improved CSR comparator, a random set of points was generated in ArcGIS that was constrained solely to the non built-up area polygon layer and Ripley's K was then calculated for this random dataset[20] and the casualty point dataset and the results plotted (Figure 14). The results indicate that spatial clustering is occurring at all distances.

The Global Moran's I statistic reveals whether, considering the dataset as a whole, attributes in nearby areas are similar or dissimilar to one another. It can be thought of as measuring the correlation coefficient between a variable in one area with the average of that variable in neighboring areas (i.e. the correlation between a variable and its spatial lag) (Wang, 2006). The extent of neighbouring areas to be considered in the calculation can be based on distance, polygon contiguity or a specified number of neighbours, and these relationships are de-

[20]Ideally a large number (>100) of these random datasets would have been generated and then upper and lower limits derived, but time did not permit this.

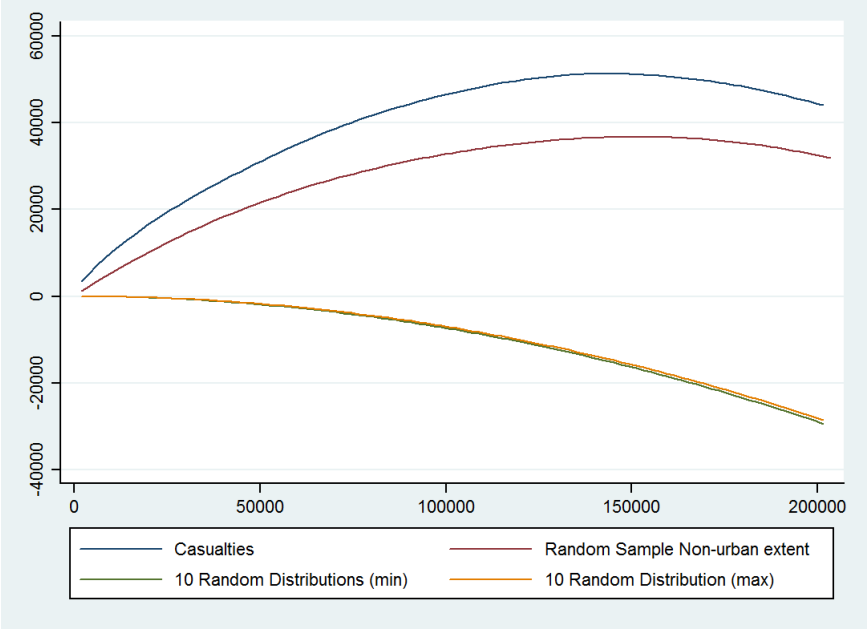

Figure 14: Ripley's K for casualties compared to CSR simulation and a random sample generated in the non built-up extent.

fined in a spatial weights matrix. The Moran's I statistic has a value that ranges from -1 (dispersed), through zero (random) to +1 (clustered). In the context of this study the areas are the segments and the attribute of interest is the aggregate casualty count for each segment. The segment polyline feature class was first converted to a point feature class based on the segment centroid which was calculated using the "Polyline Coordinates" tool in ET Geowizards. A series of row-standardized spatial weight matrices were created in ArcGIS for a range of k-nearest neighbours (2, 4, 6, 8, 10, and 12) based on euclidean distance, and these were used to calculate the Global Moran's I values. The results indicate that clustering is present and significant (P <0.00001), though not particularly strong, with Moran's I highest for two nearest neighbours (0.15) and then gradually declining (Figure 15).

The challenges posed by spatial autocorrelation and NB regression models were discussed in Section 2.6. Several studies in other fields, such as crime and socioeconomic disadvantage, have attempted to correct for spatial autocorrelation in count-based regression by introducing a spatial lag of the dependent

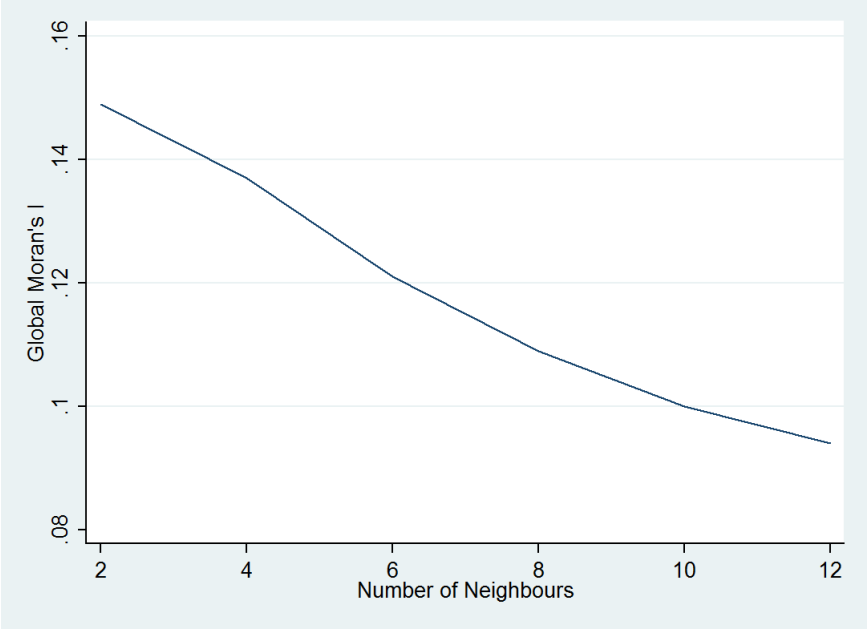

Figure 15: Global Moran's I for a range of k-nearest neighbours.

variable as an additional explanatory factor in the model (e.g. Akins et al., 2009; Kubrin and Weitzer, 2003; Nielsen et al., 2010; Hannon, 2005; Schaible and Hughes, 2008). Either the raw spatial lag is introduced (i.e. the spatial lag of the actual count), or by using a two-stage least squares technique known as the Anselin-Alternative method the spatial lag of predicted values from an initial regression is included in a final regression. Nielsen et al. (2010) note that including the raw spatial lag method produced very similar results to the two-stage method, and this simpler approach was adopted for this study. Unfortunately, the software readily available for creating spatial weights and calculating spatial lag variables was not able to cope with the size of the dataset used in this study[21]. Instead, a subset of the study area in East Anglia[22] was taken and separate all casualty regression models were run with an without the spatially lagged dependent variable to explore the effects. A spatial weights matrix based on two

[21]OpenGeoDa and two functions developed for STATA by P. Wilner Jeanty at Ohio State University (SPWMATRIX and SPLAGVAR) were investigated. ArcGIS is able to create a spatial weights matrix for a dataset this large but does not have a function for calculating spatial lag.

[22]Consisting of road segments within Peterborough Unitary Authority, Cambridgeshire, Suffolk and Norfolk.

k-nearest neighbours was used as the Global Moran's I analysis indicates that spatial autocorrelation is greatest at this scale. Table 5 lists the variables used in the East Anglia models along with summary statistics.

Variables	Type	Obs.	Mean	SD	Min.	Max.
Dependent variables						
Total casualties	Discrete	64774	0.03748	0.22982	0	8
Road characteristics						
Sinuosity	Continuous	64774	1.04572	1.51228	1	100
Steep	Binary	64774	0.00009	0.00962	0	1
Number of intersections	Discrete	64774	0.57000	0.66430	0	2
Non-motorised user interactions						
Sum of interactions	Discrete	64774	0.07994	0.27279	0	2
Demographics characteristics						
Population ('000s)	Continuous	64774	111.0921	23.07677	55.6	157.2
Spatial factors						
Distance from built-up area (m)	Continuous	64774	1733.217	1339.146	0	7490.8
Spatial lag of dependent variable		64744	0.03823	0.17491	0	6

Table 5: Summary statistics of variables used in the East Anglia models.

3.6. Hot zone identification

The kernel density method, as discussed in Section 2.4.1, was used to identify casualty hot zones based on the non built-up casualty point feature class. As others have noted, there is little guidance available on the choice of bandwidth and grid cell size, and the decision is rather subjective (Anderson, 2009). A cell size of 250m was considered appropriate, given that this matched the nominal segment size used in this study, with the bandwidth set at twice the radius (500m) as suggested by Anderson (2009). One simple approach to identifying accident hot spots would be to select the road segments with a casualty count in access of a defined threshold. The potential problem with this approach is that the boundary between segments is arbitrary and casualties may lay either side of the boundary resulting in hot spots being missed. The use of the kernel density method with a 500m search radius addresses this issue, whilst at the same time limiting the likelihood of casualties on different but nearby roads being included in the kernel - a drawback of using the Euclidean distance measure.

For each class of NMT user separately, a raster recording casualty count per

kilometre for each cell was created using the kernel density function in ArcGIS. In each case a new raster was calculated for values greater than zero, reclassified to change zero values to nodata, and then used as a mask over the original kernel density raster in the "Extract by Mask" tool, thus creating a raster of non-zero density values. This raster was then reclassified to define hot zone thresholds based on incremental multiples of the mean cell value (Eck et al., 2005), as listed in Table 6, and then converted to a polygon layer using the Spatial Analyst "Raster to Features" function. This polygon could then be used as a map overlay to highlight hot zones graphically. Finally, using a spatial join, each segment from the full segment dataset (752,792 segments) with its centroid inside a hot zone polygon was assigned the relevant hot zone threshold attribute value. For example, a segment might be in a pedestrian threshold four hot zone, a cyclist threshold three hot zone, but not in a horse rider hot zone.

	Casualty density per km (count equivalent per cell)	
Hot zone threshold	Pedestrians/Cyclists	Horse riders
1	Up to 8 (< 2)	Up to 4 (< 1)
2	8-16 (2 - < 4)	4-8 (1 - < 2)
3	16-24 (4 - < 6)	8-12 (2 - < 3)
4	24-32 (6 - < 8)	> 12 (≥3)
5	>32 (≥8)	n/a

Table 6: Hot zone classification thresholds.

4. Results and discussion

The results and discussion section begins with a review of the interpretation of NB model estimations and then presents the results of the models, discussing each explanatory factor and considering the impact of spatial autocorrelation with reference to the East Anglia subset. The findings of the hot zones analysis are then reviewed before a final section which highlights the potential limitations of the entire study.

4.1. Interpretation of NB model estimations

4.1.1. Model fit tests

A likelihood ratio chi-square statistic (LR chi^2) is used to assess the overall significance of each model. It tests the null hypothesis that the coefficients of all the independent variables are zero, and gives a probability (*P*-value) of obtaining a particular LR chi^2 value if the null hypothesis was true. If the null hypothesis is rejected, then we can be confident that at least one of the coefficients is not equal to zero and that the model itself is significant (UCLA: Academic Technology Services Statistical Consulting, 2010b).

As the NB regression model uses a maximum likelihood estimator, it does not have a "goodness of fit" test equivalent to R^2, which in OLS regression indicates the degree to which the independent variables explain the variance in the dependent variable. However, a number of "pseudo R^2" measures have been developed to help assess "goodness of fit" of NB models. Though these have a different theoretical basis from R^2, they share the same 0-1 ratio scale, with values nearer to one indicating a better model fit. In this study the Nagelkerke/Cragg & Uhler's pseudo R^2 has been used, which is calculated using the "fitstat" function within STATA. This ratio compares the log likelihood of the null model (intercept constrained to zero) with the log likelihood of the full model and indicates the extent to which the full model is an improvement over the null (UCLA: Academic Technology Services Statistical Consulting, 2010a; Long and Freese, 2001).

For all models the likelihood ratio chi-squared test of the dispersion factor (alpha) was checked to confirm that the NB model was preferred over the Poisson model.

4.1.2. Interpreting coefficients

In a NB regression the log of the expected count is modelled, and the coefficient is equal to the difference between the logs of the expected count when there

is a one unit change in the independent variable. For example, if the coefficient for the number of intersections is 0.59, for a one unit change in the number of intersections the log of the expected casualty count is estimated to change by 0.59. The coefficient sign indicates whether the independent variable is having a positive or negative effect on the dependent variable. To make interpretation easier, a coefficient can be expressed as an incidence rate ratio (IRR), and this is calculated by exponentiating the coefficient. Using the earlier example, for a coefficient of 0.59 the IRR for a one unit change would be exp(0.59) = 1.80, meaning that for a one unit increase in the number of intersections the incidence rate of casualties would increase by a factor of 1.8 (assuming other variables are kept constant). To calculate the IRR for δ-unit change in a dependent variable, the formula exp(δ x 0.59) would be used.

As the size of a coefficient or IRR is affected by the unit of measurement used for the independent variable, a standardised IRR (StdIRR) was calculated for all coefficients using the "listcoef" function within STATA. The StdIRR represents the factor by which the incident rate would change for a one standard deviation change in an independent variable (UCLA: Academic Technology Services Statistical Consulting, 2010b; New York University, 2002).

4.1.3. Dummy variables

Each road segment was designated as belonging to one of three possible road classes - A, B or minor. In order to introduce this categorical variable into the regression models, each category was converted into a separate binary variable, known as a dummy variable, with two possible values (0 or 1). If all three of these dummy variables were to be entered into the model, multicollinearity would occur, and to avoid this it is necessary to remove one of the dummy variables from the model - in this case minor roads was removed. This affects how the coefficients of the remaining dummy variables are interpreted, as they now need to be compared with reference to the excluded variable (the reference variable). For

example, if Class A had an IRR of 1.5, this implies that the casualty incidence rate for Class A segments is 1.5 times greater than that for minor road segments.

4.1.4. Non-motorised user interactions

For the total casualty models (i.e. the models not disaggregated into pedestrian, cyclist and horse rider casualties), the explanatory factors representing NMT user interactions with the road segments (National Trails, NCN routes, and dangerous road crossings) were summed to create a "sum of interactions" variable.

4.2. The models

A series of NB regression models were estimated using the full segment dataset for total casualties, fatalities, serious injuries and slight injuries (Table 7), pedestrians, cyclists and horse riders (Table 8), and using a restricted segment dataset for A-class roads (Table 13) and the East Anglia region (Table 14). The overdispersion parameter, alpha, was found to be significantly different from zero in all the models (p-Value <0.001), confirming the choice of NB model over the Poisson model, and all models were found to be statistically significant (p-Values <0.0001). Nagelkerke/Cragg & Uhler's pseudo R^2 values ranged from 0.019 for the horse rider model to 0.137 for the total casualty model. Excluding the horse rider model, which only includes 1004 casualties, all the pseudo R^2 values were greater than 0.10 and broadly in line with other studies that have used NB models for accident count data. For example, models run by Wang et al. (2009a) produced pseudo R^2 values ranging from 0.1 to 0.21, and Noland and Quddus (2004) reported values between 0.05 and 0.26[23].

Across all models the majority of explanatory variables were found to be statistically significant at the 99% confidence level. Sinuosity was the only variable

[23]The published paper provided only log-likelihood values at intercept and full model, values reported here are calculated McFadden's R^2.

not to be significant in the total casualties and slight injury models, and was significant only in the fatal injury and horse rider models. Other non-significant variables include steepness in the fatal injury, serious injury and cyclist models, and the sum of user interactions in the fatal injury model. Dangerous crossing was non-significant in both the models it was included in (pedestrian and horse rider), along with National Trail in the horse rider model.

Each explanatory factor will now be considered in more depth and where possible reference will be made to results found in other studies, although direct comparisons are difficult as there have been no other studies specifically looking at NMT road casualties in non built-up areas.

4.2.1. Road class

With the exception of the horse rider model, coefficients for the A-class and B-class variables show a strong positive effect across all models, indicating that casualties are more likely to occur on these roads than minor roads. For total casualties the expected casualty rate increases by a factor of 4.12 for A-class segments and 2.93 for B-class segments. For the fatalities model the difference between minor roads and A-class roads is particularly marked, with the incidence of fatal casualties 11.96 times greater, whilst the corresponding rates for serious and slight injuries are 4.95 and 3.51 respectively. The pedestrian and cyclist models indicate little differentiation between the response of the two groups, with a casualty rate some four and three times greater on A and B roads respectively. Interestingly, in the horse rider model the A-class variable has a negative coefficient, with an estimated casualty rate 0.74 times lower than on minor roads (-26%), although there is still a positive coefficient for B-class roads. This probably reflects lower exposure of horse riders on A-class roads.

There have been several area-level studies in the UK that have found a significant and positive relationship between class A and B roads and NMT casualties (e.g. Quddus, 2008; Wang et al., 2009a). Graham and Stephens (2005) stud-

Variable	All casualties				Fatalities				Serious injuries				Slight injuries			
	R^2 = 0.137				R^2 = 0.135				R^2 = 0.108				R^2 = 0.125			
	LR $chi^2(8)$ = 34050.21				LR $chi^2(8)$ = 3619.87				LR $chi^2(8)$ = 9671.71				LR $chi^2(8)$ = 25158.65			
	Prob > chi^2 = <0.001				Prob > chi^2 = <0.0001				Prob > chi^2 = <0.001				Prob > chi^2 = <0.0001			
	Coef.	IRR	IRRStd	*p*-Value	Coef.	IRR	IRRStd	*p*-Value	Coef.	IRR	IRRStd	*p*-Value	Coef.	IRR	IRRStd	*p*-Value
A-class road	1.4152	4.1174	1.6444	< 0.001	2.4817	11.9612	2.3921	< 0.001	1.5990	4.9479	1.7541	< 0.001	1.2566	3.5137	1.5553	< 0.001
B-class road	1.0760	2.9330	1.3665	< 0.001	1.5355	4.6435	1.5614	< 0.001	1.1696	3.2208	1.4041	< 0.001	1.0173	2.7658	1.3434	< 0.001
Sinuosity	0.0013	1.0013	1.0019	0.742^{ns}	−6.5372	0.0014	0.0001	< 0.001	0.0057	1.0057	1.0083	0.386^{ns}	−0.0002	0.9998	0.9997	0.962^{ns}
Steep	−0.1703	0.8434	0.9800	0.023	0.3749	1.4548	1.0454	0.244^{ns}	−0.1334	0.8751	0.9843	0.363^{ns}	−0.2066	0.8133	0.9758	0.018
No. of intersections	0.5928	1.8090	1.4907	< 0.001	0.2320	1.2612	1.1691	< 0.001	0.4921	1.6358	1.3930	< 0.001	0.6417	1.8997	1.5406	< 0.001
Sum of interactions	0.2158	1.2408	1.0598	< 0.001	−0.0300	0.9704	0.9920	0.775^{ns}	0.2529	1.2878	1.0704	< 0.001	0.2132	1.2377	1.0590	< 0.001
Population ('000s)	0.0017	1.0017	1.1251	< 0.001	0.0009	1.0009	1.0674	< 0.001	0.0016	1.0016	1.1186	< 0.001	0.0017	1.0017	1.1262	< 0.001
Dist. from built-up area (m)	−0.0005	0.9995	0.3253	< 0.001	−0.0004	0.9996	0.4111	< 0.001	−0.0004	0.9996	0.3737	< 0.001	−0.0005	0.9995	0.3006	< 0.001

Table 7: Estimation results of NB models for all non-motorised casualties and by severity of casualty.

Variable	Pedestrian				Cyclist				Horse rider			
	R^2 = 0.106				R^2 = 0.135				R^2 = 0.019			
	LR $chi^2(8)$ = 16008.38				LR $chi^2(8)$ = 20985.26				LR $chi^2(8)$ = 273.90			
	Prob > chi^2 = <0.001				Prob > chi^2 = <0.001				Prob > chi^2 = <0.001			
	Coef.	IRR	IRRStd	*p*-Value	Coef.	IRR	IRRStd	*p*-Value	Coef.	*IRR*	IRRStd	*p*-Value
A-class road	1.3978	4.0462	1.6343	< 0.001	1.4923	4.4475	1.6896	< 0.001	−0.3061	0.7363	0.8980	0.003
B-class road	1.0627	2.8942	1.3612	< 0.001	1.1178	3.0580	1.3831	< 0.001	0.5804	1.7868	1.1834	< 0.001
Sinuosity	0.0032	1.0032	1.0047	0.527^{ns}	−0.0004	0.9996	0.9994	0.942^{ns}	−1.4492	0.2347	0.1193	0.010
Steep	–	–	–	–	0.0057	1.0057	1.0007	0.955^{ns}	–	–	–	–
No. of intersections	0.5255	1.6913	1.4246	< 0.001	0.6731	1.9604	1.5736	< 0.001	0.2467	1.2797	1.1807	< 0.001
National Trail	0.6931	2.0000	1.0581	< 0.001	–	–	–	–	−0.2296	0.7948	0.9815	0.623^{ns}
Dangerous crossing	0.0408	1.0416	1.0013	0.808^{ns}	–	–	–	–	−0.0423	0.9586	0.9987	0.968^{ns}
Sustrans route	–	–	–	–	0.2732	1.3141	1.0710	< 0.001	–	–	–	–
Population (’000s)	0.0018	1.0018	1.1349	< 0.001	0.0015	1.0015	1.1121	< 0.001	0.0012	1.0011	1.0832	0.004
Dist. from built-up area (m)	−0.0005	0.9996	0.3511	< 0.001	−0.0006	0.9995	0.2793	< 0.001	−0.0003	0.9997	0.5538	< 0.001

Table 8: Estimation results of NB models for pedestrians, cyclists and horse riders.

ied pedestrian casualties in English wards and found that a 10% increase in the length of A-class roads in a ward was associated with a 1.5% increase in total adult pedestrian casualties, and a 2.39% increase in those killed or seriously injured. A direct comparison with this study is difficult, but it appears to be a less marked effect, perhaps because no distinction is made between urban and non-urban casualties, or because the area-level approach is inherently less sensitive to the effect due to MAUP or ecological fallacy issues as no account is taken of the road class on which casualties actually occur.

The higher incidence of fatalities on A and B class roads is expected due to the higher speed of motorised vehicles on these roads. Although many roads in all three classes are likely to have posted speed limits in excess of 50mph outside of built-up areas, average speeds are generally lower on minor roads. The relationship between vehicle speed and fatality risk for unprotected road users is well established. For example, a recent analysis into the effect of car impact speed on pedestrians based on German accident data, found that the fatality risk at 50 km/hr was twice that at 40 km/hr and five times that at 30 km/hr (Rosén and Sander, 2009).

It is surprising that the casualty rate increase for pedestrians on these roads is of a similar magnitude to cyclists (albeit sightly reduced coefficients). It is not clear under what circumstances pedestrians are being exposed to risk on these roads outside of built-up areas, and this is worthy of further investigation.

4.2.2. Sinuosity

Sinuosity is significant in the fatalities, horse rider, and A-class road models, and in all three cases a very strong negative association is apparent. For a one unit increase in road segment sinuosity the estimated casualty incidence rate decreases by 99.86% for fatal casualties, 76.53% for horse riders, and 47.56% for A-class road segments. It should be noted that a one unit increase in sinuosity from a value of one (straight line) will have a marked effect on segment bendi-

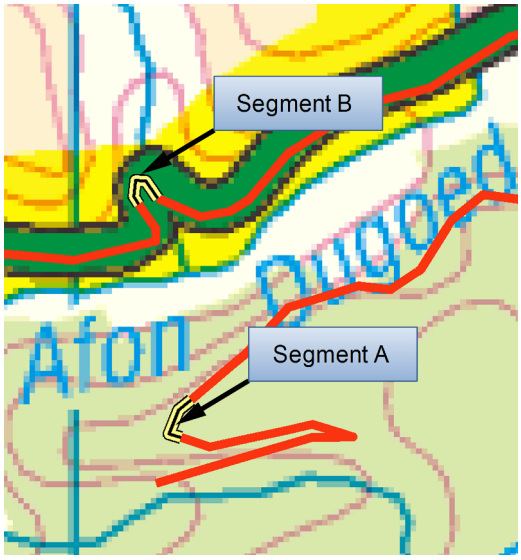

Figure 16: Illustration of road segments of different sinuosity.

ness, as shown in Figure 16, where Segment A has a sinuosity value of 1.39 and Segment B has a sinuosity value of 2.89.

A negative association between NMT casualties and various measures of road curvature has been reported by other researchers. A segment-based study of urban roads (maximum speed limit 60 km/h) in Addis Ababa, Ethiopia found a negative association between road curviness (degrees/km) and pedestrian casualties, and it was suggested that reduced vehicle speed on curved sections of road may give drivers more time to react (Berhanu, 2004). A ward-level study in England found a significant negative coefficient between bend density and seriously injured NMT road users, and noted that roads are often more curved in residential and commercial areas where drivers are more careful (Wang et al., 2009a). The same study reported a non-significant association in NMT fatality and slight injuries models.

The findings of this research are in line with these earlier studies, and suggest that higher sinuosity on A-class roads and the concomitant reduction in vehicle speed substantially reduces the likelihood of NMT road users being injured in a road traffic accident and markedly protects against the occurrence of fatal injuries. The significant negative association in the horse rider model, not seen in

the pedestrian or cyclist model, is probably influenced by the roads that horse riders prefer to use - narrow country lanes which are typically very bendy with low motorised traffic volumes, low average speeds, and low accident risk.

4.2.3. Steepness

Of the 742,255 road segments, 10,550 (1.42%) of them were designated as steep, and the significance of this factor varies between models. For total casualties there is a significant negative coefficient, with casualty incidence on steep segments reduced by a factor of 0.84 (-15.66%), and also a negative correlation in the slight injuries model, with a factor of 0.8133 (-18.67%). Steepness was introduced as an explanatory factor because it had been suggested that steep roads may increase the risk to cyclists due to a greater speed differential with motorised vehicles on uphill sections (Sustrans, 2009), but the coefficient is non-significant in the cyclist model. Steepness was not included in the pedestrian and horse rider models as there was no a priori knowledge that it would be a factor in road accidents for these groups, although the fact that the variable is significant in the total casualties and slight injuries models, but not in the cyclist model, suggests that it may be a factor for pedestrians and/or horse riders. This could be the result of lower vehicle volume or speed on steeper roads, and may reflect the fact that the vast majority of steep road segments (95.5%) are classified as minor roads, which we know from Section 4.2.1 to have lower casualty incidence (Table 9).

Road Class	Segment Frequency	
	Very Steep	Steep
A	9	107
B	17	347
Minor	982	9088
Totals	1008	9542

Table 9: Frequency of steep and very steep road segments grouped by road class.

4.2.4. Intersections

The number of intersections has a significant and positive association in all the models. A road segment can have 0, 1 or 2 intersections and the breakdown within the segment dataset is shown in Table 10. For total casualties the incidence rate increases by a factor of 1.81 (81%) for a one unit increase in the number of intersections. The coefficient is strongest for cyclists, with a factor of 1.96 for a one unit change. The estimated increase in cyclist casualty incidence from a segment with no intersections to a segment with two intersections is a factor of 3.84 (284%). A trend is evident in the incidence rates for the severity models, with the IRR for a two unit change increasing from fatalities (1.59), to serious (2.68) and slight injuries (3.61), indicating that whilst the presence of a junction increases the occurrence of a casualty of any severity, it is more likely to result in slight or serious injury than to cause a fatality.

No. of intersections	Segment frequency	%
0	390463	52.60
1	274016	36.91
2	77876	10.49
Total	742355	100.00

Table 10: Frequency of segments grouped by number of intersections.

The failure of several area-level studies to be sensitive to the positive relationship between junction presence and accident occurrence indicated by the raw STATS19 data was discussed in Section 2.3.1. The results presented here indicate that when road casualties are aggregated using a more appropriate spatial unit a strong and significant relationship is evident.

4.2.5. Non-motorised road user interactions

In the total casualties and injury severity models, the sum of interactions variable is used to test the relationship between casualty incidence and the presence of National Trails, NCN routes and dangerous crossing identified by the Ram-

blers. The sum of interactions can have a value of 0, 1, 2 or 3 depending on the number of distinct interactions present on the segment, and a breakdown of the segment frequency for this variable grouped by road class is shown in Table 11. It can be seen that the vast majority of the interactions occur on minor road segments (86.3% of the segments with at least one interaction).

No. of vulnerable user interactions	Segment Frequency			Total	%
	Class A	Class B	Minor		
0	103317	65230	518871	687418	92.60
1	3739	3643	46564	53946	7.27
2	94	37	855	986	0.13
3	5	0	0	5	< 0.01
Total	107155	68910	566290	742355	100.00

Table 11: Frequency of segments grouped by number of vulnerable user interactions and road class.

A significant positive correlation is evident in the total casualties, serious injury, and slight injury models, with the incidence rate for total casualties estimated to increase by a factor of 1.24 (24%) if one interaction is present, and by a factor of 1.54 (54%) when a segment has two interactions. The variable is not significant in the fatalities model but, given the very strong relationship between A-class segments and fatalities and the fact that only 7% of the segments with at least one interaction are located on A-class roads, this is perhaps not unexpected.

Turning to the individual interaction factors in the disaggregated user models, the presence of a National Trail is a positive and significant factor in the pedestrian model, doubling the casualty incidence, although it is not significant for horse riders. If a national or regional cycle route is present on a road segment it is significantly associated with an increased incidence of cyclist casualties, with the model estimating a factor increase of 1.31 (31%). These findings are to be expected as the variables represent increased pedestrian or cyclist exposure (i.e. the opportunity for an accident to occur) because both are well promoted and will

attract people to use them. The fact that road segments coincident with National Trails or NCN routes have higher incidence rates does not indicate that these segments are inherently more dangerous than other segments. If level of exposure was taken into account they could be found to be as safe, if not safer, than other road segments. However, it is a cause for concern that pedestrian casualties are occurring on the flagship walking routes in England and Wales, where walkers might reasonably expect not to be exposed to potentially dangerous interactions with motorised vehicles. The NCN routes are identified as either on-road or traffic-free which at least gives cyclists the opportunity to make a choice whether to expose themselves to the dangers or not. Table 12 shows the number of pedestrian and cyclist casualties occurring on National Trail and NCN route segments respectively.

The presence of a dangerous crossing identified by the Ramblers is not significant in either the pedestrian or horse rider models, and this factor will be considered further in the context of hot zones in Section 4.3.

Segment Type	Sum of pedestrian casualties			Sum of cyclist casualties		
	Fatal	Serious	Slight	Fatal	Serious	Slight
National Trail	6	53	414			
NCN route				18	291	924

Table 12: Sum of casualties associated with National Trail and NCN route segments.

4.2.6. Population

The resident population of the local authority district where a road segment is located is found to be significantly and positively associated with the number of casualties in all models. Due to the small size of the population unit (thousands), the coefficient appears to be very small, but the standardized IRR indicates that for a one standard deviation increase in the population (i.e. an increase of some 69,000), the incidence rate for total casualties increases by a factor of 1.125 (12.5%). The IRRStd is broadly similar across the models with the exception of

fatalities and horse riders where it is somewhat lower, at 1.067 (6.7%) and 1.083 (8.3%) respectively. These findings are consistent with a ward-level study by Wang et al. (2009a), which also found a positive association with resident population in separate slight, serious and fatal NMT casualty models, and suggests that population is acting, to some degree, as a proxy for exposure.

4.2.7. Distance from built-up area

A significant and positive correlation between casualty count and the distance of the road segment from a built-up area is found in all the models. As with the population variable, due to the size of unit used (metres) the coefficients appear very small, but the standardized IRR shows that a one standard deviation increase in mean distance (2.2km) reduces the incidence rate of total casualties by a factor of 0.325 (-67.5%). The reduction in incidence rate over this distance ranges from -70% for slight injuries to -58.9% for fatalities, and from -72% for cyclists to -44.62% for horse riders. Figure 17 shows a scatter plot of casualty count against distance to built-up area, clearly illustrating the pattern of casualty count decreasing with distance. These results suggest that this measure is, like population, acting to some degree as a proxy for exposure. It might reasonably be expected that more pedestrian and cyclist activity occurs closer to areas where people live and work.

4.2.8. Predicted AADF

As suitable vehicle flow data was only available for A-class roads, a separate model was run for these segments to examine the impact of traffic flow on non-motorised road user casualties. As shown in Table 13, there is a positive and significant association between predicted AADF and total casualties. The standardized IRR shows that for a one standard deviation increase in AADF (representing an additional 7,285 vehicles per day), the total casualty incidence rate increases by a factor of 1.377 (37.7%).

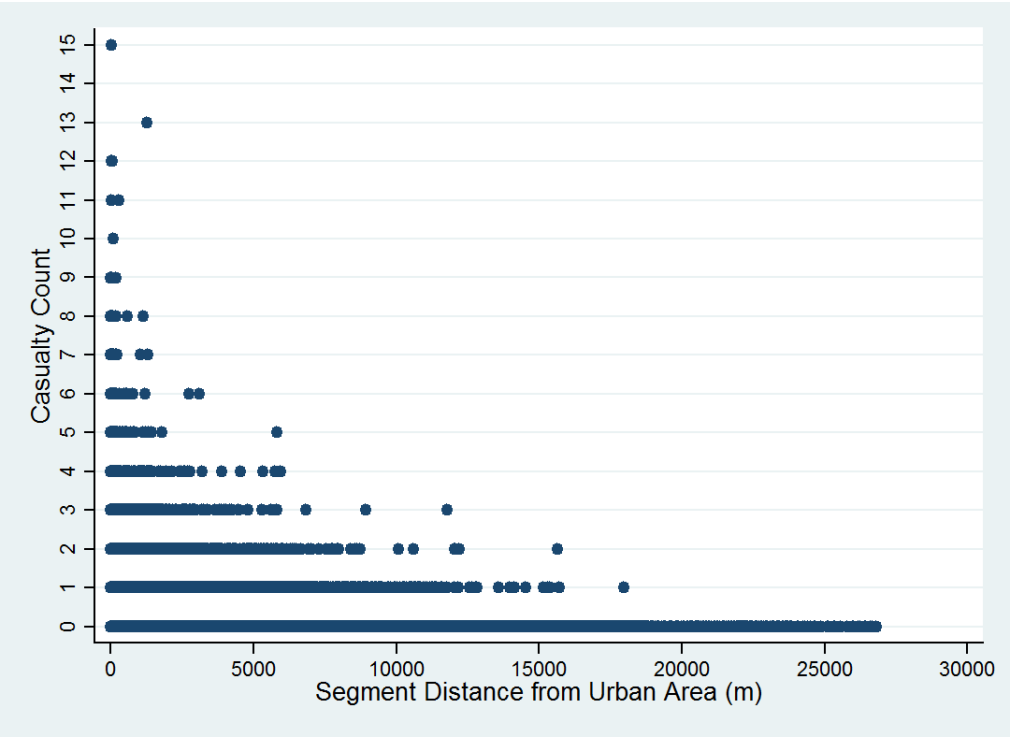

Figure 17: Scatterplot of segment casualty count against distance from built-up area.

Variable	A-class roads			
	$R^2 = 0.111$			
	LR $chi^2(8) = 7201.86$			
	Prob > chi^2 = <0.001			
	Coef.	IRR	IRRStd	*p*-Value
Sinuosity	−0.64551	0.52440	0.8205	0.020
Steep	0.43327	1.54230	1.0143	0.199^{ns}
Predicted AADF	0.00004	1.00004	1.3771	< 0.001
No. of intersections	0.63732	1.89141	1.5599	< 0.001
Sum of interactions	0.23162	1.26064	1.0459	< 0.001
Population ('000s)	0.00103	1.00103	1.0813	< 0.001
Dist. from built-up area (m)	−0.00032	0.99968	0.5079	< 0.001

Table 13: Estimation results of NB model for A-class roads.

This relationship is expected, as the more traffic there is on a road the greater the opportunity for an accident to occur and the greater difficulty pedestrians and cyclists may have in making common road manoeuvres, such as crossing or turning at junctions. Wang et al. (2009a) reported a similar positive correlation between a ward-level traffic activity measure and fatal and serious NMT casualties, but found a negative association in their slight injuries model. In contrast, in an analysis of London crash data, Quddus (2008) found traffic flow to be a non-significant factor for NMT casualties, and suggested that this may be due to the effect being accounted for by other variables, such as employment and resident population. It is interesting to note that in this study both traffic flow and population variables are found to be significant in the A-class model.

4.2.9. Impact of spatial autocorrelation

As discussed in Section 3.5.2, a small subset of the segments was taken (those located in East Anglia), to enable the introduction of a spatially lagged dependent variable into a regression model and to assess the potential impact of the presence of spatial autocorrelation. The findings could then be used to guide interpretation of the results from the main models. The model estimations (Table 14) show that the spatially lagged variable is significant, indicating that unexplained spatial autocorrelation remains which may affect coefficients and significance tests.

Indeed, it can be seen that introduction of the lagged variable has changed the sinuosity variable from being significant at the 90% confidence level to non-significant, confirming warnings referred to in Section 2.6, that type I errors are a particular problem when coefficients are close to the significance threshold. The coefficients of the other variables that are significant in the model before introduction of the lagged variable (at the 99.9% confidence level), remain significant in the spatially lagged model. The coefficients do change somewhat, generally becoming slightly smaller, but not dramatically so, and the confidence level for the significance of the population coefficient reduces slightly from 99.9% to

Variable	East Anglia (all casualties)				East Anglia (all casualties with spatial lag variable)			
	R^2 = 0.108				R^2 = 0.125			
	LR $chi^2(8)$ = 1941.69				LR $chi^2(8)$ = 2247.97			
	Prob > chi^2 = <0.001				Prob > chi^2 = <0.001			
	Coef.	IRR	IRRStd	*p*-Value	Coef.	IRR	IRRStd	*p*-Value
A-class road	1.61225	5.01406	1.7029	< 0.001	1.54697	4.69721	1.6666	< 0.001
B-class road	1.12193	3.07078	1.4104	< 0.001	1.05467	2.87102	1.3816	< 0.001
Sinuosity	−0.57100	0.56496	0.4217	0.091	−0.54423	0.58029	0.4391	0.111[ns]
Steep	−13.57650	1.27*e*−06	0.8775	0.996[ns]	−27.12725	1.65*e*−12	0.7702	1.000[ns]
No. of intersections	0.50665	1.65972	1.4001	< 0.001	0.48233	1.61985	1.3777	< 0.001
Sum of interactions	0.09944	1.10455	1.0275	0.271[ns]	0.12011	1.12762	1.0333	0.181[ns]
Population ('000s)	0.00364	1.00364	1.0875	< 0.001	0.00250	1.00250	1.0593	0.009
Dist. from built-up area (m)	−0.00047	0.99953	0.5298	< 0.001	−0.00042	0.99958	0.5705	< 0.001
Spatial lag of dependent variable	–	–	–	–	1.32795	3.77328	1.2615	< 0.001

Table 14: Estimation results of NB models for East Anglia.

99.1%.

These findings suggest that we can have fairly high confidence in the estimations produced in the main models. The significant coefficients with the lowest *p*-Values in the main models were sinuosity in the horse rider model (*p*-Value = 0.010), and steepness in the slight injuries model (*p*-Value = 0.018), in both cases not near the 90% confidence level which resulted in a Type I error in the East Anglia model.

4.3. Hot zones

Each road segment was assigned three hot zone threshold values, one for pedestrians, one for cyclists and one horse riders. The definition of the thresholds is shown in Table 6, and Table 15 shows the number of casualties occurring on road segments of each hot zone threshold. For example, there were 172 pedestrian casualties coincident with road segments that were located within pedestrian hot zones of threshold value 4, which represents zones with 24-32 casualties per square km (6-8 casualties per cell). The vast majority of casualties (80%) occur in hot zone 1, which for pedestrians and cyclists is equivalent to less than two casualties per cell, and for horse riders less than one. As the data for this study spans a ten year period it is unlikely that these locations would be

Hot zone threshold	Pedestrians				Cyclists				Horse riders			
	All	F	S	Sl	All	F	S	Sl	All	F	S	Sl
None	7	0	3	4	0	0	0	0	0	0	0	0
1	14593	1247	3785	9561	14676	458	3263	10955	879	14	148	717
2	2561	163	695	1703	2732	49	491	2192	99	0	18	81
3	472	16	115	341	782	8	110	664	23	0	4	19
4	172	5	52	115	276	7	29	240	4	0	0	4
5	105	2	26	77	249	6	34	209	-	-	-	-
Total	17910	1433	4676	11801	18715	528	3927	14260	1005	14	170	821

Table 15: Number of casualties occurring on road segments grouped by hot zone threshold of segment and road user type.

identified as requiring particular attention or be candidates for remedial action, as in many cases they will represent a single casualty. If hot zones of threshold values 4 and 5 were selected for particular attention, they would include just 2.1% of the total casualties, and 1% of the fatalities. The pedestrian casualties in hot zone threshold 5 (8 or more per cell), represent just 0.6% of the total, a proportion even smaller than the 4% of pedestrian injuries found by Morency and Cloutier (2006) in urban hot spots with 8 or more pedestrian injuries. This is probably explained by the more diffuse nature of accidents in a non-urban environment, and suggests that any approach aimed at casualty reduction which solely considers high-incidence locations will not have an appreciable effect.

It should be noted that casualties rather accidents have been considered in the identification of hot zones, and a single accident can result in more than one NMT casualty, leading to the potential for hot zones to have high casualty density but only a single accident occurrence.

A tabulation of the mean distance of segments from built-up areas grouped by hot zone threshold (Table 16) reveals that, with the exception of the horse rider hot zones, the mean distance decreases as the hot zone threshold increases. Hot zones with higher accident density occur nearer built-up areas, with the mean distance less than 1km for all except the lowest density zone (threshold value 1). This may be due to higher levels of exposure nearer the built-up areas - people walk and cycle near to where the live - but could also be the result of segments

in urban areas being included in the analysis, a possibility considered further in Section 4.4.

Hot zone threshold	Mean distance of segment from built-up polygon (m)		
	Pedestrian hot zones	Cyclist hot zones	Horse rider hot zones
0	2420.56	2437.45	2255.94
1	1191.69	1095.35	1480.27
2	636.26	335.81	1540.68
3	326.07	219.55	1778.43
4	353.00	135.02	2270.82
5	66.64	154.28	-
All segments	2248.65	2248.65	2248.65

Table 16: Mean distance of segments from built-up polygons grouped by hot zone threshold.

The regression models discussed above did not find a significant correlation between dangerous crossings and casualty count. However, the dangerous crossing dataset is very small in relation to the segment dataset, and is by no means a complete record of locations where public footpaths and bridleways are severed by busy roads. Its coverage is patchy, with no crossings identified in Devon and Cornwall for example, and some of the crossings are suppressed as they are considered so dangerous nobody would attempt to use them (e.g. six lanes of dual carriageway). To explore the relationship between casualties and the dangerous crossing further, segments containing a dangerous crossing were tabulated and grouped by hot zone threshold (Table 17). This descriptive analysis reveals that 196 of the dangerous crossings (26.6%) are coincident with segments in pedestrian hot zones, and 7 (0.95%) are coincident with horse rider hot zones, suggesting that the presence of these crossings may be contributing to accidents at these locations.

Tables 18 and 19 identify the location of each of the dangerous crossings located on segments in pedestrian hot zone threshold 2 and in horse rider hot zone threshold 1.

Hot zone threshold	Segments containing a dangerous crossing	
	Pedestrian hot zones	Horse rider hot zones
0	542	731
1	191	7
2	5	0
3	0	0
4	0	0
5	0	-
Total crossings	738	738

Table 17: Segments containing a dangerous crossing grouped by segment hot zone threshold.

Road Number	Road Name	Easting	Northing
A27(T)	Chichester By-pass	487760	105317
A27(T)	The Causeway	502495	106396
A27(T)	Arundel Road	506624	105691
A421(T)	-	501703	244536
A21(T)	Pembury Road	561310	143236

Table 18: Dangerous crossings located on pedestrian threshold 2 hot zone segments.

4.4. Limitations of analysis

4.4.1. Under-reporting and misclassification of road casualties

It has been recognised for some time that the STATS19 data does not provide a complete record of NMT road casualties. For example, Teanby (1992) compared accident records for the Merseyside police area with a trauma care database[24] and found that pedestrian accidents were under-reported by 16%. More recently, the National Audit Office found that in 2006-07 41% more pedestrians and 228% more cyclists were admitted to hospital with serious injuries than were recorded in the STATS19 data. Many of the unreported cyclist injuries were the result of accidents that did not involve collision with another vehicle or object, but even with these removed there were 18% more in the hospital data (National Audit Office, 2009). The STATS19 data is intended to include cyclists who injure themselves on a public road without any other vehicle being involved,

[24]This included data compiled from the ambulance service, hospital accident departments and coroners.

Road Number	Road Name	Easting	Northing
A61	Penistone Road	433699	395405
A61	Penistone Road	433494	395755
A3066	Crooked Oak Hill	348601	098592
A3066	-	348633	098111
A12(T)	London Road	590876	222964
A120(T)	Wix By-Pass	590876	222964
A13(T)	-	563258	180862

Table 19: Dangerous crossings located on horse rider threshold 1 hot zone segments.

and of the cyclist casualties in non built-up areas analysed in this study, 5.8% of them were the the result of such accidents. Concerns have also been raised that the police may misclassify some serious injuries as slight, and an observed increase in the proportion of casualties classified as slight rather than serious over recent years does not appear to be supported by data from hospital in-patient records (Jeffrey et al., 2009).

For under-reporting to be a concern in this study there would need to be a systematic bias in the under-reporting with respect to at least one explanatory factor used in the regression models. For example, if casualties occurring on A-class roads were less likely to be under-reported, and those on minor roads more likely to be under-reported then part of the difference in casualty incidence between the two could be explained merely by differences in reporting rates. However, there is no evidence that such a bias in reporting exists.

4.4.2. Accident location accuracy

There is some doubt about the accuracy of the accident location recorded in STATS19. The 8-figure NGR provided for each accident suggests a location precision of 10m, but it is clear from Section 3.3 that there are quality issues with this information. A review of STATS19 recording practices reveals that, in the case of the Metropolitan Police area, the attending police officer will take written notes about the accident location, including landmarks and road markings, and this information will then be passed to a central processing unit where the

NGR will be determined with reference to Ordnance Survey maps - a procedure that clearly provides an opportunity for errors to occur (Anderson, 2003). The scope for errors is also high in the case of accidents reported over the counter at police stations, where no officer has attended the scene and the only information available is that provided by a member of the public.

As discussed in Section 3.3, various steps were taken to minimise errors resulting from inaccurate NGRs, including a local authority matching process, snapping casualties to a road polyline of matching road class, and restricting snapping to a 300m tolerance. However, it is still likely that a certain proportion of the casualties will have been assigned to a road segment which has characteristics which do not reflect those of the road segment where the actual accident occurred.

4.4.3. Extent of urban areas

As outlined in Section 3.1, the polygons which were used to exclude the casualties that occurred in built-up areas were based on population data from the 2001 census and Ordnance Survey mapping current at that time. However, the casualty data used in this study is for the ten year period 1999-2008, and it can be expected that the extent of built-up areas has expanded since 2001 in many towns and cities. This is illustrated in Figure 18, which shows the built-up polygon for Ely, Cambridgeshire overlaying the Ordnance Survey 1:50,000 raster map for 2009. It is clear that there has been expansion in parts of the north and west of the city and any casualties that occurred in these areas would not have been excluded from the analysis, even though the areas may have been built-up at the time.

If built-up areas are separated by less than 200m of non built-up land then they are combined together into a single built-up polygon. However, an issue arises within some towns and cities where there are areas of land that are not built-up, either because they are disused or perhaps playing fields or common

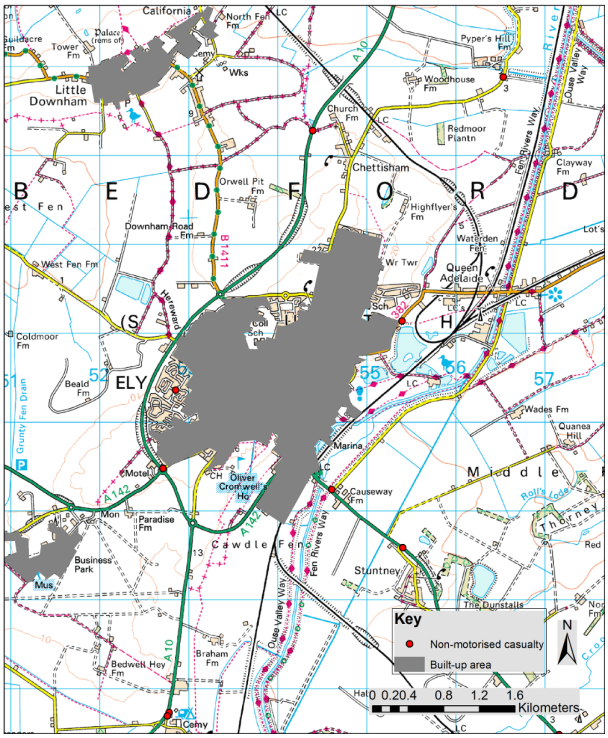

Figure 18: Built-up area polygon overlaying 2009 Ordnance Survey 1:50,000 raster map showing expansion of built-up area since 2001 in Ely, Cambridgeshire.

land, that extend beyond 200m and therefore exist outside of the built-up polygon, even though the characteristic of a road passing through the area is no different from roads within the built-up polygon. An example, where the eastern ring road passes alongside Coldham's Common in Cambridge, is shown in Figure 19.

The implication of these two issues is that a certain number of casualties that were included in the analysis actually occurred on roads that share the characteristics of built-up rather than non built-up locations. It is also possible that some of road segments with the highest casualty counts are within these locations, as many more accidents occur in built-up areas.

4.4.4. Length of study period

A relatively large study period was selected because road casualties are rare events, especially when limited to NMT casualties outside of built-up areas. A larger casualty database reduces the number of zero count segments and increases the mean casualty count, eliminating potential complications with the

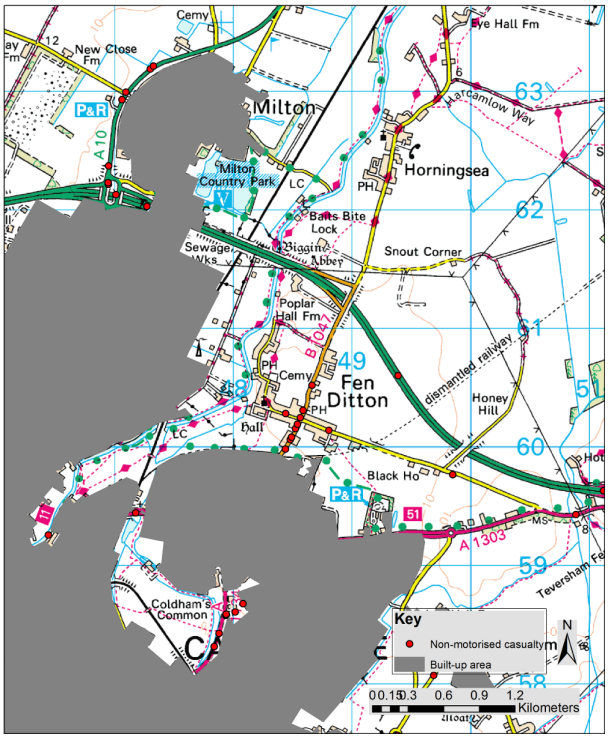

Figure 19: Built-up area polygon overlaying 2009 Ordnance Survey 1:50,000 raster map showing built-up area gap alongside Coldham's Common, Cambridge.

statistical analysis. It also enables the meaningful identification of casualty hot zones which would be unlikely to reveal themselves in a short study period due to the more diffuse nature of accidents outside of built-up areas. However, a downside of this approach is that the explanatory factors associated with the road segments may have changed during the ten year period. Roads may have been reclassified, traffic flows altered, new cycle routes established, national trails diverted and so forth. This has not been taken account of in this study, which assumes that each segment's characteristics remained static over the period and were the same if an accident occurred in 1999 or 2008.

4.4.5. Ecological fallacy issues

As the casualties have been aggregated by road segment for the count-based regression models used in this study, an assumption is made that the explanatory factors assigned to a road segment are the same as those of the actual road location where each aggregated casualty occurred. While we can be confident that

this is predominantly the case for some of the factors, for example road class and resident population, for others we can be less sure. For example, if a 250m segment has intersection nodes at both ends but an accident occurred at the centre of that segment, then the nearest road junction is 125m from the casualty location and may not be a contributory factor. As another example, if a National Trail is coincident with a segment at any point then the entire segment is identified as being part of a National Trail, when in fact the trail may just cross the road segment at a specific point or may follow the road for only a short distance, and not the full 250m. A casualty occurring on such a segment would be considered to have occurred on a National Trail, when it may actually have occurred elsewhere on the segment and not have been affected by the the trail being coincident with the segment.

5. Conclusions

This study has developed a dataset for the road network in England and Wales consisting of segments of 250m nominal length. Attributes for a range of explanatory factors and aggregated counts of reported NMT casualties occurring in non built-up areas have then been assigned to each segment. These segments have been used as the BSU in a range of disaggregated NB regression models with casualty count as the dependent variable. The potential effect of spatial autocorrelation on the regression estimations has been examined by introducing a spatial lag of the dependent variable into a model based on a small subset of the segments. While this shows that unexplained spatial autocorrelation remains, it also indicates that we can be confident in the validity of the estimations produced by the full segment models.

The results show that the expected NMT casualty rate is significantly higher on A-class and B-class road segments, and that this effect is particularly marked in the case of fatalities. Other factors that increase the likelihood of NMT ca-

sualties in non built-up areas include the presence of road junctions and greater motorised traffic flow. Road sinuosity appears to be a very important factor in reducing the likelihood of fatalities on all roads, and in reducing total casualties on A-class roads. Casualty incidence also falls as the distance from built-up areas increases and as resident population of the local authority area decreases.

These findings suggest that in order to have maximum impact in reducing NMT casualties in non built-up areas, and in particular fatalities, the focus needs to be on A-class and B-class roads, rather than minor roads. The observed effect of sinuosity in reducing casualties on the faster A-class roads, suggests that measures to reduce speed on these roads would be effective in cutting the casualty toll. In view of the inverse relationship between distance from built-up areas and casualty rates, it is possible that the introduction of stepped reductions in posted speed limits on the approach to built-up areas, rather than the sudden imposition of low speed limits at the edge of settlements, might be an effective approach. In addition, it is possible that protective speed limits on stretches of road between built-up areas could be considered where the distance is below a specified threshold. Further analysis of the data to investigate whether the rate of NMT casualties is higher on short road sections between settlement areas than it is on open roads in general would be useful. It is not clear why the factor increase in incidence of pedestrian casualties on A-class and B-class roads is of a similar magnitude as that for cyclists. Further study is needed to investigate the nature of these accidents and to understand why and how pedestrians are being dangerously exposed to motorised traffic on these roads.

Road segments with National Trail or NCN routes coincident with them have a higher casualty incidence. This is most likely to be the result of higher exposure rather than these road sections being inherently more dangerous. Cyclists have a choice between NCN routes that include on-road sections and those that are entirely traffic-free. This choice is not available for walkers, who may as-

sume that the country's flagship long distance walking routes are safe. Further analysis could be carried out to identify specific locations where accidents have occurred and this information could then be used by National Trail officers to plan a priority programme for establishing alternative walking routes that avoid roads. In addition to the National Trails there are some 1000 promoted long distance paths in the UK, and by digitising these paths a more comprehensive database of pedestrian interactions could be created. It would also be useful to obtain a much larger and more representative dataset of locations where footpaths and bridleways are severed by roads, potentially by using digitised rights of way data maintained by local authorities.

Finally, the methodology that has been developed for this study could be applied to national-scale segment-based studies of motorised casualties in the UK, either within or outside of built-up areas, and could also be adapted to carry out similar studies in other countries.

6. References

Aguero-Valverde, J., Jovanis, P. P., 2008. Analysis of Road Crash Frequency with Spatial Models. Transportation Research Record: Journal of the Transportation Research Board 2061, 55–63.
URL http://trb.metapress.com/openurl.asp?genre=article&id=doi:10.3141/2061-07

Akins, S., Rumbaut, R. G., Stansfield, R., 2009. Immigration, Economic Disadvantage, and Homicide: A Community-level Analysis of Austin, Texas. Homicide Studies 13 (3), 307–314.
URL http://hsx.sagepub.com/cgi/doi/10.1177/1088767909336814

Anderson, T. K., 2003. Review of Current Practices in Recording Road Traffic Incident Data: With Specific Reference to Spatial Analysis and Road Policing Policy.

Anderson, T. K., May 2009. Kernel density estimation and K-means clustering to profile road accident hotspots. Accident Analysis and Prevention 41 (3), 359–64.
URL http://www.ncbi.nlm.nih.gov/pubmed/19393780

Anselin, L., Syabri, I., Kho, Y., 2006. GeoDa: An introduction to spatial data analysis. Geographical Analysis 38, 5–22.
URL http://www.springerlink.com/index/K82363R125061785.pdf

Berhanu, G., 2004. Models relating traffic safety with road environment and traffic flows on arterial roads in Addis Ababa. Accident; analysis and prevention 36 (5), 697–704.
URL http://www.ncbi.nlm.nih.gov/pubmed/15203347

Beyer, H. L., 2004. Hawth's Analysis Tools for ArcGIS.
URL http://www.spatialecology.com/htools

Campaign To Protect Rural England, 1999. Rural Traffic Fear Survey.

Coxe, S., West, S. G., Aiken, L. S., March 2009. The analysis of count data: a gentle introduction to poisson regression and its alternatives. Journal of personality assessment 91 (2), 121–36.
URL http://www.ncbi.nlm.nih.gov/pubmed/19205933

Department For Transport, 2004. STATS20 - Instructions for the Completion of Road Accident Reports.
URL http://www.dft.gov.uk/pgr/statistics/datatablespublications/accidents/casualtiesgbar/s20instructionsforthecom5094.pdf

Department For Transport, 2009a. A Safer Way: consultation on Making Britain's Roads the Safest in the World - Executive summary.

Department For Transport, 2009b. Reported Road Casualties Great Britain: 2008 Annual Report.

Eck, J., Chainey, S., Cameron, J., Leitner, M., Wilson, R., 2005. Mapping crime: Understanding hot spots.
URL http://eprints.ucl.ac.uk/11291/

Edwards, J., 1996. Weather-related road accidents in England and Wales: a spatial analysis. Journal of Transport Geography 4 (3), 201–212.
URL http://linkinghub.elsevier.com/retrieve/pii/0966692396000063

Elhai, J. D., Calhoun, P. S., Ford, J. D., 2008. Statistical procedures for analyzing mental health services data. Psychiatry research 160 (2), 129–36.
URL http://www.ncbi.nlm.nih.gov/pubmed/18585790

Erdogan, S., Yilmaz, I., Baybura, T., Gullu, M., 2008. Geographical information systems aided traffic accident analysis system case study: city of Afyonkarahisar. Accident Analysis and Prevention 40 (1), 174–81.
URL http://www.ncbi.nlm.nih.gov/pubmed/18215546

ET SpatialTechniques, 2009. ET GeoWizards.
URL http://www.ian-ko.com

Flahaut, B., 2004. Impact of infrastructure and local environment on road unsafety. Logistic modeling with spatial autocorrelation. Accident Analysis and Prevention 36 (6), 1055–1066.
URL http://www.ncbi.nlm.nih.gov/pubmed/15350882

Flahaut, B., Mouchart, M., San Martin, E., Thomas, I., November 2003. The local spatial autocorrelation and the kernel method for identifying black zones. A comparative approach. Accident Analysis and Prevention 35 (6), 991–1004.
URL http://www.ncbi.nlm.nih.gov/pubmed/12971934

Flowerdew, R., Manley, D. J., Sabel, C. E., 2008. Neighbourhood effects on health: does it matter where you draw the boundaries? Social science & medicine (1982) 66 (6), 1241–55.
URL http://www.ncbi.nlm.nih.gov/pubmed/18177988

Fortin, M.-J., Dale, M. R. T., 2009. The SAGE Handbook of Spatial Analysis. SAGE, London, Ch. 6, pp. 89–103.

Geirt, F. V., Nuyts, E., 2006. Cross-sectional Accident Models On Flemish Motorways Based On Infrastructural. In: International Conference on Regional and Urban Modeling. No. 2004.
URL http://www.ecomod.org/files/papers/1278.pdf

Geurts, K., Thomas, I., Wets, G., 2005. Understanding spatial concentrations of road accidents using frequent item sets. Accident Analysis and Prevention 37 (4), 787–99.
URL http://www.ncbi.nlm.nih.gov/pubmed/15899471

Graham, D. J., Stephens, D. A., 2005. The effects of area deprivation on the incidence of child and adult pedestrian casualties in England. Accident Analysis and Prevention 37 (1), 125–35.
URL http://www.ncbi.nlm.nih.gov/pubmed/15607283

Graham, D. J., Stephens, D. A., 2008. Decomposing the impact of deprivation on child pedestrian casualties in England. Accident Analysis and Prevention 40 (4), 1351–64.
URL http://www.ncbi.nlm.nih.gov/pubmed/18606266

Gruenewald, P. J., Freisthler, B., Remer, L., Lascala, E. A., Treno, A. J., Ponicki, W. R., 2009. Ecological Associations of Alcohol Outlets With Underage and Young Adult Injuries. Alcoholism, clinical and experimental research 34 (3), 519–527.
URL http://www.ncbi.nlm.nih.gov/pubmed/20028361

Grundy, C., Steinbach, R., Edwards, P., Green, J., Armstrong, B., Wilkinson, P., 2009. Effect of 20 mph traffic speed zones on road injuries in London, 1986-2006: controlled interrupted time series analysis. BMJ 339, 1–6.
URL http://www.bmj.com/cgi/doi/10.1136/bmj.b4469

Guikema, S., Coffelt, J., 2009. Practical Considerations in Statistical Modeling of Count Data for Infrastructure Systems. Journal of Infrastructure Systems 15 (September), 172–178.
URL http://link.aip.org/link/?JITSE4/15/172/1

Haining, R., 2009. Spatial autocorrelation and the quantitative revolution. Geographical Analysis 41 (4), 364–374.
URL http://www3.interscience.wiley.com/journal/122663060/abstract

Haining, R., Law, J., Griffith, D., 2009. Modelling small area counts in the presence of overdispersion and spatial autocorrelation. Computational Statistics & Data Analysis 53 (8), 2923–2937.
URL http://linkinghub.elsevier.com/retrieve/pii/S0167947308003940

Hannon, L. E., December 2005. Extremely Poor Neighborhoods and Homicide. Social Science Quarterly 86 (s1), 1418–1434.
URL http://www.blackwell-synergy.com/doi/abs/10.1111/j.0038-4941.2005.00353.x

Haynes, R., Lake, I. R., Kingham, S., Sabel, C. E., Pearce, J., Barnett, R., 2008. The influence of road curvature on fatal crashes in New Zealand. Accident Analysis and Prevention 40 (3), 843–50.
URL http://www.ncbi.nlm.nih.gov/pubmed/18460350

Hilbe, J. M., 2008. Negative Binomial Regression. Cambridge University Press, Cambridge.

House Of Commons Transport Committee, 2008. Ending the Scandal of Complacency: Road Safety beyond 2010 Eleventh Report of Session 2007-08.

Hughes, W., 1994. Accidents on Rural Roads. AA Foundation for Road Safety Research, Basingstoke, Hampshire.

Jeffrey, S., Stone, D. H., Blamey, A., Clark, D., Cooper, C., Dickson, K., Mackenzie, M., Major, K., 2009. An evaluation of police reporting of road casualties. Injury prevention : journal of the International Society for Child and Adolescent Injury Prevention 15 (1), 13–8.
URL http://www.ncbi.nlm.nih.gov/pubmed/19190270

Jones, A. P., Haynes, R., Kennedy, V., Harvey, I. M., Jewell, T., Lea, D., 2008. Geographical variations in mortality and morbidity from road traffic accidents in England and Wales. Health & place 14 (3), 519–35.
URL http://www.ncbi.nlm.nih.gov/pubmed/18032087

Kissling, W. D., Carl, G., 2007. Spatial autocorrelation and the selection of simultaneous autoregressive models. Global Ecology and Biogeography, 1–13.

URL http://www.blackwell-synergy.com/doi/abs/10.1111/j.1466-8238.2007.00334.x

Knowles, J., Adams, S., Cuerden, R., Savill, T., Reid, S., Tight, M., 2009. Collisions involving pedal cyclists on Britains roads: establishing the causes.

Koorey, G., 2009. Road Data Aggregation and Sectioning Considerations for Crash Analysis. Transportation research record (2103), 61–68.
URL http://cat.inist.fr/?aModele=afficheN&cpsidt=22002357

Kubrin, C. E., Weitzer, R., May 2003. Retaliatory Homicide: Concentrated Disadvantage and Neighborhood Culture. Social Problems 50 (2), 157–180.
URL http://caliber.ucpress.net/doi/abs/10.1525/sp.2003.50.2.157

Levine, N., 2004. CrimeStat III: A Spatial Statistics Program for the Analysis of Crime Incident Locations.

Levine, N., 2005. CrimeStat III Manual, Chapter 5 - Distance Analysis I and II, 3rd Edition. Ned Levine & Associates, Houston, TX, and the National Institute of Justice, Washington, DC.
URL http://www.icpsr.umich.edu/files/CRIMESTAT/files/CrimeStatChapter.5.pdf

Lin, G., Zhang, T., July 2007. Loglinear Residual Tests of Moran's I Autocorrelation and their Applications to Kentucky Breast Cancer Data. Geographical Analysis 39 (3), 293–310.
URL http://www.blackwell-synergy.com/doi/abs/10.1111/j.1538-4632.2007.00705.x

Long, J., Freese, J., 2001. Scalar measures of fit for regression models. Stata Technical Bulletin 10 (56).
URL http://ideas.repec.org/a/tsj/stbull/y2001v10i56sg145.html

Lord, D., Washington, S., Ivan, J. N., 2007. Further notes on the application of zero-inflated models in highway safety. Accident Analysis and Prevention 39 (1), 53–7.
URL http://www.ncbi.nlm.nih.gov/pubmed/16949027

Lynam, D. A., 2007. Rural road safety - policy options. TRL Limited, Wokingham, Berkshire.

Morency, P., Cloutier, M.-S., December 2006. From targeted "black spots" to area-wide pedestrian safety. Injury prevention : journal of the International Society for Child and Adolescent Injury Prevention 12 (6), 360–4.
URL http://www.ncbi.nlm.nih.gov/pubmed/17170182

National Audit Office, 2009. Improving road safety for pedestrians and cyclists in Great Britain.

New York University, 2002. Event Count Models - Poisson Regression.
URL www.nyu.edu/classes/nbeck/q2/zorn.eventcount.pdf

Nielsen, A. L., Hill, T. D., French, M. T., Hernandez, M. N., 2010. Racial/ethnic composition, social disorganization, and offsite alcohol availability in San Diego County, California. Social Science Research 39 (1), 165–175.
URL http://linkinghub.elsevier.com/retrieve/pii/S0049089X09000441

Noland, R. B., Quddus, M. A., 2004. A spatially disaggregate analysis of road casualties in England. Accident Analysis and Prevention 36 (6), 973–84.
URL http://www.ncbi.nlm.nih.gov/pubmed/15350875

Okabe, A., Satoh, T., 2009. The SAGE Handbook of Spatial Analysis. SAGE, London, Ch. 23, pp. 443–464.

Openshaw, S., 1984. The modifiable areal unit problem. Concepts and Techniques in Modern Geography 38, 40.

Parida, M., Jain, S. S., Landge, V. S., 2006. Stochastic modelling for traffic crashes on non urban highways in india. In: 22nd ARRB Conference - Research into Practice. Canberra, Australia, pp. 1–13.

Pulugurtha, S. S., Krishnakumar, V. K., Nambisan, S. S., July 2007. New methods to identify and rank high pedestrian crash zones: an illustration. Accident; analysis and prevention 39 (4), 800–11.
URL http://www.ncbi.nlm.nih.gov/pubmed/17227666

Qin, X., Ivan, J., 2001. Estimating Pedestrian Exposure Prediction Model in Rural Areas. Transportation Research Record 1773 (1), 89–96.
URL http://trb.metapress.com/openurl.asp?genre=article&id=doi:10.3141/1773-11

Quddus, M. A., 2008. Modelling area-wide count outcomes with spatial correlation and heterogeneity: an analysis of London crash data. Accident Analysis and Prevention 40 (4), 1486–97.
URL http://www.ncbi.nlm.nih.gov/pubmed/18606282

Ramblers, 2003. Your Either Quick or Dead.
URL http://www.ramblers.org.uk/Resources/RamblersAssociation/Website/RightsofWay/Documents/Row_quickordead.pdf

Road Traffic Statistics Branch, 2007. How the National Road Traffic Estimates are made.
URL http://www.dft.gov.uk/matrix/estimates.aspx

Rosén, E., Sander, U., 2009. Pedestrian fatality risk as a function of car impact speed. Accident Analysis and Prevention 41 (3), 536–42.
URL http://www.ncbi.nlm.nih.gov/pubmed/19393804

Schaible, L. M., Hughes, L. A., September 2008. Neighborhood Disadvantage and Reliance on the Police. Crime & Delinquency, 1–30.
URL http://cad.sagepub.com/cgi/doi/10.1177/0011128708322531

Steenberghen, T., Aerts, K., Thomas, I., 2009. Spatial clustering of events on a network. Journal of Transport Geography, 1–8.
URL http://linkinghub.elsevier.com/retrieve/pii/S0966692309001240

Steenberghen, T., Dufays, T., Thomas, I., Flahaut, B., March 2004. Intra-urban location and clustering of road accidents using GIS: a Belgian example. International Journal of Geographical Information Science 18 (2), 169–181.
URL http://www.informaworld.com/openurl?genre=article&doi=10.1080/13658810310001629619&magic=crossref||D404A21C5BB053405B1A640AFFD44AE3

Stone, M., Broughton, J., July 2003. Getting off your bike: cycling accidents in Great Britain in 1990-1999. Accident Analysis and Prevention 35 (4), 549–56.
URL http://www.ncbi.nlm.nih.gov/pubmed/12729818

Sustrans, 2009. Chapter 7 - Rural Roads.
URL http://www.sustrans.org.uk/assets/files/guidelines/RuralRoads.pdf

Teanby, D., February 1992. Underreporting of pedestrian road accidents. BMJ (Clinical research ed.) 304 (6824), 422.
URL http://www.ncbi.nlm.nih.gov/pubmed/1637379

UCLA: Academic Technology Services Statistical Consulting, 2010a. FAQ : What are pseudo R-squareds?
URL http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm01/03/2010

UCLA: Academic Technology Services Statistical Consulting, 2010b. Stata Annotated Output Negative Binomial Regression.
URL http://www.ats.ucla.edu/stat/stata/output/stata_nbreg%_output.htm

Wang, C., Quddus, M. A., Ison, S., 2009a. The effects of area-wide road speed and curvature on traffic casualties in England. Journal of Transport Geography 17 (5), 385–395.
URL http://linkinghub.elsevier.com/retrieve/pii/S0966692308000562

Wang, C., Quddus, M. A., Ison, S. G., 2009b. Impact of traffic congestion on road accidents: a spatial analysis of the M25 motorway in England. Accident Analysis and Prevention 41 (4), 798–808.
URL http://www.ncbi.nlm.nih.gov/pubmed/19540969

Wang, F., 2006. Quantitative Methods and Applications in GIS. Taylor & Francis, Florida.

Wang, X., Kockelman, K. M., 2009. Forecasting Network Data. Transportation Research Record: Journal of the Transportation Research Board 2105, 100–108.
URL http://trb.metapress.com/openurl.asp?genre=article&id=doi:10.3141/2105-13

Warsh, J., Rothman, L., Slater, M., Steverango, C., Howard, A., August 2009. Are school zones effective? An examination of motor vehicle versus child pedestrian crashes near schools. Injury prevention : Journal of the International Society for Child and Adolescent Injury Prevention 15 (4), 226–9.
URL http://www.ncbi.nlm.nih.gov/pubmed/19651993

Wedagama, D. M. P., Bird, R. N., Metcalfe, A. V., 2006. The influence of urban land-use on non-motorised transport casualties. Accident Analysis and Prevention 38 (6), 1049–57.
URL http://www.ncbi.nlm.nih.gov/pubmed/16876100

Xie, Z., Yan, J., 2008. Kernel Density Estimation of traffic accidents in a network space. Computers, Environment and Urban Systems 32 (5), 396–406.
URL http://linkinghub.elsevier.com/retrieve/pii/S0198971508000318