Effects of geodemographic profiles of drivers on their injury severity from traffic crashes using a multilevel mixed-effects ordered logit model

Mohammed Quddus
Professor of Intelligent Transport Systems
School of Civil and Building Engineering
Loughborough University
Loughborough LE11 3TU
United Kingdom
Tel: +44(0)1509 228545
E-mail : M.A.Quddus@lboro.ac.uk

Word Count: 6,997 words text + 2 Tables*250= 7,497

Revision date: *26 February 2015*

Abstract

The purpose of this paper is to examine various geodemographic factors on the levels of driver injury severity using a statistical model. A driver's geodemographic profile with respect to the involvement in a traffic crash consists of variables from multiple hierarchical levels such as drivers who are nested within crashes and crashes that are clustered within areas. A geodemographic profile of a driver therefore contains factors such as age, gender, residence of driver, social deprivation, and the distance from home to crash locations (at the driver-level); land-use patterns of crash location, casualties per crash and vehicles involved in the crash (at the crash- level); and vehicles per 1,000 population and population density (at the area-level). This implies that driver-level observations are correlated rather than independent as assumed in many injury severity modelling. In order to capture within-group and between-group correlations among observations a multilevel mixed-effects ordered logit model has been employed in this research. Mixed-effects allows some variables to vary by observations (i.e. random parameters). The analysis is based on UK national traffic crash data between 2009 to 2011 consisting of 271,654 drivers from 217,523 traffic crashes occurring across 27,773 different census areas. Data on area deprivation, Census, and land-use patterns were collected from multiple sources and integrated using a GIS framework. The results indicate that the severity of injuries sustained by urban drivers involved in crashes increases if they travel to rural areas; the level of driver injury severity also increases if traffic crashes occur in areas with high car ownership per capita; and drivers from more disadvantaged areas would sustain, *if all else are equal*, more severe injuries. The findings from this study would be useful to the Department for Transport and Local Authorities in formulating safety policies aimed at enhancing driver education, training and licensing programmes.

Keywords: geodemographic factors, shortest path algorithm, GIS, area deprivation, multilevel mixed-effects modelling

INTRODUCTION

Identifying microscopic and macroscopic factors affecting the injury severity of drivers/riders is an important area of research in road safety as policies and regulations formed to augment driver education, training and licensing are normally based on these factors. In developing a relationship between the level of injury severity and their contributing factors, it is important that an injury severity model employs a statistical model that takes into account both within-group and between group correlations arising from risk factors taken from multiple nested levels (e.g. drivers, crashes and areas).

Research on factors affecting the severity of traffic crashes is well established and rich. Initial studies have primarily focused on identifying individual driver-level factors that influence the severity of traffic crashes or the severity of driver injury *(*e.g. *1-8)*. This is perhaps due to the fact that road crashes are a human-made 'crisis' as drivers are thought to be responsible for solely, or in interactions with roadway environment and vehicles, about 93% of the traffic crashes in the United States *(*e.g. *9).* Researchers therefore have examined various driver-level factors that influence the severity of traffic crashes. Their primary objective has been to identify the most important factors with a view of developing safety policies and regulations aimed at improving drivers' awareness, training, education and licensing. These factors mainly comprise of age, gender, nationality, experience, income and driving habits such as speeding. In 1990s and early 2000s it was established that other factors such as weather conditions (e.g. snowing, raining and sunny), road geometry (e.g. gradient, curvature), traffic characteristics (e.g. speed, flow and density) and vehicle-level (e.g. vehicle age, type such as motorcycle/truck, weight and engine size) affect driver behaviours and attitudes. These factors were then considered in many studies as contributory factors in studying the severity of traffic crashes *(*e.g. *8, 10-19).*

Age and gender have been reported as important injury severity factors in traffic crashes *(*e.g. *20, 12, 21)* Young male drivers in the 17 to 25 year age group have found to be over-represented in fatal accidents *(22-23,21)*. Older drivers aged 65+ have found to have a mixed-effect on the level of crash severity *(18)*.

Numerous studies have highlighted that single vehicle crashes tend to be more severe than multiple vehicle crashes, especially in rural areas (e.g. *12, 25*). The rate of single-vehicle fatal crashes has been found to be relatively high in rural road networks relative to urban areas *(*e.g. *26-31)*. This is perhaps due to limited medical resources, high posted speed limits and drink driving in rural areas *(*e.g. *30, 31)*.

In recent years, various macroscopic-level factors have been considered in area-wide crash frequency modelling *(*e.g. *28, 32–35*). For example, Noland and Quddus *(33)* concluded that more severe pedestrian injuries are associated with areas of income deprivation and higher per capita expenditure on alcohol. Graham et al. *(32)* stated that the occurrence of child pedestrian crashes is higher in more deprived areas. The primary area-wide factors include population density, land-use patterns, car ownership, ethnicity and area deprivation.

There is however, a dearth of research on how various area-wide macroscopic factors may affect the severity of traffic crashes or traffic casualties (e.g. drivers). Various area-wide factors can be linked with a casualty or a crash through the merging of casualty-level/crash-level data with area-wide data. Road density may influence the level of crash (or driver injury) severity. In addition, if a crash database contains information on a casualty's home postcode then a range of area-wide factors relating to the residence of casualties can also be linked with the casualties. This is to understand how area-wide socio-demographic variables (e.g. area-wide social deprivation, land-use patterns of the casualties' homes of residence) may influence the level of casualty injury severity. There is a clear gap in knowledge on how various area-wide factors, while controlling for other factors, may affect the probability of a specific injury crash occurring. Linking data from multiple nested levels would assist in answering research questions: Would urban drivers be involved in more severe crashes when they travel to rural areas? Do more severe injury crashes occur on the roads that are far away from drivers' homes?

It would be interesting to develop a driver injury severity model that includes both microscopic-level (i.e. driver or crash-level) and macroscopic-level (i.e. area-level) variables. Therefore, the primary objective of this paper is to develop a comprehensive driver injury severity model that include drivers' *geodemographic conditions* such as sociodemographic factors including a driver's place of residence, home to crash locations in terms of land-use patterns and distance and mobility patterns (urban drivers travel to rural areas and vice versa), crash characteristics and area-wide factors.

DATA COLLECTION AND VARIABLE SELECTION

Road traffic crash data between 2009 and 2011 for England were obtained from the UK Department for Transport (DfT). The database consists of three files: (1) the first data file contains data on crash characteristics such as date/time of the crash, location of the crash reported as easting and northing coordinates and other road features, (2) the second file has the data on the vehicles involved in the crash, such as vehicle type, sex/age of the driver and driver home postcode and (3) the third file holds the data on casualty characteristics such as casualty class, severity of casualty and home postcodes of casualties. From 2009 to 2011, there were 469,442 crashes in England involving 856,243 vehicles and 634,744 casualties, of which 5,973 were fatalities (0.94%), 70,472 were serious injuries (11.1%) and the remaining casualties were slight injuries 87.96%).

Among the variables appearing in the crash database, the variable representing *casualty home postcode* is confidential and therefore, not publicly available. After signing a confidentiality agreement with the DfT, home postcode data of all casualties including drivers/riders were obtained. It should be noted that home postcode data however suffer from erroneous/missing observations. After comparing the casualties' home postcodes with the national postcode database for England (obtained from the Office for National Statistics, UK), it was revealed that 24% of the home postcodes are either missing or contain mistakes. Since home postcodes of drivers is one of the most important variables for the geodemographic analysis of drivers for injury severity, casualties with only valid home postcodes were retained for further analysis. This results in a total of 482,706 casualties with 0.85% fatalities, 10.8% serious injuries and 88.35% slight injuries. The centroid of a postcode is used as the home location for all the drivers with the same postcode. This allows us to calculate the distance between home location and crash location as reported in the crash database.

In order to investigate whether the distance from home to crash locations has any impact on the severity of driver's injury, distances from home to crash locations were calculated for 271,654 traffic crashes. Although as-the-crow-flies distances can easily be calculated from the pairs of home and crash coordinates, network-level distances are more accurate. Network-level distances were then calculated using Dijkstra's shortest path algorithm based on the concept of the fastest route between home and crash locations. The average as-the-crow-flies distance from home to crash locations is 13.8km. This increases to 17.8km if network-level distances (henceforth: *distance*) are considered. The distance between home to crash location follows a log-normal distribution in which the 75th percentile of distance was found to be 23km (with a mean of 10.8km) for fatal accidents (N1=2,479) and this reduces to 16.1km (with a mean of 7.7km) for serious injury accidents (N2=31,134) and 14.8km for slight injury accidents (N3=261,216).

In order to analyse drivers' geodemographic factors (e.g. age, socio-economic status, home location) on the severity of driver injury, drivers' home locations and the corresponding crash locations were superimposed on a boundary GIS map that represents land-use patterns in England in a GIS framework. The boundary map of land-use patterns was developed by the Department for Environment, Food and Rural Affairs *(36)*. In this map, 327 local authorities in England were classified into six urban/rural classifications. They are defined (in brief) as follows *(36)*:

- Major Urban (MU): districts with either 100,000 people or 50% of their population in urban areas with a population of more than 750,000;
- Large Urban (LU): districts with either 50,000 people or 50% of their population in one of 17 urban areas with a population between 250,000 and 750,000;
- Other Urban (OU): districts with fewer than 37,000 people or less than 26% of their population in rural settlements and larger market towns (RSLMT);
- Significant Rural (SR): districts with more than 37,000 people or more than 26% of their population in RSLMT;
- Rural-50 (R-50): districts with at least 50% but less than 80% of their population in RSLMT;
- Rural-80 (R-80): districts with at least 80% of their population in RSLMT.

For each of the 271,654 crashes used in this analysis, drivers' home locations were assigned to one of the six urban/rural classifications using GIS. This allows us to estimate the *index of concentration* commonly used in geodemographic analyses to measure a population's involvement in an activity *(26)*. This index is calculated as follows:

$$Index = \left(\frac{\%\ of\ population\ subgroup\ involved\ in\ traffic\ crashes}{\%\ of\ the\ subgroup\ population\ in\ the\ whole\ population}\right) \times 100$$

Table 1 is about here

An index value of 100 representing the characteristic of interest is uniformly distributed across the population subgroups *(26)*. Table 1a shows the calculated *index of concentration* by driver's injury severity category. As can be seen, 10.2% of the population lived in Rural-80 but 18.6% of the drivers involved in fatal crashes lived in Rural-80. This results in an index of 182 indicating that rural drivers' involvement in fatal crashes is much higher than one would normally expect, with the assumption that everyone in the whole population (15 and over) had the same tendency toward involvement in fatal crashes. The effect is reversed for the case of slight injury crashes. In contrast, urban drivers exhibit a lower-than-expected involvement in fatal crashes but a higher-than-expected involvement in injury crashes.

A cross-table between rural/urban categories of drivers' homes and rural/urban categories of crash locations revealed that the observed differences between urban and rural drivers' involvement in traffic crashes are statistically significant (see Table 1b).

Some geodemographic factors of drivers are available in the crash database including gender, age and trip purpose. Based on each of the drivers' home postcodes, an Index of Multiple Deprivation (IMD) ranging from 1 to 100 was derived for each driver. If an IMD score increases then the area becomes more deprived. This can be used in the model as a good proxy for a driver's geodemographic factor. Road density around the crash location may have an impact on the injury severity. This is calculated by dividing the total road lengths within a small census tract (i.e. lower layer super output area) where the crash had occurred by the area of the same census tract. The unit is km of road length per square km of area. The average road density for 271,654 crashes was 10.4km/km^2 (with a 75th percentile of 15.7km/km^2).

The variation in drivers' severity injuries can also be explained by characteristics of areas as traffic crashes occur in clusters. Moreover, injury severity from the crashes that happen in a particular area may be correlated rather than independent, due to shared land-use patterns, drivers' socio-demographic and traffic characteristics within the area. Therefore, various area-level factors that are invariant by crashes/drivers by variant by areas can also be included in the model. A commonly employed census tract - *lower layer super output areas* (LLSOA) – is applied in this study. There are in total 32,846 LLSOAs in England. Using a GIS, each of the traffic crashes was assigned to a LLSOA based on the geocoded crash location. This process may introduce errors in mapping crashes to a specific spatial unit due to the common boundary problem. However, a matching technique considering the direction(s) of the vehicle(s) just before the crash relative to the direction of the roadway segment (either clockwise or anti-clockwise) and the distance from the crash location to the segment was used to match the crash location onto the correct roadway segment as discussed in (*16*). These include: vehicles per 1,000 population, traffic density and traffic volume. These could be used as a proxy for exposure to crash severity. In the absence of LLSOA-level traffic data, the variable - *vehicles per 1000 population* by LLSOA is employed. This is obtained from the latest UK 2011 Census data.

STATISTICAL METHODS FOR DRIVER INJURY SEVERITY MODELLING

The objective here is to examine how both microscopic and macroscopic factors (termed as geodemographic factors) influence the severity of injuries sustained by drivers involved in crashes, given that crashes had occurred. The driver injury severity in England is recorded as a discrete and ordinal categorical variable representing three ordinal levels of severity categories such as fatal, serious and slight injuries. Since there is a clear definition of fatal, serious and slight injury casualties as detailed in *(37)* and property-damage only crashes are not reported, it is envisaged that driver injury severity levels may not suffer from common unobserved effects among adjacent injury categories. The literature on methodological approaches in modelling driver injury severity is very rich and established. A range of diverse statistical and non-statistical approaches has been employed to develop a relationship between injury severity and its contributing factors. The primary statistical approaches are: (1) ordered logit/probit models *(*e.g. *4, 7, 8)* and their various extensions such as generalised ordered logit/probit models *(16)* and mixed generalised ordered logit models *(38)* (2) multinomial logit models (e.g. *(3)*) and their extensions such as mixed logit models *(*e.g. *14, 39)*. For further details of these and other methodological approaches in modelling driver injury severity, readers are referred to a recent comprehensive review article by *(40)*.

The unit of analysis is the level of injury severity of a driver resulting from a traffic crash. This implies a possibility of having multiple observations (i.e. drivers) per crash. According to a comprehensive review article by *(40)*, if the injury severity level of crash-involved individuals (i.e. drivers) is considered as an unit of observation in the analysis, then it is essential to control the potential within crash correlation among observations. This suggests that the severity of injuries sustained by drivers involved in crashes would be correlated rather than independent, suggesting that inherent data structure generates dependency. One way to address this is the use of a multilevel model in which drivers' injury outcomes from a crash are allowed to be correlated (*41, 42*). A multilevel model has the capability to explicitly model complex variances and heterogeneity. In addition to fixed parameters estimated by an ordered logit model, there is an option within a multilevel model to let a parameter vary by observations (i.e. random parameter) resulting in a mixed-effects multilevel model. By considering all the advantages and disadvantages explained above, the appropriate model chosen for this study is a multilevel mixed-effect ordered logit model. There is however an inherent assumption – *parallel regression lines or proportional odds assumption* - in an ordered logit model *(16)*. If the assumption is violated for some of the covariates then a generalised ordered logit model can be employed.

A multilevel mixed-effects ordered logit model can be expressed as follows:

Let us consider a three-level model in which drivers are nested within traffic crashes, and traffic crashes are then nested within areas (e.g. a small census tract such as lower layer super output areas). Assume that there are a series of A independent geographical areas (i.e. k=1,2,....A) where area k contains k=1,2,...n_{jk} traffic crashes and there are also a series of C independent traffic crashes (j= 1,2,....C) where traffic crash j involves j=1,2,...n_{ijk} individual drivers. Y_{ijk}^{*} is the latent continuous response representing the levels of driver injury for driver *i*, traffic crash *j* and area *k* and this can be denoted as:

$$Y_{ijk}^{*} = \boldsymbol{X}_{ijk}\boldsymbol{\beta} + \boldsymbol{W}_{jk}\boldsymbol{\delta} + \boldsymbol{V}_{k}\boldsymbol{\gamma} + u_{jk} + v_{k} + e_{ijk} \quad (1)$$

In which:

$$p_{ijk} = \Pr(Y_{ijk}^{*})$$

$$e_{ijk} = \sum_{h=0}^{m1} e_{hijk} Z_{hijk}^{(1)}$$

$$u_{jk} = \sum_{h=0}^{m2} u_{hjk} Z_{hjk}^{(2)}$$

$$v_{k} = \sum_{h=0}^{m3} v_{hk} Z_{hk}^{(3)}$$

$$Z_0 = \{1\}\ i.e.\ a\ vector\ of\ 1's$$

$\boldsymbol{X}$, $\boldsymbol{W}$ and $\boldsymbol{V}$ are the fixed part explanatory variable design matrix for the first-level (i.e. drivers); second-level (crashes) and the third-level (areas) and their corresponding coefficients are $\boldsymbol{\beta}$, $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ respectively; $u_{jk} + v_k + e_{ijk}$ is the random part of the model in which $Z^{(1)},\ Z^{(2)}\ and\ Z^{(3)}$ are the explanatory variable design matrix for the first-level, second-level and third-level respectively, representing both random intercepts $i.e. Z_0 = \{1\}$ and random coefficients; $Z^{(1)}$ may be a subset of $\boldsymbol{X}$ and likewise $Z^{(2)}$ may be a subset of $\boldsymbol{W}$; $Z^{(3)}$ may be a subset of $\boldsymbol{V}$; e_{ijk} is a set of driver-level random effects (both random intercepts and random coefficients) in which e_{0ijk} (i.e. h=0) are the errors distributed as logistic function with mean 0 and variance $\frac{\pi^2}{3}$; u_{jk} is a set of crash-level random-intercept and random coefficients; $\boldsymbol{v}_k$ is a set of area-level random-intercept and random-coefficients. It is worthwhile stating that u_{jk} and v_k are not the parameters to be estimated but their variances and covariances need to be predicted. If $\boldsymbol{\Omega}_2$ and $\boldsymbol{\Omega}_3$ are the covariance matrix for the random coefficients then,

$$u_{jk} \sim MVN(\boldsymbol{0}, \boldsymbol{\Omega}_2);\ \ v_k \sim MVN(\boldsymbol{0}, \boldsymbol{\Omega}_3);$$

If m is the number of categories of the ordinal dependent variable, then the ordered observed outcomes (Y_{ijk}) can be generated from the latent continuous response as follows:

$$Y_{ijk} = \begin{cases} 1 & if\ Y_{ijk}^{*} \le \mu_1 \\ 2 & if\ \mu_1 < Y_{ijk}^{*} \le \mu_2 \\ \ldots & \\ \ldots & \\ m & if\ \mu_{m-1} < Y_{ijk}^{*} \end{cases}$$

Equation (1) can be re-written as:

$$logit(p_{ijk}) = \log\left[p_{ijk}/(1 - p_{ijk})\right] = \boldsymbol{X}_{ijk}\boldsymbol{\beta} + \boldsymbol{W}_{jk}\boldsymbol{\delta} + \boldsymbol{V}_{k}\boldsymbol{\gamma} + u_{jk} + v_{k} + e_{ijk} \quad (2)$$

In which

$$p_{ijk} = \Pr(Y_{ijk} = m)$$

As is noticeable, larger values of Y_{ijk} are corresponding to "higher" outcomes (e.g. fatal injury). μ_{1}, μ_2 and μ_{m-1} are the ancillary parameters (also known as cut-off points or thresholds) to be estimated. The cumulative probability of the injury severity outcome being in a category higher than m is:

$$\Pr\left(Y_{ijk} > m \middle| \boldsymbol{X}_{ijk}, \boldsymbol{W}_{jk}, \boldsymbol{V}_{k}, \mu, u_{jk}, v_k\right) = F\left(\boldsymbol{X}_{ijk}\boldsymbol{\beta} + \boldsymbol{W}_{jk}\boldsymbol{\delta} + \boldsymbol{V}_{k}\boldsymbol{\gamma} + u_{jk} + v_k + e_{ijk} - \mu_m\right) \quad (3)$$

In which h>0 in e_{ijk}

From equation (2), the probability of observing driver injury severity outcome *m* can be derived as:

$$\Pr\left(Y_{ijk} = m \middle| \mu, u_{jk}, v_k\right) = \Pr(\mu_{m-1} < (\boldsymbol{X}_{ijk}\boldsymbol{\beta} + \boldsymbol{W}_{jk}\boldsymbol{\delta} + \boldsymbol{V}_{k}\boldsymbol{\gamma} + u_{jk} + v_k + e_{ijk}) \le \mu_m)$$
$$= F\left(\mu_m - \boldsymbol{X}_{ijk}\boldsymbol{\beta} - \boldsymbol{W}_{jk}\boldsymbol{\delta} - \boldsymbol{V}_{k}\boldsymbol{\gamma} - u_{jk} - v_k\right) - F(\mu_{m-1} - \boldsymbol{X}_{ijk}\boldsymbol{\beta} - \boldsymbol{W}_{jk}\boldsymbol{\delta} - \boldsymbol{V}_{k}\boldsymbol{\gamma} - u_{jk} - v_k) \quad (4)$$

Special procedures are required to obtain satisfactory parameter estimates as there are more than one residual term. In order to estimate the parameters of the model presented in equation (1), it requires approximating the multivariate normal integrals by integrating out of all random effects. One widely-used method is the numerical integration using the mean-variance adaptive Gauss-Hermite quadrature technique *(43)*.

ESTIMATION RESULTS AND DISCUSSIONS

Multilevel modelling that can address a complex data structure as well as unobserved heterogeneity (i.e. severity injuries vary crash to crash and from neighbourhood to neighbourhood) was employed so as to develop a relationship (at the micro- and macro-levels at the same time) between driver injury severity and its contributing factors from each of the three levels. Most of the factors were taken from the driver-level representing their geodemographic conditions including age, gender, level of multiple deprivations at their home location, the distance between home to crash location, whether the driver was travelling from a rural area to an urban area. At the crash level, the factors considered were: whether a crash involved a single vehicle or multiple vehicles, number of casualties from the crash and surrounding road density where the crash had occurred. Finally, the variable - vehicles per 1,000 population was considered from the area-level.

A multilevel mixed-effects ordered logit model presented in equations (1) was estimated using data consisting of 261,462 individual drivers, whereby 230,801 traffic crashes occurred on 27,501 different areas. The results are presented in Table 2. The Brant test suggested by *(43)* was performed to see whether the proportional odds assumption was valid. This assumption was violated for some explanatory variables (i.e. single vehicle, speed limit, road type and trip purpose) but the differences in coefficients of these variables between the ordered logit model and the corresponding version of generalised ordered logit model were found to be less than 10%. Therefore, the multilevel ordered logit model was chosen as the most parsimonious and appropriate model. As outlined in Table 2, variances at the crash-level and area-level are statistically significant. Moreover, the log-likelihood ratio (LR) test indicates that a multilevel ordered logit model fits the data better than that of a single-level ordered logit model. Log-likelihood value at convergence has found to be much higher in the multilevel model relative to that of the single level model (see Table 2). The interpretation of variables is briefly discussed by hierarchy level:

Table 2 is about here

<u>Driver-level (Level-1) variables</u>

All driver-level variables were tested as random-parameters. None of the standard deviations of these random effects were found to be statistically significant at the 95% confidence level, indicating that the coefficients of driver-level variables do not change from crash to crash (i.e. fixed effects). The variables are interpreted as follows:

Driver travelling from an urban to a rural area or vice-versa: an important geodemographic factor of a driver relates to where s/he lives and where s/he is involved in traffic crashes. This has been captured through a linking variable indicating a home location to a crash location (i.e. home location → crash location) by land-use patterns. Would it be more dangerous for an urban driver to travel in a rural environment? This has been tested in the model presented in Table 2. Each of the driver-level observations was associated with two land-use areas: (1) relating to a driver's home and (2) relating to the crash location where the driver was involved in a crash. There are six land-use areas representing home location and six land use areas for crash location resulting in a total of 36 different linking scenarios. The interpretation of 36 dummy variables would be difficult and somewhat impractical. Six land-use areas were then combined into four; two for urban areas and two for rural areas as urban 1= MU, urban 2 = LU + OU, rural 1= R-50 + R-80 and rural 2 = SR. Therefore, a total of 16 dummy variables that represent the location of a driver's home and where s/he was involved in a crash. The linking variable representing that a driver was travelled from urban 1 (as his home location) and was then also involved in a crash in urban 1 (as crash location) (i.e. urban 1 → urban 1) was taken as the reference case. Half of the dummy variables were found to be statistically insignificant. If all else are equal, drivers from urban areas were found to have sustained more severe injuries from the crashes when they travelled to highly rural areas (i.e. rural 1) as both variables (i.e. urban1 → rural 1; urban 2 → rural 1) were found to be statistically significant at the 95% confidence level. This may be due to a unique feature of rural roads including unfamiliar and complex rural road environments in terms of large variation in posted speed limits among adjacent roads, irregular road topography and unpredictable non-uniform road users' behaviours. Drivers from rural areas (i.e. rural 1 and rural 2) were found to suffer more severe injuries from the crashes when they travelled within rural areas. Variables rural 1 → rural 1, rural 1 → rural 2, rural 2 → rural 1 and rural 2 → rural 2 were statistically significant with rural 2 → rural 1 providing the largest value of the coefficients. Odd ratios could also be employed in interpreting the values of the coefficients. For example, when Rural 2 drivers involved in crashes in Rural 1 areas the odds are exp(0.2625)=1.3 while when Rural 2 drivers involved in crashes in Rural 2 areas the odds are exp(0.0914)=1.1. In either ways, it is concluded that the level of driver injury severity tends to increase if traffic crashes occur in rural areas where traffic speeds tend to high. This is in-line with other existing studies *(*e.g. *26, 30).* Since travelling speeds have been controlled in the model through posted speed limits, rural location can be thought of a proxy for unique characteristics of rural road as discussed above. There is no significant difference in terms of the level of injury severity between rural and urban drivers in urban areas.

Distance from home to crash locations: since no evidence was found in the literature on how the distance (from a driver's home to a crash location) affects driver injury severity, a non-linear relationship (i.e. a quadratic) between the level of injury severity and the distance was investigated. Both linear and quadratic terms were found to be statistically significant at the 95% confidence level in which the linear term shows a negative coefficient, whereas the quadratic term exhibits a positive coefficient indicating that an approximate U-shaped relationship between the distance and driver severity. The probability of sustaining a fatal injury by a driver from a traffic crash would initially decrease with the increase in distance but then increase when the distance gets longer. The point of inflection on the effect of distance on the severity level was predicted to be 30 km if all other variables are kept constant at their means. A relatively large distance would normally indicate that the driver would travel to an unfamiliar road environment resulting in more severe crashes. This however needs to be carefully interpreted as 89% of the time driver injury severity has found to fall within a 'slight injury' category.

Socioeconomic factors: both age and sex of the driver were found to be statistically significant in the multilevel model. Unlike many existing studies that specified *age of the driver* to have a linear relationship with the level of injury severity, age was included as a linear and quadratic terms in this study. Both terms were found to be statistically significant. The linear terms shows a negative coefficient whereas the quadratic term shows a positive coefficient indicating that the level of injury severity is high for young and old drivers relative to middle-age drivers. Male drivers were found to be associated with more severe injuries if involved in a traffic crash compared to female drivers and this is in-line with existing studies *(*e.g. *21, 24)*. If all else is equal, the mean predicted probability of sustaining a serious injury by a female driver is 6.5% from 264,761 traffic crashes. The probability increases to 9.7% for the case of a male driver.

Index of multiple deprivation: a small area-wide (i.e. LLSOA) index of multiple deprivation ranging from 1 to 100 associated with a driver's home location was included in the model to see whether drivers from socially deprived areas are likely to sustain more severe injuries from traffic crashes. The variable was found to be marginally significant (at the 90% confidence level) with the expected positive sign. This means that drivers from more disadvantaged areas would sustain, *ceteris paribus*, more severe injuries. This finding is also in-line with existing studies *(*e.g. *32, 33)*.

Other controlling factors: a couple of other driver-level factors were included in the model as control variables. They were: trip purpose and type of vehicle driven by the driver. Both provided expected results.

Crash-level (Level-2) variables

As can be seen in Table (2), many crash-level variables were included in the model. The primary ones were: single vehicle, number of casualties and road density. It has been found that variables *single-vehicle crash* ((e.g. run-off-the-road crashes, hitting object on the carriageway) and *number of casualties per crash* were found to have random-effects on the levels of driver injury severity. In terms of the single vehicle crash, the mean value of the coefficient is 0.7715 and the standard deviation 0.1578. This means that the impact of single vehicle crash on the levels of driver injury severity varies by observation (i.e. drivers). Since the random-effects (i.e. $\boldsymbol{v}_k$ $in\ equation$ 1) is assumed to follow a normal distribution, none of the values of the coefficient (i.e. random parameter) is less than zero. This suggests that drivers are always more likely to sustain severe injuries in a traffic crash involving a single vehicle only (relative to a multiple vehicles crash) and the effect is variable by areas.

The variable – *number of casualties per crash* - was also found to have a random effect on the levels of driver injury severity. The average value of this random-parameter is +0.147 and the standard deviation is 0.126 implying that 88.7% of the (normal) distribution is greater than 0 and 11.3% of the distribution is less than 0. Therefore, for 88.7% of the traffic crashes, the probability of sustaining a fatal injury by a driver would increase if the number of casualties per crash increases. On the other hand, for 11.3% of the crashes, the probability of sustaining a fatal injury by a driver would decrease if the number of casualties per crash increases. Using the model presented in Table 2, the probability that driver injury severity from a crash would be in the 'serious' category has been predicted to be 7.4% (i.e. Pr (Y=2) = 0.074) when there is only one casualty per crash (i.e. only the driver is injured from the crash). The probability increases to 15.1% (i.e. Pr (Y=2) = 0.151) if there are at least five casualties per crash.

A range of other crash-level factors was included in the model as control variables. They are: road type, speed limit, temporal variables such as time of day, day of week, season of year. In most cases, these variables provided expected results. The time trend variable employed as year dummies was found to be statistically significant for 2010 (relative to 2009) but marginally significant for 2011 indicating that the severity injuries of drivers sustained from a crash reduce over time.

Area-level (Level-3) variable

One area-level variable – *vehicles per 1,000 population* - was included in the model as a control variable. The variable was found to be positively associated with driver injury severity. This finding is logical as areas with high vehicle ownership rate tend to be 'rural' where the level of more severe crashes is high relative to urban areas. A quadratic relationship between this variable and the severity score was also tested but found to be statistically insignificant.

CONCLUSIONS

In this research, a statistical relationship between various geodemographic factors of a driver and the levels of injury severity sustained by the driver from a traffic crash was developed. Comparison of driver injury severity influencing factors revealed important differences in the set of statistically significant variables and coefficient values between the two modelling approaches. The statistically significant values of the random-effects (intercepts at the crash and area-level and random variables) along with the better goodness-of-fit statistics indicate that the multilevel model was more appropriate highlighting that the control of within-group and between-group correlations is important in modelling driver injury severity. Statistically significant geodemographic factors were identified as area-wide car ownership, road density, social deprivation and land-use patterns of home to crash locations. Findings from the several factors at the driver- and crash-level such as urban drivers travelling to rural areas, distance between home to crash locations, single vehicle crash could be utilised by safety policy makers to formulate new regulations and laws aimed at enhancing driver safety. For instance, engineering interventions relating to speeding and some aspects of road design may be introduced to address the occurrence of single vehicle crashes, especially in rural areas. Urban drivers may be required to take driving lessons in rural areas before they can be awarded a license to drive. Future research may focus on an in-depth study (e.g. focus groups and interviews) relating to driver behaviours and attitudes while they drive in rural areas.

REFERENCES

1. Stephen P. Shao. Estimating car driver injury severity in car/tractor-trailer collisions. *Accident Analysis & Prevention*, Vol. 19, No. 3, 1987, pp. 207–218.
2. Levy, D. Youth and traffic safety: the effects of driving age, experience, and education. *Accident Analysis & Prevention*, Vol. 22, No. 4, 1990, pp. 327–334.
3. Shankar, V., and F. Mannering. An Exploratory Multinomial Logit Analysis of Single-Vehicle Motorcycle Accident Severity. *Journal of Safety Research*, Vol. 27, No. 3, 1996, pp. 183–194.
4. Duncan, C., A. Khattak, and F. Council. Applying the ordered probit model to injury severity in truck-passenger car rear-end collisions. *Transportation Research Record*, Vol. 1635, 1998, pp. 63–71.
5. Chang, L. Y., and F. Mannering. Analysis of injury severity and vehicle occupancy in truck- and non-truck-involved accidents. *Accident; analysis and prevention*, Vol. 31, No. 5, Sep. 1999, pp. 579–92.
6. Kockelman, K. M., and Y.-J. Kweon. Driver injury severity: an application of ordered probit models. *Accident; analysis and prevention*, Vol. 34, No. 3, May 2002, pp. 313–21.
7. Quddus, M. a, R. B. Noland, and H. C. Chin. An analysis of motorcycle injury and vehicle damage severity using ordered probit models. *Journal of safety research*, Vol. 33, No. 4, Jan. 2002, pp. 445–62.
8. Abdel-Aty, M. Analysis of driver injury severity levels at multiple locations using ordered probit models. *Journal of Safety Research*, Vol. 34, No. 5, Jan. 2003, pp. 597–603.
9. Rumar, K. The Role of Perceptual and Cognitive Filters in Observed Behavior. In *Human Behavior and Traffic Safety SE - 8* (L. Evans and R. Schwing, eds.), Springer US, pp. 151–170.

10. Elvik, R., P. Christensen, and A. Amundsen. Speed and road accidents: An evaluation of the Power Model. 2004.

11. Yamamoto, T., and V. N. Shankar. Bivariate ordered-response probit model of driver's and passenger's injury severities in collisions with fixed objects. *Accident; analysis and prevention*, Vol. 36, No. 5, Sep. 2004, pp. 869–76.

12. Islam, S., and F. Mannering. Driver aging and its effect on male and female single-vehicle accident injuries: some additional evidence. *Journal of safety research*, Vol. 37, No. 3, Jan. 2006, pp. 267–76.

13. Savolainen, P., and F. Mannering. Probabilistic models of motorcyclists' injury severities in single- and multi-vehicle crashes. *Accident; analysis and prevention*, Vol. 39, No. 5, Sep. 2007, pp. 955–63.

14. Milton, J. C., V. N. Shankar, and F. L. Mannering. Highway accident severities and the mixed logit model: an exploratory empirical analysis. *Accident; analysis and prevention*, Vol. 40, No. 1, Jan. 2008, pp. 260–6.

15. Xie, Y., Y. Zhang, and F. Liang. Crash Injury Severity Analysis Using Bayesian Ordered Probit Models. *Journal of Transportation Engineering*, Vol. 135, No. 1, Jan. 2009, pp. 18–25.

16. Quddus, M. a., C. Wang, and S. G. Ison. Road Traffic Congestion and Crash Severity: Econometric Analysis Using Ordered Response Models. *Journal of Transportation Engineering*, Vol. 136, No. 5, May 2010, pp. 424–435.

17. Yasmin, S., and N. Eluru. Evaluating alternate discrete outcome frameworks for modeling crash injury severity. *Accident; analysis and prevention*, Vol. 59, Oct. 2013, pp. 506–21.

18. Kim, J.-K., G. F. Ulfarsson, S. Kim, and V. N. Shankar. Driver-injury severity in single-vehicle crashes in California: A mixed logit analysis of heterogeneity due to age and gender. *Accident; analysis and prevention*, Vol. 50, Jan. 2013, pp. 1073–81.

19. Yu, R., and M. Abdel-Aty. Using hierarchical Bayesian binary probit models to analyze crash injury severity on high speed facilities with real-time traffic data. *Accident; analysis and prevention*, Vol. 62, Jan. 2013, pp. 161–7.

20. Ulfarsson, G. F., and F. L. Mannering. Differences in male and female injury severities in sport-utility vehicle, minivan, pickup and passenger car accidents. *Accident; analysis and prevention*, Vol. 36, No. 2, Mar. 2004, pp. 135–47.

21. Gray, R. C., M. a Quddus, and A. Evans. Injury severity analysis of accidents involving young male drivers in Great Britain. *Journal of safety research*, Vol. 39, No. 5, Jan. 2008, pp. 483–95.

22. Møller, M. An explorative study of the relationship between lifestyle and driving behaviour among young drivers. *Accident; analysis and prevention*, Vol. 36, No. 6, Nov. 2004, pp. 1081–8.

23. Mathijssen, M. P. M. Drink driving policy and road safety in the Netherlands: a retrospective analysis. *Transportation Research Part E: Logistics and Transportation Review*, Vol. 41, No. 5, Sep. 2005, pp. 395–408.

24. Clarke, D. D., P. Ward, C. Bartle, and W. Truman. Young driver accidents in the UK: the influence of age, experience, and time of day. *Accident; analysis and prevention*, Vol. 38, No. 5, Sep. 2006, pp. 871–8.

25. Xie, Y., K. Zhao, and N. Huynh. Analysis of driver injury severity in rural single-vehicle crashes. *Accident; analysis and prevention*, Vol. 47, Jul. 2012, pp. 36–44.

26. Blatt, J., and S. M. Furman. Residence location of drivers involved in fatal crashes. *Accident; analysis and prevention*, Vol. 30, No. 6, Nov. 1998, pp. 705–11.

27. Brown, L. H., A. Khanna, and R. C. Hunt. Rural vs urban motor vehicle crash death rates: 20 years of FARS data. *Prehospital Emergency Care*, Vol. 4, No. 1, 2000, pp. 7–13.

28. Zwerling, C., C. Peek-Asa, P. S. Whitten, S.-W. Choi, N. L. Sprince, and M. P. Jones. Fatal motor vehicle crashes in rural and urban areas: decomposing rates into contributing factors. *Injury prevention : journal of the International Society for Child and Adolescent Injury Prevention*, Vol. 11, No. 1, Feb. 2005, pp. 24–8.

29. Donaldson, A. E., L. J. Cook, C. B. Hutchings, and J. M. Dean. Crossing county lines: the impact of crash location and driver's residence on motor vehicle crash fatality. *Accident; analysis and prevention*, Vol. 38, No. 4, Jul. 2006, pp. 723–7.

30. Chen, H. Y., R. Q. Ivers, a L. C. Martiniuk, S. Boufous, T. Senserrick, M. Woodward, M. Stevenson, a Williamson, and R. Norton. Risk and type of crash among young drivers by rurality of residence: findings from the DRIVE Study. *Accident; analysis and prevention*, Vol. 41, No. 4, Jul. 2009, pp. 676–82.

31. Chen, H., R. Ivers, A. Martiniuk, S. Boufous, T. Senserrick, M. Woodward, M. Stevenson, and R. Norton. Socioeconomic status and risk of car crash injury, independent of place of residence and driving exposure: results from the DRIVE Study. *Journal of Epidemiology and Community Health*, 2010, pp. 1–16.

32. Graham, D., S. Glaister, and R. Anderson. The effects of area deprivation on the incidence of child and adult pedestrian casualties in England. *Accident; analysis and prevention*, Vol. 37, No. 1, Jan. 2005, pp. 125–35.

33. Noland, R. B., and M. A Quddus. A spatially disaggregate analysis of road casualties in England. *Accident; analysis and prevention*, Vol. 36, No. 6, Nov. 2004, pp. 973–84.

34. Aguero-Valverde, J., and P. P. Jovanis. Spatial analysis of fatal and injury crashes in Pennsylvania. *Accident; analysis and prevention*, Vol. 38, No. 3, May 2006, pp. 618–25.

35. Lee, J., M. Abdel-Aty, and K. Choi. Analysis of residence characteristics of at-fault drivers in traffic crashes. *Safety Science*, Vol. 68, Oct. 2014, pp. 6–13.

36. DEFRA Classification of Local Authorities in England Updated Technical Guide April 2009, pp. 2–14.

37. Wang, C. *The relationship between traffic congestion and road accidents: an econometric approach using GIS*. PhD Thesis, Loughborough University, 2010.

38. Eluru, N., C. R. Bhat, and D. a Hensher. A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. *Accident; analysis and prevention*, Vol. 40, No. 3, May 2008, pp. 1033–54.

39. Malyshkina, N. V, and F. L. Mannering. Empirical assessment of the impact of highway design exceptions on the frequency and severity of vehicle accidents. *Accident; analysis and prevention*, Vol. 42, No. 1, Jan. 2010, pp. 131–9.

40. Savolainen, P. T., F. L. Mannering, D. Lord, and M. a Quddus. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. *Accident; analysis and prevention*, Vol. 43, No. 5, Sep. 2011, pp. 1666–76.

41. Dupont, E., H. Martensen, E. Papadimitriou, and Yannis, G. Risk and Protection Factors in Fatal Accidents. Accident Analysis and Prevention, Vol. 42, No. 2, 2010, pp. 645-653.

42. Vanlaar, W. Multilevel modelling in traffic safety research: two empirical examples illustrating the consequences of ignoring hierarchies. Traffic Injury Prevention, Vol. 6, No. 4,2005, pp. 311-316

43. Liu, Q., and D. A. Pierce. A note on Gauss-Hermite quadrature. *Biometrika*, Vol. 81, No. 3, 1994, pp. 624–629.

44. Brant, R. Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, Vol. 46, No. 4, Dec. 1990, pp. 1171–8.

List of Tables

Table 1: Descriptive statistics of data used in the analysis

Table 1a: Index of concentration by severity

Land-use patterns	**Total population (15 years and over)**		**Home location of drivers involved in crashes in England (2009 - 2011)**								
			Fatal			**Serious**			**Slight**		
	N	%	N1	%	Index	N2	%	Index	N3	%	Index
MU	13,518,103	33.5	484	21.1	63	7,957	27.4	82	82,933	33.8	***101***
LU	5,449,633	13.5	268	11.7	86	4,209	14.5	***107***	34,820	14.2	***105***
OU	6,107,982	15.2	354	15.4	102	4,929	17.0	***112***	42,578	17.4	***115***
SR	5,460,158	13.5	399	17.4	***128***	4,255	14.6	***108***	33,572	13.7	***101***
R-50	5,644,250	14.0	363	15.8	***113***	3,800	13.1	93	26,685	10.9	78
R-80	4,116,421	10.2	426	18.6	***182***	3,910	13.5	132	24,705	10.1	99

Table 1b: Drivers' involvement in traffic crashes by land-use patterns

		Crash Location						
		MU	**LU**	**OU**	**SR**	**R-50**	**R-80**	Total
Home Location	**MU**	78,891	753	2,292	4,235	2,249	1,372	**89,792**
	LU	985	29,506	1,612	2,484	2,557	1,474	**38,618**
	OU	3,059	1,617	31,501	3,564	3,080	3,770	**46,591**
	SR	3,242	1,779	2,147	26,093	2,314	2,128	**37,703**
	R-50	1,433	1,756	2,018	2,403	20,320	2,444	**30,374**
	R-80	550	685	2,367	1,829	2,288	20,857	**28,576**
	Total	**88,160**	**36,096**	**41,937**	**40,608**	**32,808**	**32,045**	**271,654**

Pearson chi2(25) = 9.9e+05 Pr = 0.000

Table 2: Modelling results

Variables included in the models	Single-level Ordered Logit model		Multilevel Ordered Logit model	
Severity Score: 3=Fatal, 2=Serious, 1=Slight	**Coefficient**	**t-statistic**	**Coefficient**	**t-statistic**
Area-level variables				
Cars per 1,000 people in the area where the crash occurred	0.0005329	8.59	0.0006772	8.00
Crash-level variables				
Single vehicle crash (single = 1; multiple vehicles=0)	0.5803	33.66	0.7715	32.35
Casualties per crash	0.1164	18.65	0.147	17.90
Road density (km/km2) at the crash location	-0.0355	-10.16	-0.0451	-9.58
Road density squared at the crash location	0.0006584	6.22	0.0008362	5.94
Road type:				
Roundabout				
One way street	0.3764	5.11	0.4811	5.15
Dual carriageway	0.3712	10.31	0.4702	10.21
Single carriageway	0.5145	17.79	0.6314	17.01
Slip road	0.0528	0.69	0.0606	0.62
Speed limit:				
Less than 20 mph (Reference case)				
Speed limit 30 mph	0.0611	0.71	0.0911	0.83
Speed limit 40 mph	0.3378	3.84	0.4409	3.88
Speed limit 50 mph	0.5726	6.3	0.7402	6.28
Speed limit 60 mph	0.6811	7.78	0.8776	7.75
Speed limit 70 mph	0.5690	6.11	0.7665	6.37
Time of day:				
Early morning (midnight to 6:00am) (Reference case)				
Morning (6:01am to midday)	-0.7218	-24.73	-0.9234	-23.66
Afternoon (midday to 6:00pm)	-0.7354	-25.81	-0.9529	-24.87
Evening (6:01pm to midnight)	-0.5176	-17.82	-0.6768	-17.61
Day of week:				
Sunday (Reference)				
Monday	-0.1477	-5.72	-0.1877	-5.60
Tuesday	-0.1542	-6.04	-0.1921	-5.80
Wednesday	-0.1552	-6.11	-0.1927	-5.84
Thursday	-0.1298	-5.11	-0.1680	-5.09
Friday	-0.1581	-6.32	-0.2035	-6.27
Saturday	-0.0636	-2.5	-0.0747	-2.26
Quarter of year:				
Q1 (January - March) (Reference)				
Q2 (April - June)	0.0431	2.27	0.0473	1.93
Q3 (July - September)	0.0289	1.53	0.0386	1.69
Q4 (October - December)	-0.0450	-2.34	-0.0645	-2.62
Trend:				
Accidents in 2009 (Reference)				
Accidents in 2010	-0.0433	-2.7	-0.0501	-2.43
Accidents in 2011	-0.0282	-1.77	-0.0337	-1.64
Driver-level variables				
Index of multiple deprivation at home location	0.0019	1.28	0.0024	1.97

Distance in km (home to crash location)	-0.00114	-2.63	-0.0015349	-2.77
Distance squared (home to crash location)	1.75E-06	1.27	2.63E-06	1.75
Driver age (years)	-3.29E-03	-1.19	-0.0072909	-3.12
Driver age squared	0.0001997	9.95	0.0002861	11.08
Driver gender (male = 1; female=0)	0.2882	17.31	0.3390	16.46
Type of vehicle:				
Vehicle - Cycle (Reference)				
Vehicle - Motorcycle	0.3162	15.45	0.4241	15.55
Vehicle - Car	-1.5435	-72.09	-1.9690	-59.44
Vehicle - HGV	-1.2586	-30.96	-1.6730	-30.86
Trip purpose:				
Travelling as part of work (Reference)				
Commuting	0.1956	6.86	0.2486	6.91
Travelling to/from school	-0.1686	-2.06	-0.1864	-1.88
Other purposes	0.2988	12.64	0.3813	12.80
Home location - Crash location (land-use change)				
Urban1 - Urban 1 (Reference)				
Urban1 - Urban 2	0.0022	0.03	0.0165	0.19
Urban1 - Rural 1	0.3606	4.69	0.4423	4.33
Urban1 - Rural 2	0.0249	0.55	0.0451	0.75
Urban 2 - Urban 1	-0.0762	-1.17	-0.0929	-1.15
Urban 2 - Urban 2	0.1320	6.46	0.1429	5.28
Urban 2 - Rural 1	0.1712	3.74	0.2124	3.46
Urban 2 - Rural 2	0.0313	0.85	0.0235	0.49
Rural 1 - Urban 1	0.1929	1.24	0.2773	1.43
Rural 1 - Urban 2	0.1048	1.61	0.1088	1.30
Rural 1 - Rural 1	0.1522	5.14	0.1788	4.43
Rural 1 - Rural 2	0.1797	3.45	0.1856	2.73
Rural 2 - Urban 1	0.0408	0.72	0.0475	0.66
Rural 2 - Urban 2	0.1335	3.10	0.1653	2.98
Rural 2 - Rural 1	0.2625	5.49	0.2864	4.49
Rural 2 - Rural 2	0.0914	3.84	0.0944	2.96
Cutoff Threshold 1	2.2286	1.998006	2.8980	18.7
Cutoff Threshold 2	5.0790	4.84453	6.2800	38.47
Random-parameters (for mixed-effects):				
Standard deviation of constant at area level			0.3793	10.21
Standard deviation of constant at the crash level			1.4842	19.55
Standard deviation for single vehicle crash			0.1578	4.36
Standard deviation for casualties per crash			0.1261	3.5
Number of observations	261,462		261,462	
Number of groups: areas			27,501	
Average number of observations (i.e. drivers) per area			9.51 (min=1, max = 194)	
Number of groups: crashes			230,801	
Average number of observations (i.e. drivers) per crash			1.13 (min=1, max = 19)	
Log-likelihood at convergence	-87,878.58		-87,316.78	
Pseudo R-squared	0.11		0.21	