Spatial models in transport: a review and assessment of methodological issues

Chao Wang, Mohammed A Quddus*, Tim Ryley, Marcus Enoch, Lisa Davison

Transport Studies Group
School of Civil and Building Engineering
Loughborough University
Leicestershire LE11 3TU, UK

* Corresponding Author
Email: M.A.Quddus@lboro.ac.uk

Word Count: 6960 + 750 (2 tables and 1 figure)

15 November 2011

Paper submitted for presentation to the 91st Annual Meeting of the Transportation Research Board, Washington DC

ABSTRACT

Regression models that use spatial data are widely used in transport applications. Although they have proved to be valuable in identifying and understanding the contributory factors that influence phenomena like road traffic accidents and travel demand, there are a series of methodological issues that limit their usefulness.

This paper examines some of these issues in spatial models. The Modifiable Areal Unit Problem (MAUP), occurs where the statistical inference and interpretation derived from the zones alters due to changes in zone boundaries. Evidence from both the simulated and empirical data confirms the existence of MAUP. Other methodological issues are examined, including ecological fallacy, spatial dependency, and the problem of matching individual observations to the correct spatial units.

In order to avoid these problems, some strategies are recommended for addressing these issues in an integrated way. Finally conclusions and further research avenues are discussed at the end of this paper.

INTRODUCTION

This paper intends to review and assess methodological issues and provide future research directions for spatial models that are widely applied in transport studies. In the context of this paper, the term "spatial model" refers to a regression model that involves spatial data. Spatial models have been widely used in transport research, such as in modelling road accidents and travel demand, and exploring their contributory factors. As such this paper will mainly investigate the applications of spatial analysis in the area of accident and travel demand research as examples while discussing the primary methodological issues.

There has been some advances in spatial analysis in transport area during last few years, for instance, accounting for spatial correlation in accident analysis *(1)*. There are however some important research issues that have not been fully addressed. Specifically, the paper examines the Modifiable Areal Unit Problem (MAUP), ecological fallacy, spatial dependency, and the problem of matching individual observations to the correct spatial units, before recommending strategies for addressing these issues in an integrated way, providing conclusions and suggesting future research directions.

MODIFIABLE AREAL UNIT PROBLEM (MAUP)

The Modifiable Areal Unit Problem (MAUP) refers to situations that when the boundary of zones used in a spatial analysis changes, the statistical inference and interpretation derived from the zones is also different *(2)*. It is often the case that the definition of zones used in a spatial analysis is arbitrary and modifiable. For instance, zones are often defined based on political and administrative considerations (e.g. enumeration districts and electoral wards in the UK). The definition of the zones however may not have much geographic meaning, and therefore, the statistical inference based on the zones may also be questionable.

From the literature, it has been found that spatial analysis is based on various spatial units, such as regions *(3)*, counties *(1, 4-6)*, districts *(7, 8)*, English wards *(9-11)*, and enumeration districts that are much smaller than wards *(12)*. While analysing accident data in New Zealand, Haynes et al. *(13)* used the territorial local authority (TLA) as a spatial unit which is equivalent to districts in England and Wales. In addition to administrative units, Kim et al. *(14)* used artificial spatial units, i.e. uniform grid cells (each cell being approximately 0.259 km^2) for a study in Hawaii.

It should be noted however, that the MAUP problem is essentially a problem related to data aggregation, and it does not only apply to areas but also to road segments, since road segment level analysis also exhibits a similar "scale" problem. Unlike other disciplines (e.g. biology, epidemiology), spatial analysis in transport does not only concern 'zones' but also 'roads'. Therefore, it appears that the definition of road segments may also affect the statistical results due to the MAUP in a spatial analysis. There are mainly two methods to define a road segment in existing studies: the use of either fixed-length segments (i.e. equal length segments), or variable-length segments *(15)*. With studies undertaken at different spatial units, it is sensible to question if the statistical results hold when the spatial units (either areas or road segments) for aggregation are changed.

Openshaw *(2)* showed that under different aggregations of an Iowa dataset (containing data on the percentage of Republican voters and the percentage of elderly voters), the slope coefficients estimated from a linear regression vary. The slope coefficients range from -24 to 12 when different zoning systems were used where the given area was divided into a number of 12 zones (i.e. each zoning system comprises 12 zones).

Gehlke and Biehl *(16)* found that the correlation coefficient between two variables tends to increase when the level of aggregation of census tracts are higher. Fotheringham and

Wong *(17)* performed sensitivity analyses to explore the impact of MAUP on the stability of parameter estimates. They found strong evidence of unreliability in multivariate models compared to univariate or bivariate models. Changes in the parameter estimates were complex and unpredictable when the aggregation was changed: the relationship between mean family income and the percentage of elderly people in an area is consistently insignificant for data based on 800 zones, but consistently significant for data from 200 or fewer zones. Moreover, even when a constant number of zones were used, some zoning arrangements reported a positive association and some zoning arrangements reported a negative association between the two variables. Fotheringham *(18)* concluded that it is necessary to demonstrate that the statistical results hold regardless of the type of zoning system used, otherwise the results may be artefacts of the particular zoning system and may not reflect the actual underlying process.

Although the MAUP has been known about by geographers for decades *(2)*, this problem has seemingly received little attention in spatial analysis exercises in many fields, including transport. An exception is a study by Thomas *(19)* who investigated the 'size problem' while analysing road accident data in order to determine the extent to which the length of road segment affected the statistical results. By looking at accident data in Belgium, Thomas *(19)* found that the length of road segments has an influence on statistical distributions of accident data.

For an area-wide analysis, Miller *(20)* looked at the issues of MAUP in transport, especially the impact of MAUP on travel demand analysis. Miller *(20)* argued that the choice of aggregated geographic zones, such as traffic analysis zones (TAZ), could have an impact on parameter estimates and the goodness-of-fit of a spatial model. For instance, smaller zones could result in higher levels of apparent inter-zonal flow. A subsequent study by Horner and Murray *(21)* revealed that the percentage of excess commuting changes noticeably under different zoning schemes, in terms of both level of scale (number of zones) and zone specification (i.e. how zones are defined for a given scale).

We now demonstrate the problem of MAUP using simulated data. Here, we firstly generated a sequence of 2,000 random numbers (x); and then generated another sequence of random numbers (xp) which are assumed to be Poisson distributed i.e. $xp \sim Poisson(0.1 + 0.5x)$. Next, a Poisson model was tested with xp as a dependent variable and x as an independent variable. As expected, the value of intercept is around 0.1 and the value of coefficient estimated is close to 0.5. We then aggregated the 2000 'observations' into 300 random groups (zones) and further aggregated the 300 groups into 80 super groups. The coefficients and elasticities of x under different grouping systems are presented in Table 1.

Table 1 The coefficients and elasticities of *x* under different grouping systems using simulated data

	Aggregation 1			Aggregation 2			Aggregation 3		
	Coef	z value	Elasticity	Coef	z value	Elasticity	Coef	z value	Elasticity
x	0.49	45.64	2.44	0.03	66.39	1.03	0.01	78.2	0.89
Intercept	0.18	3.19		3.44	182.53		4.89	325.05	
Number of observations	2000			300			80		
Pseudo R squared	0.16			0.65			0.82		

As shown, the coefficient of x decreases from 0.49 to 0.01 with respect to the decrease in number of 'zones' (i.e. large spatial units). The elasticity also decreases from 2.44 to 0.89 accordingly. The model goodness-of-fit however increases with respect to aggregation. This simple test thus confirms the existence of MAUP, and therefore demonstrates that the choice of aggregations can seriously affect the modelling results and the associated statistical inferences.

160 The finding above is similar to what has been observed by Park and Saccomanno
161 *(22)*: they develop two linear regression models based on both aggregate and disaggregate
162 individual level data, finding that the coefficient decreased whereas R squared increased with
163 data aggregation. They argued that this change of statistical results is due to ecological
164 fallacy. Ecological fallacy can be viewed a special case of MAUP as discussed in the
165 following section in this paper.
166 The MAUP can also be illustrated using real-world transport data. Here, we use 2009
167 London Travel Demand Survey (LTDS) as an example to illustrate the effect of MAUP on
168 regression results. Trip generation (i.e. the number of trips per person per day) was modelled
169 as a dependent variable. Several explanatory variables were then included in the model
170 relating to socio-demographic factors such as population and employment density, ethnicity
171 and age cohort, which were obtained from the UK Census 2001. Public transport stations data
172 were obtained from Transport for London; the 2007 index of multiple deprivation (IMD) was
173 obtained from the UK Communities and Local Government (CLG); the boundaries of various
174 spatial units were obtained from the EDINA UKBORDERS; and the trip rates were modelled
175 using a linear regression model. The modelling results at the Lower Layer Super Output Area
176 (LSOA), ward and district levels are presented in Table 2.

177 **Table 2 The coefficients under different aggregations using real world data**

	LSOA		Ward		District	
	Coef	z value	Coef	z value	Coef	z value
Population density (/km^2)	0.00001	1.07	0.00004	1.09	-0.0001	-0.38
Employment density (/km^2)	-0.000005	-0.22	-0.00003	-0.47	0.00017	0.30
Proportion of male	-2.551	-3.85	1.116	0.65	6.201	0.60
IMD score	-0.009	-4.73	-0.013	-3.09	-0.002	-0.09
Number of cars per person (for age 16-74)	0.769	3.90	0.894	2.26	1.087	0.77
Proportion of people who are white	0.357	3.90	0.332	1.93	0.634	0.71
Proportion of people in 0-15	Reference case					
Proportion of people in 16-24	1.065	2.09	0.646	0.5	4.687	0.45
Proportion of people in 25-44	1.020	2.33	1.350	1.19	4.201	0.61
Proportion of people in 45-64	0.165	0.29	-1.011	-0.7	1.988	0.24
Proportion of people in 65-84	-0.717	-1.31	-0.790	-0.56	-1.717	-0.16
Proportion of people in 85+	0.769	0.55	6.286	1.56	26.466	1.05
Proportion of people who mainly work at home compared to all people aged 16-74 & in employment	3.847	7.24	5.010	4.2	7.159	1.05
Number of bus stops	-0.002	-0.50	0.001	0.45	-0.00003	-0.09
Number of train stations	0.056	1.24	0.004	0.14	0.011	0.80
Number of tube stations	0.073	1.58	0.041	1.59	0.015	0.90
Total area (km^2)	-0.020	-1.94	-0.001	-0.13	-0.001	-0.28
Intercept	2.406	5.28	0.546	0.5	-4.902	-0.77
Number of observations	5285		796		50	
R squared	0.102		0.263		0.632	

From Table 2 it can be seen that, using the same sources of data, the regression results have changed after aggregation (from LSOA to district). Some of the variables with notable changes have been highlighted. For example, the coefficient for *proportion of male* is negative and significant at LSOA level, yet it becomes positive and insignificant at ward and district levels. Generally it can be observed that the coefficients become less significant at a higher aggregation level, which confirms the findings from the simulated data as shown above. This could be because at higher aggregation, the spatial units become less homogenous and thus the effect of an independent variable would be less statistically significant. It is also interesting to note that, although the results appear to be weakest at the district level in terms of model inference (as none of the coefficients were significant), the model appears to be the best in terms of goodness-of-fit (the R squared value is the highest at the district level). Once again, this phenomenon is consistent with the simulated data.

From this, both the simulated and real world data confirm the existence of MAUP. The solution to the MAUP is, however, not straightforward. Openshaw *(2)* suggested that MAUP is "a geographical problem that requires a geographical rather than a statistical solution". Arbia *(23)* also stated that "no systematic result is found, and the general belief is that aggregation variability cannot be controlled for through a statistical approach."

Fotheringham *(18)* suggested that a possible solution to the MAUP is, as demonstrated by Fotheringham and Wong *(17)*, a sensitivity analysis which tests the parameter estimates based on a variety of zoning systems. A similar technique is location-allocation modelling which is also computationally intensive. Recently, Zhang and Kukadia *(24)* employed sensitivity analysis and concluded that a grid with a size of 0.5 miles is the most suitable zoning method; while looking at commuting data, while Horner and Murray *(21)* suggested that zonal data should be "as disaggregate as possible".

Meanwhile Anselin *(25)* stated, with regards to MAUP, that "unless there is a homogeneous spatial process underlying the data, any aggregation will tend to be misleading". Thus, a better spatial clustering may be required to ease the MAUP. Yannis et al. *(26)* applied a multilevel model (i.e. counties nested within regions) in examining the regional effect of police enforcement on road accidents. While using ad hoc geographic clustering scheme, they found that there were some inconsistency in spatial pattern, which they suggest may be due to the MAUP. Alternatively they developed a mathematical clustering scheme based on spatial homogeneity in demographic and transport characteristics, and a consistent result was found. From the work of Yannis et al. *(26)*, it may be suggested that the statistical results may be more reliable if spatial units used are more homogeneous. Openshaw *(2)* also pointed that, a homogeneous zoning or grouping system would eliminate the problem of ecological fallacy which is closely related to MAUP as discussed below.

To summarise, it appears that to reduce the effect of MAUP, a range of different spatial aggregations need to be tested. Such aggregations could well be based on ad hoc geographic or mathematical clustering. It is possible that by testing different spatial aggregations (area-wide and/or road segment level), an upper and lower limit of parameter estimates could be obtained, rather than point estimates. Also a better mathematical clustering may be desirable so as to make spatial units more homogeneous. Clearly more empirical evidence is required to ascertain the suitable method of spatial aggregation.

ECOLOGICAL FALLACY

Another important issue in spatial analysis is the so-called ecological fallacy. Closely related to the MAUP, the ecological fallacy states that results based on grouped aggregate data may not be applied to the individual units that form the aggregate dataset *(2)*. The origin of the ecological fallacy was Robinson *(27)* who examined the literacy rate and proportion of

immigrants in 48 states in the USA and found that states with a higher rate of immigrants had a higher literacy rate. However, if immigrants were considered individually, then they were on average less literate than native citizens. This might be because immigrants tended to settle in a state where native citizens are more literate. Overall then, this problem occurs when data in an area is (wrongly) assumed to be homogeneous *(2)*.

In the MAUP, as stated above, the scale or size of zonal aggregations may affect the statistical modelling results. Based on this, it can be inferred that statistical results at an aggregate level may also be different from the results drawn at an individual level (i.e. the scale/size of an observing spatial unit is an individual). The 'individuals' however could be zones prior to an aggregation or non-modifiable entities *(2)*.

There are some efforts in transport studies to identify and correct ecological fallacy, notably done by Davis *(28, 29)*. Davis *(28)* looked at the link between speed variance and road accidents, and concluded that although a positive relationship between speed variance and accident rate was observed at an aggregate level, this observation is however uninformative, and thus is subject to ecological fallacy.

Davis *(29)* later investigated possible aggregation biases in modelling road accidents. By simulating pedestrian accident risk, it was determined that while there is a clear positive (increasing) relationship between accident risk and traffic speed, this relationship was weakened when data are aggregated, which is an example of the effects of ecological fallacy. As stated by Davis *(29)*, this aggregation bias can be largely avoided by using an appropriate structured sample in a statistical analysis, although this could be very difficult, if not impossible, as there is usually no prior knowledge of the underlying mechanism generating the aggregate data.

As stated above, a completely homogeneous aggregation has no risk of ecological fallacy. Therefore, as in MAUP, a better clustering scheme based on traffic and various factors may be used to reduce the potential impact. Davis *(29)* also suggested that the problem of ecological fallacy can be avoided using a better structured sample.

Because of the aggregation bias, Davis *(29)* advocated analysing at an individual accident level using a mechanism approach to develop a causal relationship. Alternatively, traditional statistical models can be applied in an individual accident level analysis using data from matched sampling of crash and non-crash cases *(30)* or ideally, using naturalistic driving event data *(31)*.

Although an individual level analysis may derive a causal relationship, there are some problems as well. Similar to ecological fallacy, a disaggregate level analysis may be subject to ***atomistic fallacy*** which refers to the fallacy of drawing inferences at aggregate level based on individual level data and is the counterpart of the ecological fallacy *(32)*. Diez-Roux *(32)* explored the possible atomistic fallacy in epidemiology studies and claimed that 'macro-level' variables can provide information that is not captured by individual level data. For instance, it is likely that a school child's travel behaviour does not only depend on the characteristics of the individual (or the household), but also on the culture of the local community and the school travel policy of the local authority. Some of the 'macro-level' variables (e.g. policies, funding and regulations by the local government) may be difficult, if not impossible, to observe at individual level.

Researchers in transport studies have also observed some evidence of the atomistic fallacy. Strong evidence supports the idea that a disaggregate level analysis cannot take into account "system-wide effects" *(5, 7, 33, 34)*. Thus, factors such as land use, population and employment density, road density, access to medical care, local drinking cultures and level of deprivation cannot be easily incorporated into a disaggregate level analysis. For example, Noland and Quddus *(35)* investigated the effect of improvements in medical care on road accidents in Great Britain, finding that improvement in medical services significantly reduces

road fatalities, which is to be expected. Yet it would be difficult to imagine incorporating the quality of local medical care (e.g. waiting times for hospital treatment; number of hospitals in an area) into an accident reconstruction approach at an individual accident level.

Therefore, regardless of the potential problem of ecological fallacy, an aggregate level analysis still has its benefits and is necessary (e.g. to control for system-wide effects). To reduce the error caused by ecological fallacy, as stated above, an improved sampling/clustering scheme is required. Similar to the possible solutions to the MAUP, different aggregations, probably based on spatial homogeneity in demographic and transport characteristics, can be tested so as to obtain a reliable parameter estimate in a spatial analysis. In terms of modelling technique, a multilevel modelling framework (discussed below) seems to provide the best balance in terms of ecological fallacy and atomistic fallacy. This is because while multilevel models can be conducted at an aggregate level, the data are explicitly defined in a hierarchical structure, so spatial units with similar characteristics can be grouped and modelled.

SPATIAL DEPENDENCY

There has been a significant effort in recent transport research to account for spatial dependency (or spatial correlation) in a spatial analysis. As pointed out by LeSage *(36)*, two problems arise when data has a locational dimension: 1) spatial correlation exists among the observations, and 2) spatial heterogeneity occurs in the relationships that are modelled. This is in-line with the first law of geography which states that "everything is related to everything else, but near things are more related than distant things" *(37)*.

Studies in other disciplines have developed methods using spatial econometrics to address the issue of unmeasured spatial correlation among neighbouring spatial units *(38)*. Such studies have primarily been based on a Bayesian framework in which conditional autoregressive (CAR) models are often employed to take into account spatial dependence among neighbouring spatial units. This method was initially used in ecological analysis and disease mapping. The basic idea of disease mapping is to locate environmental hazards and groups of people and then allocate scarce resources, so it is useful to assess environmental justice *(39)*.

The application of the CAR model in transport research is relatively new, with Miaou et al. *(1)* apparently being the first to use CAR models to deal with spatial and temporal correlation in road traffic accident data. Later studies have been devoted to using CAR models under Bayesian framework *(6, 40-48)*.

It should be noted that in addition to the CAR model mentioned above, there are several other spatial models that can take into account the effects of spatial correlation, such as the spatial filter model and the simultaneous autoregressive (SAR) model *(49)*. The spatial filter model is involved with regressing a variable (e.g. accidents) on a set of synthetic variates representing distinct map patterns that accounts for spatial autocorrelation *(49)*. This approach (spatial filtering) was used in disease mapping *(49, 50)*. It would therefore be interesting to apply the spatial filtering technique to a transport research context (such as road safety).

As for the SAR model, it has been widely used in spatial econometrics and some transport literature *(44, 51, 52)*. As discussed by Banerjee et al. *(53)*, any SAR model can be represented by a CAR model although the converse is not true. Therefore it seems that a CAR model is more flexible and powerful. Quddus *(44)* illustrated this by comparing two types of SAR models with a CAR model using London accident data. However, the SAR model used by Quddus *(44)* is a linear model which is suitable for continuous data; and the CAR used by the author is a Poisson based model which is more suitable for count data in the study, and so

it would be useful to develop a SAR count data model so that a more tightly focused comparison could be conducted.

One problem both CAR and SAR models encountered is specifying an appropriate neighbouring structure where a weight matrix is used. Different choices of weight matrix result in varying degrees of spatial associations *(25)*. There are several methods to define the weights (w_{ij}) between spatial units depending on the consideration of different neighbour structures. The weighting scheme could use contiguity based weights, for example, $w_{ij} = 1$ if spatial unit *i* and *j* are adjacent (i.e. shared border and/or vertex) and $w_{ij} = 0$ otherwise. Alternatively, distance based weights could be used. For example, the shorter the distance between *i* and *j*, the larger the weight (w_{ij}) (i.e. distance decay); or $w_{ij} = 1$ if spatial unit *i* and *j* are within 1km (i.e. distance order).

Aguero-Valverde and Jovanis *(54)* examined various neighbouring structures at a road segment level using a CAR model . They empirically tested different neighbouring structures including contiguity based, distance order, distance-exponential decay and contiguity based structures which integrate route information (i.e. higher weights assigned to segments if they belong to the same route). They found that pure distance based neighbouring models (i.e. exponential decay) did not perform as well as other neighbouring structures, and that route information also improves model performance.

In addition, as suggested by Aguero-Valverde and Jovanis *(54)*, other neighbouring structures need to be examined, such as weights that are inversely proportional to the distance, different decay functions and combining distance order and route information.

As pointed out by El-Basyouny and Sayed *(46)*, the spatially correlated effects may be simply be due to the omission of important variables, meaning that the spatial effects would diminish with inclusion of pertinent covariates in the model. Therefore, it seems spatially correlated effects may be more important to consider when a lack of appropriate explanatory variables in the models, and so it would be then interesting to test how well different neighbouring structures perform in this situation. Clearly then, more research efforts are required for investigating different neighbouring structures for both road segment and area-wide level analyses.

Another source of spatial dependency is "within group" correlation. Spatial units are often naturally clustered by various levels of spatial units, such as ward, district, county, and region. It is expected that spatial units within the same group share similar characteristics and so would be correlated. For instance, for a ward level analysis, it is expected that ward-level variables within the same county have similar properties. To address this "within group" correlation, multilevel models are proposed and applied *(26, 54, 55)*. In addition, it is possible to incorporate both "within group" correlation and spatial autocorrelation in a single model *(47, 54)*, to better control for the various spatial correlations.

MATCHING INDIVIDUAL OBSERVATIONS TO THE CORRECT SPATIAL UNITS

In a spatial analysis, it is often required to link one spatial unit to another spatial unit. For instance, in an area-wide accident frequency analysis it is essential to obtain the correct count of accidents in an area, which requires accidents to be assigned to correct areas. In an individual level travel demand analysis, it may be necessary to know in which area the individual lives, so that area-wide factors affecting individuals' travel demand (such as crime rate) can be controlled for. Matching individual observations to the correct spatial units is thus important as it ensures a high quality of data.

Geo-coded data however usually come from different sources, and there may be inconsistencies in terms of the quality of spatial data among these different data sources thus

causing a spatial mismatch problem. For instance, it was found that when accident data is overlaid onto motorway road segment data, mismatches between them are observed *(56)*.

A similar situation may also occur in an area-wide spatial analysis, an example of which is illustrated in Figure 1.

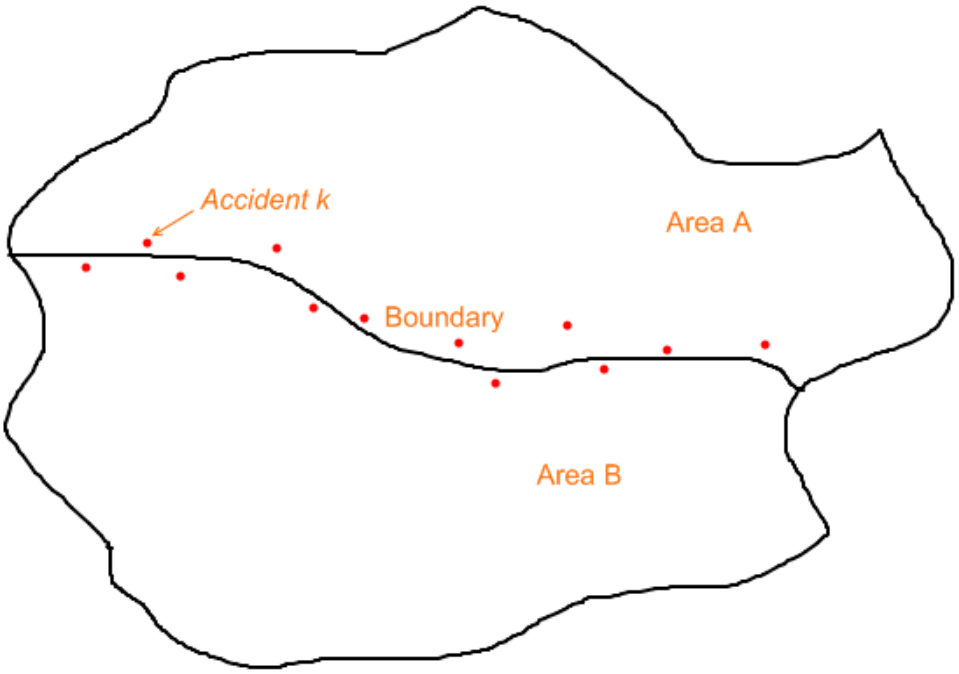

Figure 1 An example showing the accidents and boundaries

In Figure 1, the dots show the location of road accidents where there are two spatial units (i.e. area A and area B) sharing a common boundary which is the centre-line of a road. The purpose is to correctly count the total number of accidents in each of these two areas, which is important in an accident frequency analysis. However, due to possible errors (or inconsistencies) in both geo-coded accident data and the area boundary data, there may be mismatch between them and so accidents may not be assigned to the correct area. In other words, some of the accidents that have actually occurred in area A could be assigned to area B and vice-versa. Therefore, the total count of accidents for each area may be incorrect with potential implications regarding the accuracy of any use of that data in accident prediction models or severity analyses for example.

In practice though, most studies appear to have ignored this mismatch problem and have assumed that the location of different spatial units is accurate. An exception is a map-matching technique that was proposed by Wang et al. *(56)* to match the accident points onto the correct *road segment*. This mapping method used a weighting score of the perpendicular distance and the direction of the vehicle relative to a road segment to assign accidents to the correct road segments. It should be noted that this map-matching technique is more suitable for major roads such as motorways, as other roads, especially minor roads in dense urban areas, are significantly more complex in terms of road curvature and number of junctions meaning that it may be rather more difficult to apply this technique. In addition, it is yet to develop a map-matching method in the context of an area-wide analysis (e.g. assigning accidents to the correct area). A sensitivity analysis may also need to be employed to test if

modelling results vary significantly under a different matching technique, for example what if accidents near the boundary were assigned to areas randomly; or accidents were all assigned to area A/B.

Another mechanism for addressing this mismatch problem is demonstrated by Tarko et al. *(57)*, who applied a probabilistic linking technique so as to link the records of two databases. Previously this technique was mainly used in medical research *(57)*.

STRATEGIES OF ANALYSING SPATIAL DATA

It is apparent from the above discussions that a good strategy is required to take account of the various issues involved in a spatial analysis. This section aims to provide some recommendations to address the various issues raised to help enhance the robustness of spatial modelling results.

Data quality: one needs to carefully check the data and match individual observations to the correct spatial units. Some techniques that are described above or similar techniques can be used. The statistical analysis and inferences derived could be significantly improved with better quality and accurate data.

MAUP and ecological fallacy: as discussed, these issues are related to homogeneity in the data – MAUP and ecological fallacy are closely related, and their impact can be greatly eliminated if the data are homogeneous. Therefore it may be worth to check whether the data in question is homogenous or whether such assumption could be made. If this is not the case, as discussed above, a combination of different spatial aggregations can be tested. Uniform grid cells with an 'optimal' size (such as 0.5 mile as suggested by Zhang and Kukadia, *(24)*) can also be used. These aggregations can be combined with the multilevel modelling technique which groups clustered data. It should be kept in mind though, that not all spatial data are available at all spatial levels. For example, census data are often provided for spatial units with a pre-defined political boundary, meaning that uniform grid cells may not applicable to this scenario.

Spatial dependency: finally in terms of spatial dependency, one needs to look carefully at the type of data being dealt with. As El-Basyouny and Sayed *(46)* implied, it may not be necessary to employ a spatial model that accounts for spatial correlation if the observed data is sufficient to control for the spatial variations. Therefore it may be useful to test whether the data is subject to spatial correlation using techniques such as Moran's *I* test *(53)*. If such tests suggest there may be spatial correlation in the data even after controlling for all the explanatory variables, it may be necessary to employ a spatial model (e.g. a CAR model). The spatial models can be combined with the multilevel modelling techniques to best address the spatial correlation and data clustering.

CONCLUSIONS AND FUTURE RESEARCH

This paper has discussed several important methodological issues relating to spatial analyses employed in transport including the Modifiable Areal Unit Problem (MAUP), ecological fallacy, spatial dependency, and the problem of matching individual observations to the correct spatial units. For the MAUP, it has been argued that a sensitivity analysis could be employed to test the modelling results under different aggregations, to obtain upper and lower limits of parameter estimates. A better mathematical clustering based on the demographic and transport characteristics may be also be necessary to ease the MAUP. This is followed by ecological fallacy, which is closely related to the MAUP. The limitations and advantages of an aggregate level analysis have been presented. It has been argued that, although a disaggregate level analysis is often desirable because of ecological fallacy, an aggregate level

analysis is necessary as it can take into account "system-wide effects". It has been proposed that in order to reduce the error caused by ecological fallacy, a better sampling/clustering scheme based on spatial homogeneity in demographic and transport characteristics may be employed, so as to obtain a reliable parameter estimate. The multilevel modelling framework may provide the best balance in terms of ecological fallacy and atomistic fallacy.

As for spatial dependency, a significant research effort has been devoted to this research area in the past decade, notably the application of conditional autoregressive (CAR) and simultaneous autoregressive (SAR) models. There are, however, avenues to further test the performance of different specifications of neighbouring structures and weight matrix. Since spatial data is often clustered, multilevel models have been proposed and applied in previous research to accommodate the "within group" correlation.

This paper has also discussed a common problem related to matching individual observations to the correct spatial units due to inconsistency between different data sources. It is crucial to address this problem so as to improve data quality. While a weighting score method has been proposed and used for major road segments in a previous study, it has yet been investigated how to undertake map-matching for the case of minor roads and areas.

Finally, this paper has offered a range of strategies for analysing spatial data that should benefit both researchers and practitioners by means of improving the data quality, modelling results, and the statistical inferences drawn.

While the methodological issues discussed in this paper are applicable to a broad range of transport research areas, some of which however appear to be less studied using advanced spatial methods, for example estimating travel demand of demand responsive transport (DRT) or paratransit. Therefore future research is required to apply and improve spatial analysis in these research areas.

ACKNOWLEDGEMENTS

Thanks are due to the Engineering and Physical Science Research Council (EPSRC) for their funding of the project Developing Relevant Tools for Demand Responsive Transport (see www.drtfordrt.org.uk).

REFERENCES

1. Miaou, S., J. J. Song, and B. Mallick. Roadway Traffic Crash Mapping: A Space-Time Modeling Approach. *Journal of Transportation and Statistics*, Vol. 6, No. 1, 2003, pp. 33-57.

2. Openshaw, S. The Modifiable Areal Unit Problem. *Concepts and Techniques in Modern Geography*, No. 38, 1984.

3. Washington, S., J. Metarko, I. Fomunung, R. Ross, F. Julian, and E. Moran. An Inter-Regional Comparison: Fatal Crashes in the Southeastern and Non-Southeastern United States: Preliminary Findings. *Accident Analysis & Prevention*, Vol. 31, No. 1-2, 1999, pp. 135-146.

4. Amoros, E., J. L. Martin, and B. Laumon. Comparison of Road Crashes Incidence and Severity between some French Counties. *Accident Analysis & Prevention,* Vol. 35, No. 4, 2003, pp. 537-547.

5. Noland, R. B., and L. Oh. The Effect of Infrastructure and Demographic Change on Traffic-Related Fatalities and Crashes: A Case Study of Illinois County-Level Data. *Accident Analysis & Prevention,* Vol. 36, No. 4, 2004, pp. 525-532.

500 6. Aguero-Valverde, J., and P. P. Jovanis. Spatial Analysis of Fatal and Injury Crashes in
501 Pennsylvania. *Accident Analysis & Prevention*, Vol. 38, No. 3, 2006, pp. 618-625.

502 7. Haynes, R., A. Jones, V. Kennedy, I. Harvey, and T. Jewell. District Variations in Road
503 Curvature in England and Wales and their Association with Road-Traffic Crashes.
504 *Environment and Planning A*, Vol. 39, No. 5, 2007, pp. 1222-1237.

505 8. Jones, A. P., R. Haynes, V. Kennedy, I. M. Harvey, T. Jewell, and D. Lea. Geographical
506 Variations in Mortality and Morbidity from Road Traffic Accidents in England and Wales.
507 *Health & Place*, Vol. 14, No. 3, 2008, pp. 519-535.

508 9. Graham, D. J., and S. Glaister. Spatial Variation in Road Pedestrian Casualties: The Role
509 of Urban Scale, Density and Land-use Mix. *Urban Studies*, Vol. 40, No. 8, 2003, pp. 1591-
510 1607.

511 10. Noland, R. B., and M. A. Quddus. A Spatially Disaggregate Analysis of Road Casualties
512 in England. *Accident Analysis & Prevention*, Vol. 36, No. 6, 2004, pp. 973-984.

513 11. Graham, D., S. Glaister, and R. Anderson. The Effects of Area Deprivation on the
514 Incidence of Child and Adult Pedestrian Casualties in England. *Accident Analysis &*
515 *Prevention*, Vol. 37, No. 1, 2005, pp. 125-135.

516 12. Noland, R. B., and M. A. Quddus. Congestion and Safety: A Spatial Analysis of London.
517 *Transportation Research Part A: Policy and Practice*, Vol. 39, No. 7-9, 2005, pp. 737-754.

518 13. Haynes, R., I. R. Lake, S. Kingham, C. E. Sabel, J. Pearce, and R. Barnett. The Influence
519 of Road Curvature on Fatal Crashes in New Zealand. *Accident Analysis & Prevention*, Vol.
520 40, No. 3, 2008, pp. 843-850.

521 14. Kim, K., I. Brunner, and E. Yamashita. Influence of Land use, Population, Employment,
522 and Economic Activity on Accident's. *Safety Data, Analysis, and Evaluation*, No. 1953, 2006,
523 pp. 56-64.

524 15. Shankar, V., F. Mannering, and W. Barfield. Effect of Roadway Geometrics and
525 Environmental Factors on Rural Freeway Accident Frequencies. *Accident Analysis &*
526 *Prevention*, Vol. 27, No. 3, 1995, pp. 371-389.

527 16. Gehlke, C. E., and K. Biehl. Certain Effects of Grouping upon the Size of the Correlation
528 Coefficient in Census Tract Material. *Journal of the American Statistical Association*, Vol. 29,
529 No. 185, 1934, pp. 169-170.

530 17. Fotheringham, A. S., and D. W. S. Wong. The Modifiable Areal Unit Problem in
531 Multivariate Statistical Analysis. *Environment and Planning A*, Vol. 23, No. 7, 1991, pp.
532 1025-1044.

533 18. Fotheringham, A. S. GeoComputation Analysis and Modern Spatial Data. , 2000, pp. 33-
534 48.

535 19. Thomas, I. Spatial Data Aggregation: Exploratory Analysis of Road Accidents. *Accident*
536 *Analysis & Prevention*, Vol. 28, No. 2, 1996, pp. 251-264.

537 20. Miller, H. J. Potential Contributions of Spatial Analysis to Geographic Information
538 Systems for Transportation (GIS-T). *Geographical Analysis*, Vol. 31, No. 4, 1999, pp. 373-
539 399.

540 21. Horner, M. W., and A. T. Murray. Excess Commuting and the Modifiable Areal Unit
541 Problem. *Urban Studies*, Vol. 39, No. 1, 2002, pp. 131-139.

542 22. Park, Y., and F. F. Saccomanno. Evaluating Speed Consistency between Successive
543 Elements of a Two-Lane Rural Highway. *Transportation Research Part A: Policy and*
544 *Practice*, Vol. 40, No. 5, 2006, pp. 375-385.

545 23. Arbía, G. *Spatial Data Configuration in Statistical Analysis of Regional Economic and*
546 *Related Problems.* Dordrech [etc.] Kluwer Academic 1989, 1989.

547 24. Zhang, M., and N. Kukadia. Metrics of Urban Form and the Modifiable Areal Unit
548 Problem. *Transportation Research Record: Journal of the Transportation Research Board*,
549 Vol. 1902, No. -1, 2005, pp. 71-79.

550 25. Anselin, L. *Spatial Econometrics: Methods and Models.* Kluwer Academic Publishers,
551 Dordrecht, 1988.

552 26. Yannis, G., E. Papadimitriou, and C. Antoniou. Multilevel Modelling for the Regional
553 Effect of Enforcement on Road Accidents. *Accident Analysis & Prevention*, Vol. 39, No. 4,
554 2007, pp. 818-825.

555 27. Robinson, W. S. Ecological Correlations and the Behavior of Individuals. *American*
556 *Sociological Review*, Vol. 15, No. 3, 1950, pp. 351-357.

557 28. Davis, G. A. Is the Claim that 'variance Kills' an Ecological Fallacy? *Accident Analysis*
558 *& Prevention*, Vol. 34, No. 3, 2002, pp. 343-346.

559 29. ———. Possible Aggregation Biases in Road Safety Research and a Mechanism
560 Approach to Accident Modeling. *Accident Analysis & Prevention*, Vol. 36, No. 6, 2004, pp.
561 1119-1127.

562 30. Pande, A., and M. A. Abdel-Aty. Patterns in Severe Crashes on Segments of Multilane
563 Arterials with Partially Limited Access. In *Proceedings of the Paper Presented at the 88th*
564 *Annual Meeting of the Transportation Research Board*, 2009.

565 31. Jovanis, P. P., J. A. Valverde, K. Wu, and V. Shankar. Naturalistic Driving Event Data
566 Analysis: Omitted Variable Bias and Multilevel Modeling Approaches. In *Proceedings of the*
567 *Paper Presented at the 90th Annual Meeting of the Transportation Research Board*, 2011.

568 32. Diez-Roux, A. V. Bring Context Back into Epidemiology: Variables and Fallacies in
569 Multilevel Analysis. *American Journal of Public Health*, Vol. 88, No. 2, 1998, pp. 216-222.

570 33. Barker, J., S. Farmer, and M. Taylor. *The Development of Accident-Remedial Intervention*
571 *Levels for Rural Roads.* Transport Research Laboratory, Crowthorne, Berkshire, 1999.

572 34. Noland, R. B. Traffic Fatalities and Injuries: The Effect of Changes in Infrastructure and
573 Other Trends. *Accident Analysis & Prevention,* Vol. 35, No. 4, 2003, pp. 599-611.

574 35. Noland, R. B., and M. A. Quddus. Improvements in Medical Care and Technology and
575 Reductions in Traffic-Related Fatalities in Great Britain. *Accident Analysis & Prevention*,
576 Vol. 36, No. 1, 2004, pp. 103-113.

577 36. LeSage, J. P. *Spatial Econometrics.* , Department of Economics, University of Toledo,
578 1999.

579 37. Tobler, W. R. A Computer Movie Simulating Urban Growth in the Detroit Region.
580 *Economic Geography*, Vol. 46, 1970, pp. 234-240.

581 38. Clayton, D. G., L. Bernardinelli, and C. Montomoli. Spatial Correlation in Ecological
582 Analysis. *International Journal of Epidemiology*, Vol. 22, No. 6, 1993, pp. 1193-1202.

583 39. Xia, H., B. P. Carlin, and L. A. Waller. Hierarchical Models for Mapping Ohio Lung
584 Cancer Rates. *Environmetrics*, Vol. 8, No. 2, 1997, pp. 107-120.

585 40. Song, J. J. Bayesian Multivariate Spatial Models and their Applications. , 2004.

586 41. Song, J. J., M. Ghosh, S. Miaou, and B. Mallick. Bayesian Multivariate Spatial Models
587 for Roadway Traffic Crash Mapping. *Journal of Multivariate Analysis,* Vol. 97, No. 1, 2006,
588 pp. 246-273.

589 42. MacNab, Y. C. Bayesian Spatial and Ecological Models for Small-Area Accident and
590 Injury Analysis. *Accident Analysis & Prevention*, Vol. 36, No. 6, 2004, pp. 1019-1028.

591 43. Li, L., L. Zhu, and D. Z. Sui. A GIS-Based Bayesian Approach for Analyzing spatial–
592 temporal Patterns of Intra-City Motor Vehicle Crashes. *Journal of Transport Geography,* Vol.
593 15, No. 4, 2007, pp. 274-285.

594 44. Quddus, M. A. Modelling Area-Wide Count Outcomes with Spatial Correlation and
595 Heterogeneity: An Analysis of London Crash Data. *Accident Analysis & Prevention*, Vol. 40,
596 No. 4, 2008, pp. 1486-1497.

597 45. Aguero-Valverde, J., and P. Jovanis. Analysis of Road Crash Frequency with Spatial
598 Models. *Transportation Research Record: Journal of the Transportation Research Board*,
599 Vol. 2061, No. -1, 2008, pp. 55-63.

600 46. El-Basyouny, K., and T. Sayed. Urban Arterial Accident Prediction Models with Spatial
601 Effects. *Transportation Research Record: Journal of the Transportation Research Board*,
602 Vol. 2102, No. -1, 2009, pp. 27-33.

603 47. Guo, F., X. Wang, and M. A. Abdel-Aty. Modeling Signalized Intersection Safety with
604 Corridor-Level Spatial Correlations. *Accident Analysis & Prevention*, Vol. 42, No. 1, 2010,
605 pp. 84-92.

606 48. Wang, C., M. Quddus, and S. Ison. A Spatio-Temporal Analysis of the Impact of
607 Congestion on Traffic Safety on Major Roads in the UK. *Transportmetrica*, 2011 SP.

608 49. Griffith, D. A Comparison of Six Analytical Disease Mapping Techniques as Applied to
609 West Nile Virus in the Coterminous United States. *International Journal of Health*
610 *Geographics*, Vol. 4, No. 1, 2005, pp. 18.

611 50. Johnson, G. Small Area Mapping of Prostate Cancer Incidence in New York State (USA)
612 using Fully Bayesian Hierarchical Modelling. *International Journal of Health Geographics*,
613 Vol. 3, No. 1, 2004, pp. 29.

614 51. Kissling, W. D., and G. Carl. Spatial Autocorrelation and the Selection of Simultaneous
615 Autoregressive Models. *Global Ecology and Biogeography*, Vol. 17, No. 1, 2008, pp. 59-71.

616 52. Adjemian, M. K., C. -. Cynthia Lin, and J. Williams. Estimating Spatial Interdependence
617 in Automobile Type Choice with Survey Data. *Transportation Research Part A: Policy and*
618 *Practice*, Vol. 44, No. 9, 2010, pp. 661-675.

619 53. Banerjee, S., B. P. Carlin, and A. E. Gelfand. *Hierarchical Modeling and Analysis for*
620 *Spatial Data*. Chapman & Hall/CRC, London, 2004.

621 54. Aguero-Valverde, J., and P. P. Jovanis. Spatial Correlation in Multilevel Crash Frequency
622 Models: Effects of Different Neighboring Structures. In *Proceedings of the Paper Presented*
623 *at the 89th Annual Meeting of the Transportation Research Board*, 2010.

624 55. Huang, H., and M. Abdel-Aty. Multilevel Data and Bayesian Analysis in Traffic Safety.
625 *Accident Analysis & Prevention*, Vol. 42, No. 6, 2010, pp. 1556-1565.

626 56. Wang, C., M. A. Quddus, and S. G. Ison. Impact of Traffic Congestion on Road
627 Accidents: A Spatial Analysis of the M25 Motorway in England. *Accident Analysis &*
628 *Prevention*, Vol. 41, No. 4, 2009, pp. 798-808.

629 57. Tarko, A., J. Thomaz, and D. Grant. Probabilistic Determination of Crash Locations in a
630 Road Network with Imperfect Data. *Transportation Research Record: Journal of the*
631 *Transportation Research Board*, Vol. 2102, 2009, pp. 76-84.

632