Towards a descriptive model of road accident risk

SA van Heerden[a,c,*], JH van Vuuren[a,c], SS Grobbelaar[b,c]

[a]*Stellenbosch Unit for Operations Research in Engineering*
[b]*Health Systems and Innovation Group*
[c]*Department of Industrial Engineering, Stellenbosch University, South Africa*

Abstract

With the enormous losses to society resulting from *road accidents* (RAs), their prevention and severity reduction has been an active area of research focus for many decades. Over the years, researchers have proposed a wide variety of principles and models related to road safety, many of which attempt to draw insights from historical RA and exposure data. In this paper, a novel descriptive model is proposed for quantifying the risk posed by various roadway characteristics along a road network which takes inspiration from the *three dimensions of road safety* and employs network *Kernel Density Estimation* (KDE). Risk, in this context, is modelled as both the rate and severity of being involved in an RA. The proposed methodology is applied to a real-world case study based on public accident and road data so as to demonstrate its intended functionality in the context of a realistic setting.
Keywords: Road accident rate, Road accident severity, Three dimensions of road safety, Network kernel density estimation

1. Introduction

The losses to society resulting from *road accidents* (RAs) have resulted in a need to prevent and reduce the severity of these occurrences. This has induced an active area of research focus over many decades. Over the years, road safety researchers have proposed a wide variety of descriptive models for use in this context [6], many of which attempt to draw

*Corresponding author
Email addresses: 17683068@sun.ac.za (SA van Heerden), vuuren@sun.ac.za (JH van Vuuren), ssgrobbelaar@sun.ac.za (SS Grobbelaar)

insights from historical RA and exposure data — such insights provide valuable direction for governments in respect of safer road designs and countermeasures aimed at reducing the negative effects of RAs. Nevertheless, RAs still account for approximately 1.25 million deaths world-wide and an additional 50 million injuries annually [27]. According to the World Health Organisation, RAs are ranked as the 8$^{\text{th}}$ leading cause of death [21] and the leading cause of death among people aged 15–29 [27]. In the United States of America alone, the total annual cost of RAs amounts to a staggering \$836 billion [28]. The development of more informative descriptive models is, therefore, of significant importance.

In the wake of the 4$^{\text{th}}$ industrial revolution, there exists significant opportunity in the road transportation domain for the development of disruptive technologies. A solution which is able to answer the question "*To how much and what type of risk are individuals exposed on the roads?*" is one which would likely appeal to a wide variety of stakeholders. Governments may, for example, benefit from a system which is able to identify those points within its road networks that exhibit high accident rates and severities. The identification of these so-call accident "hotspots" may provide vital direction with respect to safer road designs and countermeasures aimed at reducing the number of RAs. Such a system is, however, not anticipated only to be limited to government decision support. Many vehicle insurance companies are already installing tracking devices in their customers' vehicles and quoting premiums based on how well their customers drive [15]. Consequently, vehicle insurance companies may also benefit from a system that is able to quantify and characterise the risk experienced by each of their customers on the roads so that this can be translated into monthly premiums. The application of such a system may also stretch far beyond insurance. The public at large typically relies on routing applications (*e.g.* Google Maps [16]) to prescribe the fastest or shortest route from a specified starting point to an intended destination. Hence, society may also benefit from a system that is able to provide navigation along routes which minimise the risk of being involved in an RA. Moreover, the endorsement of such a system by insurance companies may lead to incentive schemes which persuade customers to drive along safer routes.

The purpose and novel contribution of this paper is the proposition of a descriptive model for quantifying and creating a unified view of the RA risk posed by various roadway characteristics along a road network, where risk, in this context is modelled as both the rate and severity of being involved in a RA. This is achieved by taking inspiration from the *three dimensions of road safety* proposed by Rumar [22] and employing the network *kernel density estimation* (KDE) method of Okabe *et al.* [9]. Unlike traditional road safety descriptive approaches, our approach also draws upon road network geometry in order to produce both spatial and non-spatial perspectives of a given road network. Furthermore, the benefit of quantifying the rate and severity along each road segment is that it could serve as a new distance metric to prescribe safer routes to road users.

Apart from this introductory section, the paper contains a further four sections. Section 2 is devoted to a brief review of the literature related to this study and includes concise descriptions of road safety data requirements, the three dimensions of road safety and the network KDE method. An overview of the descriptive model for quantifying accident rate and severity is provided thereafter in §3. The focus then shifts in §4 to an illustrative case study which begins with a description of the data employed and an interpretation and discussion of the results obtained by means of the methodology proposed in §3. The last section, §5, contains possible avenues for future work related to this study.

2. Literature review

The purpose in this section is to provide a concise review of the literature pertaining to this study. Section 2.1 is devoted to a description of the data commonly employed during a road safety investigation. This is followed, in §2.2, by an elucidation of the three dimensions of road safety which is a popular approach adopted to describe a country's road safety situation from a wholistic perspective. Thereafter, the focus shifts in §2.3 to a review of the network KDE method, which will prove useful during the descriptive model formulation of §3.

2.1. Road safety data

In order to construct descriptive road safety models, a researcher must first acquire an appropriate set of data related to road safety. There are typically two sources of data in road safety analyses, namely *accident data* and *exposure data* [6]. The primary source of accident data is captured in official RA reports collected by police officers when responding to the scene of an accident, while secondary sources may be in the form of insurance company claims, hospital records and accident involvement surveys [26].

Although RA data are the basic information required in any descriptive analysis, significant findings are often realised only when accident data are combined with exposure data [6]. There are many different ways to define RA exposure. One may, for example, define exposure according to the distance travelled, time spent in traffic or number of trips taken by a road user. Exposure data are, typically, not commonly collected exclusively for road safety purposes, but are rather captured for road transport and economic development planning [6]. Developed countries typically employ two methods at a national level to estimate various measures of exposure to their road transport network. These may take the form of *travel habit/vehicle usage surveys* and/or employing specialised *traffic counting systems*. The former requires surveying a sample of vehicle owners about their driving patterns. The information thus obtained is then used to estimate exposure measures in person-kilometres, vehicle-kilometres, travel time and/or number of trips. The latter, on the other hand, is typically employed to estimate the *annual average daily traffic* (AADT) on individual road sections which can, in turn, be used to estimate a vehicle-kilometres measure of exposure [6].

2.2. The three dimensions of road safety

Arguably, the most popular descriptive modelling approach for describing the road safety situation of a country is according to the *three dimensions of road safety*. Rumar [22] claims that the road traffic situation can be described according to three principal dimensions which attempt to quantify: (1) The *exposure* of a road user to the road transport system,

(2) the *risk*[1] of a road user being involved in an RA, and (3) the *consequence* of a road user being involved in an RA. Given these three dimensions, road safety is often described mathematically as

$$\text{Road safety situation} = \text{Exposure} \times \text{Risk} \times \text{Consequence}, \tag{1}$$

where the risk of an RA is defined as

$$\text{Risk} = \frac{\text{Number of accidents}}{\text{Exposure}} \tag{2}$$

and the consequence of an RA is defined as

$$\text{Consequence} = \frac{\text{Number of fatalities}}{\text{Number of accidents}} \tag{3}$$

[6]. Consequently, (1) can collectively be illustrated graphically by means of a rectangular prism, as shown in Figure 1. This representation is often preferred as it simplifies the understanding of the traffic safety situation [8], where a comparatively larger rectangular prism volume indicates a more severe road safety situation.

2.3. Network kernel density estimation

Let $\mathcal{L} = \{\ell^{(1)}, \ldots, \ell^{(L)}\}$ be a set of L straight line segments, where segment $\ell^{(i)}$, having endpoints $(a_1^{(i)}, b_1^{(i)})$ and $(a_2^{(i)}, b_2^{(i)})$, is the set of points satisfying $\{(a_1^{(i)}, b_1^{(i)}) + u(a_2^{(i)}, b_2^{(i)}) \mid u \in [0, 1]\}$. The endpoints of these segments are called *vertices* and are collected in a set $\mathcal{V}$. A *network* space, embedded in the plane, is induced by the union $\tilde{\mathcal{L}} = \bigcup_{i=1}^{L} \ell^{(i)}$ of all the segments. Given a set of M points $\mathcal{P} = \{p_1, \ldots, p_M\}$, where $p_k = (m^{(k)}, n^{(k)})$, one may wish to estimate the density of these points along the network $\tilde{\mathcal{L}}$. In the GIS literature, this is traditionally achieved by means of *network histograms*, each of which captures the point density along a line segment as the number of points on the line segment divided by the line segment length. Although this approach is simple and easy to implement in practice, it often produces bias [10]. To overcome this shortcoming, Okabe *et al.* [9] proposed the

[1]Although the term *risk* is widely used in the literature when describing the three dimensions of road safety, in this paper, the term *risk* is used exclusively to represent the overall negative effects of RAs.

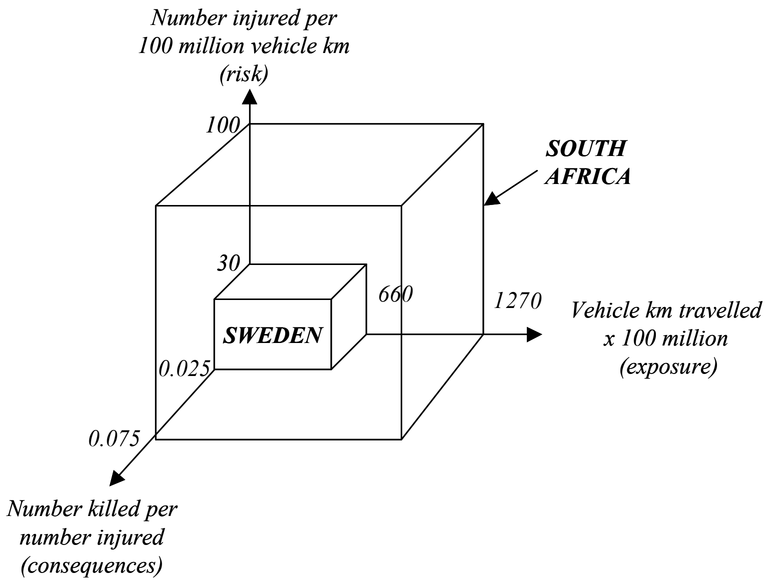


Figure 1: Comparison of the road safety problems in Sweden and South Africa according to the three dimensions of road safety [1].

network KDE method for unbiased density estimation in a network space by refining the well-known planar KDE method originally proposed by Fix and Hodges [3].

Consider a point p in a non-directed network space $\tilde{\mathcal{L}}$, as illustrated in Figure 2. A sub-network $\mathcal{L}_p \subset \tilde{\mathcal{L}}$ is constructed such that the shortest-path distance between p and any other point in $\mathcal{L}_p$ is at most w (referred to as the *buffer network* of the point p with a width

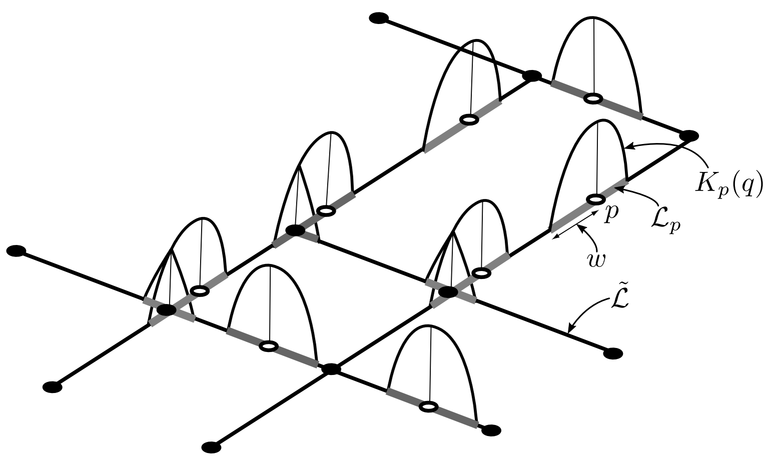


Figure 2: Multiple kernel density functions on a network embedded in the plane, where the open points indicate the kernel centres and the grey bold line segments denote the kernel supports [10].

of w). For an arbitrary point $q \in \tilde{\mathcal{L}}$, a function $K_p(q)$ is defined for p satisfying

$$K_p(q) \begin{cases} \geq 0 & \text{for } q \in \mathcal{L}_p, \\ = 0 & \text{for } q \in \tilde{\mathcal{L}} \backslash \mathcal{L}_p, \end{cases} \tag{4}$$

and

$$\int_{q \in \mathcal{L}_p} K_p(q)\, \mathrm{d}q = 1, \tag{5}$$

where $\tilde{\mathcal{L}} \backslash \mathcal{L}_p$ in (4) is the complement of the sub-network $\mathcal{L}_p$ with respect to the network $\tilde{\mathcal{L}}$, and the integral over $q \in \mathcal{L}_p$ in (5) is the one-dimensional Riemann limit sum of the function $K_p(q)$ with respect to q along $\mathcal{L}_p$. The function $K_p(q)$ is called the *network kernel density function* for which the point p defines the *kernel centre*. The sub-network $\mathcal{L}_p$ is referred to as the *kernel support*, and the value w is called the *bandwidth* of $K_p(q)$.

For the set $\mathcal{P}$ in the network space $\tilde{\mathcal{L}}$, a function $K(q)$, called the *network kernel density estimator*, is defined as the average density at a point q, expressed mathematically as

$$K(q) = \frac{1}{M} \sum_{k=1}^{M} K_{p_k}(q). \tag{6}$$

The expression in (6) therefore approximates some *true* point distribution function over the network $\tilde{\mathcal{L}}$, based on a finite set of points $\mathcal{P}$.

For the purpose of constructing an unbiased kernel density estimator, let $d(p,q)$ be the distance along the network from the point p to an arbitrary point q. Furthermore, let $k(d(p,q))$ be a function characterised by the following four properties (for all $p, q \in \tilde{\mathcal{L}}$):

1. $\int_{q \in \tilde{\mathcal{L}}} k(d(p,q))\, \mathrm{d}q = 1$,

2. there exists a real number $w > 0$ such that $k(d(p,q)) > 0$ if $d(p,q) < w$, and $k(d(p,q)) = 0$ if $d(p,q) \geq w$,

3. $k(d(p,q))$ is strictly non-increasing with respect to $d(p,q)$, and

4. $k(d(p,q))$ is a continuous function of $d(p,q)$.

Such a unimodal function is referred to as a *base kernel density function* [10]. A wide variety of choices are available for this function [5]. There is consensus in the literature that the choice of bandwidth w of the kernel function, however, has a more significant influence on the KDE results obtained than the choice of base kernel function itself [2, 11, 12, 23, 24]. Okabe *et al.* [9], warned that direct use of a base kernel function in a network space leads to biases at network vertices. They therefore proposed two kernel density functions for constructing an *unbiased* kernel density estimator when a kernel support extends beyond a network vertex, called the *equal-split discontinuous kernel density function* and the *equal-split continuous kernel density function.* Only the latter, continuous case is considered here. Although computationally more expensive, it is often preferred over the discontinuous case [10].

Consider the case where a vertex v is within a distance w from a kernel centre positioned at a point p (as illustrated in Figure 3), and let the quantity $n(v)$ denote the degree of vertex v. Then the equal-split kernel density function is expressed in a piecewise manner as a function of the base kernel density function $k(d(p,q))$, which takes the form

$$K_p(q) = \begin{cases} k(d(p,q)) & \text{for } 0 \leq d(p,q) < 2d(p,v) - w, \\ k(d(p,q)) - \frac{n(e_i)-2}{n(v)} k(2d(p,v) - d(p,q)) & \text{for } 2d(p,v) - w \leq d(p,q) < d(p,v), \\ \frac{2}{n(v)} k(d(p,q)) & \text{for } d(p,v) \leq d(p,q) < w. \end{cases}$$

3. Modelling road accident risk along a road segment

The notion of risk is, for the most part, a highly subjective concept and depends largely on the specific context under consideration. Taking inspiration from the three dimensions of road safety, we argue that risk may be expressed, in the context of RAs, as a function of two dimensions, namely: (1) The rate and (2) the severity of being involved in an RA. This formulation is arguably justified since it aligns with the two central burdens associated with road transport, namely: (1) The damage/loss of property and (2) the threat to or loss of human life, and is also in agreement with the expressions in (2) and (3) of §2.2, respectively.

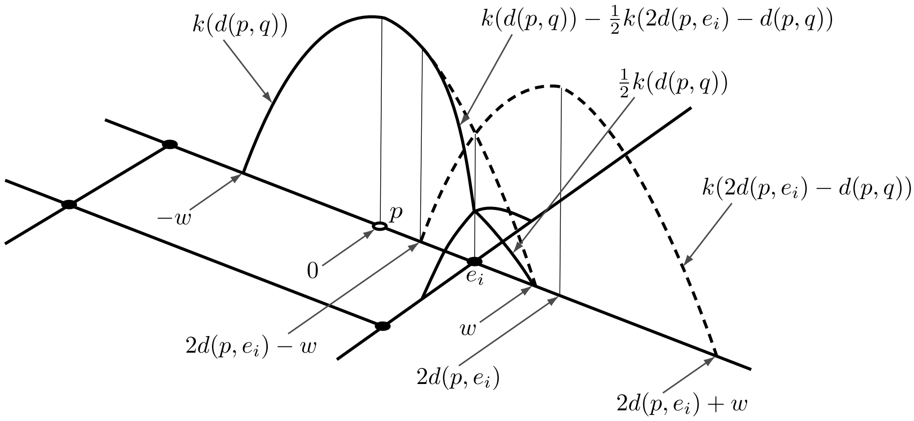


Figure 3: An equal-split continuous kernel density function centred at point p, where $k(d(p,q))$ is a base kernel density function [10].

With this notion in mind, consider a set of M historical RAs $\mathcal{P} = \{p_1, \ldots, p_M\}$, where an accident p_k is spatially represented by the geographical point $p_k = (m^{(k)}, n^{(k)})$. Most accident data sets contain "accident severity" features which capture a count of the number of victims sustaining defined injury severities. It is therefore assumed that each accident p_k possesses a vector $\boldsymbol{s}_k$ capturing the individual counts of the severities sustained by all victims involved. Furthermore, it is assumed that a further N feature values, captured in a vector $\boldsymbol{x}^{(k)} = \langle x_1^{(k)} \cdots x_N^{(k)} \rangle$, collectively describe informative characteristics of the roadway. Lastly, consider a set of L roads $\mathcal{L} = \{\ell^{(1)}, \ldots, \ell^{(L)}\}$, where a road is spatially represented by the geographical straight line segment $\ell^{(i)} = \{(a_1^{(i)}, b_1^{(i)}) + u(a_2^{(i)}, b_2^{(i)}) \mid u \in [0,1]\}$. It is also assumed that a road $\ell^{(i)}$ has a feature value t_i which indicates the number of vehicles that have passed along the road during a specified time period.

The first step in our methodology for modelling RA risk is to partition each road in $\mathcal{L}$ into a set of subintervals of approximately equal length. If a maximum subinterval length H is defined, then the road $\ell^{(i)}$ is partitioned into

$$g_i = \begin{cases} \frac{|\ell^{(i)}|}{H}, & \text{if } |\ell^{(i)}| \text{ is divisible by } H, \\ \left\lfloor \frac{|\ell^{(i)}|}{H} \right\rfloor + 1, & \text{otherwise} \end{cases}$$

subintervals, each of length

$$h_i = \frac{|\ell^{(i)}|}{g_i},$$

where $|\ell^{(i)}|$ denotes the length of line segment $\ell^{(i)}$. This ensures that the combined length of each g_i subintervals never exceeds the maximum subinterval length H for all roads in $\mathcal{L}$ (*i.e.* $h_i < H$ for all $i \in \{1, \ldots, L\}$). A *road observation* $\ell_j^{(i)}$ is defined as the set of points satisfying

$$\left\{ \left(a_1^{(i)}, b_1^{(i)}\right) + u \left(a_2^{(i)}, b_2^{(i)}\right) \mid u \in \left[\frac{j-1}{g_i}, \frac{j}{g_i}\right] \right\},$$

where all $R = \sum_{i=1}^{L} g_i$ road observations containing RAs are contained in the set $\mathcal{R} = \{\ell_j^{(i)} \mid j \in \{1, \ldots, g_i\}, i \in \{1, \ldots, L\}\}$. Notice that if $H \to 0$, then $|\mathcal{R}| \to \infty$ and $|\ell_j^{(i)}| \to H$ for all $j \in \{1, \ldots, g_i\}$ and $i \in \{1, \ldots, L\}$. The quantity H must, therefore, be chosen such that the variance in road observation lengths is sufficiently small (to ensure that the set of road observations is unbiased) and so that $\mathcal{R}$ contains an adequate number of road observations.

It follows that the frequency $F(\ell_j^{(i)})$ of RAs that have occurred along road observation $\ell_j^{(i)}$ is

$$F(\ell_j^{(i)}) = M \int_{\ell_j^{(i)}} K(q) \, \mathrm{d}q,$$

where $K(q)$ is the network kernel density estimator, as defined in (6). Taking inspiration from (2), the accident rate $R(\ell_j^{(i)})$ along road observation $\ell_j^{(i)}$ is then simply the normalised RA frequency

$$R(\ell_j^{(i)}) = \frac{F(\ell_j^{(i)})}{t_i}.$$

The next step is to define a continuous severity measure $S(\ell_j^{(i)})$ for each road observation in $\mathcal{L}$ as a combined severity effect of all the accidents occurring along a road observation $\ell_j^{(i)}$. Suppose the number of victims associated with c distinct severity types for accident k are captured in the vector $\boldsymbol{s}_k$ in descending order of severity. While the accident types assume an intrinsic order, it can of course not be said that the *true* differences between pairs of these types are the same (*e.g.* the difference between "slight" and "serious" injury is not necessarily the same as the difference between "serious" and "fatal" injury). Consequently, it is necessary that the values in $\boldsymbol{s}_k$ are redefined to reflect some cost (possibly monetary) associated with each severity type, captured by the vector $\boldsymbol{m}$.

Taking inspiration from (3), the accident severity $S(\ell_j^{(i)})$ along road observation $\ell_j^{(i)}$ is defined as

$$S(\ell_j^{(i)}) = \frac{\int_{\ell_j^{(i)}} \sum_{k=1}^{M} \boldsymbol{s}_k^T \boldsymbol{m} K_{p_k}(q)\,\mathrm{d}q}{F(\ell_j^{(i)})},$$

where each base kernel function $K_{p_k}(q)$ is weighted by the scaled severity measure $\boldsymbol{s}_k^T \boldsymbol{m}$ and the individual effects are summed and normalised for all $k \in \{1, \ldots, M\}$ along $\ell_j^{(i)}$.

The final step in our RA risk modelling approach is concerned with inferring a set of features, for each road observation in $\mathcal{R}$, which describe inherent characteristics of the roadway from the accidents that have occurred on a road observation $\ell_j^{(i)}$. Given that each RA has a set of N associated categorical features, where the n^{th} feature has a set of associated values collected in $\mathcal{C}^{(n)}$, the proportion of RAs along road observation $\ell_j^{(i)}$ describing the road according to a specific feature value is given by

$$x_n^{(c)}(\ell_j^{(i)}) = \frac{\sum_{\substack{p_k \in \mathcal{P} \\ x_n^{(k)} = c}} \int_{\ell_j^{(i)}} K_{p_k}(q)\,\mathrm{d}q}{F(\ell_j^{(i)})} \quad \text{for } c \in \mathcal{C}^{(n)}. \tag{7}$$

It should be noted that this formulation is only valid if the feature vector $\boldsymbol{x}$ describes inherent properties of the roadway (and not the accident itself), as a majority consensus must be established between the accidents occurring along a road observation $\ell_j^{(i)}$.

4. Illustrative case study

The purpose of this section is to demonstrate the effectiveness of the approach towards quantifying RA risk proposed in this paper by applying the methodology of §3 to a real-world case study involving a set of public data obtained from the UK Department of Transport. In §4.1 and §4.2, the data and model parameters are described, respectively. Thereafter, the results obtained *via* the proposed descriptive modelling approach are presented and discussed from a spatial and a non-spatial perspective in §4.3 and §4.4, respectively. Finally, in §4.5, we demonstrate how quantifying the rate and severity along each road segment can serve as a new distance metric to prescribe safer routes to road users.

4.1. The data set

The data used for this study comprises a set of publicly available accident [19] and road [18] data for Greater Manchester in the *United Kingdom* (UK) over the period 2005–2017. A geographical representation of the accident data set is shown in Figure 4(a), and consists of 35 874 accident records. The green, yellow an red dots in the figure indicate RAs in which the maximum sustained injury was "Slight," "Serious" and "Fatal," respectively. Furthermore, it has been estimated that, as of 2017, the average cost in *British Pounds* (GBP) per victim sustaining a "Slight," "Serious" and "Fatal" injury is £16 434, £213 184 and £1 897 129, respectively [25]. A total of seven categorical features which describe inherent characteristics of the roadway where identified, namely: `FirstRoadClass`, `RoadType`, `SpeedLimit`, `JunctionDetail`, `JunctionControl`, `PhysicalPedestrianCrossing` and `UrbanOrRural-Area`. Further insights into the values that each categorical attribute may assume, together with the proportion of accidents with a specific categorical feature value, may be found in Appendix A. The road data, however, only contain AADT measures for type A and A(M) roads as well as motorways. Consequently, only accidents occurring on these road types were considered during this investigation. The road data set is spatially illustrated in Figure 4(b), and comprises 755 road segments, where red and dark green segments indicate roads with high and low vehicle counts, respectively. As may be seen in the figure, motorways (*e.g.* the M60 motorway circling Manchester) experience significantly higher traffic flows, indicated by the red line segments, compared to the other road types.

4.2. Parameter choices

As described in §2, the network KDE method requires that a base kernel function $k(x)$ with associated bandwidth w is specified. We employed the well-known Epanechnikov kernel function $k(x) = \frac{3}{4}(1 - x^2)$ as the base kernel function as it is optimal in a mean-square error sense and has the lowest loss of efficiency [7]. A bandwidth of $w = 50$ metres was chosen since it is the maximum distance used by the UK Department for Transport when considering whether a road crossing may have contributed to an accident [17].

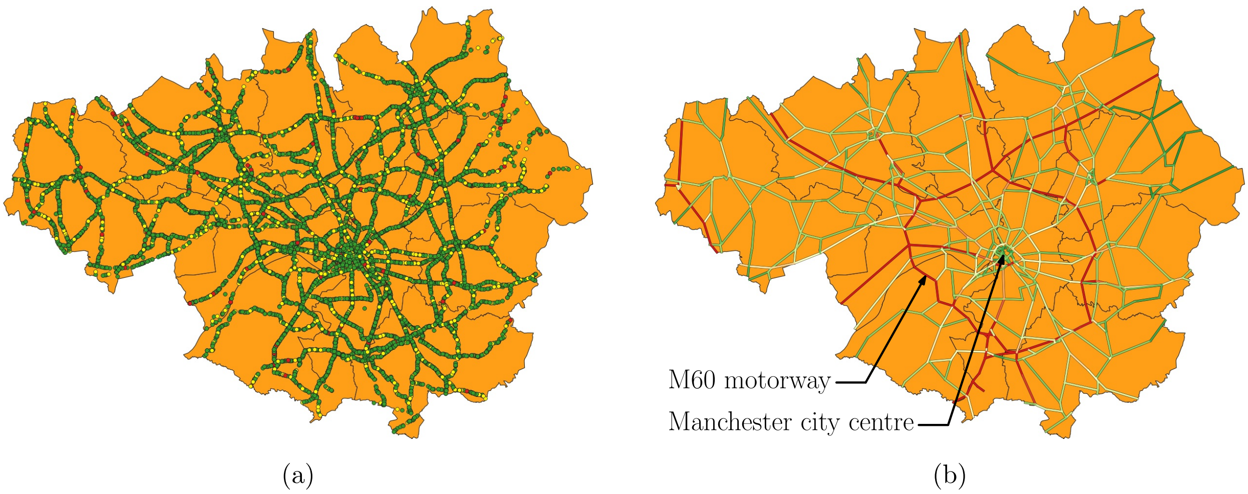


Figure 4: A spatial representation of the 2005–2017 Greater Manchester (a) accidents data set and (b) roads data set.

As described in §3, the quantity H must be chosen such that the variance in road observation lengths is sufficiently small and so that $\mathcal{R}$ contains an adequate number of road observations. The trade-off between these two considerations is shown in Figure 5 as a function of the maximum subinterval length for this case study. As may be seen in the figure, as the quantity H approaches zero, the road observation variance also approaches zero, but the number of road observations grows significantly large which would likely lead to large geoprocessing computation times. Conversely, if H is chosen too large, the variance in road observation lengths becomes significantly large (leading to biases) and too few road observations are defined. In this study, a value of $H = 20$ metres was chosen as an appropriate trade-off between these two considerations, coupled with the fact that a junction is not considered, by the UK Department for Transport, to have contributed to an accident if the accident was not within 20 metres of the junction [17].

4.3. Spatial analysis of road accident risk

The first step in this investigation was to quantify the accident rate and severity measures along the defined road segment observations. Spatial representations of these results for the Greater Manchester road network are provided in Figures 6(a) and (b), respectively, where

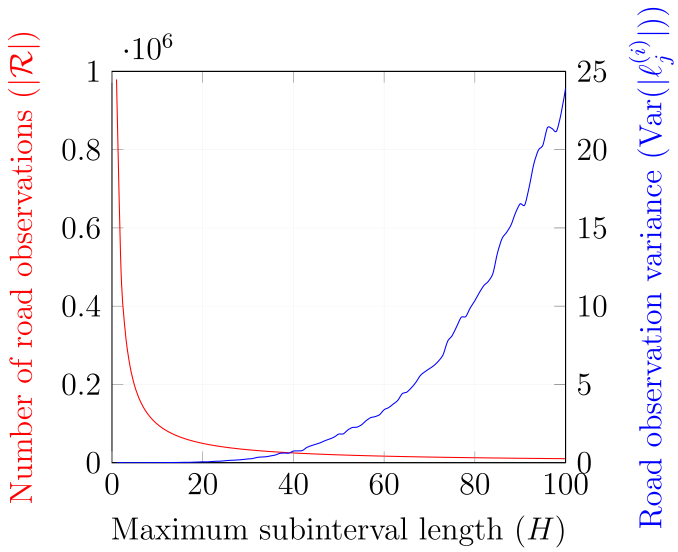
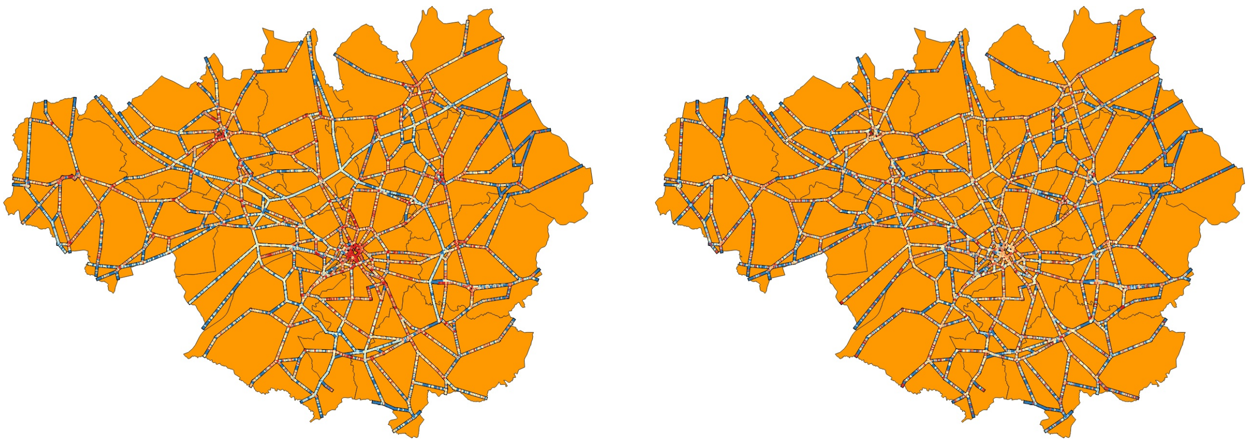


Figure 5: The effects of the maximum subinterval length choice on the number of road observations and the variance in road observation length.

red lines represent roads with higher accident rates/severities and blue lines represent roads with lower accident rates/severities[2]. From these results, one may deduce that significantly higher accident rates appear to occur in Manchester city centre, which may be due to the higher density of roads (and thus more junctions). In contrast, the M60 motorway circling Manchester appears to exhibit somewhat lower rates of road accidents, which is likely due to its high traffic volumes and lack of traffic conflict points. This finding is also consistent with the conclusion of the Road Safety Foundation which has rated the M60 motorway as one of the safest roads in Great Britain [4]. In the case of accident severity, it is not entirely clear whether one can draw any preliminary insights from the spatial representation in Figure 6(b), although there does appear to be a slightly higher accident severity in the Manchester city centre and along some sections of the M60 motorway.

A spatial representation of the roadway feature values for each road segment observation is provided in Figure 7. It is clear from the spatial results in Figure 7(a) that motorways (like the M60) are, as expected, dual carriageways, while A type roads are typically single carriageways. One can also confirm from this figure that slip roads are located at motorway

[2]Black lines in Figure 6 represent road segment observations with an accident frequency of zero.



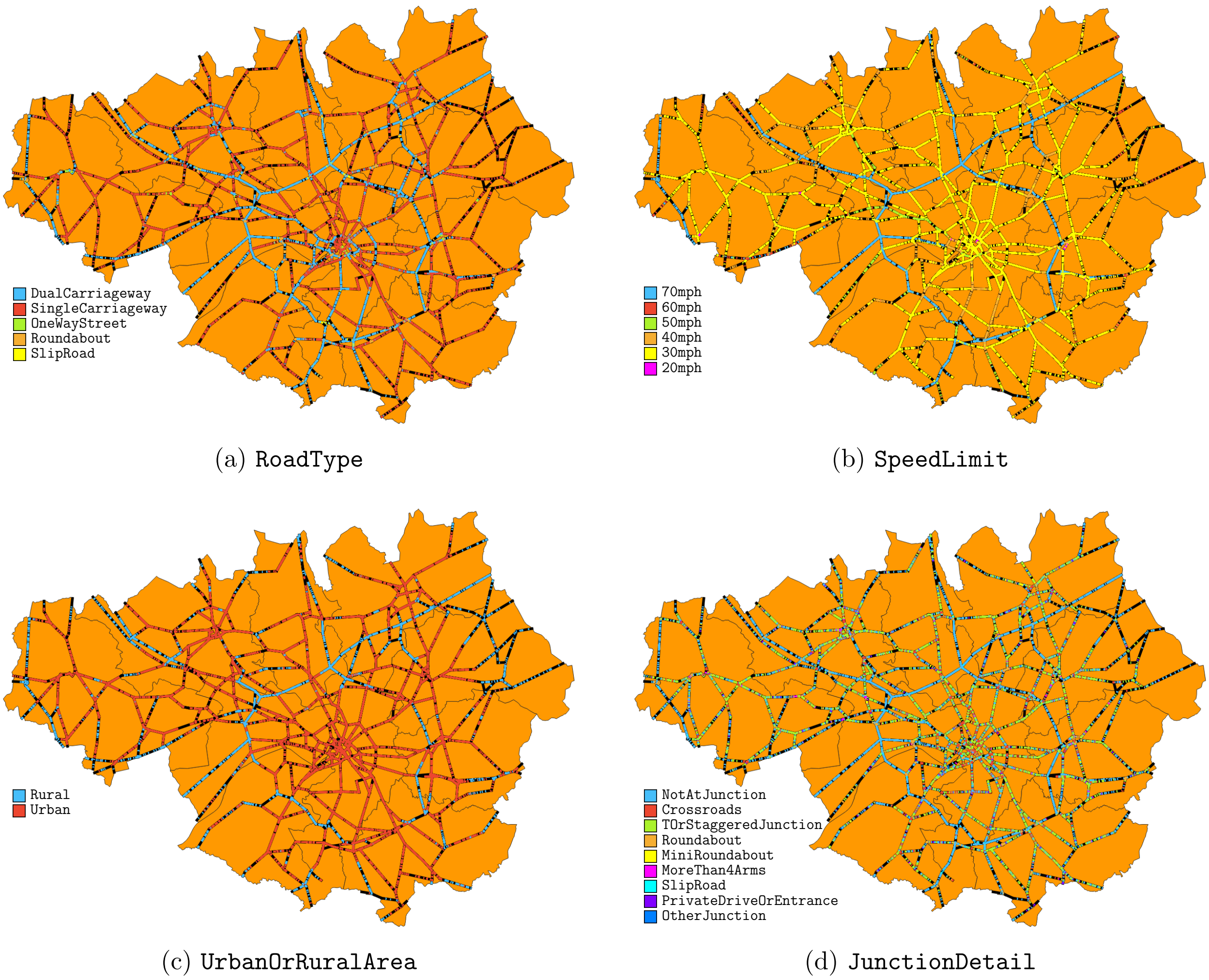
(a) (b)

Figure 6: A geographic representation of the Greater Manchester (a) accident rate and (b) accident severity measurements.

intersections, while the majority of one way streets are found in Manchester city centre. One may deduce from Figure 7(b) that motorways have a set speed limit of 70mph. Figures 7(b) and (c) also suggest that the majority of major roads in Greater Manchester are in urban settings with a set speed limit of 30mph. It may also be deduced that 60mph roads are typically single carriageways located in rural areas, while 20mph roads are found in highly urbanised settings like Manchester city centre. The spatial results of Figure 7(d) also provide some insights into the distribution of junctions in Greater Manchester. As may be seen in the figure, a large proportion of the junctions in Greater Manchester are T- or staggered junctions — the majority of which are located in urban settings (when taking Figure 7(c) into consideration).

4.4. Non-spatial analysis of road accident risk

An alternative, non-spatial representation of the results may be viewed in the form of the scatter plot shown in Figure 8, where average accident rate is plotted against the average accident severity for the `RoadType` feature's categories. In this figure, each point is plotted as a pie chart so as to illustrate what proportion of accident categories fall into the categories of the `JunctionDetail` feature and the size of a pie chart denotes the accident frequency



(a) `RoadType` (b) `SpeedLimit` (c) `UrbanOrRuralArea` (d) `JunctionDetail`

Figure 7: A geographical representation of different road segment feature values in Greater Manchester.

of the `RoadType` feature category. As may be seen in Figure 8, the average rate of RAs appear to be lower on single and dual carriageways (when compared to one-way streets and interchange roundabouts), but experience a greater accident severity, on average. A large proportion of accidents occurring on single carriageways appear to occur at some form of junction, the most prominent being a T- or staggered junctions, which corroborates the fact that the majority of junctions in Greater Manchester are, in fact, T- or staggered junction, as previously seen in Figure 7(d). The highest average accident rate occurs on one-way streets — a road type common to dense urban environments which supports the conclusions drawn from the spatial representations shown in Figures 6(a) and 7(c).

Figure 9 provides more insight into the accident rate and severity associated with each speed limit. The highest average rate of accidents occurs on 20mph roads — a speed limit common to dense urban environments, which supports the conclusions drawn from the spatial representations shown in Figures 6(a) and 7(c). The highest accident severity is observed on 60mph roads which may appear surprising when compared with the accident severity on 70mph roads. One would expect road segments with higher speed limits to result in higher accident severities. As previously deduced from Figures 7(b) and (c), however, motorways typically have a speed limit of 70mph in the UK and are commonly dual carriageways. As previously mentioned, 60mph roads, on the other hand, are typically single carriageways common to rural areas. A possible explanation for this observation may be that motorists are more likely to experience a severe accident on a rural 60mph single carriageway (than on a 70mph dual carriageway) because of the increased chance of head-on collisions on these roads (*i.e.* there is no physical barrier separating traffic flowing in opposite directions on a single carriageway).

Figure 10 provides more insight into the accident rate and severity associated with accidents which occurred at some form of junction. As may be seen in the figure, the highest average accident rate is experienced at crossroads. T- or staggered junctions, in comparison, exhibit lower accident rates but higher accident severities, on average. One possible explanation for this observation is that four-arm junctions have more conflict points between streams of traffic flow than do three-arm junctions, which may lead to higher accident frequencies [29]. Uncontrolled T- or staggered junctions, on the other hand, typically require motorists travelling on minor roads to merge with traffic on a major road. This often requires motorists turning left to make judgements based on traffic flows from both directions, which may result in T-bone collisions impacting on the driver's side of the vehicle. It may also be deduced that accidents at a roundabout are the least severe when compared to all other junction types. This result, therefore, confirms that the primary purpose of roundabouts, which is to reduce traffic speed and minimise T-bone and head-on collisions [20], is indeed being realised in Greater Manchester.

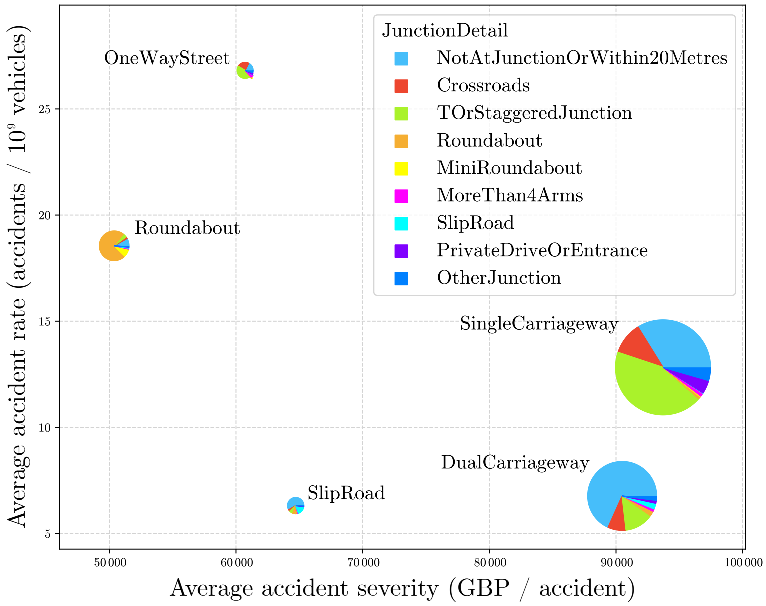


Figure 8: An accident rate against accident severity scatter plot of the `RoadType` attribute categories, where each point is plotted as a pie chart showing the road segment proportions coinciding with the `JunctionDetail` attribute categories, and the size of a pie chart denotes accident frequency for the `RoadType` attribute category.

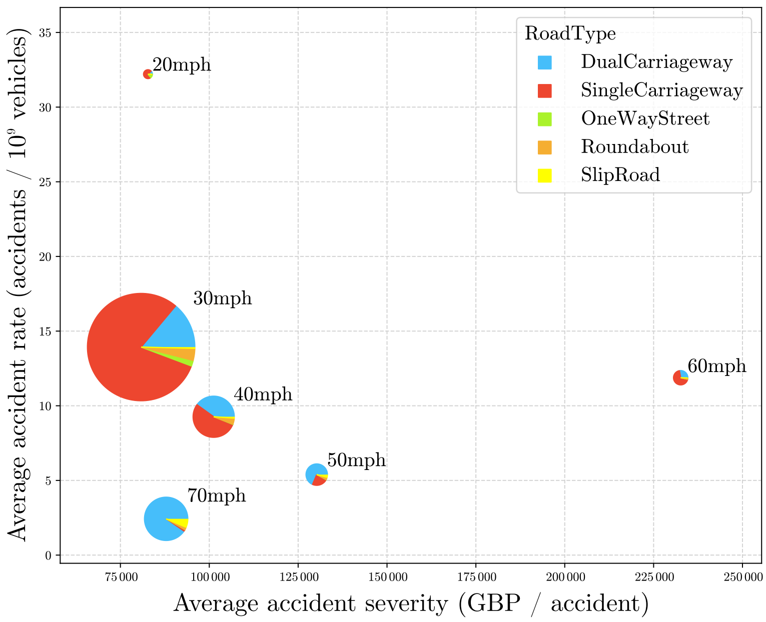


Figure 9: An accident rate against accident severity scatter plot of the `SpeedLimit` attribute categories, where each point is plotted as a pie chart showing the road segment proportions coinciding with the `RoadType` attribute categories, and the size of a pie chart denotes accident frequency for the `SpeedLimit` attribute category.

Further insights into the accident rate and severity along road segments located in urban and rural areas may be seen in Figure 11. From this figure, one may conclude that the rate of being involved in an accident on an urban road is significantly higher than that on a rural road. Conversely, the severity of an accident on rural roads is, on average, more severe than on roads in an urban setting. Presented with this information, a domain knowledge expert may form the following hypothesis: The higher rate of accidents in urban areas is due to the higher density of roads and thus an increased chance of having an accident at a road junction. The majority of accidents on rural roads, on the other hand, do not occur at any form of junction, as may be seen in Figure 11, which may also suggest that accidents on these roads typically occur at higher speeds.

4.5. Prescriptive analysis of road accident risk

As previously described, the benefit of quantifying the rate and severity along each road segment is that it may serve as a new distance metric. We demonstrate this notion by means of an example illustrated in Figure 12. Suppose a road user wishes to travel from point A to point B in the Greater Manchester road network. Drawing from the results illustrated in Figure 6(a), the route which minimises the total RA rate is given by the red line segments in Figure 12. In this case, the optimal route prescribes that a road user should rather travel along the M60 motorway thus avoiding going through Manchester city centre. Although the severity of accidents on motorways are typically greater (due to high travel speeds), there are fewer traffic conflict points and higher traffic flows which ultimately results in lower rates of RAs. If, on the other hand, one were to draw from the results illustrated in Figure 6(b) and solve for the route which minimises the total RA severity experienced, this route would be in the form of the blue line segments shown in Figure 12. In this case, the optimal route prescribes that a road user travels straight through Manchester city centre. Even though the RA rate may be higher in Manchester city centre (predominantly due to the increased density of road junctions), these accidents tend to be less severe due to the lower vehicle speeds. In this way, a road user may select either the red or blue route in Figure 12 if safety is their primary objective.

Figure 10: An accident rate against accident severity scatter plot of the `JunctionDetail` attribute categories, where each point is plotted as a pie chart showing the road segment proportions coinciding with the `JunctionControl` attribute categories, and the size of a pie chart denotes accident frequency for the `JunctionDetail` attribute category.

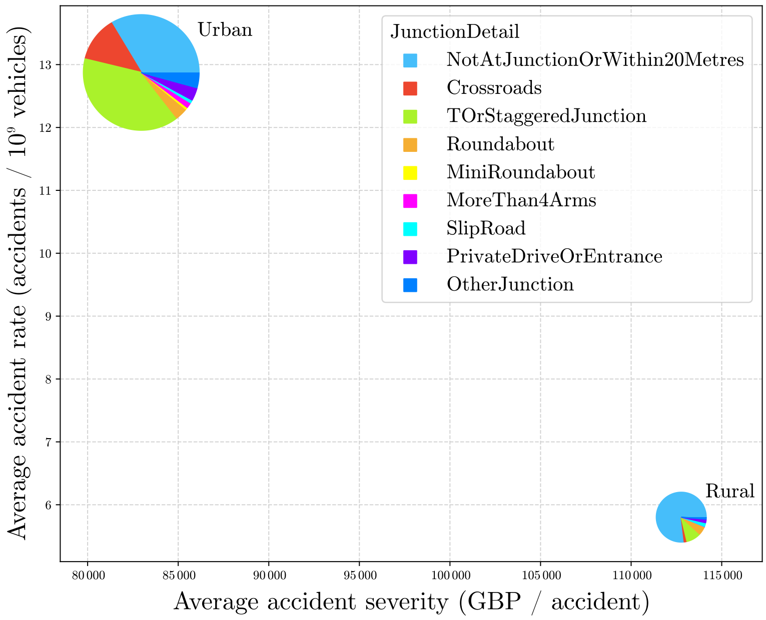


Figure 11: An accident rate against accident severity scatter plot of the `UrbanOrRuralArea` attribute categories, where each point is plotted as a pie chart showing the road segment proportions coinciding with the `JunctionDetail` attribute categories, and the size of a pie chart denotes accident frequency for the `UrbanOrRuralArea` attribute category.

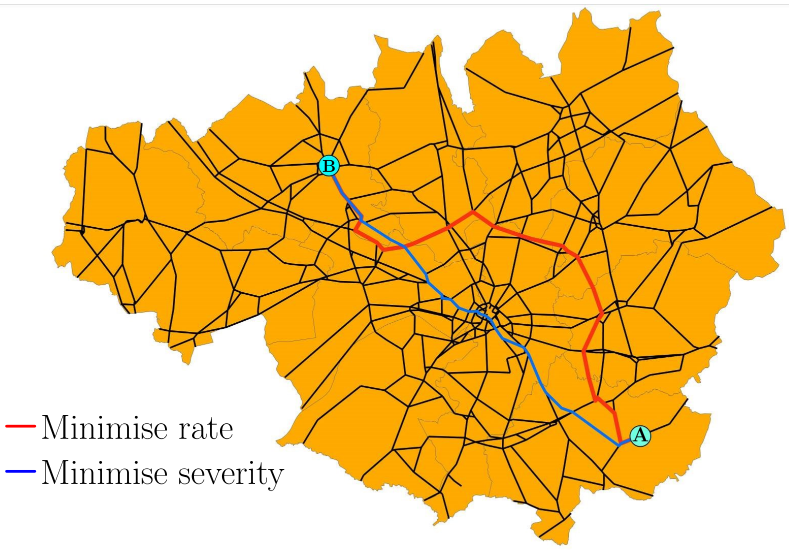


Figure 12: A geographical representation of two routes which may be prescribed to a road user travelling from point A to point B in the Greater Manchester road network. Red and blue line segments indicate the routes which minimises the total RA rate and severity experienced, respectively.

5. Conclusion

The purpose and novel contribution of this paper was the proposition of a descriptive model for quantifying and creating a unified view of the RA rate and severity posed by various roadway characteristics along a road network. The effectiveness of this approach was demonstrated in §4 in the context of a real-world case study using public data obtained from the UK department for Transport. Furthermore, it was demonstrated how one may use the quantitative measures of RA rate and severity along each road segment to prescribe safer routes to road users.

During the case study investigation, it was necessary that an appropriate base kernel function k and associated bandwidth w be specified in order to compute the rate and severity of RAs along a road segment. The choice of base kernel function and associated bandwidth is, however, not entirely obvious or justifiable. It may, therefore, be worth investigating the notions employed in multiple and infinite kernel learning [13, 14] to circumvent the functional form and shape choice of KDE base kernel. In §4.3 and §4.4, geographical and novel scatter plot representations were used to make informed interpretations about the RA risk posed by various roadway characteristics along the Greater Manchester road network, respectively. To truly demonstrate the *generic* nature of this approach, however, more case studies may be conducted in respect of real-world data from different countries. Such an undertaking would also facilitate a country comparison of the RA rate and severity posed by various roadway characteristics. Our approach outlined in §3 does, however, assume a measure of traffic flow for each road segment in the road network under investigation and a cost estimate corresponding to the various injury severity categories — both of which may not be available, especially in developing countries.

References

[1] Arries, C. S., Sewlal, R., and Mamabolo, V. (2011). The importance of exposure data for a comprehensive accident database. In *30th Southern Africa Transport Conference (SATC 2011)*, pages 254–262, Pretoria.

[2] Bailey, T. C. and Gatrell, A. C. (1995). *Interactive spatial data analysis*. Longman Scientific & Technical, Harlow.

[3] Fix, E. and Hodges, J. L. (1951). Discriminatory analysis, nonparametric discrimination: Consistency properties. Technical Report Report No. 4, Project No. 21-49-004, USAF School of Aviation Medicine, Randolf Field (TX).

[4] Foundation, T. R. S. (2014). Risk rating of britain's motorways and a roads (risk bands 2020).

[5] Guidoum, A. C. (2018). Kernel estimator and bandwidth selection for density and its derivatives.

[6] Mikkonen, V. and Peltola, H. (1997). Road safety principles and models: Review of descriptive, predictive, risk and accident consequence models. Technical Report IRRD No. 892483, Organization for Economic Cooperation and Development, Paris.

[7] Miladinovic, B. (2008). Kernel density estimation of reliability with applications to extreme value distribution.

[8] Nilsson, G. (2004). Traffic safety dimensions and the power model to describe the effect of speed on safety.

[9] Okabe, A., Satoh, T., and Sugihara, K. (2009). A kernal density estimation method for networks, its computational method and a gis-based tool. *International Journal of Geographic Information Science*, 23:7–23.

[10] Okabe, A. and Sugihara, K. (2012). *Spatial analysis along networks: Statistical and computational methods*. John Wiley & Sons, Inc., Hoboken (NJ).

[11] O'Sullivan, D. and Unwin, D. J. (2002). *Geographic information analysis*. John Wiley & Sons, Inc., Hoboken (NJ).

[12] O'Sullivan, D. and Wong, D. W. S. (2007). A surface-based approach to measuring spatial segregation. *Geographic Analysis*, 39(2):147–168.

[13] Özöğür-Akyüz, S., Üstünkar, G., and Weber, G. W. (2016). Adapted infinite kernel learning by multi-local algorithm. *International Journal of Pattern Recognition and Artificial Intelligence*, 30(4):1651004–2–1651004–21.

[14] Özöğür-Akyüz, S. and Weber, G. W. (2008). Learning with infinitely many kernels via semi-infinite programming. In *Continuous Optimization and Knowledge Based Technologies, 20^{th} EURO Mini Conference*, pages 342–348, Neringa.

[15] Compare.com (2019). How car tracking programs are changing auto insurance.

[16] Google (2019). Maps — navigate & explore.

[17] UK Department for Transport (2011). Stats20: Instructions for the completion of road accident reports from non-crash sources. Technical report, London.

[18] UK Department for Transport (2017a). GB road traffic counts.

[19] UK Department for Transport (2017b). Road safety data.

[20] Washington State Department of Transportation (2018). Roundabout benefits.

[21] World Health Organisation (2018). Global health estimates 2016: Deaths by causes age sex by country and by region. Technical report, World Health Organisation, Geneva.

[22] Rumar, K. (1999). Transport safety visions, targets and strategies: Beyond 2000. Technical report, European Transport Safety Council, Brussels. 1^{st} European Transport Safety lecture.

[23] Schabenberger, O. and Gotway, C. A. (2005). *Statistical methods for spatial data analysis.* Chapman & Hall, Boca Raton (FL).

[24] Silverman, B. W. (1986). *Density estimation for statistics and data analysis.* Chapman & Hall, London.

[25] Sönnichsen, N. (2019). Average cost of casualties and accidents on the road in great britain in 2017 (in gbp).

[26] Thulin, H. (1987). Trafikolyckor och trafikskadade enligt polis, sjukvård och försäkringsbolag. Technical report, Statens väg- och trafikinstitut, Linköping.

[27] Toroyan, T. (2015). Global status report on road safety 2015. Technical Report ISBN 9789241565066, World Health Organisation, Geneva.

[28] Vella, M. (2018). Why you shouldn‘t be allowed to drive.

[29] Yannis, G., Papadimitriou, E., and Evgenikos, P. (2011). Effectiveness of road safety measures at junctions. In *1^{st} International Conference on Access Management*, Athens.

Appendix A. Greater Manchester accident category proportions

The various category proportions of the accident data set used in the case study of §4 are summarised in Table A.1. The last column of percentage values represents the proportion of accidents with a specific categorical feature value.

Table A.1: The categories and corresponding data proportions of the various accident features.

No.	Feature	Category	%
1	FirstRoadClass	Motorway	11.99%
		AM	0.91%
		A	87.10%
2	RoadType	Roundabout	5.75%
		OneWayStreet	1.66%
		DualCarriageway	31.17%
		SingleCarriageway	59.69%
		SlipRoad	1.73%
3	SpeedLimit	20mph	0.49%
		30mph	72.79%
		40mph	10.60%
		50mph	2.86%
		60mph	1.28%
		70mph	11.97%
4	JunctionDetail	NotAtJunctionOrWithin20Metres	27.33%
		Roundabout	6.69%
		MiniRoundabout	0.70%
		TOrStaggeredJunction	34.83%
		SlipRoad	1.45%
		Crossroads	21.02%

Continued on next page

Table A.1 — *Continued from previous page*

No.	Feature	Categories	%
4	JunctionDetail	MoreThan4Arms	2.15%
		PrivateDriveOrEntrance	2.20%
		OtherJunction	3.64%
5	JunctionControl	NotAtJunctionOrWithin20Metres	27.33%
		AuthorisedPerson	0.28%
		AutoTrafficSignal	27.57%
		StopSign	0.57%
		GiveWayOrUncontrolled	44.24%
6	PhysicalPedestrian-Crossing	NoPhysicalCrossingWithin50Metres	72.52%
		Zebra	0.78%
		PelicanPuffinToucanOrSimilar	6.95%
		PedestrianPhaseAtTrafficJunction	16.25%
		FootbridgeOrSubway	0.49%
		CentralRefuge	3.00%
7	UrbanOrRuralArea	Urban	84.20%
		Rural	15.80%