Applied Spatial Analysis and Policy (2025) 18:93
https://doi.org/10.1007/s12061-025-09703-0

RESEARCH

Modelling of Traffic Collisions at Road Intersections in Cape Town, South Africa: A Bayesian Spatio-Temporal Approach

Sebnem Er[1] · Álvaro Briz-Redon[2] · Sulaiman Salau[1] · Robin Lovelace[3]

Received: 25 February 2025 / Accepted: 23 July 2025

Abstract

This paper models road traffic collision counts recorded between 2015 and 2019 in a ward located in the central part of Cape Town in South Africa, using a Bayesian spatio-temporal zero-inflated Negative Binomial approach. The method accounted for the excess zeros present in collision data by separately modeling zero and non-zero collision counts, while also capturing spatial and temporal dependencies through prior distributions. Road-level information was used as fixed-effects covariates, including speed limits, presence of traffic calming measures, traffic signals, road class, number of lanes, whether the intersection is on "Main Road", and whether a public transport route passes through the intersection. The results reveal that among the covariates included in the selected model, node degree (used as a proxy for traffic flow), the presence of traffic signals, having any major road around the intersection (road class), location along "Main Road", and the presence of a taxi route at the intersection were all associated with an increase in traffic collision counts at the intersections. The years 2018 and 2019 were associated with higher collision counts compared to the reference year, 2015. For the probability component of the model, the existence of traffic signals at the intersection and location along "Main Road"were both associated with an increase in the chances of at least one collision being observed at the intersection, whereas having any high-speed road around the intersection decreased this chance.

✉ Sebnem Er
sebnem.er@uct.ac.za

Álvaro Briz-Redon
alvaro.briz@uv.es

Sulaiman Salau
sulaiman.salau@uct.ac.za

Robin Lovelace
R.Lovelace@leeds.ac.uk

1 University of Cape Town, Rondebosch, South Africa

2 University of Valencia, Valencia, Spain

3 University of Leeds, Leeds, United Kingdom

Published online: 31 July 2025

Keywords Road traffic collisions · Random effects · Spatial random effects · Zero-inflated negative binomial · Bayesian modelling

Introduction

Sustainable transport is of utter importance in achieving some of the sustainable development goals set for 2030. However, one of the main challenges that threaten sustainable transport is the large number of traffic collisions that continue to occur (UN, 2017). Therefore, analyses and estimations aimed at improving, understanding and reducing traffic collisions are essential for sustainable transport. Moreover, since traffic collisions usually occur due to a combination of road, driver and environmental related factors, identifying these factors can support measures to reduce the frequency and injury-severity of road traffic collisions. Human behaviour is often the main cause of the majority of road collisions compared to vehicle defects which account for far fewer road traffic collisions.

According to the South African Road Traffic Management Corporation's (RTMC) road safety report (Department of Transport, 2018), South Africa recorded 12921 fatalities resulting from road traffic collisions in 2018. Human-related factors was recorded as the main cause of 89.3% of fatal road traffic collisions in South Africa in 2018, with speeding, jay-walking and hit-and run noted as the most frequent causes. Road and environmental factors were found to be the main cause of 6.5% of fatal road traffic collisions in 2018, with poor visibility, wet road surfaces, and sharp bends in the road being the most frequent causes. Vehicle-related factors accounted for 4.2% of fatal road traffic collisions, with burst tyres being the most commonly identified cause. Among all the victims of fatal road traffic collisions in South Africa in 2018, 26% were drivers, 33% were passengers, 38% were pedestrians and 2% were cyclists. The report also determined that among fatal road traffic collisions, the most frequent major collision types were head-on collisions, multiple vehicle collisions as well as single vehicles overturned.

In this paper, we focus on the road traffic collisions that occurred at road intersections in the City of Cape Town metropolitan municipality between 2015 and 2019, specifically in ward 57. The City of Cape Town (CoCT) is the second largest city in South Africa with 7.7% of the population living within the municipality according to the national census 2022 (StatsSA, 2022). It is also the second largest contributor to national employment in the country (StatsSA, 2011). Analyses that enhance understanding and reducing the collisions in the metropolitan city can inform safety efforts in similar cities in South Africa and in other developing countries.

The RTMC is the main source of South African road traffic collision data used by researchers. This database includes the attributes of each collision such as the severity, time and alleged cause of the collision. The database also contains the road and intersecting road names where the traffic collisions occured. This allows the collisions that happened at an intersection to be highly precisely geocoded compared to the collisions that happened along a road segment which often have poor, incomplete location descriptions.

The primary aim of this research is therefore to model the collision counts at road intersections using road level characteristics and understand the dynamics behind the collisions, including the differences from year to year. The broader aim is to inform policy decisions by guiding the policy makers in CoCT Transport and Data Science departments, in providing solutions for potentially reducing traffic collisions and improving the overall road safety.

Literature Review

There are numerous studies analysing road traffic collisions and the factors influencing the casualties recorded on major city road networks. Most of these studies perform statistical analyses appropriate for the nature of their data and make comparisons between the methods used. Given the diversity of the study areas, data collection techniques, aggregation methods and statistical methods employed, the results from the literature may not be directly generalisable to the CoCT. However, it is important to first understand the findings of these studies before delving into the analysis of the traffic collisions dataset for CoCT.

Road traffic collision datasets are often collated by transport departments (Levine et al., 1995) based on individual incident reports submitted in paper format by regional police departments. The datasets available to researchers often lack the precise locations of collisions, providing only the names of the road(s) where the collisions occurred. While it is relatively easier to geocode the exact location of a collision at an intersection, pinpointing the location becomes more challenging when a collision occurs along a road segment and only the road name is provided. In instances were collisions cannot be attributed to an exact location, such data is either aggregated to a road segment, or to an areal unit such as provinces, districts or wards (Erdogan, 2009). Some studies have analysed collisions at intersections separately (Lee et al., 2017), while others have built separate models for intersections and along road segments, later merging them for more comprehensive results (Briz-Redón et al., 2019b). The different levels of precision and different levels of aggregation have made the collision research literature extremely rich. However, it is important to note that aggregating the traffic collisions into collision count data at road segments, intersections or areal units leads to modifiable areal unit problem as investigated in a study by Gilardi et al. (2022).

Regardless of the chosen level of aggregation, traffic collision data aggregated at any scale may result in zero-inflated count data characterised by instances where no collisions are recorded either due to chance, or due to conditions that make collisions highly unlikely. As a result, researchers have mostly utilised zero-inflated Poisson or Negative Binomial models (Lee et al., 2017; Briz-Redón et al., 2019b; Gu et al., 2020; Yan et al., 2020; Briz-Redón et al., 2021; Gilardi et al., 2022; Wang et al., 2024) estimated either using maximum likelihood or Bayesian methods. In these models, researchers have considered various factors that may influence collision frequency, including driver characteristics, road conditions (Lee et al., 2017; Briz-Redón et al., 2019b; Liu et al., 2022) environmental factors (Liu et al., 2022) and aerial factors (Lee et al., 2017). More recently, the models have been improved with the inclusion of some

random effects and spatial correlation structures that capture the dependencies present in collision counts. Such work is captured in da Silva and de Sousa (2023) where the Negative Binomial model estimation was extended to geographically weighted regression setting in order to incorporate the spatial nature of the traffic collision count data. There is also an increasing interest in the use of spatial network analysis methods especially when the precise locations of the collisions are available (Okabe et al., 2009; Okabe and Sugihara, 2012; Baddeley et al., 2015; McSwiggan et al., 2017; Baddeley et al., 2021). Network kernel density estimation methods have been used to estimate the density (Xie and Yan, 2008; Okabe et al., 2009; Produit et al., 2010; Tang et al., 2013; Kaygisiz and Hauger, 2017) and factors affecting the density of the collisions occuring along a road network (Borruso, 2005, 2008; Xie and Yan, 2013).

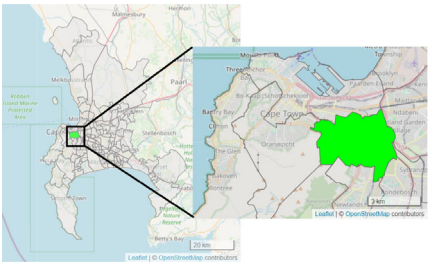
An extensive summary of the existing studies regarding the different dataset types, covariates and methods used can be found in Lord and Mannering (2010); Mannering et al. (2016) and Baddeley et al. (2021). In South Africa, there is not an extensive body of research on the analysis of traffic collisions. One of the earlier studies in South Africa focused on monthly data, collected from January 2007 to April 2009 in Tshwane municipality, and used a Poisson-Gamma model to analyse pedestrian fatalities and explore the possible infrastructural improvements that could help reduce the fatalities (Das, 2014). The analysis utilised a Bayesian approach and found that simple solutions like street lighting can reduce the number of fatalities. More recently, Marchant and Norman (2022) examined the effect of relighting of street lamps on the weekly numbers of road traffic collisions in Leeds, UK, using multilevel modelling approach. The aim of their study was to investigate whether the newly installed white light street lamps improved road safety compared to the previous orange light lamps. However, the evidence for improved safety under the new lighting was weak (Marchant and Norman, 2022).

Building on these findings from the literature and acknowledging the limitations in the data and geographic specificity, this study aims to contribute to the growing body of traffic collision research by applying a spatially-aware modeling approach to the CoCT context. The next section details the study area and the steps taken to prepare the dataset for analysis, while Section "Methodology" provides the details of the methods used. Results are summarised in Section "Results", and finally Section "Conclusions and Future Research" documents the findings and provides recommendations for future research.

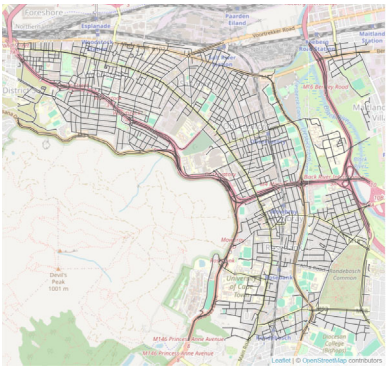
Study Area, Data Preparation and Variables

Study Area, Data Preparation

The main focus in this study is the analysis of traffic collisions that happened at the road intersections within ward 57 of CoCT. Ward 57, indicated by the green polygon in Fig. 1a, is one of the 116 wards in the metropolitan municipality. The ward includes the very central areas of CoCT, namely Gardens, Mowbray, Observatory, Salt River, Vredehoek, Woodstock and Zonnebloem where lots of economic activity take place. The ward serves as a key link between CoCT's city centre and surrounding suburbs,



(a) Ward 57 within the CoCT shown in green polygon



(b) Road network obtained from the CoCT open data portal

(c) The intersections from the road network shown in blue dots for a zoomed in area in ward 57

Fig. 1 Study area and road network

making it a major commuter corridor and an essential route for daily commuters. The heavy traffic volumes within the ward makes it one of the areas with the highest concentration of road traffic collisions in the CoCT. The road network in ward 57 shown in Fig. 1b has 2316 road segments (edges) and 1633 nodes. As shown in blue dots in the zoomed map in Fig. 1c, some of these nodes in the road network technically belong to the same intersection. The smaller roads intersecting with a major road often have two intersection points (nodes) on the map which should be treated as one intersection.

The collision dataset used in this paper was obtained from the CoCT Transport Department, following a formal data request made by the authors. Collision report forms are routinely collected by the South African Police Service and subsequently submitted to the RTMC following verification and data entry procedures (City of Cape Town, 2023). However, these records typically lack precise geolocation details such as longitude and latitude. The dataset provided for this study did include the road names and the related descriptive information, consistent with the types of entries

Table 1 Hypothetical examples of the data acquired from the CoCT

Information	Example 1	Example 2
roaddesc	BURG ST, CBD	BUITENGRACHT ST, CBD
nodedesc	STRAND ST **X** BURG ST	BUITENGRACHT ST, CBD
policestation	Cape Town Central SAPS	Cape Town Central SAPS
collisiondate	2015-01-02	2015-05-22
time	09:03	13:05
collisiontype	Sideswipe - same *direction*	Approach at angle - one or both turning
alledgedcause	Sudden stop	Drive on wrong side
vehicletype	A: Motor car/station wagon, B: Motor car/station wagon	A: Motor car/station wagon, B: Motor car/station wagon
numberofdrivers	2	2
numberofpassengers	1	0
numberofpedestrians	0	0
fatal	0	0
serious	0	0
slight	0	0
noinjuries	3	2
unknowninjuries	0	0

illustrated in the hypothetical examples shown in Table 1. It is reassuring to know that in the CoCT, there are initiatives with regards to data collection via user apps, however, currently collision data are recorded manually. The second column of Table 1 provides an example (Example 1) of a collision that occurred at an intersection. The *nodedesc* row lists the names of the intersecting roads, separated by an **X** in the collision description. The third column gives an example (Example 2) of a collision that happened along a road segment. For this study, we only focus on the collisions that happened at an intersection since both of these examples needed to be geocoded differently.

In the absence of additional location information, road collisions that are recorded in the format of Example 2 need to be geocoded to a representative location (typically middle of the road segment) as the geolocation of that particular collision. If the road traffic collision happened along a kilometer long road segment, then the lack of precise geocoding becomes crucial in the analysis stage. On the contrary, all collisions that had a node description (names of the intersecting roads separated by an **X**) can be more easily geocoded using a geocoding package. In the case of intersections, we have more precise information compared to the collisions that happened along the road segments especially for long roads. Briz-Redón (2024) suggests incorporating the uncertainty about the exact location of the collision through a Bayesian modelling approach. In this research, we do not use this approach since the intersections along the road network in the CoCT are well separated from each other.

In this study, for the collisions at the intersections, the `arcgisgeocode` R package (Josiah, 2024) was used for obtaining the longitude and latitude information of the collisions. The reliability of the results (based on the match score of the node description to which the address was matched) obtained from `arcgisgeocode` R package (Josiah, 2024) was compared with the `geocodeHERE` R package (Nissen, 2014) for the years between 2015-2017 and the results are presented in Fig. 2.

Using `arcgisgeocode` (Josiah, 2024), all of the collisions that happened at an intersection were geocoded, whereas only approximately 30% of the intersection collisions were geocoded properly with the `geocodeHERE` (Nissen, 2014) producing lots of empty cells. As an example, in 2015, road traffic collision records had 71862 collisions in total of which 27818 happened at the intersections. `geocodeHERE` (Nissen, 2014) geocoding method had 32.75% collisions succesfully geocoded for 2015. This low level of geocoding efficiency can be attributed to poorly recorded road names from traffic collision records.

In general, the `geocodeHERE` (Nissen, 2014) R package produced less reliable results compared to `arcgisgeocode` (Josiah, 2024). The average reliability score of the geocodings with `arcgisgeocode` (Josiah, 2024) was 92.57 (where 100 indicates a perfect match) with a standard deviation of 4.63 (coefficient of variation (CV) as 0.05). The score was 90.41 on average with a standard deviation of 10.92 (CV as 0.12) using `geocodeHERE` (Nissen, 2014) R package. Based on these results a decision was made to geocode the collision addresses for the years in consideration using the `arcgisgeocode` (Josiah, 2024) R package.

Since the main aim of this study is to analyse the number of collisions at intersections, it was necessary to convert the road network node information into intersections. Intersections shown in Fig. 3c were obtained by locating a 10m buffer around each of

the 1633 nodes as shown in Fig. 3a., Finally, overlapping buffers were merged to avoid double counting of the traffic collisions that happened at a particular buffer (Fig. 3b). In Fig. 3, the red circular polygons are the buffers, black points are the nodes obtained from the road network and the purple lines are the road segments (edges) obtained from the road network. The green points on this map represent the geocoded collisions which are intersected with the merged buffers around the intersections to obtain the number of collisions that happened at an intersection location.

Figure 4 provides the collision point data with respect to the intersections for each study year. The light blue colour points represent intersections with no collisions, light to dark purple coloured points represent intersections with one, two, and three collisions, respectively. Dark red points stand for four collisions, whereas intersections that had five or more collisions are coloured with bright red.

Figure 5a shows that the dataset contains a lot of intersections with zero collisions, approximately 20% of the overall collisions. This information plays a key role in identifying the method to be used to analyse the counts of collisions at intersections. Further inspections of the collision data reveal that there were two intersections, namely Offramp on Settlers Way and Onramp on Berkley Road, with more than 50 collisions per year (65-93 collisions per year). These two intersections were excluded from the analysis since they are outliers considering the range of the number of collisions at other intersections (0-35 collisions per year) and their unique characteristics. These two intersections in ward 57 are complicated off/on-ramps which are quite different from the rest of the intersections structurally. Offramp on Settlers Way is the

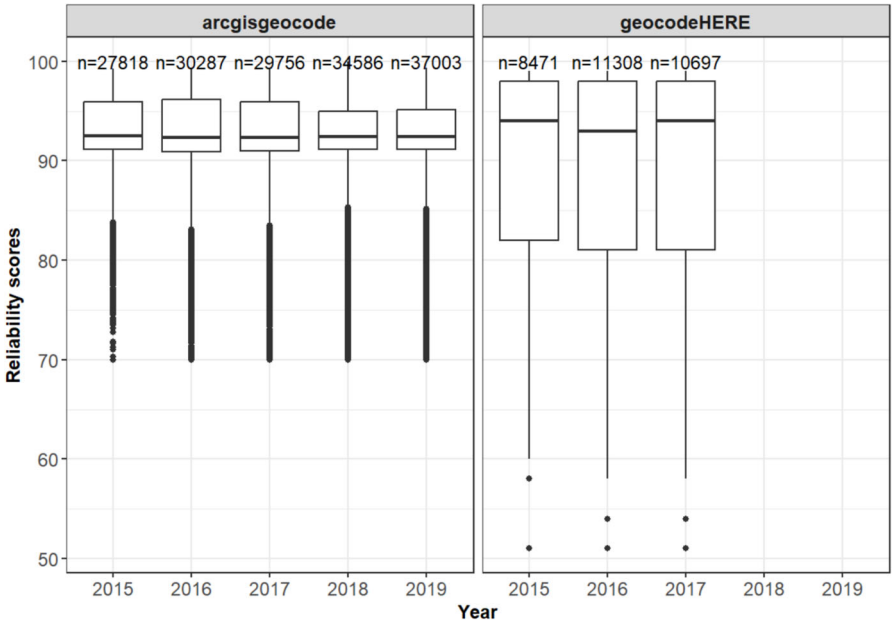
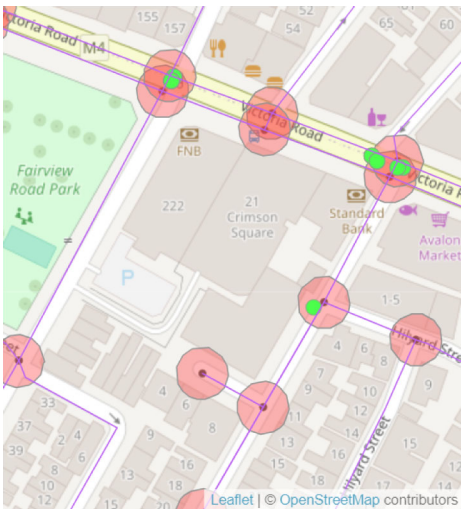
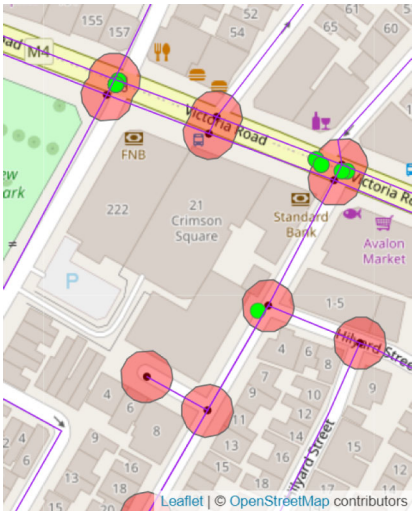


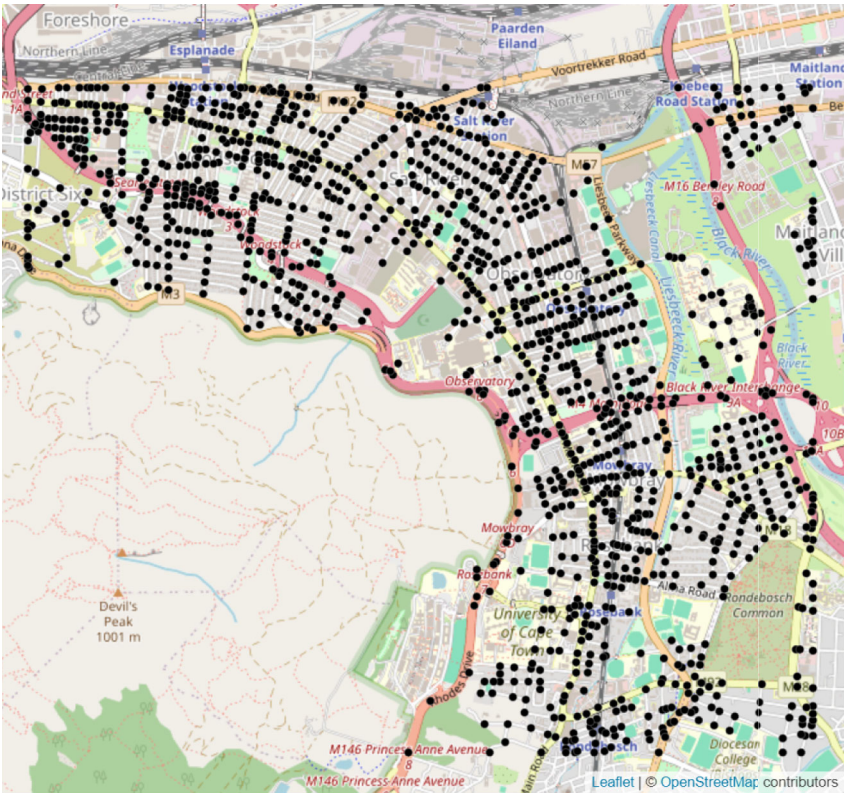
Fig. 2 Reliability score comparison between the geocoding R packages `arcgisgeocode` (Josiah, 2024) and `geocodeHERE` (Nissen, 2014) for 2015-2019 collision data for the CoCT Municipality



(a) Buffers around intersections before merging

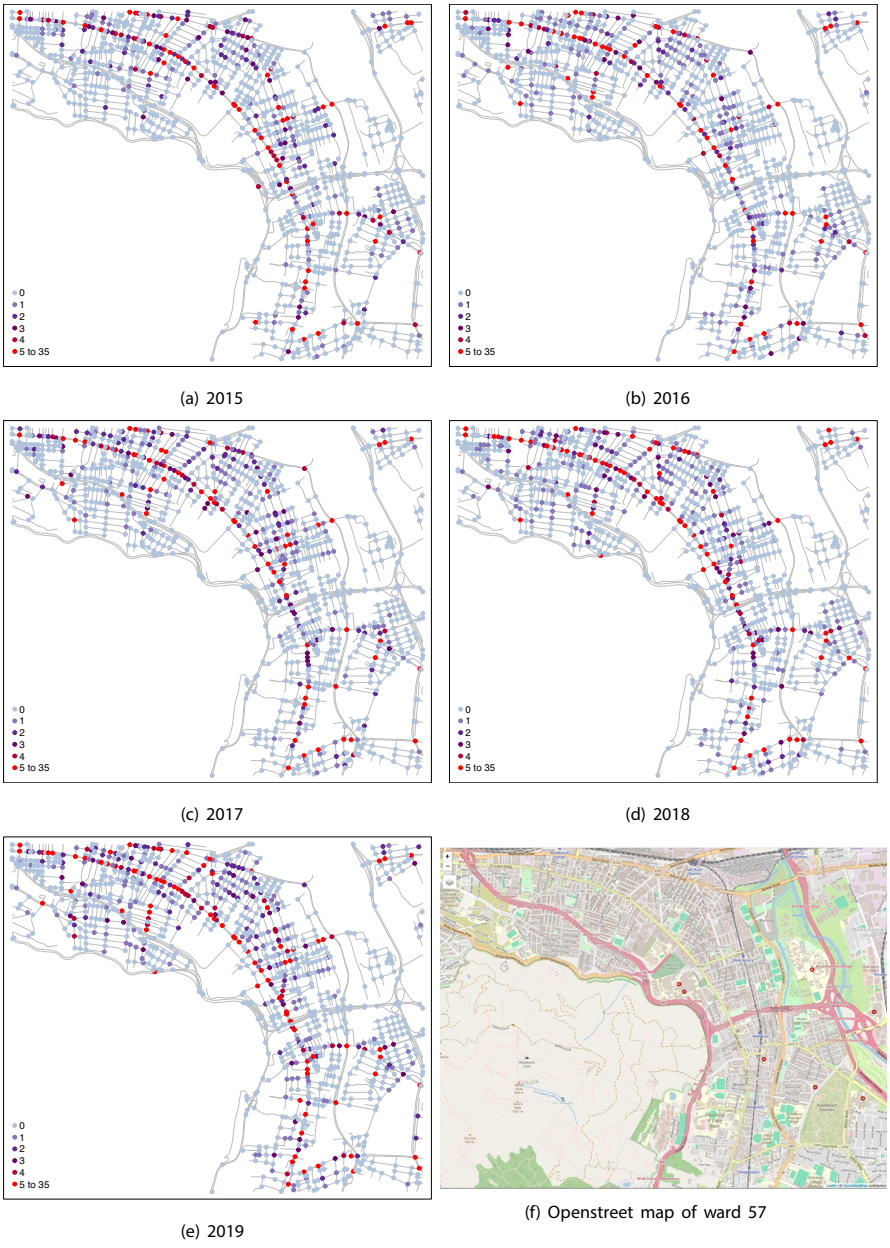


(b) Buffers around intersections after merging



(c) Ward 57 intersection locations

Fig. 3 Ward 57 in Cape Town intersection locations identified with buffers in red colour placed around the nodes of the road network

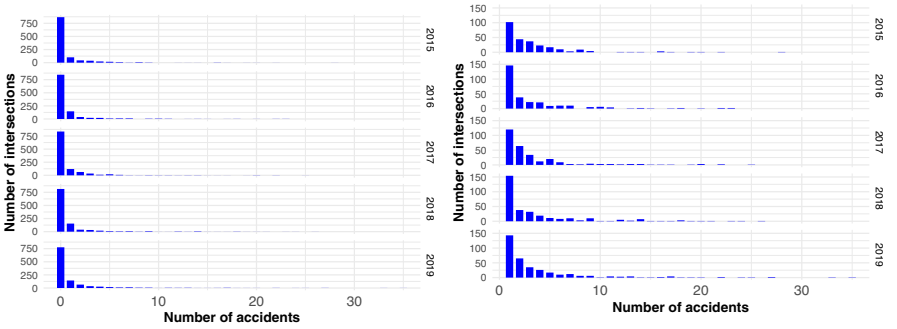


(a) 2015 (b) 2016

(c) 2017 (d) 2018

(e) 2019 (f) Openstreet map of ward 57

Fig. 4 Geographical distribution of observed collision counts at intersections in ward 57, Cape Town, South Africa between 2015-2019



(a) Frequency distribution of traffic collisions in ward 57 at intersections

(b) Frequency distribution of zero-truncated (one or more) traffic collisions in ward 57 at intersections

Fig. 5 Total number of traffic collisions per year that happened in ward 57 in 2015-2019

offramp road leading to N2 (highway) that connects the city to the airport and to the east of the country. Onramp on Berkley Road is very similar to the Offramp on Settlers Way in this context. These locations were identified as statistical outliers based on preliminary exploratory analysis, including frequency-based collision distributions.

The frequency distribution of one or more traffic collisions at all other intersections within ward 57 is given in Fig. 5b, reflecting an increase in the last two years of the study period for ten or more collisions, and a clear concentration around one collision.

Finally, for the purpose of analysing the traffic collision counts at the intersections, we consider road intersection variables that are hypothesised to have an impact on the number of collisions. Road intersection variables were also generated from the properties of the roads that pass through the intersections.

Road Intersection Variables

In this section, we describe the variables and the reasons for including them in the models. One of the important variables in traffic collision analysis is the traffic flow variable. Traffic collision counts at intersections can be affected by the amount of traffic, and therefore the count can be either adjusted with the traffic flow or the traffic flow variable can be added as a covariate in the analysis. Given the lack of traffic flow information, different authors use different types of proxies for traffic flow approximation (Briz-Redón et al., 2019a, 2021; Chaudhuri et al., 2023). Since the yearly road traffic flow informations was not available for all roads or intersections in the current study, we used the road network *nodedegree* covariate (denoted as X_1 in the models) as a proxy for traffic flow. The node degree in a network represents the number of direct road connections (i.e., edges - road segments) associated with each intersection (node) in the road network, thereby capturing its relative connectivity and potential for higher traffic flow.

All other explanatory variables considered are binary variables that take on the value of one (true) if the intersection contains that property. Specifically, the maximum speed variable (X_2) (*maxspeed*) is one (true) when any of the roads intersecting at the

intersection have a maximum speed limit of 80 km/h, 100 km/h or 120 km/h, and zero (false) for roads intersecting at the intersection have a speed limit of 60 km/h or less. This distinction is very important since in the CoCT, majority of the roads have a speed limit of 60 km/h or less which is an indication of residential roads. The number of lanes variable (X_6) (*lane*) is one (true) when any of the road segments around the intersection has three or more lanes, and zero (false) otherwise. The roads in the city contain one to six lanes where three or more lane roads are associated with major roads such as highways. Both *lane* and *maxspeed* variables are considered as an indication of high complex intersections. The road class (*roadclass*) variable (X_5) has five different categories. Category one stands for highways and major roads and the higher the category is, the less of a major road it is. Taking this into account, to be able to distinguish the highways from others, the *roadclass* variable is coded one (true) when any of the road segments around the intersection are of road class one, two or three.

Traffic signals (*trafficsignals*) and traffic calming (*trafficcalming*) variables (X_3 and X_4, respectively) were acquired using osmdata R package (Padgham et al., 2017) with the `traffic_signals` and `traffic_calming` key features which return point-level information. In the case of traffic calming, we include the total number of traffic calming measures within 40 meters of any of the roads around the intersection in the analysis. This variable is an ordinal variable ranging from a minimum of zero to a maximum of three. This means that any intersection can have between zero and three traffic calming measures, such as speed bumps, stop signs, and similar features. Therefore, the *trafficcalming* variable is converted into a binary variable, where the variable takes the value of zero (false) when there are no calming measures within the 40 meters of the intersection, and one (true) if one or more calming measures are in place. The expectation around this variable is that the more traffic calming put in place, the less the number of traffic collisions will happen at the intersections. The *trafficsignals* variable (X_3) is very similar to the *trafficcalming* variable (X_4) in the sense that the variable is ordinal with a minimum of zero and a maximum of five. Some intersections in the CoCT are very complicated and can have three, four or even five traffic signals but these intersections are very rare. As a result, this variable is also converted into a binary variable where a zero (false) indicates that the intersection has no traffic signals in place, whereas a one means that there are one or more traffic signals at that intersection. This variable shows the complexity of the intersection and it is expected that those intersections with high number of traffic signals may have more collisions.

Another variable included in the analysis is the *mainroad* variable, an indicator for an intersection located on "Main Road"in the CoCT (similarly known as "M4"according to the South African national road numbering system), one of the busiest roads in the city, extending from one end of the city centre to the other end. There are lots of foot traffic and businesses located in the surroundings of "Main Road". Figure 4 reveals the concentration of collisions along "Main Road"in the city. The (*mainroad*) variable (X_7), takes a value of zero (false) if the intersection is not along the "Main Road"and takes a value of one if it is.

In the CoCT, commuters often rely on minibus taxis as a means of public transportation. Some people also use the public busses (MyCity or Golden Arrow busses) which

have fewer route options. The existence of a minibus taxi or a public bus route makes the road busier and we expect to see more collisions if a public transport route passes through the intersection. In addition, some roads in the CoCT have on-street parking available to drivers often managed by parking payment officials or car guards. This of course can create collisions with two vehicles or with a vehicle and a pedestrian while parking or leaving the parking spot. Considering all three variables, minibus taxi route (X_8), mycity bus route (X_9) and on street parking (X_{10}), three more binary variables are created. In particular, *minibustaxi* and *mycitybus* variables take a value of one (true) if a minibus taxi or a my city bus route passes through the intersection, and a value of zero (false) otherwise. Similarly, the *onstreetparking* variable is coded as zero (false) if there are no parking facilities available at the intersection, and as one (true) if an on street parking is available at the intersection.

Lastly, four additional binary variables ($X_{11}, \ldots, X_{14}$) are created to indicate the year of the collision relative to the reference year 2015. These variables are valuable for understanding temporal trends and changes over time. Specifically, binary variables representing years relative to a reference year (such as 2015) can help identify trends in collision rates, and account for unobserved factors that vary across years, like weather patterns, road infrastructure changes, or vehicle safety advancements. All these variables, summarized in Table 2, are utilized in modeling traffic collision counts to assess their effects on collision counts. Since the number of collisions response variable is a count variable, count data modelling approaches are the most suitable models and these models are explained in the following section.

Methodology

Model Specification

A variety of statistical models have been applied to collision count datasets along various road segments and/or intersections. Considering the CoCT traffic collisions dataset, let $Y_{it} = 0, 1, \ldots$ be the number of traffic collisions that occurred at the $i-th$ road intersection ($i = 1, \ldots, n$) within ward 57 at year t ($t = 1, \ldots, T$). In this study, there are 1127 intersections in ward 57 ($n = 1127$) observed over five years between 2015 and 2019 ($T = 5$). Since we have many intersections with zero collisions, we assume that the number of traffic collisions at intersections follow a zero-inflated Negative Binomial distribution which is an extension of the Poisson distribution. The choice of the Negative Binomial instead of the Poisson allows us to account for overdispersion in the collision counts.

The zero-inflated Negative Binomial (ZINB) distribution is expressed in two components separating the zero counts from the positive counts with p_{it}, the probability of observing false zero collisions at an intersection in a particular year. In traffic collisions counts, a collision at an intersection might not be observed simply because the data was not captured due to the fact that the collision was not reported, or the intersection has some other properties that make it impossible to have a collision. This structure

Table 2 Variable specifications considered in this study

Variable	Definition	Type
Y: numberofcollisions	Total number of collisions occured at a road intersection per year	Count
X_1: nodedegree	The node degree as a proxy for road traffic flow	Numeric
X_2: maxspeed	True if the speed limit at any of the road segments around the intersection is 80km/h or more	Binary
X_3: trafficsignals	True if there is one or more number of traffic signals on the intersections	Binary
X_4: trafficcalming	True if there is one or more number of traffic calming on the roads within 40m around the intersection	Binary
X_5: roadclass	True if any of the road segments around the intersection are class 1, 2 or 3 (which indicates major roads)	Binary
X_6: lane	True if any of the road segments around the intersection has 3 lanes	Binary
X_7: mainroad	True if the intersection is on "Main Road"	Binary
X_8: minibustaxi	True if the intersection is on a minibus taxi route	Binary
X_9: mycitybus	True if the intersection is on a MyCityBus bus route	Binary
X_{10}: onstreetparking	True if there is an on-street parking available around the intersection	Binary
$X_{11} - X_{14}$: year	The year the traffic collision occured (2015-2019)	Categorical

is defined as follows (Zuur et al., 2009):

$$Y_{it} \sim \text{ZINB}(\lambda_{it}, \theta, p_{it}); \quad \text{and}$$
$$P_{\text{ZINB}}(Y_{it} = y_{it}) = \begin{cases} p_{it} + (1 - p_{it}) \left(\frac{\theta}{\theta+\lambda_{it}}\right)^{\theta} & \text{if } y_{it} = 0 \\ (1 - p_{it}) \frac{\Gamma(y_{it}+\theta)}{\Gamma(y_{it}+1)\Gamma(\theta)} \left(\frac{\theta}{\theta+\lambda_{it}}\right)^{\theta} \left(\frac{\lambda_{it}}{\theta+\lambda_{it}}\right)^{y_{it}} & \text{if } y_{it} > 0 \end{cases} \quad (1)$$

where $\Gamma(y_{it}) = (y_{it} - 1)!$, λ_{it} is the mean and θ is the shape parameter of the distribution. Here $E(Y_{it}) = \lambda_{it}(1 - p_{it})$ and $Var(Y_{it}) = (1 - p_{it}) \left(\lambda_{it} + \frac{\lambda_{it}^2}{\theta}\right) + \lambda_{it}^2(p_{it}^2 + p_{it})$.

The mean parameter for the positive counts are modelled as a linear function of the K covariates with $k = 1, \ldots, K$ via a log link function (Eq. 2), while the probability of having a false zero (p_{it}) is modelled with the same set of covariates using a logit link function (Eq. 3). Of course, the set of the covariates used do not need to be the

same. Due to the nature of our dataset and the possible spatial autocorrelation we might have, it is important to include the random effects as well as the fixed effects from the observed covariates. These random effects are included in three parts: (i) the structured spatial random effects, u_i; (ii) unstructured spatial random effects, v_i, and (iii) unstructured space-time random effects, ϵ_{it}. The spatial random effects are included in both components of the model and labelled with the superscript (λ) for the count and (p) for the probability random effects. The unstructured space-time random effects, ϵ_{it}, are only included in the count component. Considering these specifications, the count and probability components are defined as follows, with the superscript (λ) and (p), respectively:

$$\log(\lambda_{it}) = \eta_{it} = \beta_0 + \sum_{k=1}^{K} \beta_k X_{it,k} + u_i^{(\lambda)} + v_i^{(\lambda)} + \epsilon_{it}^{(\lambda)} \tag{2}$$

$$\text{logit}(p_{it}) = \gamma_0 + \sum_{k=1}^{K} \gamma_k X_{it,k} + u_i^{(p)} + v_i^{(p)} \tag{3}$$

In Eqs. 2 and 3, β_0 and γ_0 s are the respective intercepts, and β_k and γ_k are the linear fixed effects associated with the set of covariates included in the count and the probability component, respectively. To investigate the effects of the covariates on traffic collision counts, we use a Bayesian approach that enables the analysis of highly complex models, including those with spatio-temporal random effects for count data. The main difference of a Bayesian approach compared to a frequentist approach is that the unknown parameters are not assumed to be fixed (constant), rather they are assumed to be stochastic with a prior distribution before the data realisation. In the Bayesian framework, non-informative priors are specified for the intercept and fixed effects terms, β_0, β_k and γ_0, γ_k.

The unstructured spatial random effects are assumed to follow a Gaussian distribution with mean zero and constant variance, $v_i^{(\lambda)} \sim N(0, \sigma^2_{v^{(\lambda)}})$ and $v_i^{(p)} \sim N(0, \sigma^2_{v^{(p)}})$. The structured spatial random effects $u_i^{(\lambda)}$ and $u_i^{(p)}$ are specified with a conditional autoregressive (CAR) covariance specification as suggested by Besag et al. (1991) :

$$u_i^{(\lambda)} | \mathbf{u}_{-i}^{(\lambda)} \sim N\left(\sum_{j=1}^{n} w_{ij} u_j^{(\lambda)}, (s_i^{(\lambda)})^2 \right), \quad (i \neq j) \tag{4}$$

$$u_i^{(p)} | \mathbf{u}_{-i}^{(p)} \sim N\left(\sum_{j=1}^{n} w_{ij} u_j^{(p)}, (s_i^{(p)})^2 \right), \quad (i \neq j) \tag{5}$$

where, $(s_i^{(\lambda)})^2 = \sigma^2_{u^{(\lambda)}}/\mathcal{N}_i$ in Eq. 4 and $(s_i^{(p)})^2 = \sigma^2_{u^{(p)}}/\mathcal{N}_i$ in Eq. 5 are the variances for the i-th intersection depending on the number of neighbours the i-th intersections has ($\mathcal{N}_i = \#\mathcal{N}(i)$). It is clear that if an intersection has many neighbours, then the variance of that intersection will be smaller. Here, w_{ij} represents the row normalized (i, j) entry of the symmetric weights matrix $\mathbf{W}_{n\times n}$ which expresses the spatial weight between

each intersection and its neighbours on the road network. There are several different ways of defining the neighbourhood relationships. In this study, the weights matrix is calculated using a contiguity neighbourhood relationship where two intersections i and j on the road network that have a joint road segment are considered as neighbours with a non-zero w_{ij}. Additionally, the diagonal elements of the matrix are zero, $w_{ii} = 0$. Figure 6 depicts how row normalized weights matrix is obtained for an example of eight intersections chosen along the road network. Here, the *first* intersection ($I = 1$) on the road network has two neighbouring intersections, ($I = 2$) and ($I = 7$), with a shared road segment. The values of the row normalized weights matrix for those neighbouring intersections will be 0.5 (=1/2) and the rest will be zero along the *first* row. Similarly, the *second* intersection ($I = 2$) on the road network has three neighbouring intersections, ($I = 1$), ($I = 3$) and ($I = 6$), with a shared segment. The values of the row normalized weights matrix for those neighbouring intersections will be 0.333 (=1/3) and the rest will be zero along the *second* row.

(a) Map of selected intersections along the road network

	I1	I2	I3	I4	I5	I6	I7	I8
I1	0	1	0	0	0	0	1	0
I2	1	0	1	0	0	1	0	0
I3	0	1	0	1	0	0	0	0
I4	0	0	1	0	1	0	0	0
I5	0	0	0	1	0	0	0	0
I6	0	1	0	0	0	0	0	0
I7	1	0	0	0	0	0	0	1
I8	0	0	0	0	0	0	1	0

(b) Binary adjacency matrix

	I1	I2	I3	I4	I5	I6	I7	I8
I1	0	0.5	0	0	0	0	0.5	0
I2	0.33	0	0.33	0	0	0.33	0	0
I3	0	0.5	0	0.5	0	0	0	0
I4	0	0	0.5	0	0.5	0	0	0
I5	0	0	0	1	0	0	0	0
I6	0	1	0	0	0	0	0	0
I7	0.5	0	0	0	0	0	0	0.5
I8	0	0	0	0	0	0	1	0

(c) Row normalized weights matrix

Fig. 6 An example for the weights matrix construction for eight intersections along the road network

Finally, a Gaussian prior with mean zero and constant variance is considered for the unstructured spatio-temporal random effects $\epsilon_{it}^{(\lambda)} \sim N(0, \sigma^2_{\epsilon^{(\lambda)}})$. For the variances of the random effects in the count and probability components $(\sigma^2_{v^{(\lambda)}}, \sigma^2_{u^{(\lambda)}}, \sigma^2_{\epsilon^{(\lambda)}}, \sigma^2_{v^{(p)}}, \sigma^2_{u^{(p)}})$ a uniform prior was defined where all values between zero and 10 are assumed to be equally likely ($U(0, 10)$ distribution). There are different approaches for establishing a prior distribution on the variance parameter. One approach, the one we have followed, is assuming a $U(0, A)$ distribution, where A is chosen to be large enough, considering the scale and magnitude of the data. This kind of approach is supported by current literature (Martínez-Beneito and Botella-Rocamora, 2019).

Variable Selection and Model Evaluation

The models specified in Table 3 are estimated to assess the effects of our covariates as well as the effects of the random effects using Bayesian model specifications with some prior distributions. The models are compared using the Watanabe-Akaike Information Criterion (WAIC) proposed by Watanabe and Opper (2010) where a smaller WAIC measure indicates a better model performance.

The full model (Model 3) incorporates all covariates and all random effects described in Eqs. 2 and 3 for both the count and the false zero probability component (where $X^{(\lambda)} = X^{(p)}$). Table 3 presents the variations of this full model considered in this study. Model 1, Model 2 and Model 3 include all fixed effects in both the count and probability component of the models and these models only differ in their random effects specifications. Thereafter, following from the literature (Kuo and Mallick, 1998; Dellaportas et al., 2002; Sosa et al., 2023), a Bayesian variable selection method is implemented (given in Model 4) on the fixed effects part of Model 3 by introducing an inclusion probability parameter (I_k) that ranges between zero and one ($0 \leq I_k^{(\lambda)}, I_k^{(p)} \leq 1$). The index parameter has a Bernoulli prior with $I_k^{(\lambda)} \sim Bernoulli(0.5)$ and $I_k^{(p)} \sim Bernoulli(0.5)$. We have chosen a Bernoulli prior with parameter 0.5 for the inclusion probability of each covariate in order to be uninformative. Since we do not have a strong belief (a priori) about the inclusion of a covariate in the model, this represents assigning equal probability to both the inclusion/exclusion of the corresponding covariate. Here the main aim is to assess the effect of the variables ($I_k = 0$ means that the corresponding k-th variable is not included in the model) and thereafter to eliminate the ones that have an index of zero to ultimately reach to Model 5, the model with selected variables only.

Results

Before presenting the model-based results, we begin with a high-level overview of the collision data to contextualize the magnitude of the problem. Over the study period from 2015 to 2019, a total of 777228 road traffic collisions were recorded across the CoCT, averaging approximately 155000 collisions per year. Within ward 57 (the primary focus area of this study) 161869 collisions were reported during the same

Table 3 Model specifications considered in this study

Model	Fixed Effects	Model formula
Model 1	All	$\log(\lambda_{it}) = \beta_0 + \sum_{k=1}^{K} \beta_k X_{it,k}^{(\lambda)} + v_i^{(\lambda)}$
		$\text{logit}(p_{it}) = \gamma_0 + \sum_{k=1}^{K} \gamma_k X_{it,k}^{(p)} + v_i^{(p)}$
Model 2	All	$\log(\lambda_{it}) = \beta_0 + \sum_{k=1}^{K} \beta_k X_{it,k}^{(\lambda)} + u_i^{(\lambda)} + v_i^{(\lambda)}$
		$\text{logit}(p_{it}) = \gamma_0 + \sum_{k=1}^{K} \gamma_k X_{it,k}^{(p)} + u_i^{(p)} + v_i^{(p)}$
Model 3	All	$\log(\lambda_{it}) = \beta_0 + \sum_{k=1}^{K} \beta_k X_{it,k}^{(\lambda)} + u_i^{(\lambda)} + v_i^{(\lambda)} + \epsilon_{it}^{(\lambda)}$
		$\text{logit}(p_{it}) = \gamma_0 + \sum_{k=1}^{K} \gamma_k X_{it,k}^{(p)} + u_i^{(p)} + v_i^{(p)}$
Model 4	All with BVS	$\log(\lambda_{it}) = \beta_0 + \sum_{k=1}^{K} \beta_k I_k^{(\lambda)} X_{it,k}^{(\lambda)} + u_i^{(\lambda)} + v_i^{(\lambda)} + \epsilon_{it}^{(\lambda)}$
		$\text{logit}(p_{it}) = \gamma_0 + \sum_{k=1}^{K} \gamma_k I_k^{(p)} X_{it,k}^{(p)} + u_i^{(p)} + v_i^{(p)}$
Model 5	Selected from Model 4	$\log(\lambda_{it}) = \beta_0 + \sum_{k \in Sel_A} \beta_k X_{it,k} + u_i^{(\lambda)} + v_i^{(\lambda)} + \epsilon_{it}^{(\lambda)}$
		$\text{logit}(p_{it}) = \gamma_0 + \sum_{k \in Sel_B} \gamma_k X_{it,k} + u_i^{(p)} + v_i^{(p)}$

Sel_A and Sel_B represent the indices of the selected set of variables for the count and probability component of the model

period, corresponding to roughly 20.8% of the citywide total. The annual counts for both the city and the ward exhibit an increasing trend, highlighting the persistent nature of road traffic incidents in this urban context.

The models described in Table 3 are estimated using a Bayesian approach with all analyses conducted in R software v.4.4.1 (R Core Team, 2024) through the `nimble` package (de Valpine et al., 2017). The `nimble` package (de Valpine et al., 2017) is based on Markov Chain Monte Carlo (MCMC) procedures and a single chain with 200k iterations, including 100k of burn-in and a thinning of 10 is used. All plots and spatial data manipulations were done using `tmap, ggplot2, sf and igraph` R packages (Tennekes, 2018; Pebesma et al., 2018; Wickham, 2011; Csardi, 2013).

The WAIC values for the models range from 8318.14 for Model 1 to 8268.84 for Model 5 in a decreasing order, with Model 5 having the smallest WAIC value. Therefore, according to this criteria, Model 5 with selected variables, structured and unstructured spatial random effects for both the count and the probability component of the model, and unstructured spatio-temporal random effects for the count component - is chosen as the final model for the analysis of the collision data. The selection of the variables is based on the Bayesian variable selection applied on Model 4 by assessing the inclusion parameters of each of the variables included in Model 4. The posterior distribution of the inclusion parameters ($I_k^{(\lambda)}$ and $I_k^{(p)}$) provide a clear picture of which variables should be included in the final model (Posterior mean, median and standard deviations for the inclusion parameters given in Table 4). Those variables with inclusion parameter posterior means lower than 0.5 have been excluded from the final model. The posterior mean of indicator parameter for the street parking variable (X_{10}) is 0.504, which is slightly over 0.5 threshold, and therefore this variable was not included in the probability component of the model. Similarly, the indicator parameter for this variable in the count component was very close to zero and for this reason it was not included in the count component of the model. From this perspective, the inclusion of "Main Road"(X_7) variable in the probability component can be questioned as well since the inclusion parameter ($I_7 = 0.531$) for this variable is also slighly higher than the threshold value 0.5. However, we believe that "Main Road"is conceptually an important variable and therefore should be included in both the count and the probability components.

Fixed Effects

Specifically, for all categorical variables holding all else constant, the coefficient for each category is related to the log difference in the posterior mean counts for that category compared to the reference category - in this study, the reference category is the zero (false) category exhibiting that the property of that variable at the intersection does not hold.

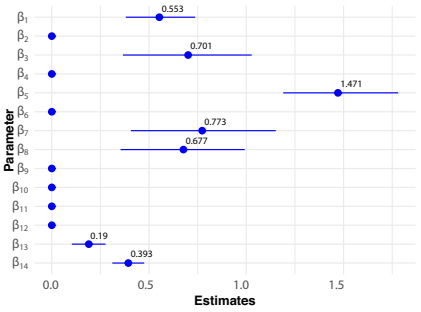
The posterior mean and the 95% credible intervals for the fixed effects are provided in Fig. 7 on the left pane with the posterior means annotated on the plot, and the density plots of the posterior distributions are provided on the right pane of Fig. 7. It is observed that all the fixed effects included in the count component of Model 5 are all positive. Amongst the covariates chosen for the count component of the model, all have positive posterior means and 95% credible intervals that do not overlap with zero. The positive

Table 4 Posterior mean, median and standard deviations for the $I_k^{(\lambda)}$ and $I_k^{(p)}$ parameters in count and probability components of the model

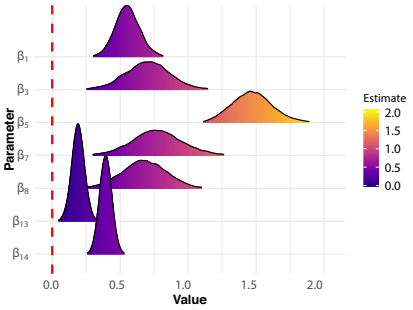
Count component				Probability component			
$I_k^{(\lambda)}$	Mean	Median	St.Dev.	$I_k^{(p)}$	Mean	Median	St.Dev.
$I_1^{(\lambda)}$	1.000	1	0.000	$I_1^{(p)}$	0.252	0	0.434
$I_2^{(\lambda)}$	0.112	0	0.316	$I_2^{(p)}$	0.854	1	0.353
$I_3^{(\lambda)}$	0.998	1	0.044	$I_3^{(p)}$	0.636	1	0.481
$I_4^{(\lambda)}$	0.011	0	0.104	$I_4^{(p)}$	0.469	0	0.499
$I_5^{(\lambda)}$	1.000	1	0.000	$I_5^{(p)}$	0.417	0	0.493
$I_6^{(\lambda)}$	0.010	0	0.098	$I_6^{(p)}$	0.397	0	0.489
$I_7^{(\lambda)}$	1.000	1	0.000	$I_7^{(p)}$	0.531	1	0.499
$I_8^{(\lambda)}$	1.000	1	0.000	$I_8^{(p)}$	0.449	0	0.497
$I_9^{(\lambda)}$	0.014	0	0.118	$I_9^{(p)}$	0.439	0	0.496
$I_{10}^{(\lambda)}$	0.013	0	0.115	$I_{10}^{(p)}$	0.504	1	0.500
$I_{11}^{(\lambda)}$	0.011	0	0.103	$I_{11}^{(p)}$	0.204	0	0.403
$I_{12}^{(\lambda)}$	0.019	0	0.136	$I_{12}^{(p)}$	0.215	0	0.411
$I_{13}^{(\lambda)}$	0.999	1	0.026	$I_{13}^{(p)}$	0.209	0	0.407
$I_{14}^{(\lambda)}$	1.000	1	0.000	$I_{14}^{(p)}$	0.216	0	0.412

sign indicates an expected increase on the collision counts with an increase on the fixed effects variables. For example, the posterior mean for the *mainroad* variable, $\hat{\beta}_7 = 0.773$, indicates that intersections on "Main Road"experience approximately twice as many traffic collisions ($e^{0.773} = 2.166$) as intersections not located on "Main Road".

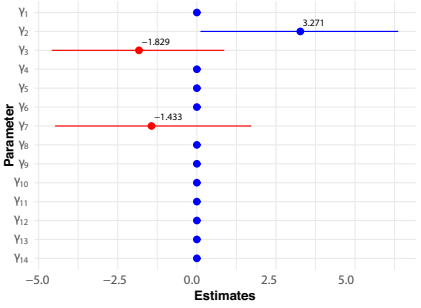
In the probability component of the model, only the *maxspeed* variable has a coefficient (γ_2) with a positive posterior mean. Two other variables included in the model, *trafficsignals*, and *mainroad* variables both have coefficients (γ_3 and γ_7, respectively) with negative posterior means. The coefficients represent the change in the log-odds of the outcome variable (p_{it}) - log-odds of observing a false zero collision - for a one-unit increase in the respective covariate, holding all other variables constant. Since all three variables are binary categorical variables, we can interpret that the positive coefficient represents an increase in observing false zero probability when that property exists at that intersection. Here, the coefficient for the speed variable ($\hat{\gamma}_2 = 3.271$), which depicts that when the intersection is on a high speed road, then the odds of a false zero probability is higher than an intersection on a road with less than 80 km/h speed limit. In particular, the finding that higher speed limits are associated with an increased probability of observing false zero collisions may initially seem contradictory to expectations. However, we interpret this in the context of the characteristics of the road network in ward 57. Roads with higher speed limits are typically major routes that are better engineered, have fewer intersections, and experience lower pedestrian activity. We believe these properties could reduce the number of conflict points. This



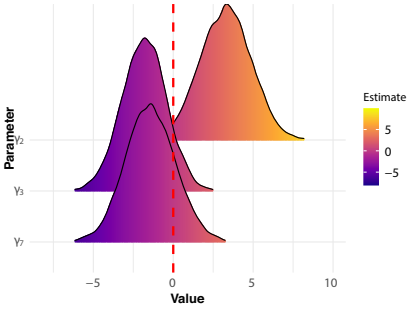
(a) Posterior mean and the 95% credible intervals of the fixed effects coefficients included in the count component. Numbers on the credible intervals are the posterior means.



(b) Density plots of the posterior distribution of the fixed effects coefficients included in the count component.



(c) Posterior mean and the 95% credible intervals of the fixed effects coefficients included in the probability component of the model. Numbers shown on the credible intervals are the posterior means.



(d) Density plots of the posterior distribution of the fixed effects coefficients included in the probability component of the model.

Fig. 7 Posterior mean, 95% credible intervals and the density distributions of the posterior distributions of the fixed effects included in Model 5. Vertical red line represents the zero posterior mean

could partly explain the observed association between higher speeds and a higher probability of false zero recorded collisions.

On the contrary, the posterior mean of the coefficients for the `trafficsignals` and `mainroad` variables are $\hat{\gamma}_3 = -1.829$ and $\hat{\gamma}_7 = -1.433$, respectively. These negative coefficients show that the existence of that road property at the intersection reduces the odds of a false zero by a factor of 0.161 (or by approximately 84%) for the *trafficsignals* and by a factor of 0.239 (or by approximately 76%). For example, if an intersection has one or more traffic signals, then the odds of a false zero is 84% less than an intersection with no traffic signals. Similarly, if an intersection is on "Main Road", then the odds of a false zero is 76% less than an intersection that is not on "Main Road".

Spatial Effects

It is often an interesting and common practice to evaluate the proportion of the variance explained by the structured spatial random effects compared to the unstructured spatial

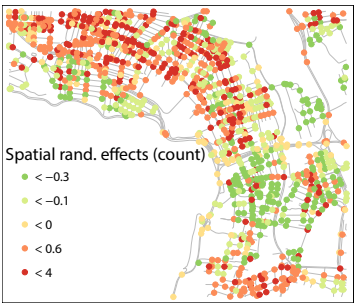
random effects in both the count ($frac_{u^{(\lambda)}}$) and probability ($frac_{u^{(p)}}$) components of the final model used for predictions. Many researchers have calculated and interpreted the proportion as follows (Blangiardo et al., 2013; Blangiardo and Cameletti, 2015):

$$\begin{aligned} frac_{u^{(\lambda)}} &= \frac{s^2_{u^{(\lambda)}}}{s^2_{u^{(\lambda)}} + \sigma^2_{v^{(\lambda)}}} \\ frac_{u^{(p)}} &= \frac{s^2_{u^{(p)}}}{s^2_{u^{(p)}} + \sigma^2_{v^{(p)}}} \end{aligned} \tag{6}$$

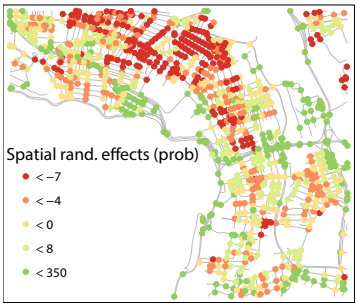
where $s^2_{u^{(\lambda)}}$ and $s^2_{u^{(p)}}$ are the estimates of the posterior marginal variances of the structured spatial random effects corresponding to the count and probability components. For our chosen model, $frac_{u^{(\lambda)}}$ is 0.626 and $frac_{u^{(p)}}$ is 0.993 indicating that the structured spatial random effects explain 62.6% of the total spatial variation in the count component, whereas almost all variation is explained by the structured spatial random effects in the probability component. The geographical distribution of these random effects is provided in Fig. 8. Figure 8a and b refer to the sum of the spatial random effects in the $\log(\lambda_{it})$ count component and $\text{logit}(p_{it}) = \log(\frac{p_{it}}{1-p_{it}})$ probability component of the model, respectively. These effects are derived by simply summing the posterior means of the spatial random effects.

Overall Model Evaluations

In order to assess how well the model predicts the observed collision counts, fitted collision counts are obtained using $\hat{Y}_{it} = (1 - \hat{p}_{it})\hat{\lambda}_{it}$. Here $\hat{\lambda}_{it}$ is the posterior mean of λ_{it} and $\hat{p}_{it}$ is the of posterior mean of p_{it}. The geographical distribution of the fitted collision counts ($\hat{Y}_{it}$) per year is given in Figure 9. These can be compared with the observed values mapped in Fig. 4. Comparison of these two sets of spatial plots reveal that the model predictions follow the observed traffic collision counts.

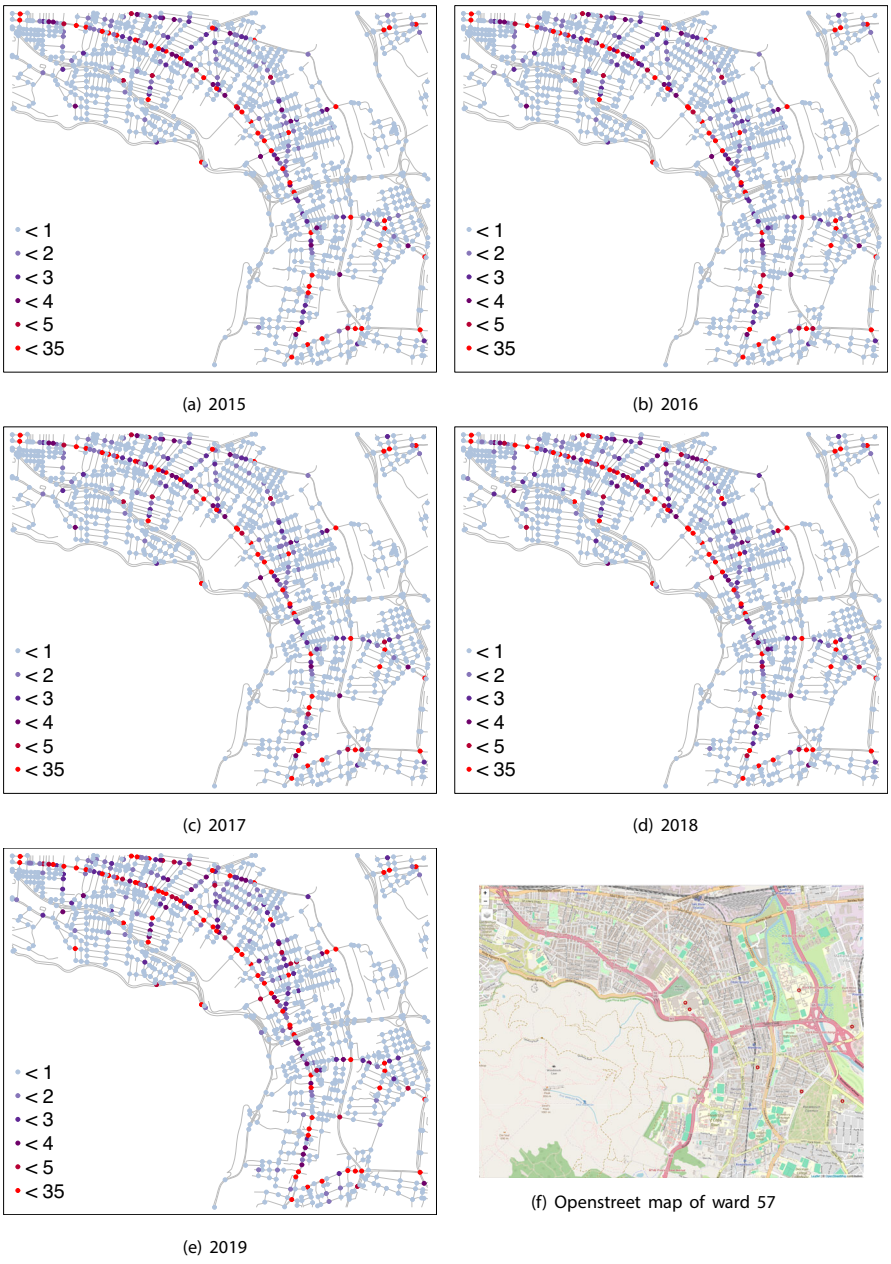


(a) Geographical distribution of spatial unstructured and structured random effects in the count component ($v_i^{(\lambda)} + u_i^{(\lambda)}$)



(b) Geographical distribution of spatial unstructured and structured random effects in the probability component ($v_i^{(p)} + u_i^{(p)}$)

Fig. 8 Geographical distribution of spatial unstructured and structured random effects in the count and probability components



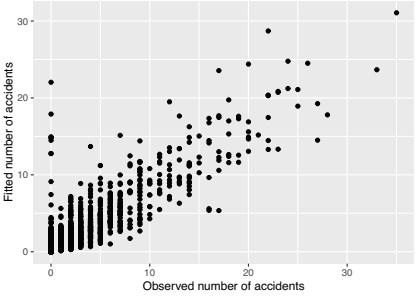
(a) 2015 (b) 2016

(c) 2017 (d) 2018

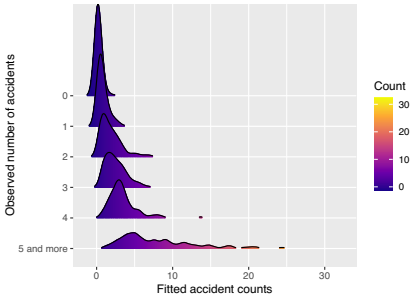
(e) 2019 (f) Openstreet map of ward 57

Fig. 9 Geographical distribution of fitted number of traffic collisions calculated using $\hat{Y}_{it} = (1 - \hat{p}_{it})\hat{\lambda}_{it}$ in ward 57, Cape Town, South Africa from spatio-temporal random effects model (Model 5). Each year predictions are provided in a separate plot

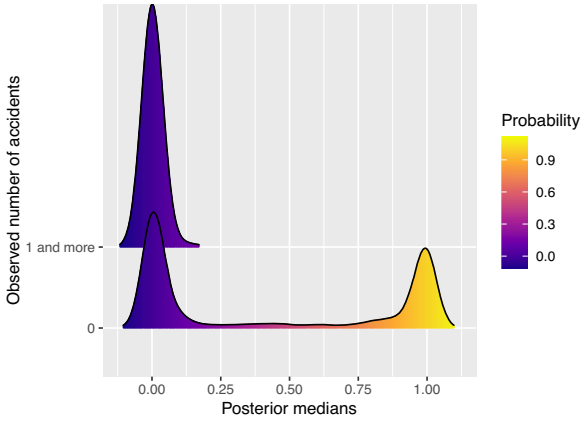
In addition, yearly aggregated posterior means for the traffic collision counts are plotted against the observed values in two different plots, Fig. 10a and b. Figure 10a shows the scatter plot of the fitted and the observed values. There is a clear agreement between the fitted and observed traffic collision counts (with Pearson and Spearman's rank correlation coefficients of $r = 0.894$, and $r_s = 0.682$, respectively, and a root mean square error estimate of $rmse = 1.495$). Moreover, the figure on the right, Fig. 10b, shows the density of the posterior means plotted with respect to the observed traffic collisions grouped into sub categories such as 0 collisions, 1 collision, 2 collisions, 3 collisions, 4 collisions and 5 and more collisions. In a model that fits the data well, we would expect to see that the density of the fitted number of collisions are distributed around their respective observed values as observed in Fig. 10b. It is worth noting that for higher observed traffic collision counts, the density distributions of the posterior means exhibit a right-skewed pattern. In the context of traffic collisions, it is better to over-predict than to under-predict since the main aim is to develop mech-



(a) Fitted ($\hat{Y}_{it}$) vs observed (Y_{it}) collision counts.



(b) Density plots of the fitted ($\hat{Y}_{it}$) collision counts with respect to observed collision counts (Y_{it} grouped as 0, 1, 2, 3, 4, 5 and more collisions).



(c) Density plots of the posterior median probability of zero collision with respect to observed collision counts (grouped as 0, and 1 or more collisions).

Fig. 10 Assessment of fitted ($\hat{Y}_{it}$) vs observed (Y_{it}) collision counts, and fitted false zero probabilities with respect to observed collision counts at intersections in ward 57, Cape Town, South Africa

anisms that would lead to reducing the total number of traffic collisions. In a similar manner, density plots of the posterior median probability of observing a zero collision versus one or more collisions is given in Fig. 10c.

We also examine the exceedance probabilities computed as $P((1 - p_{it})\lambda_{it} > \phi)$, where ϕ is a threshold value specified by the analyst in the context of the research problem. This is an approach commonly used by researchers to assess the risk maps, especially in epidemiological studies (Liu et al., 2013; Ye and Moreno-Madriñán, 2020; Madden et al., 2021; Yan et al., 2020). We explore the exceedance probabilities for a threshold value of two (Fig. 11a) and five (Fig. 11b) traffic collisions which are useful to detect potential high collision risk areas. The darker colours represent an increase in the probability of an intersection having more than the specified threshold value. The maps show that the risk of observing more collisions than the specified threshold is concentrated around mainly "Main Road". It is also not surprising to see certain areas such as around Rondebosch Common, one of the higher risk areas in this ward.

Finally, we carried out a cross-validation analysis to assess the out-of-sample predictive quality of the model. Considering that the models have been fitted within a Bayesian framework, we employed conditional predictive ordering (CPO). The CPO is a Bayesian diagnostic for detecting surprising observations according to the fitted model (Pettit, 1990). For observation y_{it}, it is defined as

$$\text{CPO}_{it} = f(y_{it}|y_{(it)}),$$

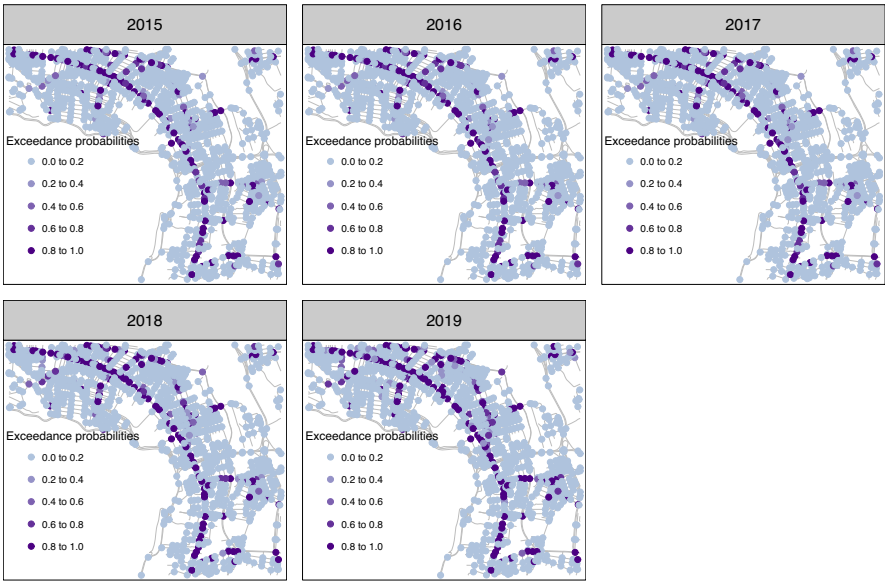
where $f(\cdot|y_{(it)})$ is the predictive distribution of a new observation given the data after deleting observation y_{it}, which are denoted by $y_{(it)}$. For this reason, this diagnostic can be considered as a type of leave-one-out cross-validation procedure. Small values (close to 0) of CPO_{it} indicate that observation y_{it} is surprising according to prior knowledge and the rest of the observations, whereas large values of CPO_{it} indicate the opposite.

The computation of CPO_{it} does not require refitting of the model without the inclusion of observation y_{it}, which would result in a massive computational cost if done for each of the observations. In particular, the following approximation can be used with the available MCMC sampled values (Gelfand and Dey, 1994)

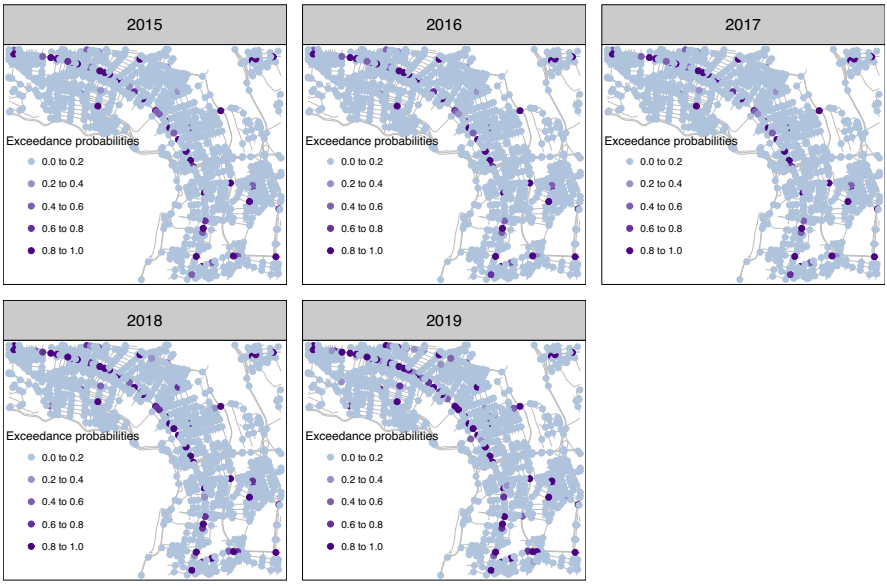
$$\widehat{\text{CPO}}_{it} = \left(\frac{1}{N} \sum_{m=1}^{N} \frac{1}{f(y_{it}|\lambda_{it}^{(m)}, \theta^{(m)}, p_{it}^{(m)})} \right)^{-1},$$

where $f(\cdot|\lambda, \theta, p)$ denotes the probability function of the ZINB distribution, superscript (m) the m-th value of the MCMC chain for the corresponding parameter, and N the length of the chain.

The results obtained from the application of this approach with our final model is summarized in Fig. 12. The distribution of the $\widehat{\text{CPO}}_{it}$ values (Fig. 12a) indicates that most observations are unsurprising for the model fitted, with most values concentrated near one. However, there is also a proportion of observations that exhibit a $\widehat{\text{CPO}}_{it}$ value close to zero. Figure 12b displays the $\widehat{\text{CPO}}_{it}$ for each observation. This kind of plot



(a) Exceedance probabilities at a threshold of φ = 2



(b) Exceedance probabilities at a threshold of φ = 5

Fig. 11 Exceedance probabilities computed as $P((1 - p_{it})\lambda_{it} > \phi)$ for intersections per year in ward 57, Cape Town, South Africa with darker colours representing probabilities between 0.6 and 1

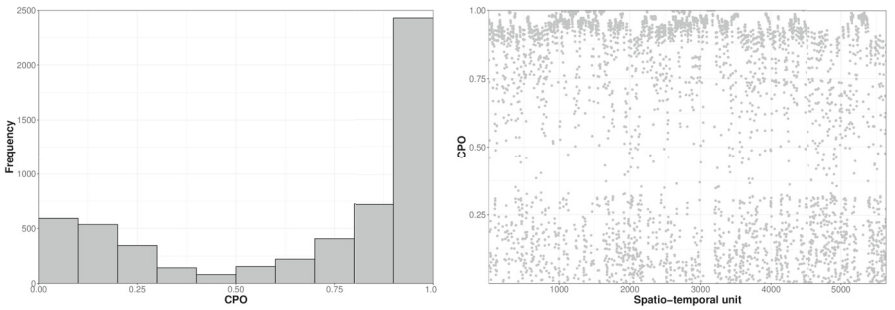


Fig. 12 Distribution of $\widehat{\text{CPO}}_{it}$ values provided in (a) and $\widehat{\text{CPO}}_{it}$ values ordered by observation number provided in (b)

allows checking in this case if the presence of surprising observations varies over time. In general, no clear pattern is observed. Further inspection of the observations with small $\widehat{\text{CPO}}_{it}$ values can be helpful for identifying typologies of intersections that might receive consistently poor predictions from the fitted model.

Conclusions and Future Research

This paper presents an approach to modelling road traffic collisions at intersections within one of the busiest wards of the CoCT, South Africa for the period of 2015-2019. The address information from the collision dataset provided by the CoCT Transport Department was converted into a georeferenced dataset and then aggregated to generate a count response variable at intersections. Traffic collision counts at intersections are then modelled with spatio-temporal zero-inflated Negative Binomial models using a Bayesian approach.

The original data obtained directly from the CoCT's Transport Department may contain some degree of inaccuracy, particularly for collisions not occurring at intersections. This is one of the main reasons we used the collisions at intersections since for those reporting the collisions, it is easier to describe the intersection with at least two street names. We used the `arcgisgeocode` package (Josiah, 2024) to geocode the reported locations based on street names and intersection descriptors such as the **X** used by the reporter and compared the results with another geocoding package. While we found that `arcgisgeocode` (Josiah, 2024) allowed for consistent spatial referencing, we acknowledge the potential for geocoding errors or biases (given that there are some intersections with 70% geocoding reliability score). This could potentially be the case where multiple similarly named streets may exist. Future work would benefit from improved geolocation practices, such as direct GPS capture at the point of collision reporting and incorporating the uncertainty about the exact location through a Bayesian modelling approach as in Briz-Redón (2024).

Two intersections in ward 57 namely, the Offramp on Settlers Way and Onramp on Berkley Road were identified as statistical outliers. Their inclusion in the models

led to significant overdispersion and influenced model estimates. To maintain model stability and focus on patterns generalisable across typical intersections, we excluded these points. However, we acknowledge that these high incident intersections may be of interest for case specific intervention studies, and therefore we recommend targeted future investigation into their unique characteristics. Better monitoring and law enforcement practices should be considered to improve road safety at these two intersections given the current high risk status.

After the data manipulation stage, we focus only on the intersection related covariates and sequentially built five different models to assess the importance of the inclusion of random effects as well as the fixed effects. According to the WAIC values of the different models, the final model chosen is the model that includes the spatial and space-time random effects with some of the fixed effects identified from the Bayesian variable selection. To the best of our knowledge, this study is the first to analyse traffic collision data in the CoCT using Bayesian methods and overall, it could be a starting point for collaboration with the CoCT Transport Department to improve safety on our roads.

The results indicate that among the covariates included in the final model, factors such as node degree, the presence of traffic signals, road classification as having any major road around the intersection, the intersection being on "Main Road", taxi routes at intersections, and dummy variables for the years 2018 and 2019 all contribute to an increase in traffic collision counts at intersections. Given these results, we believe "Main Road"could benefit from stricter speed enforcement, regular patrolling and monitoring, as well as use of traffic fines to deter violations, particularly near intersections. In the probability component of the model, the presence of traffic signals and the intersection being located on a "Main Road"are associated with negative coefficients, suggesting that these factors reduce the likelihood of a false zero collision. Conversely, higher speed is associated with an increased likelihood of a false zero. This may be due to the fact that higher speed roads in the chosen area are typically larger, highway type roads with fewer intersections and conflict points. These findings provide valuable insights for policymakers to implement necessary interventions to enhance road safety.

The study presents several opportunities for improvement. Instead of only focusing on the intersections, incorporating the road segments could be beneficial. However, this might result in data manipulation issues since the dataset obtained from the city contains only the road names where the collision occurs. Since some roads in the CoCT are quite long, it becomes difficult to assign the collision to a particular road segment. The need for more reliable data collection methods have become apparent during this study. One actionable policy recommendation is to advocate for digital recording and automated geocoding of collisions, which could support city wide adoption of spatial network analysis. The authors are excited to see that the city is already considering the development of an app that drivers can use in future for recording the collisions. This will help with the collection of precise geolocations of the collisions and will improve the process researchers need to go through to access the traffic collision data.

In addition, the study can be improved by the incorporation of reliable average annual daily traffic flow data, which was unavailable to the authors at the time of the study. As a result, node degree was chosen as a traffic flow proxy. The findings suggest

that intersections with higher traffic levels experience increased number of traffic collisions. This limitation could potentially be addressed by leveraging platforms such as Google or TomTom to access historical traffic flow data.

This study focuses on ward 57 in the CoCT over the period 2015-2019. While the results provide important insights into spatial and infrastructural correlates of road traffic collisions within this ward, it is important to acknowledge the limitations in generalising these findings to the entire city. As noted by Berrie et al. (2019), analyses based on selective subregions may capture localized dynamics that do not necessarily hold elsewhere. Ward 57 includes a diverse mix of intersections making it a valuable case study. However, some relationships found relevant for this ward may differ in wards with different socioeconomic characteristics. As such, while the patterns observed in this ward are likely to be broadly informative, further analysis across multiple wards using multilevel or hierarchical models would be necessary to explore the differences between wards, and to assess the consistency and generalisability of the identified relationships. By doing this, the city can focus on the riskiest wards to reduce the number of traffic collisions.

Moreover, the traffic collision data contains the type of the traffic collision classified as fatal, severe injury, slight injury and no injury; the time of the collision; vehicles involved and several other variables. Multivariate models are useful for comparing the different collision types and identifying the different mechanisms that generate the types of collisions. Furthermore, through our discussions with the city, interest in knowing the factors affecting the time between the occurence of collisions at intersections was expressed. This is a promising area of research, as, to the best of our knowledge, such analysis has not been explored in the CoCT.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12061-025-09703-0.

Acknowledgements We would like to thank City of Cape Town Data Science and Transport Departments, especially to Delyno du Toit and Ilse Davidson for helping with the data access.

Author Contributions S.E. was responsible for conceptualization of the models, data manipulation and curation, georeferencing the accident data, presenting the results, and writing the entire research paper. A.B.R. conducted the Bayesian data analysis, conceptualized the models, and contributed to results representation and proof reading. S.S. contributed to idea conceptualization, data cleaning and manipulation, proof reading and building collaborations with the City of Cape Town. R.L. assisted with idea conceptualization, data cleaning, proof reading and setting up the GitHub workflow.

Funding Open access funding provided by University of Cape Town. This publication is based on research that has been supported in part by xxx.

Data Availability The data that are used in this research are available from City of Cape Town Transport Department, but restrictions apply to the availability of these data, and so are not publicly available. Though a research request allows the users to access the data.

Declarations

Competing interests The authors declare no competing interests.

References

Baddeley, A., Rubak, E., & Turner, R. (2015). Spatial Point Patterns: Methodology and Applications with R. CRC Press.

Baddeley, A., Nair, G., Rakshit, S., McSwiggan, G., & Davies, T. M. (2021). Analysing point patterns on networks—a review. *Spatial Statistics, 42*, 100435.

Berrie, L., Ellison, G. T., Norman, P. D., Baxter, P. D., Feltbower, R. G., Tennant, P. W., & Gilthorpe, M. S. (2019). The association between childhood leukemia and population mixing: an artifact of focusing on clusters? *Epidemiology, 30*, 75–82.

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

Blangiardo, M., & Cameletti, M. (2015). *Spatial and Spatio-Temporal Bayesian Models with R-INLA*. John Wiley & Sons.

Blangiardo, M., Cameletti, M., Baio, G., & Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. *Spatial and Spatio-Temporal Epidemiology, 4*, 33–49.

Borruso, G. (2005). Network density estimation: Analysis of point patterns over a network. In *International Conference on Computational Science and Its Applications* (pp. 126–132). Springer.

Borruso, G. (2008). Network density estimation: A GIS approach for analysing point patterns in a network space. *Transactions in GIS, 12*, 377–402.

Briz-Redón, A. (2024). Dealing with location uncertainty for modeling network-constrained lattice data. *Spatial Statistics, 59*, 100807.

Briz-Redón, A., Martínez-Ruiz, F., & Montes, F. (2019). Estimating the occurrence of traffic accidents near school locations: A case study from Valencia (Spain) including several approaches. *Accident Analysis & Prevention, 132*, 105237.

Briz-Redón, A., Martínez-Ruiz, F., & Montes, F. (2019). Spatial analysis of traffic accidents near and between road intersections in a directed linear network. *Accident Analysis & Prevention, 132*, 105252.

Briz-Redón, A., Mateu, J., & Montes, F. (2021). Modeling accident risk at the road level through zero-inflated Negative Binomial models: A case study of multiple road networks. *Spatial Statistics, 43*, 100503.

Chaudhuri, S., Saez, M., Varga, D., & Juan, P. (2023). Spatiotemporal modeling of traffic risk mapping: A study of urban road networks in Barcelona, Spain. *Spatial Statistics, 53*, 100722.

City of Cape Town (2023). EPWP assists with road accident data capturing. https://www.capetown.gov.za/Media-and-news/EPWP%20assists%20with%20road%20accident%20data%20capturing. Accessed 22 June 2025.

Csardi, M. G. (2013). Package igraph. Last accessed, *3*, 2013.

da Silva, A. R., & de Sousa, M. D. R. (2023). Geographically weighted zero-inflated Negative Binomial regression: A general case for count data. *Spatial Statistics, 58*, 100790.

Das, S. (2014). Pedestrian fatality and natural light: Evidence from South Africa using a Bayesian approach. *Economic Modelling, 38*, 311–315.

de Valpine, P., Turek, D., Paciorek, C., Anderson-Bergman, C., Lang, D. T., & Bodik, R. (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics, 26*, 403–413.

Dellaportas, P., Forster, J. J., & Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing, 12*, 27–36.

Department of Transport, R. (2018). State of road safety report: Calendar january-december 2018. Available at https://www.rtmc.co.za/index.php/publications/reports/traffic-reports (2021/07/19).

Erdogan, S. (2009). Explorative spatial analysis of traffic accident statistics and road mortality among the provinces of Turkey. *Journal of Safety Research, 40*, 341–351.

Gelfand, A. E., & Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological), 56*, 501–514.

Gilardi, A., Mateu, J., Borgoni, R., & Lovelace, R. (2022). Multivariate hierarchical analysis of car crashes data considering a spatial network lattice. *Journal of the Royal Statistical Society Series A: Statistics in Society, 185*, 1150–1177.

Gu, X., Yan, X., Ma, L., & Liu, X. (2020). Modeling the service-route-based crash frequency by a spatiotemporal-random-effect zero-inflated Negative Binomial model: An empirical analysis for bus-involved crashes. *Accident Analysis & Prevention, 144*, 105674.

Josiah, P. (2024). A robust interface to ArcGIS geocoding services. https://github.com/r-arcgis/arcgisgeocode.

Kaygisiz, Ö., & Hauger, G. (2017). Network-based point pattern analysis of bicycle accidents to improve cyclist safety. *Transportation Research Record, 2659*, 106–116.

Kuo, L., & Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, 65–81.

Lee, J., Abdel-Aty, M., & Cai, Q. (2017). Intersection crash prediction modeling with macro-level data from various geographic units. *Accident Analysis & Prevention, 102*, 213–226.

Levine, N., Kim, K. E., & Nitz, L. H. (1995). Spatial analysis of Honolulu motor vehicle crashes: I. spatial patterns. *Accident Analysis & Prevention, 27*, 663–674.

Liu, B., Siu, Y. L., Mitchell, G., & Xu, W. (2013). Exceedance probability of multiple natural hazards: Risk assessment in China's Yangtze river delta. *Natural Hazards, 69*, 2039–2055.

Liu, F., Zheng, L., Li, M., & Tang, J. (2022). Analysis and prediction of the interval duration between the first and second accidents considering the spatiotemporal threshold. *Journal of Advanced Transportation, 2022*, 6312139.

Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice, 44*, 291–305.

Madden, J. M., McGrath, G., Sweeney, J., Murray, G., Tratalos, J. A., & More, S. J. (2021). Spatio-temporal models of bovine tuberculosis in the Irish cattle population, 2012–2019. *Spatial and Spatio-temporal Epidemiology, 39*, 100441.

Mannering, F. L., Shankar, V., & Bhat, C. R. (2016). Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic Methods in Accident Research, 11*, 1–16.

Marchant, P. R., & Norman, P. D. (2022). To determine if changing to white light street lamps improves road safety: A multilevel longitudinal analysis of road traffic collisions during the relighting of leeds, a uk city. *Applied spatial analysis and policy, 15*, 1583–1608.

Martínez-Beneito, M. A., & Botella-Rocamora, P. (2019). Disease mapping: from foundations to multidimensional modeling. Chapman and Hall/CRC.

McSwiggan, G., Baddeley, A., & Nair, G. (2017). Kernel density estimation on a linear network. *Scandinavian Journal of Statistics, 44*, 324–345.

Nissen, C. (2014). An r package that wraps the nokia here geocoding api. https://github.com/corynissen/geocodeHERE.

Okabe, A., & Sugihara, K. (2012). Spatial Analysis along Networks: Statistical and Computational Methods. John Wiley & Sons.

Okabe, A., Satoh, T., & Sugihara, K. (2009). A kernel density estimation method for networks, its computational method and a GIS-Based tool. *International Journal of Geographical Information Science, 23*, 7–32.

Padgham, M., Lovelace, R., Salmon, M., & Rudis, B. (2017). osmdata. *Journal of Open Source Software, 2*.

Pebesma, E. J., et al. (2018). Simple features for R: standardized support for spatial vector data. *R J., 10*, 439.

Pettit, L. (1990). The conditional predictive ordinate for the normal distribution. *Journal of the Royal Statistical Society: Series B (Methodological), 52*, 175–184.

Produit, T., Lachance-Bernard, N., Strano, E., Porta, S., & Joost, S. (2010). A network based kernel density estimator applied to Barcelona economic activities. In Taniar, D., Gervasi, O., Murgante, B., Pardede, E., & Apduhan, B. O. (Eds.) *Computational Science and Its Applications - ICCSA 2010, International Conference, Fukuoka, Japan, March 23-26, 2010, Proceedings, Part I, Springer* (pp. 32–45). https://doi.org/10.1007/978-3-642-12156-2_3,

R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Sosa, J., Briz-Redón, A., Flores, M., Abril, M., & Mateu, J. (2023). A spatio-temporal multinomial model of firearm death in Ecuador. *Spatial Statistics, 54*, 100738.

StatsSA, R. (2022). Census 2022. Available at https://census.statssa.gov.za (2024/12/06).

StatsSA (2011). Statistics by place 2011. Available at http://www.statssa.gov.za/?page_id=993&id=city-of-cape-town-municipality. (2021/08/27).

Tang, Y., Knodler, M. A., & Park, M. H. (2013). A comparative study of the application of the standard kernel density estimation and network kernel density estimation in crash hotspot identification. In *16th Road Safety on Four Continents Conference*.

Tennekes, M. (2018). tmap: Thematic maps in R. *Journal of Statistical Software, 84*, 1–39.

UN (2017). Road safety considerations in support of the 2030 agenda for sustainable development. Available at https://unctad.org/system/files/official-document/dtltlb2017d4_en.pdf (2021/08/27).

Wang, W., Yang, Y., Yang, X., Gayah, V. V., Wang, Y., Tang, J., & Yuan, Z. (2024). A Negative Binomial Lindley approach considering spatiotemporal effects for modeling traffic crash frequency with excess zeros. *Accident Analysis & Prevention, 207*, 107741.

Watanabe, S., & Opper, M. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research, 11*, 3571–3594.

Wickham, H. (2011). ggplot2. *Wiley interdisciplinary reviews: computational statistics, 3*, 180–185.

Xie, Z., & Yan, J. (2008). Kernel density estimation of traffic accidents in a network space. *Computers, Environment and Urban Systems, 32*, 396–406.

Xie, Z., & Yan, J. (2013). Detecting traffic accident clusters with network kernel density estimation and local spatial statistics: an integrated approach. *Journal of Transport Geography, 34*, 64–71.

Yan, Y., Zhang, Y., Yang, X., Hu, J., Tang, J., & Guo, Z. (2020). Crash prediction based on random effect Negative Binomial model considering data heterogeneity. *Physica A: Statistical Mechanics and Its Applications, 547*, 123858.

Ye, J., & Moreno-Madriñán, M. J. (2020). Comparing different spatio-temporal modeling methods in dengue fever data analysis in Colombia during 2012–2015. *Spatial and Spatio-temporal Epidemiology, 34*, 100360.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., Smith, G. M., et al. (2009). Mixed Effects Models and Extensions in Ecology with R. vol. 574. Springer.