Mapping road traffic crash hotspots using GIS-based methods: A case study of Muscat Governorate in the Sultanate of Oman

Amira K. Al-Aamri,[1] Graeme Hornby,[2] Li-Chun Zhang,[3,4] Abdullah A. Al-Maniri,[5] Sabu S. Padmadas,[3,6]

[1] Ministry of Higher Education, Sultanate of Oman
[2] Geography and Environmental Science, University of Southampton, UK
[3]Department of Social Statistics and Demography, University of Southampton, UK
[4] Southampton Statistical Sciences Research Institute, University of Southampton, UK
[5] Studies and Research Department, Oman Medical Specialty Board, Sultanate of Oman
[6] Centre for Global Health, Population, Poverty and Policy, University of Southampton, UK

ABSTRACT

Objective:

Road traffic crashes (RTCs) are a major global public health problem and cause substantial burden on national economy and healthcare. There is little systematic understanding of the geography of RTCs and the spatial correlations of RTCs in the Middle-East region, particularly in Oman where RTCs are the leading cause of disability-adjusted life years lost. The overarching goal of this paper is to evaluate the spatial and temporal dimensions, identifying the high risk areas or hot-zones where RTCs are more frequent using the geocoded data from the Muscat governorate.

Data:

This study is based on data drawn from the Royal Oman Police (ROP) sample iMAAP database and the National Road Traffic Crash (NRTC) database which is managed by the ROP and made available for research use by The Research Council of the Sultanate of Oman. The data covered the period from 1st January 2010 to 2nd November 2014. Only RTCs occurred in Muscat Governorate were included in the study. The study is based on 12,438 registered incidents, however, due to disconnections found on road network, RTCs occurred on disconnected parts were removed and the final analysis considered only 9,357 incidents.

Methods:

The analysis considered an adjacency network analysis integrating GIS and RTC data using robust estimation techniques including: Kernel Density Estimation (KDE) of both 1-D and 2-D space dimensions, Network-based Nearest Neighbour Distance (Net-NND), Network-based K-Function, Random Forest Algorithm (RF) and spatiotemporal Hot-zone analysis.

Findings:

The analysis highlight evidence of spatial clustering and recurrence of RTC hot-zones on long roads demarcated by intersections and roundabouts in Muscat. The findings confirm that road intersections elevate the risk of RTCs than other effects attributed to road geometry features. The results from GIS application of NRTC data are validated using the sample data generated by iMAAP database.

Conclusion:

The findings of this study provide statistical evidence and confirm the research hypothesis that that road intersections (roundabouts, crosses and bridges) represent higher risk of causing RTCs than other road geometric features. The results also demonstrate systematic quantitative evidence of spatio-temporal patterns associated with the crash risk over different locations on road network in Muscat. More importantly, the findings clearly pinpoint the importance and influence of the road and traffic related feature in road crash spatial analysis.

Keywords: Road traffic crashes; Road geometry features; Clustering; RTC hot-zones; spatiotemporal modelling; Kernel Density estimation

1. Introduction

Identifying the location and time of Road Traffic Crashes (RTCs) is crucial for the enforcement authorities to take effective measures to reduce the risk of RTCs (Yu et al., 2014; Benedek et al., 2016; Khanh, Pei and Liang-Tay 2019). The heterogeneity in RTC frequencies and rates is attributed to complex roadside features, traffic and weather conditions and driving behaviors (Cheng and Washington, 2005; Harirforoush, Bellalite, and Bénié, 2019). Different terminologies have been used to describe high-risk RTC locations; hazardous road locations, high-risk locations, accident-prone locations, black spots, hot spots, hot zones, black zones, sites with promise and priority investigation locations (Montella, 2010; Choudhary et al., 2015; Yao et al.,2018). Past studies have no universally standard definition for hazardous road locations, which suggests that there is no clear definition or consensus of identifying crash locations (Elvik, 2008; Anderson, 2009; Choudhary et al., 2015). The major challenge, therefore, is to make judgements on the definitions and criteria for determining RTC hotspots (Miranda-Moreno et al., 2007; Elvik, 2008; Anderson, 2009).

There is little systematic understanding of the spatial patterns and correlations of RTCs in the Middle-East region, particularly in Oman, where RTCs are the leading cause of disability-adjusted life years lost. The goal of this study is to identify the locations of hot-zones (groups of neighbouring hotspots) and spatial clustering of RTCs in the Muscat governorate. Muscat is the capital of Oman and the most densely populated (345 people per km^2) governorate in the country, covering more than 32% of the total population (NCSI, 2016; NCSI, 2018). Muscat is located in the north-eastern part of of Oman, it represents a mix of ancient cultural heritage and modern style and it is considered as the heart of the Sultanate. Spurred by rapid economic growth and urbanisation, the use of private vehicles to commute to both short and long distance to workplace, shopping and leisure centres are becoming increasingly common in Oman, especially commuting from adjoining governorates to Muscat. This has led to an increase in the concentration of daily commuting within limited major roads, which in turn has resulted in a high level of traffic congestion coupled with a high rate of traffic crashes (annually, more than 33% of RTCs in Oman occurred in Muscat) (Al-Rawas, 1993; Royal Oman Police, 2017).

Network Kernel Density (Net-KDE) estimation technique can be applied to develop an adjacency network analysis by focusing on the spatial and temporal dimensions, which is useful in identifying the high risk or hot-zone areas where RTCs are more frequent. It also identifies the significant factors affecting these spatial patterns. The identification of the so-called hot-zones or high risk areas would help transportation safety professionals and authorities to identify high-crash corridors more efficiently so that they can develop safety strategies on these hazardous locations (Harirforoush, Bellalite, and Bénié, 2019). Consequently, these hot-zones would have a priority to benefit from a systemic safety improvement program including suitable road design, proper traffic control, and effective enforcement of traffic rules (Young and Park, 2014; Achu et al., 2019).

The present research addresses the following questions:

1. Where are the high risk or hot-zone areas for road crashes in Muscat Governorate where crashes are more frequent?

2. How can we use a GIS-based spatial analysis to understand and model the patterns of road crashes integrating relevant predictors such as road geometry and traffic related features?

Additionally, it addresses two sub-questions:

a. What factors characterise the hot-zones from normal- and cold-zones?

b. Over time, which road zones represent high risk areas for road traffic crashes in Muscat?

We **hypothesise** that road intersections (roundabouts, crosses and bridges) elevate the risks to RTCs than other road geometric features.

2. Data and Methods

2.1 Data

From a statistical perspective, data of a minimum of three years are needed for any spatial analysis to obtain credible results (Benedeka et al., 2016; Yao et al.,2018). This study is based on data drawn from the National Road Traffic Crash (NRTC) database and sample iMAAP database managed by the the Royal Omani Police (ROP), and made available for research use by The Research Council of the Sultanate of Oman. iMAAP is implemented by ROP and supported by The Research Council under the National Road Safety Research programme. The data covered the period from 1st January 2010 to 2nd November 2014. Only RTCs occurred in Muscat Governorate were included in the study. The study is based on 12,438 registered incidents, however, due to disconnections found on road network, RTCs occurred on disconnected parts were removed and only 9,357 were considered in the final analysis.

The database includes information about crash date, time, sex, age and nationality of drivers, type of injuries, fatalities, type and number of vehicles involved, cause of crash, type of collision, location, type of road, weather conditions, and crash description. So the coordinates of the crashes are not available, and hence these geographical coordinates were generated by the research team. In order to generate these data, the researchers used the information given in the crash description field and Google-Maps to specify the location of the crash. The geographical coordinates were identified for each case and the researcher spent four months to generate these for the Muscat governorate. However, it is important to mention that in most cases the crash description field did not specify the direction of the road on which the crash occurred, so RTCs were geocoded to the nearest point on the road network where the road mark was located. Further refinement of the road marks was based on researchers' familiarity with the region.

Data on Muscat road network (ArcGIS shapefile format) was downloaded from the Open Street Map using ArcGIS 10.2, which included details about type of the road, road name, speed limit and the length of the road. In addition, the researchers used Google-Maps to specify the level of traffic for different segments along Muscat road network. The results from GIS application of data generated from NRTC database were validated using the pilot data generated by iMAAP network based crash analysis system developed by the UK Transport Research Laboratory. ROP in collaboration with The Research Council undertook a pilot project to establish the feasibility of iMAAP in the Sultanate.

2.2 Methods

Different methodologies have been used for the delineation of RTC hotspots (Benedek et al., 2016). The traditional statistical methods and spatial analysis using Geospatial Information Systems (GIS)-based technique are examples of the methods used for such analysis (Deshpande et al., 2011). Estimations using crash frequency, crash rate, crash density, and crash severity index are examples of simple methods that can be applied to identify RTC hotspots (Yu et al., 2014; Choudharya et al., 2015). Comparison of crash counts at different locations and ranking of these locations based on the severity of crashes has been used in the traditional methods of crash hotspots detection (Anderson, 2006). Empirical Bayes (EB) statistical method has been proven as one of the most accepted statistical methods for crash hotspot identification (Deshpande et al., 2011). EB method was found to be more superior when compared to other hotspot identification methods (Choudharya et al., 2015). However, since EB requires special statistical skills, many transportation departments still use simple methods (Choudharya et al., 2015). Although different statistical methods have been widely used for delineation of hotspots, however, combining them with spatial analysis tools will produce more efficient results RTC hotspots analysis (Deshpande et al., 2011).

Using GIS-based technique is key to estimating the crash risk at different locations and times and it helps to provide information about various risk factors and explaining variations in crash involvement rate and injury severity (Anderson, 2009; Deshpande et al., 2011; Pleerux, 2020). Such systems provide a platform to display a number of visualized geographical outputs to deeply understand the dynamics of RTCs (Mahmud et al., 1998; Deshpande et al., 2011). GIS has enabled road professionals to use sophisticated spatial-statistical methods to detect hotspots (Anderson, 2006). It enables the efficient manipulation, analysis and visualization of spatial data so that a more robust understanding can be gained by providing indications of the casual effects of RTCs (Anderson, 2009; Pleerux, 2020). Past studies used GIS for different purposes: producing maps combining different parameters of RTCs such as number of crashes, number of injuries and death, road traffic related data and demographic factors and presenting the results in a user friendly way (e.g. Gundogdu, 2010; Truong and Somenahalli, 2011; Qin et al., 2013); conducting spatial analysis to identify high risk sites (e.g. Aguero-Valverde and Jovanis, 2006; Erdogan, 2009; Vandenbulcke et al., 2014; Deshpande et al., 2011; Ivan et al., 2015; Rahman et al., 2017; Vandenbulcke et al., 2017; Harirforoush, Bellalite, and Bénié, 2019; Yao et al., 2018; Khanh, Pei, and Liang-Tay 2019); and using it to apply spatio-temporal analysis (Plug et al., 2011; Prasannakumara et al., 2011; Ivan and Haidu, 2012; Achu et al.,2019; Kaygisiz and Sümer 2019; Pleerux, 2020).

Kernel density estimation (KDE) has been widely used to analyse crime data and only recently it has been adopted to detect the spatial pattern of crash data (e.g. Anderson, 2006; Anderson, 2009;

Mohaymany et al., 2013; Kaygisiz et al., 2015; Hashimoto et al., 2016; Rahman et al., 2017; Achu et al.,2019; Harirforoush, Bellalite, and Bénié, 2019; Khanh, Pei, and Liang-Tay 2019). However, since crimes can occur anywhere (e.g. street, house, park, shopping places) while RTCs are constrained to the road network (Anderson, 2006), crime hotspots detection methods should be modified to account for this difference. There are two forms of kernel density estimation: Planar Kernel Density Estimation (PKDE) and Network Kernel Density Estimation (Net-KDE) (Loo and Anderson, 2015; Yao et al., 2018). PKDE is used to identify hotspots of point-events and it calculates the density within 2D-space using the Euclidean distance (Loo and Anderson, 2015; Yao et al., 2018). PKDE is rarely used in the study of RTCs because two points which are close in terms of Euclidian can be far away when considering the distance between them on road network, so PKDE can be misleading in this case (Mohaymany et al., 2013; Kaygisiz et al.,2015; Loo and Anderson, 2015; Benedek et al., 2016). In 2003, Flauhaut et al. proposed a method estimate density along single road segment. However, because this method was not appropriate for calculating density along road network as a whole, Xie and Yan (2008) developed Net-KDE method and when they compared the results of Net-KDE with PKDE, they found that PKDE overestimates the crash density. Therefore, because of the difference of where crimes and RTCs occur, PKDE is more appropriate for crime hotspots detection, while Net-KDE is recommended to be used to calculate density of events such as RTCs occurring along network (Xie and Yan, 2008; Yao et al., 2018). Okabe et al. (2006) developed a toolkit called SANET (Spatial Analysis along Network) which integrates with ArcGIS software, and used specifically for network spatial analysis. For the technical details of SANET, the reader is referred Okabe and Sugihara (2012).

The following section summarises the statistical methods used for the delineation of RTC hot-zones in this study.

2.2.1 Spatial analysis along network (SANET)

In this study, SANET toolkit installed within ArcGIS 10.2/ArcMap was used to implement network spatial analysis of RTCs. The formula of calculating network kernel density (Net-KDE) using SANET is the same as the formula given by Xie and Yan (2008), but instead of using Gaussian or Quartic equations to calculate the kernel function (k), SANET used following equation (Okabe at al., 2009):

$$K_y(x) = \left\{ \begin{array}{ll} k(x), & for - h \leq x \leq 2d - h \\ k(x) - \frac{n-2}{n} k(2d - x), & for\ 2d - h \leq x \leq d \\ \frac{2}{n} k(x), & for\ d \leq x \leq h \end{array} \right\} \quad (3)$$

where $k(x)$: the basic kernel function,

y: the centre of the kernel,

x: a point on the network (in this study it is a road traffic crash),

h: the bandwidth (in meters),

n: the degree of a given node (v) (see Figure 5.1),

d: the shortest network distance from y to x.

The density of RTCs on a given road segment is calculated as follows (Okabe at al., 2009):

$$D(o) = \int_{-h}^{2d-h} k(-y)dy + \int_{2d-h}^{d} \left[k(-y) - \frac{n-2}{n} k(2d+y)\right] dy + \int_{d}^{h} \frac{2}{n} k(-y)dy \qquad (4)$$

where $D(o)$: the density at the origin.

The graphical expression of the three ranks of kernel function is shown in ***Figure 2***, while the simplified graphical expression of network kernel function is given in ***Figure 1***.

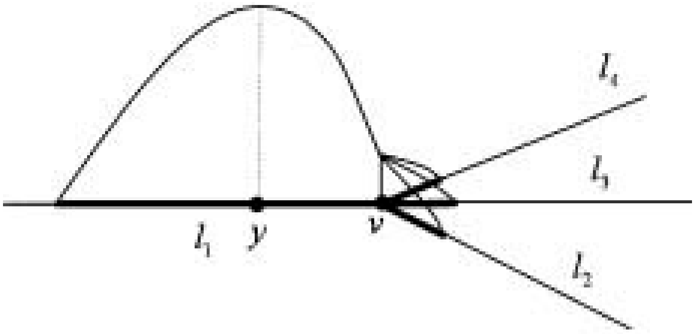

Figure 1 *The simplified example of network kernel function*
(Source : Okabe et al., 2009, P.19).

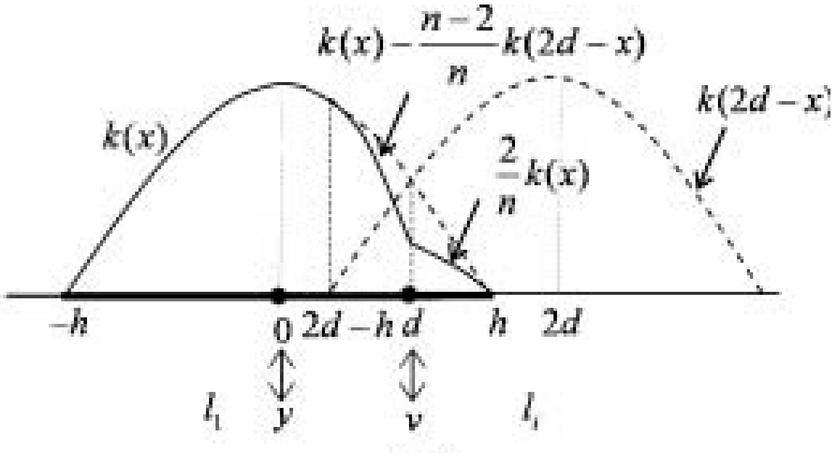

Figure 2 *The three ranges of the network kernel function*

(Source : Okabe et al., 2009, P.19).

Previous research indicates that the choice of kernel bandwidth and cell size have higher impact on the results than the choice of kernel function (Loo and Anderson, 2015; Moridpour and Toran, 2015; Khanh, Pei, and Liang-Tay 2019). In Net-KDE method, the cell size and bandwidth are specified based on the characteristics of the case study area (Loo and Anderson, 2015). Previous literature suggested using bandwidth ranging from 20 meters for urban areas to 1000 meters for rural areas so no optimal kernel bandwidth was recommended in past studies as these studies used different bandwidths (Moridpour and Toran, 2015; Khanh, Pei, and Liang-Tay 2019).

2.2.2 Network-based Nearest Neighbour Distance (Net-NND)

After applying Net-KDE, the significance of the results is validated using Network-based Nearest Neighbour Distance (Net-NND) and the K-function methods. These two methods are used to analyse the spatial patterns of incident point data (in our case is road traffic crash data). They summarise the spatial dependence (clustering or dispersion) over a range of distances.

The Net-NND test is used to examine the complete spatial randomness (CSR) hypothesis which means that RTCs crashes are independently and identically distributed over the road network (Okabe and Sugihara, 2012). Using this test, the measured distances of RTCs to their nearest neighbour are tested against the CSR hypothesis and the detected hotspots are significant if the CSR hypothesis is rejected which implies that the average nearest neighbour distance for the observed data is significantly smaller than the expected distance if they are randomly distributed (Kaygisiz et al., 2015). In this study, the Net-NND test was applied using the Monte Carlo simulation, using a cell size of 200 meters and a confidence interval of 95% to determine the statistical significance.

The Clark-Evans index, which is expressed as the ratio of the average observed Net-NND to the average expected NND, indicates whether the RTCs exhibits clustering (if the index <1.0) or dispersion (index>1.0) (Okabe and Sugihara, 2012).

2.2.3 Network-based K-Function (Net-K-Function)

Ripley's K-Function is another method to analyse the spatial patterns of RTCs data and it has an advantage of not depending only on nearest neighbour distances but on using all point-to-point distances to analyse the spatial clustering at different scales of patterns and to determine the distances where clustering or over-dispersal is significant (Bailey and Gatrell 1995 cited in Spooner et al., 2004). However, since the K-Function calculates the Euclidean distance between incident points, it cannot be used to analyse the spatial patterns of incident points along road network such as road traffic crashes as it can lead to over-detection of spatial clustering (Spooner et al., 2004). Okabe and Yamada (2001) have developed a method to conduct K-Function analysis of point patterns along network.

The univariate network K-function (Net-K-function) was used in this study to test the CSR hypothesis in terms of number of points (RTCs) in a given point set so that the shortest distance from each point is less than a parametric shortest distance. The Net-K-function calculates the shortest path distance *(t)* from each point to the other points $P = \{p_1, \dots, p_n\}$ on the road network, where n is the total number of points in network.

Let $P = \{p_1, \dots, p_n\}$ be a set of n RTCs on road network, L$_T$ be a set of road network links (road segments) and $|L_T|$ be the total length of road network. Okabe and Yamada (2001) defined the Net-K-function $K(t)$ as follows:

$$K(t) = \frac{1}{\rho} E \begin{pmatrix} \text{the number of points } P \\ \text{within network distance } t \text{ to a point } p_i \text{ of } P \end{pmatrix}, \tag{5}$$

Where E(.) is the expected value with respect to $p_1, \dots, p_n$ $(p_i \in P), \rho$ is the density of points P, so that

$$\rho = \frac{n}{|L_T|} \tag{6}$$

with an assumption that the points P (i.e. RTCs in this study) are uniformly and independently distributed over a finite road network and follow homogeneous Binomial distribution (Spooner et al., 2004; Okabe and Sugihara, 2012). The rejection of this assumption means that RTCs are spatially interacting and not forming uniform patterns (Spooner et al., 2004; Okabe and Sugihara, 2012). Thus, the observed network K-function $\widehat{K(t)}$ is given as (Okabe and Yamada, 2001):

$$\widehat{K(t)} = \frac{|L_T|}{\mathrm{n(n-1)}} \sum_{i=1}^{n} \left(\text{number of points } P \text{ on } L_P(t) \right) \tag{7}$$

Monte Carlo simulation is used to find the upper and lower critical values of a significance level α (in this study *α=5%*). From the graphical presentation of the observed Net-K-function and the expected Net-K-function together with the upper and lower confidence envelops, we can conclude at a confidence level of *1-α* (i.e. 95%) either if the RTCs tend to cluster (the graph of the observed Net-K-functions is above the upper confidence envelop) or be dispersed (the graph of the observed Net-K-functions is under the lower confidence envelop (Okabe and Yamada, 2001; Okabe and Sugihara, 2012).

Both Net-NND and Net-K-Function are employed in this study using SANET toolkit in ArcGIS10.2/ArcMap.

2.2.4 Random Forest Algorithm (RF)

The next step is to explore the differences between the cold-, normal- and hot-zones. There are several machine learning techniques which can help us in classifying road network into cold-, normal- or hot-zones. Unlike regression models, machine learning techniques do not have pre-defined relationship between the response and independent variables, so they can identify the associations between the response variable and the predictors without any assumptions about the distribution of the data or pre-defined association (Jiang et al., 2016). The main advantage of the machine learning techniques compared to traditional regression models is that they can detect the complex interactions among the predictors and they are robust to outliers (Jiang et al., 2016).

Random Forest (RF) is one of the machine learning techniques and it is a non-parametric and ensemble statistical learning algorithm, developed by Breiman (2001) to improve the Classification and Regressing Trees (CART) method by combining the results of many decision tree models (Mutanga et al., 2012; Jiang et al., 2016). It has the advantage of its capability of dealing with complex relationships and synthesizing regression and classification functions for both discrete and continuous data (Mutanga et al., 2012). RF "is a classifier consisting of a collection of tree-structured classifiers {$h(\mathbf{x},\Theta_k)$, $k=1,\ldots$ } where the {Θ_k} are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input **x**." (Breiman, 2001, P.6).

Due to its simplicity and superior performance, RF algorithm has been widely used for classification and prediction purposes in many research areas including road safety (Harb et al., 2009; Jiang et al., 2016; Yao et al., 2018).

RF algorithm works by using bootstrapping iteration and randomly selecting a number of features say *m* from all features say *T* in the dataset so that *m*<*T*. Then for each node it calculates the best split point among the *m* features and splitting the node into two daughter nodes and it repeats this step until the proposed number of nodes has been reached. Finally, all these steps are repeated for a number of times say *D* to build *D* number of trees and then by applying the majority voting technique it produces the model of highest voting rate from all the constructed trees. ***Figure 3*** presents the architecture of RF algorithm (Jiang et al. 2016) to classify road traffic accident zone TAZ_i.

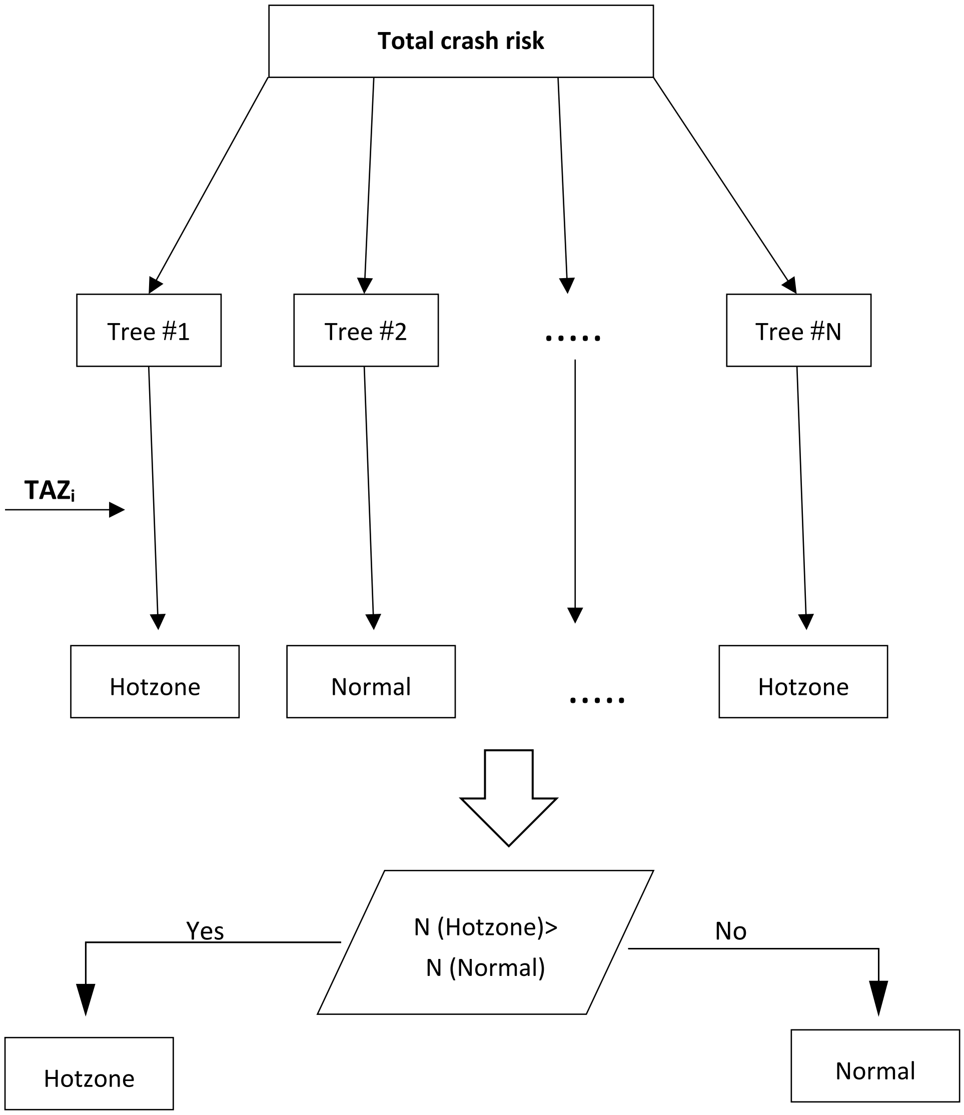

***Figure 3** The Architecture of Random Forest model to predict crash risk*

(Source : Jiang et al., 2016, P.56).

In the process of constructing RF model, the dataset is divided into two sets: the training dataset (drawn from two-third of the full dataset) and the test dataset (one-third of the observations) (Jiang et al., 2016). For each bootstrap iteration and related tree, the observations not included in that sample are denoted as out-of-bag (OOB) data to estimate prediction error (called OOB error). The OOB error "is the error rate of the out-of-bag classifier on the training set" (Breiman, 2001, P.11).

To improve the performance of a RF model, the bias of each tree and the correlation among trees should be minimized (Jiang et al., 2016). Minimizing the bias is achieved by allowing each tree to grow to its maximum depth (Jiang et al., 2016). Decreasing the correlations among trees is achieved by the following two sources of randomization (Jiang et al., 2016):

a. A bootstrap sample of the training dataset should be drawn randomly, with replacement, to build each tree in the RF model.

b. A certain number of features (mtry) should be randomly selected from all the features at each node of a tree to compete for the best split and the minimum error rate is obtained.

One of the important features of the RF algorithm is its ability to provide variable importance measure and this makes the RF one of the most popular analysis techniques used in different scientific areas (Louppe, 2014). This would help not only in making the most accurate prediction but also in identifying the most important predictors in building the model (Louppe, 2014). The importance of each predictor is obtained by calculating the percentage of the increase in the mean squared error when OOB data for each predictor permuted, while keeping other predictors unchanged (Mutanga et al., 2012). Having the values of variable importance will help in ranking the strength of the relationship between the predictors and the response variable (Mutanga et al., 2012).

To calculate the variable importance measure $\overline{D}_i$ for a variable x_i , the following procedures as presented in (Jiang et al., 2016) are shown. Let L_b be the bootstrap sample for the tree T_b, and L_b^{oob} be the corresponding OOB dataset, where $b = 1,2, \dots, B$:

i. Set $b = 1$, build the tree T_1 with bootstrap sample L_1

ii. Identify the corresponding OOB sample L_1^{oob}

iii. Use L_1^{oob} to test the tree T_1 and record the number of correct classifications R_1^{oob} in the test sample L_1^{oob}

iv. For $i = 1,2,3 \dots, N$

a. Permute the values of the variable x_i in L_1^{oob} and save the result into L_{1i}^{oob}

b. Use L_{1i}^{oob} to test the tree T_1 and record the number of correct classifications, R_{1i}^{oob}

v. Redo the steps i-iv for $b = 2,3,4 \dots, B$

vi. The variable importance measure $\overline{D}_i$ for the variable x_i is calculated as:

$$\overline{D}_i = \frac{1}{B}\sum_{b=1}^{B}(R_b^{oob} - R_{bi}^{oob}) \quad (8)$$

vii. $\overline{D}_i$ is normally distributed according to the central limit theory. By computing the standard error (SE) of the decrease in the correct classification, the $\overline{D}_i$ is standardised as:

$$\widehat{\widehat{D}}_i = \frac{\bar{D}_i}{SE} \qquad (9)$$

the higher the value of $\widehat{\widehat{D}}_i$, the greater the importance of the variable x_i (Jiang et al., 2016).

The optimum performance of RF algorithm is achieved by tuning the number of trees and the number of features (mtry) which are randomly selected at each node (Breiman, 2001; Jiang et al., 2016). Then the final decision will be based on the value of the OOB error rate, so that the lower the OOB error rate the better the performance of the model (Breiman, 2001; Jiang et al., 2016).

Before building RF algorithm in the current study, the road network in AS'Seeb[1] and Bowshar[2] (where most of road commuting occur in Muscat Governorate) were split into 82 zones. These zones represent sections of the road network where there are bridges, roundabouts, crosses, or set of T-junctions. The reason behind splitting the road network based on the location of main junctions is because results of Net-KDE indicated that the higher values of Net-KDE found on these junctions. The process of splitting the road network are as follows:

I. The centre of main junction where the zone is selected used as a centre for that given zone, and this centre is then used to determine the areas of road network covered on each zone.

II. The distance from the centre of the zone is calculated from all directions so that all road segments locating within distance not exceeding 1,000 metres are included in area covered by that given zone. However, if there are two main junctions located next to each other and the distance between the centres of the junctions is less than 2,000 metres then this shared area is divided equally between the two zones.

III. The Net-KDE for each zone is calculated by using the results of Net-KDE and adding the Net-KDE of all road segments on that given zone.

RF algorithm was applied in this study using R software version 3.3.3. The algorithm was constructed using features of 82 road zones generated by combining four different datasets: the road traffic crash data of Muscat obtained from NRTC database of Royal Oman Police, results of Net-KDE, Muscat road

[1] AS'Seeb is a *Wilayat* (administrative unit) in the northern part of Muscat where Muscat Airport, Sultan Qaboos University, many public organisations and commercial activities are located.

[2] Bowshar is a Wilayat in northeastern part of Oman, where Sultan Qaboos Grand Mosque, main buildings of most Ministries, Embassies shopping malls and big companies are located. The biggest hospital (the Royal Hospital) and Sultan Qaboos Sports complex are also located in this Wilayat.

network and the road traffic volume data generated by the researcher using Google-Maps application for each intersection in AS' Seeb and Bowshar regions in Muscat.

The features used in this algorithm include the following (most of these features are road geometry related features):

1. Net-KDE of RTCs for each zone which then classified as cold-, normal- or hot-zone (target variable), where cold, normal, and hot zones are defined based on the value of Net-KDE as follow:
Cold zones (low risk)): zones where Net-KDE < 326,

Normal zones (medium risk): zones where Net-KDE < 800,

Hot zones (high risk): zones where Net-KDE>800.

2. Level of road traffic volume (4 different level (1-4), with 1 representing the lowest level and 4 the highest level of traffic volume).

3. Type of road (one-way direction or two-way direction).

4. Number of entrances and exits on each intersection.

5. Distance (in meters) from each zone to its next nearest junction.

6. Complexity of the zone (2 levels, 0= not complex, 1=complex) and the meaning of complexity here is representing the shape of the zone, especially the shape of the bridges and roundabouts.

7. Maximum level of speed of the road on which the zone is located (five different levels 1-5, with 1 representing the lowest speed level 50km/h, and 5 the highest level 120km/h).

8. Junction Type (roundabout, bridge, cross, or set of T-junctions).

In addition, summary of the number of crashes on each zone is classified by:

a. cause of the crash: (speed, carelessness, fatigue, alcohol consumption, and non-human factor),

b. Type of the crash (Hit other vehicle, run over human, run over animal, hit fixed object on the road or overturn), and

c. Severity level of the crash (no injury, mild injury, moderate injury, severe injury and fatal crash). But these factors were not included in the construction of the RF model.

2.2.5 Wilcoxon Test (Mann-Whitney U Test)

Wilcoxon tests were employed to understand the different effect of each variable on hot-zones and normal-zones using SPSS software version 24.0.0.0. Wilcoxon Test (Mann-Whitney U Test) is a non-parametric method appropriate for testing the equality of means of of ranking of two populations (Jiang et al., 2016). The following annotation were drawn from (Jiang et al., 2016) explaining how Wilcoxon Sum Rank Test is applied.

Assume there are the two populations of size n_1 and n_2, the Wilcoxon test is conducted as follows:

I. list the observations of both samples from smallest to largest

II. Give a rank from 1 to N (where $N = n_1 + n_2$) to all observations in ascending order. If any observation is repeated more than one times, then take the average of their rank position.

III. For each sample, sum the ranks of all observations in that sample. Let R_1 and R_2be the sum of the ranks in the first and second sample respectively. The null hypothesis of Wilcoxon test is H_0 : There is no difference in the mean of ranks of the two samples, while H_1: there is a difference in the mean of ranks of the two samples.

IV. For each sample calculate U_{stat} , so that for the first sample the U_{stat} is computed as:

$$U_{stat1} = R_1 - \frac{n_1(n_1+1)}{2} \quad (10)$$

$$U_{stat2} = R_2 - \frac{n_2(n_2+1)}{2} \quad (11)$$

V. Compare the values of U_{stat} of both samples and specify the one with the lowest value.

VI. Specify the significance level α (for example, let $\alpha = 0.01$)and assume U_{stat1} has the lowest value. Identify the critical value of U_{stat} from the critical value of the Mann-Whitney U Test by taking the value where column number n_1 and row number n_2 cross each other, and call the value $U_{critical}$ at $\alpha = 0.01$.

VII. Compare the values of $U_{stat1}$and $U_{critical}$, if U_{stat1}<$U_{critical}$ at $\alpha = 0.01$, then we reject the null hypothesis and conclude that there is a difference in the mean of ranks of the two samples.

2.2.6 Spatio-Temporal Hot-zone Analysis

The spatio-temporal analysis is an investigation of correlation of a given factor over space and time (Mohaymany et al., 2013). The spatio-temporal analysis used to explore whether RTCs are more likely to cluster in same locations over a particular period of time or not (Mohaymany et al., 2013). In this study, the spatio-temporal analysis was conducted by applying Net-KDE for each year in the study period (2010-2014) using SANET toolkit in ArcGIS 10.2. Then, the road network in both AS'Seeb and Bowshar was split into same zones as those used in the in RF algorithm and the same procedures are used to find the Net-KDE of each zone in each year. After that, the following measure is conducted:

I. For each zone, the value of Net-KDE from year to year is calculated.

II. For each year of the study period, the values of the Net-KDE of the zones listed in descending order and each zone assigned a rank based on its position in the list.

III. For each zone, the sum of the annual ranking and variance in annual ranking are calculated.

This measure will help to identify the road zones which represent persistent problem areas for road traffic crashes in the selected cities.

3.Results

3.1 Net-KDE

The results of the Net-KDE are presented in ***Figure 4*** **to Figure 7.** They are coded into different colours, so the maps provide a clear visualisations of high risk areas where there is increase in the degree of redness of road section. Findings from the Net-KDE analysis demonstrate evidence of spatial clustering of RTC hot-zones on long roads, especially on Sultan Qaboos Highway, demarcated by intersections, and complex bridges and roundabouts. The crash-risk increases with higher density of intersections on the road network. This result is intuitive as this part of road network in Muscat has the highest level of traffic interactions which generate more safety problems among road users. It is also clear that since Muscat Expressway is extending outward from the core market and workplace areas, the crash risk tends to decrease in this area .

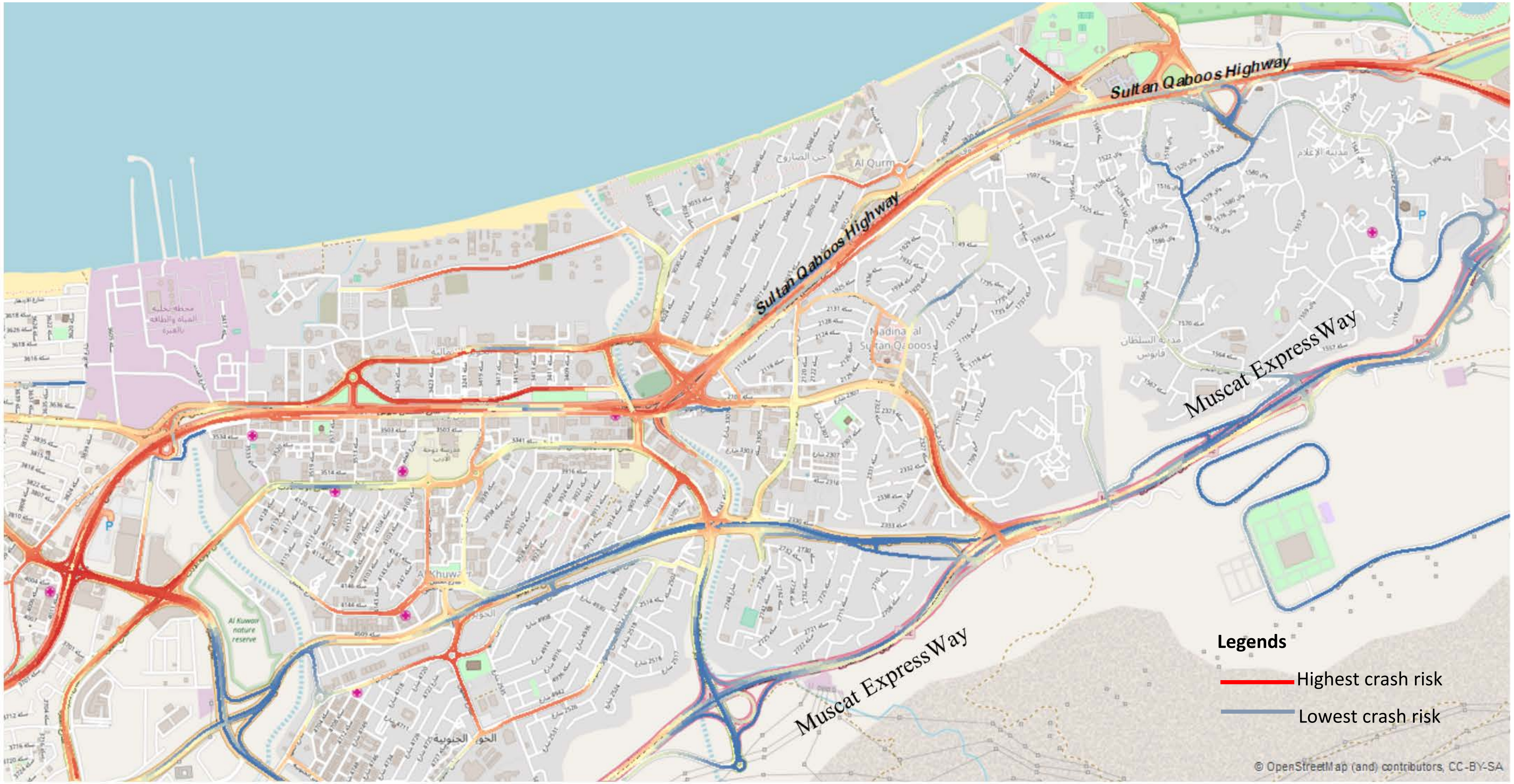

Figure 4 Results of Net-KDE on road network in Al-Khuwair (Area in Bowshar) where the main buildings of most Ministries, Embassies, shopping malls and big companies located

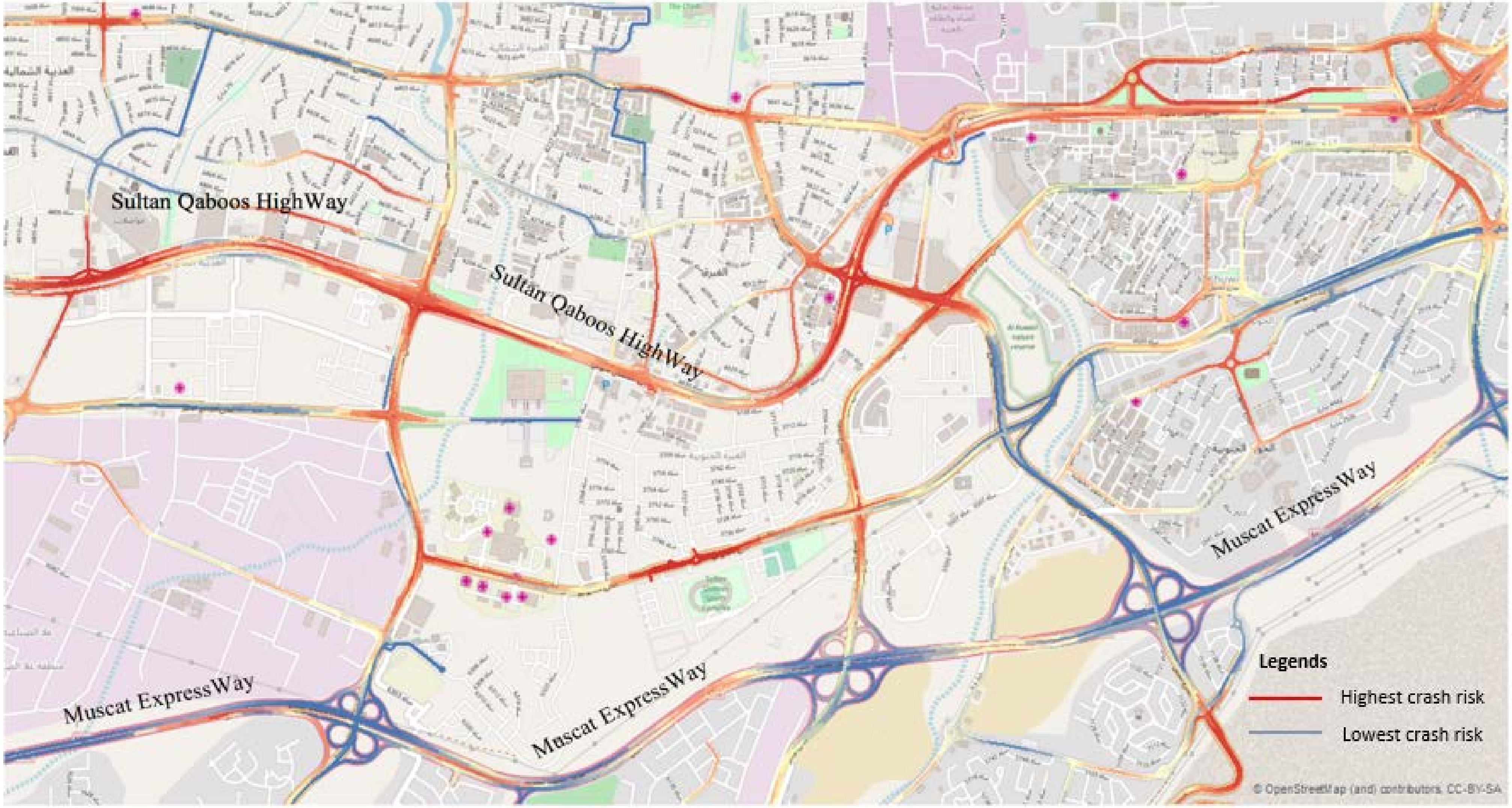

Figure 5 *Results of Net-KDE on road network from Othaibah to Al-Khuwair and Al-Ghubrah where the main buildings of most Ministries, embassies and big companies located*

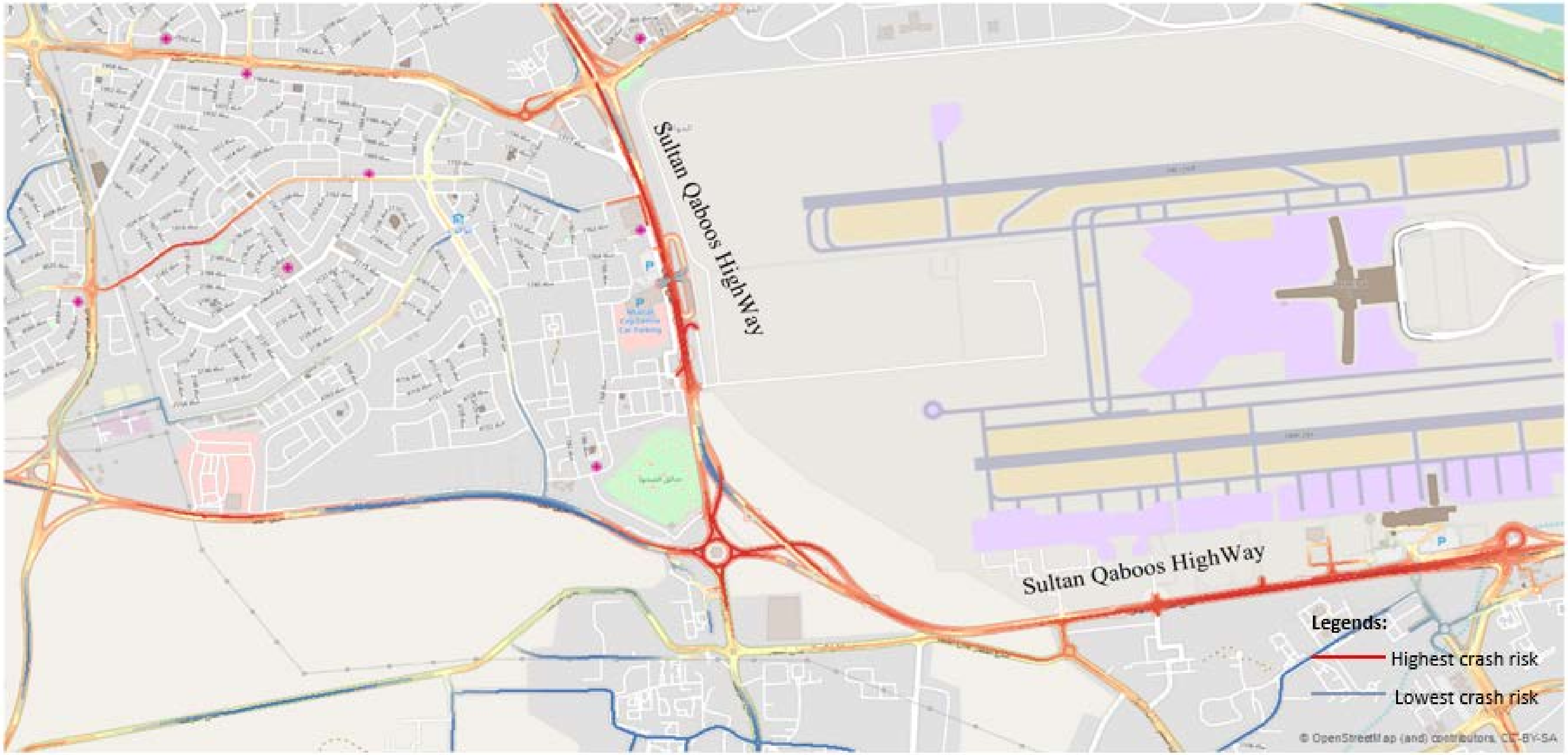

Figure 6 *Results of Net-KDE on road network from AS' Seeb City to Muscat International Airport through Burj AS' Sahwah roundabout (the biggest roundabout in Muscat from which drivers are commuting to other adjacent Governorates)*

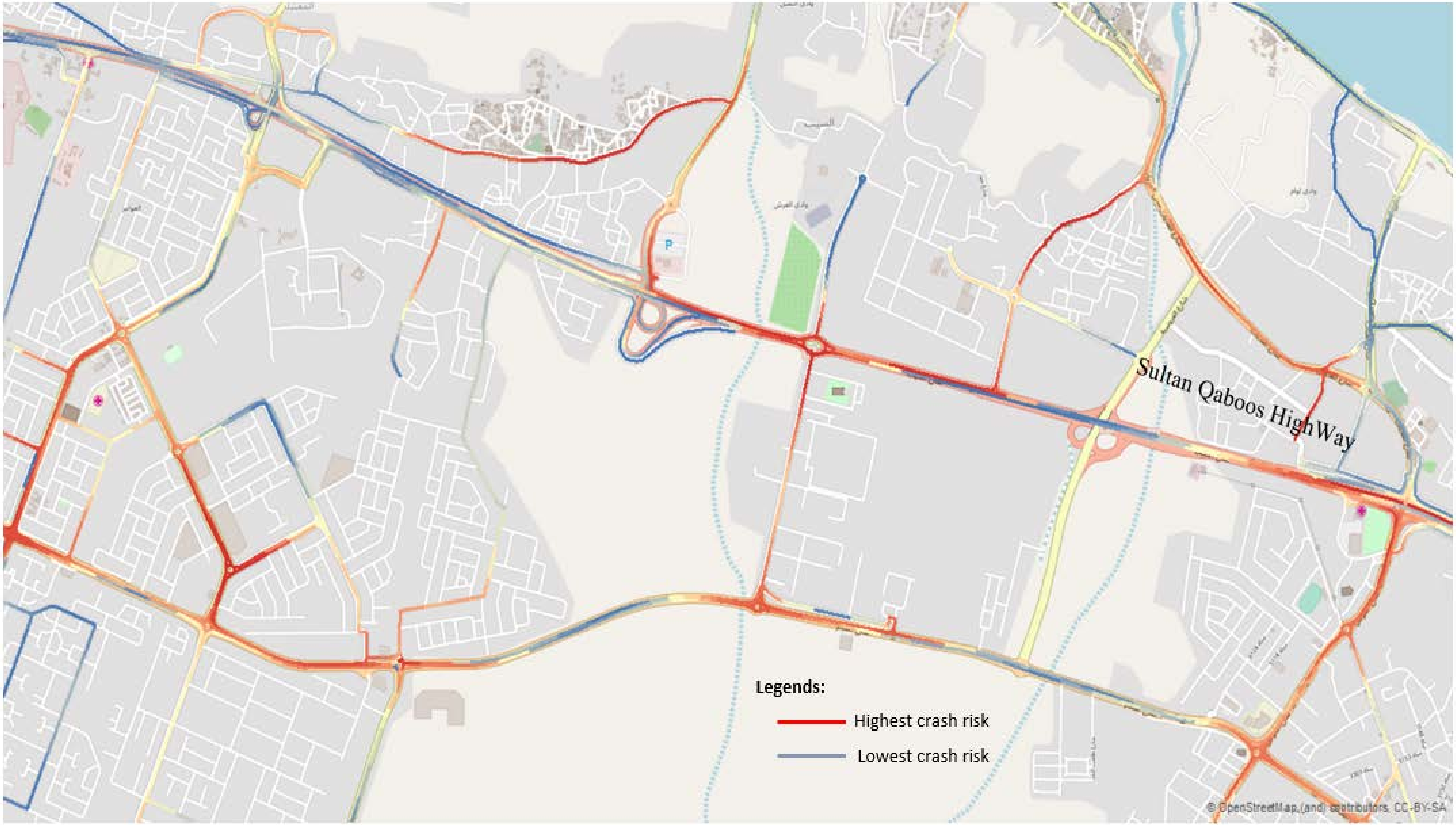

Figure 7 *Results of Net-KDE on road network in AS' Seeb Wilayat.*

3.2 Net-NND

As mentioned earlier, in order to evaluate the significance of Net-KDE of detecting the RTC hot-zones, both Net-NND and Net-K-function should be conducted to summarise the spatial dependence (clustering or dispersion) over a range of distances.

In this study, the Net-NND test was applied using the Monte Carlo simulation using a cell size of 200 meters and a confidence interval of 95% to determine the statistical level of significance. ***Table 1*** and ***Figure 8*** summarise the results of the Net-NND test (in meters).

Table 1 *Spatial dependency of RTCs in Muscat based on the Net-NND results*

Average observed Net-NND	54.371
Lower critical value for one-sided significance level	115.615
Upper critical value for one-sided significance level	119.074
Average expected Net-NND	117.327
P-value	0.0001
Clark-Evans index= Average observed Net-NND/ Average expected Net-NND	0.463

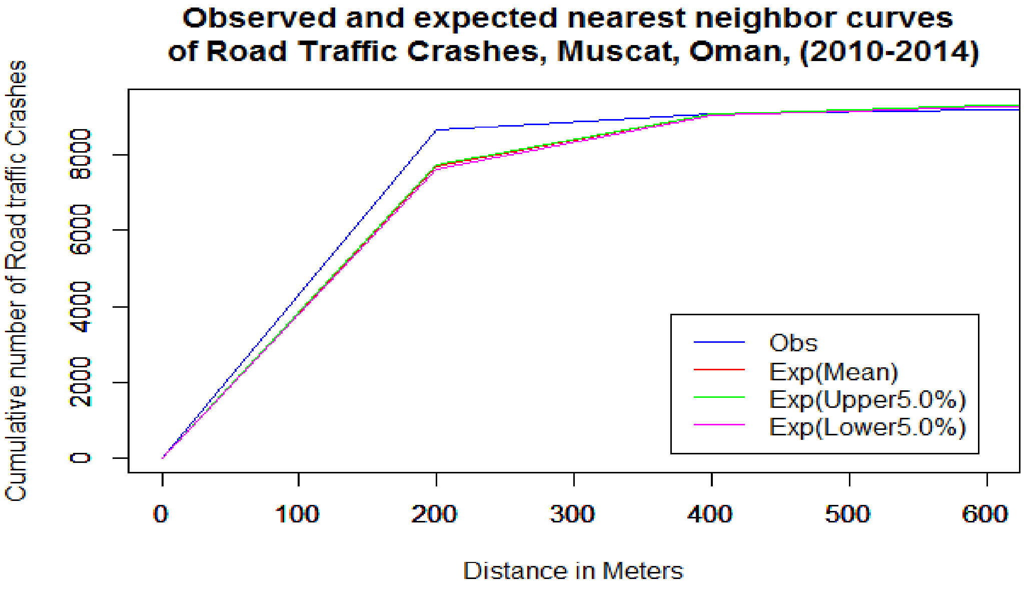

Figure 8 *Observed and Expected Nearest Neighbour curves of RTCs in Muscat based on Net-NND test*

The results show that the clustering patterns of RTCs are significant ($p<0.001$). Figure 8 displays four different curves: the observed curve (in blue), the expected curve under CSR hypothesis (in red) and the upper and lower envelop curves (in green and pink respectively) for one-sided 5% significance level. If the observed curve lays between the upper and lower envelop curve, the CSR hypothesis cannot be rejected, and if it lays above the upper envelop curve then the data exhibits clustering (Okabe and Sugihara, 2012). As it is clear from the Table, the Clark-Evans index=0.463 and the observed curve is above the upper envelop curve for distances less than 450 meters, the CSR hypothesis is rejected with 0.95 confidence level. This confirmed the significance of clustering patterns of RTCs along road network.

3.3 Net-K-Function

The K-Function method was applied in this study using Monte Carlo simulation to find the upper and lower critical values of a significance level $\alpha=5\%$. As mentioned previously, the graphical presentation of the observed Net-K-function and the expected Net-K-function together with the upper and lower confidence envelops helps to conclude at a confidence level of $1-\alpha$ (i.e. 95%) whether the RTCs tend to cluster (the graph of the observed Net-K-functions is above the upper confidence envelop) or be dispersed (the graph of the observed Net-K-functions is under the lower confidence envelop (Okabe and Sugihara, 2012). ***Figure 9*** summarises the results of the Net-K-Function method. The univariate spatial patterns of the Net-K-Function show significant deviations from the CSR hypothesis and indicate the clustering patterns (the graph of the observed K-functions is above the upper confidence envelop) of RTCs along the road network confirming the results obtained from the Net-KDE analysis. Therefore, we can conclude that the detected hotspots are significant as both the Net-NND and Net-K-Function methods reject the null hypothesis of random distribution.

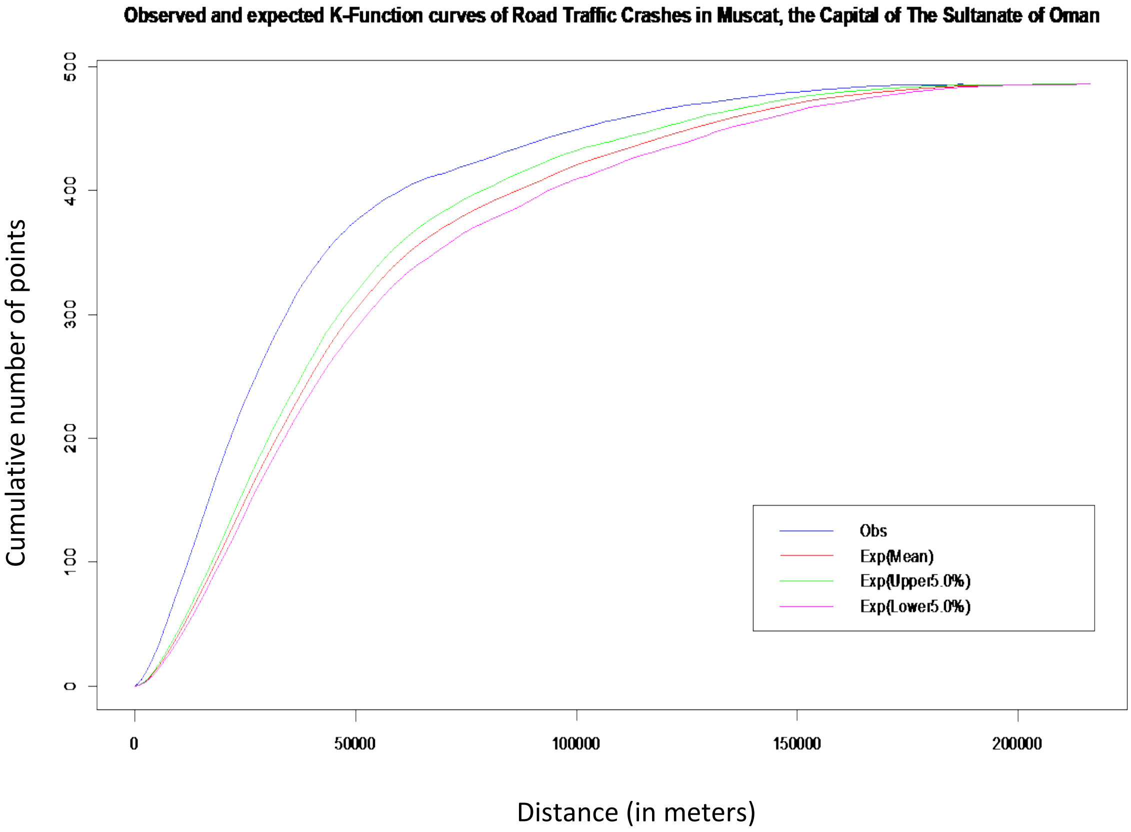

Figure 9 *Observed and Expected Net-K-function curves of RTCs in Muscat based on Net-k-function method*

3.4 RF Algorithm

Road geometry related attributes and road traffic related attributes are used in the construction of the RF Algorithm. The kernel density of every selected zone was found by summing the value of Net-KDE of all road segments located on that given zone. The average Net-KDE of all selected zones was 1128. A series of random forest models were fitted to explore the characteristics of cold-zones, normal-zones and hot-zones. Results of two models which produced the lowest percentages of classification error are presented here.

The first RF model was built using all the 82 intersections and they were classified into 2 classes (Hot-Zones= Net-KDE>800, Normal-Zones=Net-KDE <=800). A total of 500 trees with two split at each node were used to construct this model. The results of this model are summarised in ***Table 2*** and ***Figure 10***:

Table 2 *Out of Bag error for Random Forest Algorithm to classify Hot- and Normal-Zones*

Actual Class	Internal Test/Out of Bag estimate			
	N-Cases	N-accurately classified	N-misclassified	Prediction error
Normal-Zone	39	26	13	0.3333
Hot-Zone	43	32	11	0.2558
Average Out Of Bag estimate error				0. 2927

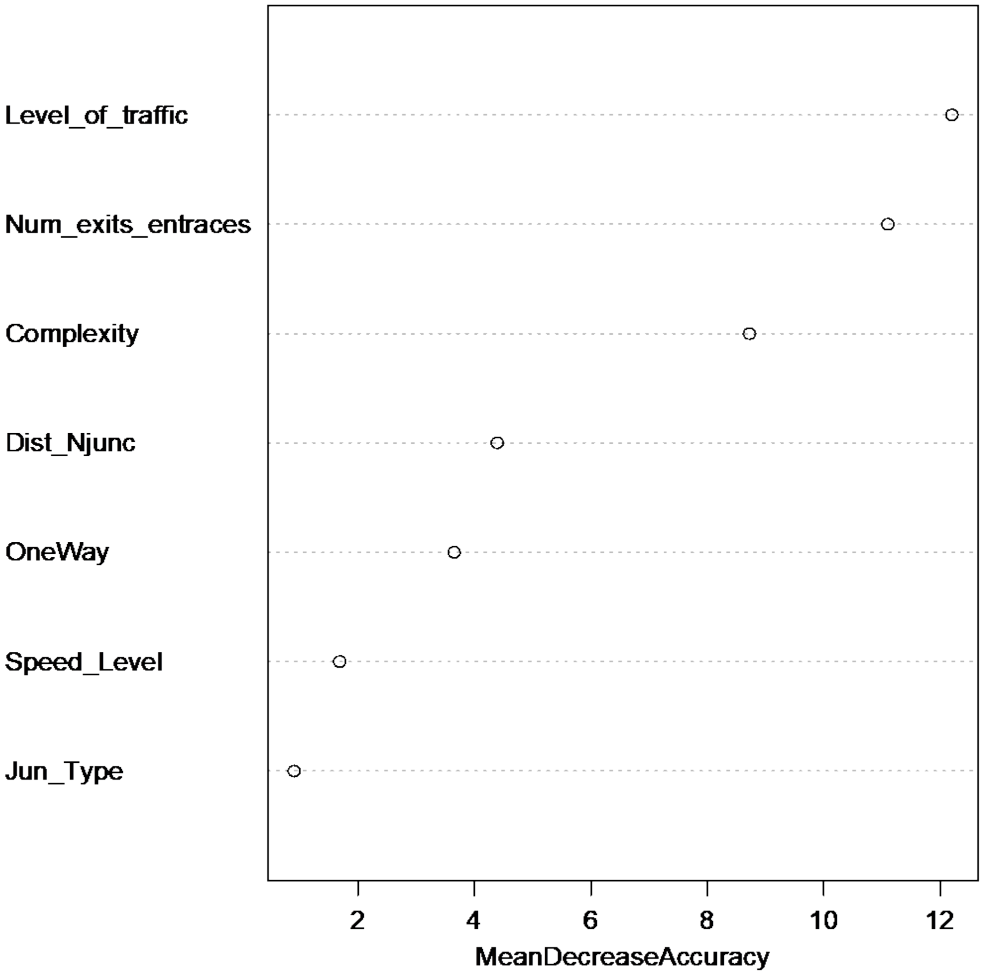

Figure 10*Mean decrease in accuracy of Random Forest Algorithm to classify Hot- and Normal-zones*

Figure 10 displays the mean decrease in accuracy which described the decrease in model accuracy from permuting the values of each variable/feature, in other words it gives indication about the variable importance in the model. The Figure indicates that level of road traffic, number of entrances and exits (num_exits_entrances) in the intersection/zone, complexity, and distance to nearest junction (Dist_Njunc) are the most important factors affecting the accuracy of classifying normal and hot-zones. Conversely, maximum level of speed (Speed_level), junction type (Jun_Type), and type of road (one-way or two-way directions) do not affect the accuracy of classification of normal and hot-zones. As it is clear from Table 2, the model was able to accurately classify 26 out of 39 normal-zones and failed to classify 13 zones in their right class. Likewise, 32 out of 43 hot-zones (74.42%) were accurately classified into their right class. Overall more than 70% of all zones were accurately classified.

Only zones which have Net-KDE above (1128) or under (326) included as hot-zones and cold-zones respectively in building the second RF model. A total of 44 zones were eligible to include in building this model and the results of this model are summarised in ***Table 3*** and ***Figure 11***. Similar to the results in Figure 10, Figure 11 indicates that level of traffic, the number of entrances and exits in the zone and type of junction are the most important factors in classifying cold and hot-zones. Unlike the first model, in this model the distance to the nearest junction and complexity level of the zone appear not affecting the accuracy of classifying cold- and hot-zones. Table 3 shows that the model was able to accurately recognise about 90% and 80% of hot- and cold-zones respectively. Interestingly, the model has an overall internal error rate of 13.64 % in classifying zones as cold or hot-zones and accordingly, more than 86.36% of these zones were correctly identified.

Table 3 *Out of Bag error for Random Forest Algorithm to classify Hot- and Cold-Zones*

Actual Class	**Internal Test/Out of Bag estimate**			
	N-Cases	**N-accurately classified**	**N-misclassified**	**Prediction error**
Cold-Zone	15	12	3	0.2000
Hot-Zone	29	26	3	0.1034
Average Out Of Bag estimate of error				0.1364

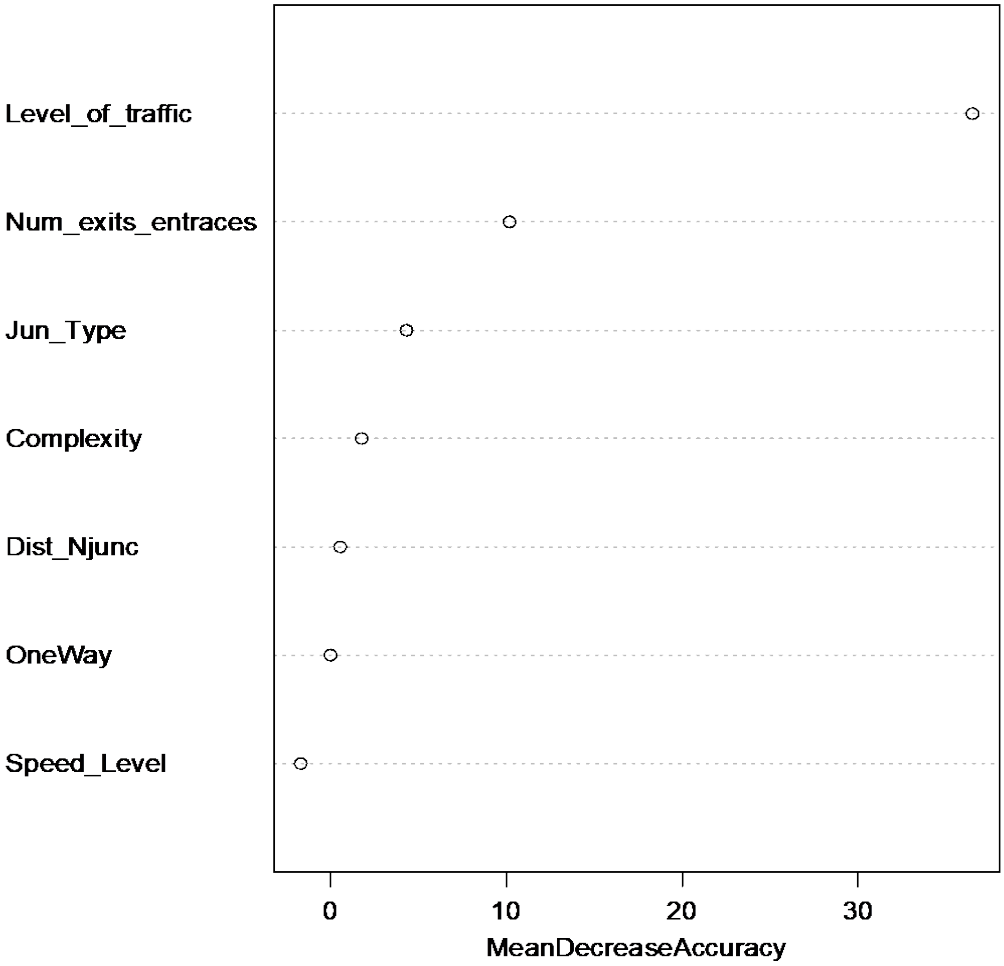

Figure 11 *Mean decrease in accuracy of Random Forest Algorithm to classify Hot- and Cold-zones*

3.5 Wilcoxon Test (Mann-Whitney U Test)

Wilcoxon tests applied in this study to quantitatively understand the different effect of each variable on hot-zones and normal-zones. Results from Wilcoxon tests are shown in ***Table 4.*** Overall, it is clear that most of the features are significantly different between normal- and hot-zones. More precisely, there is a difference in the mean of ranks of normal- and hot-zones in the number of exits and entrances, distance to nearest junction, complexity of the intersection and level of traffic at a significance level of $\alpha = 0.01$. However, it is clear that there is no significant difference between the two types of zones when considering type of junction (i.e. roundabout, cross, bridge or set of junctions), maximum level of speed, and type of road (i.e. one-way direction road, or two-ways). The highest difference can be seen in the number of exits and entrances; hot-zones appear to have higher number of exits and entrances compared with normal-zones. Likewise, hot-zones are characterised by having shorter distances to their nearest junction as compared to normal-zones. Similarly, higher level of traffic seems to increase the crash risk on road network and thus increasing the likelihood of having hot-zones. It is also clear that hot-zones are more likely to appear on areas where complex structure of road network exist when compared with normal-zones.

Table 4 *The effects of road and traffic related feature on classifying Hot- and Normal-Zones based on Wilcoxon Test*

Factor	**Hot-Zones (N=43)**		**Normal-Zones (N=39)**		**Sig. Level**
	Mean Rank	Sum of Ranks	Mean Rank	Sum of Ranks	
umber of exits and entrances	50.90	2188.50	31.14	1214.50	<0.000
istance to nearest junction	34.37	1478.00	49.36	1925.00	0.004
ype of junction	44.50	1913.50	38.19	1489.50	0.211
omplexity	48.57	2088.50	33.71	1314.50	<0.000
evel of traffic	49.73	2138.50	32.42	1264.50	<0.000
aximum level of Speed	40.98	1762.00	42.08	1641.00	0.814
ype of road	44.59	1917.50	38.09	1485.50	0.029

The effects of road and traffic related feature on classifying Hot- and Cold-Zones summarised in ***Table 5*** and it shows that there is a difference in the mean of ranks of cold- and hot-zones in the number of exits and entrances, level of traffic and complexity of the intersection at a significance level of $\alpha = 0.01$. However, there is no significant difference with the type of road, type of junction, distance to nearest junction, and maximum level of speed.

Table 5 *The effects of road and traffic related feature on classifying Hot- and Cold-Zones based on Wilcoxon Test*

Factor	**Hot-Zones (N=29)**		**Cold-Zones (N=15)**		**Sig. Level**
	Mean Ranks	Sum of Ranks	Mean Ranks	Sum of Ranks	
umber of exits and entrances	27.81	806.50	12.23	183.50	<0.000
istance to nearest junction	20.00	580.00	27.33	410.00	0.072
ype of junction	21.91	635.50	23.63	354.50	0.662
omplexity	25.90	751.00	15.93	239.00	0.004
evel of traffic	28.88	837.50	10.17	152.50	<0.000
aximum level of Speed	21.05	610.50	25.30	379.50	0.239
ype of road	22.74	659.50	22.03	330.50	0.631

Spatio-Temporal Patterns

After estimating the annual crash density for each zone on road network, the densities should be analysed temporally to identify road zones which represent persistent problem areas for road traffic crashes in the study area. The visual comparison of maps by year reveals the dependency of high-crash-density locations, however, due to limited space, these results are presented in Appendix A and B. However, the temporal analysis was conducted by applying Pearson correlation for each pair of years to assess whether there is a consistency in the locations of hot-zones or not. ***Table 6*** displays the results of the spatio-temporal Pearson correlation for each pair of years during the study period. Similarly, ***Figure 12*** shows an example of this spatio-temporal dependency, and it indicates that RTCs are inclined to cluster in the same locations within the study period. It is also clear that the crash risk had increased gradually from 2010 to 2014 in the area shown in Figure 12.

Table 6 *The spatio-temporal correlation for each pair of years during the study period (2010-2014) of RTCs in Muscat Governorate*

Year of crash	correlation				
	2010	2011	2012	2013	2014
2010		0.828**	0.627**	0.524**	0.409**
2011	0.828**		0.665**	0.581**	0.574**
2012	0.627**	0.665**		0.753**	0.541**
2013	0.524**	0.581**	0.753**		0.769**
2014	0.409**	0.574**	0.541**	0.769**	

ᵏ. Correlation is significant at the 0.01 level (2-tailed).

Results from the Pearson correlation Table indicates that there is a strong positive correlation (Pearson correlations are above 0.5) between most pairs of the study period except the pair representing the years 2010 and 2014. However, it is important to mention that for the year 2014 we have data for 10 months only (1st January-2nd November). It is also clear that the correlation becomes stronger when the pair represents two subsequent years (i.e. 2010 and 2011, 2011 and 2012 and so on). These results indicate that RTCs are more likely to cluster over the same road zones during the five years of the study period.

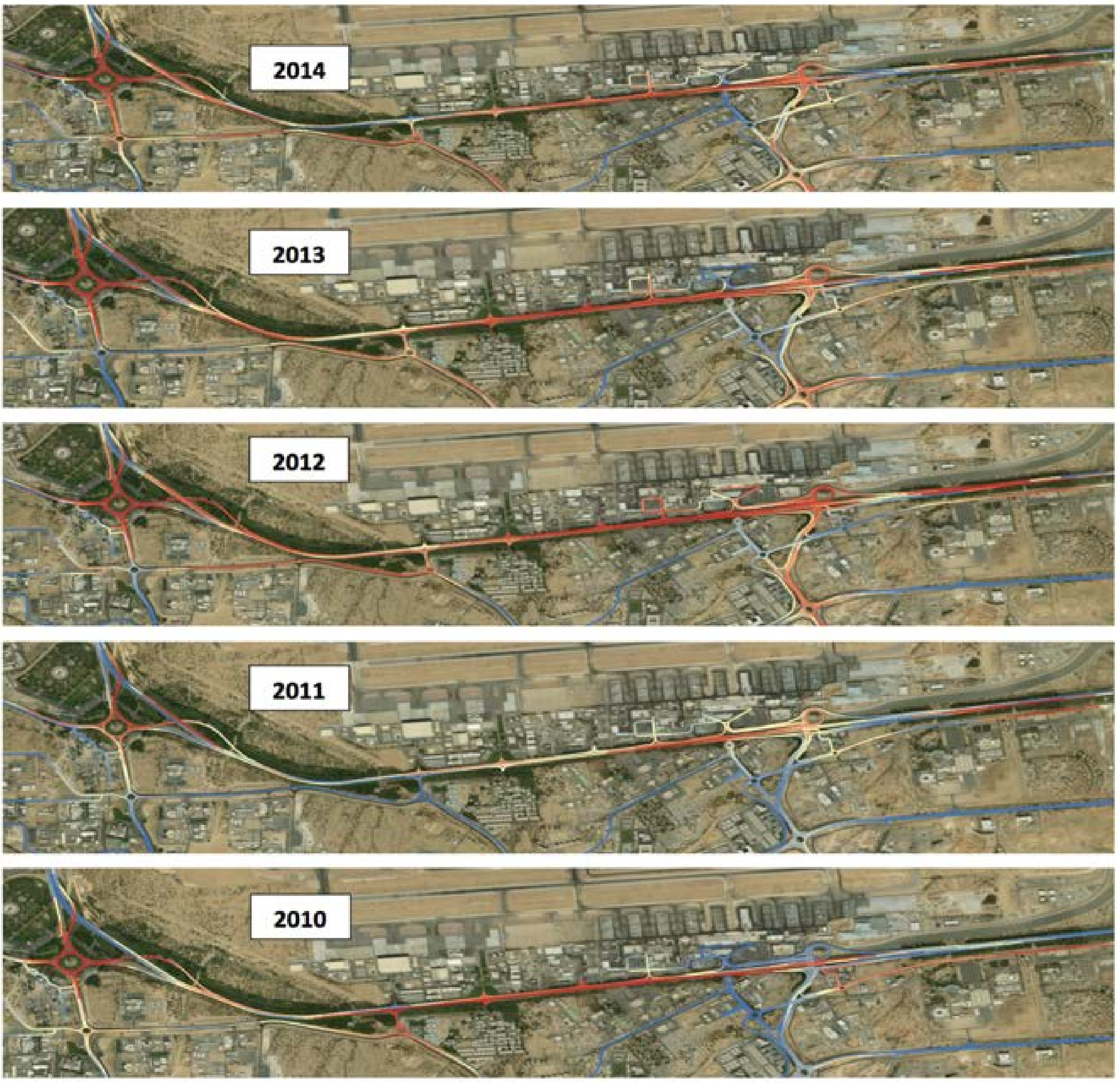

Figure 12*Spatio-temporal patterns of RTCs next to Muscat Airport in AS’Seeb City*

To identify the road zones which represent persistent problem areas for road traffic crashes, the values of the annual Net-KDE of the zones are listed in descending order and each zone is assigned a rank based on its position in the list. Then, the sum of the annual rankings, mean of annual ranking, variance and standard error (SE) of annual ranking for each zone are calculated. This step is repeated using the Olympic average approach in assigning mean annual rank for each zone. Following the Olympic average approach, the lowest and highest ranks of each zone are removed, and the remaining three ranks are used to calculate the sum, mean, variance and SE. ***Figure.13*** shows the scatter plot of mean of annual ranking of each zone resulting from both methods. The scatter plot and the variances resulting from the Olympic average approach show that this is a more robust approach and has smaller variance compared with results of using all 5-years ranks. ***Table 7*** presents the results drawn from the Olympic average, the variance and SE indicate the homogeneity of hot- and cold-zones locations. In other words, it is clear that most hot-zones have small variance and SE indicating that these locations are always having the highest densities of RTCs, and consequently, this confirms the consistency of hazardous locations over time. Although cold-zones appear to have low SE in their ranking, however, these zones are more likely to experience higher fluctuation in their annual ranking (which could also mean there is a heterogeneity in crash risk over these zones) compared with hot-zones.

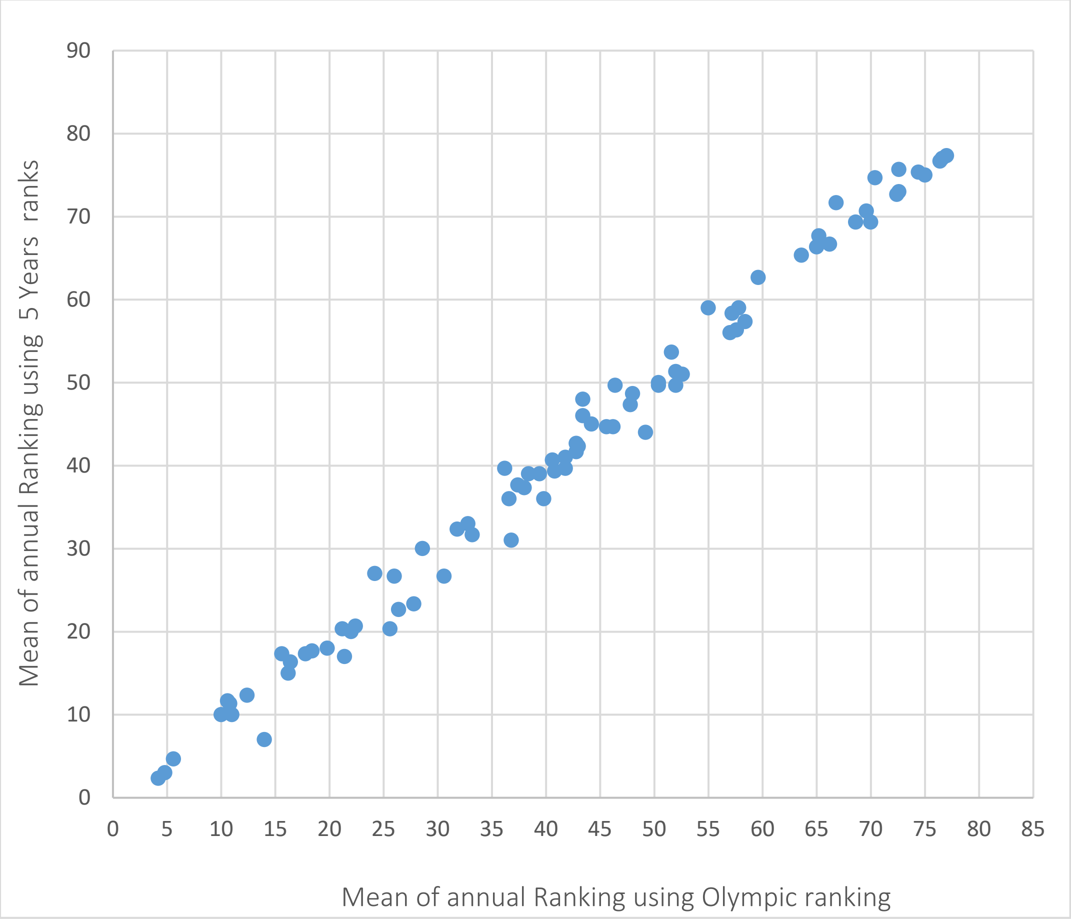

Figure 13*Scatter plot of mean of annual ranking using the 5-year and the Olympic average approach*

Table 7 Variance of ranking of road hot-zones and cold-zones based on the sum of annual ranking using Olympic average approach

Hot-Zones (10 with lowest sum of annual ranking)					**Cold-Zones (10 with highest sum of annual ranking)**				
Zone_ID	Sum of annual Ranking	Mean of annual ranking	Variance of annual ranking	S.E.	Zone_ID	Sum of annual Ranking	Mean of annual ranking	Variance of annual ranking	S.E.
12	7	2.333	5.333	2.309	2	232	77.333	41.333	6.429
22	9	3.000	3.000	1.732	42	231	77.000	7.000	2.646
19	14	4.667	14.333	3.786	15	230	76.667	25.333	5.033
21	21	7.000	1.000	1.000	39	227	75.667	10.333	3.215
10	30	10.000	4.000	2.000	60	226	75.333	25.333	5.033
17	30	10.000	13.000	3.606	41	225	75.000	9.000	3.000
14	34	11.333	40.333	6.351	37	224	74.667	8.333	2.887
24	35	11.667	4.333	2.082	51	219	73.000	52.000	7.211
79	37	12.333	20.333	4.509	3	218	72.667	12.333	3.512
76	45	15.000	4.000	2.000	33	215	71.667	25.333	5.033

3.6 Validation of Net-KDE results using iMAAP pilot data

This section compares the results of Net-KDE using RTC data of year 2014 with pilot data drawn from iMAAP, network based crash analysis system developed by the UK Transport Research Laboratory. In 2015, ROP used iMAAP in documenting a sample of RTCs in two areas in **Muscat** namely: **Al-Khoudh** and **Othaiba** and both of these areas are located in AS' Seeb Wilayat. The iMAAP pilot data includes information about the sex, age and nationality of drivers, number of casualty killed, number of casualty injured, type and number of vehicles involved, crash date, time, day of week, severity of the crash, primary and secondary cause of crash, latitude of longitude of the crash location, landmark, type of collision, type of road, type of road, road name, number of lanes, carriageway width, shoulder width, light condition, street light, weather conditions, and crash description. The pilot data has details of 255 incidents collected between January 2015 and August 2015.

Figures 14 shows the scatter plot of the Net-KDE of randomly selected locations in Al-Khoudh and Othaiba using iMAAP data and a sample of 269 observations from 2014 dataset. The scatter plot clearly indicates that there is a strong positive correlation between the density of RTCs in both datasets. However, it is important to highlight that these two dataset are not covering the same time period (i.e. iMAAP data is a sample of crashes occurring between January 2015 and August 2015). ***Figure 15*** confirms the result of the scatter plot and it provides a visual comparison of the Net-KDE values produced from both datasets using maps. The maps clearly indicate that the Net-KDE cached similar spatial patterns identifying almost the same hot-zones in both dataset, and both Figures provide evidence about the accuracy of the geo-coded data generated by the authors of this paper.

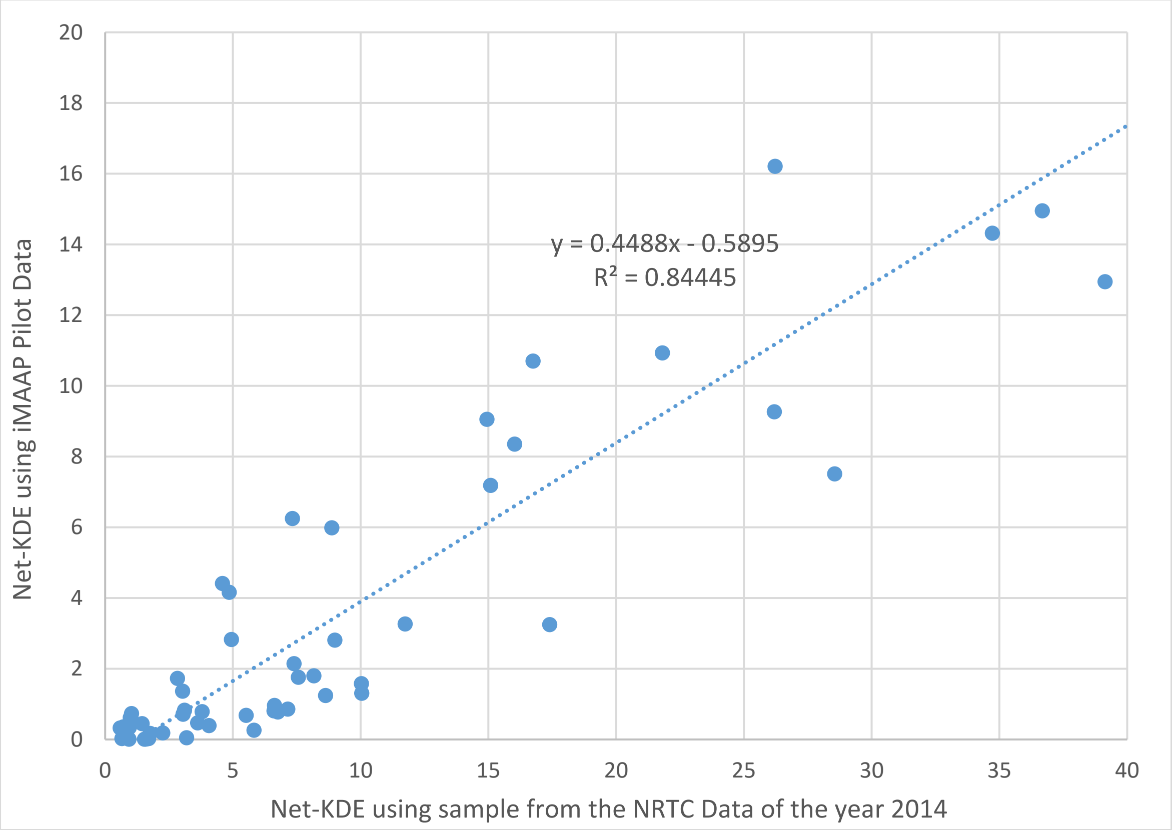

Figure 14*Scatter Plot of Net-KDE using iMAAP Pilot Data and a sample from the NRTC Data of the year 2014*

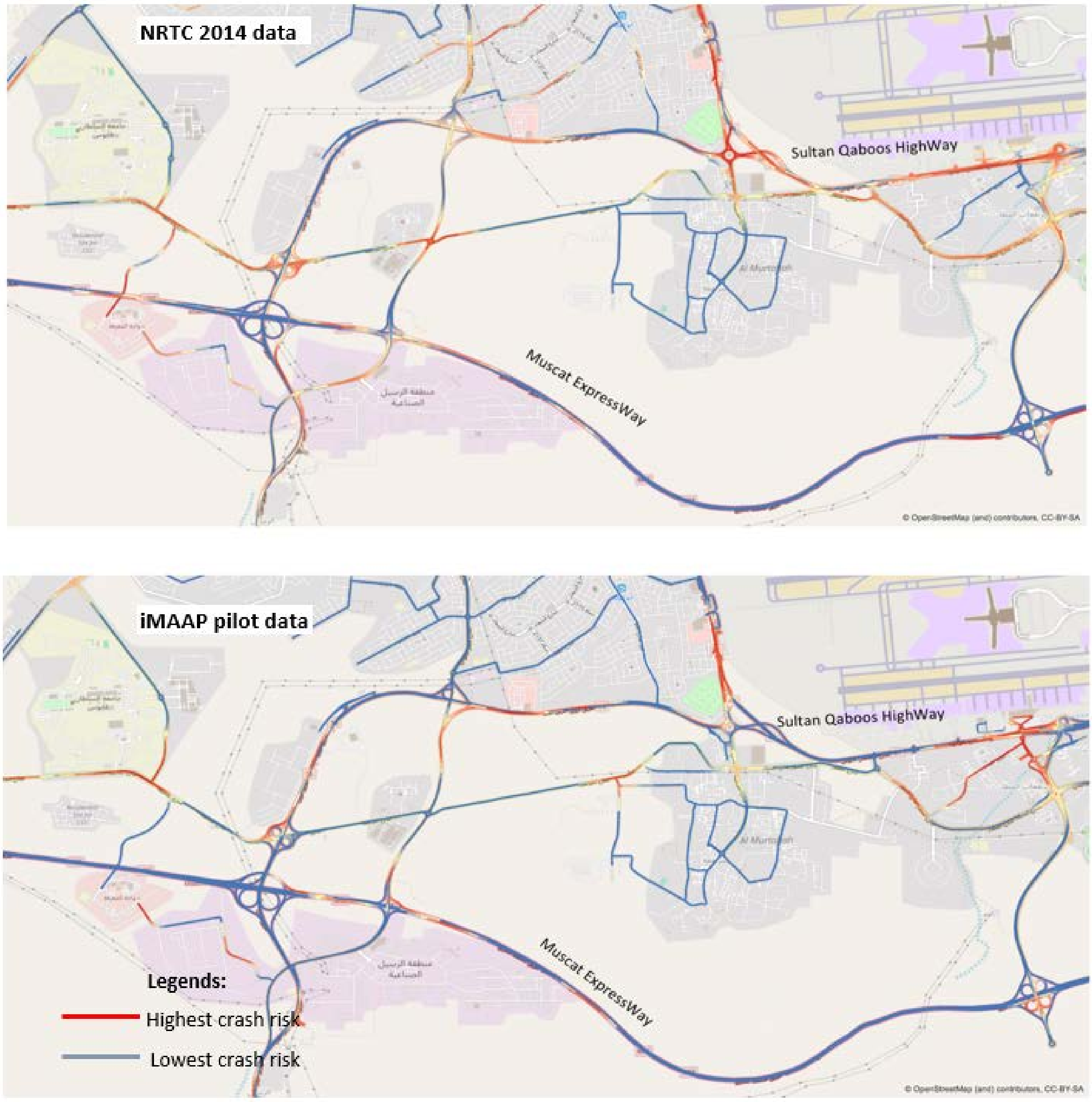

Figure 15*Comparison of Net-KDE results using the NRTC Data of the year 2014 and pilot data drawn from iMAAP System*

4. Discussion and conclusions

The identification of road crash hot-zones is pertinent to design the most effective strategies to reduce the crash density on high-risk areas. Our research aimed to: (1) identify high density crash zones in the Muscat Governorate (2) explore the characteristics of crash hot-zone, and (3) examine the spatio-temporal patterns of RTCs in the study area. It has exemplified the use of GIS in detecting hot-zones by employing a wide range of statistical techniques using data of five yeas (2010-2014) of RTCs in Muscat Governorate. The analysis highlights evidence of spatial clustering and recurrence of RTC hot-zones on long roads demarcated by intersections and roundabouts in Muscat. The study confirms that road intersections elevate the risk of RTCs than other effects attributed to road geometry features. To the best of our knowledge, the present research is the first of its kind in the Gulf Cooperation Council countries including Oman, where RTCs incur substantial impact on economic and health sectors. Our paper demonstrates the feasibility of applying robust spatio-temporal GIS modelling to limited aggregate data to identify the hot-zones and clustering of RTCs in the Muscat Governorate. The study contributes to a firmer quantitative evidence-base of RTCs hot zone maps over different locations on road networks, which can help policy decision makers to design appropriate, targeted interventions to reducing the burden of RTCs in Oman.

The hot-zones were identified with the help of Net-KDE and the significance of the results were established based on kernel density evaluated using the Net-NND and Net-K-function methods. Findings from the Net-KDE demonstrate evidence of spatial clustering of RTC hot-zones on roads demarcated by high number of intersections, complex bridges and roundabouts. Hot-zones appear to be more dominant on road segments where highest level of traffic interactions exists, especially along Sultan Qaboos highway. Conversely, low crash densities observed along roads locating outward from the core market and workplace areas, and Muscat Expressway is one example of such roads. Likewise, findings from the Net-NND and Net-K-function methods confirm the significance of clustering patterns of RTCs along road network in Muscat. These findings provide statistical evidence and confirm the research hypothesis that road intersections (roundabouts, crosses and bridges) represent the highest risk of causing RTCs than other road geometric features. These results were statistically validated using the pilot data from iMAAP network based crash analysis system.

Findings from RF algorithm and Wilcoxon tests indicate that road and traffic related features play a key role in determining locations of high crash risk. Based on the results from Wilcoxon tests, hot-zones are associated with higher level of road traffic. This result is in line with findings from past studies, which have proven that traffic level is a key factor associated with higher crash risk (Abdel-

Aty and Radwan, 2000; Caliendo et al., 2007; Jiang et al., 2016). Hot-zones also appear to be associated with the higher number of exits and entrances and shorter distance between junctions. This is consistent with results from previous literature, which concluded that higher density of intersections positively influences the number of crashes on such road sections (Siddiqui et al., 2012; Jiang et al., 2016; Khanh, Pei, and Liang-Tay 2019). Conversely, although past studies indicated that lower levels of posted speed limits associated with low crash-risk (Abdel-Aty et al., 2011; Jiang et al., 2016), results of the current study indicate that posted speed limits had no significant effect in determining the crash risk on road zones. This could be due to the low percentage of road sections with small posted speed limits (i.e. most of the road sections included in the study have posted speed limit>100 km/m) or it could be associated with existence of higher proportions of cyclists and pedestrians on road with low speed limits which in turn increase the likelihood of crash occurrence. However, the current study could not explore the interaction between speed limit and data related to number of cyclists and pedestrians due to lack of data. Overall, these findings offer new insights for road safety specialists to understanding the difference between hot-zones and other zones in Muscat Governorate, and thus helping them in adopting effective planning strategies and allocating proper resources to reduce the crash risk on these high density crash.

The spatio-temporal analysis provides evidences of the consistency in the positions of crash hot-zones in the study area. Comparing the mean of ranking of the same locations over five years of the study period, the results indicate that RTCs are inclined to cluster in the same locations within the study period. The inference of such result could be attributed to the same road and traffic related features existing on these zones. Therefore, further safety inspections and engineering studies should be carried out to investigate the possible contributing factors, and identify the potential countermeasures such as engineering improvements to reduce the crash risk at these sites.

It is important to highlight the data limitations of the present study. Unfortunately, the study could not disentangle other factors such as factors related to socio-economic, population, and land use factors due to lack of data. Socio-economic factors such as socioeconomic status, school enrolment density, number of automobile per households, number of non-retired workers per household, number of hotels are found to be significantly affecting the level of crash risk (Ng et al., 2002; Loukaitou-Sideris et al., 2007; Huang et al., 2010; Siddiqui et al., 2012; Wang et al., 2012). Having these data could help validate and improve our understanding of the spatial characteristics and the crash risk over different zones in the road network. However, the present study could not explore these factors because of lack of data.

Despite these limitations, our study demonstrates systematic quantitative evidence of spatio-temporal patterns associated with the crash risk over different locations on road network in Muscat. More importantly, the findings clearly pinpoint the importance and influence of the road and traffic related feature in road crash spatial analysis. It is recommended that future research should systematically address potential effects of the socio-economic, population, and land use factors in identifying road crash hot-zones in Oman.

References:

1. Abdel-Aty, M.A. and Radwan, A.E., 2000. Modeling traffic accident occurrence and involvement. Accident Analysis & Prevention, 32(5), pp.633-642.

2. Achu, A.L., Aju, C.D., Suresh, V., Manoharan, T.P. and Reghunath, R., 2019. Spatio-Temporal Analysis of Road Accident Incidents and Delineation of Hotspots Using Geospatial Tools in Thrissur District, Kerala, India. KN-Journal of Cartography and Geographic Information, 69(4), pp.255-265.

3. Aguero-Valverde, J. and Jovanis, P.P., 2006. Spatial analysis of fatal and injury crashes in Pennsylvania. Accident Analysis & Prevention, 38(3), pp.618-625.

4. Al-Rawas, M.A.S., 1993. Urban transportation problems in the Muscat area, Sultanate of Oman (Doctoral dissertation, University of Salford).

5. Anderson, T.K., 2009. Kernel density estimation and K-means clustering to profile road accident hotspots. Accident Analysis & Prevention, 41(3), pp.359-364.

6. Anderson, T.K., 2006. A Spatial road traffic collision hotspot typology for London. In IGU Regional Conference.

7. Benedek, J., Ciobanu, S.M. and Man, T.C., 2016. Hotspots and social background of urban traffic crashes: A case study in Cluj-Napoca (Romania). Accident Analysis & Prevention, 87, pp.117-126.

8. Breiman, L., 2001. Random forests. Machine learning, 45(1), pp.5-32.

9. Caliendo, C., Guida, M. and Parisi, A., 2007. A crash-prediction model for multilane roads. Accident Analysis & Prevention, 39(4), pp.657-670.

10. Cheng, W. and Washington, S.P., 2005. Experimental evaluation of hotspot identification methods. Accident Analysis & Prevention, 37(5), pp.870-881.

11. Choudhary, J., Ohri, A. and Kumar, B., 2015. Spatial and statistical analysis of road accidents hot spots using GIS. In 3rd Conference of Transportation Research Group of India (3rd CTRG).

12. Deshpande, N., Chanda, I. and Arkatkar, S.S., 2011. Accident mapping and analysis using geographical information systems. International Journal of Earth Sciences and Engineering, 4(6), pp.342-345.

13. Erdogan, S., 2009. Explorative spatial analysis of traffic accident statistics and road mortality among the provinces of Turkey. Journal of safety research, 40(5), pp.341-351.

14. Elvik, R., 1997. Evaluations of road accident blackspot treatment: a case of the iron law of evaluation studies?. Accident Analysis & Prevention, 29(2), pp.191-199.

15. Elvik, R., 2008. A survey of operational definitions of hazardous road locations in some European countries. Accident Analysis & Prevention, 40(6), pp.1830-1835.

16. Flahaut, B., Mouchart, M., San Martin, E. and Thomas, I., 2003. The local spatial autocorrelation and the kernel method for identifying black zones: A comparative approach. Accident Analysis & Prevention, 35(6), pp.991-1004.

17. Gundogdu, I.B., 2010. Applying linear analysis methods to GIS-supported procedures for preventing traffic accidents: Case study of Konya. Safety Science, 48(6), pp.763-769.

18. Harb, R., Yan, X., Radwan, E. and Su, X., 2009. Exploring precrash maneuvers using classification trees and random forests. Accident Analysis & Prevention, 41(1), pp.98-107.

19. Harirforoush, H., Bellalite, L. and Bénié, G.B., 2019. Spatial and temporal analysis of seasonal traffic accidents. American journal of traffic and transportation engineering, 4(1), pp.7-16.

20. Hashimoto, S., Yoshiki, S., Saeki, R., Mimura, Y., Ando, R. and Nanba, S., 2016. Development and application of traffic accident density estimation models using kernel density estimation. Journal of traffic and transportation engineering (English edition), 3(3), pp.262-270.

21. Huang, H., Abdel-Aty, M. and Darwiche, A., 2010. County-level crash risk analysis in Florida: Bayesian spatial modeling. Transportation Research Record: Journal of the Transportation Research Board, (2148), pp.27-37.

22. Ivan, K. and Haidu, I., 2012. The spatio-temporal distribution of road accidents in Cluj-Napoca. Geographia Technica, 2, pp.32-38.

23. Ivan, K., Haidu, I., Benedek, J. and Ciobanu, S.M., 2015. Identification of traffic accident risk-prone areas under low-light conditions. Natural Hazards and Earth System Sciences, 15(9), pp.2059-2068.

24. Jiang, X., Abdel-Aty, M., Hu, J. and Lee, J., 2016. Investigating macro-level hotzone identification and variable importance using big data: A random forest models approach. Neurocomputing, 181, pp.53-63.

25. Kaygisiz, Ö., Yildiz, A. and Duzgun, S., 2015. Spatio-Temporal Pedestrian Accident Analysis to Improve Urban Pedestrian Safety: The case of the Eskisehir Motorway. Gazi University Journal of Science, 28(4), pp.623-630.

26. Le, K.G., Liu, P. and Lin, L.T., 2019. Determining the road traffic accident hotspots using GIS-based temporal-spatial statistical analytic techniques in Hanoi, Vietnam. *Geo-spatial Information Science*, pp.1-12.

27. Loukaitou-Sideris, A., Liggett, R. and Sung, H.G., 2007. Death on the crosswalk: A study of pedestrian-automobile collisions in Los Angeles. Journal of Planning Education and Research, 26(3), pp.338-351.

28. Loo, B.P. and Anderson, T.K., 2015. Spatial Analysis Methods of Road Traffic Collisions. CRC Press.

29. Louppe, G., 2014. Understanding random forests: From theory to practice. arXiv preprint arXiv:1407.7502.

30. Mahmud, A.R., Sohadi, R., Umar, R. and Mansor, S., 1998. A GIS support system for road safety analysis and management. Pertanika Journal of Science & Technology, 6(1), pp.81-93.

31. Mohaymany, A.S., Shahri, M. and Mirbagheri, B., 2013. GIS-based method for detecting high-crash-risk road segments using network kernel density estimation. Geo-spatial Information Science, 16(2), pp.113-119.

32. Montella, A., 2010. A comparative analysis of hotspot identification methods. Accident Analysis & Prevention, 42(2), pp.571-581.

33. Moridpour, S., and Toran, A., 2015. Identifying crash black spots in Melbourne road network using Kernel Density Estimation in GIS.

34. Mutanga, O., Adam, E. and Cho, M.A., 2012. High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. International Journal of Applied Earth Observation and Geoinformation, 18, pp.399-406.

35. Kaygisiz, O. and Sümer, N., 2019. Effects of fixed speed cameras on spatio-temporal pattern of traffic crashes: Ankara case. Journal of Transportation Safety & Security, pp.1-19.

36. Pleerux, N., 2020. Geographic Information System-based Analysis to Identify the Spatiotemporal Patterns of Road Accidents in Sri Racha, Chon Buri. CURRENT APPLIED SCIENCE AND TECHNOLOGY, 20(1), pp.59-70.

37. NCSI, 2018. Total population registered by governorate and nationality. National Centre for Statistics and Information, Monthly Statistical Bulletin 29, March 2018.

38. NCSI, 2016. Statistical Year Book 2016: Issue 44. National Centre for Statistics and Information, Annual Statistical Bulletin, July 2016.

39. Ng, K.S., Hung, W.T. and Wong, W.G., 2002. An algorithm for assessing the risk of traffic accident. Journal of safety research, 33(3), pp.387-410.

40. Okabe, A., Okunuki, K.I. and Shiode, S., 2006. SANET: a toolbox for spatial analysis on a network. Geographical analysis, 38(1), pp.57-66.

41. Okabe, A., Satoh, T. and 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(1), pp.7-32.

42. Okabe, A. and Sugihara, K., 2012. Spatial analysis along networks: statistical and computational methods. John Wiley & Sons.

43. Okabe, A. and Yamada, I., 2001. The K-function method on a network and its computational implementation. Geographical Analysis, 33(3), pp.271-290.

44. Prasannakumar, V., Vijith, H., Charutha, R. and Geetha, N., 2011. Spatio-temporal clustering of road accidents: GIS based analysis and assessment. Procedia-Social and Behavioral Sciences, 21, pp.317-325.

45. Qin, X., Parker, S., Liu, Y., Graettinger, A.J. and Forde, S., 2013. Intelligent geocoding system to locate traffic crashes. Accident Analysis & Prevention, 50, pp.1034-1041.

46. Rahman, M.K., Crawford, T. and Schmidlin, T.W., 2017. Spatio-temporal analysis of road traffic accident fatality in Bangladesh integrating newspaper accounts and gridded population data. GeoJournal, pp.1-17.

47. Royal Oman Police, 2017. Facts and Figures GCC Traffic Week 2016, Director General of Traffic.

48. Siddiqui, C., Abdel-Aty, M. and Choi, K., 2012. Macroscopic spatial analysis of pedestrian and bicycle crashes. Accident Analysis & Prevention, 45, pp.382-391.

49. Siddiqui, C., Abdel-Aty, M. and Huang, H., 2012. Aggregate nonparametric safety analysis of traffic zones. Accident Analysis & Prevention, 45, pp.317-325.

50. Spooner, P.G., Lunt, I.D., Okabe, A. and Shiode, S., 2004. Spatial analysis of roadside Acacia populations on a road network using the network K-function. Landscape ecology, 19(5), pp.491-499.

51. Truong, L.T. and Somenahalli, S.V., 2011. Using GIS to identify pedestrian-vehicle crash hot spots and unsafe bus stops. Journal of Public Transportation, 14(1), p.6.

52. Vandenbulcke, G., Thomas, I. and Panis, L.I., 2014. Predicting cycling accident risk in Brussels: a spatial case–control approach. Accident Analysis & Prevention, 62, pp.341-357.

53. Vandenbulcke, G., Int Panis, L. and Thomas, I., 2017. On the location of reported and unreported cycling accidents: A spatial network analysis for Brussels. Cybergeo: European Journal of Geography.

54. Wang, X., Jin, Y., Abdel-Aty, M., Tremont, P. and Chen, X., 2012. Macrolevel model development for safety assessment of road network structures. Transportation Research Record: Journal of the Transportation Research Board, (2280), pp.100-109.

55. Xie, Z. and Yan, J., 2008. Kernel density estimation of traffic accidents in a network space. Computers, environment and urban systems, 32(5), pp.396-406.

56. Yao, S., Wang, J., Fang, L. and Wu, J., 2018. Identification of vehicle-pedestrian collision hotspots at the micro-level using network kernel density estimation and random forests: A case study in Shanghai, China. Sustainability, 10(12), p.4762.

57. Young, J. and Park, P.Y., 2014. Hotzone identification with GIS-based post-network screening analysis. Journal of Transport Geography, 34, pp.106-120.

58. Yu, H., Liu, P., Chen, J. and Wang, H., 2014. Comparative analysis of the spatial analysis methods for hotspot identification. Accident Analysis & Prevention, 66, pp.80-88.

Authors and Contributors

Amira Al Aamri (AAA) and Sabu S. Padmadas (SSP) designed the study and prepared the initial draft. AAA prepared the dataset for research, conducted the literature review and led the statistical analysis with support and supervision from Graeme Hornby (GH), Li-Chun Zhang (LCZ) and SSP. Abdullah Al Maniri (AAM) secured access to data, contributed to the interpretations and revised the paper for intellectual content. AAA, SSP and LCZ conducted the final review and revised the manuscript for submission. All authors read and approved the final version of the article before submission.

Declaration of interests

All authors declare that there is no conflict of interest whatsoever with regard to the submission of this manuscript.

Acknowledgements

We are grateful to The Research Council and the Royal Oman Police for providing us with access to the national road traffic accident database. We acknowledge the Ministry of Higher Education, Sultanate of Oman for funding the research project.

List of tables and figures