19th International Congress on Modelling and Simulation, Perth, Australia, 12–16 December 2011
http://mssanz.org.au/modsim2011

Data Mining of Driver Characteristics to Spatial and Temporal Hotspots of Single Vehicle Crashes in Western Australia

Jianhong (Cecilia) Xia

Department of Spatial Sciences, Curtin University, Australian
Email: c.xia@curtin.edu.au

Abstract: This paper presents innovative methods for identifying the characteristics of drivers involved in single vehicle crashes (SVCs) in Western Australia when viewed spatially and temporally. Spatial and temporal hotspots of SVCs can be defined as clusters of SVCs that have exceeded the expected value over a certain time period and at certain locations. The EM (Expectation-Maximisation) algorithm was adopted to identify the characteristics of driver involved in vehicle crashes. Drivers were divided into different segments based on their socio-demographic characteristics, i.e., age and gender and driver related crash factors such as, drive license's types, alcohol consumption, speeding, fatigue and inattention etc. The spatial hotspots of the SVCs were identified using the Kernel Density Estimation method. Comap was used to integrate space, time and characteristics of drivers into one view to better understand the nature of the SVCs.

The main conclusion draws from this study are as follows:

- Characteristics of drivers subject to SVCs Change over time with different contributing factors.
- Some subgroups of drivers at greater risk were identified.
    - Females aged around 19, with BACs around 0.03%, on probationary licences were more likely to be involved in crashes between the hours of midnight and 2:10am and require hospital attention.
    - Males aged around 61, on full licences were more likely to be involved in crashes between the hours of 7:55am and 10:55am because of inattention and require medical or hospital treatment.
    - Males, and some females, aged around 26, on full licences, and with BACs around 0.07% were more likely to be involved in crashes between the hours of 2:10am and 4:55am because of alcohol, inattention, and speeding, and require hospital treatment.
    - Males, and some females, aged around 36 were more likely to be involved in crashes between the hours of 10:55am and 17:40am because of inattention and require medical attention.
    - Males and females aged around 17 on probationary licences were more likely to be involved in crashes between the hours of 17:40pm and 20:55pm because of inattention, and require medical attention.
- Alcohol and speeding were found to be the main contributing factors in SVCs at night or early morning, and inattention was the main factor in SVCs during the day.
- Severity of SVCs is generally higher at night than during the day.
- The distribution of hotspots of SVCs varies over different time periods

The results of this research will help road safety management authorities better understand the underlying nature of why crashes happen and where and when they do so that remedial action can take into account the characteristics of drivers.

Keywords: *Vehicle crash, spatial and temporal hotspot, cluster and outlier analysis, driver characteristics, expectation-Maximisation (EM) algorithm*

1. INTRODUCTION

The rate and severity of vehicle crashes have consistently been found to be associated with certain driver characteristics. For example, Bédard *et al.* (2002) found that male drivers younger than 30 years dominate the fatality records. However, older drivers are more likely to be fatally injured as a result of the actions of other drivers than are younger drivers (Boufous *et al.*, 2008). Vaez and Laflamme (2005) extended age and gender to other demographic attributes such as class of origin and educational attainment and suggest some at-risk populations are "males, persons aged 18–19, those from households classified as "workers" and "others" (including, e.g. the long-term unemployed and those on long-term sick leave), and those with low educational attainment". Income and race were also studied by Lerner *et al*.(2001) to understand their influence on seatbelt use by adults injured in motor vehicle crashes. The results showed that older persons, women, Caucasians, and individuals with greater incomes tended to wear seat belts more. However, Dougherty, Pless and Wilkins (1990) found that Montreal neighbourhoods with the lowest quintile of income levels had four-times as many fatal vehicle crashes involving children than other more advantaged neighbourhoods. Clifton *et al.* (2009) confirmed these previous demographic studies, but also discovered that lighting conditions, the time of the day, transit access, and pedestrian connectivity are associated with injury severity as a result of crashes. This research tends to suggest that location matters in vehicle crashes. Further, Plug *et al.* (2011) applied different spatial and temporal visualisation techniques to understand the process underlying single vehicle crash patterns and their causes and showed that time as well as location influence vehicle crashes.

According to Tobler's first law of Geography: "Everything is related to everything else, but near things are more related than those far apart" (Tobler, 1970). When applied to vehicle crashes, those in close proximity to each other probably have similar causes. For example, Hall and Zador (1981) found clusters of crashes— hotspots— occurred on certain types of roads because of the geometry of the roads, such as severe horizontal curves (average 1.7 degree at crash sites). Similarly, at certain hotspots, the driver characteristics may be a factor.

Based on this theory, this paper aims to identify whether the characteristics of drivers involved in vehicle crashes change across different locations and over time. Comap will be used to identify spatial and temporal patterns of vehicle crashes, and the EM (Expectation-Maximisation) algorithm will be used to identify characteristics of drivers associated with these spatial and temporal patterns of vehicle crashes.

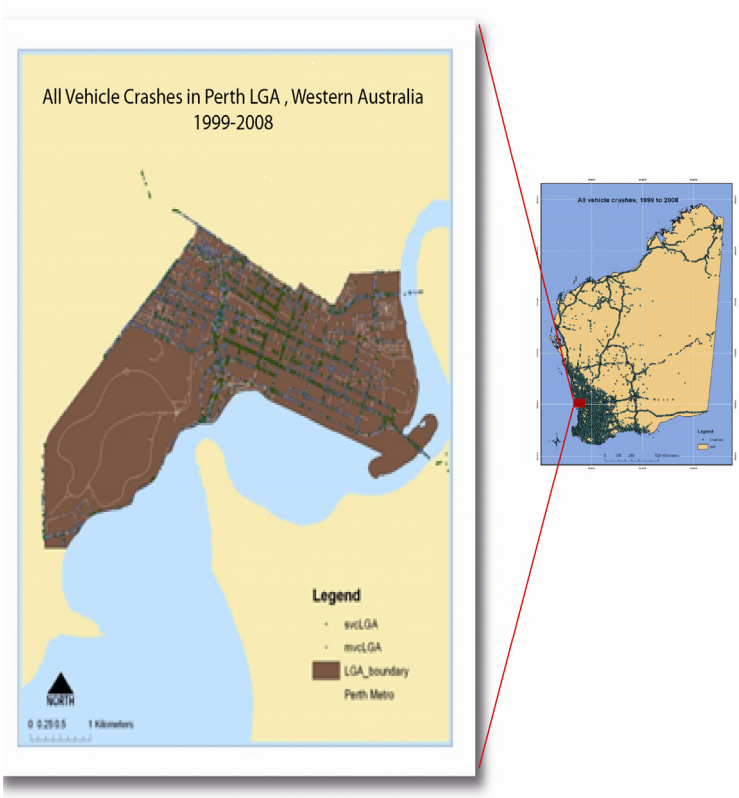


Figure 1. The study area

2. METHODS

2.1. Data collection and study area

The data used in this paper were obtained from a database maintained by the Main Roads Department in Western Australia and which contains information about all police-reported vehicle crashes in Western Australia. According to Plug *et al.* (2011), the Perth Local Government Area (LGA) is a better scale to investigate hotspots spatially. Therefore, we concentrated on the characteristics of drivers involved in single

vehicle crashes (SVCs) between 1999 and 2008 in this study area. The study area and the distribution of all crashes are illustrated in Figure 1. Table 1 summarises the percentage frequency of single vehicle crashes in the Perth LGA from 1999 to 2008. The nature values 7 to 10 represent SVCs.

Table **1**. Summary of SVCs and MVCs in Perth LGA

Value	Crashes by Nature	Frequency	% of Total
0		2089	0.09
7	Hit Pedestrian	724	0.03
8	Hit Animal	1	0.00
9	Hit Object	848	0.04
10	Non Collision	181	0.01

2.2. Characteristics of drivers

The socio-demographic variables included in the analyses are shown in Table 2.

Table 2: Socio-demographic variables.

Variable	Range and Notes.
Sex	Male or female
Age	The age range in the data set was from 9 to 94 years old.
Protection	Seat belt worn or not
Driver's Licence Type	Licenced, suspended, learner, cancelled, unlicenced, probationary
Injury type	Death within 30 days of accident; injured and requiring medical attention but not in hospital; injured and admitted to hospital; and no medical attention required (Main Roads Western Australia, 2011).
Blood Alcohol Content	The blood-alcohol-content (BAC) range in the data set was from 0.0 to 1.57. Under Western Australian law, drivers with a full licence may be issued an infringement notice if their BAC exceeds 0.05%; for drivers with a probationary licence, it is 0.02% (Western Australia Police, 2011).
Accident Cause (as assessed by the attending police officers)	Speeding, fatigue, inattention, alcohol

2.3. Kernel Density Estimation

The Kernel Density Estimation technique (KDE) is a well-established method for identifying hotspots from point data (see O'Sullivan and Unwin, 2003 for detailed examples). It creates a field representation of hotspots using different colour schemes. The density of points for each pixel in the output raster is estimated by counting the number of points in a defined region or kernel, divided by the area of the kernel. Thus, a discrete density surface is made continuous by interpolation and, for example, darker colours highlight locations of high intensity. For this study, KDE was used to estimate the density of SVCs (or hotspots) in the Perth LGA. The Spatial Analyst KDE tool in ESRI's ArcGIS 10 software was used to calculate the KDE (ESRI, 2010).

2.4. Comap

Comap, a visualisation tool, was developed from a concept of coplot (Tufte, 1990), which is based on principle of dividing a whole into subsets, and then plotting the subset data to investigate differences between individual subsets. Comap has been used to investigate spatial and temporal patterns of phenomenon such as disease, crime or vehicle crashes (Asgary *et al.*, 2010; Plug et al., 2011). A Comap is created this way:

1. Run the classification of time point data of the vehicle crashes using a natural break scheme (Jenks, 1967). This method is designed to arrange time points into a group with minimum deviation from the group mean, while seeking to maximise the deviation from mean of the other groups. Therefore the time points in the same group should have a similar nature.

2. Break the vehicle crash data into *eight* subsets based on the classified time intervals because of the structure of the Comap.
3. Use the KDE function to identify spatial patterns of vehicle crashes.
4. Create maps of the spatial pattern of vehicle crashes for each subset.
5. Illustrate this sequence of maps with defined time intervals using the Comap structure.

In this study, we applied Comap to identify changes in the SVC spatial patterns over different time intervals.

2.5. EM algorithm (Expectation and Maximisation)

The EM algorithm (Dempster *et al.*, 1977) is a model-based clustering method to build certain models for clusters based on parametric distributions, like a Gaussian (continuous) or a Poisson (discrete). Each individual cluster can be represented by a distribution. The whole data set is a mixture of these individual distributions. The goal of the algorithm is to optimise the fit between the models and data, which means that each distribution can capture dominant patterns of data. An advantage of the EM algorithm is that it can be used for both numeric (continuous) and categorical (discrete) data. Detailed information regarded to EM algorithm can be found from Xia et al (2010). This paper used the EM algorithm to identify characteristics of divers based on their social-demographic variables defined in section 2.2. Age and BAC are continuous variables and other variables are discrete. Drivers were grouped into the same class if they have similar characteristics. The number of groups of drivers will be decided by the nature of data. More than one group of drivers could be identified for a time interval.

3. RESULTS

Figure 2 illustrates the spatial and temporal pattern of single vehicle crashes by Comap. The time points of the SVCs were classified into eight time intervals using Jenks Natural break classification.

Figure 3 confirms that the classification results are correct. Each time interval covers one or two spaced peaks. In Figure 2, each bar at the top represents the length of a time interval marked by the end of a time instant. For example, the first time interval is between midnight and 2:10. Each map illustrates the spatial

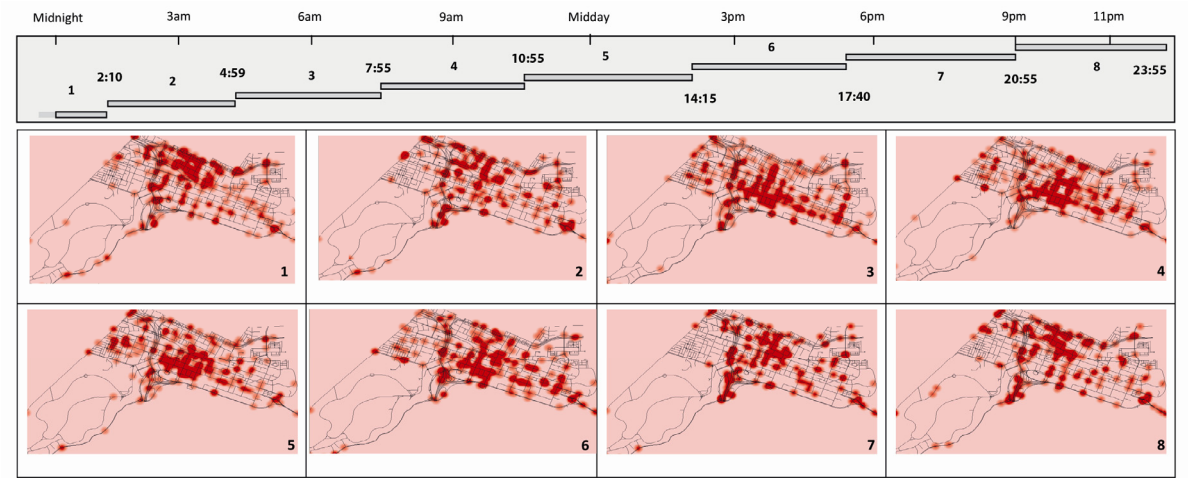


Figure 2. The Comap of the SVCs. in the Perth LGA from 1999 to 2008

distribution of the SVCs in Perth LGA within the classified time interval given by the same numbering system as time bar. For example, Map 1 shows the distribution of SVCs between midnight and 2:10am. Generally, hotspots of SVCs occur more frequently around intersections and along some major streets. However, they vary at different locations during different time intervals, which are probably due to variations of pedestrians activities.

The characteristics of drivers also change over these time intervals. For example, males are generally more likely to be involved in SVCs than females. However, more females aged around 19 tend to be severely injured at night where the crash is attributed to alcohol (BAC around 0.03%), speeding, or inattention. Males aged around 30 are more likely to speed between midnight and 2:10am and be involved in crashes that require hospital attention (see Figure 4).

Between 2:10am and 4:59am, most of those involved in SVCs were aged around 26 and alcohol, speeding, and inattention are the main causes. The injuries tend to be more serious and most drivers have to attend hospital. Males are four times more likely to be involved in SVCs than females during this time interval, but females aged around 22 and with BACs around 0.07% tend to be seriously injured.

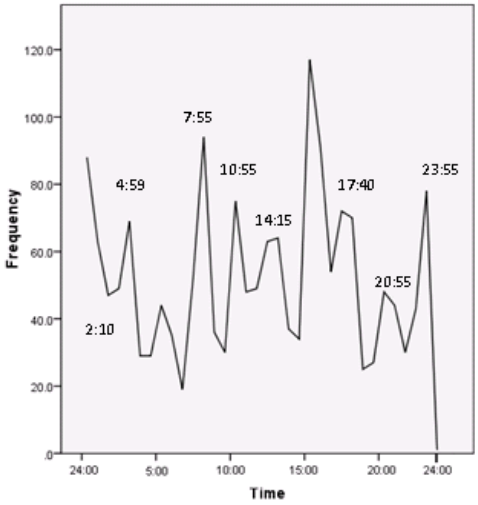
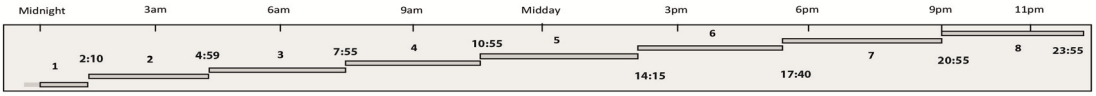


Figure 3. The frequency of the SVCs over time

During the early morning interval, between 5:00am and 7:55am, the severity of injuries decreases, but speeding and alcohol are still the main causes. However, the reported BACs were lower than for the immediately prior time interval, meanwhile, age and gender types were similar for the two periods.

Between 7:55am and 10:55am, most SVCs are caused by inattention and resulted in injuries requiring medical attention. Three vulnerable groups were identified: young driers aged around 17, male drivers aged around 61, and drivers aged around 35.



1	2	3	4
1. Age around 23, Alcohol, Medical or Hospital, Full-licenced (9, 13%). 2. Male, Age around 30, Speeding, Hospital, full-licenced (41, 60%). 3. Female, Age around 19. Alcohol (0.03), Probationary, Hospital (18, 26%). Log likelihood: -3.76187	1. Female and some Male, age around 22, Full licenced or probationary, Inattention, speeding alcohol (0.07), Hospital and Medical (18, 27%). 2. Male and some Female, Age around 26. Full-licenced, Inattention, speeding, alcohol (0.07), Hospital (43, 64%) 3. Male, age around 43, Full-licenced, inattention, alcohol (0.06), hospital (6, 9%) Log likelihood: -3.81437	1. Male, Age around 41, Speeding, Medical. (9, 16%), 2. Male and Female, Age around 23, Full licenced or probationary , Speeding, alcohol (0.07), Medical. (10 18%). 3. Male and Female, Age around 32, Full-licenced, Speeding , Alcohol (0.05), Medical (36, 65%). Log likelihood: -3.72093	1. Male and Female, age around 17, Probationary, Inattention, Medical (7, 8%) 2. Male and Female , age around 35, Full-licenced, Inattention , Medical (68, 76%) 3. Male, age around 61, Full-licenced, Inattention , Medical or hospital (15, 17%) Log likelihood: -5.44121
5	**6**	**7**	**8**
Male and some Female, Age around 36, Inattention, Medical (92, 100%). Log likelihood: -5.51377	Female and Male, Age around 37, Inattention, Medical (137, 100%). Log likelihood: -5.34139	1. Male and Female, age around 28, Inattention, Full-licenced, Medical (52,73%) 2. Male , age around 50, Full-licenced, Inattention, Medical or hospital (15,21%) 3. Male and Female, age around 17, Inattention, Probationary or Full-licenced, Medical (52,73%) Log likelihood: -4.89877	1. Male, Age around 33, full-licenced, alcohol (0.02), Medical (38, 52%). 2. Female and some male, Age around 21, Probationary or Full-licenced, Alcohol (0.02) , Medical (30, 41%). 3. Male, Age around 33, full-licenced, speeding, alcohol (0.14), Medical (5, 7%). Log likelihood: -4.33124

Figure 4. The Characteristics of drivers

Between 10:55am and 14:15pm and 14:15pm and 17:40pm, drivers involved in SVCs are mainly aged around 36 or 37, required medical attention, and the cause was attributed to inattention. As Evening wore on, younger and older drivers became involved in SVCs and alcohol and speeding re-emerged as the cause.

4. DISCUSSIONS AND CONCLUSIONS

The EM algorithm is widely used in clustering analysis (Laird, 2010; Xia et al., 2010) and has the advantage of being suited to both numeric and categorical data. So it was adopted for this study. However, there are two issues that arise when the EM algorithm is applied in practice:

- Because the initialisation is random, the algorithm may converge to local likelihood maxima instead of the global maximum. This can be avoided by running the algorithm many times, and choosing the model results that correspond to the global maximum.
- The EM algorithm cannot be used to compute the number of clusters because the likelihood of a mixture model always increases as the number of clusters (and number of parameters) increases. To select the optimal number of clusters, the likelihood must be penalised as the number of model parameters.

Comap has been used in this study to illustrate the spatial and temporal distribution of SVCs in the Perth LGA and the variation of characteristics of drivers subject to SVCs over time. Further research will look at the characteristics of drivers at some major hotspots instead of the whole region to get a better idea of the true spatial, temporal, and characteristics of drivers' 3D interactions.

The time clustering method used in this study is Jenks natural break classification scheme. It proved to be an effective classification method to break down time point into different intervals based on their frequency distribution. We limited the class number to eight because of Comap restrictions, but in practice, the number of classes should be determined by the nature of the data. Therefore, in the future, we will investigate better classification methods and modify the Comap method to make it more flexible.

The main conclusion draws from this study are as follows:

- The characteristics of drivers involved in SVCs change over space and time with different contributing factors.
- Some subgroups of drivers at greater risk were identified:
    - Females aged around 19, with BACs around 0.03%, on probationary licences were more likely to be involved in crashes between the hours of midnight and 2:10am and require hospital attention.
    - Males aged around 61, on full licences were more likely to be involved in crashes between the hours of 7:55am and 10:55am because of inattention and require medical or hospital treatment.
    - Males, and some females, aged around 26, on full licences, and with BACs around 0.07% were more likely to be involved in crashes between the hours of 2:10am and 4:55am because of alcohol, inattention, and speeding, and require hospital treatment.
    - Males, and some females, aged around 36 were more likely to be involved in crashes between the hours of 10:55am and 17:40am because of inattention and require medical attention.
    - Males and females aged around 17 on probationary licences were more likely to be involved in crashes between the hours of 17:40pm and 20:55pm because of inattention, and require medical attention.
- Alcohol and speeding were found to be the main factors associated with SVCs at night or early morning, and inattention was the main factor in SVCs during the day.
- The severity of SVCs was generally higher at night than during the day.
- The distribution of SVC hotspots moved over time.

In the future, we will conduct similar analysis for different types of SVCs, such as hit-pedestrian crashes or multi-vehicle crashes. It will be interesting to identify the characteristics of drivers in relation to vehicle

crashes at major hotspots and especially include some environmental factors, such as weather conditions, in the analysis. For this study, we chose the Perth LGA, but further research will be undertaken in other areas, such as tourist towns or residential and industrial areas, to see if the same results hold.

ACKNOWLEDGMENTS

We wish to acknowledge the assistance of Main Roads WA in providing access to the WA road crash data files needed for this study. Thanks are extended to Craig Caulfield for editing the paper, Matthew Legge for discussing some issues related to the paper and two reviewers for their comments that help improve the manuscript.

REFERENCES

Asgary, A., Ghaffari, A., & Levy, J. (2010). Spatial and temporal analyses of structural fire incidents and their causes: A case of Toronto, Canada. *Fire Safety Journal, 45*(1), 44 - 57-44 - 57.

Bédard, M., Guyatt, G. H., Stones, M. J., & Hirdes, J. P. (2002). The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. *Accident Analysis & Prevention, 34*(6), 717-727.

Boufous, S., Finch, C., Hayen, A., & Williamson, A. (2008). The impact of environmental, vehicle and driver characteristics on injury severity in older drivers hospitalized as a result of a traffic crash. *Journal of Safety Research, 39*(1), 65-72.

Clifton, K. J., Burnier, C. V., & Akar, G. (2009). Severity of injury resulting from pedestrian-vehicle crashes: What can we learn from examining the built environment? *Transportation Research Part D: Transport and Environment, 14*(6), 425-436.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. . *Journal of the Royal Statistical Society, Series B, 34*, 1-38.

Dougherty, G., Pless, I. B., & Wilkins, R. (1990). Social class and the occurrence of traffic injuries and deaths in urban children. *Canadian Journal of Public Health, 81*(3), 204-209.

ESRI. (2010). ArcGIS 10. New York: ESRI.

Jenks, G. F. (1967). The Data Model Concept in Statistical Mapping. *International Yearbook of Cartography*(7), 186-190.

Hall, J. W., & Zador, P. (1981). A survey of single vehicle fatal rollover crash sites in New Mexico. *Transportation Research Record* (819), 1-8.

Laird, N. M. (2010). The EM Algorithm in Genetics, Genomics and Public Health. *Statistical science, 25*(4), 450-457.

Lerner, E. B., Jehle, D. V. K., Billittier, A. J., Moscati, R. M., Connery, C. M., & Stiller, G. (2001). The influence of demographic factors on seatbelt use by adults injured in motor vehicle crashes. *Accident Analysis & Prevention, 33*(5), 659-662.

Main Roads Western Australia. (2011). *Vehicle crash dictionary*.

O'Sullivan, D., & Unwin, D. J. (2003). *Geographic Information Analysis*: John Wiley & Sons, Inc.

Plug, C., Xia, J., & Caulfield, C. (2011). Spatial and temporal visualisation techniques for crash analysis. *Accident Analysis & Prevention, In Press, Corrected Proof*.

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography, 46*, 234-240.

Tufte, E. R. (1990). *Envisioning information*. Cheshire, Connecticut: Graphics Press.

Vaez, M., & Laflamme, L. (2005). Impaired driving and motor vehicle crashes among Swedish youth: An investigation into drivers' sociodemographic characteristics. *Accident Analysis & Prevention, 37*(4), 605-611.

Western Australia Police. (2011). Drink driving penalities. Retrieved 2 July, 2011, from http://www.police.wa.gov.au/Traffic/Drinkdriving/Penalties/tabid/989/Default.aspx

Xia, J., Evans, F. H., Spilsbury, K., Ciesielski, V., Arrowsmith, C., & Wright, G. (2010). Market segments based on the dominant movement patterns of tourists. *Tourism Management, 31*(4), 464-469.