COMP20008
Elements of Data Processing

Assignment #2 - Data Science Project

Xing Yang Goh, Chun Yu Shih, Aryan Shahi, Xiuqi Xiao
1001969, 111839, 1170385, 1008446 respectively
May 19, 2021

1 Introduction

Over 10,000 Victorians have been killed in road accidents since 1989, with thousands more being hospitalised (Transport Accident Commission, 2020). Road safety and awareness of road hazards are crucial for avoiding casualties in the future, forming the basis of this study. The research question being addressed in this report is: *What is the appropriate speed limit for each intersection geometry in Victoria based on the distribution of severe to fatal car crashes in different speed-zones?*

The study aims to improve the livability, sustainability and health of the Victorian community. By visually presenting the correlation between speed zones and severe crashes that occurred at intersections of four main geometries (Cross, Multiple, T, Y), it highlights the existing road safety issues. The study helps to raise safety awareness and encourage better driving practices among Victorian residents. Moreover, the result is a meaningful reference for the relevant government departments such as VicRoads for addressing current road issues and improving infrastructure planning. Despite the available data on road crash incidents, this study correlates the data to provide more meaningful and intuitive information for the stakeholders.

The two datasets used in this investigation are INTERSECTION_SEVERE and ALL_CRASHES. Both of these files are in CSV format and are provided by VicRoads. The ALL_CRASHES file was reduced from 178.6MB to 47.9MB by removing columns containing excess information to comply with the file size limit in the Jupiter Hub server. INTERSECTION_SEVERE is 62MB in size and contains only details for the severe to fatal crashes (59,495 cases total) that occurred from 2014-2019. ALL_CRASHES contains details including speed zones and road geometries of all the known crashes (171,349 cases total) from 2014-2019. The study assumes that the file contains all unique intersections in Victoria due to the large number of cases it covers. Therefore, the count of unique intersections of each geometry in Victoria is estimated using ALL_CRASHES and is used to normalise the number of severe to fatal crashes in different speed zones, obtained from INTERSECTION_SEVERE.

2 Methodologies

Initially, pre-processing is performed on both INTERSECTION_SEVERE and ALL_CRASHES datasets to remove unnecessary information, such as drivers' genders. Each dataset is separated into 4 CSV files, with each file corresponding to an intersection geometry (Cross, Multiple, T, Y) creating 8 new files in total. Next, the number of unique intersections of each geometry is obtained by removal of cases with same Map IDs. Using Accident Numbers, duplicate records of accidents are removed from ALL_CRASHES's 4 sub-files.

The study investigates cases with speed zones ranging from 30 km/h to 110 km/h in intervals of 10. Other cases labelled as 'Unknown' or 'Other' speed zones are removed. The 4 sub-files from INTERSECTION_SEVERE were used to generate the count of severe to fatal crashes by extracting the speed-zone data from the data frame into a series (using .squeeze() function) and doing .value_counts(). This produced a data frame showing the number of crashes for each intersection geometry for each speed limit. Then, the 4 sub-files generated from ALL_CRASHES were used to generate a count of unique intersections for each intersection geometry for each speed limit in a similar manner.

An initial Normalised Mutual Information (NMI) calculation is performed between the speed-zone (Feature) and severity of the crash (Class label) in the ALL_CRASHES dataset with sklearn to measure the correlation. Although the speed-zones are already separated into speed-zone bins from 30-110, a further binning process into 3 equal width bins is performed to ensure there are sufficient data points in each speed-zone feature bin as speed-zones such as 30 and 110 have insufficient data points. The severity of the crash is also separated into 3 bins, with minor, severe and fatal class labels. Due to the nature of the speed-zones frequency distribution, the 'average' bin of [60, 70, 80] will have the greatest number of data points, and differences in the size of datasets between the intersections geometries could result in difficulties obtaining a clear correlation and comparisons between the intersections. To combat this, an iterative stratified sampling method to calculate an estimated NMI is performed. This method generates smaller samples using the adaptive partitioning approach

$$D = \sqrt{\frac{n}{5}}$$

that relates the number of bins in x and y as D (3 in our case) and the sample size n (Cellucci, Albano, & Rapp, 2005). This provides an optimal sample size of 45, where the samples are chosen using stratified sampling, taking 15 data-points from each speed-zone class and calculating an NMI value from this new dataset. This process is iterated 1000 times, changing the random state with each iteration and finally calculating the average NMI score.

For visualisation, the normalised number of severe and fatal cases are plotted against the speed zones in a scatter plot. The normalised points are calculated by dividing the severe to fatal crashes for each intersection geometry and speed limit by their respective unique counts, then plotting each intersection geometry as a scatter-plot with normalised crashes on the y axis and speed-zones on the x axis. A minimum threshold value of 20 severe crashes was imposed when creating the scatter-plot to avoid unreliable results from random variations. A linear regression using the least-squares method is then fitted onto the scatter-plots to visualise the correlation between them.
Other common correlation analysis techniques such as the Pearson correlation is not used as the the data are discrete. Moreover, there is no indication that the correlation is linear. In terms of visualisation, techniques like k-mean clusters were not used as the clusters will be susceptible to the raw frequency of severe crashes and will not take into account the unique counts for normalisation, which would lead to clusters in regions where the most common speed-zone like occurs.

3 Results

	Cross Intersection	Multiple Intersection	T Intersection	Y Intersection
30	1	1	1	nan
40	45	1	35	1
50	226	11	262	1
60	640	63	765	2
70	160	24	181	nan
80	366	78	427	3
90	4	nan	15	nan
100	273	19	267	7
110	13	nan	4	nan

Figure 1: Severe to fatal crashes frequency count for different speed-zones and intersections

	Cross Intersection	Multiple Intersection	T Intersection	Y Intersection
30	3	1	5	nan
40	203	20	304	8
50	1138	96	1529	20
60	1729	232	2762	35
70	252	34	523	5
80	487	95	809	15
90	10	nan	19	nan
100	319	44	526	18
110	14	2	13	nan

Figure 2: Unique Intersection frequency count for different speed-zones

	NMI Entire Dataset	NMI Iterative Stratified Sampling
Cross Intersection	0.014446306	0.080670614
T Intersection	0.005930668	0.053363889
Multiple Intersection	0.004028727	0.048942651
Y Intersection	0.033559549	0.057588665

Figure 3: NMI scores for different intersection geometries

	Cross Intersection	Multiple Intersection	T Intersection	Y Intersection
40	0.22167	nan	0.11513	nan
50	0.19859	nan	0.17135	nan
60	0.37016	0.27155	0.27697	nan
70	0.63492	0.70588	0.34608	nan
80	0.75154	0.82105	0.52781	nan
100	0.85580	nan	0.50760	nan

Figure 4: Normalised severe to fatal crashes

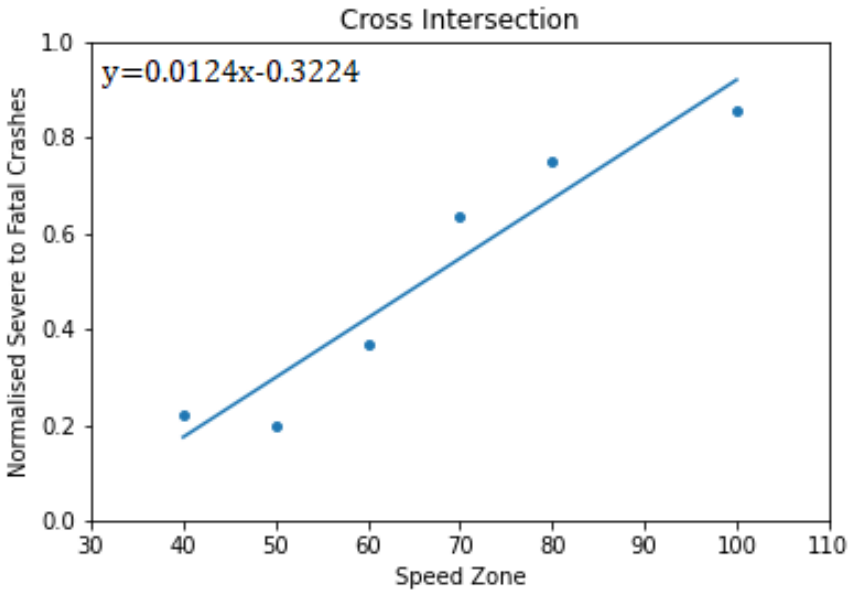

Figure 5: Scatter plot of normalised severe to fatal crashes in each speed-zone for Cross Intersection

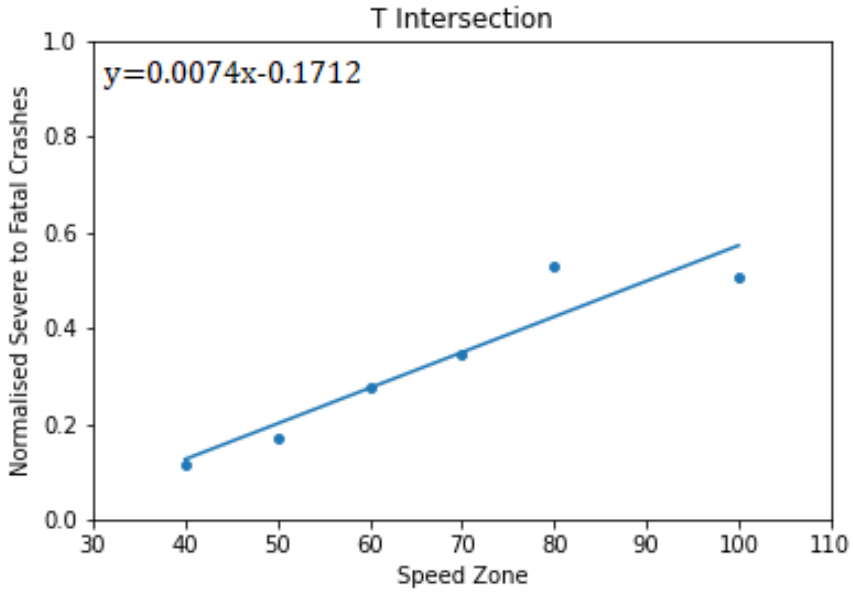

Figure 6: Scatter plot of normalised severe to fatal crashes in each speed-zone for T Intersection

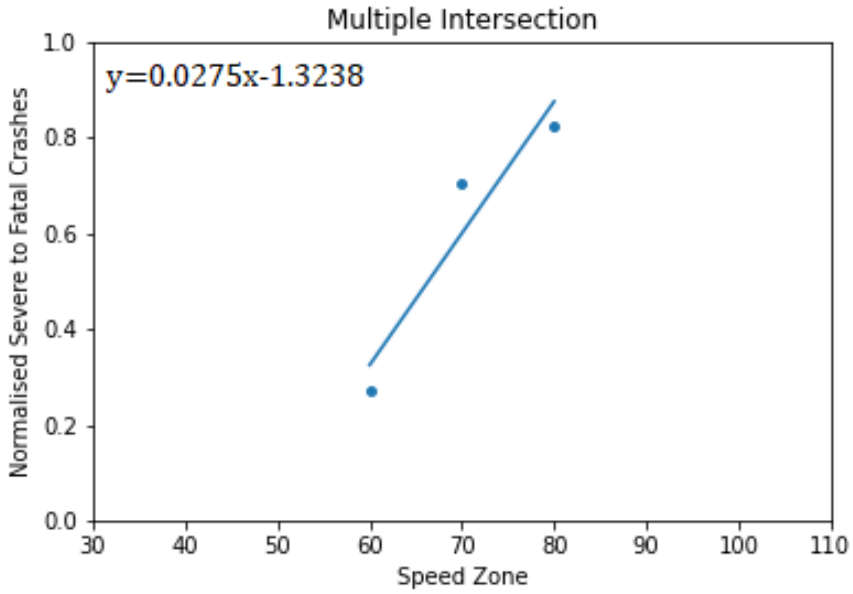

Figure 7: Scatter plot of normalised severe to fatal crashes in each speed-zone for Multiple Intersection

4 Discussion

In figure 1, the crash data is biased towards speed-zones of greater frequency, but the scatter plots accounts for this with normalisation. When looking at normalised data, the most dangerous road geometry speed-zone is cross intersections at 100km/h. Figures 5, 6, and 7 show a moderate to strong positive correlation between speed-zone and severe crashes. The most dangerous speed for the respective intersections is generally the highest speed-zone, with the safest speed being at the lowest speed-zone.

The linear regression also demonstrates a positive correlation between the speed-zone and normalised severe crash rate. However, the Y intersection has no data points as all the severe crash data-points fall under the threshold, therefore, it will not be considered in this discussion and analysis. Looking at the NMI score with iterative stratified sampling in figure 3, the cross intersection has the highest correlation, with a score of 0.0807, followed by the T intersection with 0.0534, and finally the multiple intersection with 0.0489. The relative gradients of the T and cross scatter-plots agree with this NMI score, with 0.0074 and 0.0124 respectively. Conversely, the multiple intersection does not correspond with this NMI value comparison, which is caused by the lack of severe crash data resulting in a linear regression from only 3 data points. Due to this, the multiple intersection will be omitted from analysis.

In order to determine an appropriate speed limit for the cross and T intersections, the highest speed-zone under 0.3 severe to fatal crashes against unique intersection count for that speed-zone will be taken. This metric balances trade-off between safety and commute time and will be calculated using the linear regression formula produced. For the T intersection the appropriate speed limit is:

$$y = 0.0074x - 0.1712$$

$$0.3 = 0.0074x - 0.1712$$

$$x = 63.67km/h$$

For the cross intersection, the appropriate speed limit is:

$$y = 0.0124x - 0.3224$$

$$0.3 = 0.0124x - 0.3224$$

$$x = 50.19km/h$$

Rounding down to the nearest valid VIC Roads speed band, the appropriate speed limit for **T** and **cross** intersections are **60km/h** and **50km/h** respectively.

Overall, these results are valuable as they provide insight into trends that raw data could not. For instance, the raw data did not show that an increase in speed led to an increase in severe crashes, but the normalised scatter plots did. Significant information such as this should make drivers reconsider rushing through intersections especially if the speeds are greater than the specified 'safe' speed limits for the T and cross intersections. Furthermore, comparisons about the relative safety between the intersections can be made, with the T intersection being safer than the cross intersection with a higher appropriate speed limit of 60km/h compared to 50km/h and a lower correlation (gradient) between the speed limit and severe crashes.

Limitations of this investigation are that the unique intersection counts used to normalise the data is from road data from all crashes in 2014-2019 as the team could not access map SHP files to analyse actual intersection counts in Victoria. Further, it does not account for certain intersections in school zones, meaning the speeds limits could change depending on the time of day, however, this should not be a significant error due to the size of the datasets used. Another limitation is that there is not enough data on Y and even multi-intersections to gain valid information about them. Future improvements would be to analyse SHP files to count unique intersections for different speed-zones, using larger sets of data covering large year ranges or regions, and considering datasets with information on schools to account for the road geometries that are also considered school zones.

5 Conclusion

This study aims to find the appropriate speed limit for each intersection geometry in Victoria based on the distribution of severe to fatal car crashes in different speed-zones. Through the use of iterative stratified sampling NMI scores, normalised scatter plots and least squares linear regression, the appropriate speed limit for T and cross intersections are **60km/h** and **50km/h** respectively.

References

Cellucci, C. J., Albano, A. M., & Rapp, P. E. (2005). Statistical validation of mutual information calculations: Comparison of alternative numerical algorithms. *Physical Review E*, *71*(6), 066208.

Transport Accident Commission. (2020). *Lives Lost - Year to Date.* `https://www.tac.vic.gov.au/road-safety/statistics/lives-lost-year-to-date`. (Online; accessed 19 May 2020)