FINAL REPORT 2020

Surrey Crosswalk

Alina Chalanuchpong
Roger Yu-Hsiang Lo
Julia Niebles
Sherry Zhang

Data Science for Social Good
UBC Institute of Data Science

20 December, 2020

Contents

1 Introduction

The City of Surrey has an existing process to determine whether to install a marked pedestrian crossing at a particular location. The current crosswalk assessment is informed by technical factors, such as pedestrian and traffic volume, pedestrian demand, pedestrian desired connectivity, latent demand, and resident requests submitted to the city. However, this process has a number of limitations in reflecting the actual crosswalk demands.

First, data collection for pedestrian and traffic volumes is costly in terms of time and resources. Since Surrey has over 18,000 intersections, it is infeasible to collect data at all locations. Similarly, resident service requests are a crucial component of this process because it allows the city to focus on particular locations based on residents' input. This information helps focus attention on locations where there are data gaps. However, a concern of this process is that the city receives three times as many requests from high-income areas as they do for low-income areas. At the same time, collision data shows that pedestrian collisions are approximately four times higher in lower-income areas, compared to high-income areas. Thus, the city identified a need to find a data-driven approach to determine crosswalk locations, as part of Surrey's Vision Zero safe mobility plan [1]. The goal of this project is to reduce potential data bias in service requests by improving the data collection and the analysis of technical factors to determine crosswalk locations.

The key outcomes for this project is to identify the factors important in determining crosswalk locations and find a method to rank potential intersections based on their need for a crosswalk. We identified a variety of factors important in determining crosswalk locations, including traffic speed, traffic volume, proximity to schools/community centres/hospitals, road width, and crash locations [2, 3, 4]. We built two indices and two models to predict each of these factors to supplement our intersection rankings.

An important aspect of this project is to develop a method to rank each intersection in terms of their need for a crosswalk. For this purpose we adapted two indices: the *Pedestrian Potential Index* and *Pedestrian Deficiency Index*. These indices were originally created by the City of Portland to rank locations where sidewalks were needed and have been adapted by different cities, such as North Vancouver [5], Prince George [6], and Victoria [7]. We adapted the two indices for crosswalks with the data from the City of Surrey. Additionally, based on the domain knowledge of engineers from the city and the Pedestrian Crossing Guidelines from the Transportation Association of Canada, pedestrian volumes and vehicular traffic volumes were identified as factors that are crucial for assessing the implementation of a crosswalk. Thus, we decided to build two models to predict pedestrian volume and traffic volume to complement our analysis.

1.1 Project Strategy

The aim of this project is to develop a data-driven approach for the city to determine crosswalk locations. The three main outcomes for this project are: (i) to develop a ranking system for each intersection to determine locations where there is a high need for a crosswalk, (ii) to build two models that predict pedestrian and traffic volumes for each intersection, and (iii) to create a dashboard on a Shiny app that showcases the results from the ranking system and allows the user to explore the intersection locations in terms of the factors that make up each index and the weighting of these factors in each index.

1.2 Shiny App Dashboard: Visualization Tool

Our goal for this project is to deliver a tool that the City of Surrey can use to visualize different data spatially and to analyze potential crosswalk locations. We chose to develop our visualization platform using R Shiny because of its abundant visualization tools for our map, its flexibility in the interface to change the data used globally, and its functionality to change and save the scoring matrix for the indices.

2 Data Sources

The data used for this project is derived from city-owned data, including open data available on their website (`https://data.surrey.ca`), ICBC data, and TransLink data on stop usage and locations.

2.1 Overview of Datasets

The datasets used in this project are summarized below:

- City-owned data:
  - Buildings and properties
  - Business licenses
  - Crosswalks
  - Greenways
  - Intersections
  - Pedestrian and traffic volumes
  - Places of interest
  - Road centreline
  - Schools
  - Service requests
  - Trails and pathways
- ICBC collision data: from 2012 to 2019
- TransLink data:
  - Transit stop locations
  - Detailed stop usage

3 Pedestrian Potential Index and Pedestrian Deficiency Index

3.1 Pedestrian Potential Index (PPI)

In its original formulation, the Pedestrian Potential Index (PPI) is intended as a metric to identify locations where improved physical facility is likely to result in increased walking trips, given other factors that already favor walking [8]. When first conceived by the City of Portland for their sidewalk improvement project, PPI emcompasses three sets of factors——policy factors, proximity factors, and pedestrian environmental factors. Policy factors refer to areas and corridors that have greater importance for pedestrians (e.g., pedestrian districts and main streets). Proximity factors quantify distance of a given location to pedestrian generators, such as schools, parks, and transit stops. Pedestrian environmental factors are identified as the variables that induce walk trips based on travel data. PPI was subsequently adapted by a number of cities in Canada for projects that aim to install or improve facilities for pedestrians.

3.1.1 Surrey's Pedestrian Potential Index

We adapted PPI based on the availability of data from the City of Surrey while also taking into consideration how other cities modified the index. Specifically, we tried to stick to the factors and the weighting of these factors put forth by North Vancouver whenever possible.

These considerations result in the following factors being used for computing PPI values, with notes added to factors that are not self-explanatory:

- **Proximity to the nearest walkway**: the Euclidean distance to the closest walkway, which includes greenways, trails, paths, sidewalks, and non-motorized routes.
- **Proximity to the nearest elementary school**: the Euclidean distance to the closest element school, which might be a public, a private, or a francophone one.
- **Proximity to the nearest secondary school**
- **Proximity to the nearest transit stop**
- **Proximity to the nearest park**: the part could be a usual park, a water park, a skate park, or a bike park.
- **Proximity to the nearest community centre**: community centres here include recreation centres, libraries, senior centers, community halls/centres, or youth centres.
- **Proximity to the nearest employment location with more than 20 employees**
- **Proximity to other local places of interest**: local places of interest include local attractions and movie theatres.
- **Commercial density**: this refers to the kernel density estimated using ESRI ArcGIS' density estimation tool (see `https://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-kernel-density-works.htm` for detailed description).
- **Bus usage**: this measure is based on the emerging hotspot analysis implemented in ESRI ArcGIS (`https://pro.arcgis.com/en/pro-app/tool-reference/space-time-pattern-mining/learnmoreemerging.htm`). As the output of the emerging hotspot analysis is in the format of points, we used both the nearest neighbouring point and KNN with $k = 5$ as the method to categorize each intersection into one of the output hotspot patterns.

3.1.2 Data Processing for Surrey's Pedestrian Potential Index

This section describes the datasets we used and the procedure we followed for extracting values for PPI factors.

- **Proximity to the nearest walkway**: we used the Greenways and Trails and Pathways datasets for this factor by including entries that are greenways, trails, paths, sidewalks, or non-motorized routes. We then used the R package `sf` to compute the distance from the intersection in question to the nearest walkway by the function `st_distance()` in a projected surface (`crs = 26910`).
- **Proximity to the nearest elementary school**: we used the School dataset and included only the schools that are labeled as "elementary", "private", or "francophone". The distance from an intersection to the closest elementary school is computed as described above for walkway.
- **Proximity to the nearest secondary school**: this is the same as the proximity to the nearest elementary school in terms of the dataset and procedure, except that now we only included secondary schools.
- **Proximity to the nearest transit stop**: for this measure, we used the Transit Stop Locations Data and computed the distance using the `st_distance()` function.

- **Proximity to the nearest park**: for this factor, as well as the ones involving community centres and local places of interest, relevant entries from the Places of Interest dataset were extracted, and again the distance is estimated as above.
- **Proximity to the nearest employment location with more than 20 employees**: Building and Property Data was used, and we only kept the entries that have 20 or more employees. The distance estimation procedure was then repeated.
- **Commercial density**: The Business License dataset was first imported into ArcMap and the kernel density was estimated using the Kernel Density tool, with the search radius parameter set to 500 meters.
- **Bus usage**: We first computed the hourly average rates for weekday and non-weekdays from the detailed bus usage data. We then imported the data into ArcGIS Pro for running the emerging hotspot analysis.

3.1.3 Calculating PPI Scores

The scoring criteria for PPI are based on the report from North Vancouver and through counseling with the personnel from the City of Surrey. The detail scoring criteria are summarized in Table 1.

3.2 Pedestrian Deficiency Index (PDI)

The deficiency index measures how critically pedestrian improvements are needed. Similar to the Potential Index, a value is assigned to each intersection based on the deficiency factors which originally were:

- Automobile-Pedestrian Crash (number of incidents in the last 3 years close to an intersection)
- Traffic Speed
- Traffic Volume
- Roadway Width
- Street Segment Length

The PDI score is aimed at helping determine whether a particular intersection has enough deficiency factors that would suggest the need for a crosswalk to be installed.

3.2.1 Surrey's Pedestrian Deficiency Index

The PDI score is used to determine whether a particular intersection in Surrey has enough deficiency factors that would suggest the need for a crosswalk to be installed. In order to adapt the index for the City of Surrey, the following factors were included in its calculation: Crash factor, traffic speed, number of lanes, road width, traffic volume and proximity to other crosswalks. Each factor is given points based on their value, and then summed up to calculate a PDI score.

Surrey's PDI Factors

Given the data we had available, the following factors were included to calculate a deficiency score for Surrey:

Table 1: PPI Score

Factor	**Feature**	**Points given**	**Max. points possible**
Distance to nearest walkway (m)	$x < 100$	2	2
	$x \geq 100$	0	
Distance to nearest elementary school (m)	$x < 500$	4	4
	$500 \leq x < 1000$	3	
	$1000 \leq x < 1500$	2	
	$1500 \leq x < 2000$	1	
	$x \geq 2000$	0	
Distance to nearest secondary school (m)	$x < 500$	4	4
	$500 \leq x < 1000$	3	
	$1000 \leq x < 1500$	2	
	$1500 \leq x < 2000$	1	
	$x \geq 2000$	0	
Distance to nearest transition stop (m)	$x < 500$	2	2
	$x \geq 500$	0	
Distance to nearest park (m)	$x < 500$	2	2
	$x \geq 500$	0	
Distance to nearest community center (m)	$x < 500$	2	2
	$x \geq 500$	0	
Distance to nearest employment location (m)	$x < 500$	2	2
	$x \geq 500$	0	
Distance to nearest local interests (m)	$x < 500$	2	2
	$500 \leq x < 1000$	1	
	$x \geq 1000$	0	
Commercial density within 500 m	$x \geq 70$	7	7
	$60 \leq x < 70$	6	
	$50 \leq x < 60$	5	
	$40 \leq x < 50$	4	
	$30 \leq x < 40$	3	
	$20 \leq x < 30$	2	
	$10 \leq x < 20$	1	
	$x < 10$	0	
Bus usage pattern (KNN)	Intensifying hotspot	3	3
	Consecutive hotspot	2	
	New hotspot	1	
	No pattern	0	
Bus usage pattern (nearest)	Intensifying hotspot	4	4
	Consecutive hotspot	3	
	New hotspot	2	
	Sporadic hotspot	1	
	No pattern	0	
Total			34

- **Number of lanes**: The number of lanes was included as a factor to complement the data on road width and traffic volume, since there can be positive correlation between the number of lanes and both traffic volume and road width. A road with several lanes will likely have a higher road width than a road with 2 lanes, and the more lanes there are the more potential for a high traffic volume. The range is from 1 to 6 lanes.
- **Road width**: the road width measures the distance from road centreline to the curb multiplied by two.
- **Speed**: indicates the speed limit for a particular intersection. Values range from 20 to 100 km/h.
- **Predicted vehicle AADT (volume)**: The predicted annual average daily traffic from the AADT model we built (see Section 4). Values range from 0 to 70,000 vehicles per day.
- **Distance to existing crosswalks**: distance in meters to a crosswalk in Surrey's crosswalk inventory. Values range from 0 to 3400 m.
- **Crash pattern**: The pattern is the corresponding hotspot category for an intersection based on an emergent hotspot analysis of ICBC crash data from 2012 to 2019. These categories were created by the hotspot analysis tool in ArcGIS, and integrated into this indicator (see Appendix for detailed explanation of the hotspot analysis performed).

  The resulting categories are:

  - Intensifying hotspot
  - Persistent hotspot
  - Consecutive hotspot
  - New hotspot
  - Sporadic hotspot
  - No pattern

3.2.2 Data Processing for Surrey's Pedestrian Deficiency Index

In order to create the PDI, the data for each factor was extracted from different datasets, and joined together to create a data frame with all the factors, and a PDI score for each intersection that has a unique ID number.

- **Speed, number of lanes, and road width**: The dataset Road Centrelines was used, which contained the fields of speed, and number of lanes for each road in Surrey. Since there was no field for road width, the datasets Road Boundary and Intersections were used to calculate it. The road width was calculated by creating a buffer zone of 1 meter around the intersection point and taking the nearest distance from the intersection point to the road boundary and multiplying by 2.
- **Crash pattern**: The ICBC Collision raw data was transformed to a layer in ArcGIS for an emergent hotspot analysis. The data was filtered to include only collisions where pedestrians were involved. "Pattern", indicated the hotspot trend and was extracted and joined to the PDI dataset by using the "Intersection_ID" field to match all each intersection with its corresponding "Pattern" value.
- **Predicted Annual Average Daily Traffic (AADT)**: The values were taken from the model results, (see Section 4, for more details on the model). The prediction from the model produced

the data fields: “FIT”, “AADT”, “LOWER”, “UPPER”, “AADT_LOWER”, “AADT_UPPER”. The “AADT” values were taken and any NA values were replaced with values from the “FIT” column. Then they were joined to the PDI dataset

- **Distance to crosswalk**: The distance to an existing crosswalk was calculated by measuring the euclidean distance of each Intersection ID to an existing crosswalk in Surrey’s crosswalk inventory (3007 in total).

3.2.3 Calculating PDI Scores

In order to provide a guideline for scoring each factor, we created default PDI scores. The scoring criteria were based on the methodology used by the City of Portland [8], Victoria (BC) [7], and North Vancouver [5], for calculating a sidewalk deficiency index. The factors distance to existing crosswalks and crash indicator were scored in consultation with the team at the city and based on the emergent hotspot analysis data of automobile-pedestrian collisions (see Appendix A for more information).

In the visualization app, the user will have the flexibility to change the weight and scoring of each factor (i.e., decide which factor is more influential for that specific location), based on their own criteria. Additionally, they will be able to save the scoring values and download a dataset with the calculated PDI scores. Table 2 shows the default scores that will initially appear in the tab “PDI Weight” in the visualization. These default values are included for reference and to exemplify how a PDI score is built.

3.3 Combining the Potential Index and the Deficiency Index

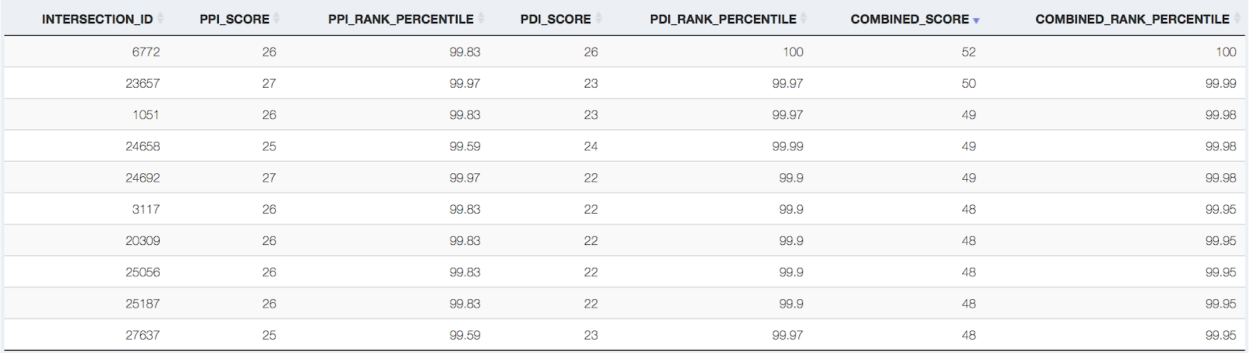
The overarching goal of building both indices was to combine them and create a combined score to enable the city to prioritize and rank intersections that need a crosswalk based on both potential and deficiency factors. Individually the indices convey different perspectives:

- **Potential Index**: measures the potential of locations where new pedestrian improvements—such as crosswalks—would increase the opportunities for pedestrian trips, given that the location already has factors that favour pedestrian activity. Thus, a high score would imply that a crosswalk would improve overall pedestrian mobility. A low score, might indicate that pedestrian trips or demand is low and the surroundings of that area do not generate much pedestrian traffic.
- **Deficiency Index**: measures the need for pedestrian improvements, thus a high value would indicate that a location might need a crosswalk since there are not enough environmental factors present that would facilitate pedestrian activity. This index could also be an indication of locations with a high-risk for pedestrians due to the lack of infrastructure.

By combining both indices, there is a better understanding of each intersection’s needs and potential for a crosswalk. Although the weighting of each factor is ultimately at the discretion of the user, combining the scores of the Potential Index and the Deficiency Index to create a combined scores can help in focusing and prioritizing certain locations out of the 18,000 intersections. The combination of both indices is a simple summation of the final scoring as seen in Figure 1. Ideally, the focus would be in those intersections with a high combined score (high potential index and a high deficiency index). Different results would also be informative, for example a high deficiency score but a low potential score. In Section 5.3, we give an example of how to analyze an intersection according to its combined score.

Table 2: PDI Score

Factor	**Feature**	**Points given**	**Max. points possible**
Speed (kph)	$x \geq 80$	5	5
	$70 \leq x < 80$	4	
	$60 \leq x < 70$	3	
	$50 \leq x < 60$	2	
	$40 \leq x < 50$	1	
	$x < 40$	0	
Lanes	$x \geq 4$	4	4
	$x = 3$	3	
	$x = 2$	2	
	$x = 1$	1	
	$x < 1$	0	
Predicted vehicle AADT (vehicles/day)	$x \geq 20000$	5	5
	$15000 \leq x < 20000$	4	
	$10000 \leq x < 15000$	3	
	$5000 \leq x < 10000$	2	
	$2000 \leq x < 5000$	1	
	$x < 2000$	0	
Road width (m)	$x \geq 27$	6	6
	$24 \leq x < 27$	5	
	$21 \leq x < 24$	4	
	$18 \leq x < 21$	3	
	$15 \leq x < 18$	2	
	$12 \leq x < 15$	1	
	$x < 12$	0	
Distance to crosswalk (m)	$x \geq 1000$	5	5
	$500 \leq x < 1000$	4	
	$400 \leq x < 500$	3	
	$300 \leq x < 400$	2	
	$200 \leq x < 300$	1	
	$x < 200$	0	
Crash pattern	Intensifying hotspot	10	10
	Persistent hotspot	8	
	Consecutive hotspot	6	
	Sporadic hotspot	4	
	New hotspot	2	
	No pattern	0	
Total			35



INTERSECTION_ID	PPI_SCORE	PPI_RANK_PERCENTILE	PDI_SCORE	PDI_RANK_PERCENTILE	COMBINED_SCORE ▾	COMBINED_RANK_PERCENTILE
6772	26	99.83	26	100	52	100
23657	27	99.97	23	99.97	50	99.99
1051	26	99.83	23	99.97	49	99.98
24658	25	99.59	24	99.99	49	99.98
24692	27	99.97	22	99.9	49	99.98
3117	26	99.83	22	99.9	48	99.95
20309	26	99.83	22	99.9	48	99.95
25056	26	99.83	22	99.9	48	99.95
25187	26	99.83	22	99.9	48	99.95
27637	25	99.59	23	99.97	48	99.95

Figure 1: Table of combined scores

3.3.1 A Note on Weighting

The methodology used in the literature and our project, the Potential and Deficiency Indices do not have the same amount of possible points (in this case it would depend on the parameters the users give to each factor within the scores). If using the default scores, the Deficiency Index would have 35 possible points whereas the Potential Index default scores would create a maximum of 34 points. Thus, one of the indices would contribute more to the combined score. This could be a limitation, but if the parameters used are similar to default scores (based on the literature) then the methodology should be accurate.

3.3.2 Limitations of the Indices

The indices provide a measurement to quantify the need for a crosswalk, but there are some limitations that have to be addressed. First, in adapting the index from different cities to Surrey, the indices' variables changed for two reasons: the data available for the City of Surrey was different, and the index methodology was originally used to determine sidewalk locations not crosswalks. Thus, in some cases we could not follow the exact methodology based on the literature.

Second, for the factors that measured distance (e.g., in PDI distance to crosswalks) the euclidean distance was used which can differ from the actual walking distance changing the accuracy of some values For some variables we had to missing data such as road width and the AADT values which were derived from a predictive model.

Despite these limitations the indices can inform future data collection in the city as well as shed light on factors that need more attention. Additionally, the indices have the potential to be adapted and used in other contexts such as measuring accessibility needs for pedestrians with disabilities and improving the walkability of different places.

4 Annual Average Daily Traffic Modeling

Pedestrian volume and vehicle volume are two criteria currently used by the City of Surrey when evaluating locations needing a crosswalk (`https://www.surrey.ca/city-services/8756.aspx`). In addition, daily vehicle volume is a factor considered in PDI. To evaluate all potential crosswalk locations, we need pedestrian volume and vehicle volume for all intersections, but collecting traffic volume data at all 18568 intersections is clearly impractical. Therefore, the Annual Average Daily Traffic (AADT) models aim to provide estimates of pedestrian and vehicle volumes for every intersection to the most recent year possible. In our AADT models, pedestrian volume include both pedestrian and bicycle traffic flows, while vehicle volume included counts of cars, buses, and trucks.

4.1 Data Preprocessing

4.1.1 Response Variable: AADT Conversion

From 2006 to 2019, traffic volumes were recorded manually for all four directions (i.e., northbound, southbound, eastbound, westbound) and all three movements (i.e., left, right, through) in 15-minute intervals for a total of 2, 4, or 7 hours. In 2019, the city transitioned to automatic counting via video; traffic volume were therefore typically counted for 24 or 48 consecutive hours for 144 intersections in 2019 and 2020. On average, two years' worth of traffic volume data was available for 1335 count locations.

To make these short-term traffic counts more comparable, we performed a series of conversions, adjusting for different hours at different times of the year, based on the 2018 AADT adjustment factors provided to us. The factors recognize four categories of traffic volume pattern: *significant school impact*, *some school impact*, *commercial*, and *typical*. For each category, two sets of adjustment factors were used: (1) factors that convert less-than-24-hour counts to daily traffic volume, differentiating weekdays from weekends, and (2) factors that convert daily traffic volume to AADT values based on the day of the week and the month of the year. Even though the document provides some characteristics of each category, there were no rigorous specifications to categorize an intersection. Therefore, we decided to use only the factors from the `typical` category, which should be the correct category for about two thirds of the intersections.

Before using the adjustment factors, the first step is to aggregate the 15-minute interval counts into hourly counts. For this, we first summed the counts for all directions and movements in the same time interval for each type of road users. Although the majority of count data is recorded in 15-minute intervals, we also have data from intervals of 5, 30, and 45 minutes. For example, the first entry of the hour could be from 8 AM, but the following entry was from 8:45 AM. In processing the raw data, we relied on the assumption that the interval is consistent for the same count ID, that is, we computed the interval as the minimum time difference between two consecutive entries from the same hour of that count ID. The second assumption we made was that if the first and final intervals exist for the hour, the missing entries in between corresponded to 0 counts; otherwise, we deemed the hour incomplete and discarded the data. These assumptions reduced the number of count locations from 1335 to 1328 intersections.

The next step made use of the first set of adjustment factors to convert hourly counts to a daily value. Due to the uncertainty introduced by a substantial difference in hourly counts from day to day and from location to location, we also included an upper and lower bound that constitute a 95% confidence interval in the conversion. The final step was to convert daily values to AADT values using the second set of the adjustment factors. We used the same adjustment factors for the daily estimates as well as their lower and upper bounds. The same steps were taken for both pedestrian AADT conversion and vehicle AADT conversion.

The final step was to convert the daily value to an AADT value using the second set of the adjustment factors. We used the same adjustment factors for the daily estimates as well as their lower and upper bounds. The same steps were taken for both pedestrian AADT conversion and vehicle AADT conversion.

4.1.2 Spatial and Temporal Outliers

Even though the AADT conversion took the short term traffic count data at different intersections to a more comparable level, there is a chance that an exceptional event, like extreme weather, happened on the day of the count, causing artificial inflation or deflation to the AADT values. Hence, we examined the AADT values spatially (see Figure 2) and temporally (see Figure 3).

According to Tobler's first law of geography, which states that "near things are more related than distant things", we expected AADT values close in space to have similar values. By visually inspecting the spatial distribution of pedestrian AADT year by year (see Figure 2a), four spatial outliers (circled in red in Figure 2a) were removed. On the other hand, no spatial outliers were removed for vehicle AADT.

For intersections with traffic volume data spanning multiple years, we expected the AADT values to not fluctuate too much across the years. For this reason, we examined their time series and flagged the extreme AADT values more than three interquartile ranges below the first quartile or above the third quartile. The time series of flagged intersections are shown in Figure 2a and Figure 2b. One entry circled in red was removed for pedestrian AADT while three were removed for vehicle AADT.

4.1.3 Explanatory Variables

As PPI measures how likely people will walk near an intersection and PDI measures how lacking the pedestrian facilities are near an intersection, we expected pedestrian volume to be positively correlated with PPI factors and negatively correlated with PDI factors. Additionally, based on data availability, variables used in the state-of-the-art AADT models (see Table 1 at `https://www.sciencedirect.com/science/article/pii/S2046043019301108`) were included to form a total of 48 variables from 22 categories. Both the pedestrian AADT model and the vehicle AADT model used the same set of explanatory variables.

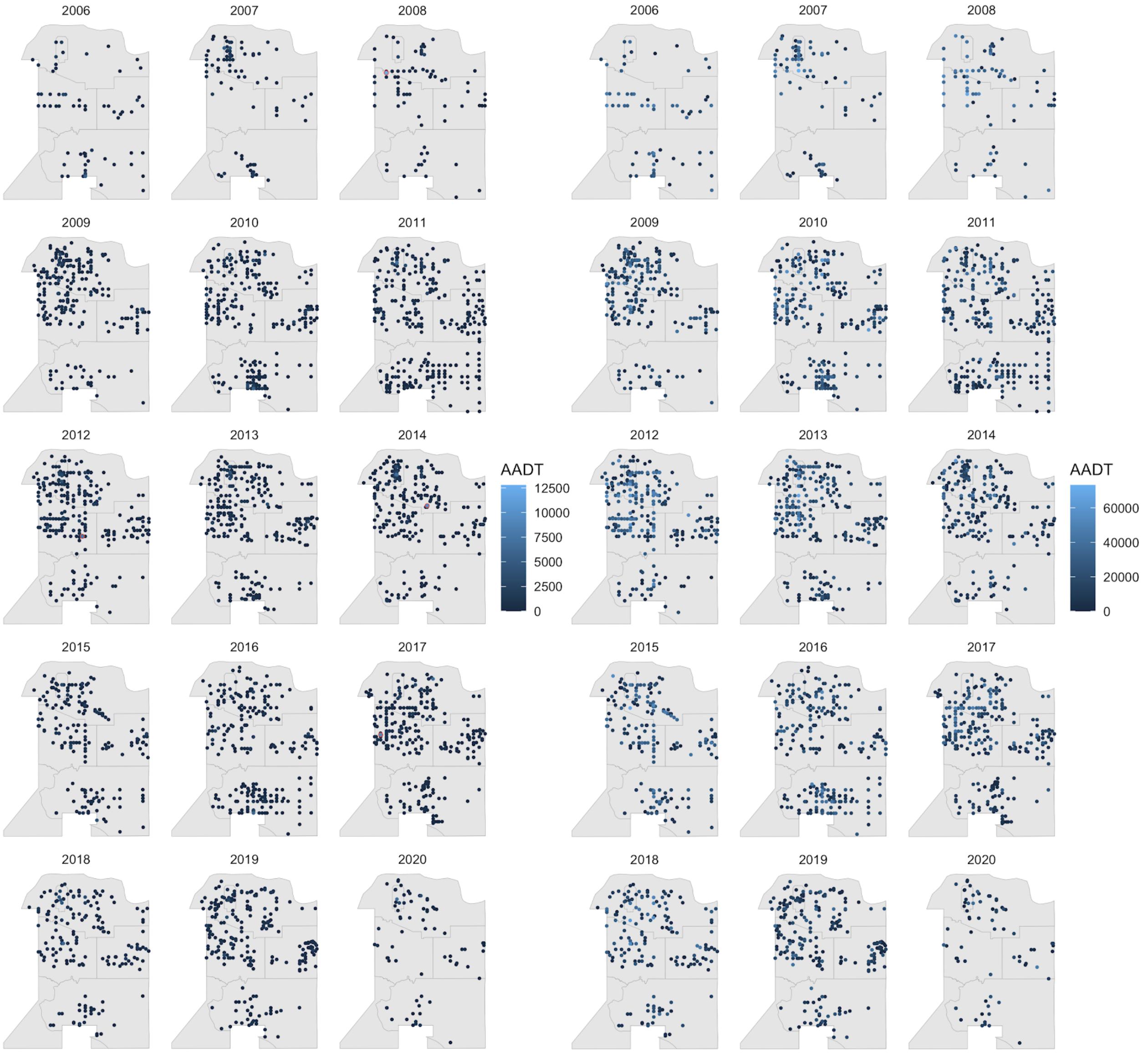
Spatial Explanatory Variables

Latitude and longitude were included in accordance with Tobler's first law of geography. Apart from that, summary statistics of the other spatial variables for the count locations are shown in Table 3 and Table 4. It is worth mentioning that while PPI and PDI took a distance-based approach, the AADT model variables were mostly density-based.

Temporal Explanatory Variables

The response variable spanned years from 2006 to 2020; at the same time, the census data, ICBC collision data, and bus stop usage data also covered multiple years. In addition to using the year as a temporal explanatory variable, we used the 2006 census data for years 2006 to 2008, the 2011 census data for years 2009 to 2013, and the 2016 census data for years 2017 to 2020. However, the ICBC collision data was only available from 2010 to 2019; the bus stop usage data was only available from 2014 to 2019.

We extracted two topics from the census data: household income and mode of commuting. Income was included not only because it was in the literature (`https://www.sciencedirect.com/science/article/pii/S2046043019301108`). Suppose income was selected by the pedestrian AADT model in a negatively correlated fashion so that lower-income areas actually had higher pedestrian volume. In that case, the disproportion between the service request and pedestrian collision in low-income and high-income areas would be even more alarming. On the other hand, if



(a) Spatial distribution of pedestrian AADT

(b) Spatial distribution of vehicle AADT

Figure 2: Spatial distribution of AADT

Table 3: Continuous spatial explanatory variables

Variable name		Mean	1st quantile	Median	3rd quantile	Minimum	Maximum
Population	100m	156.1	63.2	122.2	185.3	0	2806.32
	500m	3609.3	2287.6	3339.6	4538.2	3.6	15531.7
	1km	12946.4	8797	12522.2	16125.5	145.9	42856.3
Number of elementary schools	100m	0.08961	0	0	0	0	2
	500m	0.7666	0	1	1	0	4
	1km	2.48	2	2	3	0	7
	1.5km	4.942	3	5	6	0	12
	2km	8.01	6	8	10	0	18
Number of secondary schools	100m	0.01205	0	0	0	0	1
	500m	0.1258	0	0	0	0	1
	1km	0.442	0	0	1	0	2
	1.5km	0.8893	0	1	1	0	3
	2km	1.44	1	1	2	0	4
Number of places of interest	100m	0.1212	0	0	0	0	7
	500m	2.209	0	1	3	0	20
	1km	8.209	2	6	12	0	33
Number of bus stops	100m	0.8855	0	0	2	0	10
	500m	7.346	4	6	10	0	31
Distance to nearest bus stop (m)		240.79	42.3	125.66	322.42	10.47	4839.64
Distance to nearest crosswalk (m)		149.95	0	48.22	210.03	0	3207.66
Road width (m)		15.246	9.224	14.5	19.302	2.803	41.789
Number of lanes		2.523	2	2	2	2	6
Speed limit		52.95	50	50	60	30	100

Table 4: Categorical spatial explanatory variables

Existence of walkways

10m		50m		100m		200m		400m	
Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
1070	258	1304	24	1309	19	1312	16	1320	8

Road class

Arterial	Collector	Local
775	343	210

Intersection type

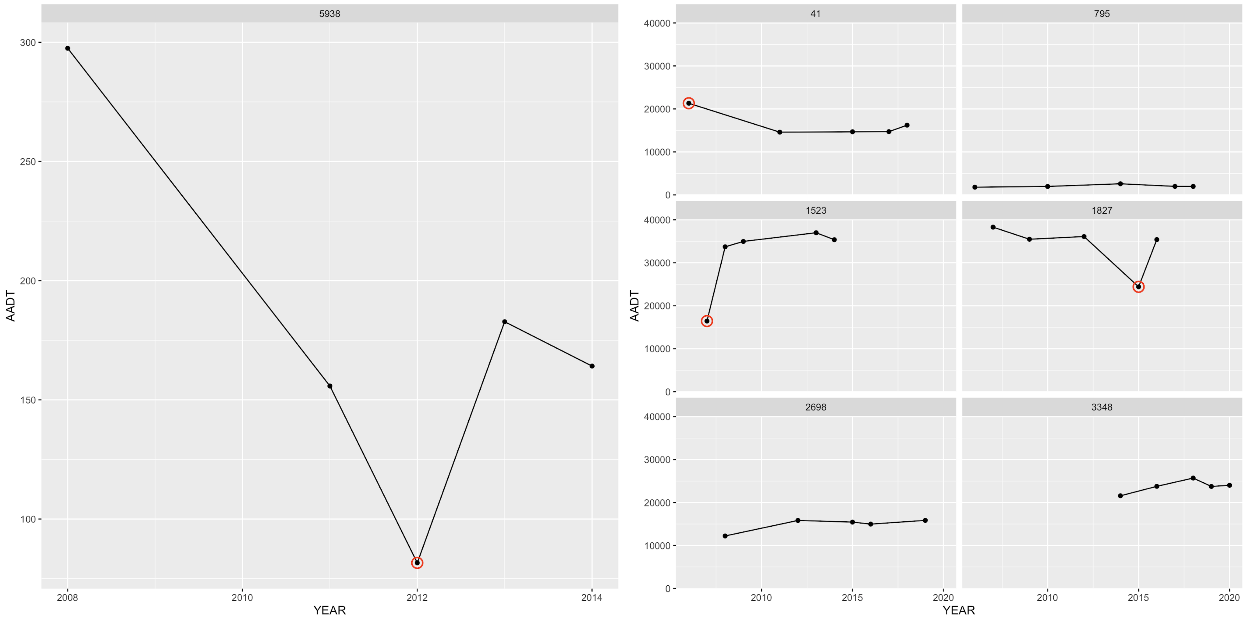
Real	Midblock	Other
1213	110	5

Traffic control

None	Traffic signal	Pedestrian signal	All-way stop	Beacons signal	Other
879	323	46	30	13	37

Land use zoning

Single family residential zone	Comprehensive development zone	One acre residential zone	Community commercial zone	General agricultural zone	Other
473	200	169	46	45	395



(a) Pedestrian AADT time series with a temporal outlier

(b) Vehicle AADT time series with temporal outliers

Figure 3: AADT time series

income had a significant positive correlation with pedestrian volume, then the higher number of service requests submitted from high-income areas could be justified. However, some dissemination areas (DAs) did not have income information but had intersections. In those cases, we filled in the missing income with the mean income from adjacent DAs. Furthermore, we separated the ICBC collision data into two categories, collisions that involved pedestrians or cyclists, and collisions that did not, to more precisely target the two models. Summary statistics of these temporal variables for the count locations are shown in Table 5.

4.2 Generalized Additive Model

It is important to account for spatial autocorrelation when modelling spatial data. Tobler's first law of geography states that "everything is related to everything else, but near things are more related than distant things." If an intersection has a high traffic volume, it is likely that the next intersection 100 meters away also has a high traffic volume. Spatial autocorrelation measures just the degree to which near things are related to each other.

Many models assume independent observations, but spatial autocorrelation puts that assumption in question. One solution is to incorporate the spatial dependency as a systematic part of the model, and one such model is the generalized additive model.

Table 5: Temporal explanatory variables

Variable name		Mean	1st quantile	Median	3rd quantile	Minimum	Maximum
Year		2013	2011	2013	2017	2006	2020
Household income	Median	77107	59648	77278	91902	15259	186874
	Average	89601	68929	86609	104014	28259	351769
Proportion of commuting mode	Bicycle or walk	0.02799	0	0.01445	0.04255	0	0.5333
	Bus	0.12544	0.05882	0.10976	0.17678	0	0.53968
	Car	0.8388	0.7818	0.8605	0.9167	0.4	1
Collision involving pedestrian or cyclist	50m	0.237	0	0	0	0	9
	100m	0.2627	0	0	0	0	9
	250m	0.6083	0	0	1	0	14
Collision not involving pedestrian or cyclist	50m	9.34	0	2	8	0	227
	100m	10.2	0	3	9	0	227
	250m	21.72	3	8	24	0	282
Annual average daily bus stop usage	100m	109.21	0	0	54.18	0	16247.79
	500m	1107.74	73.01	285.4	730.24	0	31050.86

Table 6: 10-fold CV results

	% within AADT bounds	% overlap	% AADT bounds covered	% relative order preserved
Gaussian pedestrian AADT	29.8	75.8	80.1	75.7
Negative binomial pedestrian AADT	28.1	63.2	67.3	75.8
Gaussian vehicle AADT	67.5	93.6	71.8	88.8
Negative binomial vehicle AADT	61.3	85.8	58.6	87.8

4.2.1 Model Structure

Generalized additive models (GAMs) are an extension of generalized linear models. They allow the linear predictor to depend linearly on unknown smooth functions of the explanatory variables:

$$g(E(Y_i)) = X_i\beta + f_1(x_{i,(p+1)}) + f_2(x_{i,(p+2)}) + \ldots + f_q(x_{i,(p+q)})$$

where g is a link function, X_i^* is the i-th row of the model matrix for the strictly parametric model components, β is the $(p+1)$-dimensional parameter vector, and $f_1, \ldots, f_q$ are unknown smooth functions to be estimated by non-parametric means, and can be functions of multiple explanatory variables.

In our case, the smooth functions were represented with penalized regression splines, which allowed us to capture the spatial dependency (and spatiotemporal dependency, if there was any) by modelling the interaction between latitude and longitude (and year) in a tensor product smooth.

In practice, a modestly large value is chosen as the basis dimension for the smooth function, then a penalty is applied to each smooth function to regularize the smoothness, thereby reducing overfitting. The trade-off between the smoothness and the model fit is controlled by the smoothing parameter, which in our case was optimized by minimizing the out-of-sample prediction error, as measured by the generalized cross-validation score.

4.2.2 Model Selection

As mentioned before, the ICBC collision data and the bus stop usage data did not cover all years for which the response AADT value was available. Including the ICBC collision data as an explanatory variable would exclude roughly 20% of the observations, which corresponds to approximately 500 observations earlier than 2010 or later than 2019. Similarly, including the bus stop usage data would exclude about 54% of the observations. The decision of whether to include them was made based on the 10-fold cross-validation results discussed later. To summarize, the models fitted with and without these two categories of explanatory variables did not have a staggering difference. Considering that more than half of the observations would need to be excluded for the bus stop usage data to be included and that observations prior to 2010 might not be as relevant, we decided to include the ICBC collision data but exclude the bus stop usage data. As a result, only observations from 2010 to 2019 were used in training.

Since our response variable is over-dispersed count data, we considered fitting a negative binomial model as well as a square root transformed Gaussian model, because taking the square root of the response AADT values is a variance-stabilizing transformation. As the AADT models' primary goal was to predict rather than explain, the Akaike information criterion (AIC) was used for stepwise variable selection for the parametric part of the model, to select the subset of most relevant explanatory variables that minimizes prediction error.

The adjusted R-squared estimates the proportion of variance in the response AADT values explained by the selected explanatory variables; the deviance explained measures the proportion of the null deviance explained by the model. They both assess the goodness-of-fit for a model. Table 7 shows these measures for the candidate models.

Table 7: Goodness-of-fit measures for Gaussian and negative binomial AADT models

	Adjusted R-squared	**Deviance explained**
Gaussian pedestrian AADT	0.584	0.664
Negative binomial pedestrian AADT	0.381	0.490
Gaussian vehicle AADT	0.895	0.921
Negative binomial vehicle AADT	0.811	0.829

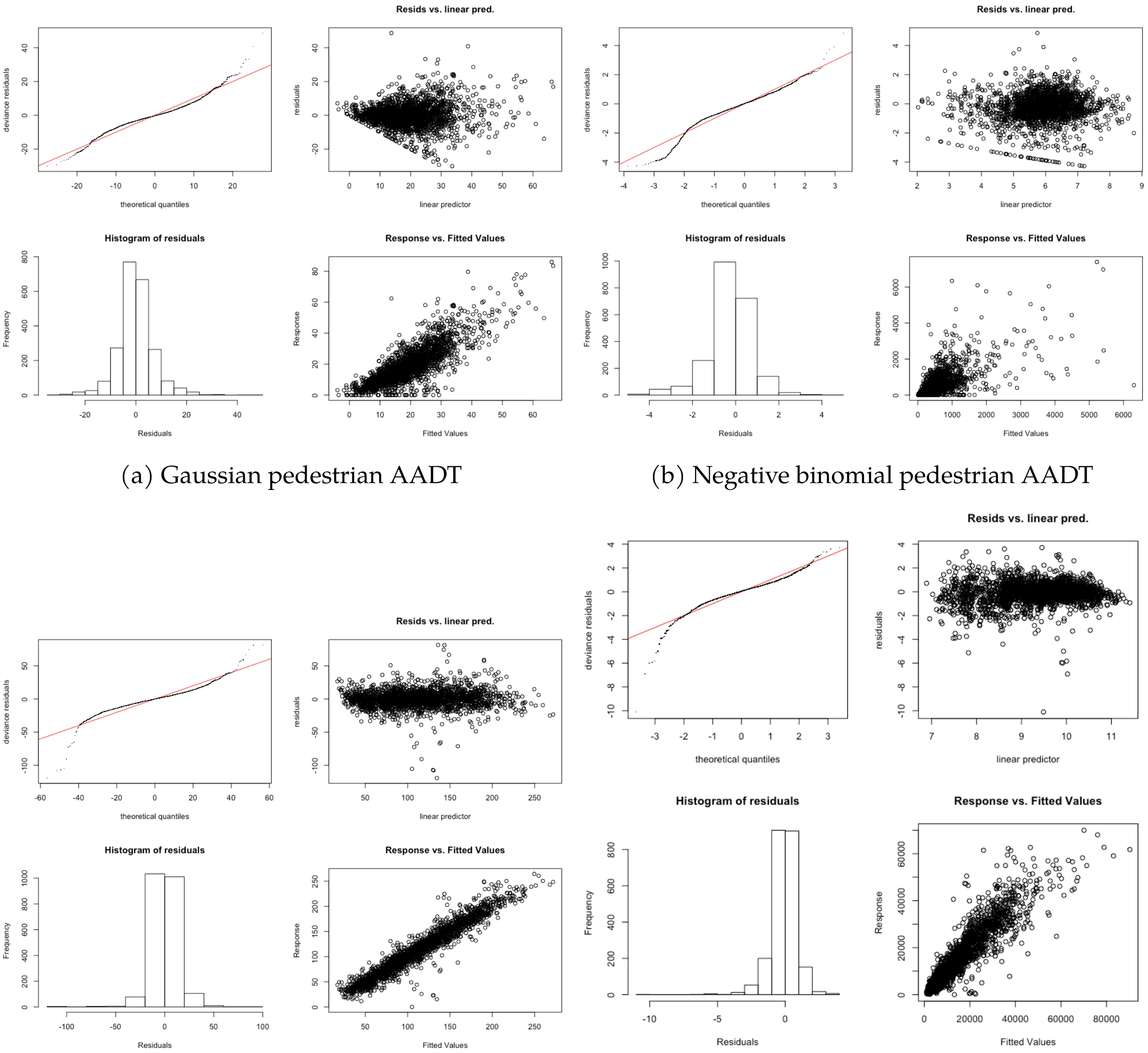
The Gaussian models obtained higher adjusted R-squared and higher deviance explained than the negative binomial models. However, one of the model assumptions was that the residuals should exhibit roughly constant variance. From the diagnostic plots in Figure 4, this constant variance assumption was grossly violated by the Gaussian pedestrian AADT model but more acceptable in the negative binomial pedestrian AADT model. For vehicle AADT, the constant variance assumption was complied arguably to the same degree by both the Gaussian and negative binomial models.

As the primary goal of the AADT models was to predict, we also performed 10-fold cross-validation to assess

1. the proportion of times the model estimate was within the lower and upper AADT bounds,
2. the proportion of times the 95% confidence interval (CI) overlapped with the AADT bounds; and when they overlapped, the percentage of AADT bounds covered by the 95% CI, and
3. the proportion of pairs of intersections for which the relative order of AADT was preserved.

The results are summarized in Table 6. Overall, the Gaussian models had better performance in terms of the three CV measures; the vehicle AADT models had a significantly better fit and prediction accuracy than the pedestrian AADT models. Still, in terms of our project goal—ranking intersections for the need of a crosswalk——the proportion of relative order preserved was a more critical metric than the accuracy of the estimates themselves. For that particular metric, the difference between the Gaussian model and the negative binomial model was virtually indistinguishable; the difference between the pedestrian AADT model and the vehicle AADT model was also less striking.

Being the more natural choice for over-dispersed count data, the negative binomial models had the advantage of ensuring that the predictions are non-negative, which was a problem faced by the Gaussian models. The Gaussian pedestrian AADT model had negative predictions for 537 intersections, and the Gaussian vehicle AADT model had 358. Instead of forcing these negative predictions to be 0 and collapsing their corresponding confidence intervals, taking into account all



(a) Gaussian pedestrian AADT

(b) Negative binomial pedestrian AADT

(c) Gaussian vehicle AADT

(d) Negative binomial vehicle AADT

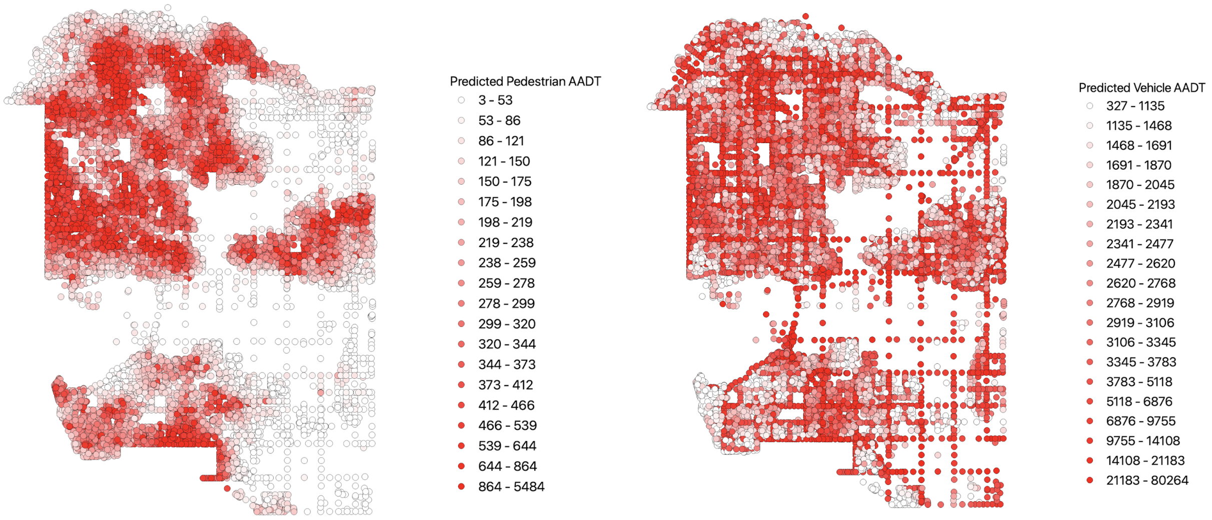
Figure 4: Diagnostic plots of models

factors discussed above, we decided to use the negative binomial models for both the pedestrian AADT and the vehicle AADT.

The selected variables for the strictly parametric part of the model are listed in Table 8. Latitude and longitude were included in the non-parametric smooth function. Interestingly, year was only selected by the vehicle AADT model and the degree of freedom was only 1, meaning no temporal smooth was needed. Also, average income was negatively associated with vehicle AADT, in agreement with the fact that pedestrian collisions happen more frequently in low-income areas. However, the average income was no longer a significant explanatory variable after adding the spatial tensor product smooth.

4.2.3 Model Prediction

The distributions of predicted pedestrian and vehicle AADT are shown in Figure 5. Intuitively, the pedestrian volumes clustered spatially, while the vehicle volumes clustered along the major roads. Intersections with both high pedestrian and high vehicle AADT would be more in need of a crosswalk.



(a) Predicted pedestrian AADT

(b) Predicted vehicle AADT

Figure 5: Distribution of predicted pedestrian and vehicle AADT

Despite the reasonable-looking AADT distributions, we encountered some difficulties in producing the predictions——some levels of the categorical explanatory variables were not in the training set. Traffic control was selected by both models, but 326 intersections had traffic controls like diverter, fire signal, and raised crosswalk, that were not seen in the training set. As 95% of the intersections had “none” traffic control, and the most important types of traffic controls, for example, traffic signals, pedestrian signals, and all-way stops, were included in the training set, we just grouped the unseen traffic controls into the “none” category. Likewise, we had 149 intersections with unseen land use zonings. This time since land use zoning is a spatial attribute, we took an approach similar to KNN where we made predictions using adjacent zone values seen in the training data and took the mean of those predictions. The corresponding standard errors were adjusted as well to form the appropriate confidence intervals.

It is worth mentioning that the predictions were made for 2019, because the ICBC collision data was used as explanatory variables in the models and it was available up to 2019.

Table 8: Selected variables in the final models

Variable name		**Negative binomial pedestrian AADT**	**Negative binomial vehicle AADT**
Spatial explanatory variables			
Population	100m		✓
	500m	✓	✓
	1000m		✓
Number of elementary schools	100m	✓	✓
	500m	✓	✓
	1km	✓	
	1.5km		✓
	2km	✓	✓
Number of secondary schools	100m	✓	
	500m	✓	
	1.5km	✓	✓
	2km		✓
Number of places of interest	500m		✓
Number of bus stops	100m	✓	✓
Distance to nearest bus stop (m)		✓	
Distance to crosswalk (m)			✓
Existence of walkway	100m		✓
	200m	✓	
	400m	✓	✓
Number of lanes	✓	✓	
Speed limit	✓	✓	
Road class	✓	✓	
Intersection type		✓	
Traffic control	✓	✓	
Land use zoning	✓		
Temporal explanatory variables			
Year		✓	
Household income	Average		✓
Proportion of commuting mode	Bus	✓	
	Car		✓
Collision involving pedestrians or cyclists	50m	✓	✓
	250m	✓	✓
Collision not involving pedestrians and cyclists	50m	✓	
	100m	✓	✓
	250m	✓	✓

4.3 Limitations

The AADT models had some limitations and areas for improvement for both the response variable and the explanatory variables.

We had assumed missing in-between traffic count entries to be 0 and discarded incomplete hours. We would recommend recording the traffic count data in a more robust format, making a clear differentiation between the missing data and the 0 counts, but the problem should be solved by transitioning to automatic traffic volume counting through video. The transition to automatic counting would solve another problem about distinguishing cyclists on road and cyclists on crosswalk, making the classification of crosswalk users more accurate.

Also, the AADT conversion factors were first calculated in 2013 and later updated in 2018, but only the 2018 AADT conversion factors were available to us. If possible, we recommend using the conversion factors from the closest year. On a lighter note, only factors from the typical category were used. Although this problem was partially addressed by including distance to schools, distance to places of interest, and population as explanatory variables, thereby mimicking the school and commercial impact, the accuracy of the response variable would still benefit from a more dedicated conversion.

Most importantly, the AADT conversion factors were meant for vehicles, but the pedestrian AADT model used the same set of conversion factors. Looking at Figure 6, we would expect more fluctuation in pedestrian volume across the months of the year, and probably higher on weekends and holidays. If a dedicated set of conversion factors was calculated for pedestrians, the pedestrian AADT model should potentially achieve a better model fit and higher prediction accuracy.

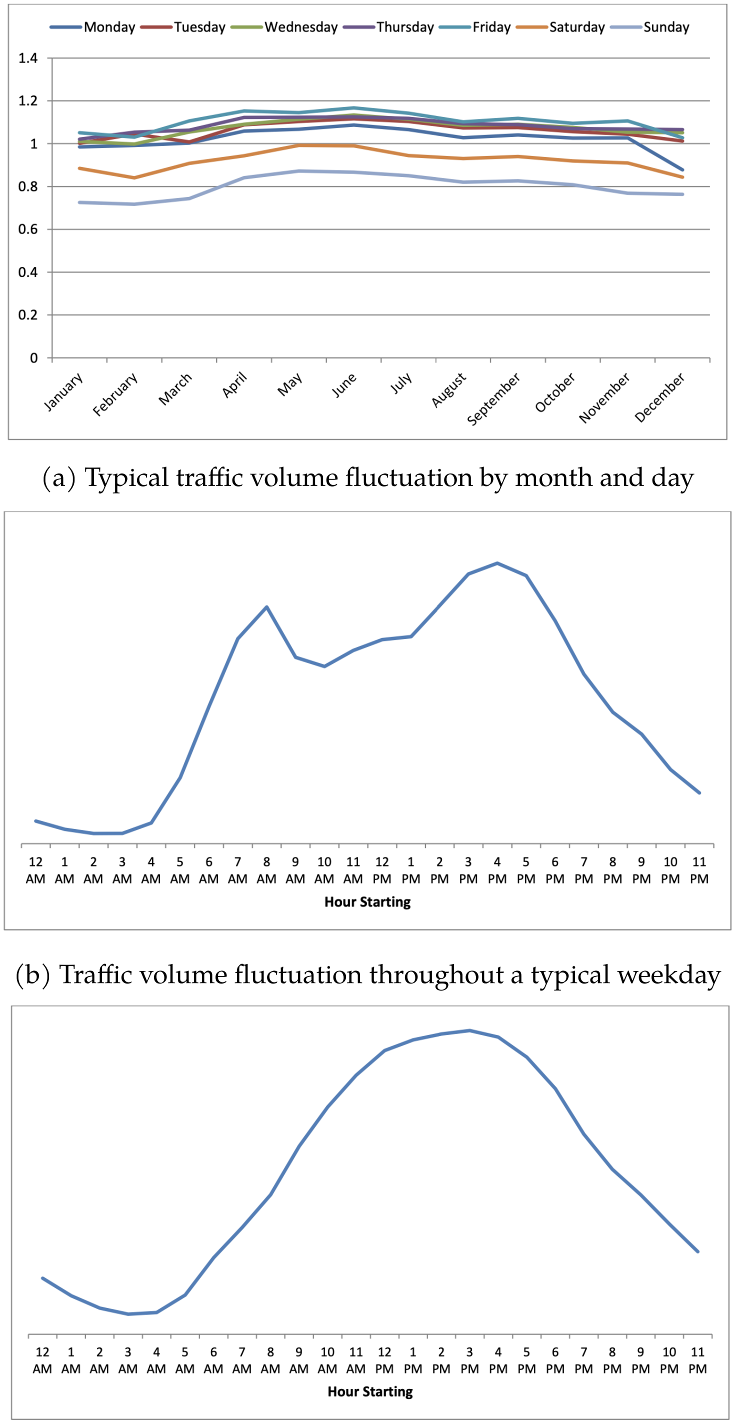
Year was only selected by the vehicle AADT model as an explanatory variable and the relationship was only linear. While this could be the true relationship, it was more likely due to the fact that only two years of data was available on average for each intersection. The final recommendation we would make regarding the response variable is to take consistent observations over more years for each count location, so that the temporal autocorrelation can be examined more in depth.

As for explanatory variables, bus stop usage was selected by variable selection but excluded from the final model primarily because including it would cut the size of the training set to less than a half. It may be worthwhile to reconsider this decision when bus stop usage data becomes available for a longer period of time. Furthermore, we would recommend adding the year built information to the spatial explanatory variables where possible, essentially making them spatiotemporal. For instance, if an increase in pedestrian volume was due to a new nearby bus stop, the model would not be able to know, and may falsely attribute the increase to year, when in fact it is the distance to the nearest bus stop.

5 Shiny App Dashboard

5.1 Purpose

Our goal in making the Shiny App is to create a platform where our analysis can be explored and visualized. The interactive web app allows the user to view the provided default and adjust weighting criteria in a stepwise function by customizing the score for each variable within both indices. This new weighting criteria can be saved for future reference. Additionally, in terms of visualization, the user is able to view the interaction between the intersection and its surroundings on a map, in a donut chart, and summary plot of our AADT models.



(a) Typical traffic volume fluctuation by month and day

(b) Traffic volume fluctuation throughout a typical weekday

(c) Traffic volume fluctuation throughout a typical weekend or holiday

Figure 6: AADT conversion factors

5.2 Features

5.2.1 Dashboard Layout

The Shiny App Dashboard, shown in Figure 7, consists of the main panel, initially shown in full screen, and a sidebar, which is accessible by clicking on the icon beside the title bar. The sidebar contains a menu to access a map with intersections (PPI/PDI Map), a map for the crash hotspot analysis (Ped Crash Map), the scoring matrix (PPI Weight and PDI Weight), and finally the scores in tabular form (Weight Tables).

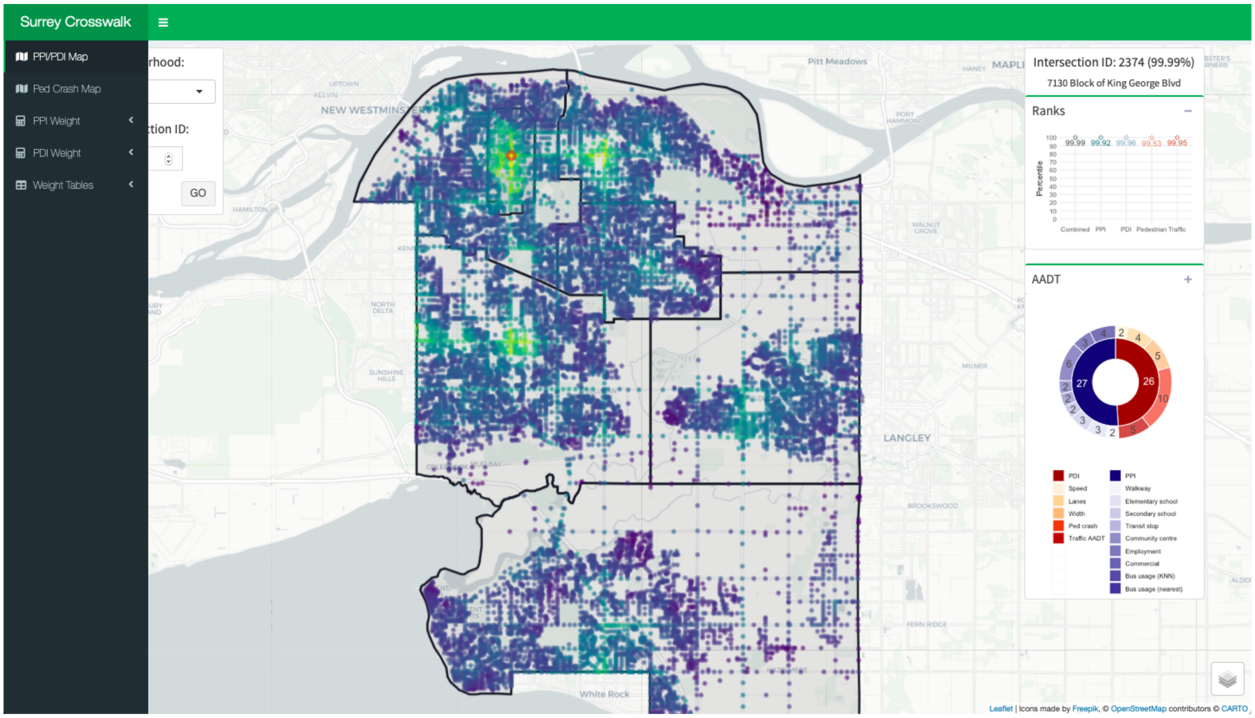

Intersection ID: 2374 (99.99%)
7130 Block of King George Blvd

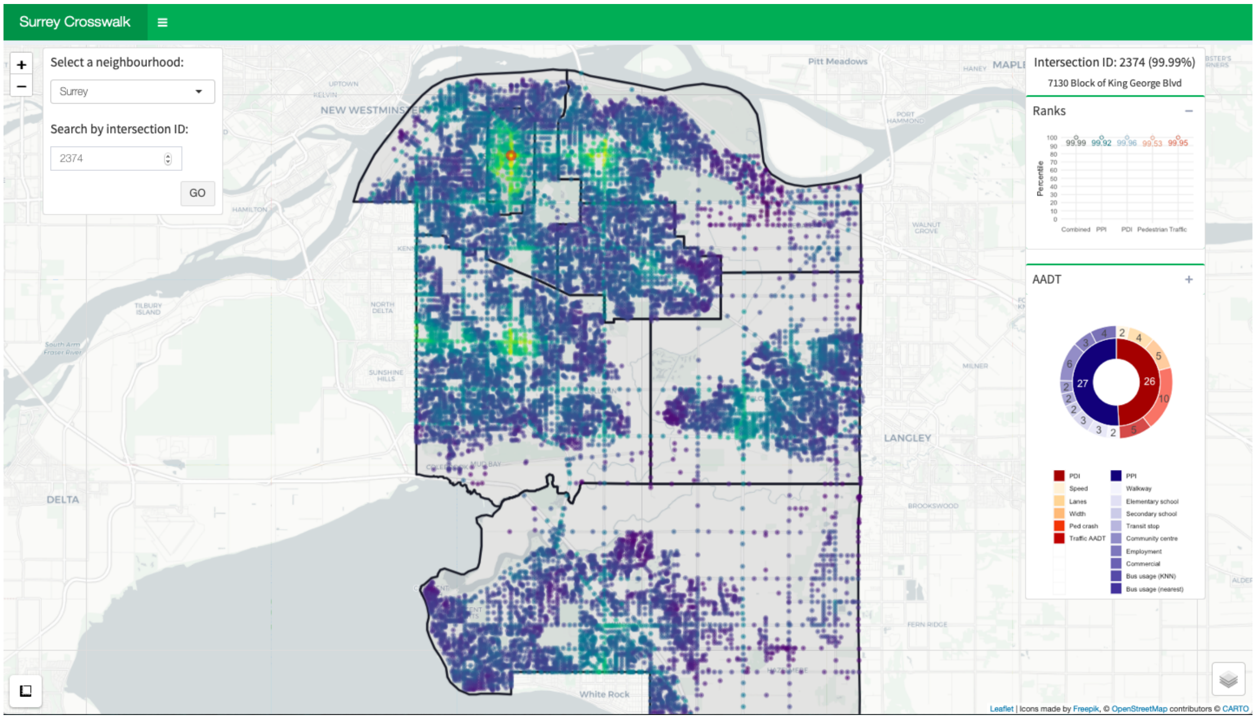
Figure 7: Main interface

5.2.2 PPI/PDI Map

The first map (see Figure 8) we see is the map that plots all intersections in Surrey. The dark border determines the border in which the intersections are included. Each circular marker on the map are intersections that are color-coded according to the score. The legend to the colors is provided in the bottom left corner—the lighter the color, the higher the score. The user is able to click on each intersection to see an updated panel with the corresponding AADT numbers and breakdown of the score and rankings. The bottom right corner is a legend to various layers of places of interest.

Filtering out Intersections

With the selection widget (see Figure 9) on the left hand side panel, the user is able to choose to see all intersections or a subset of intersections that currently do not have a crosswalk, which excludes intersections with traffic signals and existing crosswalks. The percentile ranking (seen in the right panel and in the Weight Tables tab) will be reflective of the chosen set of intersections. We chose to give this option because I hope that this app can be used for other purposes, not exclusive to


Surrey Crosswalk

Select a neighbourhood: Surrey

Search by intersection ID: 2374

GO

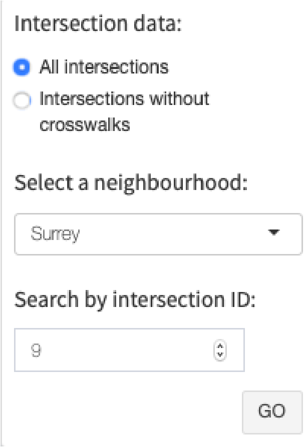
Intersection ID: 2374 (99.99%)

7130 Block of King George Blvd

Figure 8: PPI/PDI map

crosswalk building. By providing the index scores and AADT data for each intersection, other patterns can be identified and more road safety features can be implemented.

Additionally, the user is able to select a neighborhood in order to filter out intersections outside of the area of interest. Below that is the intersection ID search bar, which allows the user to search for a specific intersection by ID number. This search bar is not able to recognize non-integer input, therefore cannot search via intersection name.


Intersection data:

● All intersections

○ Intersections without crosswalks

Select a neighbourhood: Surrey

Search by intersection ID: 9

GO

Figure 9: Selection widget

Intersection Details and Plots

The panel that starts on the right hand side of the map is movable and has the intersection ID, name, and percentile ranking as a header. The ranking provided is a global ranking of all intersections in the City of Surrey. This panel updates to every intersection the user clicks (see Figure 10).

There are two expandable sub panels and a donut chart. The first sub panel is the ranking panel that plots the global ranking and the percentile of each component: PPI score, PDI score, and Pedestrian traffic. The second sub panel is the summary of our AADT model with the actual AADT value, if provided. Refer back to Section 4 to explore our AADT model. The blue bars are the confidence intervals for the pedestrian AADT and the orange bars are the confidence intervals for the vehicle AADT.

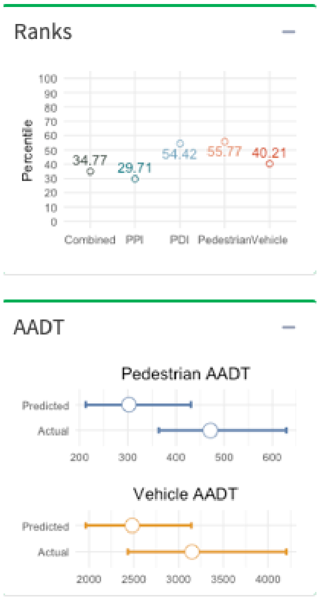


Figure 10: Intersection rank and AADT

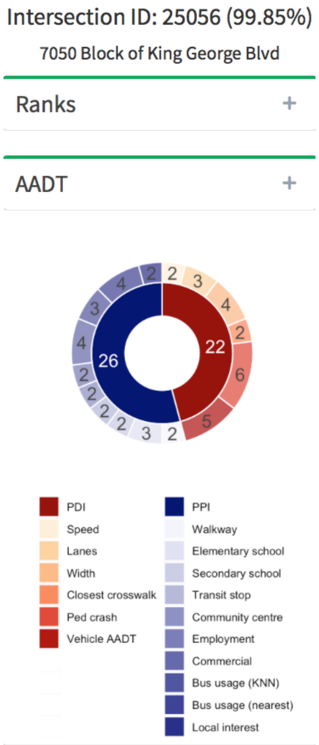
The donut chart (see Figure 11) shows the proportion of the score that each index contributes to. The inner donut is the PPI and PDI score, and the outer donut is the proportion of the variables of the respective index. The legend only shows the variables present, hence each legend is specific for each intersection. For ease of matching the colors to the value, the outer donut variables are listed on the legend clockwise.

5.2.3 Pedestrian Crash Map

While this map (see Figure 12) does not directly serve the objective of this project, it sheds light on the pedestrian crash patterns in Surrey. This map allows the user to visualize our emerging hotspot analysis by exploring the color coded markers with the legend located in the bottom right corner. The intersections with no pattern detected are not shown on this map.

5.2.4 PPI Weight and PDI Weight

The third and fourth submenu in the sidebar is PPI Weight and PDI Weight, shown in Figure 13. In these two tabs, the user is able to explore the data structure of each variable in the index. The


Intersection ID: 25056 (99.85%)

7050 Block of King George Blvd

Figure 11: Donut chart of factors

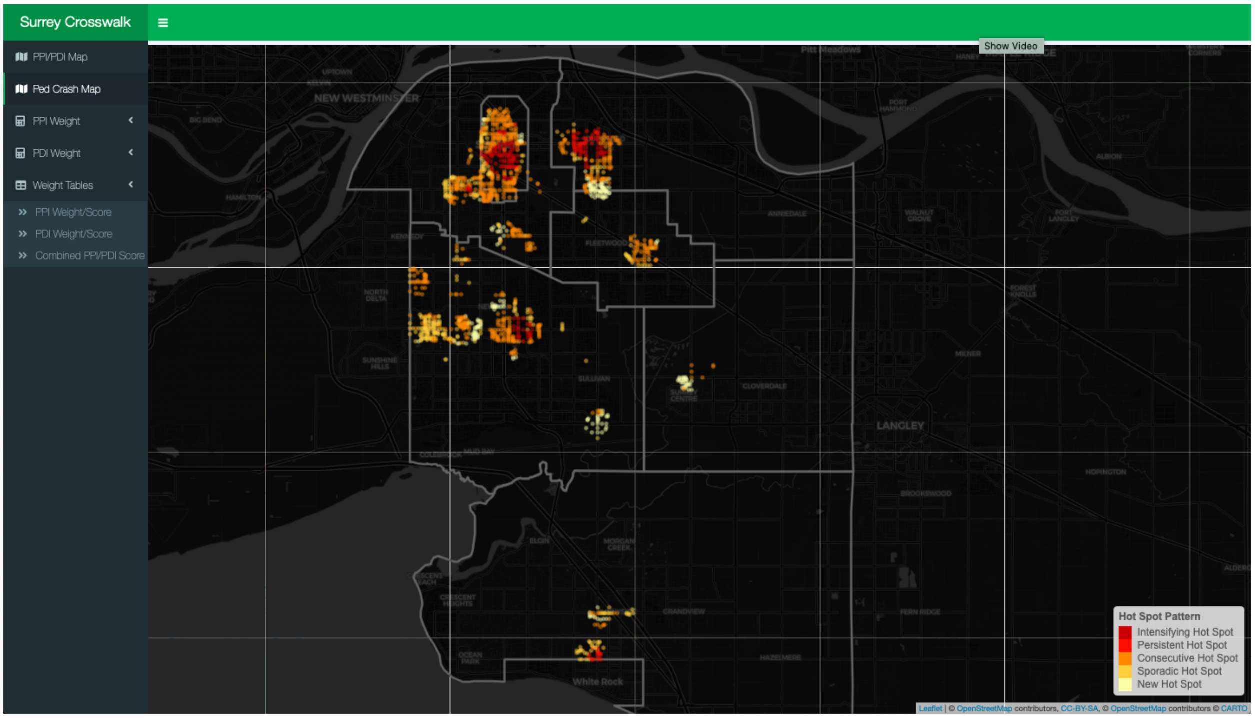
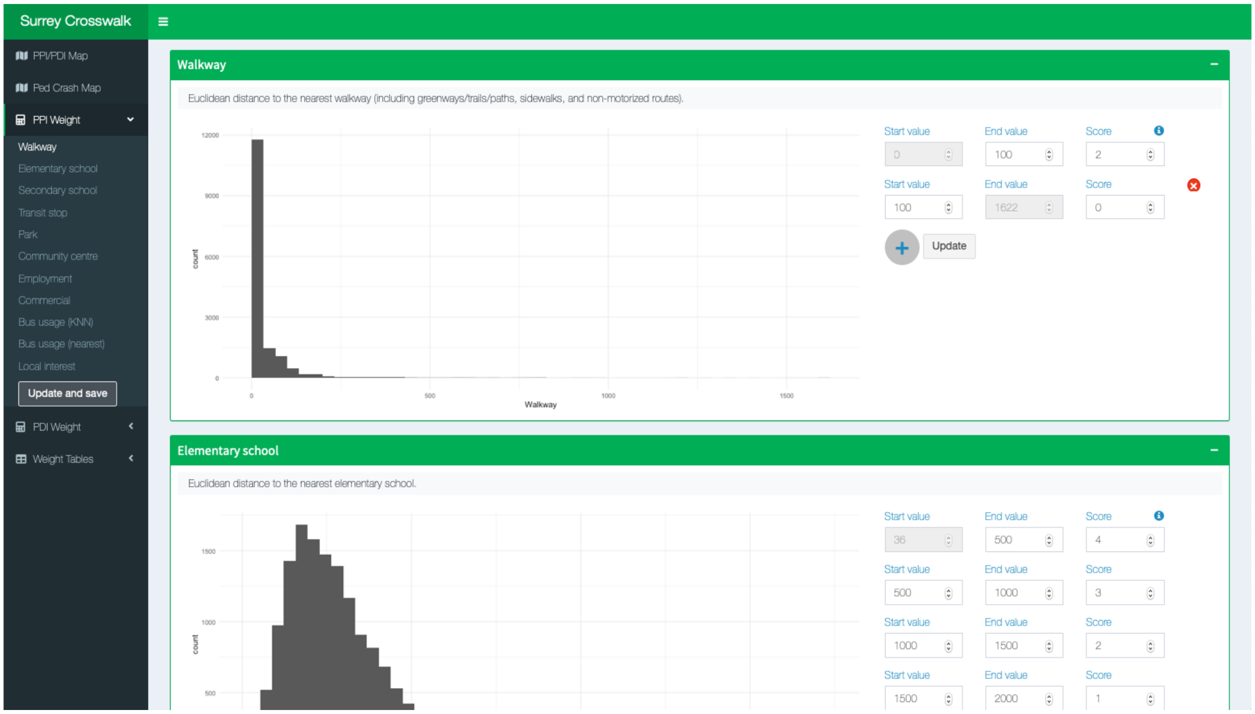


Figure 12: Crash map

shortcut link on the sidebar will bring the user to the specific cell. For example, PPI Weight has 11 cells and PDI has 6 cells, which corresponds to the number of variables that make up the respective scores. In each cell, the variable name, description, data overview and scoring matrix is available. The histogram and bar charts provided are the raw data for each variable, with the description describing the x axis. The counts (y-axis) for the plots are inclusive of the subset of intersections that are not within 100 meters of existing pedestrian traffic signals and/or crosswalks. We chose to exclude intersections with pedestrian traffic signals and crosswalks because the index scores serve to find potential locations to build a new crosswalk, thus the scores should be assigned relative to only those locations without a crosswalk.



Walkway

Euclidean distance to the nearest walkway (including greenways/trails/paths, sidewalks, and non-motorized routes).

Start value	End value	Score
0	100	2
100	1622	0

Elementary school

Euclidean distance to the nearest elementary school.

Start value	End value	Score
36	500	4
500	1000	3
1000	1500	2
1500	2000	1

Figure 13: PPI/PDI weight

Scoring Matrix

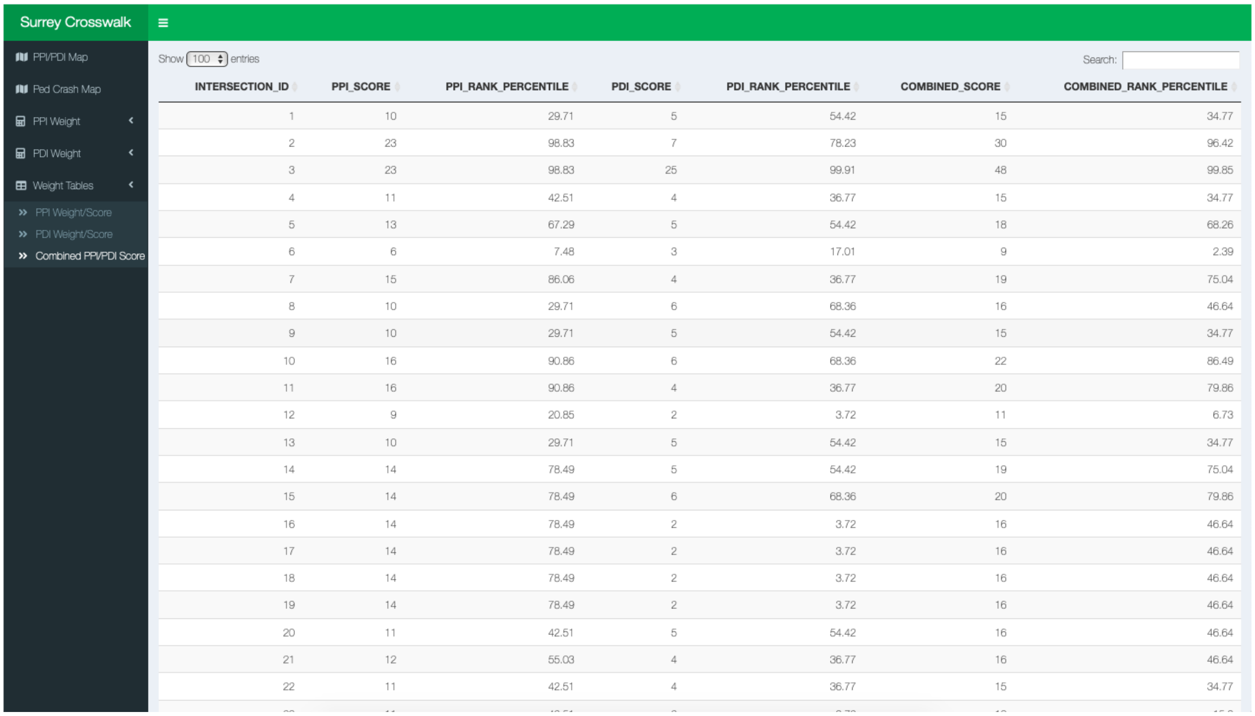
The scoring matrix starts with the default value provided by the team through referencing other cities' and consulting with members from the City of Surrey. The matrix is provided in a stepwise function and the user has the flexibility to add or remove intervals. The scores are inclusive of the start value and exclusive of the end value. The minimum and maximum value reflects the dataset's distribution. For example, the cell for walkway under PPI Weight has grey cells for 0 and 1622, the starting and ending value respectively. The two values correspond to the minimum rounded down to the nearest integer, 0 meters, and maximum rounded up to the nearest integer, 1622 meters, of the euclidean distance to the nearest walkway from an intersection that does not have a crosswalk or a pedestrian traffic signal.

The update button will overlay the scoring plot on top of the data's bar chart to assist with visualizing the step-wise scoring matrix. Once the user is confident with their new scoring matrix, the save button located in the sidebar allows the user to (1) save the scoring matrix in a CSV file and (2) see the updated score on the map, PPI/PDI Map, and rankings, Weight Tables.

It is important to note that the scoring is not normalized between PDI and PPI. Therefore, PPI may be overrepresented due to the larger number of variables contributing to the score. We chose to leave the normalization up to the user, for we provide information on the percentile ranking of each index independently as well as the combined score.

5.2.5 Weight Tables

The weight tables, shown in Figure 14, include three sub tabs: PPI Score, PDI Score, and Combined PPI/PDI Score. PPI Score and PDI Score summarizes the scores in a tabular format. All the tables are updated once the user changes the scoring matrix and clicks save under the PDI Weight or PPI Weight tab. Additionally, the tables will show the intersections selected by the user in the PPI/PDI Map: all intersections or intersections without crosswalks. The Combined PPI/PDI Score shows the rank percentile of each intersection by the individual score and the combined score. We chose to include the rank percentile to give a sense of the relative priority compared to other intersections in Surrey.



INTERSECTION_ID	PPI_SCORE	PPI_RANK_PERCENTILE	PDI_SCORE	PDI_RANK_PERCENTILE	COMBINED_SCORE	COMBINED_RANK_PERCENTILE
1	10	29.71	5	54.42	15	34.77
2	23	98.83	7	78.23	30	96.42
3	23	98.83	25	99.91	48	99.85
4	11	42.51	4	36.77	15	34.77
5	13	67.29	5	54.42	18	68.26
6	6	7.48	3	17.01	9	2.39
7	15	86.06	4	36.77	19	75.04
8	10	29.71	6	68.36	16	46.64
9	10	29.71	5	54.42	15	34.77
10	16	90.86	6	68.36	22	86.49
11	16	90.86	4	36.77	20	79.86
12	9	20.85	2	3.72	11	6.73
13	10	29.71	5	54.42	15	34.77
14	14	78.49	5	54.42	19	75.04
15	14	78.49	6	68.36	20	79.86
16	14	78.49	2	3.72	16	46.64
17	14	78.49	2	3.72	16	46.64
18	14	78.49	2	3.72	16	46.64
19	14	78.49	2	3.72	16	46.64
20	11	42.51	5	54.42	16	46.64
21	12	55.03	4	36.77	16	46.64
22	11	42.51	4	36.77	15	34.77

Figure 14: Weight table

5.3 Analysis Example

An example of an analysis that can be done using this tool is as follows. First, the user can check the highest ranking intersection, for which the default will be the intersection ID 2374. The user is then able to go to the map and search for the intersection using the search bar. The intersection is between 104 Avenue and King George Blvd. The user is able to see that the AADT for both pedestrians and traffic is relatively high and the percentile for all indexes is above 90%. The double donut chart tells the user that many variables are contributing to the high score, predominantly the pedestrian crash score. Next, the user can look at different layers to visualize the interaction between this intersection and its environment. In this case, there is already a crosswalk built in this intersection, which may call for a closer look at this intersection to improve the safety of the road and decrease the pedestrian crash numbers. The user can then flag this intersection to their team members and carry forward a program to reduce the safety-risk in intersection 2374.

5.4 Limitations

A limitation of the app is that it does not have a feature that will update the raw data for the PDI and PPI raw values. For instance, as crosswalks are built and crash numbers change it is important to reflect an updated score for the surrounding intersections. Currently, a dataset update will call for a manual data clean-up and wrangling, therefore further work in implementing the infrastructure to update the data used to determine the score for each intersection is a priority.

6 Conclusions

We have developed a tool to prioritize and describe the potential intersections that might need a crosswalk installed.The city aims to improve the safety of the city's most vulnerable road users, including pedestrians and cyclists as part of Surrey's Vision Zero Safe Mobility Plan. This includes minimizing the discrepancy between higher and lower income areas for the numbers of pedestrian crashes.

Using the AADT models, the Deficiency and Potential Index will enable the City of Surrey to make data-driven decisions regarding crosswalk implementation as opposed to relying solely on resident service requests. We hope that our project will help create more equity by using data to inform decision-making regarding pedestrian safety features, therefore lowering the inequalities created by focusing only on areas with resident requests. However, to use the scoring and visualization tool most effectively, there has to be an understanding of the most important variables that drives safety concerns at each intersection. The Shiny App provides an online tool to explore the locations of high-risk intersections determined by the scoring matrix which the user can change and customize by setting scoring criteria for each factor within the Deficiency and Potential indices. Moreover, the app has a feature that allows the user to apply different statistical transformations (logarithmic, stepwise, etc.) to the data to create more continuous and nuanced scores. This can help with the decision-making process in terms of differentiating and prioritizing intersections by having less discrete values.

6.1 Social Impact and Data Bias

This project is part of the Data Science for Social Good program at the Data Science Institute. As researchers it is important to acknowledge the strengths and limitations of our methodology and results from our project. Although the results were satisfactory and fulfilled the initial project goals it is important to consider the inherent data bias. The data we were given access to for the project, influenced our ability to model and analyze. For instance, an important consideration for building crosswalks is the pedestrian volume measured in "Equivalent Adult Units" which are used to account for pedestrian age and physical ability. However, we did not have disaggregated data for age or physical ability thus our analyses, models and results excluded these important considerations which can obscure the mobility needs of youth, senior citizens, and pedestrians that have different accessibility needs such as visual, walking, and hearing aid. Despite this flaw in our project, we hope that our work can inspire further research on using data-driven tools to minimise data bias, create more diversity on data sources and better inform decision-makers.

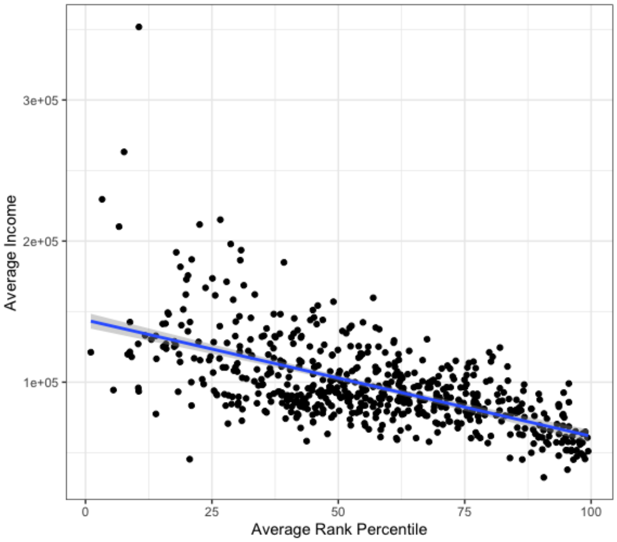


Figure 15: Intersections were grouped into dissemination areas. A negative relationship between the average income and average rank percentile suggests that the dissemination area with lower average income is likely to have higher need for crosswalks.

An analysis on the relationship between average income and average rank percentile is shown in Figure 15. We found that intersections that belong to dissemination areas with lower average income are significantly more likely to have a higher need for road safety features implementation. This finding supports the city's observation that higher income areas tend to have more attention, but our scoring matrix is able to shed light beyond the service requests by providing a data-driven priority index.

Acknowledgments

We would like to thank our partners from the City of Surrey Nathan Baugh,Kristen Tiede, and Shabnem Afzal for providing us guidance and support throughout the project.

We would also like to thank members of The University of British Columbia's Data Science Institute, Data Science for Social Good program, particularly Raymond Ng, Kevin Lin, Jean-François Rajotte, and Joe Watson for sharing their expertise in research methodology, data science and for their ongoing support of the project.

A Emerging Hotspot Analysis

This type of analysis identifies trends in the data—in this case, ICBC collision data. The analysis finds new, intensifying, diminishing and sporadic hot and cold spots. First, each collision data point is converted into space-time bins: "a three-dimensional cube made up of space-time bins with the x and y dimensions representing space and the t dimension representing time. Every bin has a fixed position in space (x, y) and in time (t). Bins covering the same (x, y) area share the same location ID. Bins encompassing the same duration share the same time-step ID."[1]

[1] https://pro.arcgis.com/en/pro-app/latest/tool-reference/space-time-pattern-mining/create-space-time-cube.htm

Table 9: Hotspot categories. For a full list of categories, please visit `https://pro.arcgis.com/en/pro-app/latest/tool-reference/space-time-pattern-mining/learnmoreemerging.htm`.

Category	**Description**
No pattern detected	Does not fall into any of the hot or cold spot patterns defined below.
New hotspot	A location that is a statistically significant hotspot for the final time step and has never been a statistically significant hotspot before.
Consecutive hotspot	A location with a single uninterrupted run of statistically significant hotspot bins in the final time-step intervals. The location has never been a statistically significant hotspot prior to the final hotspot run and less than ninety percent of all bins are statistically significant hotspots.
Intensifying hotspot	A location that has been a statistically significant hotspot for ninety percent of the time-step intervals, including the final time step. In addition, the intensity of clustering of high counts in each time step is increasing overall and that increase is statistically significant.
Persistent hotspot	A location that has been a statistically significant hotspot for ninety percent of the time-step intervals with no discernible trend indicating an increase or decrease in the intensity of clustering over time.
Sporadic hotspot	A location that is an on-again then off-again hotspot. Less than ninety percent of the time-step intervals have been statistically significant hotspots and none of the time-step intervals have been statistically significant cold spots.

Once the bins are created, a statistic called Gi* is calculated for each value in the data set. The Gi* number returns a z-score and a p-value.[2] These values indicate where the features—either high or low values—cluster spatially. For statistically significant positive z-scores, the larger the z-score is, the more intense the clustering of high values (hotspot). For statistically significant negative z-scores, the smaller the z-score is, the more intense the clustering of low values (cold spot).[3] The result are the categories in Table 9, which we used to classify the automobile-pedestrian collision patterns in the Deficiency Index and the bus usage pattern in the Potential Index.

References

[1] City of Surrey. Vision Zero Surrey safe mobility plan 2019–2023. Technical report, City of Surrey, 2019.

[2] Alexandra Frackelton. Pedestrian transportation project prioritization incorporating app-collected sidewalk data. Master's thesis, Georgia Institute of Technology, 2013.

[2] Z-scores are standard deviations. If, for example, a tool returns a z-score of +2.5, one would say that the result is 2.5 standard deviations.

[3] `https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/h-how-hot-spot-analysis-getis-ord-gi-spatial-stati.htm`

[3] Cuiping Zhang, Xuedong Yan, Lu Ma, and Meiwu An. Crash prediction and risk evaluation based on traffic analysis zones. *Mathematical Problems in Engineering*, 2014:1–9, 2014.

[4] Transport Canada. Transportation in Canada 2019: Overview report. Technical report, Minister of Transport, 2019.

[5] District of North Vancouver. Pedestrian master plan: Final report. Technical report, District of North Vancouver, 2009.

[6] Erica Geddes and Terry Fjellstrom. Pedestrian network study. Technical report, City of Prince George, 2004.

[7] City of Victoria. Pedestrian master plan: Final report. Technical report, City of Victoria, 2008.

[8] Office of Transportation. Portland pedestrian master plan. Technical report, City of Portland, 1998.