Pedestrian Collision Prevention Model – Literature Review
Jason Kim

Introduction

Just a few months ago, the City of Toronto approved $22 million in additional funding for accelerating its Vision Zero Road Safety Plan. The goal of this plan is to eliminate pedestrian deaths from car collisions by 2021. Meanwhile, 2018 is on pace to be one of the worst years for pedestrian and cyclist fatalities and 2017 was also a peak year at 162 deaths or major hospitalizations.

The purpose of this data analysis project is to identify and prioritize zones within Toronto that have a high risk of pedestrian and cyclist collisions so that safety improvements could be more efficiently implemented by the City. I propose to use various clustering algorithms in order to identify and prioritize these high risk zones. These cluster attributes will then be analyzed and profiled to give insight on what features increase the probability of a collision. Finally, a visualization of the high risk zones and their priority level will be created to provide the City with a clear plan of action.

Literature Review

Although automobile collisions with pedestrians and cyclists have long ailed large cities across the globe, the use of data analysis in collision prevention and transportation planning is a relatively recent phenomenon. Much of the academic interest in this area has been in using various clustering algorithms to profile and predict high collision zones. Notably, Anderson's oft-cited study uses both kernel density estimation (KDE) and k-means clustering to identify collision hotspots in London, UK, and following her lead, many other scholars use either one or both methods in other locales (for Seattle see Quistberg et al., 2015; for Japan see Hashimoto et al., 2016; for Tunisia Ouni & Belloumi, 2018).

KDE is commonly used in GIS software to identify spatial patterns on a 2D map divided into cells. It remains a popular choice in the field of transportation studies due to the widespread use of GIS software in the discipline, though increasingly the tools of data science are being used. In KDE, a search radius is selected and the density of collision points in each cell that fall within the area of the circle are grouped into "humps" or kernels. These kernel densities are often visualized as heat maps. So as one might expect, areas with a high density of collisions would have a taller hump and are easily discernable visually. One benefit of this method is that unlike k-means, the number of clusters does not need to be set ahead of time, but only the search radius (also known as bandwidth). One downside to this method is that it provides no information about what characteristics influence the density of collisions. Further, changing the bandwidth could drastically effect what is considered to be a hot zone.

Alternatively, k-means clustering is also commonly used in collision prediction studies. Besides its simplicity, its prevalence could be explained by its ability to identify unknown groups by shared attributes. These shared cluster attributes are then used by data analysts and traffic researchers to profile collision zones and thus provide concrete features to target for intervention. For example, Vasavi uses k-means clustering to identify hot zones then uses the a

priori algorithm to find association rules for attributes that lead to a high probability of collision (Vasavi, 2018).

Many researchers pick either KDE or k-means clustering, followed by a classification algorithm of their choice – the simplest studies opting to use variations of multiple linear regression and especially Poisson regression (collisions are count data so it should always follow Poisson distribution if the sample is of sufficient size), (Quistberg et al., 2015). Another study on collisions involving schoolchildren in Toronto used negative binomial regression, which is essentially Poisson regression (Rothman, Macarthur, To, Buliung, & Howard, 2014).

In summary, the field of transportation research and accident prevention could be much enhanced with more sophisticated data analysis approaches and tools, but there are signs that the field is headed in this direction with the prevalence of k-means clustering. My project will contribute further to this field by testing various clustering techniques against each other as well as by including more detailed categorical data due to the several join operations I have performed. In particular, I believe my use of density-based clustering or Gaussian mixed method should significantly outperform k-means since it combines the best aspects of the KDE method without the downside of k-means' poor performance on noisy data.

Dataset Description

The dataset used in this study was created by transforming and joining various open access datasets to the primary dataset.

These datasets were joined because I wanted to associate each collision in the original collisions dataset to a wider sociodemographic context. So for example, each collision is now associated with one of the 140 unique neighbourhoods, and each of these neighbourhoods in turn have unique sociodemographic profiles. This enables us to see whether high collision zones could be profiled and appropriately targeted for intervention.

The primary dataset is a heavily cleaned and processed version of the Collisions - Events and Collisions - Involved dataset provided during the 2018 Toronto Vision Zero Challenge.

Column Name	Type	Description
collision_id	Integer	The unique identifier of the dataset, linking this dataset with the Collisions Involved dataset. This is not the identifier used by Toronto Police Services, and is unique to this dataset.
collision_date	Date	Date when the collision took place (yyyy-mm-dd).
collision_time	Integer	Time when the collision occurred. Format is in "hhmm".
px	Integer	Primary identifier of all signalized intersection. Please see Traffic Signals.
street_1	Text	Street/address that the collision occurred on.
street_2	Text	Cross-street/address that the collision occurred on.
street_type_2	Text	Suffix/type of the second street.

direction_2	Text	Direction of traffic on the second street that the collision occurred on.
street_3	Text	Third street, if any, the collision occurred on.
street_type_3	Text	Suffix/type of the third street.
direction_3	Text	Direction of traffic on the third street that the collision occurred on.
location_class	Text	Class of the location on the road where the collision occurred.
location_desc	Text	Location on the road where the collision occurred.
collision_type	Text	Class/severity of the collision.
impact_type	Text	Type of impact that occurred in the collision.
road_class	Text	Class of road the collision occurred on.
visibility	Text	Visibility conditions at the time and location of the collision.
light	Text	Lighting conditions at the time and location of the collision.
road_surface_cond	Text	Condition of the road surface at the time and location of the collision.
longitude	Numeric	Longitude.
latitude	Numeric	Latitude.
collision_id	Integer	Collision unique identifier.
traffic_control	Text	Type of traffic control at the intersection/road (e.g. traffic light, stop sign etc.).
vehicle_class	Text	Class of the vehicle involved in the collision.
initial_dir	Text	Direction the vehicle/cyclist/pedestrian was travelling.
event1	Text	Object/vehicle involved in the collision.
event2	Text	Action the vehicle undertook.
event3	Text	Fixed object involved in the collision.
involved_class	Text	Type of involved individual.
involved_age	Text	Age of involved individual.
involved_injury_class	Text	Level of injury sustained by the individual.
safety_equip_used	Text	Type of safety equipment used by the vehicle.
driver_action	Text	Apparent driver action.
pedestrian_action	Text	Apparent pedestrian action.
pedestrian_collision_type	Text	Categorization of pedestrian collision type.
cyclist_action	Text	Apparent cyclist action.
cyclist_collision_type	Text	Categorization of cyclist collision type.
manoeuver	Text	Vehicle manoeuver.
ksi_check	binary	"1" if the collision was a KSI; "0" if not
intersection_check	binary	"1" if the collision was at an intersection
is_senior	binary	"1" if the collision involved someone 65+ in age
is_child	binary	"1" if the collision involved someone under 18 years old
daylight_check	binary	"1" if the collision took place in day light
visibilily_check	binary	"1" if the collision took place in clear conditions
road_surface_cond	binary	"1" if the collision took place with dry road conditions
arterial_check	binary	"1" if the collision took place on an arterial road
pedestrian_check	binary	"1" if the collision involved a pedestrian; "0" if cyclist

total_length	Numeric	The length of the street the collision took place on in kilometres

JOINED DATASETS

The following datasets were joined to the above primary dataset. Unless otherwise noted, the join was performed on the Neighbourhood ID associated with each collision.

Toronto Centreline (GIS)

We used the Centreline shapefile as-is and joined relevant attributes to the Collisions dataset using the Collisions dataset using a join on street names in Tableau Prep. We could have used k-NN and a spatial join as below, but this is computationally very expensive so we opted to join on street name rather than on GPS coordinates (spatial join).

This join made it possible to calculate the length of streets in kilometres.

Neighbourhood Boundaries (GIS)

We used the Neighbourhood shapefile as-is and joined relevant attributes to the Collisions dataset using a k-Nearest Neighbours to snap out-of-bounds points to neighbourhood polygons and a spatial join in QGIS.

Civics

This is a transformed version of the Wellbeing Toronto - Civics dataset available through the City of Toronto's Open Data Catalogue. We used 2011 measures when available, 2008 if not.

Column Name	Type	Description
Neighbourhood	Text	Neighbourhood name
Neighbourhood Id	Integer	Neighbourhood Id
City Grants Funding $	Integer	The amount in Canadian dollars that the neighbourhood received in city grant funding
Diversity Index	Float	Ethnic Diversity Index compiled from Statistics Canada Census 2006. This indicator reflects the ethnic diversity of a neighbourhood, by comparing how many ethnicities (such as Chinese, Scottish, Italian, etc.) there are in a given neighbourhood and how the proportions are distributed (for example, 20%/40%/40% or 5%/5%/90%). The entropy method

		of determining heterogeneity was used with Census 2006 Ethnic Origin Total Responses data. Higher values indicate greater ethnic diversity (more heterogeneity), lower values indicate lesser ethnic diversity (more homogeneity).
Voter Turnout	Float	Percentage of population that voted in 2008
Walk Score	int	Measures walkability on a scale from 0 - 100 based on walking routes to destinations such as grocery stores, schools, parks, restaurants, and retail. More information can be found on the walkscore.com website. Original raw data has been transformed from a 0-100 scale to Wellbeing Toronto's 1-100 scale.
Neighbourhood Equity Score	float	Composite indicator of 15 neighbourhood outcomes using data from Urban HEART@Toronto. Indicators measure outcomes related to economic opportunities, social development, participation in decision-making, physical surroundings, and healthy lives.
Salvation Army Donors	int	Number of Salvation Army Donors by Neighbourhood, computed from 6-digit postal codes from the Active Donors database for 2011 and aggregated to neighbourhoods.
equity_check	binary	If the equity score is *below* average, "1". If not, "0".
walk_check	binary	If the walk score is *below* average, "1". If not, "0".
diversity_check	binary	If the diversity score is *below* average, "1". If not, "0".
turnout_check	binary	If the voter turnout rate is *below* average, "1". If not, "0".

Economics

This is a transformed version of the Wellbeing Toronto - Economics dataset available through the City of Toronto's Open Data Catalogue. 2011 data is used where available; 2008 if not.

Column Name	Type	Description
Number of Businesses	int	Total number of licensed business establishments.
Child Care Spaces	int	Licensed child care spaces
Social Assistance Recipients	int	Count of recipients of aid qualifying for Ontario Works, Temporary Care (OW), Ontario Disability Support Program (ODSP) or Special Assistance (ODSP/OW) programs

Local Employment	int	Total local employment (jobs), persons aged 15+ years.
Debt Risk Score	int	The Risk Score is a proprietary index value provided by TransUnion Canada that indicates the likelihood of missing three consecutive loan payments. Low-value scores (<707) indicate a High Risk of missing 3 consecutive loan payments; High-value scores (769+) indicate a low risk. These risk scores are calculated for non-mortgage consumer debt such as lines of credit, credit cards, automobile loans and installment loans. TransUnion data is provided by postal code and covers approximately 92% of all Canadians with credit files. For privacy reasons, postal codes with fewer than 15 credit files are suppressed. TransUnion dataset provided by the Community Data Program.
Home Prices	int	Average price for residential real estate sales during the period 2011-2012, in Canadian dollars. Data collated by Realosophy.com.
businesses_check	binary	If the number of businesses is above city average, "1". If not, "0".
childcare_check	binary	If the number of child care spaces is above city-wide average, "1". If not, "0".
homeprice_check	binary	If the average house price is above the city-wide average, "1". If not, "0".
localemployment_check	binary	If the number of local jobs is above the city average, "1". If not, "0".
socialasst_check	binary	If the number of social assistance recepients is above city average, "1". If not, "0".

Neighbourhood Profiles

This is a processed and transformed version of the Neighbourhood Profiles dataset available through the City of Toronto's Open Data Catalogue. It contains data based on the 2016 Census, making it far more up-to-date than related datasets with similar measures like the Wellbeing Toronto datasets.

Column Name	Type	Description
Hood Name	string	Name of the neighbourhood
Hood ID	int	Unique id of neighbourhood

Population 2016	int	Population of neighbourhood according to 2016 census
Population density per square kilometre	float	Population density per sq km
Land area in square kilometres	float	Land area in sq km
Children (0-14 years)	int	Number of children living in area
Youth (15-24 years)	int	Number of youth living in area
Working Age (25-54 years)	int	Number of working aged people living in area
Pre-retirement (55-64 years)	int	Number of pre-retirement aged people living in area
Seniors (65+ years)	int	Number of seniors living in area
Older Seniors (85+ years)	int	Number of older seniors living in area
% Immigrants	float	Percentage of neighbourhood population that is an immigrant
% Visible Minority	float	Percentage of neighbourhood population that identify as a visible minority
Unemployment rate	float	The unemployment rate in percentage
Commute within census subdivision (CSD) of residence	int	Number of people who work and are age 15+ who commute within the area in which they live
% Commute in Car truck	float	Percentage of people who work and are age 15+ who commute to work primarily by car or truck
% Commute in Public transit	float	Percentage of people who work and are age 15+ who commute to work primarily by public transit
% Commute in Walked	float	Percentage of people who work and are age 15+ who commute to work primarily by walking
% Commute by Bicycle	float	Percentage of people who work and are age 15+ who commute to work primarily by bicycle
% Commute by Other	float	Percentage of people who work and are age 15+ who commute to work primarily by other means

Commuting duration - Less than 15 minutes	int	Number of people whose commute lasts the listed duration
Commuting duration - 15 to 29 minutes	int	Number of people who work and are age 15+ whose commute lasts the listed duration
Commuting duration - 30 to 44 minutes	int	Number of people who work and are age 15+ whose commute lasts the listed duration
Commuting duration - 45 to 59 minutes	int	Number of people who work and are age 15+ whose commute lasts the listed duration
Commuting duration - 60 minutes and over	int	Number of people who work and are age 15+ whose commute lasts the listed duration
% Time leaving for work - Between 5 a.m. and 5:59 a.m.	float	Percentage of people who work and are age 15+ who leave for work during the listed time
% Time leaving for work - Between 6 a.m. and 6:59 a.m.	float	Percentage of people who work and are age 15+ who leave for work during the listed time
% Time leaving for work - Between 7 a.m. and 7:59 a.m.	float	Percentage of people who work and are age 15+ who leave for work during the listed time
% Time leaving for work - Between 8 a.m. and 8:59 a.m.	float	Percentage of people who work and are age 15+ who leave for work during the listed time
% Time leaving for work - Between 9 a.m. and 11:59 a.m.	float	Percentage of people who work and are age 15+ who leave for work during the listed time
% Time leaving for work - Between 12 p.m. and 4:59 a.m.	float	Percentage of people who work and are age 15+ who leave for work during the listed time
child_check	binary	If the neighbourhood has an above average number of children living in it a "1" is assigned; if not, "0".
senior_check	binary	If the neighbourhood has an above average number of seniors living in it a "1" is assigned; if not, "0".
minority_check	binary	If the neighbourhood has an above average number of visible minorities living in it a "1" is assigned; if not, "0".
immigrants_check	binary	If the neighbourhood has an above average number of immigrants living in it a "1" is assigned; if not, "0".

commute_car_check	binary	If the neighbourhood has an above average number of people who commute to work by car a "1" is asiigned; if not, "0".

Income

This is a transformed version of the Toronto Health Profiles - Income dataset. This dataset uses income measures from the 2016 Census and also uses Neighbourhoods as the unit of analysis, making it uniquely well-suited to join to our collisions dataset in order to see if there is a relationship between income and collision or injury risk.

Column Name	Type	Description
HOOD ID	int	Unique ID for neighbourhood
HOOD NAME	string	Name of neighbourhood
Total % In LIM-AT	float	Percentage of households who fall into the "low income measure - after tax" category according to the 2016 Census. This measure takes into account the reduced spending power of larger households. You can learn more about the LIM-AT from Statistics Canada.
lim_check	binary	If an above average number of households fall under the LIM-AT category, "1", If not, "0".

Language

This is a transformed version of the Toronto Health Profiles - Language Spoken Most Often at Home dataset. This dataset uses income measures from the 2016 Census and also uses Neighbourhoods as the unit of analysis, making it uniquely well-suited to join to our collisions dataset. Since the collision reports do not record the ethnic, linguistic, or newcomer status of collision victims, we used the prevalence of non-official languages spoken at home in the area the collision occurred as one of our proxy variables to see if there is a relationship between newcomer status and collisions.

Column Name	Type	Description
HOOD ID	int	Unique ID for neighbourhood
HOOD NAME	string	Name of neighbourhood
MOST SPOKEN NON-OFFICIAL LANG	string	The most spoken non-official language at home

2ND MOST SPOKEN NON-OFFICIAL LANG	string	The 2nd most spoken non-official language at home
3RD MOST SPOKEN NON-OFFICIAL LANG	string	The 3rd most spoken non-official language at home
% OFFICIAL LANG AT HOME	float	Percentage of people who primarily speak one of Canada's official languages at home
% NON-OFFICIAL LANG AT HOME	float	Percentage of people who primarily speak a non-official language at home
language_check	binary	If an above average percentage of people speak a non-official language at home, "1". If not, "0".

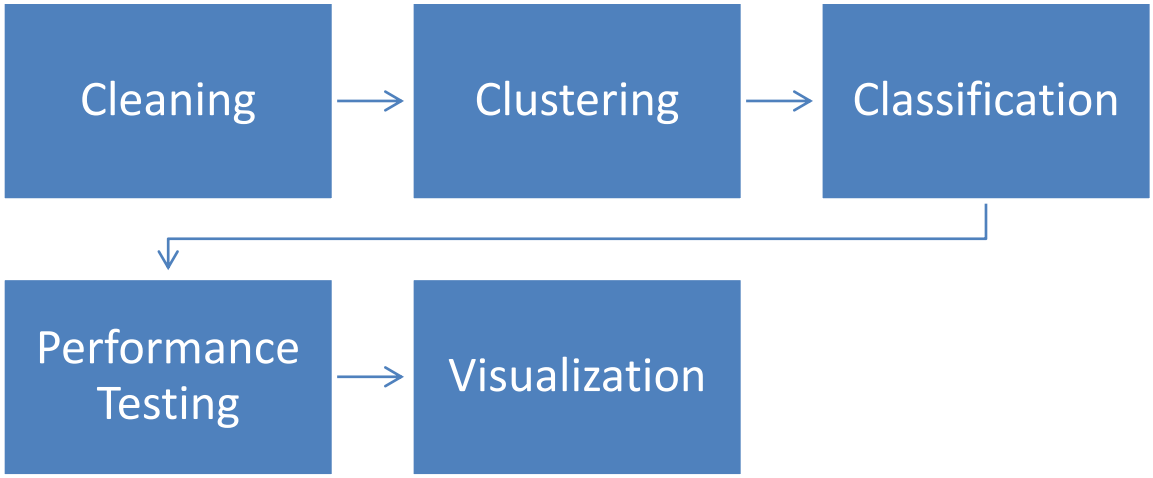
Transportation

This is a transformed version of the Wellbeing Toronto - Transportation dataset available through the City of Toronto's Open Data Catalogue. We used 2011 measures when available, 2008 if not.

Column Name	Type	Description
Neighbourhood	Text	Neighbourhood name
Neighbourhood Id	Integer	Unique Neighbourhood Identifier
TTC Stops	Integer	Total number of TTC stops
Pedestrian/Other Collisions	Integer	Total number of pedestrian, cyclist and private property collisions
Traffic Collisions	Integer	Total number of traffic collisions
Road Kilometres	Integer	Total number of road kilometres
Road Volume	Integer	Total 24-hour volume (collector roads only)
ttc_check	binary	If the number of TTC stops is above the average, "1". If not, "0".
road_km_check	binary	If the total road kilometrage is above the average, "1". If not, "0".
road_vol_check	binary	If the total 24-hour volume of traffic on collector roads is above average, "1". If not, "0".

Pedestrian Collision Prevention Model – Literature Review
Jason Kim

Approach



Step 1: Cleaning

Attributes with near zero variance or at least 50% missing values were removed. A very high pair-wise correlation was found between two attributes – collisions per km and Killed or Seriously Injured (KSI) per km (Pearson's R of 0.93, p-value < 0.05). Due to this, these variables are likely not independent of each other and so collisions per km was retained and KSIs per km was dropped since there are far more collisions than KSIs in the dataset.

Step 2: Clustering

Three clustering methods will be tested:

- Density-based spatial clustering (DBSCAN), epsilon selected based on k-distance threshold
- K-means clustering, k selected based on elbow method
- Gaussian mixed method

Step 3: Classification

Based on the cluster attributes, high risk zones will be identified using:

- DBSCAN (?)
- Random Forest
- Logistic Regression

I'm not completely sure yet how I'd classify clusters, but I imagine it would involve appending the cluster ID as a column to the original dataset and remove records that do not fall within any clusters since those points would be considered noise by an algorithm like DBSCAN.

The goal of classification here is to identify not only the high risk zones (clusters), but also what cluster characteristics influence the risk. For example, a certain section of road may be located in a neighbourhood with high numbers of seniors – a feature that could increase the probability of a collision.

Step 4: Model Evaluation

The classification model will be tested against the 4-year period preceding the one the model was trained on. In other words, the training set will be based on pedestrian collisions from 2013-16, while the test set will be based on 2009-2012.

OR

The classification model will be tested using 10-fold cross validation based on pedestrian collisions from 2013-2016.

Step 5: Mapping and Visualization

The high risk zones will be mapped to their actual locations and an interactive visualization will be created in Tableau. This tool could then be used by the City to prioritize intervention sites and communicate and explain their plan of action to both government and citizen stakeholders.