STATE OF CALIFORNIA • DEPARTMENT OF TRANSPORTATION

TECHNICAL REPORT DOCUMENTATION PAGE

TR0003 (REV 10/98)

Lock Data on Form

ADA Notice

For individuals with sensory disabilities, this document is available in alternate formats. For information call (916) 654-6410 or TDD (916) 654-3880 or write Records and Forms Management, 1120 N Street, MS-89, Sacramento, CA 95814.

1. REPORT NUMBER	2. GOVERNMENT ASSOCIATION NUMBER	3. RECIPIENT'S CATALOG NUMBER
CA19-3110		

4. TITLE AND SUBTITLE

Prioritizing High Collision Concentration Locations (HCCLs) identified using Pedestrian Safety Monitoring Report Tool

5. REPORT DATE

August 24, 2019

6. PERFORMING ORGANIZATION CODE

7. AUTHOR

Jiajian Lu, Aditya Medury, and Offer Grembek

8. PERFORMING ORGANIZATION REPORT NO.

9. PERFORMING ORGANIZATION NAME AND ADDRESS

UC Berkeley Safe Transportation Research & Education Center
2614 Dwight Way, #7374
Berkeley, CA 94720-7374

10. WORK UNIT NUMBER

11. CONTRACT OR GRANT NUMBER

65A0690

12. SPONSORING AGENCY AND ADDRESS

California Department of Transportation
Division of Research and Innovation, MS-83
1227 O Street
Sacramento CA 95814

13. TYPE OF REPORT AND PERIOD COVERED

Final Report

14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

16. ABSTRACT

The Pedestrian Safety Improvement Program is an effort of the California Department of Transportation (Caltrans) to identify and address problems with regard to pedestrian safety in California, with the long-term goal of substantially reducing pedestrian fatalities and injuries in California. The research presented in this report explores different crash frequency-based prioritization techniques to provide more robust metrics to identify high collision concentration locations. In particular, we evaluate methodologies to statistically quantify the presence of recurring crash patterns, as well as methods to mitigate regression-to-the-mean phenomena in the absence of safety performance functions. These comparisons were conducted using both empirical and simulated datasets. The empirical findings indicate that metrics proposed for pattern recognition may potentially mimic pattern identification process of investigators. However, the use of simulation reveals that the accuracy of these methods may depend on the amount of overdispersion present in the crash population, which needs further exploration.

17. KEY WORDS	17. DISTRIBUTION STATEMENT
	No restrictions. This document is available to the public through the National Technical Information Service, Springfield, VA 22161

19. SECURITY CLASSIFICATION *(of this report)*	20. NUMBER OF PAGES	21. COST OF REPORT CHARGED
Unclassified	32	N/A

DISCLAIMER STATEMENT

This document is disseminated in the interest of information exchange. The contents of this report reflect the views of the authors who are responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the State of California or the Federal Highway Administration. This publication does not constitute a standard, specification or regulation. This report does not constitute an endorsement by the Department of any product described herein.

For individuals with sensory disabilities, this document is available in Braille, large print, audiocassette, or compact disk. To obtain a copy of this document in one of these alternate formats, please contact: the Division of Research and Innovation, MS-83 California Department of Transportation, P.O. Box 942873, Sacramento, CA 94273-0001

FINAL REPORT

CONTRACT #65A0690

Prioritizing HCCLs Identified using Pedestrian Safety Monitoring Report Tool

Jiajian Lu, Dr. Aditya Medury, and Dr. Offer Grembek

University of California at Berkeley Safe Transportation Research and Education Center (SafeTREC)

August 24, 2019

Contents

Chapter 1

Introduction

This research aims to identify and compare methodologies for prioritizing pedestrian hotspots, or high collision concentration locations (HCCLs) in the absence of pedestrian volume information and accompanying safety performance functions. In particular, we seek to evaluate the following methods:

1. Methods evaluating presence of significant patterns among crash types:
   (a) Direct diagnostics approach [9],
   (b) Probability of specific crash types exceeding threshold proportion (HSM 4.4.2.9)[1],
2. Methods controlling for regression-to-the-mean (in the absence of exposure):
   (a) Excess predicted average crash frequency using method of moments (HSM 4.4.2.6)[1].

In order to evaluate the above-mentioned methods, chapter 2 first presents two alternative pedestrian crash typology alternatives for pattern recognition. Chapter 3 describes the two pattern recognition-based approaches. Chapter 4 describes the method of moments approach for mitigating regression-to-the-means phenomenon, wherein the observed crash frequencies may contain outliers. Chapter 5 compares the performance of the potential HCCL prioritization alternatives using the 2009-2013 pedestrian safety monitoring report (PSMR) investigation results. Finally, we provide the conclusions and recommendations from this research effort in Chapter 6.

Chapter 2

Crash Typology Development

A prerequisite for identifying pedestrian crash patterns is to develop a crash typology which can cover a substantial number of pedestrian crashes. In this chapter, we present two alternatives: (i) a fixed crash typology, and (ii) a dominant crash typology. A fixed crash typology is defined using a limited number of crash descriptors that are common to all crashes of a given type. In comparison, a dominant crash typology is defined by summarizing the dominant crash characteristics of all crashes assigned a given crash type.

2.1 Fixed crash typology

A fixed crash typology requires defining salient crash characteristics that can potentially span a large percentage of crashes are also meaningful across multiple types of facilities (i.e., intersections, ramps and segments). In the case of automobile collisions, the collision dynamics are succinctly summarized by categories such as broadside, sideswipe, head-on, etc. In order to create a similar pedestrian crash typology, we considered the following variables: movement preceding collisions of the pedestrian and the other party type involved in the collision. The motor vehicle movements were categorized as traveling straight, turning right or left, or making an alternate type of movement prior to the collision (e.g., backing up, lane changing, stopping). In the case of pedestrian movements, we considered differentiated them by whether the pedestrian was crossing at/away from the crosswalk, or not crossing at all. After evaluating different combinations of the two types of movements, we proposed a fixed crash typology comprising of 5 crash types as shown in table 2.1.

The proposed crash types are generic in the sense that crashes belonging to a given type can occur across intersections, ramps or segments, as well as in restricted access-control locations (like freeways or expressways) or along arterial streets. We sought to ensure that the sample size of each crash type is substantially large so as to maintain adequate representation across multiple HCCLs. As a result, under some crash types (e.g., 4 and 5), we grouped different types of crossing movements while keeping the turning movements distinct, as specific turning-related patterns may lead to different operational countermeasures. In comparison, instances of pedestrians getting hit by vehicles traveling straight when crossing away from the crosswalk can be numerous and occur more randomly. As a result, we chose to keep crash types 1 and 3 as distinct events.

A limitation of the fixed crash typology is that some crashes which do not exactly meet the desired crash characteristics do not get assigned to any crash type (e.g., crashes with alternate

Table 2.1: Fixed crash typology

Crash Type	Vehicle Movement	Pedestrian Movement	Percentage of Total Crashes (2013-2017)
1	Straight	Crossing at crosswalk	10%
2	Straight	Not crossing	30%
3	Straight	Crossing not at crosswalk	15%
4	Right turn	Crossing (at either location)	10%
5	Left turn	Crossing (at either location)	10%
-	Other		25%

motor vehicle movements). Moreover, while we considered additional variables such as lighting, time of day, etc. with the crash typology, the inclusion of more variables puts additional constraints on the crash typology resulting in smaller sample sizes for each crash type.

2.2 Dominant crash typology

Clustering methods such as K-Means and Latent Class Analysis (LCA) are commonly used for generating crash typologies in the traffic safety literature [2, 8, 3, 5, 4]. These methods either deterministically or probabilistically partition the crash data into a user-defined number of clusters. Cluster methods have the advantage of efficiently incorporating multiple variables while also ensuring that each collision can be assigned to one cluster or another.

In order to develop a dominant crash typology, we considered five variables: access control, facility type, lighting, vehicle movement and pedestrian movement. Since the pedestrian crash data contain variables with discrete values, we first did principal component analysis (PCA) on the data and took the first 14 principal components which account for 90% variance of the data for further clustering. PCA not only transforms the discrete values to continuous values which are suitable for K-mean clustering but also reduces the dimensionality and noise of the data.

Then we performed K-mean clustering on those 14 principal components. The idea of K-mean is to assign each data point to a closest center and then update all the centers until centers converge. We need to select the number of centers as the only parameter of this algorithm and we chose five clusters as the number of clusters according to elbow plot shown in Figure 2.1. Finally, we transformed the principal components back to the data and calculated the majority of values in each variable of each crash type.

Table 2.2 presents the 5 crash types identified using k-means along with the dominant crash characteristics for each variable considered. For example, 61% of the crashes in crash type 1 have freeways under access control, so freeway is identified as the dominant crash characteristic. However, because the dominant crash characteristics are not necessary for a crash to be included within a given type, we are not guaranteed that all crashes of a given type within a HCCL share the same

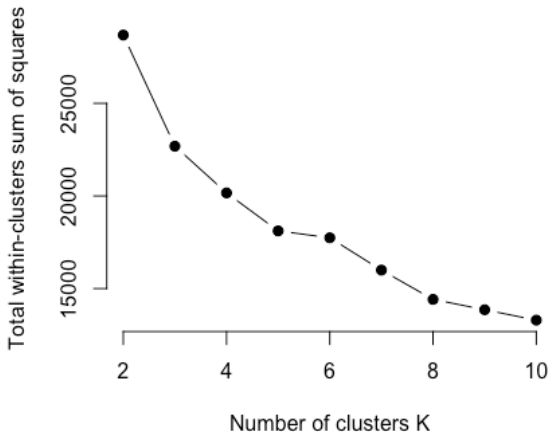


Figure 2.1: Identifying appropriate number of clusters using an elbow method

Table 2.2: Dominant crash typology

Crash Type	Access Control	Facility Type	Lighting	Vehicle Movement	Pedestrian Movement	Percentage of Total Crashes (2013-2017)
1	Freeway (61%)	Segment (92%)	Dark no street light (96%)	Straight (69%)	Roadway (70%)	11
2	Arterial (59%)	Segment (100%)	Daylight (95%)	Straight (58%)	Not crossing (64%)	32
3	Arterial (58%)	Segment (87%)	Dark with street light (98%)	Straight (71%)	Crossing not at crosswalk (71%)	17
4	Freeway (100%)	Ramp (100%)	Daylight (68%)	Turning (46%)	Crossing at crosswalk (60%)	12
5	Arterial (95%)	Intersection (77%)	Daylight (74%)	Turning (63%)	Crossing at crosswalk (80%)	28

attributes. Finally, as Table 2.2 indicates, the crash coverage of the dominant crash typology is 100%.

In summary, in this chapter, we proposed two crash typologies that can be used for identifying patterns within HCCLs. In the next chapter, we shall describe the methodologies of two pattern recognition approaches that have previously been proposed in the traffic safety literature.

Chapter 3

Pattern Recognition Methods

In this chapter, we examine two methods for pattern recognition in HCCL prioritization. The first method is direct diagnostic approach proposed by [9]. The second method is defined under the Highway Safety Manual [1] as the probability of specific crash types exceeding threshold proportion (HSM 4.4.2.9).

3.1 Direct Diagnostics

Direct diagnostics is a pattern recognition-based approach to assess whether a location is observing a high proportion of crashes relative to the mean proportion derived from the reference population. The methodology assumes that collision occurrence follows a binomial distribution, wherein each Bernoulli trial is determined by the mean probability of crash occurrence. The mean proportion of occurrence of a targeted crash type, c, $\bar{p}_c$, is calculated as:

$$\bar{p}_c = \frac{\sum_i x_{ic}}{\sum_i n_i} \tag{3.1}$$

Where x_{ic} is the number of type-c crashes at site i and n_i is the total number of crashes at site i.

Since the direct diagnostics approach assumes that all locations share the same underlying probability of occurrence of a given crash type, it follows that that a HCCL with n_i crashes is result of a series of Bernoulli trials which follows a binomial distribution. The probability of observing x_{ic} crashes of type c out of a total n_i crashes is given by, :

$$P(x_{ic}|n_i, \bar{p}_c) = \binom{n_i}{x_{ic}} \bar{p}_c^{x_{ic}} (1-\bar{p}_c)^{n_i - x_{ic}} \tag{3.2}$$

Finally, the probability of observing at least x_{ic} out of n_i crashes, p-val$_{ic}^b$, is given by:

$$\text{p-val}_{ic}^b = P\left(x \geq x_{ic}|n_i, \overline{p}\right) = 1 - P\left[x \leq (x_i - 1)\right] \tag{3.3}$$

$$= 1 - \sum_{x=0}^{x_{ic}-1} \frac{n_i}{x!\,(n_i - x)!} (\overline{p})^x (1-\overline{p})^{n_i - x} \tag{3.4}$$

A low value of p-val$_{ic}^{b}$ indicates an overrepresentation of collisions of a given crash type, relative what is observed in the reference population.

However, this method assumes that all facilities for a given reference population all have the same underlying propensity to observe a crash type which is restrictive. In order to relax this constraint, a beta-binomial distribution can be used, which utilizes a beta distribution to capture the uncertainty in the underlying probability of a observing a given crash type. In addition, the beta-binomial distribution can accommodate a greater amount of variation in the crash data, defined as overdispersion, than a binomial distribution.

3.2 Probability of specific crash types exceeding threshold proportion

This approach, as defined by the Highway Safety Manual [1] utilizes a beta-binomial (BB) distribution to model the crash proportions. Herein, the beta distribution,$Beta(p|\alpha,\beta)$, models the prior distribution of the true proportion of the target crash type, p, among all the facilities in the reference population:

$$Beta(p|\alpha,\beta) = \frac{p^{\alpha-1}(1-p)^{\alpha-1}}{B(\alpha,\beta)}, \quad \text{where } 0 \leq p \leq 1 \tag{3.5}$$

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{3.6}$$

Herein, Γ represents the gamma function, and $\alpha(>0)$ and $\beta(>0)$ are two parameters that determine the shape of the Beta distribution. More specifically, the mean and variance of the underlying true proportion for a given crash type can be described as follows:

$$E(p) = \frac{\alpha}{\alpha+\beta} \tag{3.7}$$

$$Var(p) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{3.8}$$

Equations 3.7 and 3.8 explain how α and β quantify the uncertainty in the prior distribution. While the relative magnitudes of alpha and beta inform the skewness of the distribution, high values of α and β lead to a small variance in the true proportion within the reference population.

The BB distribution combines the $Beta$ prior with the empirical evidence of the total number of crashes as well as the crashes observed for the target crash type.The probability of observing x_{ic} crashes of the target crash type, given n_i total crashes, is defined by a binomial distribution:

$$P(x_{ic}|p,n_i) = \begin{pmatrix} n_i \\ x_{ic} \end{pmatrix} p^{x_{ic}}(1-p)^{n_i-x_{ic}}, \tag{3.9}$$

$$p \sim Beta(\alpha,\beta), \tag{3.10}$$

the resulting compound distribution can be defined as follows:

$$P\left(x_{ic}|n_i,\alpha,\beta\right) = \int_0^1 P(x_{ic}|p,n_i)Beta(p|\alpha,\beta)dp \tag{3.11}$$

$$= \frac{n_i!}{x_{ic}!\left(n_i - x_{ic}\right)!}\frac{B\left(\alpha + x_{ic}, \beta + n_i - x_{ic}\right)}{B(\alpha,\beta)}. \tag{3.12}$$

We used Maximum Likelihood to solve for $\hat{\alpha}, \hat{\beta}$. Thereafter, the posterior distribution of $\tilde{\alpha}, \tilde{\beta}$ are given by the parameters:

$$\begin{aligned}\tilde{\alpha} &= \hat{\alpha} + x_{ic} \\ \tilde{\beta} &= \hat{\beta} + n_i - x_{ic}\end{aligned}$$

Finally, the BB test evaluates the proposition that the estimated proportion of the target crash type at a given location (given the posterior distribution) is greater than the median proportion of the reference population (as defined by the prior distribution) [7, 10]. The median proportion is derived from the prior distribution as follows:

$$\int_{p_m}^1 Beta(p|\hat{\alpha},\hat{\beta})dp = 0.5 \tag{3.13}$$

Heydecker and Wu (1991) propose an alternate hypothesis to evaluate the posterior distribution while Lyon et al (2007) test the implication of choosing threshold proportions that are different from median. However, since the median proportion is popularly chosen threshold for the BB test in the literature (ref needed), our assessment of this methods shall be restricted to it.

Given the chosen threshold proportion, p_m, we test the proposition that the location's proportion, p_i, exceeds the threshold proportion as follows:

$$\begin{aligned}P\left(p_i > p_m|x_{ic},n_i\right) &= \int_{p_m}^1 \text{Beta}(p|\tilde{\alpha},\tilde{\beta})dp \\ &= 1 - Beta\left(p_m,\tilde{\alpha},\tilde{\beta}\right)\end{aligned} \tag{3.14}$$

Equation 3.14 returns a value close to 1 if the posterior probability of observing a crash of a given type is much greater than median proportion of the reference population. However, to make this performance measure behave similarly to p-val$_{ic}^{b}$ for the purposes of comparison, we utilize the following metric, p-val$_{ic}^{bb}$:

$$\text{p-val}_{ic}^{bb} = 1 - P\left(p_i > p_m|x_{ic},n_i\right) = Beta\left(p_m,\tilde{\alpha},\tilde{\beta}\right) \tag{3.15}$$

Chapter 4

Methods controlling for Regression-to-the-Mean

4.1 Excess predicted average crash frequency using method of moments

This approach is to adjust a site's observed crash frequency to partially account for regression to the mean and calculate the "potential improvement". We then rank sites based on the potential improvement.

According to HSM, we have five steps for this approach. First, we need to find reference sites for each HCCL. Reference sites should have similar attributes compared with a specific HCCL. For intersection or ramp HCCL, we find a location as reference site when it has the same intersection type, intersection control type, similar AADT (± 2500) and median width (± 2). For segment HCCL, we find a location as reference site when it has the same number of lanes, similar AADT (± 2500) and median width (± 2).

All the HCCLs are less than 0.1 mile so the reference site should be in comparable length. For intersection and ramp references, they are around 0.1 mile. But for segment references, some can be very long since the entire segment has homogeneous attributes which may contain lots of crashes. Therefore, we need to find a representative for the long segment. We use sliding windows to find the 0.1-mile window that has the highest number of crashes over other windows within the segment as the representative.

For a given reference population of size R and T years of crash data, we calculate the average crash frequency ($\bar{N}$) and the crash frequency variance (var) [6]:

$$\bar{N} = T\bar{y}, \tag{4.1}$$

$$var = T^2\left(s_y^2 - \bar{y} + \bar{y}/T\right), \tag{4.2}$$

$$\bar{y} = \sum_{i=1}^{R} \sum_{t=1}^{T} y_{it}/RT, \tag{4.3}$$

$$s_y^2 = \sum_{i=1}^{R} \sum_{t=1}^{T} \left(x_{it} - \bar{y}\right)^2 /RT, \tag{4.4}$$

where,

- $\bar{y}$: represents the annual estimate of number of collisions
- s_y^2: variance of the annual crash estimate

Thereafer, we calculate the adjusted crash frequency using empirical Bayes (EB) for site i:

$$N_{i,ad} = N_i + \frac{\bar{N}}{var}(\bar{N} - N_i) \tag{4.5}$$

Finally, we calculate the potential improvement for site i and rank sites according to PI:

$$\text{PI}_i = N_{i,ad} - \overline{N}$$

In the absence of safety performance functions, method of moments-based estimators such as $N_{i,ad}$ and PI_i, provide the capability to the mitigate concerns of regression-to-the-mean by weighting the observed number of crashes with the average number of crashes observed in the reference population. Since these metrics are derived for the total number of crashes, these estimates do not reveal if specific crash types are recurring these locations. However, EB-adjusted estimates can also be derived for specific crash types.

In comparison, the pattern recognition methods may be subject to the regression-to-means phenomenon if an observed pattern is driven by outliers as opposed to a true underlying crash pattern. However, the methods described in Chapter 3 are targeted towards querying whether specific crash types are over-represented at a given location or not.

Chapter 5

Empirical Comparison of Methods

In this chapter, we present the results of conducting an empirical assessment of the the methods described in chapters 3 and 4, using pedestrian crash data from 2009-2013, which was also analyzed and investigated as part of the first round of the pedestrian safety monitoring program pilot. In the subsequent sections, we first compare the performance of using the the direct diagnostic approach vs the beta-binomial test, followed by comparing pattern recognition-based detection with method of moments. In order to assess the quality of the prioritization, we analyzed the recommendations emanating from the investigations conducted at specific high collision concentration locations. While the comparison does not reveal if a given prioritization method is indicative of a true hotspot, it provides insight into which methods may align well with the investigation process.

5.1 Comparison between the pattern recognition-based approaches

The outputs for pattern recognition-based approach and probability of specific crash types exceeding threshold proportion approach are both probability matrix where each row represent one site and each column represent one crash type. For any probability p_{ic} in the matrix, it represents the probability of N_{ic} crashes happen in site i where N_{ic} is the number of type-c crashes in site i. If this probability is small but it did happen, we should pay more attention to this type of crashes in site i so we can prioritize sites based on different crash type probabilities.

Figure 5.1 shows parts of probability matrices from direct diagnostics and probability of specific crash types exceeding threshold proportion approach. The probabilities less than 0.1 are highlighted in red to indicate that they likely contain a significant pattern.

Moreover, we can see that the results from two different methods are similar and highlight almost the same locations. Then we calculated the similarity between two results, and they are 99.6% similar. Since the computation of the probability of specific crash types exceeding threshold proportion approach is complex, we utilized the direct diagnostic method as the desirable pattern recognition-based approach for comparison with method of moments.

District	Route	County	Route Name	PM START	PM END	Number of F+I	Number of crashes in					Pattern Recognition					Exceed Threshold Proportion				
							C1	C2	C3	C4	C5	C1 (%)	C2 (%)	C3 (%)	C4 (%)	C5 (%)	C1 (%)	C2 (%)	C3 (%)	C4 (%)	C5 (%)
10	120	STA		5.28	5.35	11	0	2	2	0	7	100	75	57	100	13	100	78	60	100	11
4	101	SF	T	4.86	4.948	10	0	2	1	0	7	100	70	84	100	7	100	74	85	100	6
4	185	ALA		2.55	2.64	10	0	3	1	0	6	100	41	84	100	20	100	48	85	100	18
1	101	HUM		77.23	77.31	9	0	3	1	0	5	100	34	80	100	31	100	41	82	100	27
1	101	HUM		77.48	77.55	9	0	1	1	0	7	100	90	80	100	3	100	91	82	100	3
12	1	ORA		8.49	8.59	9	0	2	1	0	6	100	64	80	100	13	100	69	82	100	11
12	39	ORA		3.11	3.12	9	0	1	2	0	6	100	90	45	100	13	100	91	50	100	11
12	39	ORA		6.58	6.65	9	0	2	0	0	7	100	64	100	100	3	100	69	100	100	3
1	101	HUM		78.78	78.87	8	0	3	1	0	4	100	27	77	100	46	100	35	78	100	41
4	29	SOL		2.26	2.34	8	0	4	0	0	4	100	9	100	100	46	100	14	100	100	41
4	101	SF		5.54	5.62	8	0	1	5	0	2	100	87	0	100	91	100	88	1	100	88
4	101	SF		5.95	6.05	8	0	2	1	0	5	100	58	77	100	21	100	63	78	100	18
4	123	ALA		0.37	0.47	8	0	0	2	0	6	100	100	39	100	6	100	100	44	100	5
7	2	LA		10.896	10.99	8	0	2	4	0	2	100	58	3	100	91	100	63	5	100	88
12	39	ORA		11.64	11.681	8	0	2	4	0	2	100	58	3	100	91	100	63	5	100	88
12	39	ORA		12.63	12.72	8	0	1	1	0	6	100	87	77	100	6	100	88	78	100	5
4	29	SOL		2.12	2.21	7	0	1	0	0	6	100	84	100	100	3	100	85	100	100	2
4	101	SF	T	5	5.076	7	0	2	1	0	4	100	50	72	100	33	100	57	74	100	29
4	101	SF		5.74	5.81	7	0	0	3	0	4	100	100	9	100	33	100	100	13	100	29
4	101	SF		5.86	5.94	7	0	1	0	0	6	100	84	100	100	3	100	85	100	100	2
7	1	LA		6.511	6.591	7	0	3	0	0	4	100	20	100	100	33	100	28	100	100	29
7	1	LA		25.43	25.46	7	0	0	1	0	6	100	100	72	100	3	100	100	74	100	2
12	1	ORA		24.551	24.61	7	0	1	3	0	3	100	84	9	100	63	100	85	13	100	57

Figure 5.1: Results from Direct Diagnostics and Exceeding Threshold Proportion

5.2 Results from Method of Moments

For excess predicted average crash frequency using method of moments approach, we calculated a list of potential improvement (PI) shown as Figure 5.2.

District	Route	County	PM Begin	PM End	Length	Number of F+I	Method of Moment				
							# of sites	AVES	VARS	ADJ	PI
10	120	STA	5.28	5.35	0.07	11	356	0.83	2.65	7.82	7.00
4	101	SF	4.86	4.95	0.088	10	307	2.19	7.84	7.82	5.63
4	185	ALA	2.55	2.64	0.09	10	30	2.27	9.03	8.06	5.79
1	101	HUM	77.23	77.31	0.08	9	71	0.85	3.53	7.05	6.20
1	101	HUM	77.48	77.55	0.07	9	131	2.49	8.14	7.01	4.52
12	1	ORA	8.49	8.59	0.1	9	217	1.22	4.08	6.68	5.46
12	39	ORA	3.11	3.12	0.01	9	75	2.93	10.77	7.35	4.41
12	39	ORA	6.58	6.65	0.07	9	37	0.51	2.37	7.16	6.65
1	101	HUM	78.78	78.87	0.09	8	42	2.10	5.41	5.71	3.62
4	29	SOL	2.26	2.34	0.08	8	164	1.20	3.67	5.78	4.57
4	101	SF	5.54	5.62	0.08	8	307	2.19	7.75	6.36	4.18
4	101	SF	5.95	6.05	0.1	8	307	2.19	7.75	6.36	4.18
4	123	ALA	0.37	0.47	0.1	8	219	1.95	4.73	5.51	3.57
7	2	LA	10.9	10.99	0.094	8	131	2.48	8.05	6.30	3.82
12	39	ORA	11.64	11.68	0.041	8	368	2.17	7.18	6.24	4.07
12	39	ORA	12.63	12.72	0.09	8	368	2.17	7.18	6.24	4.07
4	29	SOL	2.12	2.21	0.09	7	164	1.20	3.59	5.07	3.87
4	101	SF	5	5.08	0.076	7	307	2.18	7.72	5.64	3.46
4	101	SF	5.74	5.81	0.07	7	307	2.18	7.72	5.64	3.46
4	101	SF	5.86	5.94	0.08	7	307	2.18	7.72	5.64	3.46
7	1	LA	6.51	6.59	0.08	7	307	2.18	7.72	5.64	3.46
7	1	LA	25.43	25.46	0.03	7	219	1.94	4.68	4.90	2.96
12	1	ORA	24.55	24.61	0.059	7	307	2.18	7.72	5.64	3.46

Figure 5.2: Results from Method of Moments

The potential improvements greater than 4 are highlighted in red.

District	Route	County	Number of F+I	Number of crashes in					Pattern Recognition					Potential Improvement				
				C1	C2	C3	C4	C5	C1 (%)	C2 (%)	C3 (%)	C4 (%)	C5 (%)	C1	C2	C3	C4	C5
10	120	STA	11	0	2	2	0	7	100	75	57	100	13	0.00	0.69	0.61	0.00	3.94
4	101	SF	10	0	2	1	0	7	100	70	84	100	7	0.00	0.42	0.25	0.00	3.45
4	185	ALA	10	0	3	1	0	6	100	41	84	100	20	0.00	0.94	0.34	0.00	2.65
1	101	HUM	9	0	3	1	0	5	100	34	80	100	31	0.00	1.26	0.31	0.00	3.13
1	101	HUM	9	0	1	1	0	7	100	90	80	100	3	0.00	0.18	0.28	0.00	3.34
12	1	ORA	9	0	2	1	0	6	100	64	80	100	13	0.00	0.32	0.10	0.00	3.53
12	39	ORA	9	0	1	2	0	6	100	90	45	100	13	0.00	-0.03	0.53	0.00	2.59
12	39	ORA	9	0	2	0	0	7	100	64	100	100	3	0.00	0.43	0.53	0.00	5.26
1	101	HUM	8	0	3	1	0	4	100	27	77	100	46	0.00	0.66	0.05	0.00	1.23
4	29	SOL	8	0	4	0	0	4	100	9	100	100	46	0.00	1.63	0.01	0.00	2.06
4	101	SF	8	0	1	5	0	2	100	87	0	100	91	0.00	0.16	1.94	0.00	0.31
4	101	SF	8	0	2	1	0	5	100	58	77	100	21	0.00	0.42	0.25	0.00	2.18
4	123	ALA	8	0	0	2	0	6	100	100	39	100	6	0.00	-0.10	0.34	0.00	2.21
7	2	LA	8	0	2	4	0	2	100	58	3	100	91	0.00	0.48	1.71	0.00	0.18
12	39	ORA	8	0	2	4	0	2	100	58	3	100	91	0.00	0.46	1.30	0.00	0.31
12	39	ORA	8	0	1	1	0	6	100	87	77	100	6	0.00	0.18	0.22	0.00	2.76
4	29	SOL	7	0	1	0	0	6	100	84	100	100	3	0.00	0.27	0.01	0.00	3.42
4	101	SF	7	0	2	1	0	4	100	50	72	100	33	0.00	0.42	0.25	0.00	1.56
4	101	SF	7	0	0	3	0	4	100	100	9	100	33	0.00	-0.10	1.00	0.00	1.56
4	101	SF	7	0	1	0	0	6	100	84	100	100	3	0.00	0.16	-0.11	0.00	2.81
7	1	LA	7	0	3	0	0	4	100	20	100	100	33	0.00	0.72	-0.11	0.00	1.56
7	1	LA	7	0	0	1	0	6	100	100	72	100	3	0.00	-0.10	0.13	0.00	2.21
12	1	ORA	7	0	1	3	0	3	100	84	9	100	63	0.00	0.16	1.00	0.00	0.93

Figure 5.3: Results from Direct Diagnostics and Method of Moments for each crash type

We also applied method of moments for each crash in the crash typology. The idea is that

instead of calculating the PI with total F+I, we use the number of crashes in each crash type and calculate PI for each crash type as in Figure 5.3.

5.3 Combining results from method of moments and direct diagnostics

The correlation coefficient between Potential improvement (PI) and fatal & Injury crash frequency is 0.83 which means PI is highly correlated with F+I crash frequency. Therefore, PI contains the information about number of crashes in an HCCL.

Information from pattern recognition can help the investigators find out the recurring trend and the corresponding countermeasure and thus reduce the false positive rate.

Thus, we considered combining PI and the p-value from the pattern recognition to create a weighted PI for crash type, c:

$$\text{PI}_{ic}^{w} = \text{PI}_{ic} * \left(1\text{–p-val}_{ic}^{b}\right) \tag{5.1}$$

Herein, $(1\text{–p-val}_{ic}^{b})$ can be seen as a discounting factor reflecting how confident we are in observing a recurring crash type c. So the higher PI_{ic}^{w}, the more potential improvement and confidence for a HCCL with a recurring crash type c relative to the reference population.

5.4 Potential prioritization metrics

Based on pattern recognition and method of moment, we have five available prioritization criteria:

1. Observed crash frequency (fatal(F) + injury(I)), N_i
2. EB-adjusted crash frequency, $N_{i,eb}$
3. Potential for safety improvement, PI_i
4. Weighted sum of potential improvement by crash types:

$$\sum_{c} \left(1\text{–p-val}_{ic}^{b}\right) * PI_{i,c}$$

5. Weighted max of potential improvement by crash types:

$$\max_{c} \left(1\text{–p-val}_{ic}^{b}\right) * PI_{i,c}$$

In order to compare these metrics, we evaluate the the HCCL list from PSMR Round 1 where we compare the consistency between the recommendation column and the ranking from each criterion.

5.5 Empirical data based on PSMR Round 1

There are 129 HCCLs investigated from PSMR Round 1 and all HCCLs are provided with recommendation from site investigators. Since the recommendations are made by the site investigators who are experienced and have been to the sites, we can think of them as "ground truth" ranking for the HCCLs. In order to compare the recommendation with the five criteria, we will need to convert them from the text format into some quantitative format. Therefore, we read all the recommendations and label them in the following way:

Table 5.1: Classification of recommendations

Label	**Recommendation**
0	No action
1.2	Planned/implemented projects do not align with crash type patterns
1.3	Planned/implemented projects align with crash type patterns
2	Improvements made which did not align with crash type patterns
3	Improvements made which aligned with crash type trends

Some examples pertaining to each label, recommendation and the associated direct diagnostic outputs are shown in Table 5.2. The direct diagnostic p-values for the fixed and dominant crash typology shown in the examples indicate that the fixed typology might reveal more consistent patterns.

After labelling the recommendation, we sorted the HCCLs by a given performance metric (e.g., F+I). We divided the list of 129 sites to top 43, middle 43, and last 43 sub-groups and aggregate the number of recommendations within each sub-group. Performance metrics which mimic the investigator's process best should yield higher 3/1.3 (or fewer 0's) in the top and middle sub-group relative to the bottom 43 HCCLs. We were specifically interested in the following two questions:

- How is the distribution of patterns identified upstream aligned with a given performance metric's prioritization?
- Is their correlation between "no action taken" and a prioritization?

Table 5.3 provides the mean estimates of the various prioritization metrics for all the 129 HCCLs distributed across the 5 recommendation categories. We observe the following trends:

1. The average number of crashes are greatest among the type 3 recommendations, followed by type 2 recommendations and finally no action taken.
2. The weighted PIs are on average greater in the type 3 recommendations than in the other cases. This implies that recommendations with patterns also yield high scores in the pattern recognition-based metrics. In addition, it is also likely that the investigators require a substantial number of crashes to deem a pattern to be present during their investigations which would explain the presence of greater number of crashes within type 3 recommendations.
3. Type 3 recommendations also have fewer fatalities in comparison to Type 2 recommendations and recommendations with no actions.

Table 5.2: Examples of recommendations

Recommendation	**Label**	**Pattern Recognition**
Project EA 10-0K150 just ended July 7, 2015. This project is a cold plane and resurface roadway which includes to place high visibility cross walks with PED XING marking in key locations.	1.2	No significant crash type
Pedestrian enhancements, including corner bulb-outs, at most intersections, ADA compliant curb ramps, and **pedestrian countdown timers (PCT)** and audible pedestrian signals (APS) at all intersections Traffic signal infrastructure for real-time traffic management, including traffic signal replacement, fiber interconnect, transit signal priority, **protected left-turn phases at intersections**, variable real-time message signs and real-time bus arrival information displays (NextMuni)	1.3	• Fixed typology: **left-turning crashes** at intersection (type 5)(pval: 0.05) • Dominant typology: turning-movement crashes at intersection (type 5) (pval: 0.12)
"Install rectangular rapid flashing beacon system and continental style crosswalk markings for SB Sunrise to WB US-50 on ramp. In addition, install SW24-2 signs and continental style crosswalk at the SB Sunrise to EB US-50 loop onramp."	2	• Fixed typology: vehicle going straight, pedestrian crossing not at crosswalk (type 3) (pval: 0.008) • Dominant typology: none of them is significant
"Since the collisions are due to **illegal crossing**, the Vallejo Police Department has been contacted by our field investigator and and asked to aid in enforcement. No additional improvements are proposed."	3	• Fixed: vehicle going straight, **pedestrian crossing not at crosswalk** (type 3) (pval: 0.003) • Dominant: conventional segment, dark with street light, straight, **pedestrian crossing not at crosswalk** (type 3) (pval: 0.03)

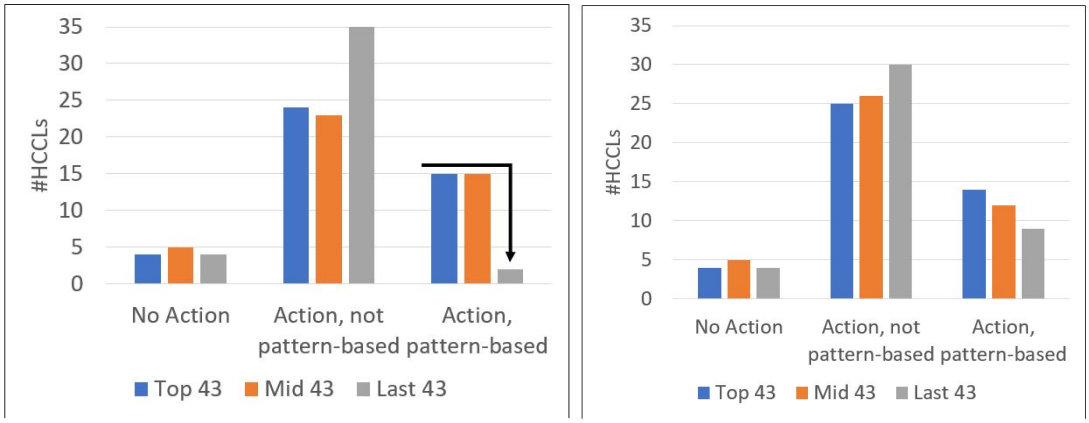
Table 5.3: Mean estimates of different prioritization metrics across recommendations

Recommendation Type	**HCCLs**	**F+I**	**F**	**ADJ**	**PI**	**Weighted max PI (Fixed)**	**Weighted max PI (Dominant)**	**Weighted sum PI (Fixed)**	**Weighted sum PI (Dominant)**
0 (No action)	13	4.2	0.7	3.8	2.7	1.1	1.3	1.6	1.8
1.2 (Projects planned, no pattern)	24	4.7	0.8	4.3	2.8	1.0	1.3	1.7	1.9
1.3 (Projects planned, pattern)	12	6.7	0.1	6.3	4.5	2.7	3.1	2.8	3.2
2 (Recommendations, no pattern)	60	4.0	0.8	3.6	2.5	0.9	1.0	1.4	1.6
3 (Recommendations, based on pattern)	20	5.2	0.5	4.8	3.3	2.7	2.8	2.4	2.5
Total	129	4.6	0.7	4.2	2.9	1.4	1.6	1.8	2.0

To further breakdown the performance of the different metrics, we analyze the trends in the types of recommendations when sorting the list of HCCLs by each metric. We group the recommendations into (i) no action, (ii) action that are not pattern-based and finally (iii) actions that are pattern-based.

5.5.1 Observed crash frequency(F+I)

The results of the F+I metric are shown in Figure 5.4. We observe no major trends in pattern-based HCCLs observed in dominant crash types (figure 5.5a). In contrast, there is sharp drop-off in pattern-based recommendations in the last 43 HCCLs for fixed crash types (figure **??**). Finally, we do not observe any significant patterns for the case of no action taken when we progress from the top 43 to the last 43 sub-groups.



(a) Patterns identified using fixed typology

(b) Patterns identified using dominant typology

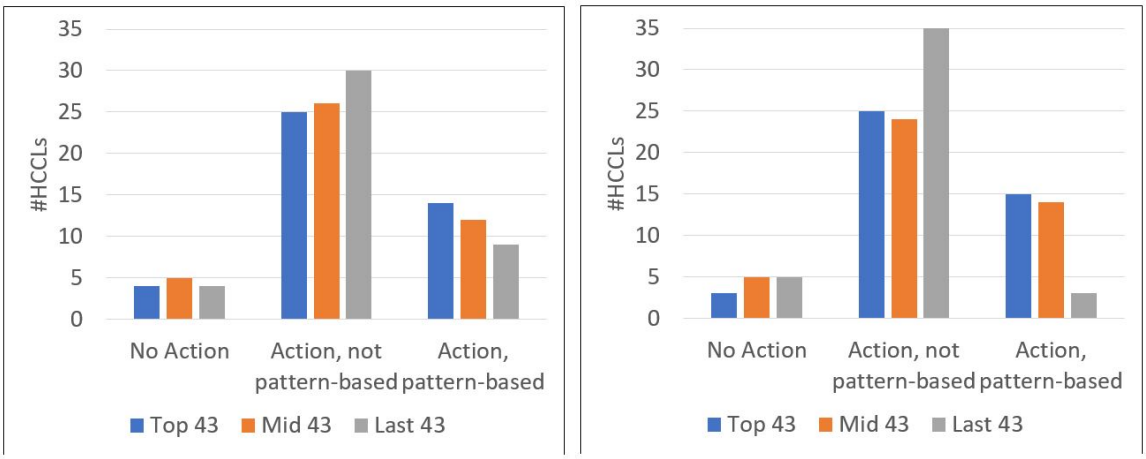
Figure 5.4: Empirical assessment of F+I metric

5.5.2 EB-adjusted crash frequency

The results of the EB-adjusted crash frequency metric are shown in Figure 5.5. The performance is similar to F+I, which is not unexpected given that we are adjusting the observed crash frequencies within EB using a 5-year evaluation period. In theory, we expect EB to converge to F+I as the number of years being considered are increased. However, as the number of years increase, we also expect the traffic and built environment conditions to change for the location which may make the historical crash data less meaningful.

5.5.3 Potential for Improvement (PI)

The results of PI-based prioritization are shown in Figure 5.6. The performance is similar to F+I and EB. However, we observe a greater drop-off in the pattern-based recommendations after the top 43 HCCLs.



(a) Patterns identified using dominant typology

(b) Patterns identified using fixed typology

Figure 5.5: Empirical assessment of EB-adjusted crash frequency metric

5.5.4 Weighted Sum PI

Figure 5.7 shows the trends for weighted sum PI for both dominant and fixed crash types. As observed earlier in the summary statistics, the weighted sum PI shows a consistent drop in pattern-based recommendations across the different sub-groups for both crash types.

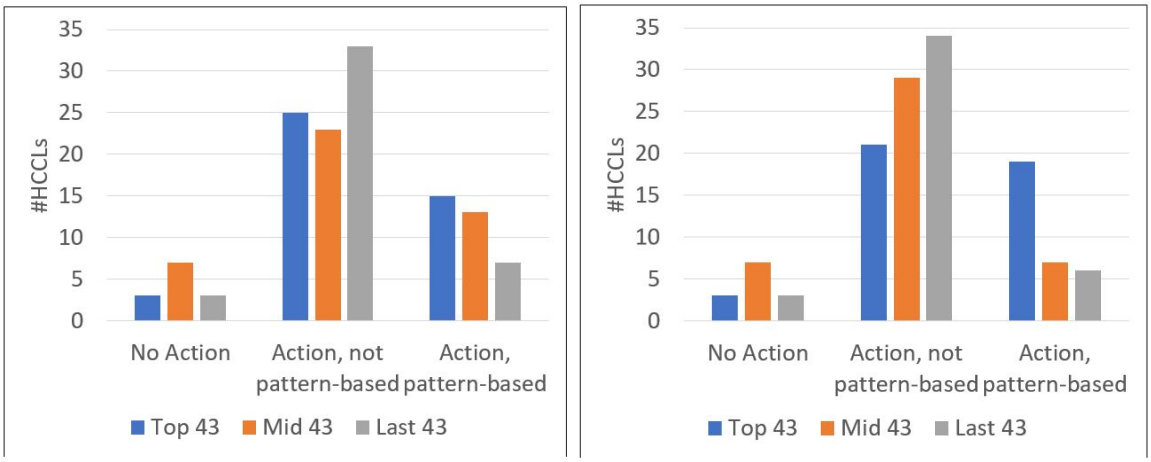
5.5.5 Weighted Max PI

Figure 5.8 shows the trends for weighted max PI for dominant and fixed crash types. The results for the weighted max case are similar to the weighted sum PI. However, since we are only selecting the weighted PI estimate of the crash type which has the highest value, the use of max over sum amplifies the presence of dominant patterns. This phenomenon is visible from the relatively high number of pattern-based recommendations (∼30) in the top 43 HCCLs when prioritized using the weighted max metric.

In summary, based on the empirical comparison of the different prioritization metrics proposed in this research, we find the weighted sum PI and weighted max PI to be the most effective in prioritizing pattern-based recommendations. We also note that the fixed crash typology showed more consistent trends in pattern-based recommendations. However, none of the metrics show any strong correlations vis-a-vis predicting fewer no action scenarios.

It is possible that since these investigations were being conducted for the first time under the PSMR pilot, there was a greater push to implement some pedestrian countermeasures (such as upgrading signal controls or re-painting crosswalks) even when there wasn't a significant pattern within the crashes to drive the countermeasure identification. Moreover, the list that was identified for the PSMR pilot was not based on single crash prioritization metric (e.g., F+I) and included some considerations for pedestrian fatalities. As a result, the findings may differ when evaluating a larger list of potential HCCLs for the 2009-2013 crash data.

Finally, we note that the performance of the weighted sum/max PI metrics indicates that



(a) Patterns identified using dominant typology

(b) Patterns identified using fixed typology

Figure 5.6: Empirical assessment of potential for improvement metric

these metrics can prioritize HCCLs with patterns higher up on the list. However, since the direct diagnostic metric is dependent on the observed number of collisions, presence of outliers within the 5-year evaluation period can predict patterns even though the site may not have an underlying long-running pattern of producing crashes of the given crash type. Thus, more research needs to be done to investigate if regression-to-the-mean is present within the direct diagnostic method.

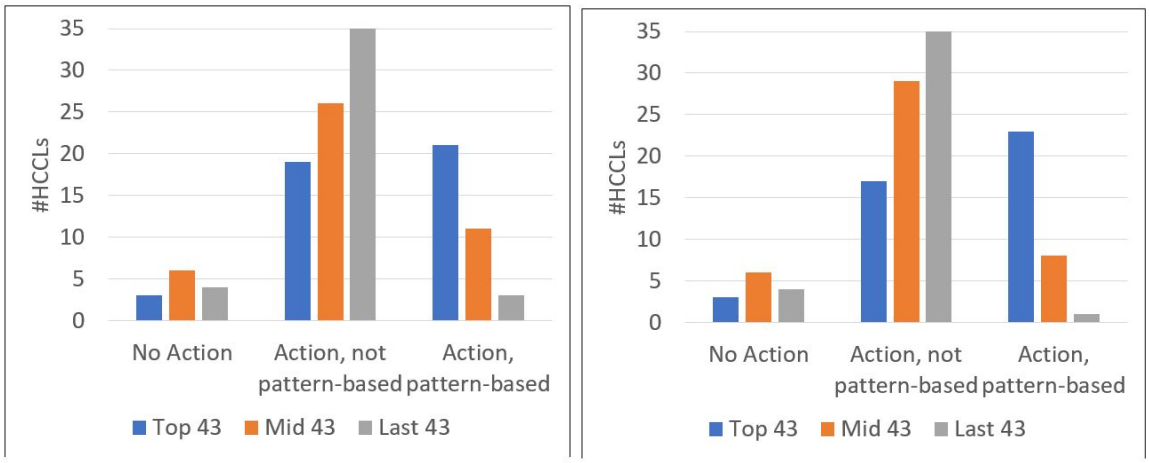
5.6 Practical Considerations

As part of the analysis, we also made some observations regarding the implementations of the proposed methods as well as the current network screening process:

1. Method of moments: Finding suitable reference sites for method of moments is challenging (some site characteristics may be more common than others). As a result, using a model to estimate the expected number of crashes for a given set of site characteristics (e.g., SPFs), even in the absence of pedestrian volume data, may be more desirable.

2. Planned projects: 2/3 of the HCCLs at the time of investigation indicated prior projects that were already implemented or planned. Such a high proportion may be likely due to the large time gap in the period of evaluation (2009-2013) and the period of site investigations (2016-2017).

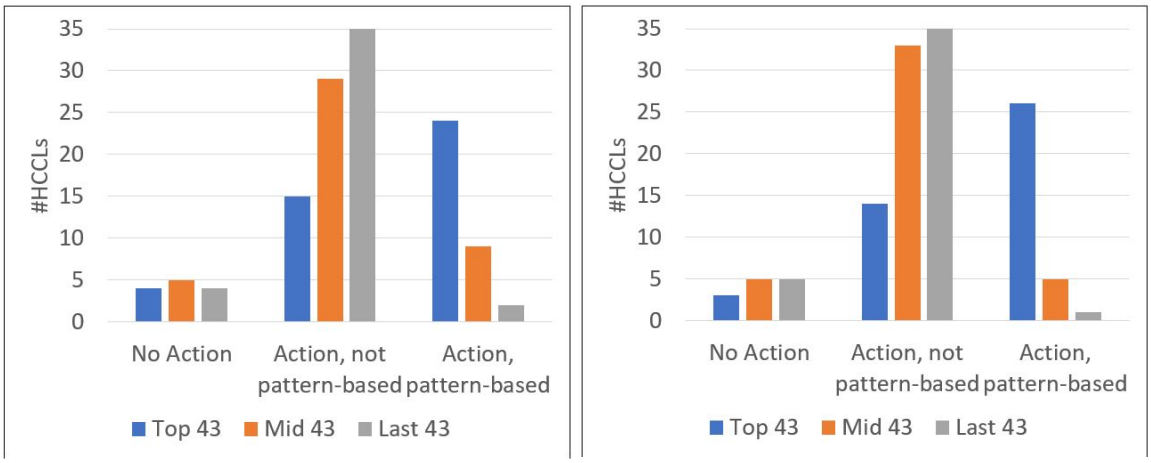
3. Need for segmentation: the use of a sliding window-based method for identifying pedestrian HCCLs provides the capability to maximize the total number of crashes. However, they also lead to HCCL definitions that are not ideal for providing well-defined site characteristics for implementing SPFs or method of moments. Thus, using a segmentation-based approach may be desirable for the network screening of pedestrian collisions.



(a) Patterns identified using dominant typology

(b) Patterns identified using fixed typology

Figure 5.7: Empirical assessment of weighted sum PI metric



(a) Patterns identified using dominant typology

(b) Patterns identified using fixed typology

Figure 5.8: Empirical assessment of weighted max PI metric

Chapter 6

Conclusions, Recommendations and Future Work

6.1 Summary

The objective of this research was to identify metrics to prioritize crash frequency-based pedestrian high collision concentration locations, that went beyond prioritizing sites based on total number of fatal and injury collisions. To this end, we analyzed methods that controlled for the regression-to-the-mean phenomenon in the total number of crashes in the absence of safety performance functions (Chapter 4). We also evaluated methods that used information present within the coded, party-level crash data to identify recurring crash patterns further upstream in the decision-making process (chapter 3). More specifically, we analyzed the following three methods:

- Direct diagnostic approach ([9]),
- Probability of specific crash types exceeding threshold proportion (HSM 4.4.2.9),
- Excess predicted average crash frequency using method of moments (HSM 4.4.2.6).

In order to implement these pattern recognition-based methodologies, we also proposed two alternative pedestrian crash typologies (chapter 2). Once the crash typologies were defined, based on the methods described above, we considered five potential prioritization metrics:

1. Observed crash frequency (F + I), N_i
2. EB-adjusted crash frequency, $N_{i,eb}$
3. Potential for safety improvement, PI_i
4. Weighted sum of potential improvement by crash types:
5. Weighted max of potential improvement by crash types:

Finally, we assessed the quality of the aforementioned metrics with regards to (i) mimicing the pattern identification process of traffic safety investigators and (ii) minimizing cases where an

HCCL investigation yields no action. To this end, we used the investigation results of round 1 of the pedestrian safety monitoring report (PSMR) which analyzed 129 pedestrian HCCLs. We analyzed these HCCLs to assess if the any recommendations were made, and if so whether the recommendations were in alignment of either the fixed of the dominant crash typology. Thereafter, we sorted 129 HCCLs using each metric and evaluated whether the investigations with no actions showed an monotonically increasing trend (fewer at the top, more at the bottom), and if the investigations with patterns yielded a monotonically decreasing trend (more at the top, fewer at the bottom). The results of the empirical analysis provided the following findings:

- None of the results showed any significant trends with regards to minimizing false positives, i.e. investigations with recommendations to take no action.
- Weighted max and weighted sum PIs showed a consistently decreasing trend in pattern-based HCCLs. These metrics combined information from both method-of-moments and the pattern recognition-based methods.
- The performances of EB-adjusted crash frequency, potential for improvement and unadjusted crash frequency (F+I) were similar.

6.2 Recommendations

Based on the findings from this research, we recommend the following:

1. Metrics derived from significance of pattern recognition may potentially mimic pattern identification process of investigators. Thus, incorporating pattern recognition-based metrics further upstream in the decision-making process can potentially help increase the likelihood ensuring that the investigators can find a recurring safety concern at a location.
2. Fixed crash typologies provide more consistent results along with the added the advantage of being easy to interpret.
3. Since the round 1 of PSMR did not contain many false positives, we encourage continuing further research into the methods and metrics analyzed in this project to assess if they are prone to predicting presence of recurring pattern in response to outliers.
4. The results of the investigations also contain multiple cases of sites where countermeasures were implemented prior to the investigations as part of other capital projects. Thus, we recommend identifying sites that have undergone relevant design/operational changes prior to the HCCL list generation, so as to consider excluding them from the final list of recommended sites for investigations.
5. We also recommend revising the existing network screening process to adopt a segmentation-based approach so as to make it easier to implement methodologies that require well-defined site characteristics (e.g., SPFs and method of moments). We also recommend considering model-based alternatives to method of moments as defining the reference populations for the latter approach is not straightforward.

6.3 Future Work

To further investigate the accuracy of the methods considered in this research, we propose conducting a simulation-based study, where we can generate synthetic HCCLs with a true crash mean and crash type distributions, and analyze the performance of the various metrics for varying numbers of reference sites, overdispersion, total number of crashes, etc.

For instance, The aforementioned parameters can interact with each other to create scenarios where direct diagnostics and the beta-binomial tests may produce different results. Consider a case where the mean proportion of observing a given crash type is $p = 0.49$. Subsequently, let's say we observe that a specific site produces 4 out 4 collisions of the corresponding crash type. Under such a scenario, the direct diagnostic/binomial test would indicate that a pattern exists ($\text{p-val}_{ic}^{b} = 0.058$), as illustrated in Figure 6.1.

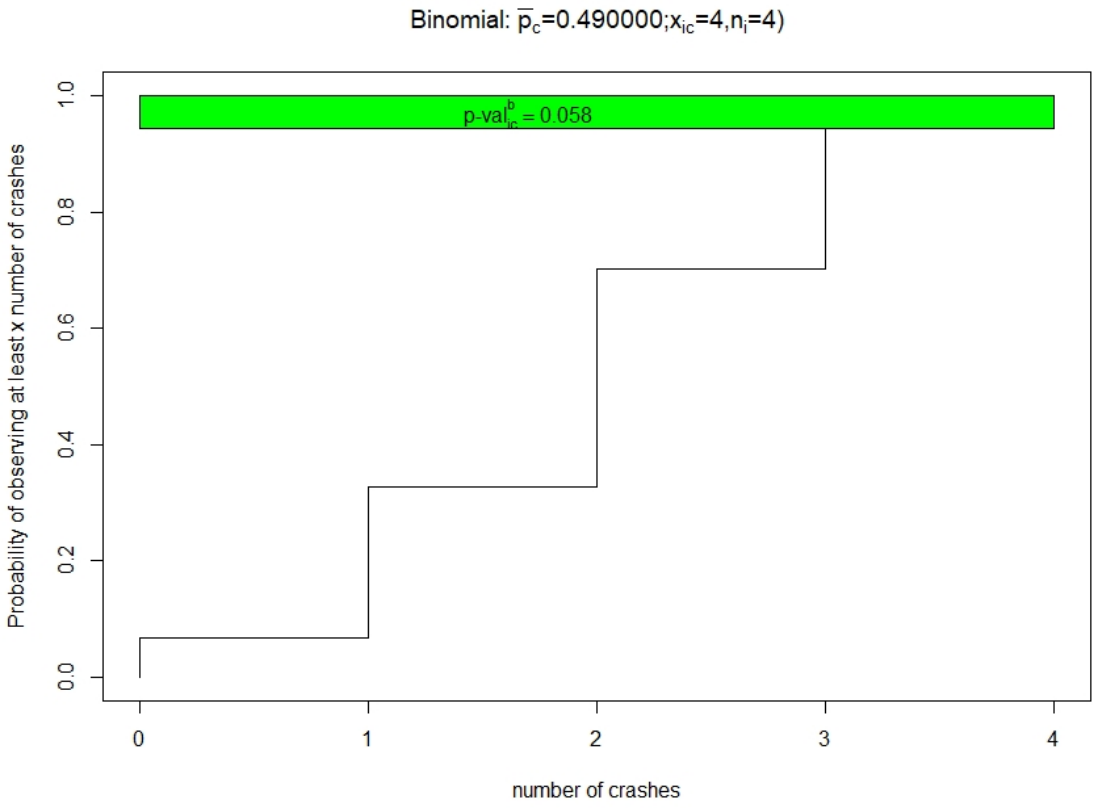


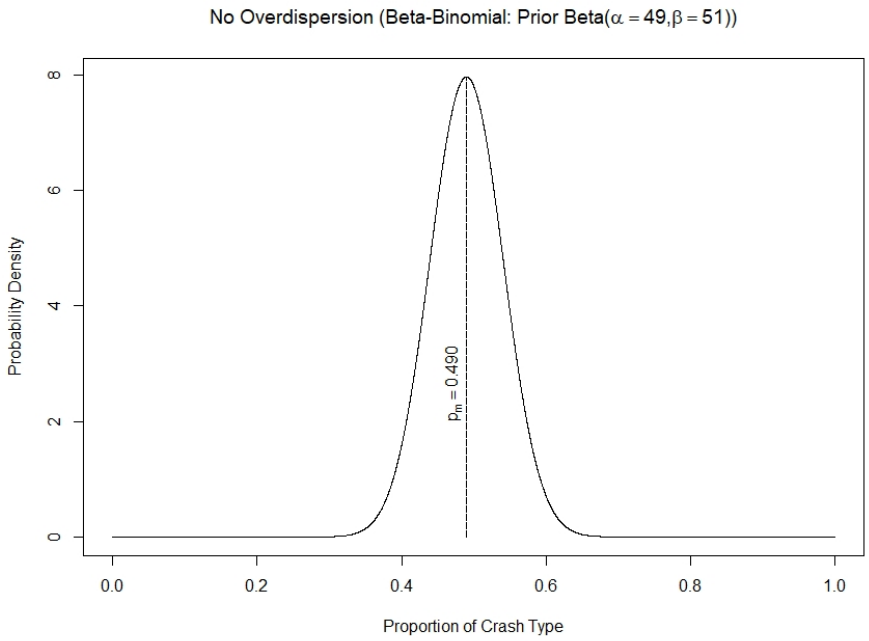
Figure 6.1: Direct diagnostic/binomial test visualization

For the beta-binomial test, the detection of a pattern is contingent on how much overdispersion there is in the reference population. Overdispersion is a statistical measure of the amount of variation there is the sample relative to the mean estimate. In the context of proportions, it can mean that the sites in the reference population may be homogeneous (leading to less overdisperison) or

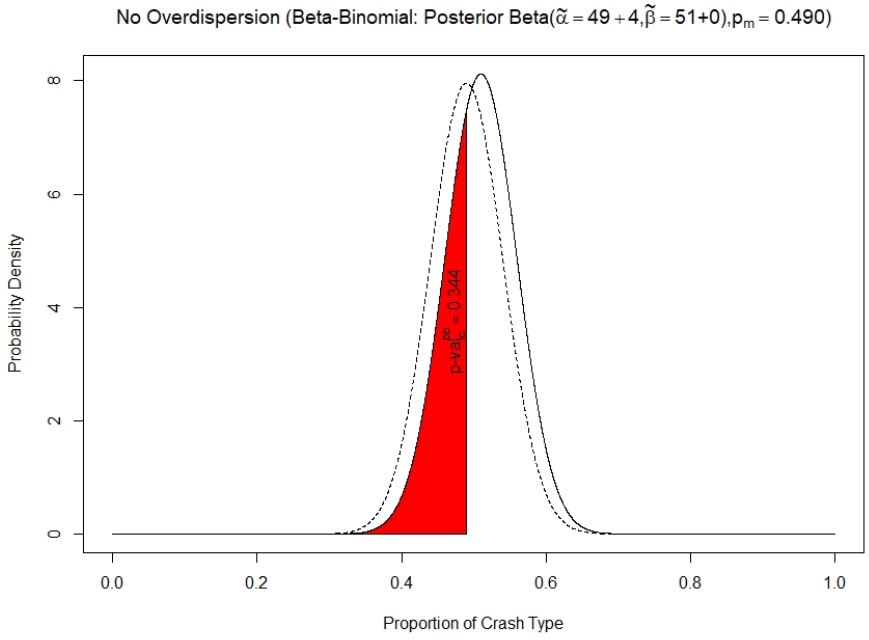
extremely varied (leading to high overdispersion). A low-to-no overdispersion case can be simulated by using high α and β values (see Figure 6.2a). Assuming $\alpha = 49$ and $\beta = 51$, we observe the median proportion to be $p_m = 0.49$. The lack of variation in the proportions is indicated by the high concentration of the distribution shown in Figure 6.2a around the mean and median estimate. Given the information about the number of crashes about the given site, the posterior beta distribution under the beta-binomial test does not change much (as shown in Figure 6.2b), which results in p-val$^b_{ic} = 0.344$. Thus, as per BB test, we cannot say that there is a significant pattern present.

In comparison, if there is high overdispersion (i.e., lots of variation in the crash proportions observed in the reference population), we can simulate this scenario using low parameter estimates for the prior Beta distribution such as $\alpha = 0.49$ and $\beta = 0.51$ (see Figure 6.3). The high overdispersion implies that even though the mean proportion of crashes in the population remains 0.49, a significant percentage of sites in the population actually have a true proportion of nearly 0 or 1 as reflected in the bathtub shape of the distribution shown in Figure 6.3a around the mean and median estimate. This also implies that we have low confidence in our prior information provided by the reference population. As a result when accounting for the fact that the specific site observes 4 out of 4 crashes for the given crash type, the posterior beta distribution under the beta-binomial test changes dramatically to skew towards the right (as shown in Figure 6.3b), which results in p-val$^b_{ic} = 0.013$. Thus, as per BB test, given high overdispersion, the BB test says that there may be a significant underlying crash pattern.

Thus, in the presence or absence of overdisperion, the performance of the direct diagnostic test may differ from the beta-binomial test. A simulated dataset can provide more insights into developing a more robust pattern recognition framework.

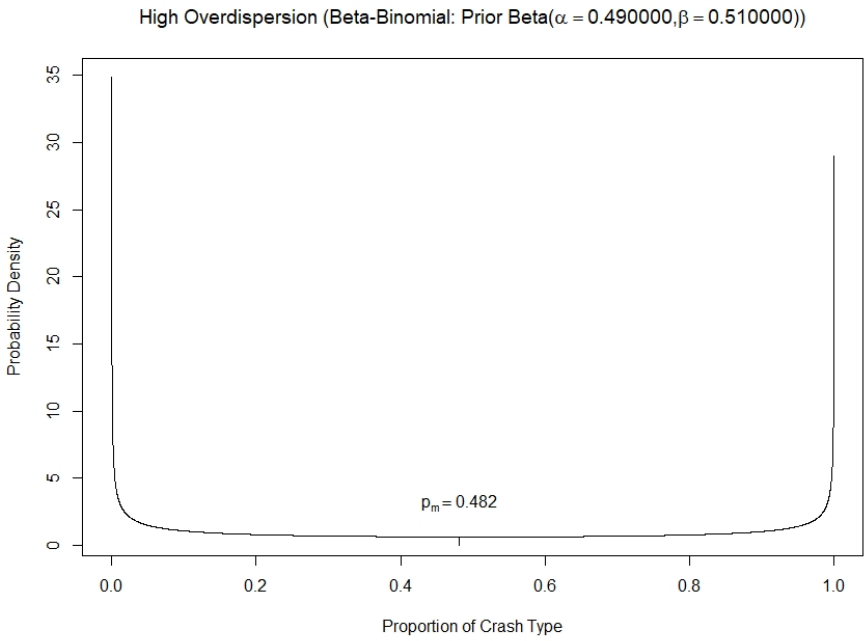


(a) Assumed beta distribution prior to represent low/no overdispersion

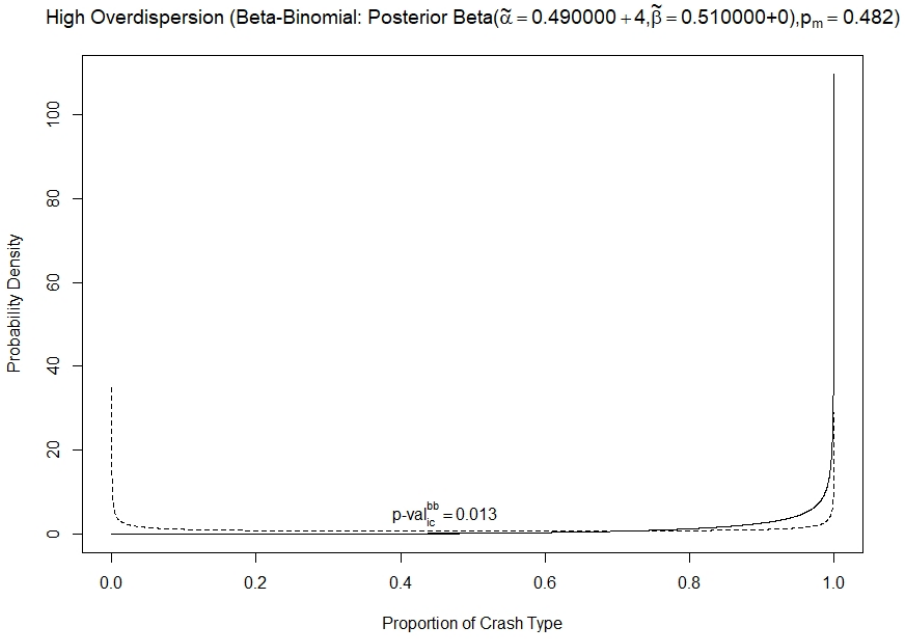


(b) Beta-Binomial test given the posterior distribution
(prior distribution in dotted lines)

Figure 6.2: Beta-binomial test given low/no dispersion



(a) Assumed beta distribution prior to represent low overdispersion



(b) Beta-Binomial test given the posterior distribution
(prior distribution in dotted lines)

Figure 6.3: Beta-binomial test given high overdispersion

Bibliography

[1] AASHTO. "Highway Safety Manual". In: *Washington, DC* 529 (2010).

[2] Tessa K Anderson. "Kernel density estimation and K-means clustering to profile road accident hotspots". In: *Accident Analysis & Prevention* 41.3 (2009), pp. 359–364.

[3] Carola A Blazquez and Marcela S Celis. "A spatial and temporal analysis of child pedestrian crashes in Santiago, Chile". In: *Accident Analysis & Prevention* 50 (2013), pp. 304–311.

[4] Juan De Oña et al. "Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks". In: *Accident Analysis & Prevention* 51 (2013), pp. 1–10.

[5] Benoit Depaire, Geert Wets, and Koen Vanhoof. "Traffic accident segmentation by means of latent class clustering". In: *Accident Analysis & Prevention* 40.4 (2008), pp. 1257–1266.

[6] Ezra Hauer. "On the estimation of the expected number of accidents". In: *Accident Analysis & Prevention* 18.1 (1986), pp. 1–12.

[7] BJ Heydecker and J Wu. "Using the information in road accident records". In: *Proc., 19th PTRC Summer Annual Meeting, London*. Vol. 166. 1991.

[8] Karl Kim and Eric Y Yamashita. "Using a k-means clustering algorithm to examine patterns of pedestrian involved crashes in Honolulu, Hawaii". In: *Journal of advanced transportation* 41.1 (2007), pp. 69–89.

[9] Jake Kononov. "Identifying locations with potential for accident reductions: Use of direct diagnostics and pattern recognition methodologies". In: *Transportation research record* 1784.1 (2002), pp. 153–158.

[10] Peter Y Park and Rajib Sahaji. "Safety network screening for municipalities with incomplete traffic volume data". In: *Accident Analysis & Prevention* 50 (2013), pp. 1062–1072.