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 CA21-2451	2. GOVERNMENT ASSOCIATION NUMBER	3. RECIPIENT'S CATALOG NUMBER

4. TITLE AND SUBTITLE Strategies for Reducing Pedestrian and Bicyclist Injury at the Corridor Level: Phase 3	5. REPORT DATE May 30, 2021
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR Julia Griswold, Joy Pasquet, Jiajian Lu, Aditya Medury, and Offer Grembek	8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Safe Transportation Research and Education Center University of California, Berkeley Berkeley, CA 94720	10. WORK UNIT NUMBER
	11. CONTRACT OR GRANT NUMBER 65A0678
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 Strategies for Reducing Pedestrian and Bicyclist Injury at the Corridor Level project is an effort of the California Department of Transportation (Caltrans) to identify and address systematic and specific problems with regard to pedestrian and bicycle safety in California, with the long-term goal of substantially reducing pedestrian and bicycle fatalities and injuries in the state. This project focused on working closely with Caltrans to translate the tools and models of previous pedestrian efforts into an implementable program that supports activities to reduce bicycle crashes, and also continuing to develop additional methods that support improvements in pedestrian safety to assist Caltrans in meeting its pedestrian and bicycle safety goals. Specifically, the core of this phase included four overarching objectives: develop a bicycle safety monitoring tool, support the pedestrian safety monitoring tool, develop a systemic approach and tool for bicycles, and develop an approach to model bicycle exposure for the state highway system. The project provides a foundation for spot, corridor, and systemic safety programs to identify and address safety problems with regard to bicyclist and pedestrian safety in California.

17. KEY WORDS bicycles, bicyclist safety, systemic safety, hotspot, bicycle exposure	17. DISTRIBUTION STATEMENT	
19. SECURITY CLASSIFICATION *(of this report)* Unclassified	20. NUMBER OF PAGES 77	21. COST OF REPORT CHARGED N/A

Reproduction of completed page authorized.

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

Strategies for Reducing Pedestrian and Bicyclist Injury at the Corridor Level

Phase 3

FINAL REPORT

Task Order 65A0678

by

Julia Griswold, Joy Pasquet, Jiajian Lu, Aditya Medury, and Offer Grembek

Safe Transportation Research and Education Center
University of California, Berkeley

Berkeley SafeTREC
SAFE TRANSPORTATION RESEARCH AND EDUCATION CENTER

for

California Department of Transportation

Acknowledgments

This research was sponsored by the Division of Research, Innovation and System Information (DRISI) of the California Department of Transportation. We would like to thank Caltrans Division of Safety Programs and DRISI staff for supporting the research and providing feedback throughout the study.

Table of Contents

Executive Summary

The Strategies for Reducing Pedestrian and Bicyclist Injury at the Corridor Level project is an effort of the California Department of Transportation (Caltrans) to identify and address systematic and specific problems with regard to pedestrian and bicycle safety in California, with the long-term goal of substantially reducing pedestrian and bicycle fatalities and injuries in the state.

This project focused on working closely with Caltrans to translate the tools and models of previous pedestrian efforts into an implementable program that supports activities to reduce bicycle crashes, and also continuing to develop additional methods that support improvements in pedestrian safety to assist Caltrans in meeting its pedestrian and bicycle safety goals.

Specifically, the core of this phase included four overarching objectives:

Develop a bicycle safety monitoring tool - The traditional, dominant approach used by state agencies to allocate safety resources is framed around the identification of safety hotspots, where agencies prioritize locations eligible for safety improvements based on historical collision concentrations. Corridor approaches result in safety improvements along entire corridors, which include some hotspots as well as locations with lower collision concentrations. The bicycle safety monitoring tool is based on the corresponding pedestrian tool and methodologies developed under 65A0547 to identify bicycle-related HCCLs on the state highway system. In addition, the research team developed a methodology for bicycle crash corridor identification using the DBSCAN algorithm. This tool will be used to support a pilot bicycle monitoring program as proposed by Caltrans.

Support the pedestrian safety monitoring tool - The research team responded to enhancements that need to be added to the functionality of the existing PSMR tool developed in a previous project.

Develop a systemic approach and tool for bicycles - The systemic approach consists of targeting blanket improvements at sites across a road network based on specific roadway features that are associated with a particular crash type. It uses historical crash data to identify the types of roadways that suffer from recurring safety concerns and provides a mechanism to make improvements also at sites that do not have many (or any) crashes. Research team developed a systemic approach for bicycles as well as a spreadsheet-based systemic tool. The systemic approach is intended to be complementary to the hotspot and corridor approaches.

Develop an approach to model bicycle exposure for the state highway system - The existing Caltrans TASAS-TSN highway database does not include any bicycle volume data, which are an important tool for understanding bicycle exposure and risk. To meet this challenge the research team developed an approach to modeling bicycle volumes across the state highway system. The modeling approach includes a Poisson mixture model, which allows for separate formulation of relationships for different types of trips—i.e., to have a utilitarian component and a recreational component within the same model. The research team performed data collection and processing to support implementation of the model in a future study.

Chapter 1. Introduction

The Strategies for Reducing Pedestrian and Bicyclist Injury at the Corridor Level project is an effort of the California Department of Transportation (Caltrans) to identify and address systematic and specific problems with regard to pedestrian and bicycle safety in California, with the long-term goal of substantially reducing pedestrian and bicycle fatalities and injuries in the state.

This focus on improved pedestrian and bicycle safety in California dovetails well with efforts already underway. For example, the Pedestrian and Bicycle Challenge Areas of the Strategic Highway Safety Plan have worked for several years to represent the needs of pedestrians and bicyclists at the State level and to develop achievable goals for improved safety. Furthermore, Caltrans underwent an external evaluation by the State Smart Transportation Initiative (SSTI) in 2014 to understand how it can improve its performance going forward. While rightly pointing out the leadership Caltrans has displayed in the past, the SSTI report also highlighted the need for Caltrans to modify its efforts and programming to better reflect statewide goals of improved safety and mobility for non-motorized modes (SSTI, 2014). The earlier phases of this project show that Caltrans has already made progress in this direction, and provides avenues to further the progress through suggested future research.

Current road safety management practices can be assessed against a proactiveness continuum that goes from fully reactive approaches to truly proactive ones (See Figure 1-1). The traditional, dominant approach used by state agencies to allocate safety resources is framed around the identification of safety hotspots, where agencies prioritize locations eligible for safety improvements based on historical collision concentrations. This focus constitutes a reactive approach, where the possibility of a safety improvement for a location is tied to the that fact that crashes already occurred at that location. Corridor approaches are slightly less reactive in that they result in safety improvements along entire corridor, which include some hotspots as well as locations with lower collision concentrations. Conversely, new approaches like "Vision Zero" and "safe system" that attempt to prevent the very possibility of severe crashes anticipate the occurrence of crashes and target improvements at locations regardless of their historical collision profile. The systemic approach finds itself somewhat in-between these two extremes, with both reactive and proactive components to it. It consists of targeting blanket improvements at sites across a road network based on specific roadway features that are associated with a particular crash type. It uses historical crash data to identify the types of roadways that suffer from recurring safety concerns, which qualifies it as a partly reactive approach. But on the other hand, the fact that it provides a mechanism to make improvements also at sites that did not have many (or any) crashes yet makes it a partly proactive approach. The systemic approach is typically used in parallel to the hotspot approach and is considered a complement rather than an alternative.

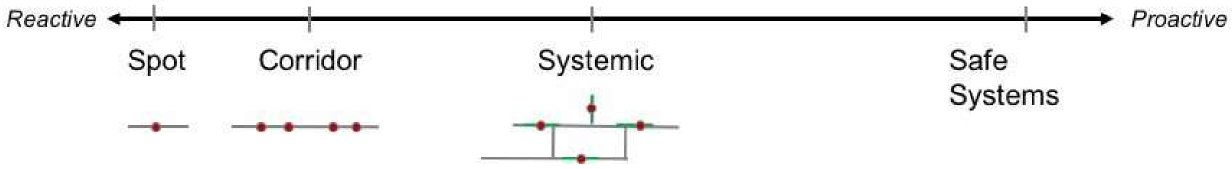


Figure 1-1. The Systemic Approach on the Reactive-Proactive Continuum

Review of past and on-going research

Phase II of the project (65A0509) was completed on January 31, 2017. Under that phase, Caltrans, in partnership with the University of California, Berkeley Safe Transportation Research and Education Center, accomplished several tasks and activities: (i) studied where systemic efforts fit alongside other road safety management efforts in terms of reactive vs. proactive approaches; (ii) identified the core components of the systemic approach which led to the development of the systemic matrix; (iii) developed and populated the systemic pedestrian crash matrix using available crash and roadway data; (iv) customized matrices for intersection and roadway sections; (v) developed lists of relevant countermeasures for each matrix cell; and (vi) developed a user-friendly prototype tool in MS Excel that can conduct such an analysis and produce a list of attributes of relevant countermeasures.

The outcome is a methodology to support systemic pedestrian efforts across the California state highway system. The methodology is incorporated into a user-friendly MS Excel prototype tool to conduct systemic pedestrian efforts analyses and safety improvements.

Moreover, as part of Phase II and III of the Pedestrian Safety Improvement Program Project (65A0547 and 65A0712), several other relevant activities are being developed that can support this project and include: (i) a model to estimate pedestrian volumes for different facility types across the state highway system; and (ii) methods to identify pedestrian hotspots across the California State Highway System.

Objectives

The current phase for this project focused on working closely with Caltrans to translate the tools and models of pedestrian efforts into an implementable program that supports activities to reduce bicycle crashes, and also continuing to develop additional methods that support improvements in pedestrian safety to assist Caltrans in meeting its pedestrian and bicycle safety goals.

Specifically, the core of this phase included four overarching objectives:

Develop a bicycle safety monitoring tool - The tool is based on the pedestrian tool and methodologies developed under 65A0547 to identify bicycle-related HCCLs on the state highway system. This tool will be used to support a pilot bicycle monitoring program as proposed by Caltrans.

Support the pedestrian safety monitoring tool - The research team responded to enhancements that need to be added to the functionality of the existing PSMR tool developed in a previous project.

Develop a systemic approach and tool for bicycles - In addition to the corridor approach developed in this project the research team also developed a systemic approach that seeks blanket improvements that can be implemented at sites across the road network, based on specific roadway features that are associated with a particular crash type.

Develop an approach to model bicycle exposure for the state highway system - The existing Caltrans TASAS-TSN highway database does not include any bicycle volume data. To meet this challenge the research team developed an approach to modeling bicycle volumes across the state highway system. The research team performed data collection and processing to support implementation of the model in a future study.

The rest of this report is structured as follows. Chapter 2 describes the methodology developed for bicycle crash corridor prioritization. Chapter 3 describes the work to develop systemic matrices for bicycle crashes on highways and at intersections. Chapter 4 describes the work completed towards the development of a pilot bicycle exposure model for the state highway system.

Chapter 2. Bicycle Safety Monitoring Tool

Bicycle Crash Corridor Prioritization

Transportation safety professionals strive to build a system on which no street user can be severely or fatally injured. To accomplish such a safe system, it is necessary to effectively harness all of the core protective opportunities provided by the system. For bicycle safety this includes safe street design with adequate separation from motorized traffic, safe vehicles, safe bicycles, safe cyclist behavior, safe behavior of other street users, all of which are governed by safe speeds, and supported by the medical emergency system when needed. Due to insufficient efforts to prevent such catastrophic outcomes, bicycle crashes occur. A total of 155 fatal bicycle crashes were reported in California in 2018. Across the United States 857 fatal bicycle crashes were reported in 2016. The number of bicycle fatalities in 2018 reflects an increase of 55 percent relative to 2010 (the end of the most recent financial crisis), and an increase of 38 percent across the country.

Each crash involves a series of undesirable events that results in an instantaneous and violent transfer of kinetic energy occurring in a specific time-space spot. The fact that these preventable outcomes occur in specific spots, warrants the justifiable response of investigating what went wrong and what changes need to be implemented so it does not happen at this spot again. While such spot efforts are a necessity and can save lives, the approach addresses the problem in a reactive manner and on a very small scale. Moreover, the shortage of data to sufficiently reconstruct the events that led to the crash cannot facilitate considerations across all the core protective opportunities.

To complement spot identification methods, transportation safety practitioners have shown an increased interest in developing approaches that can also lead to the implementation of improvements in a proactive manner. In this context, the distinction between reactive and proactive lies in the ability to make improvements exactly where crashes have occurred versus the ability to make improvements also at locations that have not yet experienced a crash, or have had relatively fewer crashes. To accomplish this, the scope of the potential implementation should have a common thread that goes beyond the number of crashes in a spot. For example, if a continuous set of segments (or intersections) is defined as a long segment and labeled as a corridor, it is possible to compare the safety of corridors across the network. The outcome of this can generate a subset of corridors that exhibit a level of safety that is below a certain threshold. If an improvement is identified and implemented across all the segments (or intersections) of the corridor it is implemented in a proactive manner. The conjecture here is that when crashes occur at a significant number of spots along the corridor, intermediate spots that have not had a crash yet, might be exposed to similar safety issues that can be addressed in a proactive manner. The outcome is that, in comparison to spot approaches, corridors present opportunities for larger scale operational and safety improvements along a route. Systemic safety is another category of proactive safety management, under which the common thread is not spatial but rather a set of locations that share common design attributes that are associated with a specific risk.

In the context of bicycle activity, corridors provide opportunities for improving both bicycle safety and mobility, through corridor-level improvement projects such as installation of bicycle lanes, providing additional protection through the removal or modification of parking lanes, installing traffic-calming measures, etc. Examples of intersection-specific modifications to better accommodate cyclists include bicycle boxes, lead bicycle crossing intervals, and protected intersections.

Methodology

Corridor identification

Density Based Spatial Clustering of Application with Noise (DBSCAN) is used to identify bicycle crash corridors. The DBSCAN algorithm is illustrated in Figure 2-1. Given a set of points in two-dimensional space, the algorithm can detect core points (A in Figure 2-1), border points (B in Figure 2-1), and outliers (N in Figure 2-1) under two predefined parameters: (i) a searching radius; and (ii) the minimum number of points within the area. Points that satisfy the above two criteria are designated as core points, while points that do not satisfy these criteria but are within a radius of these core points are designated as border points, while others are designated as outliers. A cluster is formed by at least one core point and a few border points. Several continuous core points are density reachable to each other, allowing a cluster to be extended. Therefore, since the length of a cluster is not restricted, DBSCAN can be used to identify bicycle crash corridors which also have flexible lengths.

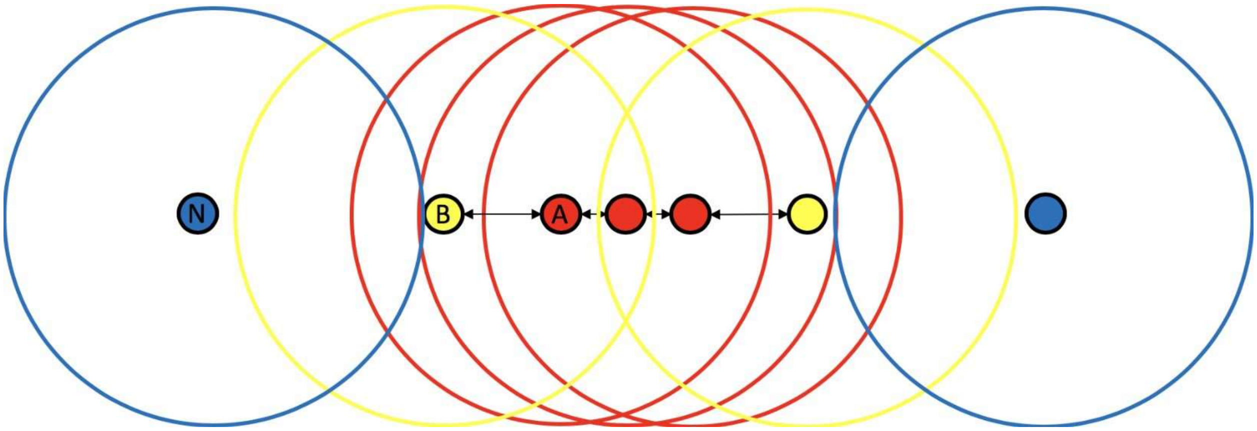


Figure 2-1. DBSCAN Illustration

In the case of bicycle crash corridor identification, each point in the figure represents a crash location, while each cluster represents a crash corridor. The searching radius was set to 0.25 mile and the minimum number of points within the area was set to 5.

Corridor Analysis

Several attributes of crash corridors are defined in order to help safety investigators for prioritization:

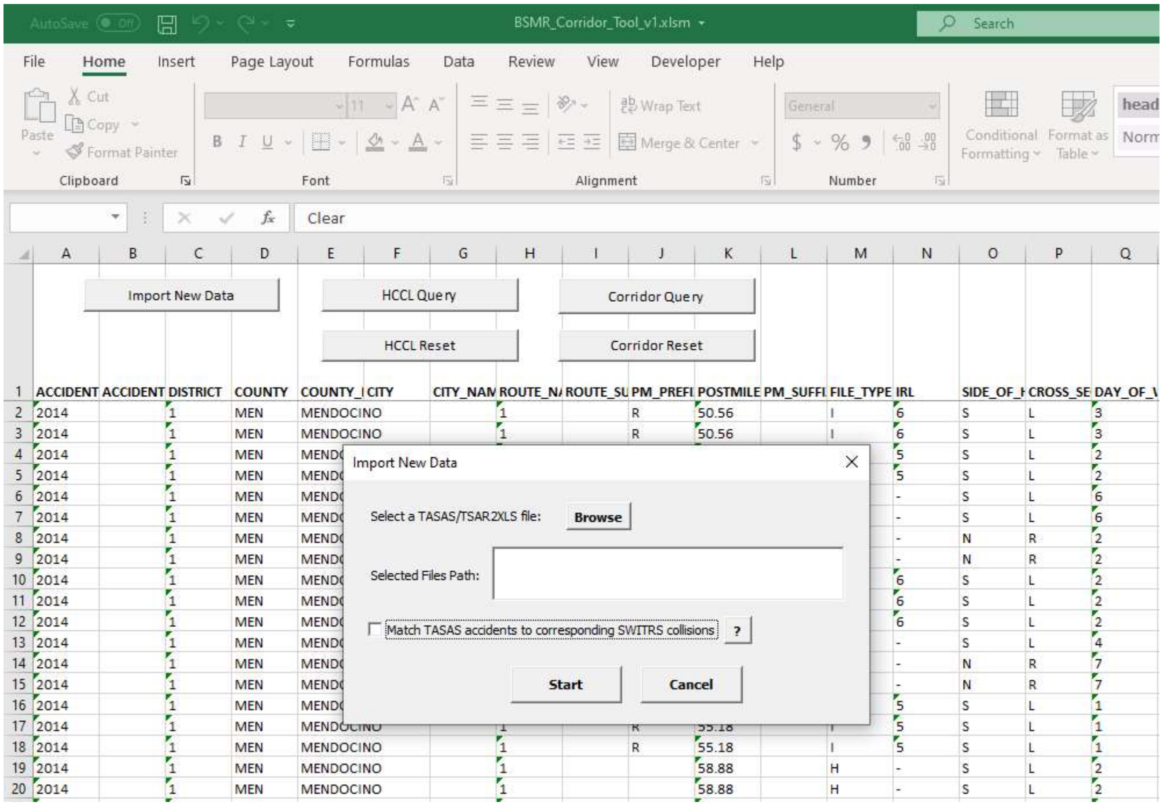
1. Corridor Length L: the distance between the postmile of the start point and the end point of a corridor.
2. Total number of crashes in a corridor N: the number of crashes that occurred within a corridor.
3. Crash per mile $CPM = \frac{N}{L}$: the density of the crashes for a corridor.

Corridors can be prioritized based on the ranking of the above attributes.

BSMR Tool

We implemented the above algorithm in a MS Excel Macro and built a revised Bicycle Safety Monitoring Report (BSMR) Tool for high collision concentration location (HCCL) and corridor analysis. Below are some functionalities of the tool.

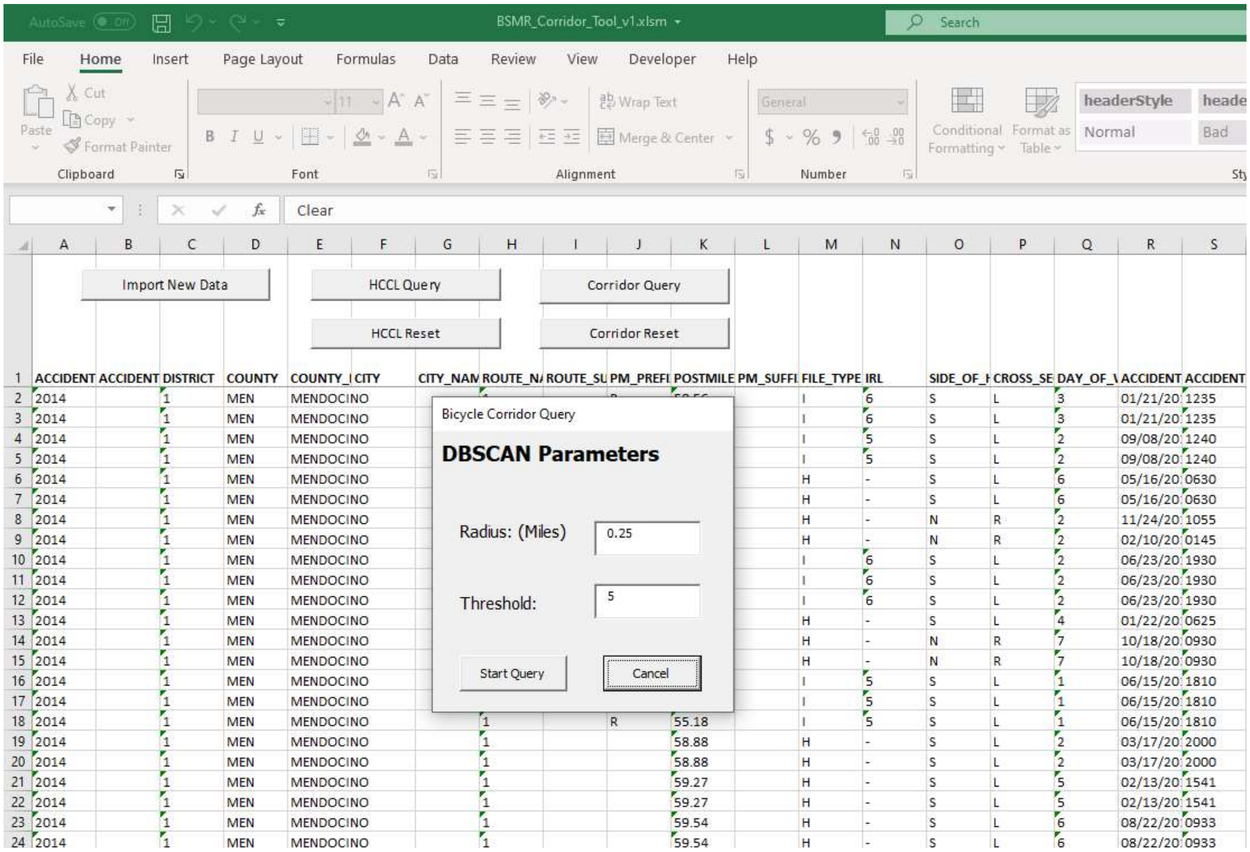
The tool uses collision data from TASAS tables in CSV format. Figure 2-2 shows the interface for importing crash data and the option to match TASAS data to the corresponding SWITRS data.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	ACCIDENT	ACCIDENT	DISTRICT	COUNTY	COUNTY_I	CITY	CITY_NAM	ROUTE_N/	ROUTE_SU	PM_PREFI	POSTMILE	PM_SUFFI	FILE_TYPE	IRL		SIDE_OF_H	CROSS_SE	DAY_OF_V
2	2014		1	MEN	MENDOCINO			1		R	50.56		I	6	S	L	3
3	2014		1	MEN	MENDOCINO			1		R	50.56		I	6	S	L	3
4	2014		1	MEN	MEND									5	S	L	2
5	2014		1	MEN	MEND									5	S	L	2
6	2014		1	MEN	MEND									-	S	L	6
7	2014		1	MEN	MEND									-	S	L	6
8	2014		1	MEN	MEND									-	N	R	2
9	2014		1	MEN	MEND									-	N	R	2
10	2014		1	MEN	MEND									6	S	L	2
11	2014		1	MEN	MEND									6	S	L	2
12	2014		1	MEN	MEND									6	S	L	2
13	2014		1	MEN	MEND									-	S	L	4
14	2014		1	MEN	MEND									-	N	R	7
15	2014		1	MEN	MEND									-	N	R	7
16	2014		1	MEN	MEND									5	S	L	1
17	2014		1	MEN	MENDOCINO			1		R	55.18		I	5	S	L	1
18	2014		1	MEN	MENDOCINO			1		R	55.18		I	5	S	L	1
19	2014		1	MEN	MENDOCINO			1			58.88		H	-	S	L	2
20	2014		1	MEN	MENDOCINO			1			58.88		H	-	S	L	2

Figure 2-2. Import crash data

The bicycle corridor query (Figure 2-3) requires two parameters, the search radius in miles and the threshold number of collisions. The defaults are 0.25 miles and 5 collisions, respectively, but the safety investigator may adjust the values to modify the prioritization.



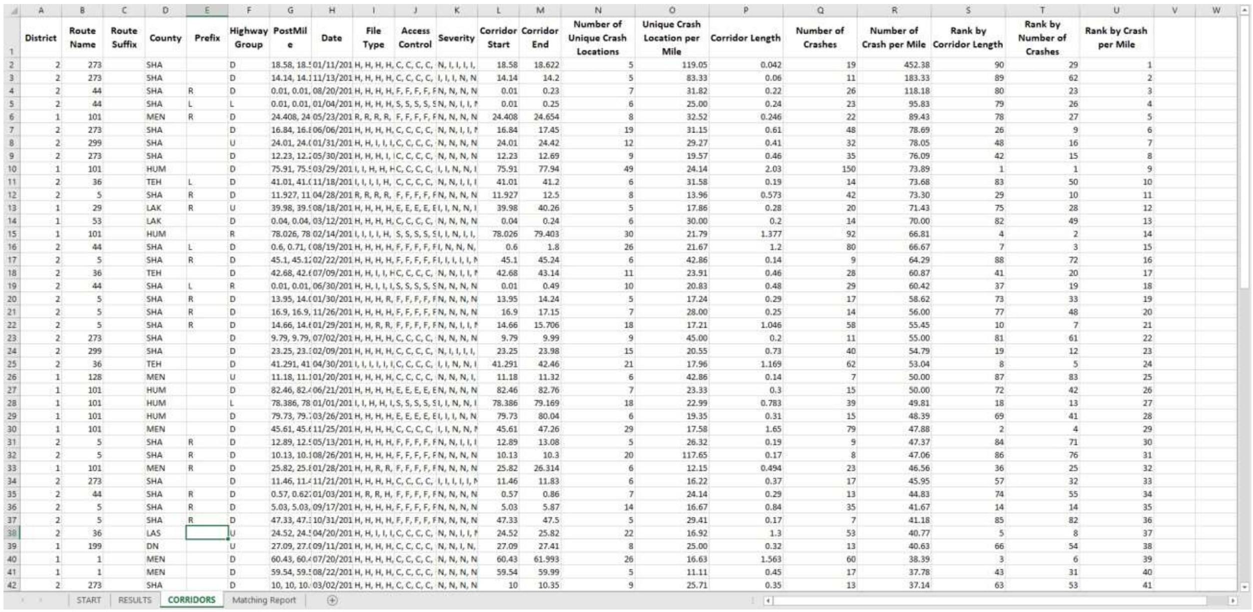
Radius: (Miles) 0.25

Threshold: 5

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	ACCIDENT	ACCIDENT	DISTRICT	COUNTY	COUNTY_I	CITY	CITY_NAM	ROUTE_N/	ROUTE_SU	PM_PREFI	POSTMILE	PM_SUFFI	FILE_TYPE	IRL	SIDE_OF_I	CROSS_SE	DAY_OF_V	ACCIDENT	ACCIDENT
2	2014		1	MEN	MENDOCINO								I	6	S	L	3	01/21/20	1235
3	2014		1	MEN	MENDOCINO								I	6	S	L	3	01/21/20	1235
4	2014		1	MEN	MENDOCINO								I	5	S	L	2	09/08/20	1240
5	2014		1	MEN	MENDOCINO								I	5	S	L	2	09/08/20	1240
6	2014		1	MEN	MENDOCINO								H	-	S	L	6	05/16/20	0630
7	2014		1	MEN	MENDOCINO								H	-	S	L	6	05/16/20	0630
8	2014		1	MEN	MENDOCINO								H	-	N	R	2	11/24/20	1055
9	2014		1	MEN	MENDOCINO								H	-	N	R	2	02/10/20	0145
10	2014		1	MEN	MENDOCINO								I	6	S	L	2	06/23/20	1930
11	2014		1	MEN	MENDOCINO								I	6	S	L	2	06/23/20	1930
12	2014		1	MEN	MENDOCINO								I	6	S	L	2	06/23/20	1930
13	2014		1	MEN	MENDOCINO								H	-	S	L	4	01/22/20	0625
14	2014		1	MEN	MENDOCINO								H	-	N	R	7	10/18/20	0930
15	2014		1	MEN	MENDOCINO								H	-	N	R	7	10/18/20	0930
16	2014		1	MEN	MENDOCINO								I	5	S	L	1	06/15/20	1810
17	2014		1	MEN	MENDOCINO								I	5	S	L	1	06/15/20	1810
18	2014		1	MEN	MENDOCINO			1		R	55.18		I	5	S	L	1	06/15/20	1810
19	2014		1	MEN	MENDOCINO			1			58.88		H	-	S	L	2	03/17/20	2000
20	2014		1	MEN	MENDOCINO			1			58.88		H	-	S	L	2	03/17/20	2000
21	2014		1	MEN	MENDOCINO			1			59.27		H	-	S	L	5	02/13/20	1541
22	2014		1	MEN	MENDOCINO			1			59.27		H	-	S	L	5	02/13/20	1541
23	2014		1	MEN	MENDOCINO			1			59.54		H	-	S	L	6	08/22/20	0933
24	2014		1	MEN	MENDOCINO			1			59.54		H	-	S	L	6	08/22/20	0933

Figure 2-3. Bicycle crash corridor query

Figure 2-4 shows the output results from the corridor query. Each record describes one corridor, and specific collision attributes such as postmile, date, facility type, access control, and severity are listed in series in the respective fields. Other fields include summaries and rankings of the corridor for use in prioritization.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	District	Route Name	Route Suffix	County	Prefix	Highway Group	PostMile	Date	File Type	Access Control	Severity	Corridor Start	Corridor End	Number of Unique Crash Locations	Unique Crash Location per Mile	Corridor Length	Number of Crashes	Number of Crash per Mile	Rank by Corridor Length	Rank by Number of Crashes	Rank by Crash per Mile
2	2	273		SHA		D	18.58, 18.5	01/11/201	H, H, H, H,	C, C, C, C,	N, I, I, I, I,	18.58	18.622	5	119.05	0.042	19	452.38	90	29	1
3	2	273		SHA		D	14.14, 14.1	11/13/201	H, H, H, H,	C, C, C, C,	I, I, I, N, N	14.14	14.2	5	83.33	0.06	11	183.33	89	62	2
4	2	44		SHA	R	D	0.01, 0.01,	08/20/201	H, H, H, H,	F, F, F, F, F	N, N, N, N	0.01	0.23	7	31.82	0.22	26	118.18	80	23	3
5	2	44		SHA	L	L	0.01, 0.01,	01/04/201	H, H, H, H,	S, S, S, S, S	N, N, I, I, N	0.01	0.25	6	25.00	0.24	23	95.83	79	26	4
6	1	101		MEN	R	D	24.408, 24	05/23/201	R, R, R, R,	F, F, F, F, F	N, N, N, N	24.408	24.654	8	32.52	0.246	22	89.43	78	27	5
7	2	273		SHA		D	16.84, 16.8	06/06/201	H, H, H, H,	C, C, C, C,	N, N, I, I, N	16.84	17.45	19	31.15	0.61	48	78.69	26	9	6
8	2	299		SHA		U	24.01, 24.0	01/31/201	H, H, I, I, I,	C, C, C, C,	N, N, N, N	24.01	24.42	12	29.27	0.41	32	78.05	48	16	7
9	2	273		SHA		D	12.23, 12.2	05/30/201	H, H, H, I, I	C, C, C, C,	N, N, N, N	12.23	12.69	9	19.57	0.46	35	76.09	42	15	8
10	1	101		HUM		D	75.91, 75.9	03/29/201	I, I, H, H, H	C, C, C, C,	I, I, N, N, I	75.91	77.94	49	24.14	2.03	150	73.89	1	1	9
11	2	36		TEH	L	D	41.01, 41.0	11/18/201	I, I, I, I, H,	C, C, C, C,	N, N, I, I, I	41.01	41.2	6	31.58	0.19	14	73.68	83	50	10
12	2	5		SHA	R	D	11.927, 11	04/28/201	R, R, R, R,	F, F, F, F, F	N, N, N, N	11.927	12.5	8	13.96	0.573	42	73.30	29	10	11
13	1	29		LAK	R	U	39.98, 39.9	08/18/201	H, H, H, H,	E, E, E, E, E	I, I, N, N, I	39.98	40.26	5	17.86	0.28	20	71.43	75	28	12
14	1	53		LAK		D	0.04, 0.04,	03/12/201	H, H, H, H,	C, C, C, C,	N, N, N, N	0.04	0.24	6	30.00	0.2	14	70.00	82	49	13
15	1	101		HUM		R	78.026, 78	02/14/201	I, I, I, I, H,	S, S, S, S, S	I, I, N, I, I,	78.026	79.403	30	21.79	1.377	92	66.81	4	2	14
16	2	44		SHA	L	D	0.6, 0.71, 0	08/19/201	H, H, H, H,	F, F, F, F, F	I, N, N, N,	0.6	1.8	26	21.67	1.2	80	66.67	7	3	15
17	2	5		SHA	R	D	45.1, 45.12	02/22/201	H, H, H, H,	F, F, F, F, F	I, I, I, I, I, N	45.1	45.24	6	42.86	0.14	9	64.29	88	72	16
18	2	36		TEH		D	42.68, 42.6	07/09/201	H, H, I, I, H	C, C, C, C,	N, N, I, I, N	42.68	43.14	11	23.91	0.46	28	60.87	41	20	17
19	2	44		SHA	L	R	0.01, 0.01,	06/30/201	H, H, I, I, I,	S, S, S, S, S	N, N, N, N	0.01	0.49	10	20.83	0.48	29	60.42	37	19	18
20	2	5		SHA	R	D	13.95, 14.0	01/30/201	H, H, H, R,	F, F, F, F, F	N, N, N, N	13.95	14.24	5	17.24	0.29	17	58.62	73	33	19
21	2	5		SHA	R	D	16.9, 16.9,	11/26/201	H, H, H, H,	F, F, F, F, F	N, N, N, N	16.9	17.15	7	28.00	0.25	14	56.00	77	48	20
22	2	5		SHA	R	D	14.66, 14.6	01/29/201	H, H, R, R,	F, F, F, F, F	N, N, I, I, N	14.66	15.706	18	17.21	1.046	58	55.45	10	7	21
23	2	273		SHA		D	9.79, 9.79,	07/02/201	H, H, H, H,	C, C, C, C,	N, N, N, N	9.79	9.99	9	45.00	0.2	11	55.00	81	61	22
24	2	299		SHA		D	23.25, 23.3	02/09/201	H, H, H, H,	C, C, C, C,	N, I, I, I, I,	23.25	23.98	15	20.55	0.73	40	54.79	19	12	23
25	2	36		TEH		D	41.291, 41	04/30/201	I, I, I, I, I, I,	C, C, C, C,	I, I, N, N, I	41.291	42.46	21	17.96	1.169	62	53.04	8	5	24
26	1	128		MEN		U	11.18, 11.1	01/20/201	H, H, H, H,	C, C, C, C,	N, N, N, I,	11.18	11.32	6	42.86	0.14	7	50.00	87	83	25
27	1	101		HUM		D	82.46, 82.4	06/21/201	H, H, H, H,	E, E, E, E, E	N, N, N, N	82.46	82.76	7	23.33	0.3	15	50.00	72	42	26
28	1	101		HUM		L	78.386, 78	01/01/201	I, I, H, H, I,	S, S, S, S, S	I, I, N, N, I	78.386	79.169	18	22.99	0.783	39	49.81	18	13	27
29	1	101		HUM		D	79.73, 79.7	03/26/201	H, H, H, H,	E, E, E, E, E	I, I, I, N, N	79.73	80.04	6	19.35	0.31	15	48.39	69	41	28
30	1	101		MEN		D	45.61, 45.6	11/25/201	H, H, H, H,	C, C, C, C,	I, I, N, N, N	45.61	47.26	29	17.58	1.65	79	47.88	2	4	29
31	2	5		SHA	R	D	12.89, 12.9	05/13/201	H, H, H, H,	F, F, F, F, F	N, N, I, I, I	12.89	13.08	5	26.32	0.19	9	47.37	84	71	30
32	2	5		SHA	R	D	10.13, 10.1	08/26/201	H, H, H, H,	F, F, F, F, F	N, N, N, N	10.13	10.3	20	117.65	0.17	8	47.06	86	76	31
33	1	101		MEN	R	D	25.82, 25.8	01/28/201	H, H, R, R,	F, F, F, F, F	N, N, N, N	25.82	26.314	6	12.15	0.494	23	46.56	36	25	32
34	2	273		SHA		D	11.46, 11.4	11/21/201	H, H, H, H,	C, C, C, C,	I, I, I, I, I, N	11.46	11.83	6	16.22	0.37	17	45.95	57	32	33
35	2	44		SHA	R	D	0.57, 0.62	01/03/201	H, R, R, H,	F, F, F, F, F	N, N, N, N	0.57	0.86	7	24.14	0.29	13	44.83	74	55	34
36	2	5		SHA	R	D	5.03, 5.03,	09/17/201	H, H, H, H,	F, F, F, F, F	N, N, N, N	5.03	5.87	14	16.67	0.84	35	41.67	14	14	35
37	2	5		SHA	R	D	47.33, 47.3	10/31/201	H, H, H, H,	F, F, F, F, F	N, N, N, N	47.33	47.5	5	29.41	0.17	7	41.18	85	82	36
38	2	36		LAS		U	24.52, 24.5	04/20/201	H, H, I, I, I,	C, C, C, C,	N, N, I, I, N	24.52	25.82	22	16.92	1.3	53	40.77	5	8	37
39	1	199		DN		U	27.09, 27.0	09/11/201	H, H, H, H,	C, C, C, C,	N, N, I, N,	27.09	27.41	8	25.00	0.32	13	40.63	66	54	38
40	1	1		MEN		D	60.43, 60.4	07/20/201	H, H, H, H,	C, C, C, C,	N, N, N, N	60.43	61.993	26	16.63	1.563	60	38.39	3	6	39
41	1	1		MEN		D	59.54, 59.5	08/22/201	H, H, H, H,	C, C, C, C,	N, N, N, N	59.54	59.99	5	11.11	0.45	17	37.78	43	31	40
42	2	273		SHA		D	10, 10, 10.	03/02/201	H, H, H, H,	C, C, C, C,	N, N, N, N	10	10.35	9	25.71	0.35	13	37.14	63	53	41

Figure 2-4. Corridor results and prioritization

Chapter 3. Systemic Safety Tool for Bicycles

Road safety researchers and professionals have identified the value of the systemic approach to safety. The Federal Highway Administration Office of Safety has acknowledged four benefits of the systemic approach and developed the Systemic Safety Project Selection Tool (Figure 3-1): (i) solves an unmet need in transportation safety; (ii) uses a risk-based approach to prevent crashes; (iii) results in a comprehensive road safety program; and (iv) advances a cost-effective means to address safety concerns (FHWA 2013).

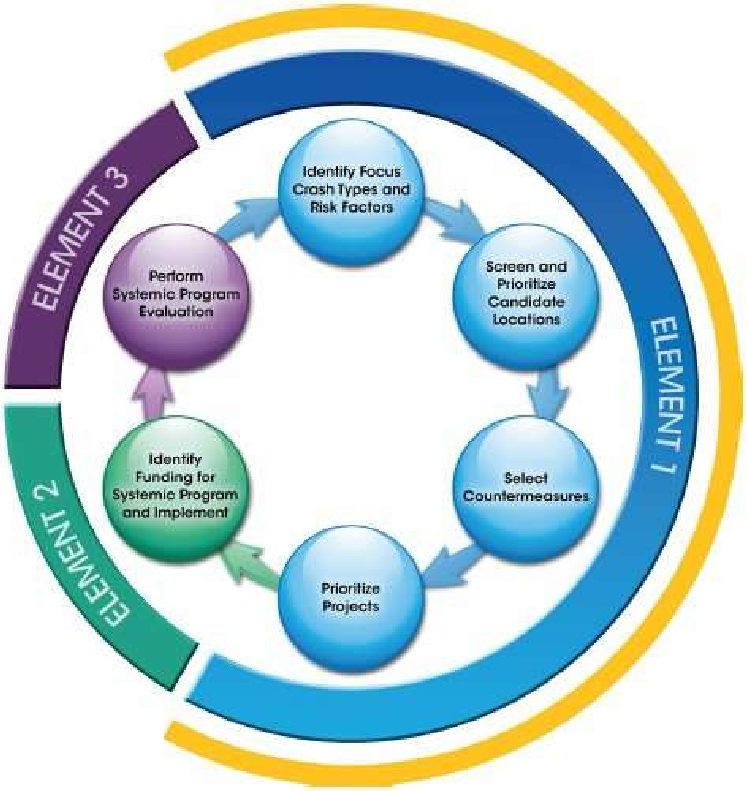

ELEMENT 1
ELEMENT 2
ELEMENT 3
Identify Focus Crash Types and Risk Factors
Screen and Prioritize Candidate Locations
Select Countermeasures
Prioritize Projects
Identify Funding for Systemic Program and Implement
Perform Systemic Program Evaluation

Figure 3-1. FHWA's Systemic Safety Project Selection Tool. Element 1: The Systemic Safety Planning Process (from FHWA 2013)

This chapter describes the overall task and findings. The next section includes a description of the proposed systemic approach. The subsequent sections detail the core elements of the systemic crash elements and the resources that were used to establish the bicycle crash matrix structure, present the development of the countermeasure matrix, and describe the development of the prototype Excel tool. Finally, we present conclusions and recommendations for further development and implementation of the proposed approach and tool.

The Systemic Matrix Approach

The systemic approach to road safety originated at the intersection of two distinct strategies of road safety management that have emerged over the past two decades—the traditional, reactive approach, and the proactive approach.

The first is the traditional approach, for which sites with a higher than expected occurrence of crashes are identified. Appropriate countermeasures are then adopted for these specific sites, which are commonly called hotspots. Whether the high-crash locations are isolated from one another in a *spot approach* or considered along corridors with recurring safety concerns in a *corridor approach*, both schemes utilize a *reactive* rationale. This type of approach is problematic because addressing safety issues requires waiting for a crash occurrence, and in addition because underreporting issues can negatively impact the accuracy of the data upon which safety analyses are based. Indeed, not all crashes are reported, and not all those that are reported can be found in a single database—police, hospitals, road administrations, and insurance companies each have their own reporting systems and cannot individually capture all previous crashes (Turner et al. 2015). Furthermore, in some contexts these high-risk location approaches are no longer relevant due to the increase of crashes at locations with fewer crashes (SWOV2010).

The second is the *proactive* approach, of which the most emblematic program is *Vision Zero*. This approach was first introduced in Sweden in 1997, when it was passed into the national legislation. Vision Zero maintains that no loss of life is acceptable for users of the transportation system and assigns the responsibility for traffic deaths and permanent injuries on the designer of the system. From this perspective, human error is considered and the system's features should make it impossible under any circumstances for anyone to be killed or severely injured as a result of road traffic. This approach has been proven to generate quite satisfying results in Sweden, with traffic deaths having plummeted since the new policy was implemented. The Swedish Vision Zero program has been translated into the *safe systems approach*, which according to involves "building a system in which people cannot be fatally or severely injured on despite human error." (Jobs et al. 2016a).

The systemic approach is found between these two extremes. Defined by the Federal Highway Administration (FHWA) as making "improvement[s] that [are] widely implemented based on high-risk roadway features that are correlated with particular crash types," the systemic approach intersects reactive and proactive strategies. Indeed, it uses historical crash data to target road facilities that have experienced higher incidences of crashes. However, it goes beyond identifying clusters of crashes, as it does not consider specific locations, but rather high-risk road features, and ultimately would also apply countermeasures to low or no-crash sites.

The rationale behind the adoption of this approach is that transportation agencies moved away from approaches trying to address all levels of crash severity and chose to focus on reducing the occurrence of the most severe crashes (Turner et al. 2015). Considering the low density and wide distribution of such crashes over the road network—in 2013, 53% of fatal crashes within the federal road network were located in rural areas—adopting a traditional hotspot approach would not efficiently identify potential safety investments. In addition, adopting a systematic approach, that is, implementing countermeasures across the entire network, is not realistic in a budget-constrained environment. The systemic approach therefore appears to be best suited to address the occurrence of severe crashes across road networks.

Measuring Safety Using the Systemic Approach: Choosing the Right Safety Indicators

Ultimately, the systemic approach is about improving road safety, by better identifying safety needs—that is, by better spotting unsafe features of the road network. But measuring road safety is not an easy task because the concept of safety itself is hard to define, though the term is broadly used among both experts and the public (Oppe 1994). The core problem lies in the fact that safety problems are brought to light when unsafe situations occur in the form of crashes and subsequent injury or fatality. This is why Ezra Hauer felt the need to state at the beginning of his *Observational before-after studies in road safety* that "road safety is manifest in the occurrence of accidents and their harm," as opposed to the subjective feeling of security (Hauer 1997). Crashes are manifestly correlated with road "unsafety," and crash counts have therefore been widely used as metrics for road safety, especially by policy makers because figures of road fatalities and injuries make a stronger case for road safety than complex measurements.

However, using crash data as a direct measure of road safety has its caveats. The overarching goal of safety analysis as formulated by Leonard Evans is to "examine factors associated with crashes with the aim of identifying those that can be changed by countermeasures (or interventions) to enhance future safety" (Evans, 1991). This implies the need for large sample sizes for significant statistical observations, which is not always possible when it comes to crash data. Additionally, relying solely on crash data ignores the fact that crashes themselves are a result of the emergence of hazardous situations—some of which resulted in a crash, while the others do not. This distinction is essential, because it recognizes an essential dilemma: what do we consider to be a safer system, a lower number of crashes or a lower risk of getting into a crash? Traffic events can be represented as a continuum of situations in pyramidal layers (see Fig. xx), whose volumes corresponds to an event's frequency (Hyden 1987; Hauer 1997; Tarko 2012). The connection between these events and road "unsafety" make the case for the use of surrogate measures of safety, which: (i) are correlated with the occurrence of crashes, and (ii) capture the effects of safety countermeasures (Hauer 1997; Gettman and Head 2003; Tarko et al. 2009; Tarko,2012). These two features make surrogate measures of road safety valuable because they deepen the understanding of factors leading to failure mechanisms in the road system. Furthermore, as shown in Fig. xx, more frequent events are easier to measure, which would call for a wider reliance on surrogate measures. Still, as mentioned by Hauer, using surrogate measures to quantify safety "rests on the observation that where there is smoke, there is fire." Such an assumption, true or not, reinforces the fundamental link between road safety and crashes. Nonetheless, some nuanced interpretation should be made between fluctuating crash counts and the permanent idea of the safety of a road entity: facilities with zero past occurrences of crashes should not be considered perfectly safe, since that only roads with no traffic at all have a zero chance of a crash.

Recognizing this last point calls for taking into account levels of traffic when measuring whether a traffic facility is safe. The busier a roadway is, the more likely it is, all else being equal, that vehicles will collide. Therefore, some studies have relied on crash exposure rather than crash frequency to measure road safety. And furthermore, why should agencies worry about fixing facilities that are not predominant within their network? This is the concern that could be addressed by relying on a third road safety indicator: crash density—the ratio between the number of crashes and the "size" of the road network, whether in terms of the mileage of roadways or the number of intersections. These last two indicators are quite similar, in the sense that they can be respectively seen as an activity-based measures of exposure and an infrastructure-based one. The Dutch Institute for Road Safety Research summarizes the concept of exposure measures as capturing a unit amount of risk—a unit that can express duration, distance, population, expected number of encounters, or other factors, depending on the intended use of the measure—mainly useful for making situations comparable (Oppe 1994; Hakkert et al. 2002). While they are in no way the only indicators for measuring safety on a road network, they share important advantages over more elaborate ones, the first one being their ease of calculation based on crash data from police reports and basic infrastructure data. Depending on data availability, it may be relevant for some agencies to consider mixed safety indicators that would go beyond the dichotomy between crash numbers and rates. An example is the ratio between crash frequency and vehicle-miles traveled, a combined infrastructure- and activity-based exposure rate that takes into account both the length of the road network and the traffic flow on the infrastructure. Many other safety indicators could be built, some of them more direct measures of safety, some more surrogate measures. Ultimately, the purpose of the present study is not to outline in a definitive manner *the* right way of measuring road safety, and *the* unique safety indicator to be used when implementing the systemic approach. Each and every indicator responds to different safety concerns, and choosing one over the others constitutes nothing less than a political choice. It is therefore the responsibility of each safety agency to decide which indicator is most appropriate in identifying systemic safety concerns. In this study, for illustrative purposes, the level of safety of a traffic facility is measured as the number of crashes, by kind and severity, that occurred on this facility during a specified period—considering that crash counts constitute the most directly available information on crashes based on any police reports, regardless of the local jurisdiction.

The Systemic Matrix Scheme

At the core of the proposed approach is an easy-to-interpret systemic crash matrix that shows what types of crashes occur on what types of facilities. Matrix rows represent crash types, while columns correspond to facility types. The cells of the matrix are referred to as crash profiles and include aggregate information on crashes that occurred for each crash profile. The way in which this information is aggregated depends on the chosen safety indicator—in the case of crash frequencies, each cell contains the number of type X crashes that occurred on type α roadways. The cells with the highest value represent systemic hotspots, which are systemic challenges on the roadway network in which a particular crash profile is consistently associated with a particular type of road infrastructure.

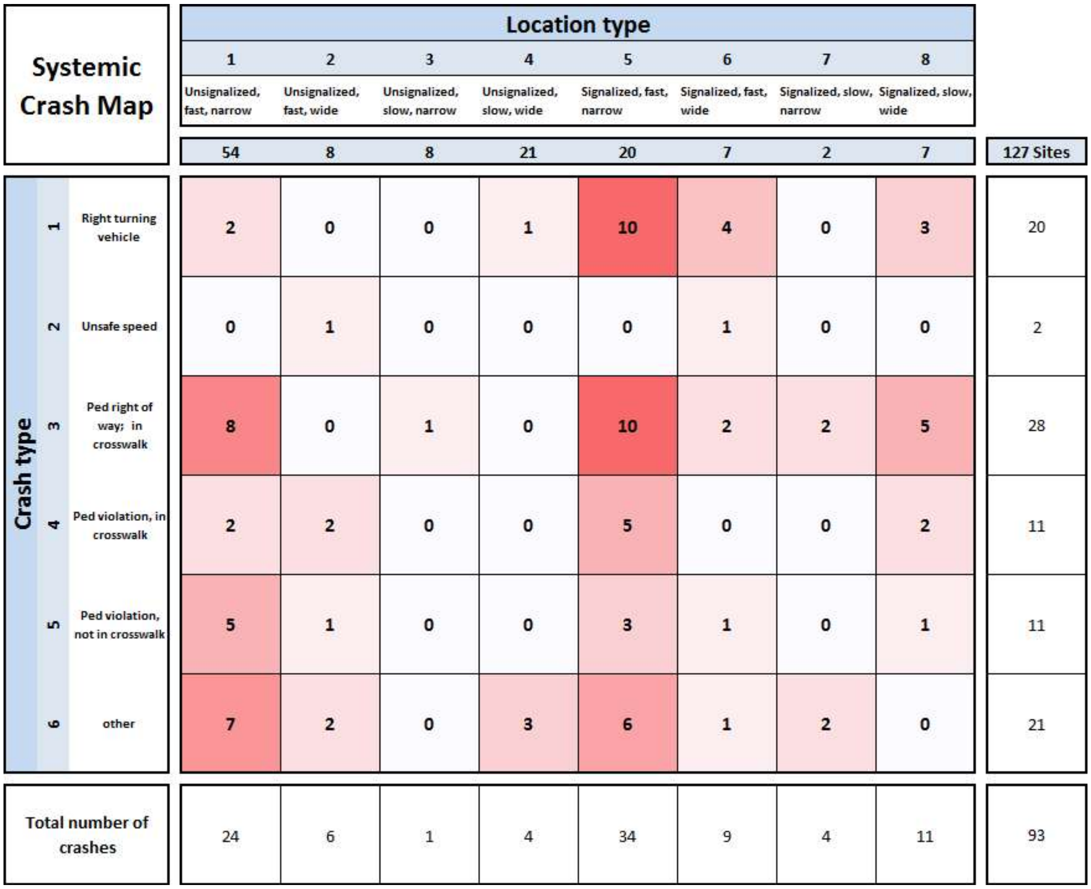
Using such a matrix provides agencies with a snapshot of any systemic problems on their networks that are both easy to assemble and to interpret. The advantage of this scheme is that it is compatible with the data-driven rationale of the systemic approach, offering enough flexibility to allow agencies with varying degrees of data availability to implement it. The approach mainly expands on two previous initiatives in the United States: FHWA's Systemic Safety Project Selection Tool, and California's Systemic Pedestrian Safety Analysis. Both approaches involved building a matrix, the rows and columns of which were determined to best illustrate the infrastructure-related dynamics behind road collisions. The FHWA tool has been regularly used to guide road safety analyses across the nation and to help prioritize locations. The process developed by FHWA starts with the identification of focus crash types and facility types based on crash data and infrastructure information. This principle was adopted by the Californian analysis, in which the crash matrix uses columns representing locational characteristics understood to influence the collisions and bases on data availability, and rows corresponding to crash types, understood as primary collision factors and behaviors thought to influence the crash. The following section will guide road safety professionals and researchers through the process of creating a systemic safety matrix. While it builds on findings from a methodology elaborated with data from the Highway Safety Information System (HSIS), this data-driven process was considered to be generalizable to other data sources and road networks beyond the seven HSIS member states, and the following will provide guidance on how to assess various matrix structures and select the most appropriate one.

The Crash Matrix

Once the matrix categories are identified, each category will result in a different pair of matrices (a crash matrix and a countermeasure matrix) being developed, with each their own structure, determined based on the dataset that falls within that category.

At the core of the proposed approach is a transparent systemic crash matrix that shows what types of crashes occur on what types of facilities. Matrix rows represent crash types, while columns correspond to facility types. The cells of the matrix are referred to as crash profiles and include aggregate information on the crashes that occurred for each crash profile. The way in which this information is aggregated depends on the chosen safety indicator: in the case of crash frequencies, each cell contains the number of type A crashes that happened on type α roadways. The cells with the highest value represent so-called systemic hotspots, i.e. systemic challenges on the roadway network where a particular crash profile is consistently associated with a particular type of road infrastructure. Using such a matrix provides state agencies with a snapshot of any systemic problems on their network that is both easy to assemble and to interpret. The benefits of this scheme are that it is compatible with the data-driven rationale of the systemic approach, offering enough flexibility to allow agencies with varying degrees of data availability to implement it.

Figure 3-2 shows the concept of the crash matrix. Total mileage of roadways or total number of intersections are also included. Each cell indicates the number of a specific type of crash (in the rows) happening at a specific type of location (in the columns). Systemic hotspots then get identified based on a screening criterion that either uses crash frequency or exposure to determine what crash profiles to prioritize in the implementation of systemic safety improvements.



Systemic Crash Map			Location type								
			1	2	3	4	5	6	7	8	
			Unsignalized, fast, narrow	Unsignalized, fast, wide	Unsignalized, slow, narrow	Unsignalized, slow, wide	Signalized, fast, narrow	Signalized, fast, wide	Signalized, slow, narrow	Signalized, slow, wide	
			54	8	8	21	20	7	2	7	127 Sites
Crash type	1	Right turning vehicle	2	0	0	1	10	4	0	3	20
	2	Unsafe speed	0	1	0	0	0	1	0	0	2
	3	Ped right of way; in crosswalk	8	0	1	0	10	2	2	5	28
	4	Ped violation, in crosswalk	2	2	0	0	5	0	0	2	11
	5	Ped violation, not in crosswalk	5	1	0	0	3	1	0	1	11
	6	other	7	2	0	3	6	1	2	0	21
Total number of crashes			24	6	1	4	34	9	4	11	93

Figure 3-2. Structure of systemic crash matrix

Developing the Crash Matrix

Data requirements

Building a systemic matrix requires the collection of historical data on road collisions as well as infrastructure characteristics – ideally of the entire road network, but at the very least of the locations where crashes occurred. The present section will detail data needs when developing a systemic crash matrix for bicycles. Each of the subsequent sub-sections will detail our guidelines for what data elements are needed to create systemic matrix, followed by the decisions we made regarding the development of systemic matrices with accident and infrastructure data from Caltrans Traffic Accident Surveillance and Analysis System (TASAS).

Crash data

The singularity of the systemic approach resides in both its reactive and proactiveness. The use of historical crash data to identify systemic road safety challenges is at the center of the approach. While not all characteristics of a specific crash are used to identify systemic hotspots, it is important to meet some minimum requirements when undertaking the systemic approach.

For the California State Highway System (SHS), collision data comes from TASAS data. Five years of data were used for the present project, with a first iteration using 2010-2014 data and the final version of the systemic bicycle tool being based on more recent data, from 2013 to 2017. In general, increasing the number of years of collision data represents one way of mitigating the lack of data points, which is critical in a data-driven process, as generalizability is key. However, the caveat of settling for too many years is that road infrastructure is not permanent. There should therefore be a balance between maximizing data points and considering potential roadway network changes, or travel pattern modifications evolutions. Assuming that the database structure is conservative/solid, this could be solved by ensuring the appropriate pairing of crashes with roadway characteristics that were prevailing at the time of the crash, but this requires regular internal updates to the roadway elements of the database and prevents from populating the matrix with exposure-based rates instead of crash frequency. This is the reason why using five years of data is recommended. The distribution of SHS crashes for the years 2013 to 2017 for California is shown on Figure 3-3.

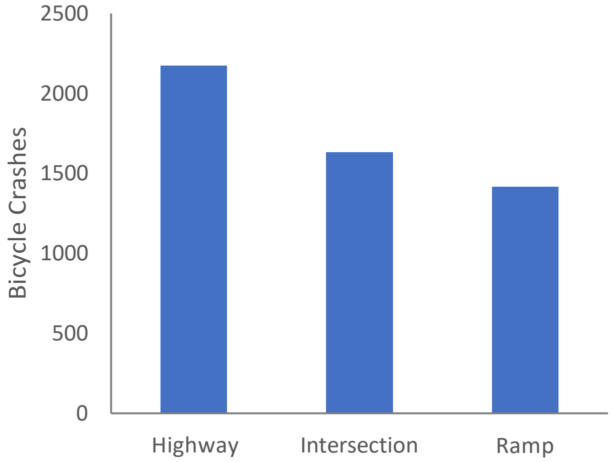


Figure 3-3. Distribution of 2013-2017 bicycle crashes by location type (Source: TASAS)

The systemic approach requires at a minimum the elements listed in Table 3-1. TASAS data includes most of these elements but is missing some information: for instance, the violation code is not present, and only the derived “primary cause” of the crash (which is equivalent to the “primary collision factor” in Table 3-1) is included. Information on the actual violation is more informative when it comes to implementing the systemic approach than having information on a broad primary cause, because it can dictate more closely what engineering countermeasure could address the corresponding violation. TASAS records were therefore matched with information for the same collisions included in the Statewide Integrated Traffic Records System (SWITRS) database to include the violation code as well as the party type at fault. This last variable has also proved to be very valuable information to determine what countermeasure would more appropriate: for example, a failure to yield by a motorized vehicle would not call for the same safety measures as a failure to yield by a bicycle.

Table 3-1. Minimum crash data requirements for the systemic approach

ID	Variables	Description	Data type	Data values	Comments
C1	Crash identifier	Unique identifier within a given year that identifies a particular crash.	Numeric or character string	0123456789	Value usually assigned by the police, as the first entity recording the incident at the crash scene.
C2	Crash date	Date on which the crash occurred.	Numeric	DDMMYYY	Useful for seasonal comparisons and time series analyses, among others.
C3	Crash time	Time at which the crash occurred, using the 24-hour clock format.	Numeric	HHMM	Useful for comparisons between periods (e.g. AM, PM, nighttime).
C4	Crash location	Exact location at which the crash occurred.	Character string		Various referencing methods are possible and include: (1) latitude/longitude coordinates; (2) linear referencing system; (3) link-node system. Ideally, a combination of GPS coordinates with the route name or another designation is desired to best relate geographic coordinates to roadway elements listed in the road infrastructure directory. If not available, the crash location should at the very least document the street or road name, a reference point, and the distance and direction from that reference point. The accuracy of the crash location documentation is critical for the identification and implementation of engineering countermeasures on crash sites.
C5	Crash type	Other party or object that led to the injury or damage-producing event of the crash.	Categorical	Moving vehicle; parked vehicle; pedestrian; bicycle; fixed object; non-fixed obstacle; animal; train; no object; etc.	Collisions can include more than one event. However, the main triggering element of the collision should be listed, and is key to identifying countermeasures.

ID	Variables	Description	Data type	Data values	Comments
C6	Primary collision factor	Principal cause of the collision.	Categorical	Alcohol; failure to yield; improper turn; following too closely; speeding; etc.	Similarly, there may be multiple factors at play in a single crash. Knowing the primary cause is key to identifying countermeasures.
C7	Violation code	If applicable, legal code of the traffic violation that led to the crash.	Categorical	22107	Provides more flexibility in the grouping of crashes by traffic violation types (e.g. control violation) than the standard primary collision factor (C6) categories listed above, at the discretion of the matrix developer. Provides more details on the specific primary causes of a collision.
C8	Impact type	Manner in which the motorized vehicle(s) involved initially collided with another vehicle, object or person.	Categorical	Single-vehicle crash; rear-end; head-on; Sideswipe; broadside; etc.	Useful for suggesting the trajectory of the vehicles involved in the collision.
C9	Movement prior to the collision	Type of movement of the primary vehicle preceding the first impact.	Categorical	Proceeding straight; left turn; right turn; U turn; backwards; changing lanes; unknown; etc.	Useful for suggesting the trajectory of the vehicles involved in the collision.
C10	Number of parties involved	Number of parties involved, including motorized and non-motorized vehicles	Numeric		Informs on the overall scale of the crash.
C11	Party type	Type of parties involved in the crash, in addition to the motorized vehicle(s).	Categorical	Auto-involved; pedestrian-involved; bicycle-involved.	Informs on the involvement of non-motorized individuals in the collision. Considering that some collisions may involve vehicles, bikes and pedestrians, they would be flagged as both pedestrian and bike-involved, and thus included in more than one of the matrix categories listed in the previous section.

ID	Variables	Description	Data type	Data values	Comments
C12	Crash severity	Most severe injury of any person involved.	Categorical	Fatal; severe injury; slight injury; property damage only.	Facilitates the grouping of crashes by severity level, thus enabling different policy focuses (e.g. reducing traffic deaths and severe injuries).
C13	Number of fatalities	Number of deaths resulting from the crash.	Numeric		Count includes all vehicles and individuals involved in the crash.
C14	Number of non-fatal injuries	Number of non-fatal injuries resulting from the crash.	Numeric		Count includes all vehicles and individuals involved in the crash.
C15	Weather conditions	Prevailing atmospheric conditions at the crash location, at the time of the crash.	Categorical	Clear; rain; snow; fog; strong winds; unknown; etc.	Unveils potential causes of vision impairment or challenging conditions of the road pavement surface.
C16	Light conditions	Level of natural and artificial light at the crash location, at the time of the crash.	Categorical	Daylight; dusk; dark; dark with streetlights; unknown; etc.	Unveils potential issues of visibility.

Infrastructure data

The systemic approach links crash profiles and infrastructure types, in order to unveil linkages between specific types of crashes and specific features of roadways, thus allowing the implementation of blanket improvement across an entire facility type. Infrastructural elements at the location of a collision are therefore central in the development of a systemic matrix.

In the systemic matrix, columns represent locational attributes of the infrastructure that help predict the likelihood of the occurrence of a crash. For the California State Highway System, infrastructure data comes from TASAS-TSN.

Table 3-2 lists the minimum infrastructure data requirements to allow the purposeful selection of the final set of columns for the systemic matrix. Not all matrix categories require the same information on infrastructure: quite simply, medians are not present at intersections, and the corresponding information irrelevant for an intersection systemic matrix, but still very relevant for a highway systemic matrix. Therefore, the table differentiates between location types.

Table 3-2. Minimum infrastructure data requirements for the bicycle systemic approach

Attributes	Label	Variables	Data type	Data values	Intersection	Highway
Lanes	I1	Number of through lanes – both directions – mainline	Numeric		x	x
Lanes	I2	Number of through lanes – both directions – cross-street	Numeric		x	
Lanes	I3	Number of left turn lanes – mainline	Numeric		x	
Lanes	I4	Number of right turn lanes – mainline	Numeric		x	
Lanes	I5	Number of left turn lanes – cross-street	Numeric		x	
Lanes	I6	Number of right turn lanes – cross-street	Numeric		x	
Median	I7	Presence and type of median	Categorical	Raised; striped; etc.		x
Speed	I8	Speed limit – mainline	Numeric or categorical	25 mph	x	x
Speed	I9	Speed limit – cross-street	Numeric or categorical	25 mph	x	
Traffic control	I10	Presence and type of intersection control	Categorical	No control; four-way stops; etc.	x	
Traffic counts	I11	Traffic volumes along mainline	Numeric		x	x
Traffic counts	I12	Traffic volumes along cross-street	Numeric		x	

It is important to note that these variables correspond to information needed to determine the systemic matrix structure and will not all be included as columns. Using the systemic approach for an extensive, diverse road network can lead to the temptation of trying to be too exhaustive when describing facility types in the systemic matrix columns. A roadway can be described with its number of lanes in each direction, the presence of traffic controls, its traffic volumes, the presence of a median, of a bike lane, its design speed, as well as many other attributes. And yet, the more roadway characteristics are included in the matrix columns, the more the matrix expands. This leads to a much larger number of cells, and therefore lowers significantly the "size" of systemic hotspots by spreading road crashes among a greater number of cells. As a consequence, interventions on the road network following the identification of relatively small systemic hotspots would have a smaller scale, thus reducing the impact of the systemic approach. It is therefore important to thoughtfully select the attributes that will be used to describe the roadway infrastructure.

Data Needs Specific to Exposure Indicators

Regardless of the structure of the crash and infrastructure data at hand, crash frequency constitutes the most straightforward safety indicator because it does not require the combination of infrastructure information – like the number of roads falling under a certain facility type, or the traffic volumes of a certain facility – to the raw number of collisions, as required crash density or crash exposure.

Populating the systemic matrix with crash exposure instead of crash frequency would require to divide the number of collisions that fall under a specific crash profile and happened in a specific facility type by the traffic levels experienced by that facility. There are some nuances in this calculation depending on which location category is considered. For intersections, there is the option to either take into account the traffic flow on the primary road only, or a combination of both the mainline and cross-street traffic flows. One important challenge arising from the calculation of exposure is that it imposes the inclusion of traffic counts as one of the column attributes – or else, facilities with different volumes would belong to the same facility and call for the use of volume averages, which would defeat the purpose of illustrating the singularities of facility types. Second, when it comes to bicycle-involved collisions, relying solely on car traffic counts only addresses part of the problem: a comprehensive approach would require taking into account jointly car traffic volumes and bicycle traffic volumes. Nevertheless, this ambition is thwarted by the challenge of having access to network-wide volume counts for bicycles.

When it comes to crash density, the matrix cells should contain the ratio between the number of collisions falling under the corresponding crash profile and facility type, and the number of corresponding facilities on the network being studied – in the case of intersections – or the total combined length of said facilities on the entire network – in the case of road segments. This implies that unlike frequency and exposure, the use of density rates does not allow direct comparisons between intersection and roadway segments, since the denominator is different, which can be an issue for an agency willing to consider the systemic safety challenges of its network as a whole.

Data Processing

Data collection was followed by thoughtful data processing, to both pair the two data sources and ensure the validity of the conclusions coming from the systemic matrix. Crash data was first filtered to retrieve only bicycle records, then the few duplicates of a same crash were removed, and the cleaned bicycle crash dataset was then matched with infrastructure data based on crash location (postmile), which resulted in one dataset with bicycle intersection crashes and the corresponding intersection features, and one dataset with bicycle highway crashes and the corresponding highway features. The matching was followed by the cleaning of the matched records. It included reformatting some variables, combining some into a single variable (e.g. mainline and cross-street number of lanes), adding labels, and creating ad hoc categories (e.g. individual traffic volumes into ADT categories). The rationale behind the creation of new categories will be detailed in the subsequent sections, and definitions for these new variables are listed in Appendix 1.

The resulting clean dataset is a table with a combination of crash and infrastructure data for the corresponding location type (intersection or highway), ready for use in the systemic matrix.

Crash Matrix Generation

Generating a systemic matrix for a given matrix category is a data-driven process, based on numerous successive trial-and-error iterations with the years of crash records available. It is important to note that a different dataset (even for the same road network but with different years of collisions) may lead to a different arbitration between options for selecting the matrices' rows and columns. Defining rows and columns therefore happens concurrently with populating the matrix. Identifying these rows and columns is equivalent to defining the crash profiles and creating the framework for the systemic matrix. As mentioned in the data collection section, this crucial step should consider the tradeoffs between the desire to include as many crash profiles as possible and the need for a compact and legible matrix structure. Judging whether a particular structure is fit for the dataset under study requires to find the just balance between the personal logic of the matrix developer and some objective measure of fitness. Adding a variable in the rows, deleting one in the columns, every single choice on the variables to be included affects not only the size of the matrix, but more importantly, the story told by the crash profiles they define, as will be detailed in the case studies to follow.

Rows and columns are defined separately, in no particular order: starting with one or the other does not affect the final matrix structure. The following will therefore detail each consecutively – starting with the definition of rows – for intersections and highways respectively, though they follow the same overall logic.

Determining the intersection systemic matrix structure

Defining Rows

In order to illustrate how crashes are influenced by the built environment, the rows of the matrix need to represent crash dynamics. These dynamics are specific to each party type involved, as crashes between a car and a cyclist show many dissimilarities with crashes between two cars. The row arrangement therefore was therefore tailored to bicycles for the present project.

Primary collision factor

As emphasized before, unlike mere crash counts for hotspots, a systemic matrix tells a story – a story about the entire road network. What the systemic approach intends to unveil is the underlying causes of typical collisions, so that their causes can be addressed in a comprehensive way and their future occurrence be prevented on all suspect road locations. The primary cause of a collision allows to "explain" its occurrence better/more concisely than a long combination of its individual characteristics, especially in the perspective of keeping the number of rows reasonable. However, with a database like TASAS-TSN, the values listed in the corresponding categories are too broad to be insightful, as shown on Figure 3-4.

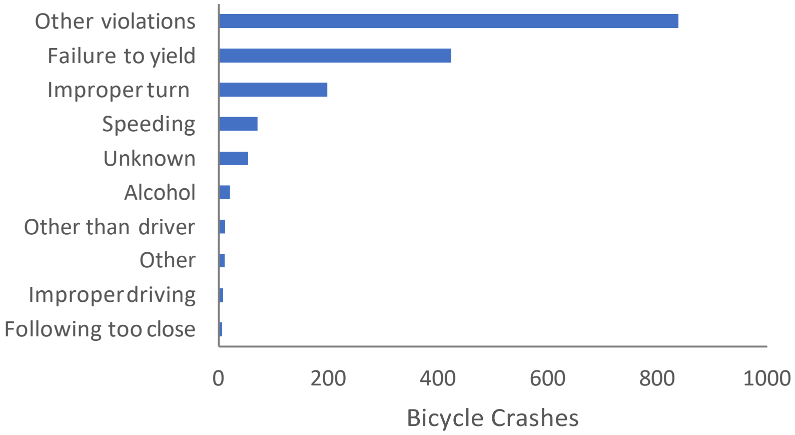


Figure 3-4. Bicycle intersection crashes by primary collision factor (2013-2017)

Crashes are therefore best described by the underlying violation code that were assigned to them by the reporting officer that registered the collision.

Violation category

As mentioned in the section on data requirements for a bicycle systemic matrix, TASAS-TSN data was supplemented with information from SWITRS to include the violation code for each collision. Using violation codes presents the opposite challenge of using primary collision factors: there are too many possible values. This is not desirable when developing a systemic matrix, as too much detail results in crash profiles being too narrow and limits the scope of the resulting safety improvements. One way of obtaining a compact matrix is to group variable values into larger overarching categories: for example, several violation codes may correspond to different instances of the same violation type, such as a failure to yield under various circumstances. One easy way of defining these groupings is to use sections of the California driving code, but ultimately, this new categorization is equivalent to manually redefining primary collision factors in a way that is deemed logical.

Because this process requires the review of individual violation codes and their manual categorization in a logical way, the abovementioned grouping cannot be exhaustive. It is therefore acceptable to focus on the most represented crash types. For the present project, the following rule was established: violation categories will be created to cover violation codes that account for at least 95% of all crashes (excluding crash records with missing violation codes) for that particular matrix (i.e. here, intersection crashes only), as long as each violation code accounts for at least 10 crashes. The remaining collisions are categorized into a default violation category named “Other”, as are records with missing violation codes. This rule is meant to reduce the burden for the analyst while being representative enough of the violations present in the dataset. It is up to the discretion of the analyst to include additional violation codes in the categories, although there is limited interest in expanding them since violation codes that are not included only cover very few crashes in the crash dataset under consideration. In the case of the present project, the research team decided to include additional codes that had been used for a systemic safety analysis for a local agency in California in addition to the top 95% violation codes. Table 0-1 in Appendix 1 shows the violation categories that resulted from this process for 2013-2017 TASAS-TSN and SWITRS data.

Note that in order to allow for meaningful comparisons when trying to describe the dynamics behind any particular hotspot, it is recommended to homogenize the resulting violation crash types across matrix categories (i.e. in the present case, for the intersection crash matrix and the highway one). When homogenizing the violation categories across matrix categories, no more than 5% of crashes with a non-blank violation code should fall within the "Other" violation category for each matrix category. The distribution of intersection crashes across the new violation categories is shown in Figure 3-5 and gives a better breakdown of crashes than when using the TASAS-TSN primary collision factor. The fact that the most represented category is "Other" is due to missing violation codes, but the breakdown of the subsequent prevailing categories is satisfactory.

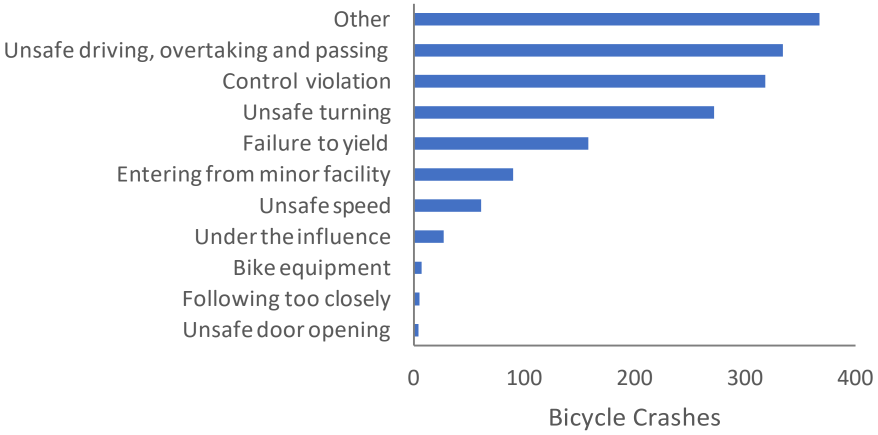


Figure 3-5. Bicycle intersection crashes by violation category (2013-2017)

Bicycle movement type

The TASAS-TSN database provides party-level information for each collision, which means that several records correspond to a single crash. However, it does not specify which party was the one at fault, which limits how the level of information provided by the variable on movement types, considering that the analyst cannot infer which record to prioritize for a given crash. For the present project, movement types were looked at for records that corresponded to bicycles, considering that bicycle safety was the focus of the analysis. For a given collision, only bicycle records were retrieved. Whenever there were several bicycles involved, only one record was kept in order to avoid double-counting a single crash. This led the research team to leave out less than 20 bicycle records, which is minimal in comparison with the over 1,600 records for 2013-2017 TASAS-TSN data. Bicycle movements for these crashes are displayed on Figure 3-6. It appears that the vast majority of cyclists involved in collisions at intersections on the California state highway system were proceeding straight. Such an imbalanced distribution is not desirable for a systemic matrix, as the excessive way of the first category would lead to too many crashes being captured in one row, which results in safety resources being devoted to tackle a problem that is too large.

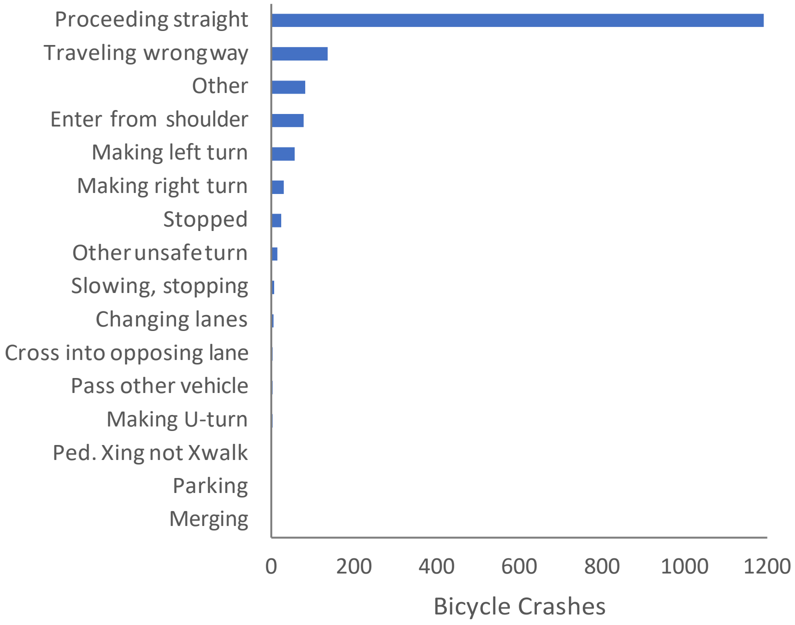


Figure 3-6. Bicycle intersection crashes by movement type (2013-2017)

Collision type

The collision type describes the impact between parties involved in a crash and whether it was broadside, rear-end, sideswipe, etc. Similar to the type of movement, in the case of bicycle intersection SHS crashes, it appeared that there was an imbalance between different categories (See Figure 3-7), which indicates that this variable is not the most ideal descriptor of crash dynamics. Broadside crashes are the most represented, followed by "other" types of collision.

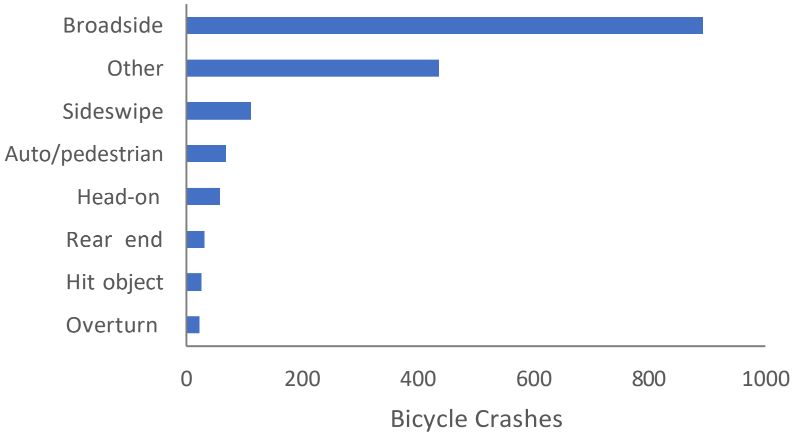


Figure 3-7. Bicycle intersection crashes by collision type (2013-2017)

Party at fault

For bicycle-involved collisions, the story behind a particular crash changes dramatically depending on which party is at fault, because cyclists are more vulnerable than motorized vehicles. The safety measures taken from identifying the party at fault will also be different: different countermeasures will be taken if in a crash with a failure to yield, the bike was at fault, or if it was the car, regardless of the location. In the case of 2013-2017 bicycle intersection crashes, bicycles were at fault in most cases (See Figure 3-8).

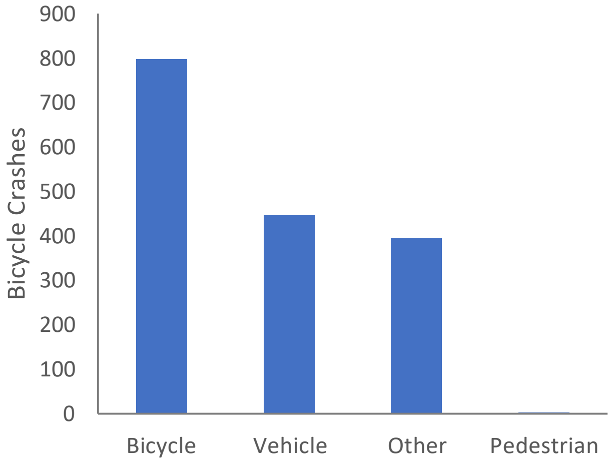


Figure 3-8. Bicycle intersection crashes by party at fault (2013-2017)

Combining row levels

Having information from only one of the abovementioned variables is not enough to create meaningful crash profiles: for instance, knowing that bicycle intersection crashes where the bicycle was at fault happened at intersections with high traffic volumes is not informative enough to determine what engineering countermeasures should be applied to the corresponding systemic hotspot. This is the reason why combinations of variables are considered after having explored individual distributions.

Several combinations resulted in the top categories not being insightful because they included "other" values. This was the case when combining primary collision factor and collision type or violation category with collision type. On the other hand, combining violation category and party at fault proved interesting, as illustrated by Figure 3-9.

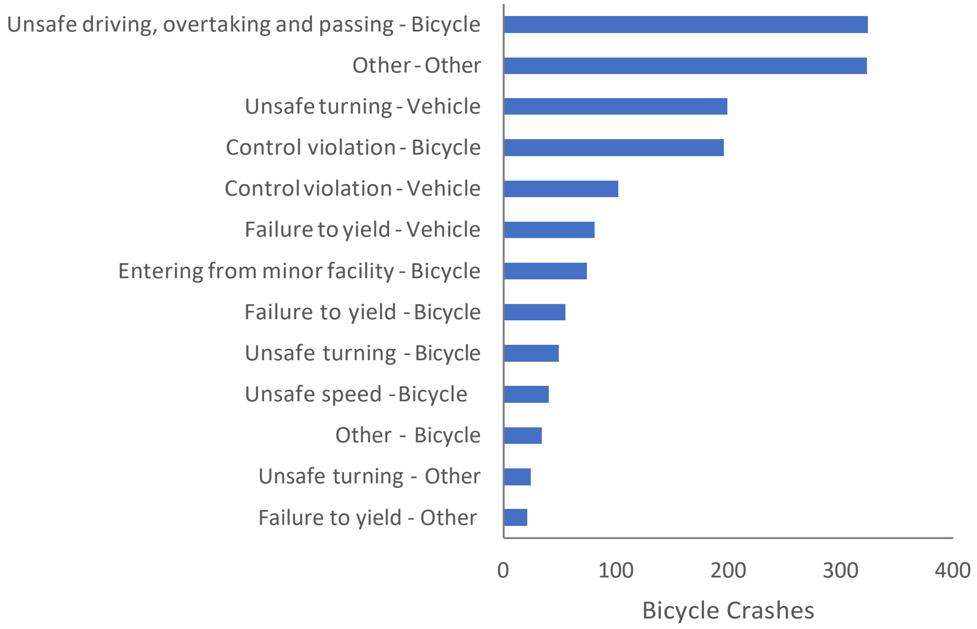


Figure 3-9. Bicycle intersection crashes by violation category and party at fault (2013-2017)

Considering that the top violation was still flagged "other", a third variable was added to break it down more, which resulted in a final row structure made of the following variables:

(i) violation category
(ii) party at fault
(iii) collision type.

Figure 3-10 shows the frequency of crashes by combined category, where unsafe driving, overtaking and passing, at-fault bicycle, and broadside is the most common type.

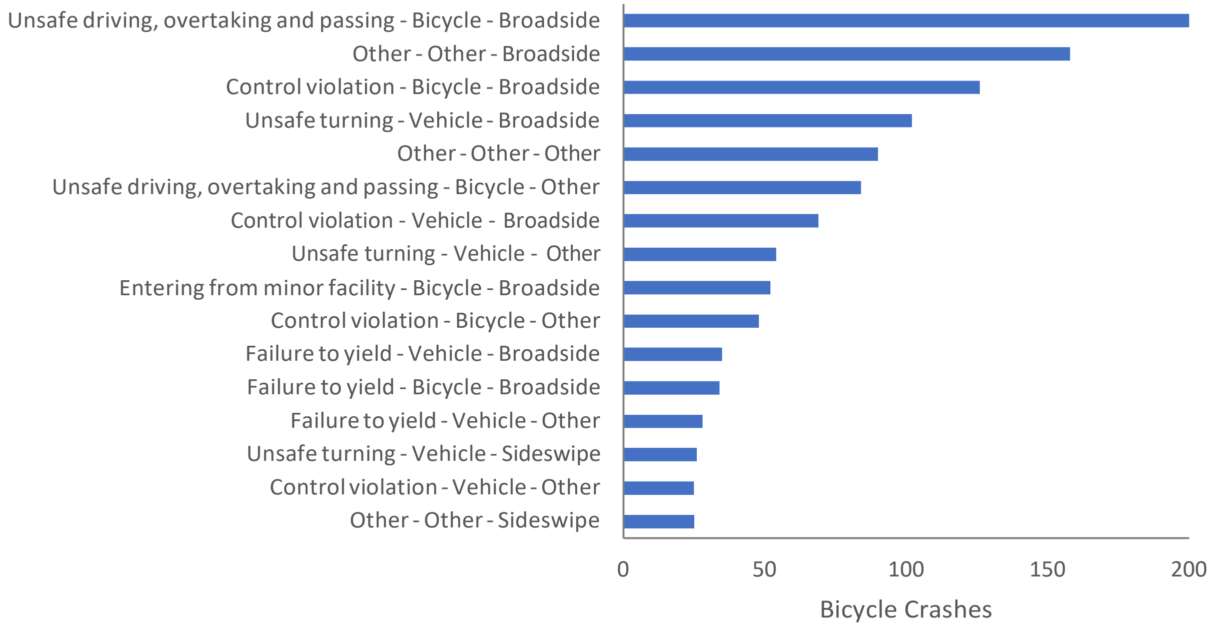


Figure 3-10. Bicycle intersection crashes by violation category, party at fault and collision type (2013-2017)

As will be discussed later, this row structure was enough to flag systemic hotspots once the column structure was finalized, and it kept the row size within reasonable bounds.

The research team did not include as rows other variables that are important but relate more to the overall context or the policy focus and should therefore be used as filters for the entire matrix. This includes considerations on the severity of the crash (e.g. to focus on fatal and severe injury crashes only). Similar decisions were made for very broad infrastructure information regarding the type of area (e.g. to focus on urban crashes), roadway class (e.g. to focus on freeways), district, etc.

Defining columns

Traffic control

It is common to implement the systemic approach at intersections by differentiating between types of signalization, if any. Many variations exist in terms of signalization, but several bear resemblance and therefore call for groupings. The research team created new categories listed in Appendix 1 that resulted in differentiating between intersections that had no traffic control, traffic signals, four-way stops, two-way stops, yield signs or something else. The large imbalance between categories of traffic control shown on Figure 3-11 indicates the need to add more detail to characteristics of intersections with traffic signals or with two-way stop signs. It is also interesting to note the total absence of bicycle crashes at intersections with yield signs in 2013-2017.

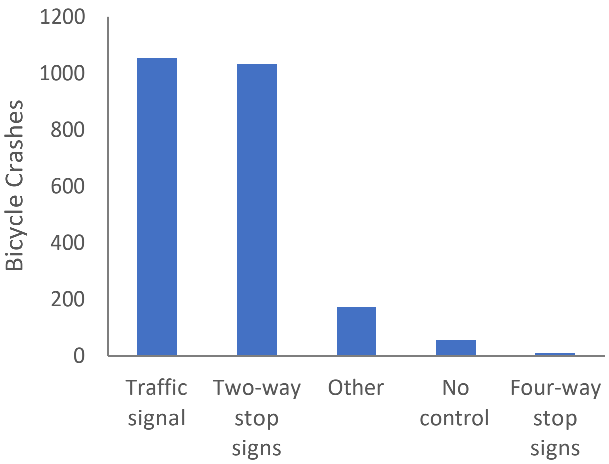


Figure 3-11. Bicycle intersection crashes by type of traffic control (2013-2017)

Number of lanes

The number of lanes at an intersection is a good indication of the geometrical attributes of an intersection. It also gives an indication of the distance that bicycles need to cross to get to the other side. Considering the number of lanes on the mainline road as well as the cross-street is considered best practice, as it is not possible to know which leg of the intersection each party was coming from without looking at the police crash reports one by one. However, a major downside of using the number of lanes as is in the systemic approach is that it results in a large number of categories, thus expanding significantly the number of columns – and not always in a meaningful way: there is little difference between a 7+2 and a 8+2 intersection, and yet, each would get a separate column, thus complicating the inference of appropriate engineering countermeasures. Figure 3-12 shows the number of intersection bicycle crashes by number of lanes on the mainline and cross-street.

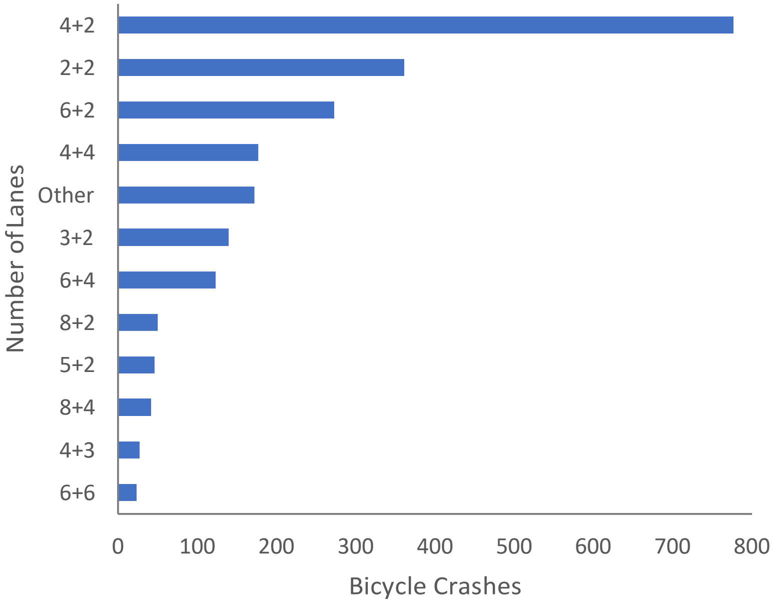


Figure 3-12. Bicycle intersection crashes by number of lanes (mainline and cross-street) (2013-2017)

Vehicular traffic volumes

Including vehicular traffic volumes in a systemic matrix for bike safety is challenging because as mentioned earlier, it only partly captures the use of infrastructure and does not describe the intensity of use by cyclists. In the case of bicycle safety, considerations about traffic volumes are only relevant when deciding between bicycle lane classes, and as such, do not apply to intersections. This information was therefore not included in the systemic intersection matrix.

Turn channelization

The prevalence of unsafe turns in the row categorization described above makes understanding what turning movements are allowed at intersections that had crashes critical for the analysis. This is why turn channelization was considered. The TASAS-TSN database includes detailed information on left and right-turn channelization respectively, for the main road and for the cross-street. For compactness purposes, different types of channelization where grouped (see Appendix 1) to only describe the presence or absence of a left or right turn channelization, for the mainline or cross-street respectively. Considering the importance of permissive left turns in transportation planning in California, only left turn channelization was considered. Nearly two thirds of bicycle intersection crashes occurred at intersections where left turns were channelized (See Figure 3-13 and Figure 3-14).

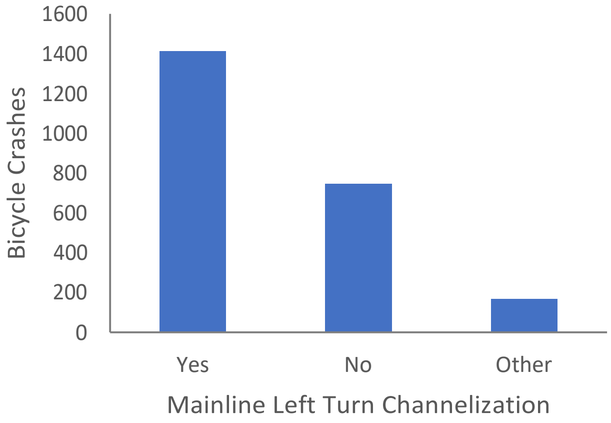


Figure 3-13. Bicycle intersection crashes by left turn channelization on mainline (2013-2017)

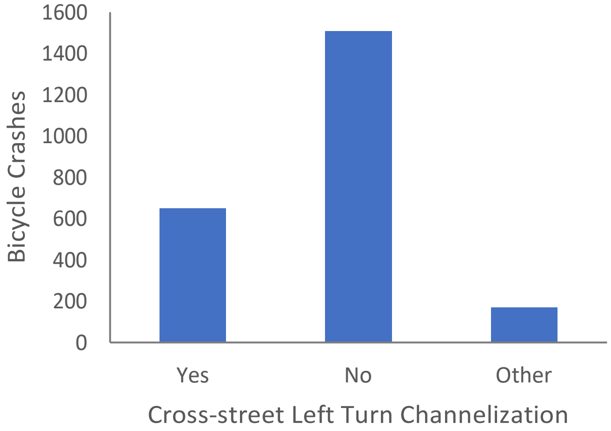


Figure 3-14. Bicycle intersection crashes by left turn channelization on cross-street (2013-2017)

Combining column levels

Based on the previous sections, the most promising column categories to describe intersections on the California State Highway System that had bicycle crashes between 2013 and 2017 are the type of traffic control and left turn channelization (mainline and cross-street). Figure 3-15 displays the distribution of crashes when combining them. The result is a well-balanced distribution that describes well different types of facilities in a compact way.

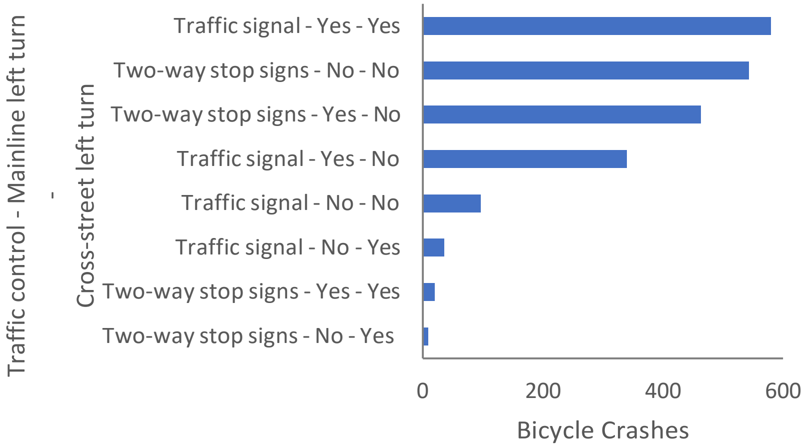


Figure 3-15. Bicycle intersection crashes by type of traffic control and left turn channelization (mainline & cross-street) (2013-2017)

Determining the highway systemic matrix structure

Defining rows

Similar to the intersection systemic matrix, historical crash data from 2013-2017 was used to determine the most appropriate variables that should be used as row in the highway systemic matrix.

Primary collision factor

The primary collision factor presented the same issues of representativeness as explained in the previous section: the values listed in the corresponding categories are too broad to be insightful, as shown on Figure 3-16.

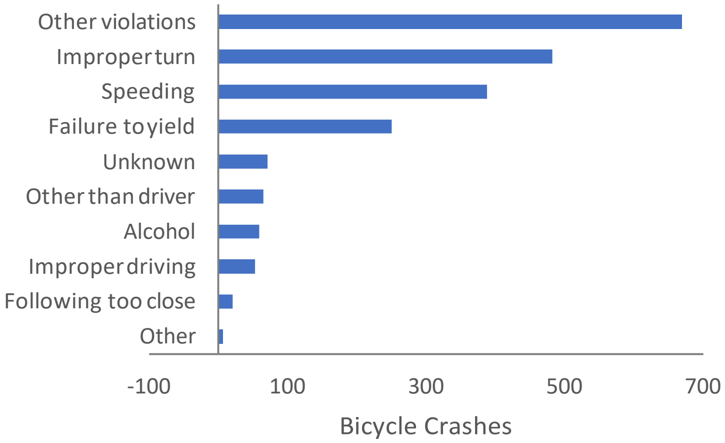


Figure 3-16. Bicycle highway crashes by primary collision factor (2013-2017)

Crashes are therefore best described by the underlying violation code that were assigned to them by the reporting officer that registered the collision.

Violation category

Following the same logic as previously, violation categories were created based on the most represented violation codes. This resulted in the distribution of intersection crashes shown in Figure 3-17, which gives a better breakdown of crashes than when using the TASAS-TSN primary collision factor. The fact that the most represented category is “Other” is due to missing violation codes, but the breakdown of the subsequent prevailing categories is satisfactory.

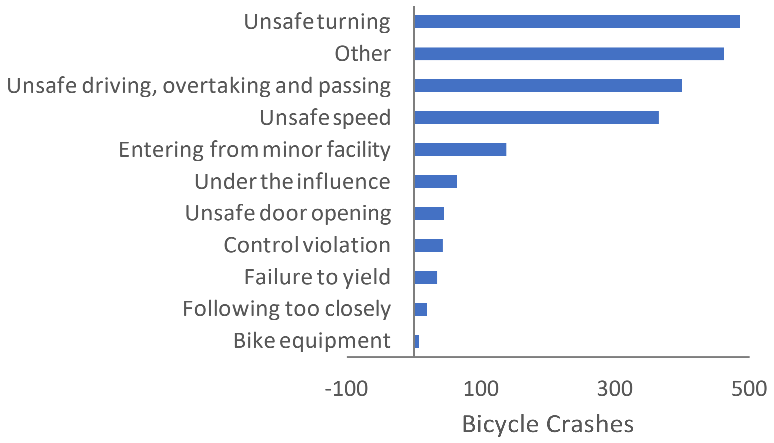


Figure 3-17. Bicycle highway crashes by violation category (2013-2017)

Bicycle movement type

The vast majority of cyclists involved in collisions on road segments were proceeding straight – which is logical for a road segment, unless one of the parties was entering from a minor facility. As a consequence, the type of bicycle movement was not included in the highway matrix.

Collision type

Unlike for intersection crashes, the types of collision for highway crashes were well balanced. Broadside crashes were still among the most represented, although “other” types of collision were most represented – most likely because of missing data (See Figure 3-18).

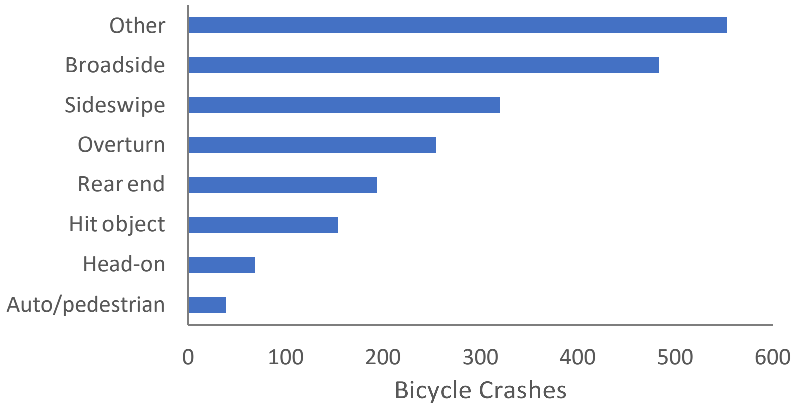


Figure 3-18. Bicycle highway crashes by collision type (2013-2017)

Party at fault

As explained in the previous section, for bicycle-involved collisions, the story behind a particular crash changes dramatically depending on which party is at fault, because cyclists are more vulnerable than motorized vehicles. In the case of 2013-2017 bicycle intersection crashes, bicycles were at fault in most cases (See Figure 3-19).

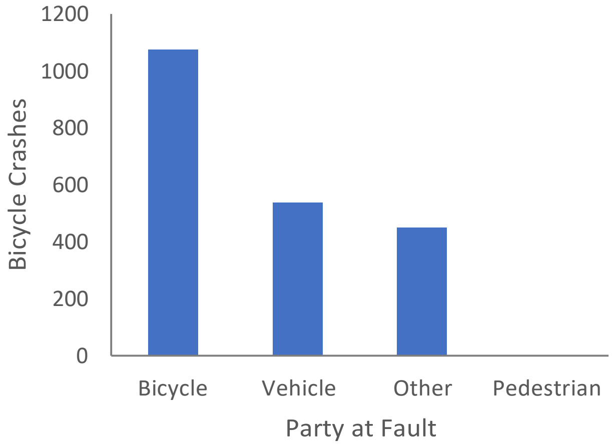


Figure 3-19. Bicycle highway crashes by party at fault (2013-2017)

Combining row levels

As for intersection crashes, combining violation category and party at fault and collision type allowed to obtain an interesting, well-balanced breakdown for highway crashes, as illustrated by Figure 3-20. This was selected as the final row structure for highway matrices.

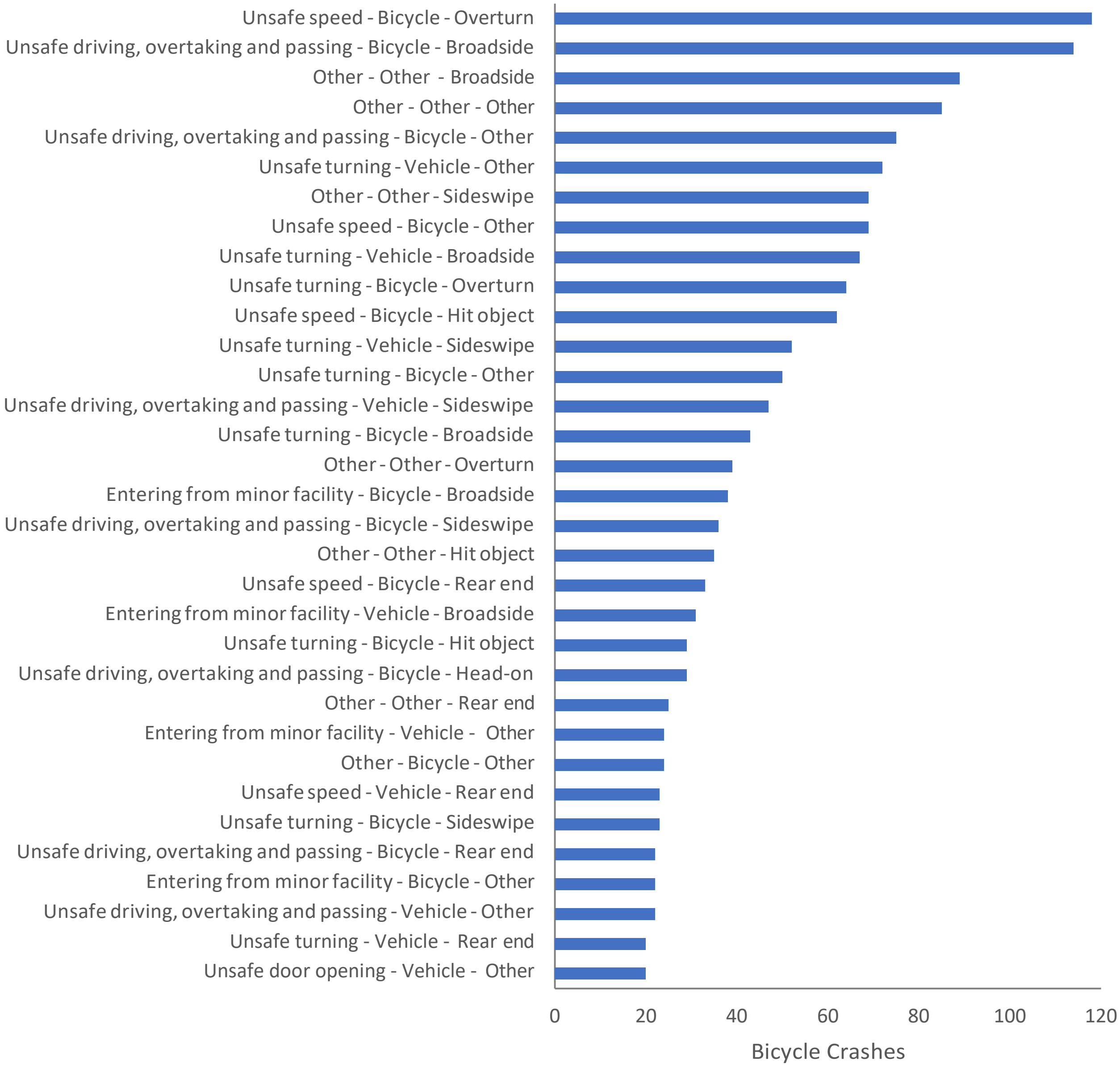


Figure 3-20. Bicycle highway crashes by violation category, party at fault and collision type (2013-2017)

Defining columns

Median

The presence of a physical delimitation between opposite flows of traffic can prevent the occurrence of some types of crashes, like head-on collisions. However, medians can be designed in various ways, which makes the presence of a median and the type of barrier important characteristics to study in the systemic approach. Only 4 SHS bicycle crashes happened on undivided highways between 2013 and 2017, against 121 on divided highways with a raised median, and more than 1,400 on divided highways without a physical barrier (See Figure 3-21). This great imbalance may be related to the overall makeup of the SHS network – which could be explored as part of exposure calculations.

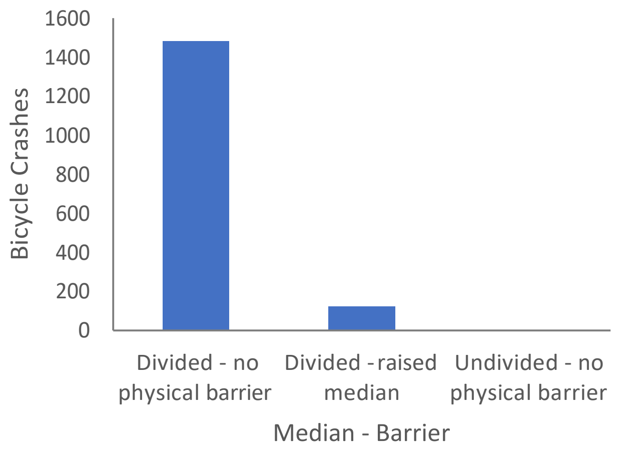


Figure 3-21. Bicycle highway crashes by median and barrier type (2013-2017)

Number of lanes

For highway segments, the number of lanes gives a slightly different information than for intersections. While it also describes the geometry of a highway, it mainly suggests the potential intensity of vehicular traffic happening to the left of bicycles riding on that highway, but does not relate to a distance that bikes would have to cross. Again, a major downside of using the number of lanes as is in the systemic approach is that it results in a large number of categories. It appears that most crashes happened on narrow, symmetrical highways, with one or two lanes in each direction (See Figure 3-22).

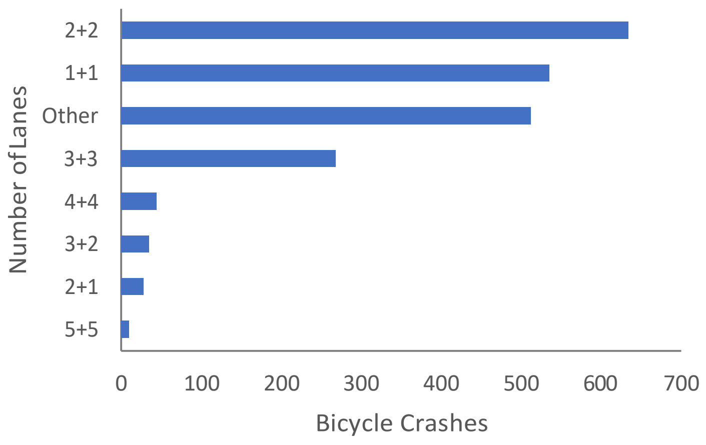


Figure 3-22. Bicycle highway crashes by number of lanes (left & right) (2013-2017)

When combining the number of lanes with information on median and barrier type, it results in Figure 3-23, which still appears quite imbalanced since the top two categories account for about three quarters of highway crashes, which calls for the addition of another variable.

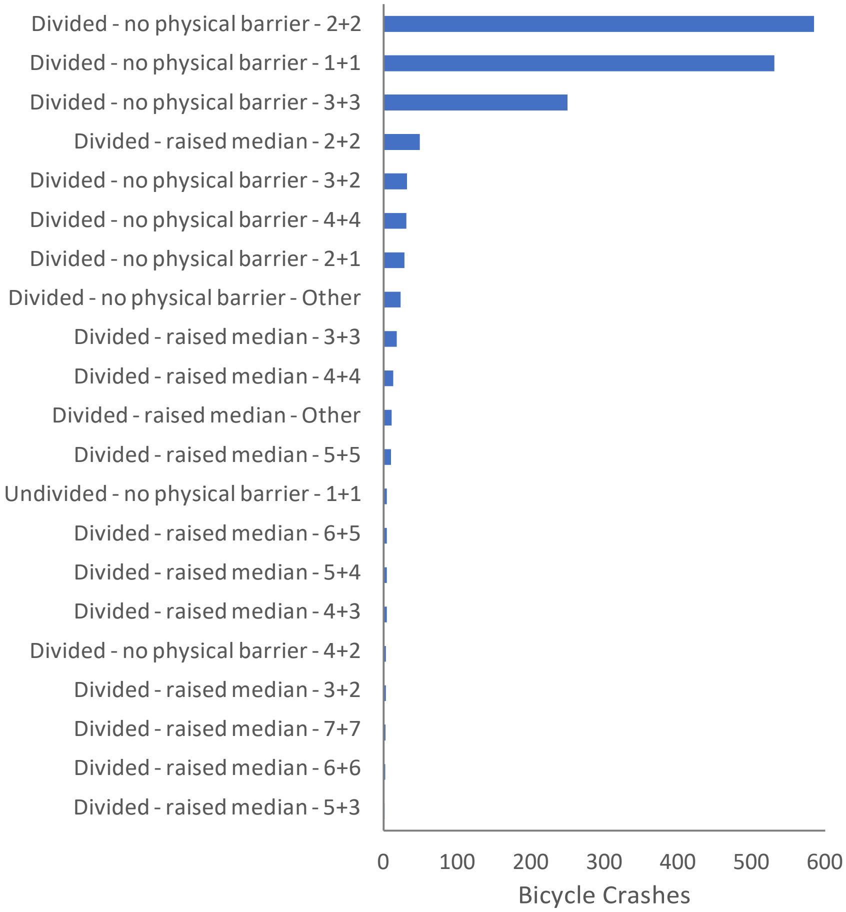


Figure 3-23. Bicycle highway crashes by median type and number of lanes (left & right) (2013-2017)

Vehicular traffic volumes

As explained for the intersection matrix, including vehicular traffic volumes in a systemic matrix for bike safety is challenging because it does not describe the intensity of roadway use by cyclists. However, it remains a common descriptor of highway segments, especially when deciding between bike lane classes. The difficulty with numerical variables like ADT (average daily traffic) is to choose the appropriate categorization that will yield a balanced yet meaningful breakdown. Using the median as the cutoff is not an option because too many bicycle crashes happened on very low traffic highways, and considering roads with volumes below 1000 ADT separately would not be insightful regarding the appropriate engineering countermeasures. The research team therefore used a breakdown that had been adopted for the systemic approach for a local agency in California, with ADT thresholds at 7,000 – 15,000 – 25,000. When combined with the type of median and the number of lanes, it resulted in a balanced, acceptable breakdown of crashes (See Figure 3-24).

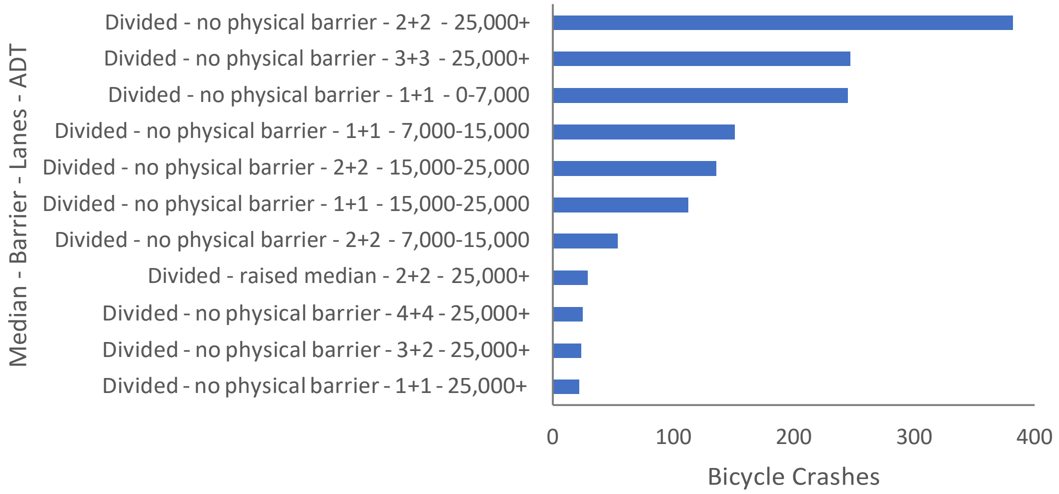


Figure 3-24. Bicycle highway crashes by median type, number of lanes (left & right) and vehicular traffic volumes (2013-2017)

Validating crash matrix structures

The final structure for the systemic intersection matrix is as follows:

- Rows: descriptors of crash dynamics:
  - Violation category (categorical)
  - Party at fault (categorical)
  - Collision type (categorical)
- Columns: descriptors of infrastructure features:
  - Traffic control type (categorical)
  - Mainline left turn channelization (dummy)
  - Cross-street left turn channelization (dummy)

The final structure for the systemic highway matrix is as follows:

- Rows: descriptors of crash dynamics:
  - Violation category (categorical)
  - Party at fault (categorical)
  - Bicycle movement type (categorical)
- Columns: descriptors of infrastructure features:
  - Median presence (dummy)
  - Median barrier type (categorical)
  - Number of lanes – left & right (categorical)
  - Vehicular traffic volume – ADT (categorical)

The structures outlined above were chosen after successive iterations, based on the five years of crash records available. They were selected because they were best aligned with the goals of the present project and best suited to enable the actual implementation of engineering countermeasures on the identified systemic hotspots. And yet, it is important to keep in mind that the systemic approach is a data-driven one. Therefore, while the previous section outlined decisions mostly based on professional judgment and expertise, the present section will detail some tools that can support a data-driven decision-making process when looking for an acceptable systemic matrix structure.

First, regardless of its size, it is useful to keep in mind the possible extreme forms that a systemic matrix can take. On one hand, it can be almost empty, with all crashes being clustered in a single hotspot, i.e. a unique pair of collision characteristics and infrastructure characteristics. On the other hand, it can be almost full, with each crash profile having occurred at least once in the period of study. Both cases are to be avoided, as they prevent the identification of an appropriate number of systemic hotspots. Having too few means that a couple crash profiles with high numbers of collisions will be the only systemic hotspots. In that case, the crash profile appears to be too general, meaningless, and defining the numerous corresponding countermeasures will result too burdensome and expensive. Conversely, having too many systemic hotspots, that is, if the matrix is too balanced between crash profiles, too many cells will have similar numbers of crashes, meaning that too many infrastructure types will require engineering countermeasures, which is not realistic financially. In summary, finding the appropriate balance in terms of matrix structure is key to the success of the implementation of the systemic approach, which culminates with countermeasures and ought to be feasible.

What this strive implies, is that ultimately, the share of empty cells in a systemic matrix does not matter. It is only important that it does not fall into one extreme or the other. As shown in Table 3-3, the actual share of blank cells for the 2013-2017 matrices for California described in the case study stand relatively high, above 80%. As emphasized throughout this report, the focus of the systemic matrix is cells with the highest weights in terms of numbers of collisions, not crash profiles with only few occurrences. This is why the kurtosis of the matrix constitutes an interesting index to aid choice-making between different matrix structures. Indeed, the kurtosis indicates the extent to which a distribution is peaked or flat. Knowing that a normal distribution has a kurtosis of 3, matrices with an overall kurtosis superior to 3 have systemic hotspots that significantly stand out. Conversely, a kurtosis below 3 would indicate a less than ideal matrix structure, where collisions are too evenly distributed. Though this index validates in an absolute manner the structure of a systemic matrix, since kurtosis is non-linear, it does not allow to directly measure the advantages of switching from one structure to the other. Doubling the kurtosis does not divide by two the number of systemic hotspots, or increase their weight by two. In short, the kurtosis only indicates how much a particular distribution of crash profiles peaks; it only says whether the matrix structure is acceptable as is, not whether it is inherently better than another “acceptable” matrix structure. As illustrated by Table 3-3, it is the case for the highway and intersection matrices.

The emergence of systemic hotspots is best enabled by the presence of peaks in rows and columns as well. This is indicated by the row and column totals and allows to use their respective kurtosis as informative indicators as well. Logically, the row kurtosis of an almost empty matrix will be greater than 3: since almost all of its crashes will be contained in a single hotspot, the row on which this hotspot is located will also contain the wide majority of the crashes. The same goes for the columns’ kurtosis. On the contrary, an evenly-distributed matrix will have balanced rows and columns, respectively, and therefore both kurtoses will stand below 3. When using these two indices, it looks like the two final matrices have a moderately acceptable structure. However, based on the previous sections, the chosen variables were the best options based on data availability and road safety expertise.

Another useful and straightforward index is the size of the matrix. It allows to assess the legibility of the matrix, which shall remain in reasonable proportions to allow better navigation between crash profiles. However, the compactness of a systemic matrix is very relative, since it depends tremendously on the number of variables included in the structure, and the number of categories that these include. Sometimes, variables are just binary (e.g. median presence), while others need to be broken down in many categories to be comprehensive (e.g. types of violations).

Finally, the last quantitative index is the ratio of the cell maximum (i.e. the number of collisions pertaining to the first systemic hotspot) to the 95th percentile. This index goes beyond the kurtosis in that it not only indicates how acute the peak of the distribution is, but it also allows more fathomable comparisons between two distributions, since it is a percentage. Additionally, the higher the ratio, the easier it is to set the threshold for hotspot identification, since the top five percent of the crash profiles stand out so much.

Table 3-3. Summary indices on various systemic matrix structures.

Matrix	Syst_H	Syst_I	Blank	Even	Random
Share of blank cells	91%	83%	99%	20%	49%
Overall kurtosis	46.23	20.31	17.00	-2.00	0.84
Column kurtosis	6.52	0.40	17.00	1.15	1.67
Row kurtosis	-1.20	20.32	175.00	-0.05	-0.58
Max to 95th pctile	482%	209%	400%	0%	62%
Number of rows	183	175	175	175	175
Number of columns	34	17	17	17	17
Table size	6222	2975	2975	2975	2975
Mean 95th pctile	38	45	1394	1	28

Using the abovementioned indices does not disqualify the case-by-case decision-making detailed previously. The two processes are complementary and should both be applied when considering how many variables should be included in the systemic matrix, which ones, and in what order. As emphasized before, these structures are by no means the only valid ones. They only correspond to "acceptable", meaningful matrix structures that fit the data for 2013-2017 bicycle crashes well and told a story about systemic safety hazards on the Californian state highway system for these years. Ultimately, the systemic matrix is nothing but a decision-making tool to inform agencies about the flaws of their road network and the potential improvements they could make in order to improve safety outcomes for some subsets of the population (drivers, bicyclists and/or pedestrians). This thoughtful approach allows the emergence of systemic hotspots, which then call for another decision to be made: how should systemic hotspots be defined? Should there be a fixed cut-off number of collisions? The next section will detail how to identify systemic hotspots once the structure of the systemic matrix has been finalized.

Screening systemic hotspots

Developing the quantitative criteria for systemic hotspots

Once the systemic crash matrices are generated, systemic hotspots are identified, using quantiles. For the vehicle matrices the criteria for high-priority systemic hotspots is matrix cells with crashes counts that are above the 99% percentile. For both matrices, the percentile was rounded down to prevent situations where a systemic hotspot was missed due to a fraction of a crash.

High-priority systemic hotspots are defined as matrix cells that require attention from the agency managing the road network, and represent the primary output of the systemic matrices. Considering the material constraints experienced by public agencies in charge of road infrastructure, there is the need for a metric that can efficiently and reliably alert these agencies of systemic safety concerns. If the metric is too restrictive, it can miss valuable safety-improving opportunities. However, if the metric is too inclusive, it can reduce the agency's ability to respond effectively. In light of this, the trade-off between the desire to have an inclusive list and the efficiencies of a restrictive list was taken into account by the research team.

Descriptive statistics and data visualizations were used to assess several approaches for criteria-setting. This includes average-based confidence intervals, signal-to-noise ratios, triangular distributions, and quantiles. The quantile method was determined to provide the best fit across the different types of matrices. This is partly driven by the empirical distributions across matrices. More specifically, although the data in each of the matrices is zero-inflated the behavior at the upper extremities varied quite a bit and a quantile-based method provided the most transparent and consistent outcome.

To determine the actual criteria, the data for each matrix was broken down to 1000 quantiles. The data was then plotted and reviewed to look for change-points. Figure 3-25 shows a graph of the quantiles for the highway and intersection matrices. The quantiles go from left to right. At the far left is the 0.001 percentile, and at the far right is the 99.9 percentile. Figure 3-26 is a zoomed in graph of the upper part of the data for each matrix, between the 90th and 99.9th percentile. Using 1000 quantiles allows to identify a criterion that is at a lower resolution than the plotted quantiles. After reviewing the data, a threshold of 99th percentile is indicated as a reasonable threshold for the two matrices. While the inflexion point is clear for the highway matrix, the plot for intersection crashes did not demonstrate a clear inflexion point, at a resolution of 0.01 percentile, a criterion of 90th percentile emerged as a reasonable threshold (See Figure 3-27). This threshold was deemed too low, as it would have resulted in systemic hotspots of as little as 5 crashes for the highway matrix. This is the reason why the research team decided on a threshold of 99th percentile.

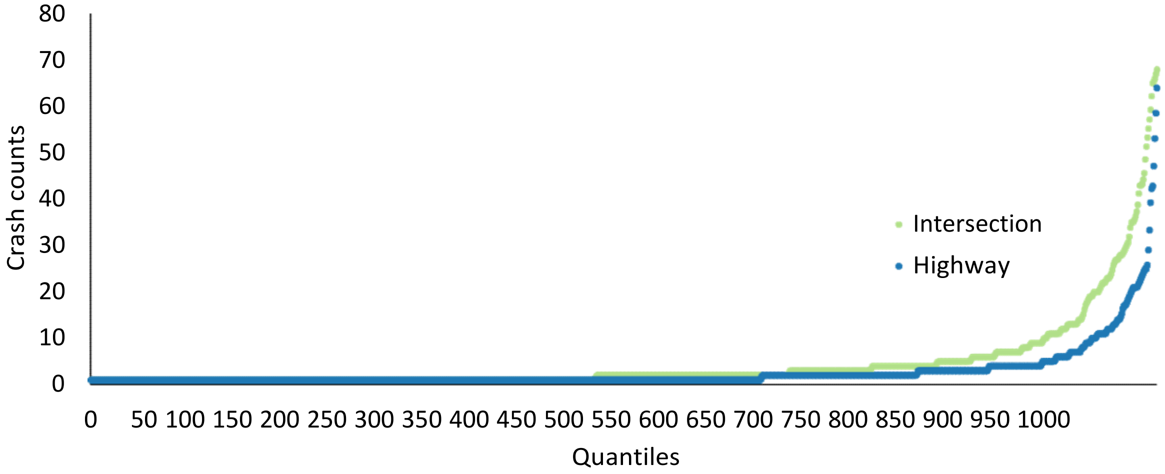


Figure 3-25. Bicycle highway and intersection crash quantiles (0.001 percentile)

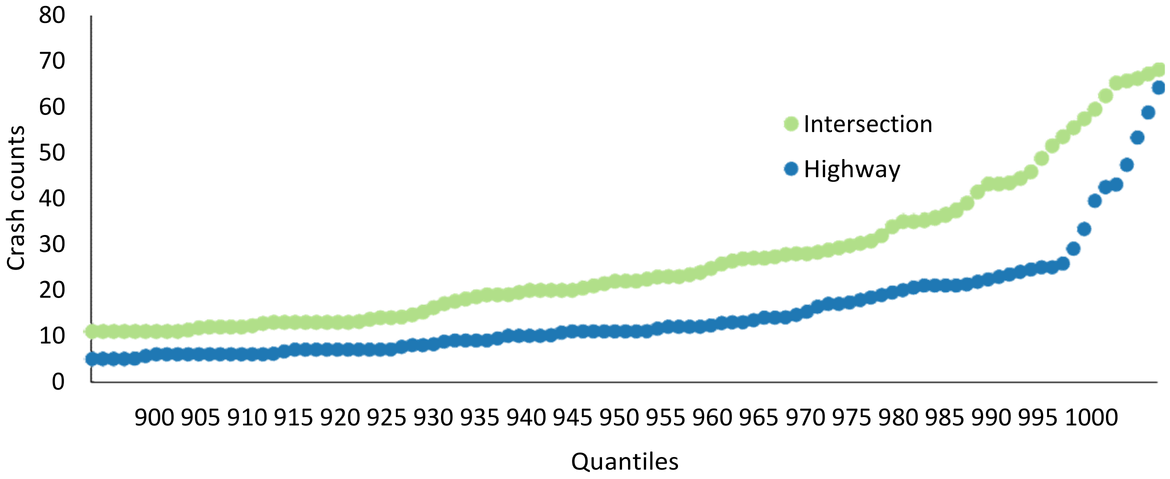


Figure 3-26. Bicycle highway and intersection crash quantiles (0.001 percentile) on zoomed axis

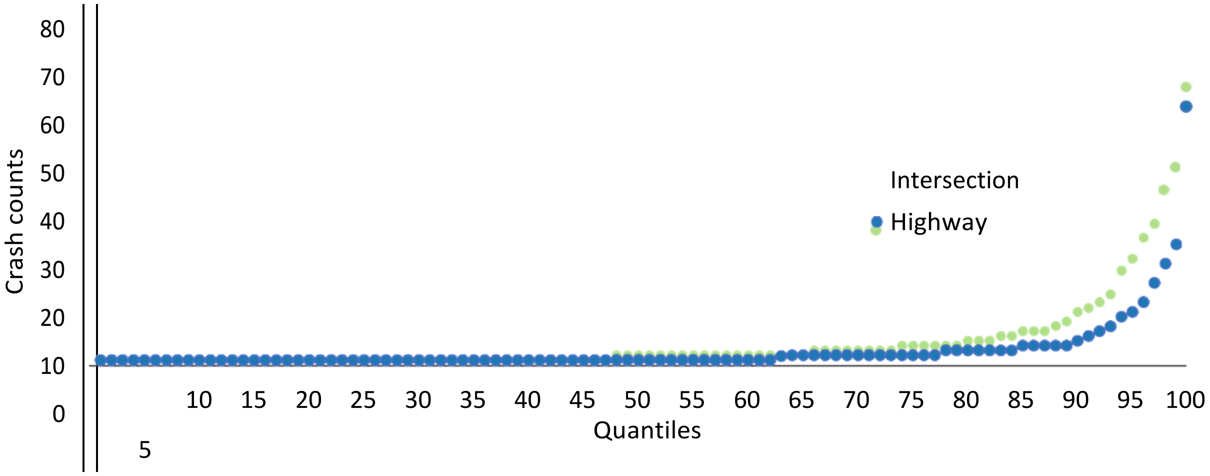


Figure 3-27. Bicycle highway and intersection crash quantiles (0.01 percentile)

Prioritizing among systemic hotspots

Once the high-priority crash profiles are identified, they are considered systemic hotspots and are labeled as first, second, and third priority. The priorities are determined by the order of each systemic hotspot within a column: the top systemic hotspot in a column is labeled as first priority, the second highest is labeled as second priority, and any additional systemic hotspots are labeled as third priority. The final ranking of the systemic hotspots is by descending order (in the number of crashes) of all of the first-priority systemic hotspots, followed by descending order of all of the second-order priority ones, and ends with all of the third-priority hotspots, by descending order. Finally, the total number of crashes for an entire facility type (or column) are used as tie-breakers between hotspots with the same crash counts: the hotspot corresponding to the "most dangerous" facility type will come first. If the tied crash profiles pertain to the same facility type, then it is the row totals that help decide between them.

Table 3-4. Bicycle intersection collisions: (systemic hotspot threshold of 51 crashes for the 99th percentile).
Row Structure: Violation category – Party at fault – Collision type
Column structure: Traffic control type – Mainline left turn channelization – Cross-street left turn channelization

Crashes	Crash type	Roadway type	Priority	Tie-breakers
68	Control violation | Bicycle | Broadside	Traffic signal | Yes | Yes	1	n/a
66	Unsafe driving, over taking and passing | Bicycle | Broadside	Two-way stop signs | Yes | No	1	n/a
55	Unsafe driving, over taking and passing | Bicycle | Broadside	Two-Way stop signs | No | Np	1	n/a
65	Unsafe driving, over taking and passing | Bicycle | Broadside	Traffic signal | Yes | Yes	2	n/a
51	Unsafe turning | Vehicle | Broadside	Two-way stop signs | Yes | No	2	n/a
59	Other | Other | Broadside	Traffic signal | Yes | Yes	3	n/a

Table 3-5. Bicycle highway collisions: (systemic hotspot threshold of 25 crashes for the 99th percentile).
Row Structure: Violation category – Party at fault – Bicycle movement type
Column structure: Median presence – Median barrier type – Number of lanes (left & right) – ADT

Crashes	Crash type	Roadway type	Priority	Tie-breakers
64	Other | Other | Proceeding straight	Divided | No physical barrier | 2+2 | 25,000+	1	n/a
54	Unsafe speed | Bicycle | Proceeding straight	Divided | No physical barrier | 1+1 | 0-7,000	1	n/a
42	Unsafe driving, overtaking, passing | Bicycle | Proceeding straight	Divided | No physical barrier | 3+3 | 25,000+	1	n/a
25	Unsafe turning | Vehicle | Proceeding straight	Divided | No physical barrier | 1+1 | 7,000-15,000	1	n/a
43	Unsafe turning | Vehicle | Proceeding straight	Divided | No physical barrier | 2+2 | 25,000+	2	n/a
31	Other | Other | Proceeding straight	Divided | No physical barrier | 3+3 | 25,000+	2	n/a
25	Unsafe turning | Vehicle | Proceeding straight	Divided | No physical barrier | 1+1 | 0-7,000	2	n/a

This logic allows a first-level systemic hotspot to be ranked above a second-level hotspot with more crashes, and is established to provide more opportunities to develop systemic improvements across multiple facility types. Indeed, a first-level hotspot will be the first type of facility that will benefit from engineering countermeasures to prevent the occurrence of certain crash types. By treating this facility type, it is possible that some co-benefits will result and also reduce the occurrence of a difference type of crashes for this facility (i.e. lower the number of crashes in different cells of the same column). Treating it before other columns therefore allows to solve systemic concerns across the road network in a more comprehensive way than focusing too much on a specific facility type.

Chapter 4. Bicycle Exposure Modeling Approach

Background

This section summarizes existing research on bicycle volume models, highlighting variables that can potentially be used to estimate bicycle volumes at specific locations on the California State Highway System. The estimated bicycle volumes can ultimately be added to the Caltrans transportation system information database for planning, design, and safety analysis purposes. Existing literature includes bicycle volume studies from California as well as other parts of North America.

Bicycle volume data are important for safety analysis because they can be used as a basic measure of exposure at a specific location. For example, the relative risk of bicycle crashes for people traveling along state highways can be estimated as the number of bicycle crashes per million bicycles. Further, using bicycle volume as a variable in safety performance functions can show which roadway design features or other characteristics of a location should be modified to reduce bicycle crashes and injuries. Volume data can also be used to identify how common bicycle activity is on the State Highway System, indicating the importance of designing roadways for safe and convenient bicycle access.

It is impractical to count bicycles at every intersection and along every segment of the 15,000-mile State Highway System on a routine basis. This problem can be addressed by applying statistical models to estimate volumes at specific locations.

Previous Bicycle Demand Models

There are generally two approaches for bicycle volume estimation. One is choice based and the other is facility based (Proulx 2016).

- Choice-based models. Traditional activity-based modeling estimates large scale travel demand. Trips by different modes of transportation are summarized by an Origin-Destination Matrix (ODM) where each entry indicates the estimated number of trips between each OD pair by time of day, trip purpose etc. With ODM, a route choice model must be built to determine link or intersection-level volumes. This modeling framework is limited in estimating nonmotorized travel because of the relatively high occurrence of recreation trips where there's no destination. What's more, it's also not applicable in estimating intersection-level bicycle activity because extra data (e.g. stated preference data) need to be collected for route choice model construction.
- Facility-based models. This approach relies on volume counts at discrete locations. It can be decomposed into temporal extrapolation and spatial interpolation. (Proulx 2016) Temporal extrapolation tries to find out how traffic is distributed across time in order to map short duration observed volumes into Annual Average Daily Bicyclists (AADB) whereas spatial interpolation tries to figure out how volume counts vary across different locations. For our task, we are going to focus on spatial interpolation of facility-based model, or "direct demand" models.

Direct demand models are usually based on regression modeling to explain the relationship between volume counts and 'measured characteristics of the adjacent environment' (Kuzmyak et al. 2014). That is, the model explains the spatial variation in bicycle demand in terms of characteristics of surrounding environments like land use, transportation network, sociodemographics, etc. The model lacks in behavioral realism compared with choice-based model, but it is generally simple and easy to apply, which makes it the most widely used tool for bicycle volume estimation modeling. Typical steps used in the direct demand approach are listed below.

- First, bicycle counts are taken at a sample of locations in a community. These counts are often collected manually over short periods of time, but automated detection techniques that collect data over weeks, months, or even years can also be used.
- Second, short-period counts may be expanded to represent annual volume estimates (annual volume estimates can be compared with crash data that is reported on a yearly basis).
- Third, the estimated annual bicycle volumes are used as the dependent variable. A predictive model is built to establish the relationships between the bicycle volumes at each study location and explanatory variables describing the characteristics of the surrounding environments.
- Finally, the preferred prediction model can be used to estimate pedestrian volumes in other locations throughout the community.

Table 4-1 summarizes several recent direct demand bicycle volume models. Count data are obtained from different sources including manual count, automated count, public database and Strava tracking data. Manual counts are usually short counts (hourly or peak period). Models built based on manual counts usually make prediction for hourly or peak period demands. Some (Chen & Sun 2017) included both spatial and temporal (non-winter/winter, weekend etc.) variables in model so that the model can estimate bicycle demands through a full year but it required the count data to cover an entire year. Automated counts can have duration of a day or even a year and thus models built on these could produce hour-specific demand (Lu et al. 2018) or annual average daily bicyclist (AADB) (Roy et al. 2019). Hochmair et al. (2019) used Strava tracking data to build a model and estimated bicycle exposure in bicycle kilometer traveled rather than volumes. Most of the models considered land use, transportation system, and socioeconomic characteristics as independent variables. Some models that output hourly specific estimation also included temporal and weather variable for correction. Model structures include log-linear, stepwise linear, generalized linear mixed model, negative binomial, and geographically weighted regression. For example, Strauss et al. (2013) used hourly, weekly and monthly expansion factor to get average seasonal daily volumes (ASDV).

Table 4-1. Direct Demand Bicycle Volume Models

Model Location	Source	Locations Used for Model	Bicycle Count Description	Type of Count Sites	Count Period(s) Used for Model	Land Use	Transportation System	Socioeconomic Characteristics	Other	Model Output	Model Type
Alameda County, CA	UC Berkeley SafeTREC (Griswold et al. 2011)	81	Bicycle was logged according to the movement that was made. Only ridden bicycles were included	Intersections along arterial or collector roadways	Weekdays 12am -6 pm; weekends 9 am – 6 pm	• Number of Commercial properties. (0.1 mi) • Natural log of network dist. to UC Berkeley Campus Edge	• Connected node ratio (0.5 mi) • Presence of bicycle markings on any approach		• Average slope (degrees) of terrain (0.5 mi)	Total bicycle demand at intersections in Alameda for 2 hours	Log linear
Montreal, Quebec, Canada	McGill University (Jillian Strauss & Luis F Miranda-Moreno 2013)	758	8-h manual bicycle counts collected by the city of Montreal; Counts were then normalized to get average seasonal daily bicycle flows using expansion factors estimated from permanent automatic bicycle count stations	Signalized intersection	6 - 9 am, 11 am - 1 pm, 3:30 - 6:30 pm	• Employment (400m) • Presence of school (400m) • Land mix (800m)	• Number of bus stops (150 m) • Presence of a bicycle lane (15 m) • Presence of a cycle track (15 m) • Presence of parking entrance	• Mean income (50 m)	• Humidity • Presence of precipitation	Hourly bicycle flow adjusted for weather conditions, average seasonal daily volumes (ASDV)	Log-linear; negative binomial
Montreal, Quebec, Canada	McGill University (Jillian Strauss et al. 2013)	647	8-h manual bicycle counts collected by the city of Montreal	Signalized intersections	6 - 9 am, 11 am - 1 pm, 3:30 - 6:30 pm	• Employment (400m) • Presence of school (400m. • Land mix (800m) • Area of commercial land use (50m)	• Number of metro stations (150 m) • Length of bicycle facilities (800 m)			Average annual daily bicycle flows	Bivariate Poisson model
Calgary, Canada	University of Calgary (Maryam Tabeshian & Lina Kattan 2014)	34	6-h counts provided by the City of Calgary	Signalized and unsignalized intersections located on major arterials	7 - 9 am, 11 am - 13 pm, 16 - 18 pm	• Area of institutional space (0.5 mi) • Area of low-density residential space (0.1mi) • Area of commercial space (0.1 mi)	• Number of bus stop (0.25 mi) • Total number of street lanes reaching the intersection in all directions.			Total bike flow during the p.m. peak	Multiple linear and Poisson regression model

Model Location	Source	Locations Used for Model	Bicycle Count Description	Type of Count Sites	Count Period(s) Used for Model	Land Use	Transportation System	Socioeconomic Characteristics	Other	Model Output	Model Type
Seattle, WA	University of Utah & University of Texas at Austin (Daniel J Fagnant 2016)	251	Cyclist count data obtained from the Puget Sound Reginal Council	Signalized and unsignalized intersections	6 – 9 am, 3 – 6 pm, Tu & Th	• Population density • Employment density • recreational area access • residential	• Number of lanes • curb-lane width • bike-lane width • separated path sharrows • speed limit • bicycle-trail access • AADT		• mean temperature • morning period	Peak period bicyclist count	Negative binomial model
Minneapolis, MN	Virginia Polytechnic Institute and State University Steve Hankey & Greg Lindsey 2016)	471	Count data set is part of an ongoing effort by the Minneapolis Department of Public Works	Signalized and unsignalized intersections	4 – 6 pm	• Employment density • Population density (1250 m) • House density (200 m) • Retail area (100 m) • Industrial area (3000 m) • open space area (3000 m)	• Principal arterials (750 m) • Local roads (750 m) • Off-street trail (200, 2000 m) • On-street facility (100 m) • Intersection density (100 m) • Number of transit stops (1250 m)	• Household income (2500 m)	• Precipitation • Temperature	Peak period bicycle volumes	Stepwise linear varying the spatial scale of independent (land use & transportation) variables
Seattle, WA	Harbin Institute of Technology & University of Washington (Chen & Sun 2017)	50	Count data collected by SDOT according to the National Bicycle and Pedestrian Documentation methodology	Signalized and unsignalized intersections	Jan, May, July, and Sept. 10 - 12 am, 4 – 6 pm, 5 – 7 pm on weekdays; 12 – 2 pm on Sat.	• Percentage of steep areas (1 mi) • Percentage of water bodies (1, 0.5, 0.25 mi) • Percentage of offices (0.25 mi)	•Presence of bike lane (1, 0.25 mi)	• Percentage of white (1, 0.5, 0.25 mi) • median age (0.5, 0.25 mi)	• Non-winter/winter • Peak hour • Weekend	Total bicycle flows at intersections through a full year	Generalized linear mixed model (GLMM). Variables are tested for (1, 0.5, 0.25 mi buffer separately.
Chittenden County, Vermont	Southwest Jiaotong University (Yang et al. 2017)	346	Multiple comprehensive sets of field data collected in Chittenden County. An assembly of walking and biking traffic counts	Signalized and unsignalized intersections	4 -6 pm peak period	• Number of buildings (1000 ft) • Percentage of educational buildings (1000 ft) • Distance to downtown	• Number of car lanes (1000 ft) • Number of transit stops (1000 ft)			Total active mode trip volume during pm peak hour on Tu, Wed, Th during summer	Geographically weighted regression model

Model Location	Source	Locations Used for Model	Bicycle Count Description	Type of Count Sites	Count Period(s) Used for Model	Land Use	Transportation System	Socioeconomic Characteristics	Other	Model Output	Model Type
20 MSAs across U.S.	Virginia Tech (Le et al. 2018)	4593	Data from various publicly available sources. Intersection counts with turning movements are separated into segment counts for each leg of the intersection	Street segment or intersection	7 – 9 am and 16 – 18 or 17- 19 pm	• Area of water and green space (200, 300, 400, 500 m) • Number of jobs (200, 300 m) • Proximity to university or college campus	• Off-street facilities (bike trails, shared-use paths) • Multimodal network density (100, 500, 3000 m) • Intersection density (400, 2500 m)			Morning and afternoon peak-period bicycle traffic volumes	Stepwise linear
Blacksburg, VA	Virginia Tech (Lu et al. 2018)	101	Use automated count devices to collect counts of bicycle and then calculate annual average hourly traffic (AAHT)	Count locations along road, off-street trials and bike lanes	24 hours	Hour-specific: • Population density (100, 1250 m) • Number of Residential/non-residential addresses (500 m) • Industrial area (100 m) Spatiotemporal: • Population density (1250 m)	Hour-specific • Length of local roads (100 m) • Number of intersections (100 m) • Number of bus stop (100, 250, 1000 m) • Length of sidewalks (750m) • Centrality (point) • Length of on-street facility (100, 250, 500, 750, 1000 m) Spatiotemporal • Length of on-street facility (250 m) • Centrality	Hour-specific: • Income (100, 250, 750, 1750 m)	Spatiotemporal • Time of day	Bicycle demand for each hour of the day (hour-specific model); Demands in a 4-hour time interval (spatiotemporal)	Stepwise linear
Miami-Dade County area, FL	University of Florida (Hochmair et al. 2019)		Strava tracking data	Road segment level activity count		• Presence of bicycle park • Presence of bay bridge • Distance to bay/ocean (Within a block group)	• Length of local road • Length of local road with bike lane • Length of collector road with paved shoulder • Length of bike trail (Within a block group)	• Population • Household income • African American • male (Within a block group)		Bike Kilometer Traveled in a census block group	Linear

Model Location	Source	Locations Used for Model	Bicycle Count Description	Type of Count Sites	Count Period(s) Used for Model	Land Use	Transportation System	Socioeconomic Characteristics	Other	Model Output	Model Type
Maricopa County, AZ	Arizona State University (Roy et al. 2019)	44	Automated bicycle counts (annual average daily bicyclist AADB) completed by Association of Governments (MAG)	Signalized & non-signalized intersections	Eight continuous 2-week periods in April, Map, Oct, and Nov.	• Distance to residential areas • Distance to green spaces	• Average segment speed limit on street segment	• Percentage of white population • Median household income (Within census block group)	• Average number of Strava riders	Average annual daily bicyclist (AADB)	Linear

To conduct safety analysis of cycling, we need bicycle volume data, usually annual bicycle volume as an exposure measure. However, from the previous section we know that direct demand models usually give predictions about short term counts (e.g. hourly count). To get the total exposure across a year, short term counts must be converted to long term counts, which is usually done by using count expansion methods. Expansion factors are developed using long-term count data in the following approach:

- Aggregate hourly counts to get daily total volumes and compute overall daily average
- Calculate average daily volume for each day of the week and for each month of the year. Then obtain daily expansion factor by dividing each daily average by overall daily average and monthly expansion factor by dividing each monthly average by overall daily average.
- Calculate average hourly totals for each hour of the day. Then divide each hourly average by the overall average to get hourly expansion factors.

In addition to temporal factors, weather can also have a significant effect on bicycle volumes. Nosal et al. (2014) summarized four methods to estimate average annual daily bicycle volumes. The first two methods only account for temporal variation while the third and fourth methods try to also take weather variation into account. And it turns out that the weather and disaggregate models, which are capable of capturing weather variation, outperformed traditional methods.

- Traditional Method. This method accounts for daily and seasonal variation. It converts observed short-term daily bicyclists at short-term site i on day j in year y, which falls on day of week d in month m into AADBT, using day-of-the week factor for day of week d and month factor for month m
- Day-by- Month Method. This method account for daily and seasonal variation within one factor. It computed AADBT from short-term daily bicyclists at short-term site i on day j in year y, which falls on day of week d in month m directly using day-by-month factor for day of week d in month m.
- Weather Model. The model tries to capture the effect of weather on bicycle volumes by using the expected cyclist count deviation. The short-term count is first adjusted on the basis of the predicted deviation from the 21-day moving average due to weather and then the adjusted count is converted into AADBT using moving average factor which reflects how 21-day average varies from the AADBT value.
- Disaggregate Factor Method. This method is simple and straightforward. An expansion factor is created for each day of the year. Both weather and temporal variation could be considered as long as the weather is the same at long-term and short-term sites.

Since temporal pattern of traffic volumes could also vary in space, the Traffic Monitoring Guide (TMG) recommends to classify patterns into "factoring groups". There are two approaches to group long-term count patterns, one is called land use classification, which is based on surrounding land use pattern. The other one is empirical clustering approach, which uses statistics to group sites with similar activity pattern. For example, Miranda-Moreno et al. (2013) classified sites into four categories: utilitarian, mixed utilitarian, mixed recreational and recreational sites, based on hour-of-day pattern on weekday and weekend, and day-of week pattern.

Empirical clustering approach could yield more accurate estimates but it's not easy to map unobserved sites into factor groups. Proulx (2016) proposed a decomposition method for link characterization. Latent Dirichlet Allocation was used to get weights for each identified latent bicycle trip types on observed links and recognize the temporal patterns for each trip type. Then a fully spatio-directionally autoregressive fractional logit model is used to infer the proportion of each trip type on an unobserved link, based on which the overall temporal profile of trips on a given link is reconstructed. To bridge the gap between land use and empirical clustering approach, Medury et al. (2019) fitted a multinomial logit model to identify the relationship between empirical clusters and land use patterns. The model is then applied to match a location to certain factor group using land use data.

In addition to the expansion factor approach, which needs long-term count data to calculate, a Seasonal Adjustment Regression Model (SARM) was developed by Roll & Proulx (2018) using only short duration counts from multiple years of data. Daily conditions like maximum daily temperature, total daily precipitation, minutes of daylight were used to predict annual bicycle traffic volumes. This approach does not require permanent counters but relies on a large set of short duration counts at each location where volumes are desired.

Potential Explanatory Variables

Using the direct demand modeling approach, bicycle volumes are assumed to be a function of the characteristics at and around specific locations on the California State Highway System. These characteristics will be represented by a set of explanatory variables. Previous research suggests explanatory variables representing land use, transportation system, socioeconomic, and several other characteristics are associated with bicycle volumes. While there are many possible bicycle model inputs, some explanatory variables will be easier than others to gather statewide. For example, population density is provided by the U.S. Census Bureau's American Community Survey at the block level for the entire country, so this information would be relatively easy to obtain for any location along the State Highway System. In contrast, there are no statewide databases of commercial property locations (this information has been gathered in previous studies through special requests to county tax assessors). Lists of potential explanatory variables and the assumed ease of collecting these variables are provided in Table 4-2, Table 4-3, and Table 4-4. Ease of collecting each variable is classified into the following categories:

- Easy. Data are available statewide from an existing data source. The variable can be created through basic GIS analysis.
- Moderate. Data are available for most or all of the state from existing data sources, but the data may be in different formats in different jurisdictions. The variable may require more sophisticated GIS analysis to create.
- Difficult. Data are not available from existing data sources. Field data collection or manual data collection from aerial or street-level imagery may be needed to create the variable.

Table 4-2. Potential Bicycle Volume Model Inputs and Ease of Data Collection: Land Use Variables

Variable	Study (buffer area used)	Ease of Collection
Population density within a given distance	Minneapolis, MN (Hankey & Lindsey 2016) (1250 m); Blacksburg, VA (Lu et al. 2018) (100, 1250 m); Seattle, WA (Fagnant & Kockelman 2016)	Easy (American Community Survey block data)
Employment density within a given distance	Montreal, QC (Strauss et al. 2013; Strauss & Miranda-Moreno 2013) (400 m); Seattle, WA (Fagnant & Kockelman 2016); Minneapolis, MN (Hankey & Lindsey 2016);	Easy (Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics block data)
Number of jobs within a given distance	20 MSAs across U.S. (Le et al. 2018) (200, 300 m)	Easy (Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics block data)
Network distance to campus edge	Alameda County, CA (Griswold et al. 2011)	Easy (California Department of Education GIS data)
Presence of school within a given distance	Montreal, QC (Strauss et al. 2013; Strauss & Miranda-Moreno 2013) (400 m);	Easy (California Department of Education GIS data)
Percentage of educational buildings within a given distance	Chittenden, Vermont (Yang et al. 2017) (1000 ft);	Easy (California Department of Education GIS data)

Area of commercial land use	Montreal, QC (Strauss et al. 2013) (50 m);
	Calgary, Canada (Tabeshian & Kattan 2014) (0.1 mi);
		Moderate (County tax assessor parcel data; jurisdiction-specific)

Number of commercial properties within a given distance	Alameda County, CA (Griswold et al. 2011) (0.1 mi)	Moderate (County tax assessor parcel data; jurisdiction-specific)
Area of institutional space within a given distance	Calgary, Canada (Tabeshian & Kattan 2014) (0.5 mi);	Moderate (County tax assessor parcel data; jurisdiction-specific)
Area of low-density residential space within a given distance	Calgary, Canada (Tabeshian & Kattan 2014) (0.1 mi);	Moderate (County tax assessor parcel data; need to look to each jurisdiction)
Number of residential addresses within a given distance	Blacksburg, VA (Lu et al. 2018) (500 m);	Moderate (County tax assessor parcel data; jurisdiction-specific)
House density within a given distance	Minneapolis, MN (Hankey & Lindsey 2016) (200 m);	Moderate (County tax assessor parcel data; jurisdiction-specific)

Variable	Study (buffer area used)	Ease of Collection
Distance to residential areas	Maricopa, AZ (Roy et al. 2019);	Moderate (County tax assessor parcel data; jurisdiction-specific)
Land mix within a given distance	Montreal, QC (Strauss et al. 2013; Strauss & Miranda-Moreno 2013) (800 m);	Difficult (County tax assessor parcel data; jurisdiction-specific and also requires complex calculation)
Area of retail land use within a given distance	Minneapolis, MN (Hankey & Lindsey 2016) (100 m)	Moderate (County tax assessor parcel data; need to look to each jurisdiction)
Area of industrial land use within a given distance	Minneapolis, MN (Hankey & Lindsey 2016) (3000 m); Blacksburg, VA (Lu et al. 2018) (100 m);	Moderate (County tax assessor parcel data; jurisdiction-specific)
Area of open space within a given distance	Minneapolis, MN (Hankey & Lindsey 2016) (3000 m)	Moderate (County tax assessor parcel data; jurisdiction-specific and also need to define "open space")
Percentage of water bodies within a given distance	Seattle, WA (Chen et al. 2017) (0.25, 0.5, 1 mi);	Easy (US Census Bureau GIS data)
Area of water and green space within a given distance	20 MSAs across U.S. (Le et al. 2018) (200, 300, 400, 500 m)	Easy (US Census Bureau GIS data)
Distance to green spaces	Maricopa, AZ (Roy et al. 2019)	Easy (US Census Bureau GIS data)
Distance to bay/ocean	Miami-Dade, FL (Hochmair et al. 2019)	Easy (US Census Bureau GIS data)
Percentage of steep areas within a given distance	Seattle, WA (Chen et al. 2017) (1 mi)	Easy (US Geological Survey National Elevation Dataset)
Percentage of offices within a given distance	Seattle, WA (Chen et al. 2017) (0.25 mi)	Moderate (County tax assessor parcel data; jurisdiction-specific)
Number of buildings within a given distance	Chittenden, Vermont (Yang et al. 2017) (1000 ft)	
Distance to downtown	Chittenden, Vermont (Yang et al. 2017)	Moderate (US Census Bureau GIS data; need to define CBD location(s) within each region)
Presence of bicycle park within a block group	Miami-Dade, FL (Hochmair et al. 2019)	
Presence of bay bridge	Miami-Dade, FL (Hochmair et al. 2019)	
Recreational area access	Seattle, WA (Fagnant & Kockelman 2016);	

Table 4-3. Potential Bicycle Volume Model Inputs and Ease of Data Collection: Transportation System Variables

Variable	Study (buffer area used)	Ease of Collection
Connected node ratio within a given distance	Alameda County (Griswold et al. 2011) (0.5 mi)	
Presence of bicycle markings	Alameda County (Griswold et al. 2011)	Difficult (Pedestrian and bicycle facility inventories do not exist statewide)
Presence of bicycle lane within a given distance	Montreal, QC (Strauss & Miranda-Moreno 2013) (15 m); Seattle, WA (1) (0.25, 1 mi)	Difficult (Pedestrian and bicycle facility inventories do not exist statewide)
Presence of cycle track within a given distance	Montreal, QC (Strauss & Miranda-Moreno 2013) (15 m)	Difficult (Pedestrian and bicycle facility inventories do not exist statewide)
Length of bicycle facilities within a given distance	Montreal, QC (Strauss & Miranda-Moreno 2013) (800 m)	Difficult (Pedestrian and bicycle facility inventories do not exist statewide)
Number of bus stops within a given distance	Montreal, QC (Strauss & Miranda-Moreno 2013) (150 m); Calgary, Canada (Tabeshian & Kattan 2014) (0.25 mi); Blacksburg, VA (8) (100, 250, 1000 m);	Moderate (Metropolitan Planning Organization or Regional Transit Authority; jurisdiction-specific)
Number of metro stations within a given distance	Montreal, QC (Strauss et al. 2013) (150 m)	
Number of transit stops within a given distance	Minneapolis, MN (Hankey & Lindsey 2016) (1250 m); Chittenden, Vermont (Yang et al. 2017) (1000 ft)	Moderate (Metropolitan Planning Organization or Regional Transit Authority; jurisdiction-specific)
Number of street lanes reaching the intersection in all directions	Calgary, Canada (Tabeshian & Kattan 2014) Chittenden, Vermont (Yang et al. 2017) (1000 ft)	Moderate (Caltrans TASAS data; needs to be connected from segments to intersections and may not be available for all roads)
Length of principal arterials within a given distance	Minneapolis, MN (Hankey & Lindsey 2016) (750 m)	Moderate (US Census GIS data or Caltrans TASAS data; need reliable classification of arterial roadways)

Variable	Study (buffer area used)	Ease of Collection
Length of local roads within a given distance	Blacksburg, VA (Lu et al. 2018) (100 m); Minneapolis, MN (Hankey & Lindsey 2016) (750 m)	Moderate (US Census GIS data or Caltrans TASAS data; need reliable classification of arterial roadways)
Length of sidewalks within a given distance	Blacksburg, VA (Lu et al. 2018) (750 m); Miami-Dade, FL (Hochmair et al. 2019) (within block group)	Difficult (Pedestrian and bicycle facility inventories do not exist statewide)
Length of local roads with bike lane within a given distance	Miami-Dade, FL (Hochmair et al. 2019) (within block group)	Difficult (Pedestrian and bicycle facility inventories do not exist statewide)
Length of collector road with paved shoulder within a block group	Miami-Dade, FL (Hochmair et al. 2019) (within block group)	Moderate (US Census GIS data or Caltrans TASAS data; need reliable classification of arterial roadways)
Length of on-street facility within a given distance	Minneapolis, MN (Hankey & Lindsey 2016) (100 m); Blacksburg, VA (8) (100, 250, 500, 750, 1000 m)	
Length of off-street trail within a given distance	Minneapolis, MN (Hankey & Lindsey 2016) (200, 2000 m); 20 MSAs across U.S. (Le et al. 2018); Miami-Dade, FL (Hochmair et al. 2019) (within block group)	Difficult (Pedestrian and bicycle facility inventories do not exist statewide)
Bike-lane width	Seattle, WA (Fagnant & Kockelman 2016)	Difficult (Pedestrian and bicycle facility inventories do not exist statewide)
Curb-lane width	Seattle, WA (Fagnant & Kockelman 2016)	
Presence of separated path sharrows	Seattle, WA (Fagnant & Kockelman 2016)	
Speed limit	Seattle, WA (Fagnant & Kockelman 2016); Maricopa, AZ (Roy et al. 2019)	Moderate (Caltrans TASAS data; needs to be connected from segments to intersections and may not be available for all roads)
Bicycle-trail access	Seattle, WA (Fagnant & Kockelman 2016)	Difficult (Pedestrian and bicycle facility inventories do not exist statewide)
Multimodal network density within a given distance	20 MSAs across U.S. (Le et al. 2018) (100, 500, 3000 m)	
Intersection density within a given distance	20 MSAs across U.S. (Le et al. 2018) (400, 2500 m)	Easy (US Census GIS data or Caltrans GIS data)

Variable	Study (buffer area used)	Ease of Collection
Number of intersections within a given distance	Blacksburg, VA (Lu et al. 2018) (100 m)	Easy (US Census GIS data or Caltrans GIS data)
Centrality	Blacksburg, VA (Lu et al. 2018) (100 m);	
Presence of parking entrance	Montreal, QC (Strauss & Miranda-Moreno 2013)	
Link-level recreational volume	Miami-Dade County area, FL (Hochmair et al. 2019); Maricopa County, AZ (Roy et al. 2019)	Easy (Strava Metro)

Table 4-4. Potential Bicycle Volume Model Inputs and Ease of Data Collection: Socioeconomic and Other Variables

Variable	Study (buffer area used)	Ease of Collection
Household income within a given distance	Montreal, QC (Strauss & Miranda-Moreno 2013) (50 m); Minneapolis, MN (Hankey & Lindsey 2016) (2500 m); Blacksburg, VA (Lu et al. 2018) (100, 250, 750, 1750 m); Miami-Dade, FL (Hochmair et al. 2019) (block group); Maricopa, AZ (Roy et al. 2019) (block group)	Easy (American Community Survey block data)
Percent of white within a given distance	Seattle, WA (Chen et al. 2017) (0.25, 0.5, 1 mi); Maricopa, AZ (Roy et al. 2019) (block group)	Easy (American Community Survey block data)
Median age within a given distance	Seattle, WA (Chen et al. 2017) (0.25, 0.5 mi)	Easy (American Community Survey block data)
Average slope within a given distance	Alameda County, CA (Griswold et al. 2011) (0.5 mi.)	Easy (US Geological Survey National Elevation Dataset)
Humidity	Montreal, QC (Strauss & Miranda-Moreno 2013)	Easy (National Oceanic and Atmospheric Administration weather data)
Presence of precipitation	Montreal, QC (Strauss & Miranda-Moreno 2013); Minneapolis, MN (Hankey & Lindsey 2016)	Easy (National Oceanic and Atmospheric Administration weather data)
Temperature	Seattle, WA (Fagnant & Kockelman 2016); Minneapolis, MN (Hankey & Lindsey 2016)	Easy (National Oceanic and Atmospheric Administration weather data)
Morning period	Seattle, WA (Fagnant & Kockelman 2016)	
Non-winter/winter	Seattle, WA (Chen et al. 2017)	
Peak hour	Seattle, WA (Chen et al. 2017)	

Variable	Study (buffer area used)	Ease of Collection
Weekend	Seattle, WA (Chen et al. 2017)	
Time of day	Blacksburg, VA (Lu et al. 2018)	

Proposed Modeling Approach

Overall, our understanding is that this project will develop a link-level bicycle volume model based on count data. The focus is on estimating volumes for the state highway system, but count data from local roads will also likely be incorporated in the model estimation, given limited count data availability from state highway locations.

Some of the key challenges we see in developing a model of this scale include:

- The applicability of various datasets will depend heavily on roadway context (urban/suburban/ rural).
- Both travel along the highways and travel across highways (especially freeways) should ideally be accounted for.
- Utilitarian trips may be more related to surrounding land uses/destinations, whereas recreational trips probably are more complicated and could be a combination of local destinations, attractiveness of route, and proximity to population centers. These recreational trips are expected to make up a greater share of trips on rural highways, but there is limited documentation of the patterns in these areas.

Overall model form: Poisson Mixture Model

One approach to consider is to formulate the model as a Poisson Mixture Model, where volume on link *l* is:

$$v_l \sim \pi_{rec} Poisson(\lambda_l^{rec}) + (1 - \pi_{rec}) Poisson(\lambda_l^{util})$$

Where

$$v_l = volume\ on\ link\ \mathrm{l}$$
$$\pi_{rec} = probability\ of\ an\ observation\ being\ drawn\ from\ the\ \mathrm{recreational}\ group$$
$$\lambda_l^{rec} = Poisson\ rate\ parameter\ for\ link\ \mathrm{l}\ for\ recreational\ trips$$
$$\lambda_l^{util} = Poisson\ rate\ parameter\ for\ link\ \mathrm{l}\ for\ utilitarian\ trips$$

Mixture models like this are suited to situations where an observation is hypothesized to be drawn from one of multiple subpopulations, but each observation's membership in those subpopulations is not observed. This formulation could be useful because it allows us to separately formulate different relationships for these two types of trips—i.e., to have a utilitarian component and a recreational component within the same model. Karlis (2005) provides a method for estimating Poisson mixture models: https://www.casact.org/library/astin/vol35no1/3.pdf.

Recreational Trip Component

For recreational trips, we could consider a model of the form:

$$v_l^{rec} \sim Poisson(\lambda_l^{rec})$$

$\log(\lambda_l^{rec})$ =fn(distance to nearest population center, density of population, other factors)

Essentially, the rate parameter for these trips would be modeled as a function of proximity to population centers. This could be augmented with the bicycle commute mode share of the region and possibly other factors that might indicate the propensity for biking in the region and the suitability/popularity of a given route. Alternatively, if recreation-focused crowdsourced data were available (e.g., Strava Metro), these trips could be modeled with that data.

Utilitarian Trip Component

For utilitarian trips, there are multiple approaches that could be taken. One idea relies on working with O-D data, such as from the CA statewide travel demand model. In this formulation, the goal would be to relate the potential trips associated with each O-D pair to the links they might use.

The goal of formulating the utilitarian model in this way is to subvert some of the following issues with previous work in thisarea:

- Bicycle route choice models are difficult to estimate, subject to data limitations, and can be overly deterministic in predicting the route that a given cyclist will take, ignoring heterogeneity in preferences between cyclists.
- Some direct-demand models have related bicycle traffic volumes to land uses immediately surrounding the observation location. However, travel happens along the corridor between the origin and destination, not simply in the immediate surroundings of the origin and destination.

Pilot Model Development

Figure 4-1 describes the workflow for direct demand modeling, including development of dependent and independent variables, model estimation, and model application.

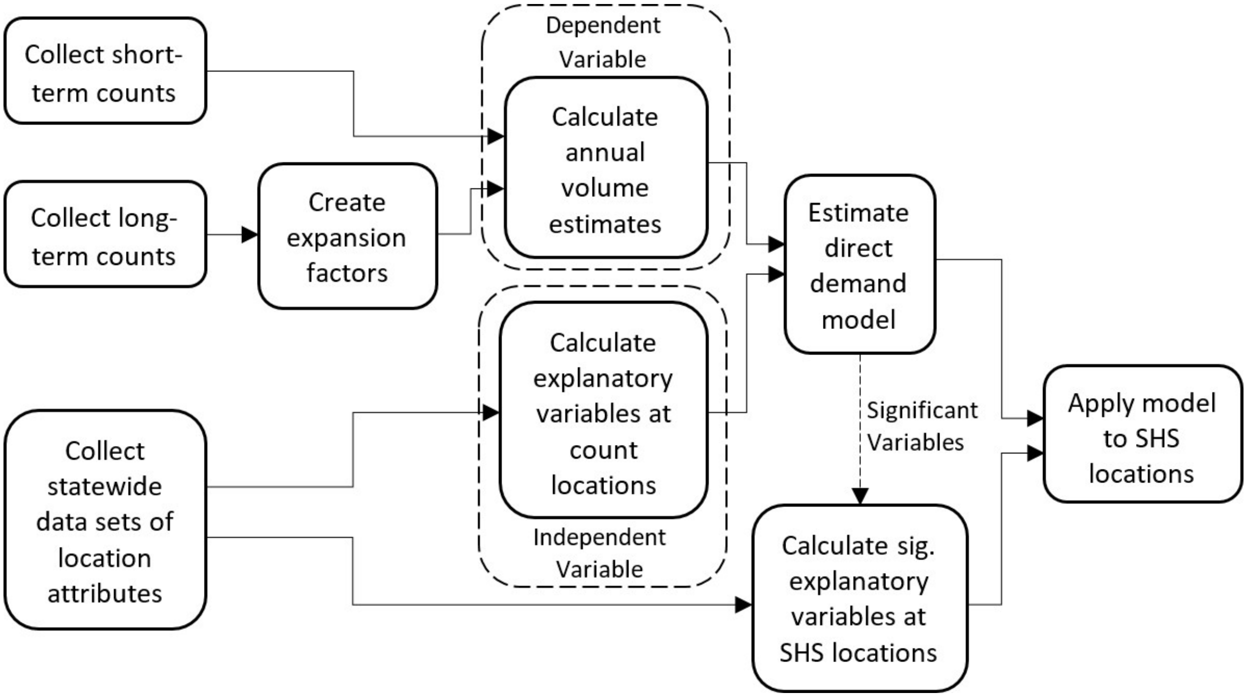

Collect short-term counts
Dependent Variable
Calculate annual volume estimates
Collect long-term counts
Create expansion factors
Estimate direct demand model
Calculate explanatory variables at count locations
Significant Variables
Apply model to SHS locations
Collect statewide data sets of location attributes
Independent Variable
Calculate sig. explanatory variables at SHS locations

Figure 4-1. Direct demand modeling process

Dependent Variable

The dependent variable for the bicycle exposure model will be the annual bidirectional link- or segment-level flow. Because agencies cannot afford to conduct long-term counts, either manually or using automated counting technology, we began the process to create annual count estimates by expanding short-term crossing counts using expansion factors.

This approach required that we compile large amounts of short-term count data, as the dependent variable for the model, and long-term count data to create the expansion factors. The count data processing involved two main tasks. First, we used the long-term count data to develop expansion factors, and second, we applied the expansion factors to the short-term counts to create the annual volume estimates.

Count Data Compilation

Each Caltrans district has a budget for collecting video-based count data through Miovision, and these data, collected at several hundred locations, formed the basis of the short-term bicycle count data. Among the 428 studies with bicycle count data, durations ranged from 1 hour to 96 hours. Count durations longer than 12 hours were generally multiple daytime counts, such as 7AM to 7PM on consecutive days.

To acquire more data, Caltrans Local Assistance emailed a count data request to a list of previous applicants for Active Transportation Program (ATP) grant funding as they had done for a previous pedestrian study. A number of agencies shared their pedestrian and bicycle count data sets. Table 4-5 describes the short-term bicycle count studies that we received in these outreach efforts. Most counts from local agencies were conducted during morning and afternoon 2-hour peaks and some also included midday peak.

Table 4-5. Count studies by Caltrans district and location type

District	Intersection	Segment
1	25	6
2	14	0
3	193	4
4	272	0
5	258	97
6	517	0
7	1294	2
8	1	0
9	103	0
10	1	0
11	80	0
12	80	250

Because we intend to model bicycle segment flow, we translated the intersection turn movement counts to approach counts. Figure 4-2 shows an example of the turning movements that we sum to estimate the volumes on the north approach. We applied a corresponding method for the rest of the approaches. To match the correct segment approach to each intersection point in GIS, we wrote a Python script that calculated the bearing of each segment and associated it with the nearest direction according to the thresholds in Figure 4-3.

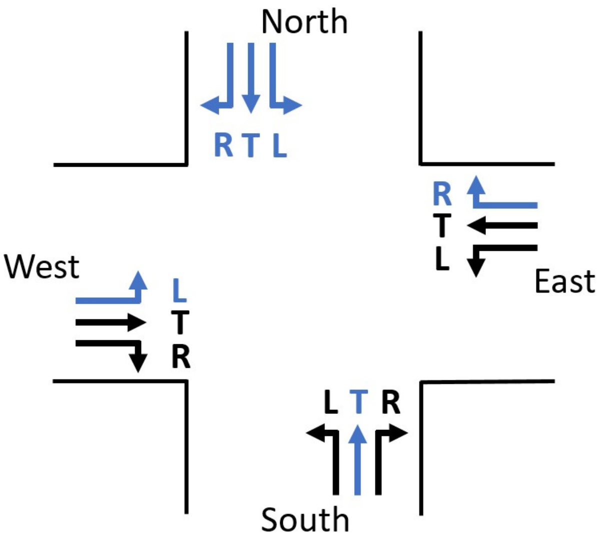


Figure 4-2. Intersection turning movements with north approach counts in blue

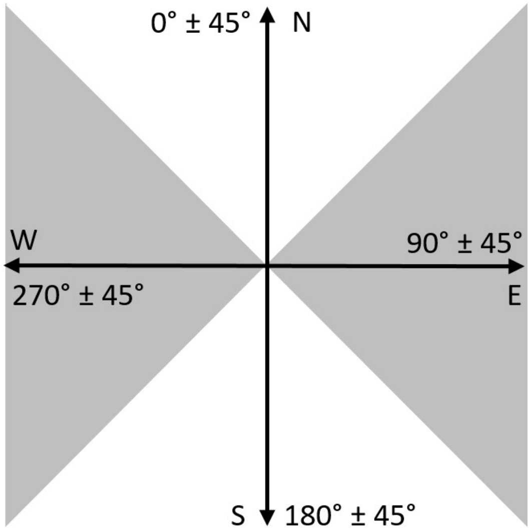


Figure 4-3. Bearings matched with each direction for assignment of direction to segment approaches

The next step in the dependent variable development will be to expand the short-term segment counts to annual counts using expansion factor groups. Given the limited research available on the connection between activity patterns and land use, we will likely use simple factor groups based on urban and rural locations. The urban factor group will demonstrate utilitarian activity patterns and the rural group will demonstrate recreational ride patterns.

Independent Variables

The complete list of independent, or explanatory, variables is shown in Table 4-6. For variables that pertain to an area around a segment, such as population, the value was calculated at three different buffer distances–half-mile, quarter-mile, and tenth-mile. The scale of these variables is described as "buffer" in the scale column of Table 4-6. We used density instead of total for some of the buffer variables for consistency due to the variable length of street segments, and thus, the variable areas of the segment buffers. Other variables are related to the specific attributes of the segment location and are described as "segment" in Table 4-6. Explanation for how we calculated each variable is described below.

Table 4-6. Explanatory (independent) Variables

Description	Scale	Data Source	Status
Demographics			
Population density	Buffer	U.S. Census ACS	Complete
Household density	Buffer	U.S. Census ACS	Complete
Population that is white alone density	Buffer	U.S. Census ACS	Complete
Density of bicycle commuters	Buffer	U.S. Census ACS	Complete
Density of households with no vehicle	Buffer	U.S. Census ACS	Complete
Density of college degree holders	Buffer	U.S. Census ACS	Complete
Percent of population that is white alone	Buffer	U.S. Census ACS	Complete
Bicycle commute mode share	Buffer	U.S. Census ACS	Complete
Percent of households with no vehicle	Buffer	U.S. Census ACS	Complete
Percent of population with a college degree	Buffer	U.S. Census ACS	Complete
Infrastructure			
Bicycle facility	Segment	ATAIP and manual	Partially complete
Functional classification	Segment	CRS	Complete
Speed limit	Segment	CRS	Complete
Adjacent intersection has a signal	Segment	Open Street Map	Planned
Network Connectivity			
Number of meters of streets per area	Buffer	U.S. Census TIGER	Planned
Street segments density	Buffer	U.S. Census TIGER	Planned
Employment/Land Use			
Employment square footage of foot traffic land uses	Buffer	ESRI Business Analyst	Planned
Number of employees	Buffer	ESRI Business Analyst	Complete
Other			
Strava segment volumes	Segment	Strava Metro	planned
Distance to University	Segment	ESRI Business Analyst	Complete
Slope	Segment	Mapquest API	Complete

Population and demographics

We used U.S. Census American Community Survey data to develop the demographic variables. The five-year dataset, collected from 2012 to 2016, provides sample-based estimates at the block group level; block groups are smaller than tracts but larger than blocks. The spatial resolution worked well for our tenth, quarter-mile, and half-mile buffer distances.

We collected demographic data on race, education, households, and commute mode. For each attribute, we calculated both the density and the percent of the block group population.

Our analysis required us to take the Census data and analyze it spatially, near the count locations. We used the Census GIS shapefile and joined the columns in Table 4-7 to it.

Table 4-7. Census variables used in analysis

Variable	Census Description
B02001e1	Race: Total: Total population -- (Estimate)
B02001e2	Race: White alone: Total population -- (Estimate)
B08301e1	Means of Transportation to Work: Total: Workers 16 years and over -- (Estimate)
B08301e19	Means of Transportation to Work: Walked: Workers 16 years and over -- (Estimate)
B08301e10	Means of Transportation to Work: Public transportation (excluding taxicab): Workers 16 years and over -- (Estimate) (includes bus, streetcar, subway, railroad, and ferryboat)
B11001e1	Household Type (Including living alone): Total: Households -- (Estimate)
B15003e1	Educational Attainment for the Population 25 Years and Over: Total: Population -- (Estimate)
B15003e22	Educational Attainment for the Population 25 Years and Over: Bachelor's degree -- (Estimate)
B25044e1	Tenure by Vehicles Available: Total: Occupied housing units -- (Estimate)
B25044e3	Tenure by Vehicles Available: Owner occupied: No vehicle available: Occupied housing units -- (Estimate)
B25044e10	Tenure by Vehicles Available: Renter occupied: No vehicle available: Occupied housing units -- (Estimate)

We wrote a Python script using the ArcPy library and an R script using the tidycensus package to process the variables. For each buffer distance, the Pythonscript clipped the block groups by the appropriate buffer, calculated the area of the clipped block groups, and then divided that area by the total area of the original block groups to determine the percentage of block groups that fall within the buffer. The calculations for the different types of variables in R were as follows:

- For the density variables, like population density, the percentage was used to scale down the total block group population to an area-based estimate of the population that falls within the clipped block group. Summing the population estimates of the block groups by buffer and dividing by the area of buffer, produced estimates of the population density falling within each buffer.

- For the percentage variables, like percent of population that is white alone, we followed the steps described above for both the numerator variable, white-alone population, and the denominator variable, total population. The final variable was the ratio of the two.

Table 4-8 below lists the input variables used to make the calculations for each demographic variable.

Table 4-8. Input variables used for calculation for each Census variable

Variable Name	Numerator Variable	Denominator Variable
Population	B02001e1	Buffer area
Number of households	B11001e1	Buffer area
Population that is white alone	B02001e2	Buffer area
Number of walk commuters	B080301e19	Buffer area
Number of transit commuters	B08301e10	Buffer area
Number of households with no vehicle	B25044e3 + B25044e10	Buffer area
Number of college degree holders	B15003e22	Buffer area
Percent of population that is white alone	B02001e2	B02001e1
Walk commute mode share	B08301e19	B08301e1
Transit commute mode share	B08301e10	B08301e1
Percent of households with no vehicle	B25044e3 + B25044e10	B25044e1
Percent of population with a college degree	B15003e22	B15003e1

Employment

We selected two employment metrics: employment density and square footage density of traffic-generating commercial uses. The first metric attempts to capture the contribution to pedestrian exposure from people working near the relevant intersections. The second measure will capture the scale of businesses that generate walking trips by attracting customers. The data source for both metrics is ESRI Business Analyst software. ESRI sourced the data from Infogroup. The dataset mapped every business in the United States, complete with the number of employees that work there and the approximate size, in square feet, of the business. To determine the employment density near relevant segments, we conducted a GIS analysis that selected all businesses within our chosen buffer distance, summed the number of employees in those businesses, and divided it by the area of the buffer. We did not discriminate based on the type of business.

We will use the same dataset for business square footage density, but we will filter the businesses by type. Warehouses, for example, do not generate significant foot traffic outside of their employees, and that foot traffic is captured in the metric above. Each of the businesses in the ESRI Business Analyst dataset has a corresponding North American Industry Classification System (NAICS) code that categorizes the business by type. For our analysis, we will only consider businesses from the following categories:

- 44-45: Retail Trade
- 522: Banks
- 54: Professional, Scientific, and Technical Services
- 62: Health Care and Social Assistance

- 71: Arts, Entertainment, and Recreation
- 72: Accommodation and Food Services
- 812: Personal and Laundry Services
- 813: Religious, Grantmaking, Civic, Professional, and Similar Organizations

The ESRI Business Analyst data do not provide exact square foot measurements for each business, but instead categorizes them into one of four ranges:

- A: 1 - 2,499 square feet
- B: 2,500 - 9,999 square feet
- C: 10,000 - 39,999 square feet
- D: 40,000 square feet and above

We will use the middle of each of the A-C ranges and the lowest value for range D when summing the total amount of square feet within the buffer distance. Therefore, we will apply the value 1,250 for A, 6,250 for B, 25,000 for C, and 40,000 for D. We will sum all of the square footages for the businesses within the buffers and divide it by the area of the buffer for our metric.

Infrastructure

Bicycle facility information is available at state highway locations through data collected by Caltrans as part of the Active Transportation Asset Inventory Program. This information will need to be manually collected at off-network locations. We will generate several variable options, including dummy variables or scaled variables, to represent this information for testing in the model. We plan a similar approach for functional classification and speed limit which are already available statewide from the California Road System dataset. We can gather signalization information from Open Street Map data.

Network Connectivity

We plan to develop network connectivity measures that count the number of meters of streets per area within the buffer and the density of street segments, the count of street segments within the buffer divided by the area of the buffer).

Strava Segment Volumes

Strava Metro aggregates and de-identifies trip data by segment to create segment-level bicycle volume estimates. SafeTREC's application for access to these data is pending with Strava. Inclusion of these data will enhance the recreational portion of the model.

Distance to University

This variable is based on the ESRI Business Analyst data category NAICS code 61131013. We calculated the distance to the nearest entity with this category.

Slope

The slope is the average slope between the beginning and end points of each street segment. We gathered the elevation from the beginning and end points of each segment, and calculated the difference divided by the length of the segment.

Next Steps

Completion of the model will occur in the next phase of the study. This task will include expansion of the short-term segment counts to annual segment flows to be used as the dependent variables, gathering additional data and processing for additional explanatory variables, exposure model estimation, and application of the exposure model to state highway segments (and cross streets where possible).

Chapter 5. Conclusion

This report aimed to create a comprehensive picture of bicycle safety in California, as well as to continue and support the efforts for implementing the Bicycle Safety Monitoring Program in California. Each chapter in this report describes an activity that contributes to the overall strategy to enhance bicycle safety in California. A concise summary, important insights, and some recommendations for each chapter are provided below:

Chapter 2 – Bicycle Safety Monitoring Report Tool described the bicycle crash corridor methodology which has been incorporated into the set of tools previously developed for pedestrian hotspot identification.

Key insights:

- The new Bicycle Safety Monitoring Report Tool takes into consideration the unique differences between pedestrian and bicyclist collisions.
- The bicycle crash corridor methodology utilizes the Density Based Spatial Clustering of Application with Noise (DBSCAN) algorithm for identification of corridors.
- Attributes for corridor prioritization include corridor length, total number of crashes, and crashes per mile.

Recommendations:

- Develop a methodology to incorporate bicycle exposure into the BSMR Tool to select and prioritize crash hotspots and corridors.

Chapter 3 – Systemic Safety Tool for Bicycles described the core elements of the systemic crash analysis and the resources used to establish the bicycle crash matrix structure as well as the development of the prototype Excel tool.

Key insights:

- Existing Caltrans databases include most, but not all, of the minimum required elements for application of the systemic approach.
- The final structure of the systemic intersection matrix included crash dynamics of violation category, party at fault, and collision type and infrastructure features of traffic control type, mainline left turn channelization, and cross-street left turn channelization.
- The final structure of the systemic highway matrix included crash dynamics of violation category, party at fault, and collision type and infrastructure features of median presence, median barrier type, number of lanes, and vehicular traffic volume.
- A threshold of 99th percentile is indicated as reasonable criterion for selecting systemic hotspots.

Chapter 4 – Bicycle Exposure Modeling Approach described the process towards developing a state-scale pedestrian exposure model for the California State Highway System (SHS). The report explains the data that were collected, the processing and analysis of those data, and the modeling approach planned for the next phase.

Key insights:

- Local agencies have data that are beneficial for larger scale modeling projects. The project team was able to utilize such counts for the purpose of this project.
- The differences between bicyclist and pedestrian travel make a direct demand modeling approach more complex.
- The team developed Poisson mixture model approach that will account for both recreational and utilitarian bicycle trips.
- There are data processing challenges when working on a statewide scale. Future modeling work will be more efficiently completed with a GIS-based dataset containing Caltrans infrastructure data.

Recommendations:

- As the quality of big data sources for bicycle trips improves, these data may be used to improve the quality of bicycle exposure predictions using data fusion techniques.
- Systematic collection of bicycle count data, including a statewide network of permanent bicycle counters, would reduce the error in bicycle count expansion methods improve accuracy of the dependent variable for modeling purposes.

References

Chen, P., Zhou, J., & Sun, F. (2017). Built environment determinants of bicycle volume: A longitudinal analysis. Journal of Transport and Land Use, 10(1).

Fagnant, D. J., & Kockelman, K. (2016). A direct-demand model for bicycle counts: the impacts of level of service and other factors. Environment and Planning B: Planning and Design, 43(1), 93-107.

Griswold, J. B., Medury, A., & Schneider, R. J. (2011). Pilot models for estimating bicycle intersection volumes. Transportation research record, 2247(1), 1-7.

Hakkert, A. S., Braimaister, L., & Van Schagen, I. (2002). *The uses of exposure and risk in road safety studies* (Vol. 2002, No. 12). SWOV Institute for Road Safety.

Hankey, S., & Lindsey, G. (2016). Facility-demand models of peak period pedestrian and bicycle traffic: comparison of fully specified and reduced-form models. Transportation research record, 2586(1), 48-58.

Hochmair, H. H., Bardin, E., & Ahmouda, A. (2019). Estimating bicycle trip volume for Miami-Dade County from Strava tracking data. Journal of Transport Geography, 75, 58-69.

Karlis, D. (2005). EM algorithm for mixed Poisson and other discrete distributions. ASTIN Bulletin: The Journal of the IAA, 35(1), 3-24.

Kuzmyak, J. R., Walters, J., Bradley, M., & Kockelman, K. M. (2014). Estimating bicycling and walking for planning and project development: A guidebook (No. Project 08-78).

Le, H. T., Buehler, R., & Hankey, S. (2018). Correlates of the built environment and active travel: evidence from 20 US metropolitan areas. Environmental health perspectives, 126(07), 077011.

Lu, T., Mondschein, A., Buehler, R., & Hankey, S. (2018). Adding temporal information to direct-demand models: Hourly estimation of bicycle and pedestrian traffic in Blacksburg, VA. Transportation Research Part D: Transport and Environment, 63, 244-260.

Medury, A., Griswold, J. B., Huang, L., & Grembek, O. (2019). Pedestrian Count Expansion Methods: Bridging the Gap between Land Use Groups and Empirical Clusters. Transportation Research Record, 0361198119838266.

Miranda-Moreno, L. F., Nosal, T., Schneider, R. J., & Proulx, F. (2013). Classification of bicycle traffic patterns in five North American Cities. Transportation research record, 2339(1), 68-79.

Nosal, T., Miranda-Moreno, L. F., & Krstulic, Z. (2014). Incorporating weather: comparative analysis of annual average daily bicyclist traffic estimation methods. Transportation Research Record, 2468(1), 100-110.

Oppe, S. (1994). *Guidelines for retrospective safety analysis* (Vol. 93, No. 75).

Proulx, F. R. (2016). Bicyclist Exposure Estimation Using Heterogeneous Demand Data Sources (Doctoral dissertation, UC Berkeley).

Roll, J. F., & Proulx, F. R. (2018). Estimating annual average daily bicycle traffic without permanent counter stations. Transportation research record, 2672(43), 145-153.

Roy, A., Nelson, T. A., Fotheringham, A. S., & Winters, M. (2019). Correcting Bias in Crowdsourced Data to Map Bicycle Ridership of All Bicyclists. Urban Science, 3(2), 62.

State Smart Transportation Initiative. (2014) “The California Department of Transportation: SSTI Assessment and Recommendations”

Strauss, J., & Miranda-Moreno, L. (2013). Spatial modeling of bicycle activity at signalized intersections Jillian Strauss, Luis F Miranda-M.

Strauss, J., Miranda-Moreno, L. F., & Morency, P. (2013). Cyclist activity and injury risk analysis at signalized intersections: A Bayesian modelling approach. Accident Analysis & Prevention, 59, 9-17.

SWOV (2010) The high risk location approach. SWOV Fact sheet, January 2010. www.swov.nl/publicatie/high-risk-location-approach

Tabeshian, M., & Kattan, L. (2014). Modeling nonmotorized travel demand at intersections in Calgary, Canada: use of traffic counts and Geographic Information System data. Transportation Research Record, 2430(1), 38-46.

Turner, B., Breen, J., & Howard, E. (2015). Road safety manual: a manual for practitioners and decision makers on implementing safe system infrastructure. PIARC.

Yang, H., Lu, X., Cherry, C., Liu, X., & Li, Y. (2017). Spatial variations in active mode trip volume at intersections: A local analysis utilizing geographically weighted regression. *Journal of transport geography*, *64*, 184-194.

Appendix 1. New variable categorizations for systemic approach

Table 0-1. Violation categories for bicycle crashes (2013-2017)

Violation Category	Violation Code
Bike equipment	21201
Control violation	21451, 21453, 21456, 21460, 21461, 22450, 22453
Entering from minor facility	21663, 21804
Failure to yield	21800, 21802, 21803, 21950, 21952, 22106
Following too closely	21703
Unsafe door opening	22517
Unsafe driving, overtaking and passing	21202, 21650, 21658, 21750, 21755, 21760
Unsafe speed	22350, 22400
Unsafe turning	21717, 21801, 22100, 22101, 22102, 22103, 22107
Under the influence	21200, 23152

Table 0-2. Categories for party at fault

Party at fault	Statewide type of party at fault
Bicycle	L
Pedestrian	N
Vehicle	A, B, C, D, E, F, G, H, I, J, K, M, O

Table 0-3. Categories for traffic control

Signalization presence	Traffic control category	Traffic control types
Unsignalized	No control	A
Signalized	Two-way stop signs	B, C
Signalized	Four-way stop signs	D, E, F, G
Signalized	Yield signs	H, I
Signalized	Traffic signal	J, K, L, M, N, P
Other	Other	Z, other

Table 0-4. Categories for median barrier type

Median barrier type	Median barrier code
No physical barrier	Z
Raised median	A, B, C, D, E, F, G, H, I, J, K, L, M, N, P, Q, R, S, X
Other	Y, other

Table 0-5. Categories for median presence

Median presence	Median type code
Divided	B, D, E, F, G, H, I, J, K, L, M, N, P, Q, R, S, T, U, V, Z
Undivided	A, C