Multiscale Crash Analysis: A Case Study of Integrating FARS, Maryland's Crash Data, and Montgomery County's Traffic Violation Data

Xiao-Feng Xie

WIOMAX LLC
PO Box 540
Rockville, MD 20848-0540
xie@wiomax.com

Zunjing Jenipher Wang

WIOMAX LLC
PO Box 540
Rockville, MD 20848-0540
wang@wiomax.com

4997 words + 8 figures + 2 tables

November 6, 2017

ABSTRACT

Road safety is a serious issue raising increased public concerns. In this paper, we analyze road safety with an integration of multiple data sources on multiple scales. As a case study, we consider three datasets, including the nationwide Fatality Analysis Reporting System (FARS), the statewide traffic crashes in Maryland (MDCrash), and the countywide traffic violations in Montgomery County, MD (MoCoVio). For data integration, we first exploit basic common characteristics among all the datasets. The time interval statistics of the datasets are found stable and can be modeled into parametric statistical distributions. We then check essential features of the datasets corresponding to road safety and the relationship among them. We also compare the patterns of six common risk factors across all the three datasets. It is found that despite the difference in the features of the datasets, the patterns of DUI/DWI are very similar. Next, we explore practical values of the multiple data integration on road crash analysis. The crash risk patterns extracted from data fusion is shown to be rather valuable. By identifying determinant risk factors in the patterns, we can better understand the effects of other risk factors. In addition, conditional risk matrix can be computed from data integration to measure the probability of the injury levels and to evaluate the impact of each individual risk factor on injuries. Finally, we conduct a multi-source data integration to discover the safety factors for pedestrians, where we obtain temporal patterns from FARS but acquire spatial patterns from the traffic crash and violation data. The results indicate that, in comparison with only using FARS, integrating multiple data has the power of showing more insights of the patterns on risk factors for traffic crashes, which allows us to not only better optimize limited resources but also realize more effective countermeasures for enhancing road safety.

INTRODUCTION

Road safety is a serious threat to public health. Worldwide, there were 1.25 million road traffic deaths in 2013, and up to 50 million more per year suffering injuries as a result of road traffic crashes (*1*). In USA, 35,092 people died and 1.7 million people injured in 6.3 million police-reported traffic crashes in 2015 (*2*). Road crashes cause a significant loss on a national economy output, gross domestic product (GDP), and employment (*3, 4*). In USA, the road crashes in 2010 led to an estimated economic and societal cost of $836 billions (*3*). To reduce road crashes that result in enormous losses to society and economy, we must gather the corresponding risk factors and understand them as they affect the probabilities of road crash, which is also important to establish methods for quantitative estimation and prediction, especially for making policies and evaluating the policy efficiency with countermeasures.

Road accidents are often caused by complex interactions between a variety of risk factors, ranging from human factors, equipment of vehicles, surrounding traffic, roadway and environmental conditions. According to the Fatality Analysis Reporting System (FARS) [1] by NHTSA and other studies (*5*), most fatal crashes are due to some factors based on human choices. Speeding or driving too fast for conditions (*6, 7*) continues to be as a dangerous driving behavior resulting in a large number of fatalities. A higher driving speed increases both the probability to be involved and the injure severity in a crash (*7*). Road crashes often occur at intersections (*8*), where typical risk factors include red-light running (RLR) (*8, 9*) and other sign or signal violation behaviors. It was estimated that there are 260,000 annual RLR crashes in USA (*8*). Many cities have installed red-light cameras as a countermeasure, but whether the installation is an effective means of reducing RLR crashes is still at the research stage (*10*). Another significant risk factor is driving under the influence (DUI) or driving while impaired (DWI) by alcohol, drugs, and medications (*11, 12, 13*). DUI/DWI can impair the driver's cognitive capability and muscle coordination. Crash risk grows exponentially with the increase of blood alcohol concentration (BAC) levels (*14*). Illegal drug usage has been detected in an increasing number of fatal road crashes (*13*). In statistic, DUI/DWI is shown to significantly increase driver injury severity (*12*). Seat belt use (*2, 15*) is proved to significantly reduce both fatal and non-fatal injuries in road crashes for front and rear seat occupants. The nationwide seat belt use rates are 88.5% and 90.1% respectively in 2015 and 2016 (*16*). The remaining 10% unrestrained people face a significant high risk to be easily injured in traffic crashes. There are some other human factors concerning road safety. In distracted driving (e.g., texting while driving) (*17*), crashes are caused as drivers divert their attention to some other activity. In fatigue driving (or drowsy driving) (*18*), crashes are caused as drivers fall asleep even shortly. In addition, some drivers are involved into crashes due to their aggressive driving behaviors (*19*).

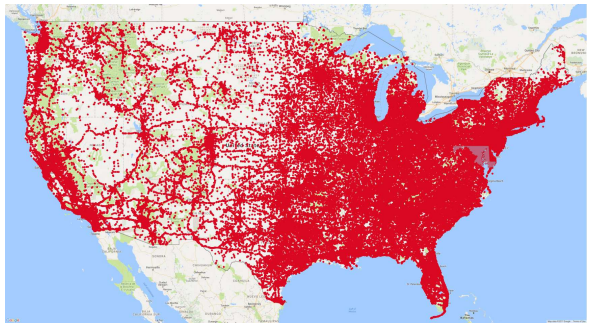
Acquisition of sufficient data resources is one of critical prerequisites for any statistical data analysis, which includes traffic safety studies, to be right. The FARS is a nationwide dataset that has been used in many studies (*2, 11, 13*) for extracting insights about risk factors of fatalities in traffic crashes. Some studies have used statewide dataset (*20, 21*) and countywide dataset (*22*) for exploring local patterns and implementing countermeasures regarding traffic safety. In addition, other data sources like social media have been analyzed to help identifying crashes (*23*). It turns out that each individual dataset has its own limit of application in practice. For example, FARS does not have unbiased sampling on non-fatal crashes, and both statewide and countywide data have low sampling of crashes for certain risk factors.

[1] `https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars`

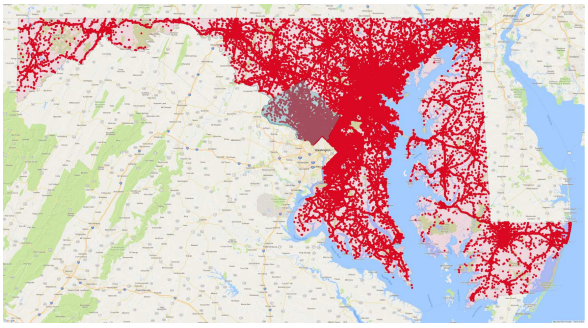
In this paper, we focus on investigating the feasibility of conducting a multi-source data integration for the data resources on national, state, and county scales. A key challenge of data fusion (*24*) or integration lies in building the connections across the different datasets which do not hold a unified terminology. As a case study, here we will integrate three datasets of FARS, the statewide crash data in the State of Maryland (MD), and the countywide moving violations in Montgomery County, MD. Notice that, aiming to improve road safety, Maryland has adopted AASHTO's Toward Zero Deaths (TZD) vision [2], and Montgomery County has adopted the Vision Zero policy recently [3]. Integration of various useful data sources is certainly crucial for acquiring a proper understanding on spatio-temporal patterns of road crashes as a function of different risk factors, which is a premise for implementing countermeasures to reduce traffic crashes and for optimizing the resources to enhance road safety.

FRAMEWORK AND DATA DESCRIPTION

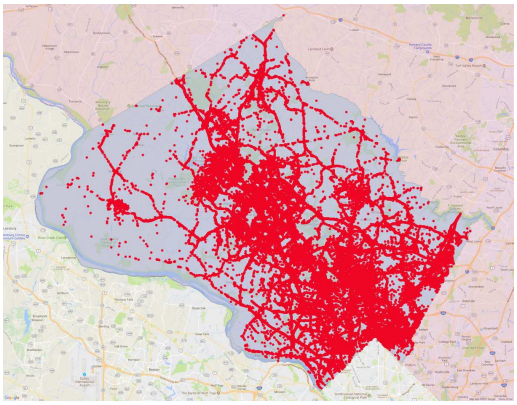
Fig. 1 gives the multi-scale datasets and the data analysis framework following fundamental principles of multi-source data fusion (*24*). The data fusion framework includes three levels of processes. First, multiple data sources are taken as inputs. Next, the data structures are respectively and patterns identified at the intermediate level. Finally, road safety is analyzed through an overall data integration with a combination of all the patterns across various datasets.



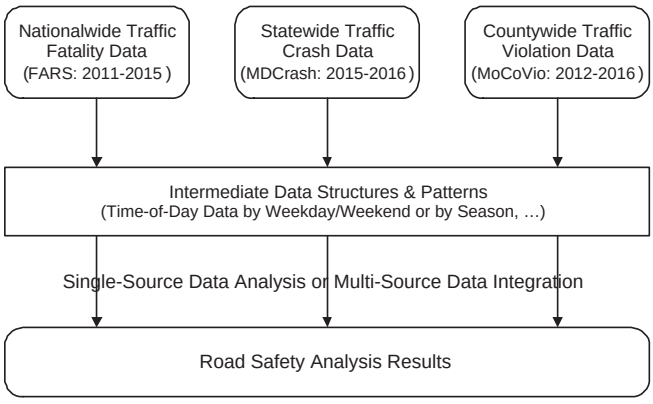
(a) Fatal Traffic Crashes in FARS (2011-2015).



(b) Traffic Crashes in Maryland (2015-2016).



(c) Traffic Violations in MoCo (2012-2016).



(d) Multiscale Data Integration Framework for Crashes.

FIGURE 1 : Multi-Scale Datasets and Integration Framework.

[2] http://towardzerodeathsmd.com

[3] https://montgomerycountymd.gov/COUNCIL/Resources/Files/res/2016/20160202_18-390.pdf

We consider three datasets, which are the Fatality Analysis Reporting System (FARS), the Maryland Statewide Vehicle Crash Data (MDCrash), and the Montgomery County Traffic Violation Data (MoCoVio), respectively on the scales of nationwide, statewide, and countywide. Maintained by the National Highway Traffic Safety Administration (NHTSA), FARS is a nationwide database regarding fatal injuries suffered in motor vehicle traffic crashes. The data is available since 1975. For FARS, we use the recent data in the years from 2011 to 2015. Within the five years, 378,429 persons in total (75,686 per year) involved in 153,297 crashes (30,659 per year), where 166,990 (33,398 per year) persons lost their lives. Maintained by the Department of Maryland State Police, MDCrash contains all approved crash reports in Maryland. The data is available since 2015. For MDCrash, we use all the data of the two years, 2015 and 2016. Within the two years, 500844 persons in total involved in 228710 crashes, where 38031, 52577, 5715, and 977 people respectively suffered non-incapacitating, possible incapacitating, incapacitating/disabled, and fatal injuries. Maintained by the Department of Police of Montgomery County (MoCo) in Maryland, MoCoVio includes information of all the electronic traffic violations issued in the county. For MoCoVio, we use the data in the recent five years from 2012 to 2016. Within the five years, the dataset contains 1.01 million traffic violation records of the county.

Among the three datasets, the FARS is the most well-documented and well-studied data, with which many significant insights about fatal crashes have been extracted by previous studies. NHTSA has an agreement with each state to share information in a standard format on fatal crashes in the state, but without a standard data format on non-fatal crashes. On county scale, traffic violation dataset is not designed for crash studies, and some outliers could be included in the county data. During the preprocess of data, we removed 7401 records of violations outside of the county, and corrected 77915 null records of geolocations using the Google Places API [4].

The three datasets are different in collected information and samples. FARS is designed mainly for fatal crash data. Although some non-fatal crash data are included in FARS, their samples are biased towards the fatal ensemble as each crash in FARS must have at least one fatal person. On the contrary , the samples of the traffic crash data in MDCrash are unbiased across all injury levels, including both non-fatal and fatal crashes. MoCoVio is a dataset on traffic violations, where some of the violations led to traffic crashes, but not all of them.

RELATIONS BETWEEN DIFFERENT DATASETS

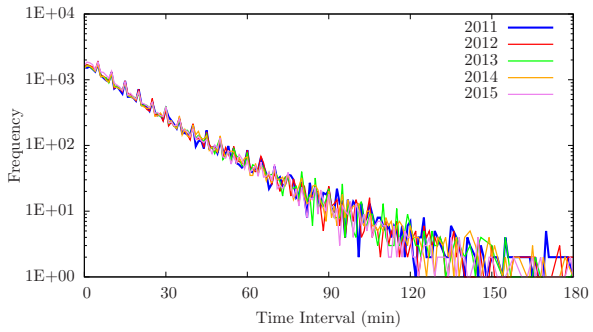
In this section, we explore the connections among different datasets. For convenience, we call each instance in the three datasets as an *record* of fatality crash, traffic crash, and traffic violation respectively.

Time Interval Estimation for Traffic Crashes/Violations

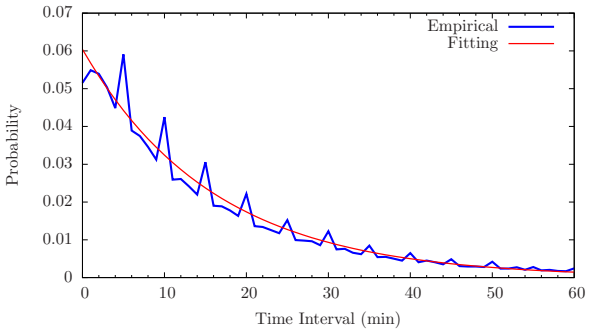
Fig. 2 gives the statistical analysis on time intervals between two continuous record occurrences for FARS, MDCrash, and MoCoVio respectively. For each dataset, all its crash/violation time records can be shown with an ordered set of time series (x_i) from earliest to nearest time, where x_i is the time at which the ith record occurs. We define t_i as the time interval from the $(i-1)$th record to the ith record, i.e. $t_i = x_i - x_{i-1}$. The left column of Fig. 2 gives the frequency of t_i for all the three datasets, where we show statistical results yearly for each dataset to check if the distribution of t_i is a function of index of year. It turns out that the distribution of t_i is almost independent of

[4] https://developers.google.com/places/

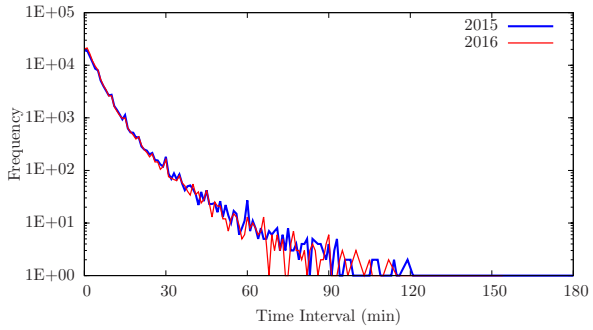
year index for all the three datasets. Therefore we can compute average distribution of t_i over all the years for each dataset. The right column of Fig. 2 gives the calculated (empirical) probability density functions (PDF) of t_i for all the three datasets (blue). It shows that the distribution of t_i can be well fitted to an exponential function with a single parameter (red), using the Cooperative Group Optimization (CGO) (25, 26) minimizing the least squares. Notice that the jigsaw curves in Fig. 2b result from the customized rounding treatment by 5-minutes on record time x_i rather than from any essential features of crashes, therefore overlooking them for fitting is suitable. Acquiring time interval distribution for traffic crashes/violations is important, as it enables us to estimate the probability of a next record occurring for a given time period, or the probability of time intervals from this crash/violation to the next crash/violation occurring in the region.



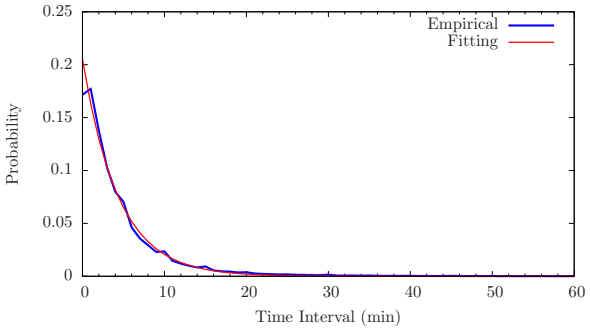
(a) FARS: Frequency of Crash Time Intervals.



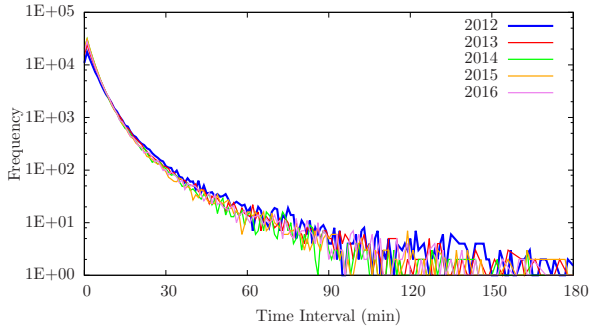
(b) FARS: Empirical PDF & Fitting.



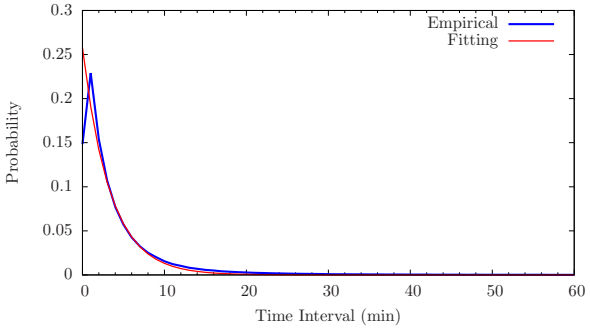
(c) MDCrash: Frequency of Crash Time Intervals.



(d) MDCrash: Empirical PDF & Fitting.



(e) MoCoVio: Frequency of Record Time Intervals.



(f) MoCoVio: Empirical PDF & Fitting.

FIGURE 2 : Statistics of Time Intervals of FARS, MD Crash Data, and MoCoVio Data.

Time-of-Day (ToD) Crash Patterns

We first study Time-of-Day (ToD) patterns of road crashes and violations at the intermediate interface of framework. Two types of hourly averaged Time-of-Day (ToD) rates are used, where one is organized by weekday/weekend, and another by seasons. The seasons by the months are respectively Spring (March, April, May), Summer (June, July, August), Autumn (September, October, November), and Winter (December, January, Feburary), defined according to the monthly average temperature, which are respectively 56.33°F, 78.67°F, 60.83°F, and 37.83°F in Maryland [5].

Concerning the analysis at the intermediate level of framework, in order to compare results among the three datasets with different periods of time, we calculate the hourly ToD rates, which is defined as the hourly count of road crashes/violations averaged by the total number of associated days, for each dataset. The left and right columns of Fig. 3 give the hourly ToD rates for each database respectively by weekday/weekend and by seasons.

Overall, the patterns are very different among the three datasets. Fig. 3a from the FARS data shows a dramatic increase in road crashes between 0-4 AM for a weekend, and a crash peak around the time period at sunset both for a weekday and weekend. Instead, Fig. 3c from the MDCrash data shows a pattern regarding the commuting between home and workplace, where two peaks of crashes at AM and PM respectively can be clearly identified for a weekday. This is because that Figs. 3c and 3d contain all crashes including no injury, non-incapacitating, possible incapacitating, incapacitating/disabled, and fatal injuries, while Figs. 3a and 3b include only fatal crashes. Notice that Figs. 3e and 3f give the patterns for all violations, rather than limited to crashes. The difference in the ToD patterns among the three datasets in fact reflects the distinction in risk patterns among the three traffic datasets, which are respectively fatal crashes, all crashes and traffic violations.

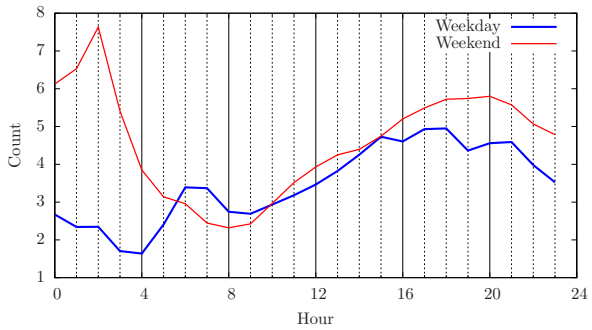
Details in Data Relations

In order to discover more patterns at the intermediate interface of framework, we need analyze details of data to obtain the connections among different datasets.

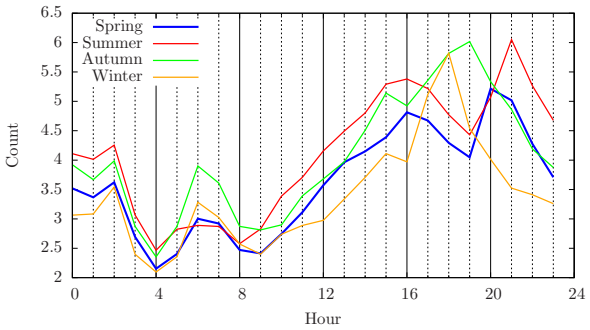
Traffic Crashes in MDCrash

Fig. 4 gives the average hourly crash rates in MDCrash at different injury levels. It is found that more people were involved in non-incapacitating or possible incapacitating injury than in incapacitating/disabled and fatal injuries. The numbers of crash counts with non-incapacitating or possible incapacitating injury differ by one order of magnitude from incapacitating/disabled injury, and by around two orders of magnitude from fatal injury. Although both Figs. 4a and 4b show a commuting pattern as found in Fig. 3c, their peak crash counts are different. Both Fig. 4d and Fig. 3a describe the fatal crash pattern, though using different datasets, MDCrash and FARS respectively. Notice that the total number of samples in the State of Maryland is much smaller than that of nationwide dataset of FARS, therefore, the results show apparent noises in both Figs. 4c and 4d. By taking into account the difference in sampling noises, we can see that the fatal crash pattern shown by Fig.4d has some similarities with the result in Fig. 3a. In addition, we check detailed numbers of data to explore the connections between MDCrash and FARS datasets. Both FARS and MDCrash have collected data for the fatal crashes in 2015. It is found that the numbers of fatal injuries in Maryland are consistent between the datasets (513 in FARS and 517 in MDCrash).

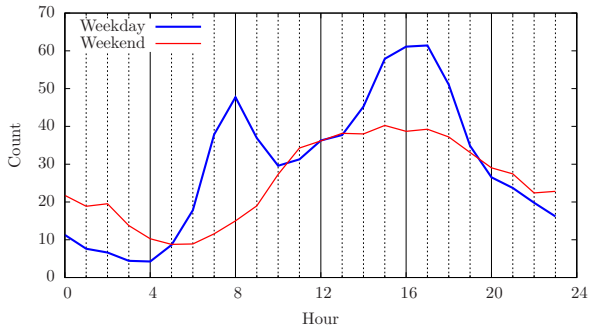
[5] `http://www.usclimatedata.com/climate/maryland/united-states/1872`



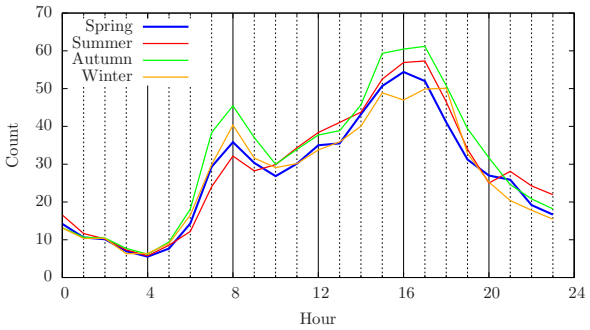
(a) FARS Fatal Rate by Weekday & Weekend.



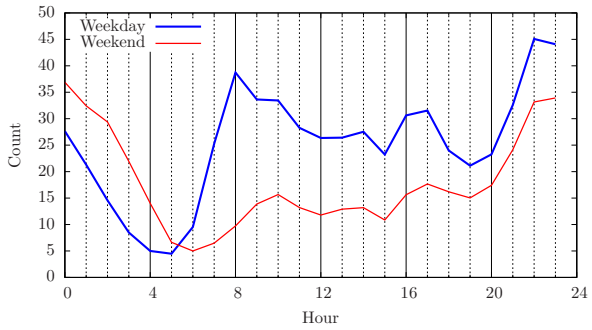
(b) FARS Fatal Rate by Season.



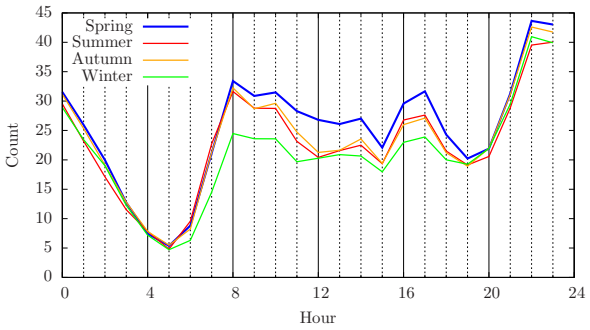
(c) MD Crash Rate by Weekday & Weekend.



(d) MD Crash Rate by Season.



(e) Violation Rate in MoCo by Weekday & Weekend.



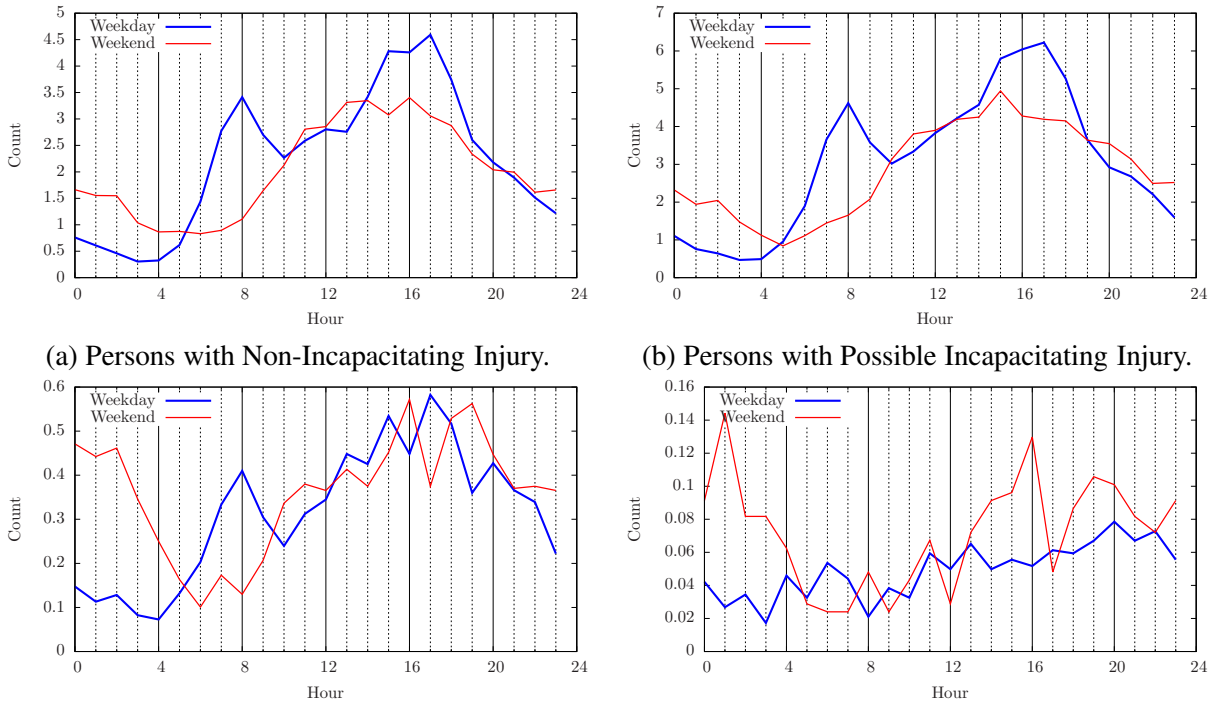
(f) Traffic Violation Rate in MoCo by Season.

FIGURE 3 : Temporal Distributions of Records in FARS, MDCrash, and MoCoVio.

Traffic Violations in MoCoVio

In MoCoVio, the dataset of traffic violations in Montgomery County of MD, each traffic violation is defined by the Transportation Article of the Code of Maryland [6]. Table 1 gives the description and statistics of traffic violations of significant charge classes. Class 21 is a main class regarding moving violations. Classes of 13, 16, and 17 are respectively the violations corresponding to certificates of title and registration of vehicles, driver's license, and insurance. Class 22 is concerning the unsafe equipment of vehicles, e.g., brake, lighting, and seat belt use. Notice that the MoCoVio dataset does not include a complete set of traffic crashes, as many police-reported crashes do not involve any traffic violation (*27*). For example, according to FARS, among 32,166 fatal crashes in 2015, only 5220 crashes were reported to involve traffic violations. The MoCoVio dataset only provides a small subset of the traffic crashes occurring in the county, which contain the features of

[6]Code of Maryland: `http://mgaleg.maryland.gov/webmga/frmStatutes.aspx`



(a) Persons with Non-Incapacitating Injury.

(b) Persons with Possible Incapacitating Injury.

(c) Persons with Incapacitating/Disabled Injury.

(d) Persons with Fatal Injury.

FIGURE 4 : Crash Rates for Different Injury Levels in MDCrash.

"Property Damage", "Contribute To Accident", "Personal Injury", and "Fatal Injury".

As shown in Table 1b, the classes of 21-2 and 21-8 both have a large number of violation counts. Both of the two classes of enforcement could use cameras, which are widely installed red-light cameras speed cameras respectively. One common example of class 21-2 violation is red-light running (RLR) (*9, 28*) at signalized intersections. In Table 1b, the violation classes of 13, 16, and 22 also have high numbers of counts. The classes of 20-8, 20-9 have high crash ratios on injury and fatal crashes. Notice that 65.72% (36925/56182) violations in Class 20-9 are also under the subclass 21-902, which is regarding driving under the influence (DUI) or driving while impaired (DWI) by alcohol, drugs, or medication. Inside Class 21, the sub-classes of 21-9, 21-8, 21-4, 21-2, and 21-3 often have the highest crash ratios. In addition, 69.31% (25259/36442) of violations in Class 21-11 are corresponding to texting while driving, which is a typical behavior of distraction while driving (*17*). Class 21-5, the charge class for pedestrians, accounts for only 0.72% of violations, but 5.45% of fatal injuries. Few violations were issued on the classes of 21-12 and 21-13 in MoCoVio, which are respectively for bicycles and motorcycles.

Fig. 5 gives the comparison of the crash rates between a weekday and weekend from MoCoVio for all types of crash features. Figs. 5a, 5b and 5c show the results of the violations with the crash features of "Contribute To Accident", "Property Damage", "Personal Injury" respectively, which all display an apparent commuting pattern that is similar to the crash patterns from MDCrash data as shown in Fig. 4. The result of the violations associated with "Fatal Injury" crash in Fig. 5d, however, is very noising due to its small sample size in MoCoVio dataset. During 2015-2016, there were 27419 crashes in Montgomery County, according to MDCrash data, but there were only 17969 moving violations associated with traffic crashes, according to MoCoVio

TABLE 1 : Description and Statistics of MoCo Traffic Violations in Charge Classes.

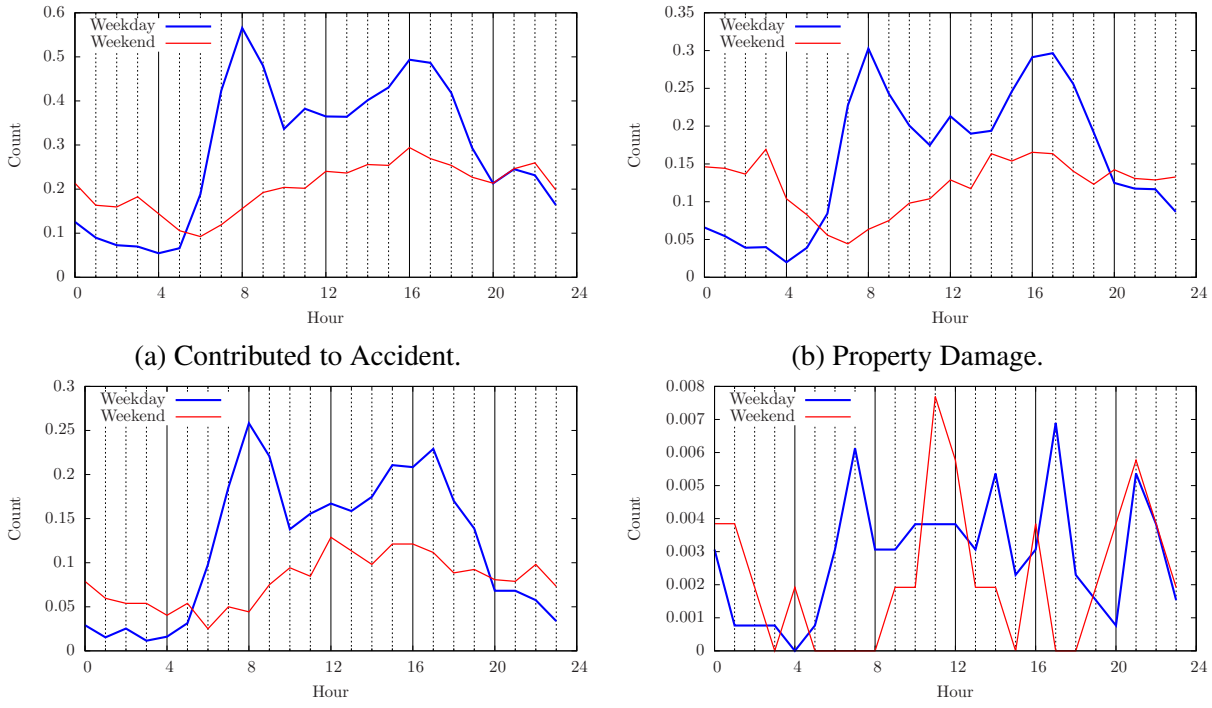
(a) Description of Charge Classes

Charge	Description
21-2	Traffic Signs, Signals & Markings
21-3	Driving on Right Side Of Roadway; Overtaking & Passing; Use of Roadway
21-4	Failed to Yield Right of Way (ROW)
21-5	Pedestrian's Rights And Rules
21-6	Turning & Starting; Signals on Stopping, Turning & Starting
21-7	Special Stops Required
21-8	Speed Restrictions
21-9	1) Reckless, Negligent, Aggressive; 2) Impaired Driving; 4) Fleeing or Eluding Police
21-10	Stopping, Standing, And Parking
21-11	Miscellaneous Rules (e.g., Texting While Driving, Windshield Obstructions)
21-12	Operation of Bicycles & Play Vehicles
21-13	Operation of Motorcycles
13	Certificates of Title and Registration of Vehicles
16	Driver's License
17	Required Security (Insurance)
20	Accidents & Accident Reports
22	Equipment of Vehicles (Brake, Lighting, Seat Belt Use, $\cdots$)

(b) Data Statistics in Charge Classes

	All Violations		Contrib. To Accident		Personal Injury		Fatal Injury	
Charge	Count	Ratio (%)	Count	Ratio (%)	Count	Ratio (%)	Count	Ratio (%)
21-2	**125896**	12.43	1637	6.97	791	6.74	**16**	7.27
21-3	38741	3.82	**2091**	8.90	832	7.09	14	6.36
21-4	20376	2.01	**2494**	10.61	**1305**	11.12	**16**	7.27
21-5	7303	0.72	358	1.52	335	2.85	12	5.45
21-6	8145	0.80	180	0.77	86	0.73	3	1.36
21-7	34772	3.43	254	1.08	111	0.95	4	1.82
21-8	**167462**	16.53	**5972**	25.42	**2815**	23.99	**35**	15.91
21-9	56182	5.54	**4740**	20.17	**2265**	19.30	**41**	18.64
21-10	5938	0.59	31	0.13	6	0.05	0	0.00
21-11	36442	3.60	483	2.06	120	1.02	2	0.91
21-12	151	0.01	21	0.09	18	0.15	2	0.91
21-13	395	0.04	11	0.05	15	0.13	0	0.00
13	**171813**	16.96	669	2.85	498	4.24	**23**	10.45
16	**147060**	14.51	1854	7.89	**1389**	11.84	**23**	10.45
17	8784	0.87	79	0.34	66	0.56	4	1.82
20	7787	0.77	**1974**	8.40	**895**	7.63	7	3.18
22	**103316**	10.20	375	1.60	119	1.01	12	5.45
Others	72640	7.17	274	1.17	69	0.59	6	2.73

data.



(a) Contributed to Accident.

(b) Property Damage.

(c) Personal Injury.

(d) Fatal Injury.

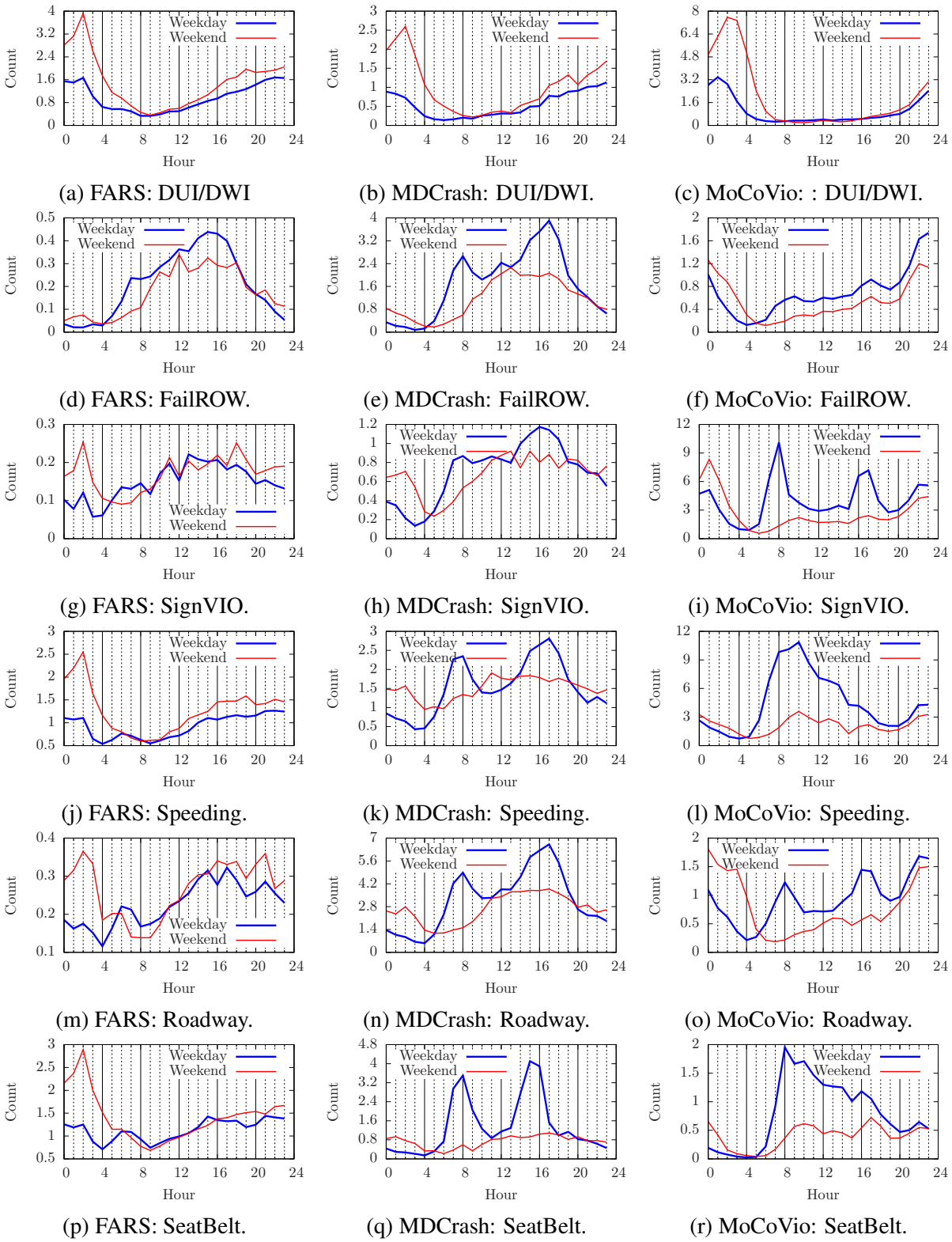
FIGURE 5 : MoCoVio Traffic Violations Related to Crashes.

Major Common Risk Factors

Here we discuss and compare common risk factors across all three datasets to understand the similarity and difference in the impacts of these risk factors. Notice that some datasets might do not contain a sufficient large size of sampling data for analyzing certain risk factors. On MoCoVio, we consider the following six common risk factors based on their charge classes (see Table 1a):

- DUI/DWI: It is corresponding to Class 21-902, a main subclass of Class 21-9. This factor means the driving under the influence (DUI) or driving while impaired (DWI).
- FailROW: It includes both Class 21-4 and Class 21-6.
- SignVIO: It includes both Class 21-2 and Class 21-7.
- Speeding: It means violating speed restrictions in Class 21-8.
- Roadway: it means being illegal on roadway as defined in Class 21-3.
- SeatBelt: It is Class 22-412 (a subclass of Class 22), which is about unsafe seat belt use.

On the other two datasets, FARS and MDCrash, we first identify the correlations between their crash features and the risk factors defined by the Transportation Article of MoCoVio, then classify them accordingly into the most correlated one of the six common risk factors in use for MoCoVio. For example, the crashes featuring “Under Influence of Drugs”, “Under Influence of



(a) FARS: DUI/DWI (b) MDCrash: DUI/DWI. (c) MoCoVio: : DUI/DWI.

(d) FARS: FailROW. (e) MDCrash: FailROW. (f) MoCoVio: FailROW.

(g) FARS: SignVIO. (h) MDCrash: SignVIO. (i) MoCoVio: SignVIO.

(j) FARS: Speeding. (k) MDCrash: Speeding. (l) MoCoVio: Speeding.

(m) FARS: Roadway. (n) MDCrash: Roadway. (o) MoCoVio: Roadway.

(p) FARS: SeatBelt. (q) MDCrash: SeatBelt. (r) MoCoVio: SeatBelt.

FIGURE 6 : Common Risk Factors for FARS, MDCrash and MoCoVio.

Alcohol", "Under Influence of Medication", "Under Combined Influence Physical/Mental Difficulty" in MDCrash are all considered as DUI/DWI, and the fatal crashes featuring "Under the Influence of Alcohol, Drugs or Medication" in FARS are also classified as DUI/DWI.

The left, middle and right columns of Fig. 6 give the hour-of-day distributions of the crashes/violations featuring the six risk factors respectively for FARS, MDCrash, and MoCoVio. It turns out that the traffic crashes or violations featuring DUI/DWI show very similar patterns in their distributions for all the three datasets. However, the crashes/violations featuring the other five risk factors show rather different patterns for the three datasets. For MDCrash, a commuting pattern with the AM and PM peaks is observable on all the other five risk factors. For MoCoVio, a commuting pattern is only shown on two factors of SignVIO and Roadway. It is worth to be aware that a significant amount of traffic crashes/violations happened during night.

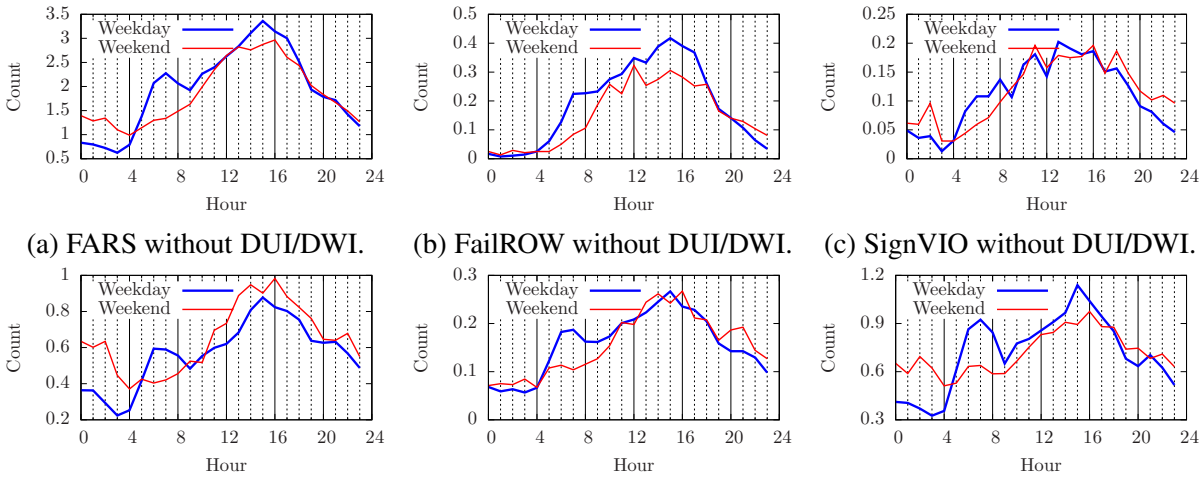
PRACTICAL VALUES OF THE MULTIPLE DATA INTEGRATION

In this section, we discuss the values of the multiple data integration for improving traffic safety. In details, we explore how we take advantage from multiple data fusion to gain the crash patterns under various or certain traffic conditions and then to use the knowledge to reduce traffic crashes.

Determinant Risk Factors

Recognition of determinant risk factors is important for getting solutions to reduce traffic crashes. However, it is difficult, as each crash often involves multiple risk factors, i.e. risk factors are often entangled together while inducing the crashing. In order to separate the entangled effects of multiple risk factors on crashes or to identify the determinant risk factors, we need study the dependence of crash patterns on different risk factors. A straightforward way to judge if one single risk factor is entangled with others in data analysis is through removing the effects of the risk factor on crashes. By checking whether the crash distributions featuring the other risk factors would change with the removed factor, we can judge if the other risk factors are entangled with the removed one. In addition, if the removed factor is a determinant risk factor, most crash patterns of all the other risk factors would be significantly changed by its exclusion, and the changed crash patterns would display the true crash peak time induced by the their featuring risk factors. Moreover, the crash pattern of a determinant risk factor is little changed by other risk factor, therefore a unitary pattern is often shown on a determinant risk factor even across different datasets.

As shown in Fig. 6, the crash patterns of DUI/DWI are all similar across the three datasets. Thus, we hypothesis that DUI/DWI is likely to be a determinant risk factor. To check if it is right, we remove the effects of DUI/DWI (as shown in Fig. 6a) from the total FARS data (as shown in Figs. 3a) and the data of other five risk factors in FARS (as shown in 6d, 6g, 6j, 6m, and 6p in the first column of Fig. 6). We show the results in Fig. 7. As shown in the figure, all the crash patterns of the other risk factors, including that with all risk factors and those respectively featuring one of the other five risk factors, have significant changes after the exclusion of DUI/DWI effects. This confirms that DUI/DWI is indeed a determinant risk factor for road crashes. The crash count peaks are all significantly reduced as shown in Fig. 7, between 0-4AM in the Fig. 3a for all fatal crashes, Fig. 6g for SigVIO crashes, Fig. 6j for Speeding crashes, Fig. 6m for Roadway crashes, and Fig. 6p for SeatBelt crashes. This means that the crash peaks are all caused by the determinant risk factor DUI/DWI. The pattern in Fig. 6d for FailROW crashes has no any apparent change with the exclusion of DUI/DWI effects, referring to an independent relation between FailROW and DUI/DWI in FARS crash data.



(a) FARS without DUI/DWI. (b) FailROW without DUI/DWI. (c) SignVIO without DUI/DWI.

(d) Speeding without DUI/DWI. (e) Roadway without DUI/DWI. (f) SeatBelt without DUI/DWI.

FIGURE 7 : Total FARS Data (a) and Other Five Risk Factors in FARS (b-f) without DUI/DWI.

Conditional Risk Matrix

Although FARS provides rather complete information for fatal crashes, it only contains a very small number of biased samples for non-fatal crashes, as each crash in FARS must involve fatalities. As a complement, MDCrash includes unbiased information for all crashes in Maryland. Integrating the information from both FARS and MDCrash, we can compute *Conditional Risk Matrix* as a function of the two variables of risk factor and injury level, where each value of matrix for certain injury level and risk factor represents the probability of crashes occurring at the injury levels and with the risk factor as a leading cause.

Table 2 gives the conditional risk matrix for 3 serious injury levels and 6 risk factors. For each injury level, the factors with the top three highest ratios (i.e. probabilities) are marked in bold. It is shown that SeatBelt, DUI/DWI and Speeding, are the top three risk factors resulting in fatal or serious injuries. In 2015 and 2016, the seatbelt use rates of Maryland were respectively 92.9% and 90.8%, slightly higher than the nationwide rates of 88.5% and 90.1% (*16*).

As shown in Table 1b, DUI/DWI and Speeding are also two significant risk factors causing crashes in MoCoVio. The sample size of crashes is often small at a county level. Data fusion would be very useful for extracting conditional risk matrix when we aim to construct a safety performance function (SPF) (*29*) evaluating the dependence of the crash probabilities on certain significant risk factors, especially for a city or county with a limited number of samples.

Pedestrian Safety

In this section, we perform multi-source data fusion to gain a better understanding on the safety factors for pedestrians. In USA, there were 5,376 pedestrians killed (15.3% of total deaths) and an estimated 70,000 injured in traffic crashes in 2015. In Maryland, there were 175 pedestrians killed, which is 17.9% of total 977 deaths in 2015-2016. In Montgomery County, there were 19 pedestrians killed, which is 22.0% of total 85 deaths in 2015-2016.

Figs. 8a and 8b give the hour-of-day distributions of pedestrian deaths in FARS, respectively by weekday/weekend and seasons. For both a weekday and weekend, the highest main count peak locates at 17-22 PM. The second count peak appears at 0-4AM for a weekend while 4-8 AM

TABLE 2 : Conditional Risk Matrix for Three Injury Levels and Six Risk Factors.

Risk Factor	Total Persons	Fatal		Incapacitating/Disabled		Possible Incapacitating	
		Count	Ratio	Count	Ratio	Count	Ratio
DUI/DWI	11799	57	**0.0048**	409	**0.0347**	1424	**0.1207**
FailROW	27110	44	0.0016	326	0.0120	2698	0.0995
SignVIO	12128	42	0.0035	239	0.0197	1466	**0.1209**
Speeding	26468	164	**0.0062**	605	**0.0229**	3016	0.1139
Roadway	52816	238	0.0045	883	0.0167	4462	0.0845
SeatBelt	20674	243	**0.0118**	718	**0.0347**	2624	**0.1269**

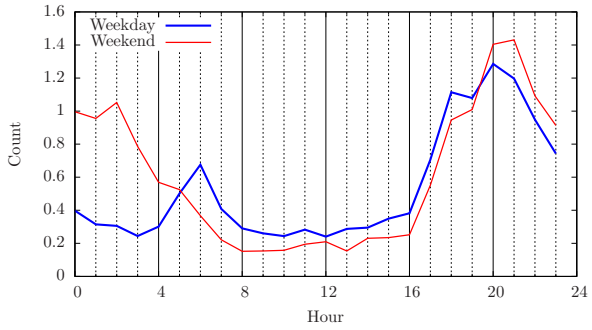
for a weekday. The distributions by seasons indicate that the main highest count peak starts around the sunset time for all the seasons, i.e. when drivers and pedestrians start to have difficulty in seeing each other. In addition, by analyzing MDCrash, we find that the fatal crash rate for the pedestrian wearing dark clothing is much higher than wearing light clothing, which are respectively 4.69% (94/2004) and 1.14% (13/1139).

Fig. 8c shows the spatial distributions of pedestrian crashes in MDCrash (yellow triangles) and of pedestrian-related traffic violations in MoCoVio (blue dots). On pedestrian crashes, 585 crashes at all serious injury levels are shown in Fig 8c, including possible incapacitating, incapacitating/disabled, and fatal injuries. The MoCoVio dataset follows the law of Maryland, where Class 21-5 is specifically used to define Pedestrian's Rights and Rules, also some rules in other classes such as 21-2 are related to pedestrian safety. In MoCoVio, 3969 and 5734 traffic violations in total are respectively for drivers and pedestrians. Fig. 8c indicates that pedestrian crashes and traffic violations are largely spatial overlapped, referring to a strong correlation between them.

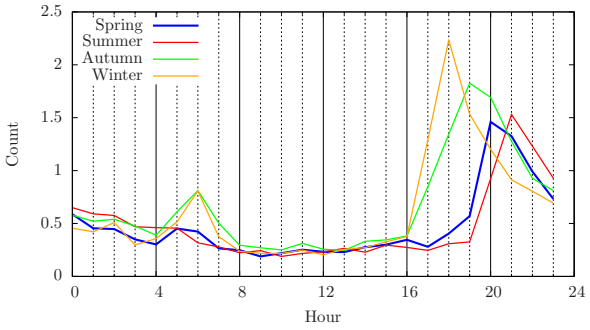
Fig. 8d gives both the heatmap of pedestrian crashes in MDCrash and the hot spots of pedestrian-related traffic violations in MoCoVio, where the heatmap is obtained using Kernel Density Estimation, and the hot spots with 99% confidence are obtained using Getis-Ord Gi* statistics (*30*). Again, it shows that all the hot spots of traffic violations overlap with the major peaks in the heatmap of crashes. Moreover, the hot spots of pedestrian-related traffic violations as shown in Fig. 8d have covered all the 10 formal High Incidence Areas (HIA) identified by Pedestrian Safety Initiative of Montgomery County with a countywide pedestrian crash data (*22*). To see more details on the spacial distribution of the pedestrian-related traffic violations, we zoom in and show them for two communities of the county, Silver Spring and Bethesda respectively, in Figs 8e and 8f. Notice that there are four traditional HIAs respectively locating at Georgia Avenue and Colesville Road in Silver Spring, and at Wisconsin Avenue and Old Georgetown Road in Bethesda. Enforcement efforts for reducing traffic accidents were extensively applied in HIAs, such as the HIAs in the two communities (*22*) shown in Figs 8e and 8f. However, traffic crashes are still often found to happen close to HIAs, especially for locations with less enforcement. It would help reducing crashes by optimizing the locations for enforcement and engineering efforts in or near some HIAs.

As the noncompliance behaviors of road users concerning the execution of law enforcement, traffic violations bring significant risks to traffic safety, thus have been regarded as a surrogate safety factor in statistical analysis on probability of crashes (*31, 32*). Enforcement, education and engineering efforts are shown to be most effective three ways for reducing traffic crashes (*22*). Integrating traffic violation and crash data in analysis would help us optimize our resources for

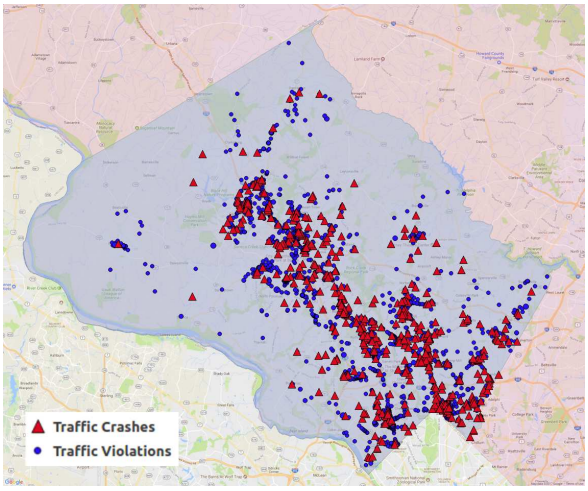
enforcement, education, and engineering corresponding to various time and locations. As a typical engineering effort, smart multimodal intersection control is able to reduce pedestrian wait time while no interruption to vehicle flow (*33, 34*), therefore very useful for improving pedestrian safety.



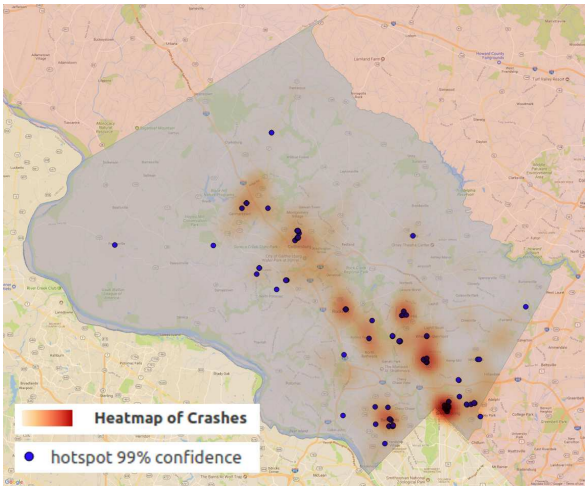
(a) FARS Pedestrian Deaths by Weekday/Weekend.



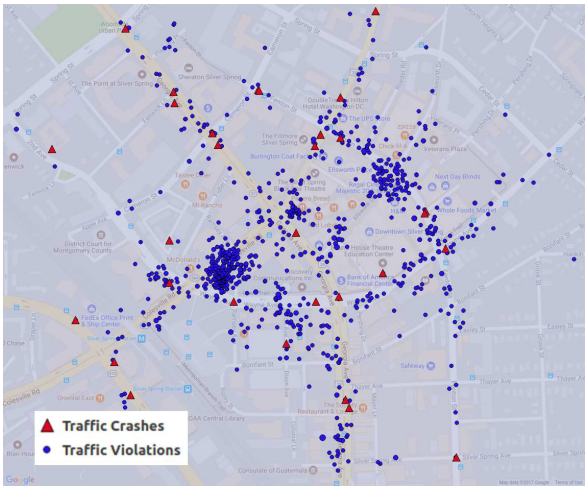
(b) FARS Pedestrian Fatalities by Seasons.



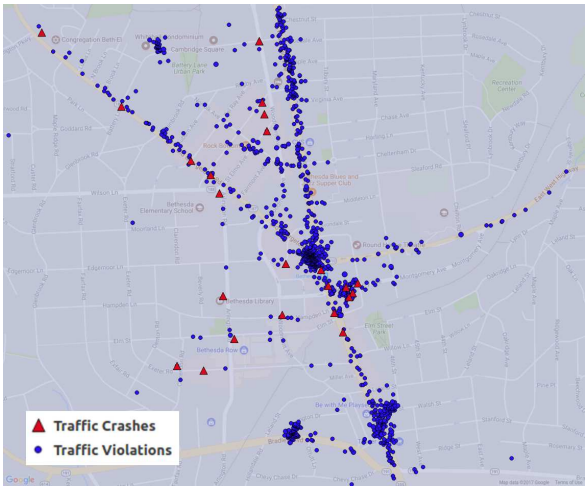
(c) Pedestrian-Related Traffic Crashes & Violations.



(d) Traffic Crash Heatmap and Violation Hot Spots.



(e) Pedestrian Related Violations in Silver Spring.



(f) Pedestrian Related Violations in Bethesda.

FIGURE 8 : Pedestrian-Related Traffic Crashes and Violations.

CONCLUSION

In this paper, we conducted a multi-scale data integration to discuss improving road safety, where the analyzed crash datasets include national FARS, the statewide crash data in Maryland (MDCrash), and the countywide moving violation data in Montgomery County, MD (MoCoVio).

We first examined basic characteristics of all the three datasets. The statistical analysis on time intervals between two continuous crash occurrences showed that the distributions are independent of year index for all the three datasets, and each distribution can be fitted to an exponential function with a single parameter. Next, we checked the data ensembles and their relations among the three datasets. FARS includes all national fatal crashes and a small number of records on the non-fatal crashes biased towards fatal injuries. Compared to FARS, MDCrash contains unbiased records on all traffic crashes for all injury levels in Maryland, but it has a very small sample size for both incapacitating and fatal injuries since it is a statewide rather than nationalwide dataset. MoCoVio contains all traffic violation records regarding traffic crashes in Montgomery County, but only a portion of total crashes in the county, as compared with MDCrash. Finally, we investigated the patterns of six common risk factors across the three datasets via data fusion. The results showed that despite the difference in characteristics and ensembles among the three datasets, their patterns are all similar in the crash distributions on DUI/DWI.

Furthermore, we explored practical values of the multiple data integration for helping road crash reduction. First, we showed that the crash patterns extracted from data integration is able to reveal determinant risk factor of road crashes, and allow us to disentangle the other risk factors from the determinant one for their true effects. Second, we computed conditional risk matrix from data fusion, which is a probability matrix of crashes. As a function of the two parameters of injury level and risk factor, conditional risk matrix enables us to evaluate the significance of each risk factor. DUI/DWI, speeding and seat belt use were found to be top important factors leading to serious injuries. Finally, we used the multi-source data integration to analyze safety factors for pedestrians. Temporal analysis on pedestrian crashes in FARS disclosed that fatal crashes occur in the highest probability when drivers and pedestrians have difficulty in seeing each other. Spatial analysis on pedestrian crashes in MDCrash and pedestrian-related traffic violations in MoCoVio showed that the hot spots of traffic violations cover all the high incidence areas of crashes. By zooming in to check the locations of pedestrian crashes and pedestrian-related traffic violations, we found that enforcement efforts did reduce pedestrian crashes in some areas but not optimized to arrive the best performance. Optimization of locations for enforcement and engineering efforts would be very helpful for reducing crashes. Compared to use only FARS, integration of multiple data empowers us to gain deeper and more complete insights about crash patterns, traffic risk factors, and their relations, which is extremely valuable for the optimization of limited resources and for the realization of countermeasures regarding road safety enhancements.

REFERENCES

[1] World Health Organization, *Global Status Report on Road Safety 2015*. WHO, 2015.

[2] NHTSA. *Traffic Safety Facts 2015*. Tech. Rep. HS812382, U.S. Department of Transportation, 2015.

[3] Blincoe, L., T. R. Miller, E. Zaloshnja, and B. A. Lawrence. *The Economic and Societal Impact of Motor Vehicle Crashes, 2010*. Tech. Rep. DOT-HS-812-013, NTSHA, Washington, DC, 2015.

[4] Zaloshnja, E., T. R. Miller, and B. A. Lawrence. Economics of alcohol-involved traffic crashes in the USA: an input-output analysis. *Injury prevention*, Vol. 22, No. 1, 2016, pp. 19–24.

[5] Petridou, E. and M. Moustaki. Human factors in the causation of road traffic crashes. *European Journal of Epidemiology*, Vol. 16, No. 9, 2000, pp. 819–826.

[6] Viallon, V. and B. Laumon. Fractions of fatal crashes attributable to speeding: Evolution for the period 2001–2010 in France. *Accident Analysis & Prevention*, Vol. 52, 2013, pp. 250–256.

[7] Aarts, L. and I. Van Schagen. Driving speed and the risk of road crashes: A review. *Accident Analysis & Prevention*, Vol. 38, No. 2, 2006, pp. 215–224.

[8] Retting, R. A., R. G. Ulmer, and A. F. Williams. Prevalence and characteristics of red light running crashes in the United States. *Accident Analysis & Prevention*, Vol. 31, No. 6, 1999, pp. 687–694.

[9] Xie, X.-F. and Z.-J. Wang. Integrated in-vehicle decision support system for driving at signalized intersections: A prototype of smart IoT in transportation. In *Transportation Research Board (TRB) Annual Meeting*, Washington, DC, 2017, 17-0671.

[10] Langland-Orban, B., E. E. Pracht, J. T. Large, N. Zhang, and J. T. Tepas III. Explaining differences in crash and injury crash outcomes in red light camera studies. *Evaluation & the Health Professions*, Vol. 39, No. 2, 2016, pp. 226–244.

[11] Slater, M. E., I.-J. P. Castle, B. K. Logan, and R. W. Hingson. Differences in state drug testing and reporting by driver type in US fatal traffic crashes. *Accident Analysis & Prevention*, Vol. 92, 2016, pp. 122–129.

[12] Behnood, A. and F. L. Mannering. The effects of drug and alcohol consumption on driver injury severities in single-vehicle crashes. *Traffic Injury Prevention*, Vol. 18, No. 5, 2017, pp. 456–462.

[13] Brady, J. E. and G. Li. Trends in alcohol and other drugs detected in fatally injured drivers in the United States, 1999–2010. *American Journal of Epidemiology*, Vol. 179, No. 6, 2014, pp. 692–699.

[14] Compton, R. P. and A. Berning. Drug and alcohol crash risk. *Journal of Drug Addiction, Education, and Eradication*, Vol. 11, No. 1, 2015, p. 29.

[15] Høye, A. How would increasing seat belt use affect the number of killed or seriously injured light vehicle occupants? *Accident Analysis & Prevention*, Vol. 88, 2016, pp. 175–186.

[16] Pickrell, T. M. *Seat Belt Use in 2016-Use Rates in the States and Territories*. Tech. Rep. DOT-HS-812-417, National Highway Traffic Safety Administration (NHTSA), 2017.

[17] Klauer, S. G., F. Guo, B. G. Simons-Morton, M. C. Ouimet, S. E. Lee, and T. A. Dingus. Distracted driving and risk of road crashes among novice and experienced drivers. *New England Journal of Medicine*, Vol. 370, No. 1, 2014, pp. 54–59.

[18] Pack, A. I., A. M. Pack, E. Rodgman, A. Cucchiara, D. F. Dinges, and C. W. Schwab. Characteristics of crashes attributed to the driver having fallen asleep. *Accident Analysis & Prevention*, Vol. 27, No. 6, 1995, pp. 769–775.

[19] Paleti, R., N. Eluru, and C. R. Bhat. Examining the influence of aggressive driving behavior on driver injury severity in traffic crashes. *Accident Analysis & Prevention*, Vol. 42, No. 6, 2010, pp. 1839–1854.

[20] Guo, F., X. Wang, and M. A. Abdel-Aty. Modeling signalized intersection safety with corridor-level spatial correlations. *Accident Analysis & Prevention*, Vol. 42, No. 1, 2010, pp. 84–92.

[21] Anarkooli, A. J. and M. H. Hosseinlou. Analysis of the injury severity of crashes by considering different lighting conditions on two-lane rural roads. *Journal of Safety Research*, Vol. 56, 2016, pp. 57–65.

[22] Dunckel, J., W. Haynes, J. Conklin, S. Sharp, and A. Cohen. Pedestrian Safety Initiative in Montgomery County, Maryland: Data-Driven Approach to Coordinating Engineering, Education, and Enforcement. *Transportation Research Record*, 2014, pp. 100–108.

[23] Xie, X.-F. and Z.-J. Wang. An empirical study of combining participatory and physical sensing to better understand and improve urban mobility networks. In *Transportation Research Board (TRB) Annual Meeting*, Washington, DC, 2015, 15-3238.

[24] Hall, D. L. and J. Llinas. An introduction to multisensor data fusion. *Proceedings of the IEEE*, Vol. 85, No. 1, 1997, pp. 6–23.

[25] Xie, X.-F., J. Liu, and Z.-J. Wang. A cooperative group optimization system. *Soft Computing*, Vol. 18, No. 3, 2014, pp. 469–495.

[26] Xie, X.-F. and Z.-J. Wang. Cooperative group optimization with ants (CGO-AS): Leverage optimization with mixed individual and social learning. *Applied Soft Computing*, Vol. 50, 2017, pp. 223–234.

[27] Curry, A. E., M. R. Pfeiffer, R. K. Myers, D. R. Durbin, and M. R. Elliott. Statistical implications of using moving violations to determine crash responsibility in young driver crashes. *Accident Analysis & Prevention*, Vol. 65, 2014, pp. 28–35.

[28] Baratian-Ghorghi, F., H. Zhou, and W. C. Zech. Red-light running traffic violations: A novel time-based method for determining a fine structure. *Transportation Research Part A*, Vol. 93, 2016, pp. 55–65.

[29] El-Basyouny, K. and T. Sayed. Safety performance functions using traffic conflicts. *Safety Science*, Vol. 51, No. 1, 2013, pp. 160–164.

[30] Getis, A. and J. K. Ord. The analysis of spatial association by use of distance statistics. *Geographical Analysis*, Vol. 24, No. 3, 1992, pp. 189–206.

[31] Ayuso, M., M. Guillén, and M. Alcañiz. The impact of traffic violations on the estimated cost of traffic accidents with victims. *Accident Analysis & Prevention*, Vol. 42, No. 2, 2010, pp. 709–717.

[32] Factor, R. The effect of traffic tickets on road traffic crashes. *Accident Analysis & Prevention*, Vol. 64, 2014, pp. 86–91.

[33] Xie, X.-F., S. Smith, L. Lu, and G. Barlow. Schedule-driven intersection control. *Transportation Research Part C*, Vol. 24, 2012, pp. 168–189.

[34] Xie, X.-F., S. Smith, T.-W. Chen, and G. Barlow. Real-time traffic control for sustainable urban living. In *IEEE International Conference on Intelligent Transportation Systems*, 2014, pp. 1863–1868.