Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2018) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

World Conference on Transport Research - WCTR 2019 Mumbai 26-31 May 2019

Data recording patterns and Missing Data in Road Crashes: Case study of five Indian cities

Alok Nikhil Jha[a*], Geetam Tiwari[b], Niladri Chatterjee[c]

[a,b]TRIPP, IIT Delhi, New Delhi 110016, India
[c]Department of Mathematics, IIT Delhi, New Delhi 110016, India

Abstract

Decisions towards improvement of road safety are strongly based on accident data, and data of the vehicles and victims involved in accidents. Non-availability of these data or false data can lead to wrong estimates and planning, making road safety exercise ineffective. The present study reports the parameters collected in road traffic accidents data from five mid-sized Indian cities namely Agra, Amritsar, Bhopal, Ludhiana and Visakhapatnam. Relative to road traffic crash, the data is captured focusing details of all relevant parameters of all dimensions in accident covering participating vehicle, impacted victims, pedestrian and accident related details. The data collection technique for traffic crashes across all the cities are similar but on an average approximate 20% to 70% of data are not being recorded or missing. The observation at city level data is extracted and cleaned from raw information and found that it shows a different kind of mixed pattern for recorded data and data not recorded (or missing data). Further, it presents the various classifications of missing data formats. Analysis of the data patterns gives a quantum of missing data and enables right categorization of the missing data. It further provides suggestions on various possibilities of filling the gaps for the missing data which can be used in further decision making. The data thus categorized will be used for designing a better traffic safety system.

Keywords: Crash data recording, Missing data in Road Accident, Road Accident data in Indian Cities, Missing Data patterns

1. Introduction

Safety has significant relevance in transportation systems. Road transport accounts for last mile connectivity and hence it has a special relevance. Governments, across the world, are increasingly establishing protocols and regulations to improve safety on roads including the United Nations declaring the present decade as the Decade of Action for Road Safety (2011–2020) (Sleet et. al., 2011) with the target is to reduce the fatality toll by 50% in low and middle-income countries (Bliss et. al.,2012).

When an accident occurs, it is either reported by the victim entity with some information or recorded by the on-duty Police department personal or by both of them and cluster of these information & data are known as the accident data. Most of the decisions taken to plan and improve road safety are based on accident data, which makes it the back bone of the country's road safety system. It is used for designing policy, processes and safety parameters for any city, town, territory and overall country. For all accident data collected from various sources, correctness of data has an important role and to ensure correctness of observations recorded at micro level from accident site with no incomplete or inaccurate recording of crash related variables. If there is any mistake in preliminary data, it will be carried forward and affect the subsequent data, analysis and results. The first & foremost important pre-requisite, to monitor the progress, is high-quality reliable data, without which any observed changes would be un-interpretable (Chokotho et. al.,2013). Hence consistent and correct data has a vital important role.

Incomplete & inaccurate recording or missing data errors is one of the major problem in the recorded data of road crash data leading to many planning and management related issues making the entire road safety exercise ineffective. The study showcases the situation in five Indian cities of different geographical and demographics, the data being captured for road accidents and variations in various key parameters having incomplete recording of data (missing data) and erroneous recording, there by impacting directly and indirectly in many decisions. The output of this research study is to summarize the discrepancies & classify various types of missing data within the accident data.

From 1990 to 2013, road accident has moved up to fifth position (Global burden of disease study; 2015) from tenth for the cause of mortality due to increase in fatalities across globe. Road traffic accidents (RTA) are the prime factor in injuries & leading to fatalities further wasting resources of the nation. As road traffic safety is one of the major concerns, a lot of conclusion can be derived from past data, a careful analysis of traffic data and records can collectively provide lots of details and information that can be utilized in designing a safety system. It is an important aspect to know the type of issues in recording the information in the accident data and the factors around it. Any mistake occurring at the time of reporting and recording cannot give same result at later stage. However there could be situations and possibilities where at the time of recording ambiguity may occur leading to inaccurate recording. In the past, studies have been conducted to analyze the results of fatal road traffic crashes and one of such also focuses on Indian cities (Mohan et. al., 2016) where in the percentage of vulnerable road user deaths in various cities for different modes of transport has been presented. These analysis has discrepancies which will be shown later in the paper.

2. Literature

The growth of vehicles has led to roads with higher traffic density. The immediate effect of this situation is the dramatic increase of traffic accidents and increase in number of fatalities due to this Road Traffic Incidents (RTI) on the road. This has become a serious problem in many countries. For example, 2478 people died on Spanish roads in 2010, which means one death for every 18,551 inhabitants (Fogue et. al., 2013). Additionally, 34,500 people died in the whole European Union as a result of a traffic accident in 2009 (Statistical office of the European Communities; 2012). In the United States (according to the Department of Transportation, United States), in 2012, 33,561 people died in motor vehicle traffic crashes. Global situation worsened in 2013 with RTI moving up to the fifth position after many diseases e.g. heart, respiratory infections, cerebrovascular disease etc (Global burden of disease study; 2015).

The conditions varied in case of many of the OECD (Organizations for Economic Co-operation and Development) countries (IRTAD, 2014) however, lowest and middle income countries (WHO, 2013) were affected. India accounts for a large share of the deaths and disabilities due to RTI among the low and middle income countries, owing partly to the country's large share to the world population (WHO, 2013) and a lack of appropriate road safety measures (Mohan et. al., 2009).

Road crashes resulted in 141,526 fatalities in India in 2014, resulting in a rate of 11 deaths per 100,000 population as compared to rates of in more successful countries, where rates tended to be around 3 or 4 deaths per 100,000 population (NCRB, 2014). The status report published by National Crime records bureau, India, around 477,731 injuries have been reported in 2014 road traffic crashes in India (NCRB, 2015) but the data will be higher as approx. 3-4 lacs persons visited hospitals for road traffic injuries and most of the cases went unreported and for every such incident(s) an epidemiology (Mani et. al., 2013) approach is followed and data is collected from available sources. The high rate of Road crashes is also evident with the varying fatality rates between 3 and 35 per 100,000 as per global burden of disease study by lancet in 2015. The fatality rate has increased over the decade, however, as details of RTI and crashes are not in the public domain for most of the cities, it is difficult to ascribe reliable reasons & any kind of dependency relationship for these increases in fatalities over time.

2.1. Missing Data and Errors

Missing data are frequently encountered in almost all areas of nonlinear phenomena in science and engineering. Missing data poses a big threat towards validity, reliability, and generalizability of data. In the medical researches, the missing data has been part of research in trauma care planning (Shivasabesan et. al., 2018) and it was observed the missing value in trauma care completely impacts the healthcare approach. Missing data issues have been in domain of image processing and applications and researchers have been working on estimation and imputation of those missing values. Missing data are ubiquitous in quantitative research studies. Methodologists also described missing data as one of the most important statistical and design problems in research (Azar et. al., 2002) due to its pervasive nature.

Missing data errors in accident data have not been the subject of extensive research; however, scientists have paid attention to this important area of road safety in the last few decades. The data in road accident once recorded can have missing values when no data value is stored for a variable in an observation. Due to absence of data in the data set, many meaningful conclusions are left over. There are researches carried out on missing data methods and classifying in various categories and predicting methods using mining and other algorithms.

Occurrence of a Road Accident is a random activity. Though we may have digital & technology devices, however the data recording and reporting has manual interventions & procedures which is hence prone to errors. Errors in accident data can be broadly classified into two types- error in reporting and error in recording, as shown in Fig. 1.

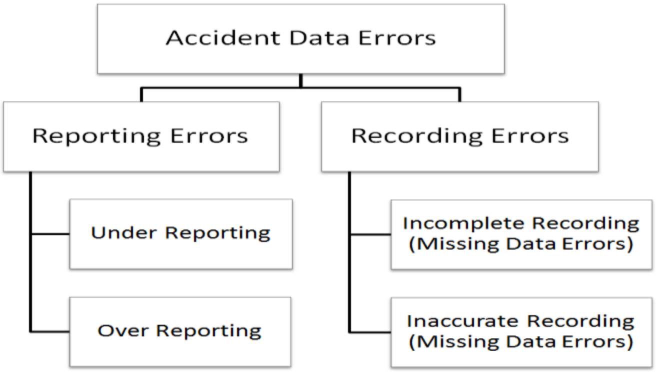

Accident Data Errors
Reporting Errors
Recording Errors
Under Reporting
Over Reporting
Incomplete Recording (Missing Data Errors)
Inaccurate Recording (Missing Data Errors)

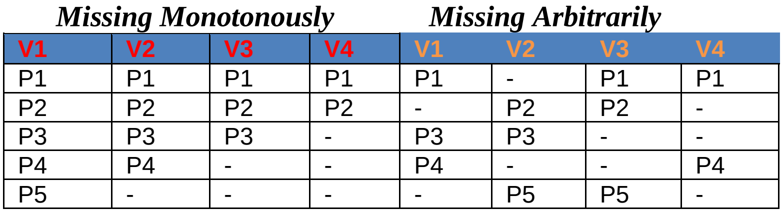
Fig. 1: Type of Errors in Accident Data

The reporting errors lead to incorrect statistics of accidents for town or country. It occurs due to under reporting or over-reporting of accidents. The recording errors results in ambiguities in evaluation and calculations for factors responsible for the occurrence of an accident, such as type of road, type of accident, the environment, the driver, the vehicle, the location, classification of injury severity etc. It occurs due to incomplete or inaccurate recording or no recording of information for the accidents. The incomplete or no recording of variables are Missing data errors or simply missing data. The recording errors are measured w.r.t. location of accident, victims' information, vehicles' information and the information pertinent to the road and environment. The location of accident could be in the form of geographic coordinates (Miler et. al., 2016), local road map coordinates (Imprialou et. al., 2015) (Qin et. al., 2013), route number (Burns et. al., 2015). The victims' information include number of victims, their gender, age, injury severity, type (driver/occupant/pedestrian), safety belt usage, helmet usage, alcohol and drug use (Hauer et. al., 1988). The vehicles' information include number of vehicles, their registration number, type, make, colour, extent of damage (Nguyen et. al., 2011) (Razzak et. al., 1998). The road and environment is a broad category which contain information related to date of accident, time of accident, intersection type, road features, type of accident, cause of accident, weather, lighting and traffic conditions.

2.2. Missing Data Interpretation & Analysis

Missing data as encountered in almost all areas of science and engineering tends to have inaccurate results of the observation. The inaccurate and incomplete records of accident data make it unusable for any planning and activities. These missing data, as it stated, are either not at all captured at the time of recording or partially recorded or otherwise inaccurate to consider. The missing data problem can occur due to many reasons e.g. generalized missingness or arbitrarily missing data; occurs when no obvious patterns is observed, monotone missing is the one when missing values can be ordered in such a way that once a missing value appears, all subsequent variables also have missing values. Wave missing is specific to longitudinal data with repeated measures. In this case if one of the sampling times is missed, all data from that particular time point will be missing (Dohoo. et. al., 2015.). The two main patterns of missingness are presented in table 1.

Table 1 : Missing data patterns



Missing Monotonously				***Missing Arbitrarily***			
V1	V2	V3	V4	V1	V2	V3	V4
P1	P1	P1	P1	P1	-	P1	P1
P2	P2	P2	P2	-	P2	P2	-
P3	P3	P3	-	P3	P3	-	-
P4	P4	-	-	P4	-	-	P4
P5	-	-	-	-	P5	P5	-

There are three mechanisms by which data may be missing: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR – also called MNAR).

MCAR means, as its name implies, that the probability of a value being missing in a variable is a completely random event. It is unrelated to any other record in the database, completely random behavior. This type of missingness is unsystematic. An example of this might be a road accident happens between two vehicles and details of injured or fatalities are not present. The probability of being missing is the same for all observations in the dataset and is not related to any other data recorded. For example, in table 3, in type of accident is known in 50% cases (excluding missing data). Now based on other record we may predict that missing data would have not been a Hit-run accident as speed is 50+ like other two records. Though it may not be correct and could have been a hit and run accident. When we make this assumption, we are assuming that missing data is completely unrelated to the other information in the database. (Raghunathan, et. al., 2004), (Muthen. Et. al., 1987). The missingness is completely unsystematic.

Table 2: Missing Data Types

S.No	Time	Victim's Speed	Hit-Run
1	0900	40-50	Yes
2	1600	40-50	Yes
3	2030	50+	No
4	2130	50+	Missing /Unknown
5	0100	50+	No

MAR means that the probability of being missing depends only on the observed data (i.e. the probability of being missing can be fully explained by variables recorded in the dataset). An MAR mechanism is not random at all and describes systematic missingness where the propensity for missing data is correlated with other study-related variables in an analysis. The missingness can be predicted and prediction is based on dependent on the data that is given, hence it is not random at all. In the example above, we can calculate based on time and speed & can propose a probabilistic relationship establishing that missing variable will be 'No'. In MAR we have a better probability of predicting the true value of the missing data. As a second example, suppose that a school administers a math aptitude exam, and students that score above a certain cut-off participate in an advanced math course. The math course grades are MAR because missingness is completely determined by scores on the aptitude test (e.g., students that score below the cut-off do not have a grade for the advanced math course). However, if the cut-off is not specified and condition would have been open, then missing data type will be MCAR. For example if availability of a road divider is not recorded in accident information sheet, it could be due a reason that the road would be a one way & has no divider, but entire analysis is not recorded then it will be MAR.

Missing not at random (MNAR or NMAR) (also known as non-ignorable nonresponse) is data that is neither MAR nor MCAR (i.e. the value of the variable that's missing is related to the reason it's missing). Data are NMAR if the probability of being missing depends on unobserved data and are missing based on the would-be values of the possible type of missing values. In this case, the probability of being missing is related to a value which is itself missing. The probability of being missing may also be related to some completely unknown (and unmeasured) factor which will produce NMAR data if this unknown factor is related to the exposure and/or outcome. The missingness analysis depends on would-be values and independent of those given in analysis. Since the values missing are based on possible would be values and hence it is kind of impossible to calculate.

It is important to point out that the missing data mechanisms are not characteristics of an entire data set, but they are assumptions that apply to specific analyses. Consequently, the same data set may produce analyses that are MCAR, MAR, or MNAR depending on which variables are included in the analysis.

3. Proposed Scheme of Study

This research study reports the parameters collected in road traffic accidents data from five mid-sized Indian cities namely Agra, Amritsar, Bhopal, Ludhiana and Visakhapatnam for over a period of five years viz. 2007 to 2011. Relative to road traffic crash, the data is captured focusing details of all relevant parameters from perspectives relevant to participating vehicle, impacted victims and pedestrian and accident related details

3.1. Selection of cities

Five cities with populations between 1.0 and 2.0 million and different RTI(Road traffic incident) fatality rates were selected from different locations in India: Agra, Amritsar, Bhopal, Ludhiana, and Vishakhapatnam. Table 3 shows the population and fatality rate for each city (Hauer, et. al., 1988) (Shinar. et. al., 1983). These cities exemplify the type of growing urban agglomerations that observers expect to experience high growth rates

Table 3: Population and road traffic fatality data for five cities selected for the study (Hauer et.al.; 1988 and Shinar et. al;1983)

	City				
	Agra	Amritsar	Bhopal	Ludhiana	Vishakhapatnam
2011 population	1,574,542	1,132,761	1,795,648	1,613,878	1,730,320
2011 road traffic fatalities	653	70	254	294	416
Fatalities per 100,000 persons	41	6	14	18	24

3.2. Data Gathering

Primary data manually obtained from Agra, Amritsar, Bhopal, Ludhiana and Vishakhapatnam (Vizag) on vehicle registration, road traffic fatality and injury cases, and other data available from city police stations and First Information Reports (FIRs) detailing road traffic crashes for the period 2007 to 2011. Table 4 shows the number of records obtained from each city. Vehicle registration data for Amritsar was not available.

Table 4: Accident Records in 5 Indian Cities

City	Accident Data	Vehicle Data	Victim Data	Pedestrian Data
Agra	674	1061	635	358
Amritsar	265	450	223	78
Bhopal	685	1027	1027	317
Ludhiana	651	968	490	223
Vizag	1164	1670	1048	553

The data collected over the period of five years from five cities were segregated separately in categories of accident related, vehicles and victim related information from the centrally available information. Each category viz, Accident, vehicle & victim was reviewed separately for these cities on data not recorded or error in the recorded data. The error is defined as the data which should fall in some criteria but was not mentioned within limits of identified criteria or any outlier value. The data from the police files for respective cities were organized.

The variables for all the cities are segregated in a database with an intention to study all the parameters in respective functionally dependent domain of variables for the ease of understanding. And the data set was thus organized as Accident related variables, vehicle related variables and victim related variables. Below variables are considered-

- Month, day and time of crash
- Type of road & related attributes at accident site
- Type of accident, road user type and crash vehicles
- Accident severity, impact- injuries and fatalities,
- Location based parameters e.g. holidays, spot,
- Victim and Vehicle related information and variables

All the variables were studied for the missingness in the recorded data along with the possible category of missing data. The data is taken up in percentage nomenclature of available data & missing data.

4. Analysis

Data for the five years period for the five Indian Cities i.e. Agra, Bhopal, Ludhiana, Amritsar and Vizag; were reviewed for patterns of data recording in the police records. As stated, the data is segregated in three main categories viz. accident related data, vehicle variables and their data and victim variables and their data. Victim variables are divided into two subsets of pedestrian & NMT (Non-Motorized Transport) victims and non-pedestrian victims which involved one or more vehicles in the crash. Each segment of data head is observed separately and their patterns and average missingness along with category of missingness is discussed in subsequent sections.

4.1. Accident Related Data

In accident related data, it was found that certain parameters were not at all recorded like GPS location of accident location, lighting condition for all these cities. However, over the other parameters, which are regularly recorded, some of them were missing. Considering the variables like time of accident, type of accident, Hit-Run status, Total injuries, collision details, road layout, any landmark specified, it was identified that Amritsar has 34% data missing, Ludhiana is on 23%, Bhopal has 28% missing, Agra has 32% missing & Vizag had 34% missing data. The missing data is calculated by averaging the missing weightage of each city for corresponding parameters. Fig 2 shows the variation of missing data for accident related parameters.

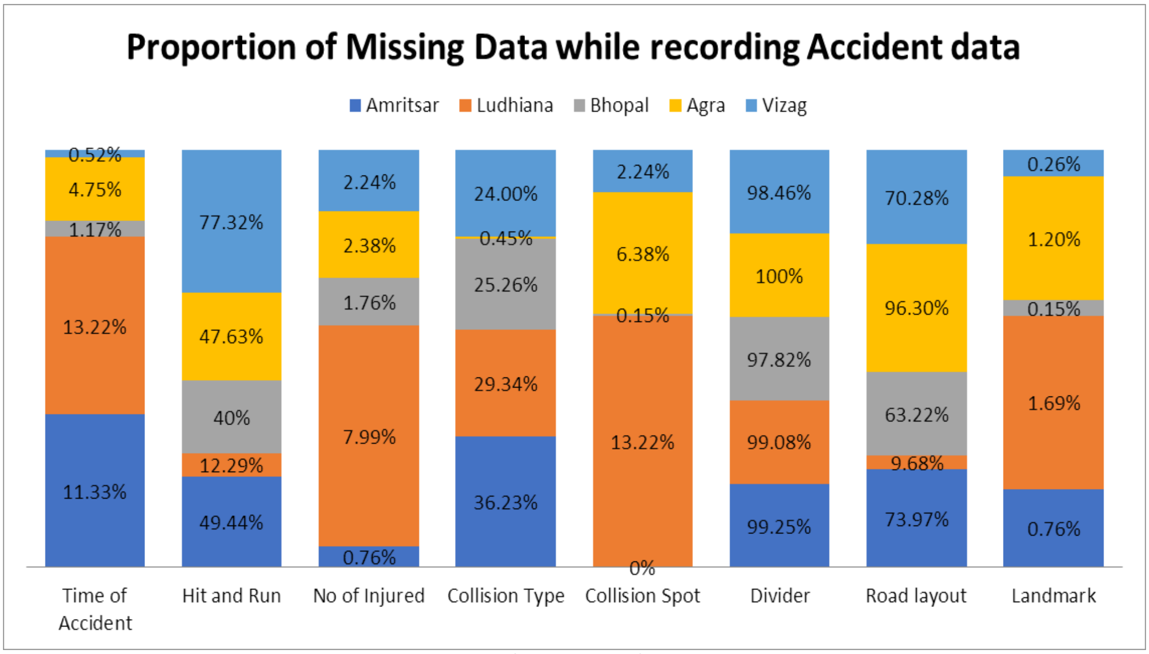


Fig. 2: Proportion of missing data for accident variable

While observing the data, it was also observed that some of the parameters were not recorded at all in all cities and some parameters were completely missing in a specific city and we have tried to consider those values which have impact and doesn't behave as outliers. The record if the city Ludhiana belonging to same states as of Amritsar, has around 98% data missing or unknown about the holiday status on the day of accident. However that's not the case in Amritsar or any other city considered in the study. This variable was not considered while evaluating average for other parameters. The missingness for Holiday status and as defined, this falls into MAR type of Missing data. Similarly the parameters no of vehicles, road category and location are missing only in Ludhiana with 9%, 12% and 98% in all, and hence not considered in overall evaluation. These missingness are NMAR, MAR and NMAR respectively. Table 4 shows the Missing data type for the accident related variables. Road layout 2 depends and evaluated based on road layout 1 and this dependency makes it to fall in NMAR category. Collision type & spot could be human error or depends on vehicles colliding and severity. Other variables are showing a random behavior, hence they are in MCAR category.

Table 5: Missing type from recorded data for Accident related variables

Variables	**Amritsar**	**Ludhiana**	**Bhopal**	**Agra**	**Vizag**
Time of Accident	MCAR	MCAR	MCAR	MCAR	MCAR
Hit and Run	MAR	MAR	MAR	MAR	MAR
No of Injured	MCAR	NMAR	MCAR	NMAR	NMAR

Collision Type	MAR	MAR	MAR	MCAR	MAR
Collision Spot	NMAR	NMAR	NMAR	NMAR	NMAR
Type of Road	MAR	MAR	MAR	MAR	MAR
Divider	MAR	MAR	MAR	MAR	MAR
Location	NA	NMAR	NA	NA	NA
Road Category	NA	MAR	NA	NA	NA
Road layout	MAR	MAR	MAR	MAR	MAR
Landmark	MCAR	NMAR	MCAR	NMAR	MCAR
No of Vehicle	NA	NMAR	NA	NA	NA

4.2. Vehicle Related Data

The vehicle related variables being recorded at accident site was selected and observed to have data missing in selected parameters. The other parameters considered in the evaluation as shown in Fig. 3, for missing & unrecorded data. Considering these variables as shown vehicle type, manoeuver at time of accident if it was proceeding straight, or overtaking, turning, reversing, parked or not known or otherwise not recorded, loading if vehicles was overloaded or not, disposition required, if any mechanical failure or fire caused crash, any object like tree, stone, animal, pole or other vehicle caused the crash, total impacted, hazardous cargo if applicable, make and model of vehicle, it was identified that Amritsar has 61% data missing, Ludhiana has 27%, Bhopal has 60% missing, Agra has 39% missing and Vizag had 58% missing data; by averaging out the missing information. It is an uttermost important to understand that a missing data may also imply that non relevance. For example, hazardous cargo, loading is mostly relevant for cargo trucks and not suitable for smaller vehicles.

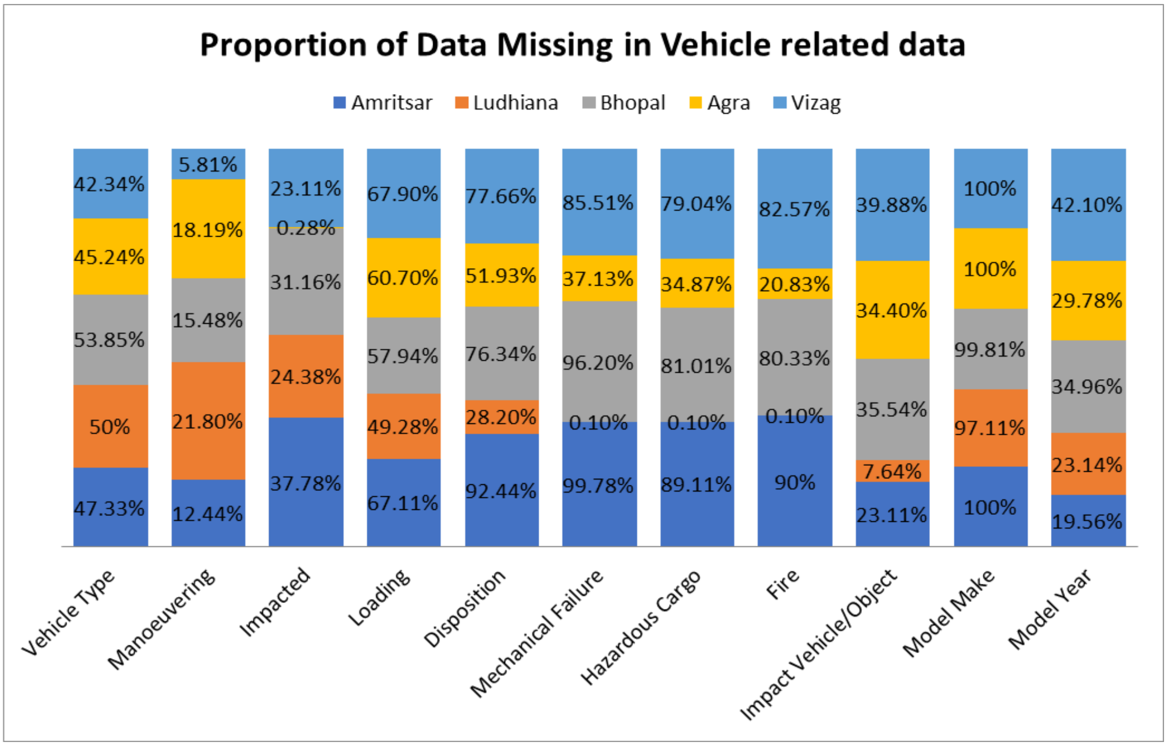


Fig. 3: Proportion of missing data in vehicle related attributes

The variables under the category after observation, were segregated and placed under respective missing error type. The Vehicle type and manoeuvring variable, which talks about vehicle participated in accident and whether it was a back hit or during overtaking or parking etc. And per definition of MAR, we have the field values from the available domain values. The impacted count of people, model and make of vehicle are NMAR category as their values depends on real time values, not on some domain specific. Mechanical failure, Hazardous cargo, Fire are MAR type as per definition of MAR , however it is MCAR for Ludhiana city because of very small (less than 1%) data missing, which could be due of some special case. The loading and disposition which specifies overloading status and the crash vehicle condition post-accident will be MAR category as per definition of MAR. The Table 6 details all the aspects discussed about various possible Missing data type for the collected data.

Table 6: Missing type from recorded data for Vehicle related variables

Parameters	**Amritsar**	**Ludhiana**	**Bhopal**	**Agra**	**Vizag**
Vehicle Type	MAR	MAR	MAR	MAR	MAR
Manoeuvring	MAR	MAR	MAR	MAR	MAR
Impacted	NMAR	NMAR	NMAR	NMAR	NMAR
Loading	MAR	MAR	MAR	MAR	MAR
Disposition	MAR	MAR	MAR	MAR	MAR
Mechanical Failure	MAR	MCAR	MAR	MAR	MAR
Hazardous Cargo	MAR	MCAR	MAR	MAR	MAR
Fire	MAR	MCAR	MAR	MAR	MAR
Impact Vehicle/Object	MAR	MAR	MAR	MAR	MAR
Model Make	NMAR	NMAR	NMAR	NMAR	NMAR
Model Year	NMAR	NMAR	NMAR	NMAR	NMAR

4.3. Victim Related Data

The crash victims (injured and fatalities) are segregated as pedestrians (& non-motorized transport or NMT) and non-pedestrians involving vehicles. The same variables were considered viz. road user type, occupant vehicle, location, Age, Gender, mode of treatment, no of days in hospital, Injury severity except for pedestrian crashes, NMT location was considered. For the five cities, the varying records for non-motorized traffic victims and motorized traffic victims were available and studied separately. For pedestrian victims, it was observed; as shown in Fig. 4, Injury details, no of days in hospital and movement of injured are missing for all cities. This data is not recorded at all, which could have been a manual error or a variable which might not be in use. Analyzing further, it was identified that Amritsar has 47% data missing, Ludhiana has 63% of data missing, Bhopal has 61% missing, Agra has 59% missing and Vizag had 57% missing data. The missing records are computed by simple mean of the missing data for the variables evaluated under MCAR, MAR and NMAR Category.

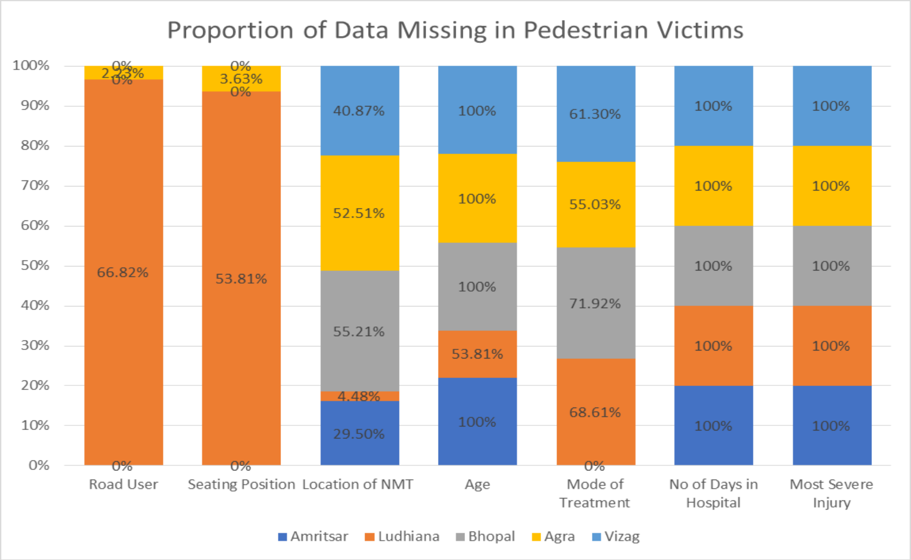


Fig. 4: Proportion of missing data in Pedestrian (NMT) Victims

The different type of missingness in data falling in various category are specified in table 7. Road user type defines from within the source of identified category. However for Agra its only 8 records missing making it to 2% hence could be a random behavior or human error. Seating position, talks about how many people impacted in motorized vehicle is MAR but MCAR in cases when possibly it is not known or cases like overturning of bus. Age can be anything and per definition it will be NMAR. Treatment mode, Injury Severity and days in hospital are having specific domain limits and thus MAR. Missing of location is MCAR.

Table 7: Missing type from recorded data for Pedestrian Victims related variables

Parameters	**Amritsar**	**Ludhiana**	**Bhopal**	**Agra**	**Vizag**
Road User	NA	MAR	NA	MCAR	NA
Seating Position	MAR	MAR	MAR	MAR	MAR
Location of NMT	MCAR	MCAR	MCAR	MCAR	MCAR
Age	NMAR	NMAR	NMAR	NMAR	NMAR
Mode of Treatment	MAR	MAR	MAR	MAR	MAR
No of Days in Hospital	MAR	MAR	MAR	MAR	MAR
Most Severe Injury	MAR	MAR	MAR	MAR	MAR

Analysing the non-pedestrian victims, by taking a mean of unrecorded data, it was identified that Amritsar has 71% data missing, Ludhiana has 54% of data missing, Bhopal has 74% missing, Agra has 63% missing and Vizag had 68% missing data. It was observed that location variable is completely missing in Amritsar however that's not the case in Ludhiana, both belonging to same state as seen in Fig. 5.

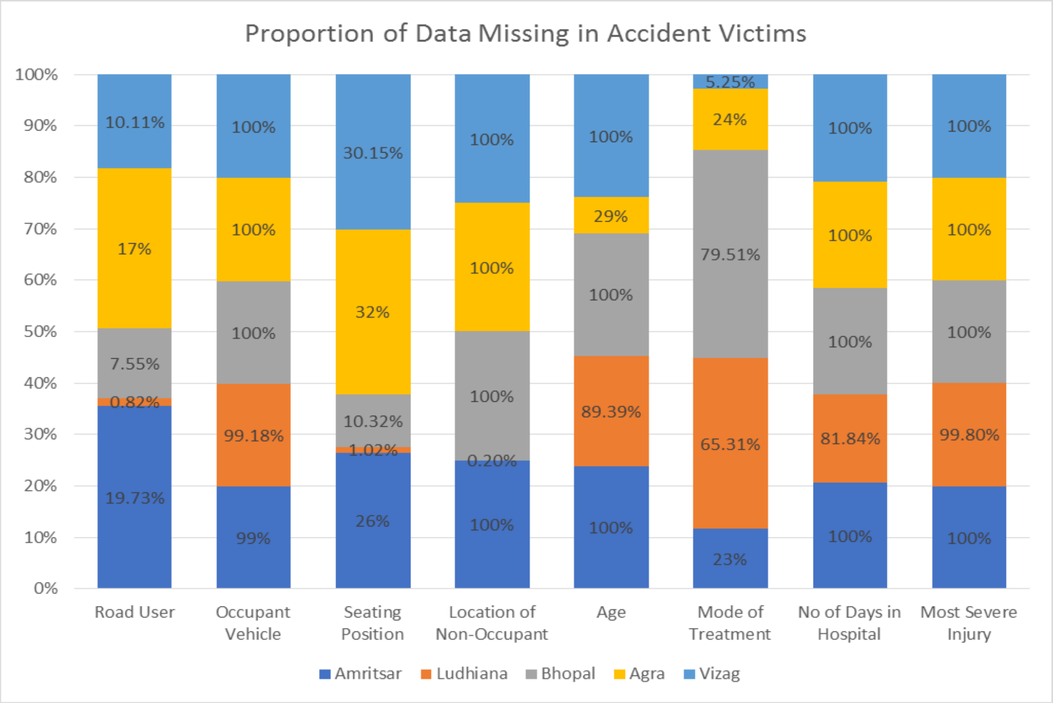


Fig 5: Proportion of missing data in Motorized Victims

The Table 8 depicts similar behavior for missing categorization for non-pedestrian victims as explained above. The direct missingness for Occupant Vehicle are marked MAR and those based on possible values are NMAR as in case of Age here.

Table 8: Missing type from recorded data for Victims related variables

Parameters	Amritsar	Ludhiana	Bhopal	Agra	Vizag
Road User	MAR	MCAR	MAR	MAR	MAR
Occupant Vehicle	MAR	MAR	MAR	MAR	MAR
Seating Position	MCAR	MCAR	MCAR	MCAR	MCAR
Location	MAR	MAR	MAR	MAR	MAR
Age	NMAR	NMAR	NMAR	NMAR	NMAR
Mode of Treatment	MAR	MAR	MAR	MAR	MAR
No of Days in Hospital	MAR	MAR	MAR	MAR	MAR
Most Severe Injury	MAR	MAR	MAR	MAR	MAR

After reviewing the crash data recorded at the site and later after other evidences, for five cities, we observed categories of these missingness. Collectively observing all the variables under their respective superset which we have observed under subsets we get the behaviour as shown in graph (Fig. 6). The city Bhopal having highest population density & an average fatality rate has highest number of unrecorded variables for victims. This clearly indicates, that if the records would not have been missing then data for fatality rate per 100000 person and overall fatalities could have been something else. And these kinds of flaws are creating a big gap in analyzing overall data & designing traffic safety plans based on them. To have a flawless, accurate & complete recording is an essence.

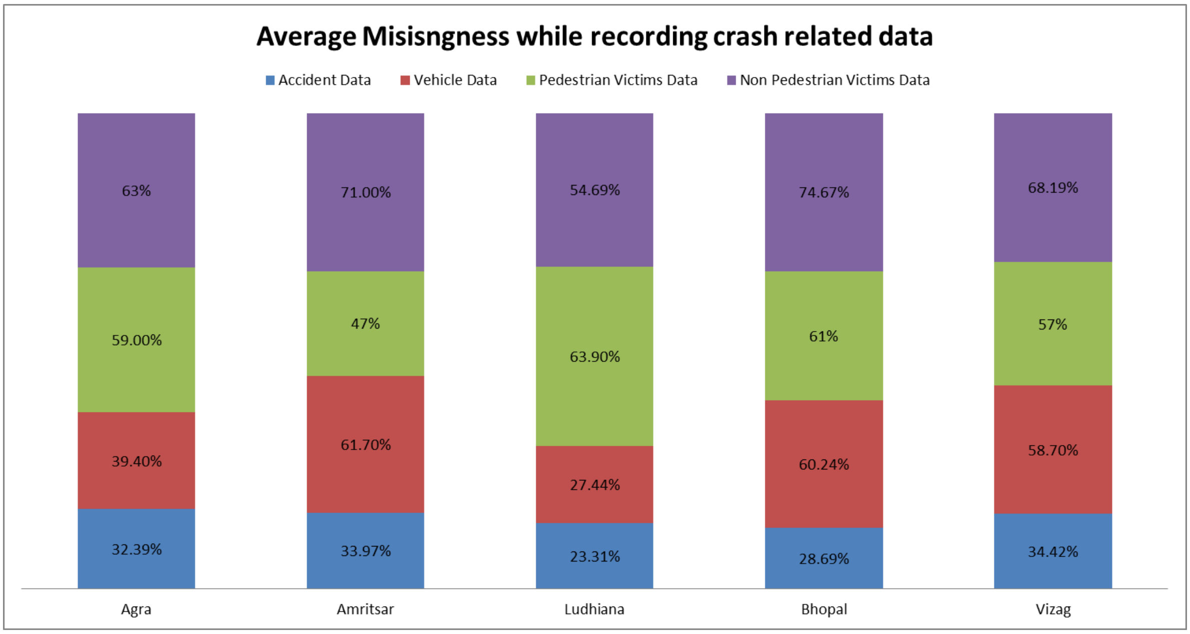


Fig. 6: Average missing data in each city

Similarly if we observe Agra which has highest number of fatality of 653 (table 3), it can be visualized that they have 59%-63% going unrecorded for victims and around 30% of accident related data is not recorded, and hence these data, which are interlinked remains unresolved. For example, the accident related data captures the prime data of site etc and victim's, vehicle's data explains the impact and combined all can be useful in planning a safety policy or system, understand and solve the accident cases, identify the improvement areas but due to missing records, it that percentage of cases go out of hand. Around 40% of vehicle related information is not available. On an average of 48% data is not recorded for Agra and hence overall fatality rate could be something else than what is available in records. The Vizag with population density similar to Bhopal but having more of industrial belt and sea ports, has 24 fatality per 1lac population (table 3) and has around 58% of vehicle data not available in records and more than 65% data missing for victims, thus making a doubtful in considering the actual records. The actual record may vary from what is quoted if we have all the data available. Amritsar and Ludhiana belongs to same state Punjab and Amritsar has population of 11.32 Lacs in 2011 census (table 3) and has least number of fatalities per one lac and least number of overall fatalities. However the data shows that more than 40% victims' data are not recorded or missing and more than 60% data for impacting vehicles are missing, making more than 50% missing data in Amritsar leaving a question mark on biasness of reports being calculated and referred. The Ludhiana has an overall missing of 42% with more than 50% missing in victim's data.

5. Conclusion

Missing data and errors in road accident data are inevitable because it is a passive system in which the incident is required to be reported. However, its extent varies from country & geographical locations with many other factors contributing. A total of five Indian cities with different geographical locations and population density were studied and various dependent parameters were identified with analysis on categories of missingness and allotting the city based parameters in respective causes. The average missing values for each city is evaluated and it was found that city with highest population density has lowest number of victim details available. For the ease of analysis and review, the crash data set was divided on 4 type of data sets viz. accident related, vehicle related and victim related which had pedestrian and non-pedestrians. Evaluating each city separately it was found that on an average, the missingness varies approximately from 20% to 70% in the cases around various parameters and overall it was found to be varying from 40% to 56% range. It was observed that vehicle related data set were missing highest in terms of overall average considering all the cities. The victims are accounting to 45% missingness and accident related data has 22%

unrecorded data. An overall of approximate 50% data is contributing to missing variables in road crashes making it a serious problem. The possible missing category was also identified and observed was also discussed in the sections as how Missing at Random (MAR), Not-Missing at Random(NMAR) and Missing Completely at Random(MCAR) are justifying the missingness.

This 50% missingness of data leads to lack of proper logistical and operational planning for safety. It is extremely important to motivate the personnel responsible for accident data collection to record all information as correctly as possible, specifically accident location and level of injury severity. The accident reporting forms should be improved taken into consideration the limitations of data collection on an active accident site.

Knowledge of the true picture of the road accident problem will be useful for the purpose of negotiating adequate funding for road safety activities. Estimating the actual percentage of unreported data in road accidents and casualties is very useful for any further research in road accidents and safety. It will also enable it to be used as an indicator in the hands of decision makers when dealing with the official data or any results from previous studies. Further we are working in fulfilling the gaps by creating appropriate models using interpolation and regression models.

6. Acknowledgement

Assistance on the data of various cities of India was provided by TRIPP, IIT Delhi.

7. References

Sleet, D.A., Baldwin, G., Dellinger, A., Dinh-Zarr, B., 2011. The decade of action for global road safety. J. Safety Res. 42, 147–148

Bliss, T., Breen, J., 2012. Meeting the management challenges of the decade of action for road safety. IATSS Res. 35, 48–55.

Chokotho, L.C., Matzopoulos, R., Myers, J.E., 2013. Assessing quality of existing data sources on road traffic injuries (RTIs) and their utility in informing injury prevention in the western cape province, South Africa. Traffic Inj. Prev. 14, 267–273.

IRTAD, IRTAD 2014 Annual Report, OECD/International Transport Forum, Paris, 2014 526.

W.H.O., Global Status Report on Road Safety 2013: Supporting a Decade of Action, World Health Organization, Geneva, 2013

Mohan. D, Tsimhoni O, Sivak M., Flannagan M.J., Road Safety in India: Challenges and Opportunities, The University of Michigan Transportation Research Institute, Ann Arbor, MI, 2009 1–57.

NCRB, Accidental Deaths and Suicides in India 2014, National Crime Records Bureau, Ministry of Home Affairs, New Delhi, 2015 Page -320.

Mani, A., Tagat, A., 2013. Safety Assessment of Auto-rickshaws in Mumbai, Embarq India, Mumbai, 1–39.

G. B. D. Mortality and Causes of Death Collaborators., 2013. Global, regional, and national age–sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013, The Lancet 385 (2015) 117–171.

Mohan. D, Tiwary G., Mukherjee S., Urban traffic safety assessment: A case study of six Indian cities, 2016

Miler, M., Todić, F., Ševrović, M., 2016. Extracting accurate location information from a highly inaccurate traffic accident dataset: a methodology based on a string matching technique. Transp. Res. Part C: Emerg. Technol. 68, 185–193.

Imprialou, M.-I.M., Quddus, M., Pitfield, D.E., 2015. Multilevel logistic regression modeling for crash mapping in metropolitan areas. Transp. Res. Record: J. Transp. Res. Board 2514, 39–47

Qin, X., Parker, S., Liu, Y., Graettinger, A.J., Forde, S., 2013. Intelligent geocoding system to locate traffic crashes. Accid. Anal. Prev. 50, 1034–1041.
Burns, S., Miranda-Moreno, L., Stipancic, J., Saunier, N., Ismail, K., 2014. Accessible and practical geocoding method for traffic collision record mapping. Transp. Res. Record: J. Transp. Res. Board 2460, 39–46.

Hauer, E., Hakkert, A., 1988. Extent and some implications of incomplete accident reporting. Transp. Res. Rec. 1185, 1–10.

Shinar, D., Treat, J.R., Mcdonald, S.T., 1983. The validity of police reported accident data. Accid. Anal. Prev. 15, 175–191

Razzak, J.A., Luby, S.P., 1998. Estimating deaths and injuries due to road traffic accidents in Karachi, Pakistan, through the capture-recapture method. Int. J. Epidemiol. 27, 866–870.

Nguyen, H.D.U.C., Duong Thi, M.H.O.A., Nguyen, T.H., Nguyen, N.B.A.O., 2011. Study on quality of existing traffic accident data in Vietnam. Proceedings of the Eastern Asia Society for Transportation Studies, 2011 374

Dohoo, I.R., 2015. Dealing with deficient and missing data and Preventive Vet. Medicine.

Raghunathan, T. E. (2004). What do we do with missing data? Some options for analysis of incomplete data. Annual Review of Public Health, 25, 88−117

Muthen, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. Psychometrika, 51, 431−462.

Austin, K., 1995b. The identification of mistakes in road accident records: part 2: casualty variables. Accid. Anal. Prev. 27, 277–282

Fogue M., Garrido P., Martinez F.J., Cano J.-C., Calafte C.T., A novel approach for traffic accidents sanitary resource allocation based on multi-objective genetic algorithms, Expert Syst. Appl., 40 (1) (2013), pp. 323-336

Eurostat: Statistical Office of the European Communities (2012). Transport statistics in the European Union

Azar B., 2002, Finding a solution for missing data, Monitor on Psychology, 33 (2002), p. 70

Shivasabesan G., Mitra B., 2018, G. M. O’Reilly, Missing data in trauma registries: A systematic review, 2018