Accident Analysis and Prevention 42 (2010) 548–557

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

A genetic programming approach to explore the crash severity on multi-lane roads

Abhishek Das*, Mohamed Abdel-Aty

Department of Civil & Environmental Engineering, University of Central Florida, 4000 Central Florida Blvd., Orlando, FL 32816-2450, United States

ARTICLE INFO

Article history:
Received 27 April 2009
Received in revised form
14 September 2009
Accepted 24 September 2009

Keywords:
Crash severity
Multi-lane roads
Genetic algorithm
Genetic programming
Discipulus

ABSTRACT

The study aims at understanding the relationship of geometric and environmental factors with injury related crashes as well as with severe crashes through the development of classification models. The Linear Genetic Programming (LGP) method is used to achieve these objectives. LGP is based on the traditional genetic algorithm, except that it evolves computer programs. The methodology is different from traditional non-parametric methods like classification and regression trees which develop only one model, with fixed criteria, for any given dataset. The LGP on the other hand not only evolves numerous models through the concept of biological evolution, and using the evolutionary operators of crossover and mutation, but also allows the investigator to choose the best models, developed over various runs, based on classification rates. Discipulus™ software was used to evolve the models. The results included vision obstruction which was found to be a leading factor for severe crashes. Percentage of trucks, even if small, is more likely to make the crashes injury prone. The 'lawn and curb' median are found to be safe for angle/turning movement crashes. Dry surface conditions as well as good pavement conditions decrease the severity of crashes and so also wider shoulder and sidewalk widths. Interaction terms among variables like on-street parking with higher posted speed limit have been found to make injuries more probable.

1. Introduction

Crashes on high-speed (speed limit greater than 45 mph), multi-lane arterial corridors (more than one lane in each direction of travel) with partially limited access account for a significant proportion of traffic fatalities. In the state of Florida; crashes on high-speed (speed limit equal or greater than 45 mph), multi-lane arterial corridors with partially limited access account for 45.36% (NHTSAT, 2007) of the total number of fatalities related to speeding. Changing traffic conditions and environmental settings make highway safety and traffic operations a perennial field of concern. Numerous state-of-the-art methods to improve safety of the roadways are available to the practicing engineer; hence the challenge is not only to identify which methods suits best but also to explore what new insight could be added to the existing body of knowledge.

In the field of transportation safety, it is not only important for us to identify the contributing factors but also to understand their contribution to the problem at hand. To understand the contribution for better assessment of the safety situation, innovative methodologies are being adopted. Since the data used in this study is observational (i.e. collected outside the purview of a designed experiment); an information discovery approach has to be adopted. Pande and Abdel-Aty (2008) in their work on association rules, point out that data mining techniques remain underutilized for analysis of crashes. The underutilization is especially noteworthy since most studies use observational data collected outside the purview of an experimental design. Apart from using a new methodology, the authors have also approached the roadway elements in a more unified way. In this study the corridors have been treated in its entirety, i.e. putting both the segments and intersections together. They then have been clustered into four groups based on the length of corridor. Abdel-Aty and Wang (2006) have shown a spatial correlation between crash patterns of successive signalized intersections, which may be attributed to the characteristics of the segments joining them. In Florida, all the crashes occurring within 250 ft from the center of an intersection are categorized as intersection related crashes. Recently Das et al. (2008) showed that proximity only is not the best way to assign crashes. Wang et al. (2008) used frequency modeling for crashes with fixed as well as varying influence distance and found different set of significant factors. These recent research justified the treatment of the corridor as a whole and not breaking them into segments and intersections.

In the present study the authors set up a classification problem for the injury occurrence as well as the severity of crashes. In a typical classification problem the algorithm develops a set of rules which when followed leads to a particular category of the target variable. For example, in crash severity analysis when the

* Corresponding author. Tel.: +1 407 823 1056; fax: +1 407 823 3315.
E-mail address: abhishekdas.das@gmail.com (A. Das).

0001-4575/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2009.09.021

binary target variable represents severe/non-severe crashes, the classification rule developed will lead to either severe crashes or non-severe crashes. The variables that enter the rule are significant and their directionality is critical for understanding the contribution of the variable in the analysis. Abdel-Aty and Pande (2006) have used the classification trees and neural networks in detecting the relationship between real time freeway traffic conditions and rear-end crashes. However according to Deschaine and Francone (2004), genetic programming (GP) is observed to perform better than classification trees in terms of lower error rates and also outperforms neural networks in regression analysis. GP is a heuristic search technique that iteratively evolves better programs which could either be the best solutions or lead to the best solutions. The innovative evolutionary computation, GP, is based on the genetic algorithms (GA). In GA, the optimum solution is reached by using the well established techniques of evolutionary biology. In a recent work by Makkeasorn et al. (2006) in the field of water resources management, soil moisture estimation models were developed by the use of Discipulus™ Linear Genetic Programming (LGP) software and were applied to the soil moisture distribution analysis. The work shows that LGP, a type of GP, helps in the development of excellent nonlinear multivariate regression models. The work also compared the LGP model developed with the linear regression and nonlinear regression models independently and the LGP model was found to be the best for the data. The linear regression model overestimated the soil moisture while the nonlinear regression models tend to underestimate it. According to Chang and Chen (2000) the regression models generated by GP is also independent of any model structure. Use of GA in transportation is not new. They have been used widely in traffic signal system optimization and network optimization (Park et al., 2000; Ceylan and Bell, 2004; Teklu et al., 2007). The use of GA or GP in transportation safety studies is relatively new and hence the authors intend to test the method and observe its potential. A set of roadway geometric variables were chosen to understand the classification of injury/non-injury crashes as well as severe/non-severe crashes. The authors also use Discipulus™ for the classification problem. Aspects of the software critical to the study in hand will be discussed in Section 3 of the paper.

The focus in this study is to evolve the best possible classification rules that are developed by the LGP methodology. The best program developed by LGP is essentially a set of line-by-line instructions. When the instructions are read from top-to-bottom, they lay out a classification rule. The use of LGP to detect the classification rule is an improvement from all other existing methodologies. In LGP, the heuristic approach to reach the best program goes through a process of evolving numerous programs. The process terminates when no further improvement in classification or decrease in misclassification is observed. The details of the selection process are given in Section 3.

The following section deals with the intricate details of the data preparation, which includes the creation of the dummy variables. The approach to modeling and the dependent variable set up are to be discussed. The section after that deals with the modeling methodology, i.e. explain the GP in more detailed manner and also explain the set up of the classification problem. Disadvantages of GA which led to the development of GP are discussed in this section. The results and analysis section primarily focus on the significant variables and their relationships, discovered by the best evolved programs and their interpretation relevant to roadway safety.

2. Data collection and preparation

2.1. Study area and available data

The crash data available were from the Crash Analysis and Reporting (CAR) system of the Florida Department of Transportation (FDOT). The Roadway Characteristics and Inventory (RCI) data was also made available through FDOT. The data used are for the years 2004 through 2006 for all the state roads of Florida. The datasets have information regarding traffic, roadway geometric and traffic crashes. The datasets were merged and the parameters were modified to suit the genetic programming methodology being implemented in the study.

2.2. Data preparation

The corridors were logically combined to form continuous sections based on design standards. The corridors are of variable lengths. Hence, it was logical to cluster them based on length before further analysis on severity could take place. The optimum number of clusters was found based on the Partitioning around the Medoids (PAM) algorithm proposed by Kaufman and Rousseeuw (1990). In the PAM algorithm the objective function is the sum of the dissimilarities of the objects in a group. The algorithm terminates when the minimization of the objective function is not possible with an interchange of objects across groups. The optimum number of clusters was found to be 4. The following are the length of the corridors in each cluster: Cluster 1 (1.009–2.89 miles); Cluster 2 (2.898–5.729 miles); Cluster 3 (5.762–10.556 miles) and Cluster 4 (10.644–78.293 miles).

The types of crashes used in the study are: (i) angle/turning movement (44,088 crashes); (ii) head-on (3709 crashes); and (iii) rear-end (57,155 crashes). The other type of crashes could not be used as insufficient data failed to produce any classification rule for any cluster. Continuous variables like ADT, percentage of trucks, and k-factor (design hour volume as a percentage of ADT) and skid (friction resistance multiplied by a factor of 100) were also divided into categories. Their relationships with severe/fatal crash occurrence may not be monotonic in nature. Nominal variables such as median types, access management, shoulder types, surface types, etc. were also used in the data set. In most statistical applications the nominal variables can be defined and the dummy variables are created internally. In the present study however, the researcher will have to create dummy variables for all the nominal variables with three or more categories. Otherwise the LGP will treat it as an ordinal variable. A total of 58 variables have been used.

Table 1 presents all the independent variables along with the dependent severity variable. The present analysis deals with roadway geometric and design factors. The authors would like to reiterate that the objective of the study is to understand the classification of injury/non-injury crashes as well as severe/non-severe crashes. Apart from that, the researchers wanted to investigate the usefulness of using the heuristic LGP methodology in the classification problem to identify significant variables and their relationship. As an initial approach the authors have used specific roadway geometric and design factors in this particular study, information for which were completely available. A broad spectrum of variables is always available and open to investigation. However, in this study only certain variables (Table 1) have been included which broadly belongs to roadway geometric and design category. These variables are generally used in engineering studies to develop safety countermeasures. Many of these variables have been collected and are unique to this study. As discussed in Section 4, the results highlight intuitive observations and also help in discovering of interactions among variables. All other variables that have not been included are beyond the scope of the present study.

Most of the variables as can be observed are binary with a few continuous variables. Most of the binary variables are dummy variables which uniquely represent a particular aspect of the original nominal variable and hence, the results of the classification could be directly interpreted. The descriptions for the variables 16 through

Table 1
Dependent/independent variables used in the analysis.

Variable name	Variable number	Description
Target or dependent variable		
Injury		Binary target variable
Severity		Binary target variable
Environmental and roadway geometric parameters		
Surface_width	0	Width of the surface (continuous)
Max_speed	1	Maximum posted speed limit (continuous)
Road_cond	2	Road condition at time of crash (binary (1 = no defects; 2 = defects))
Vision	3	Vision obstruction (binary (1 = no; 2 = yes))
shld_side	4	Shoulder + sidewalk width (continuous)
surf_cond	5	Surface condition (binary (1 = dry; 2 = other))
light	6	Daylight condition (binary (1 = daylight; 2 = other))
k_fact	7	Average k-factor (k_fact ≤9.85, k_fact > 9.85)
trfcway	8	Vertical curvature (binary (1 = level; 2 = upgrade/downgrade)
park	9	Presence of parking (binary (1 = no; 2 = yes))
surf_type	10	Type of surface (binary (1 = black top surface; 2 = any other surface))
shld_t	11	Type of shoulder (binary (1 = paved; 2 = unpaved))
LIGHTCDE_1	12	No street light (binary)
LIGHTCDE_2	13	Presence of street light (binary)
LIGHTCDE_3	14	Partial lighting (binary)
ACMANCLS_num_0	15	No median opening (binary)
ACMANCLS_num_2	16	Presence of restrictive median with service roads (binary)
ACMANCLS_num_3	17	Presence of restrictive median (binary)
ACMANCLS_num_4	18	Presence of non-restrictive median (binary)
ACMANCLS_num_5	19	Presence of restrictive median with shorter directional openings (binary)
ACMANCLS_num_6	20	Presence of non-restrictive median with shorter signal connection (binary)
ACMANCLS_num_7	21	Presence of both restrictive and non-restrictive median types (binary)
curvclass_1	22	Presence of curve < 4° (binary)
curvclass_2	23	Presence of 4° ≤ curve ≤ 5° (binary)
curvclass_3	24	Presence of 5° < curve ≤ 8° (binary)
curvclass_4	25	Presence of 8° < curve ≤ 13° (binary)
curvclass_5	26	Presence of 13° < curve ≤ 27° (binary)
curvclass_6	27	Presence of curve >27° (binary)
ADT_1	28	ADT ≤31,000 (binary)
ADT_2	29	31,000 < ADT ≤ 40,000 (binary)
ADT_3	30	40,000 < ADT ≤ 52,500 (binary)
ADT_4	31	ADT > 52,500 (binary)
t_fact_1	32	t_fact ≤4.05 (binary)
t_fact_2	33	4.05 < t_fact ≤ 5.895 (binary)
t_fact_3	34	t_fact > 5.895 (binary)
dayandtime_1	35	Afternoon Peak Weekday (binary)
dayandtime_2	36	Morning Peak Weekday (binary)
dayandtime_3	37	Friday or Saturday Night (binary)
dayandtime_4	38	Off-peak (binary)
pavecond_1	39	Poor condition (binary)
pavecond_2	40	Fair condition (binary)
pavecond_3	41	Good condition (binary)
pavecond_4	42	Very good condition (binary)
skid_f_1	43	Skid ≤34
skid_f_2	44	34 < skid ≤ 38
skid_f_3	45	Skid >38
median_0	46	No median (binary)
median_1	47	Presence of painted (binary)
median_2	48	Presence of median curb ≤6″ (binary)
median_3	49	Presence of median curb >6″ (binary)
median_4	50	Presence of lawn (binary)
median_5	51	Presence of paved median (binary)
median_6	52	Presence of curb ≤6″ and lawn (binary)
median_7	53	Presence of curb >6″ and lawn (binary)
median_8	54	Other median (binary)
ele_1	55	Segment related crashes (binary)
ele_2	56	Intersection related crashes (binary)
ele_3	57	Access related crashes (binary)

20 described in Table 1 have *restricted median* or *non-restrictive* median types. The *restrictive medians* are those medians which provide a physical barrier between the opposing traffic lanes; where as the *non-restrictive medians* are those which are painted medians or center lines that do not provide a physical barrier. The variables 55 through 57 in Table 1 provide some new innovative variations to the traditional parameters. Traditionally the site location variable has been used by researchers to assign crashes to the three roadway elements (segments, intersections and access points). However a detailed review of several hundred crash reports, suggested that the 'site location' variable by itself was a weak indicator for the location. For example, it was observed that it is possible for a crash to be not attributed to a signalized intersection even if it may have occurred very close to one. In fact, 'traffic control' in combination with the 'site location' along with the information of the presence or absence of signal, did a superior job in attributing crashes to one of the three roadway elements. Based on these three independent parameters, the variables *ele_1*, *ele_2*, *ele_3* were created to assign the crashes to

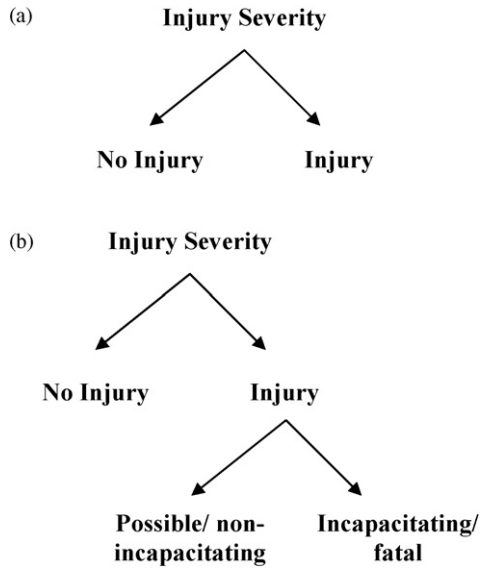

Fig. 1. (a) Binary classification of non-injury and injury related crashes and (b) nested modeling concept.

the three roadway elements, namely segments, intersections and access points, respectively.

The first analysis that was carried out was a binary classification problem between injury crashes and non-injury crashes. Fig. 1(a) shows the primary binary classification problem. It must be noted that a major proportion of non-injury crashes are primarily PDO crashes which are known to be under-reported (Abdel-Aty and Keller, 2005; Yamamoto et al., 2008). A correction factor has not been included as that will over represent PDO crashes at many sites. It is not believed that this issue would affect the results and objectives of this study.

However, this will be just a part of the analysis. Since the injury related crashes represent all types of injuries and the degree of severity ranges from possible injury to death, it should be further be split. Keeping in view the nature of the injury two possible grouping of the injury related crashes is possible. The crashes with fatalities and incapacitating injuries have been grouped together. They are put together into one level as the crashes that involve incapacitating injury could easily have been fatal and vice-versa possibly due to vulnerability of the subjects involved (Das et al., 2008). The other level includes the crashes with possible injuries and non-incapacitating injuries. A similar argument that a possible injury could easily have been a non-incapacitating injury and vice-versa depending on the subjects involved leads us to group the two categories together. Fig. 1(b) shows the complete picture of the modeling concept adopted in the paper. The first step in the analysis compares injury related crashes with no-injury. The second step (nested) analyzes the two broad groups of injury related crashes. This essentially carries out the classification of moderate injuries versus severe injuries.

3. Modeling methodology

3.1. Problems in genetic algorithm

GP, which is a class of evolutionary algorithms, has its roots in the GA. GA is a method to grow from one population to a new population through the process of evolution. For a detailed review of conventional GA the readers are directed to the classical work by Holland (1975) and Goldberg (1989). For the more advanced learners, typically, in GA the representation is generally fixed length representation of length 'l' and the alphabet size is 'k'. In the search space of a fixed length representation of length 'l' and alphabet size 'k' the available candidate solutions are k^l. The initial selection of string length limits the search space and puts restrictions on the learning process of the GA. Thus traditional GA sometimes converges on suboptimal solution. Suboptimal performance may also occur when there is no hill to climb, i.e. if there is a single fitness criterion. For example, in binary classification the criterion is to check whether it goes to the right bin or not. Hence the GA may fail during classification. This observation is critical for the choice of GP over GA in classification problems.

3.2. Genetic programming

According to Koza (1992), "the most natural representation for a solution is a hierarchical computer program rather than a fixed-length character string". The size and shape of the computer program, in other words the complexity of it, is not known apriori. The restrictions in the traditional genetic algorithms has led to the use of the more powerful and versatile genetic programming which takes into account the complexity of problem solving. They use other forms of representations like the tree structure or straight forward one line instructions to the machine. The authors direct the inquisitive reader to the well documented work of Koza (1992) on GP.

In traditional GP, the programs, in the memory, are stored as tree structures. Every tree node has an operator and every leaf node is an operand. This makes the evolution as well as the evaluation of the tree much uncomplicated. The evolutionary biological operations like crossover and mutation are also fairly easy to implement. Typically during crossover there occurs an interchange of sections between two homologous chromosomes at a certain splice point. On the other hand mutation means the alteration of any particular point in a chromosome. Chromosome here refers to the program instructions. With a tree based structure replacing a node, which occurs during the crossover, the whole branch is replaced. The resultant individual is very much different from the parent. In mutation, either the node's information is replaced or the node is removed.

However, in LGP the crossover will occur between two or more instructions' set whereas mutation will occur on a single instruction set. Fig. 2(a) and (b) shows the crossover and mutation occurring in LGP. For example two functions, $g(0)$ and $h(0)$, be two instructions that has to be crossed over. The process of crossover between the two instructions is illustrated in Fig. 2(a). The part of the instructions shown within the ovals will swap places.

As can be observed from Fig. 2(a) that crossover takes place between the branches, along with the operand, resulting in two daughter instructions, $g'(0)$ and $h'(0)$. In Fig. 2(b) the process of mutation in LGP is illustrated. The operand, in this example the division sign, '/', has been circled. This operand can undergo mutation to any other mathematical operand. In this particular example it mutates to the multiplication symbol, '*'.

Typically evolution or development occurs through generation and the fitness of the population, which is typically the evalua-

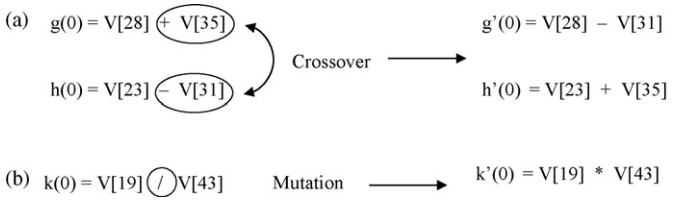

Fig. 2. (a) Crossover between two instructions in LGP and (b) mutation of an operand in an instruction in LGP.

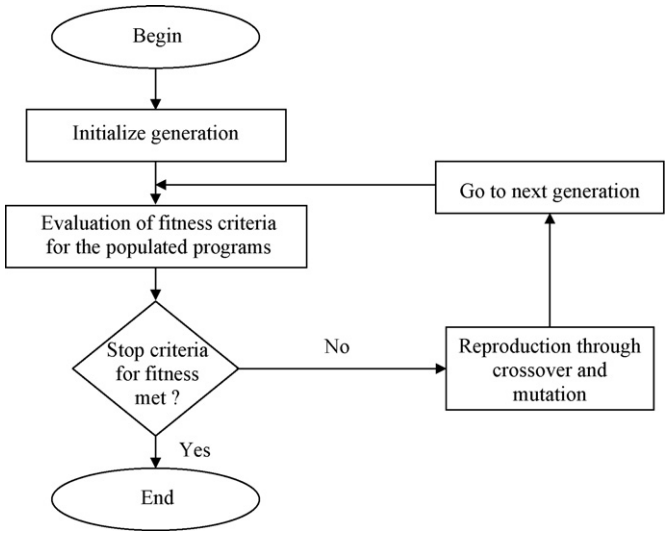

Fig. 3. Typical steps in one generation of GP.

tion criteria, is examined in every generation. Fig. 3 represents the flowchart for a typical generation in GP. The fitness function in this study is the average of the squared errors, where error is the difference between the evolved output and the target output.

However, in LGP the representation of the computer programs is a set of instructions written in the machine language (Brameier and Banzhaf, 2007). The software Discipulus™, which has been used in this study, implements LGP to develop best programs evolved for the problem at hand. Please note that from here on the authors will use LGP term as that is the specific form of GP used. It must be noted that any reference to the term GP, in this particular study, always means the broader category of the heuristic approach.

3.3. *Discipulus*™

Since it is based on LGP, the population is comprised of linear computer programs. From an initial pool of computer programs, a random "tournament" selection from a set of four randomly chosen programs is conducted. The tournament then chooses the two best programs based on the performance on the task designated. These programs are then copied and the standard crossover and mutation operators are applied. The new "child" programs replace the two loser programs and the process repeats till the LGP finds the best program suited for the given task. The software is a multiple-run genetic programming system. The fact that the genetic programming is a stochastic algorithm, running it multiple times yields a wide variety of results. In this particular study for every run 80 generations must pass without improvement for the run to be terminated. Fig. 4 represents the process undergoing in a typical run in the LGP. In each run the population undergoes evolutionary changes through generations.

In the classification problem at hand, the software takes into consideration all the variables that have been fed to it as input. However, all the variables are not included in all the programs as it searches for the programs best fit for the classification under study. The selection of variables is essentially analogous to any regression model where only the significant variables enter the final model from a host of input variables. In LGP too, the various models (programs) have only a select subset of variables and each program has a different classification rule. The Discipulus™ software produces a series of 30 best programs evolved over the runs. The model development process continues till no further minimization of error is observed through further runs. The one with the best classification rate or lowest misclassification rate is chosen by the researcher. The

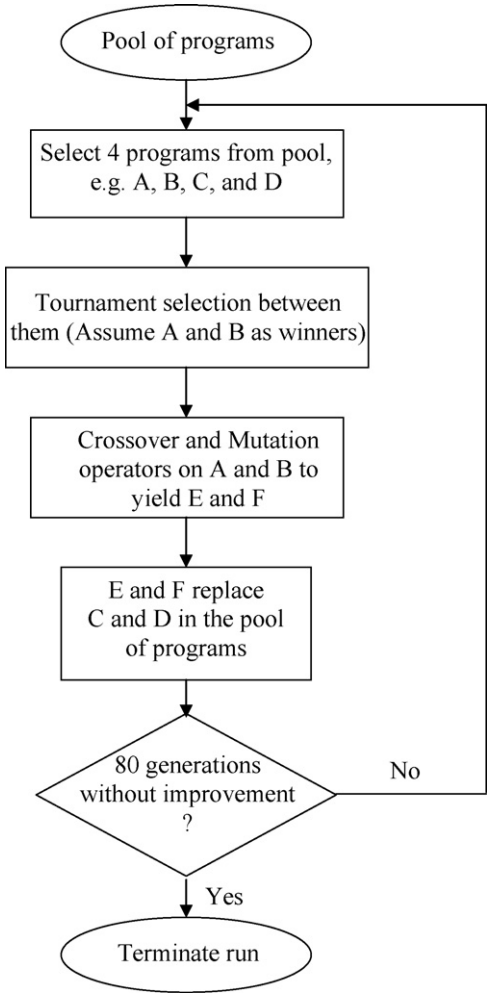

Fig. 4. Flowchart for processes in a typical run.

program contains a series of effective register instructions along with introns (non-effective instructions). In order to find the simplest set of linear instructions, the researcher has to get rid of all the introns. Once the introns are removed the fitness of the program remains unchanged. The final set of instructions is read line-by-line to get the final form of the program. The final program is the classification rule for the particular dataset.

4. Analysis and results

Each of the best programs chosen for the analysis in hand is a set of effective instructions which lead to the final classification rule. Typically for the classification problem the "Class 1 Hit Rate", "Class 0 Hit Rate" and the "Weighted Hit Rate (WHR)" for each of the best programs are provided. Once the criterion is chosen, the set of effective instructions (after the removal of introns) form the classification rule for that particular program. In the present study the WHR has been used as the criteria to select the classification model. The WHR is the analogous to the correct "hits" as described by Swets (1964) in the classical work on signal detection theory. For development of models the primary dataset was split into training and validation datasets consisting of 70% and 30% of the data, respectively. The WHR reported is always for the validation dataset.

As mentioned in the previous section and as illustrated by Fig. 1(a) and (b), the first step in the analysis begins with the classification problem of injury related crashes and no-injury crashes. The second analysis is specifically for the injury severity in which

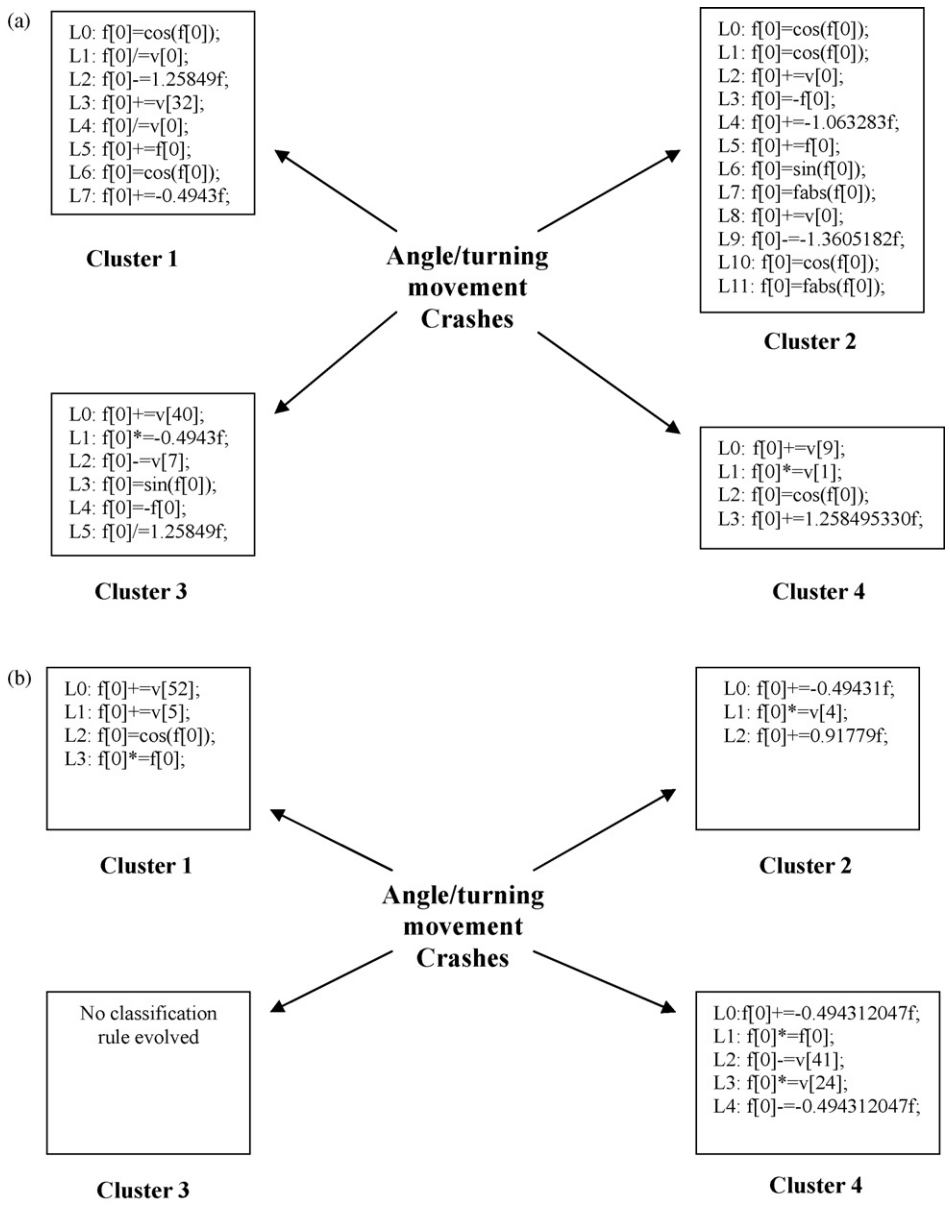

Fig. 5. (a) Injury vs. non-injury analysis for angle/turning movement crashes and (b) severe vs. non-severe analysis for angle/turning movement crashes.

the levels of the binary target variable represent the possible/non-incapacitating injuries and incapacitating injuries/fatalities. Due to data constraints only three types of crashes are considered, namely: (i) angle/turning movement; (ii) rear-end; and (iii) head-on. The results are discussed according to the crash type.

4.1. Angle/turning movement crashes

This particular category of crashes includes all the angle crashes and also the left and the right turn crashes. As previously mentioned the corridors have been categorized into four clusters. Hence, the authors try to explain the results in light of the corridor clusters. This is critical to the understanding of the results; especially the inclusion of the variables which enter the program's set of instructions.

The boxes in Fig. 5(a) indicate the set of instructions (classification rules) that were developed for the particular cluster for the angle/turning movement crashes for the injury and no-injury analysis. The classification rule (represented by '$f(0)$' in the set of instructions) is developed line-by-line. The value of the function '$f(0)$' is initialized to zero. At every step the information is updated through any arithmetic or trigonometric modification with either a variable (refer to Table 1 for all the variables appearing in the results) or a constant. The final value of $f(0)$ is then used to conduct the appropriate classification, based on a threshold value (in this case 0.5). The authors report also the WHR for all the programs mentioned here in the study.

To elaborate more, the authors explain one of the results, for example the result for Cluster 1, from Fig. 5(a). The WHR for the program is 60.4106 which imply that 60.4106% of the cases were classified correctly. As mentioned earlier, at the start of the function the $f(0)$ is initialized to zero. In the first line the cosine value of $f(0)$ is computed. The resulting function is then divided by $V[0]$ (surface width) followed by subtracting a constant and subsequently adding $V[32]$ (truck factor <4.05) again to the function. The value of $f(0)$ is thus calculated at every step and the final value is used for classification. In this study if the final value of $f(0)$ is less than

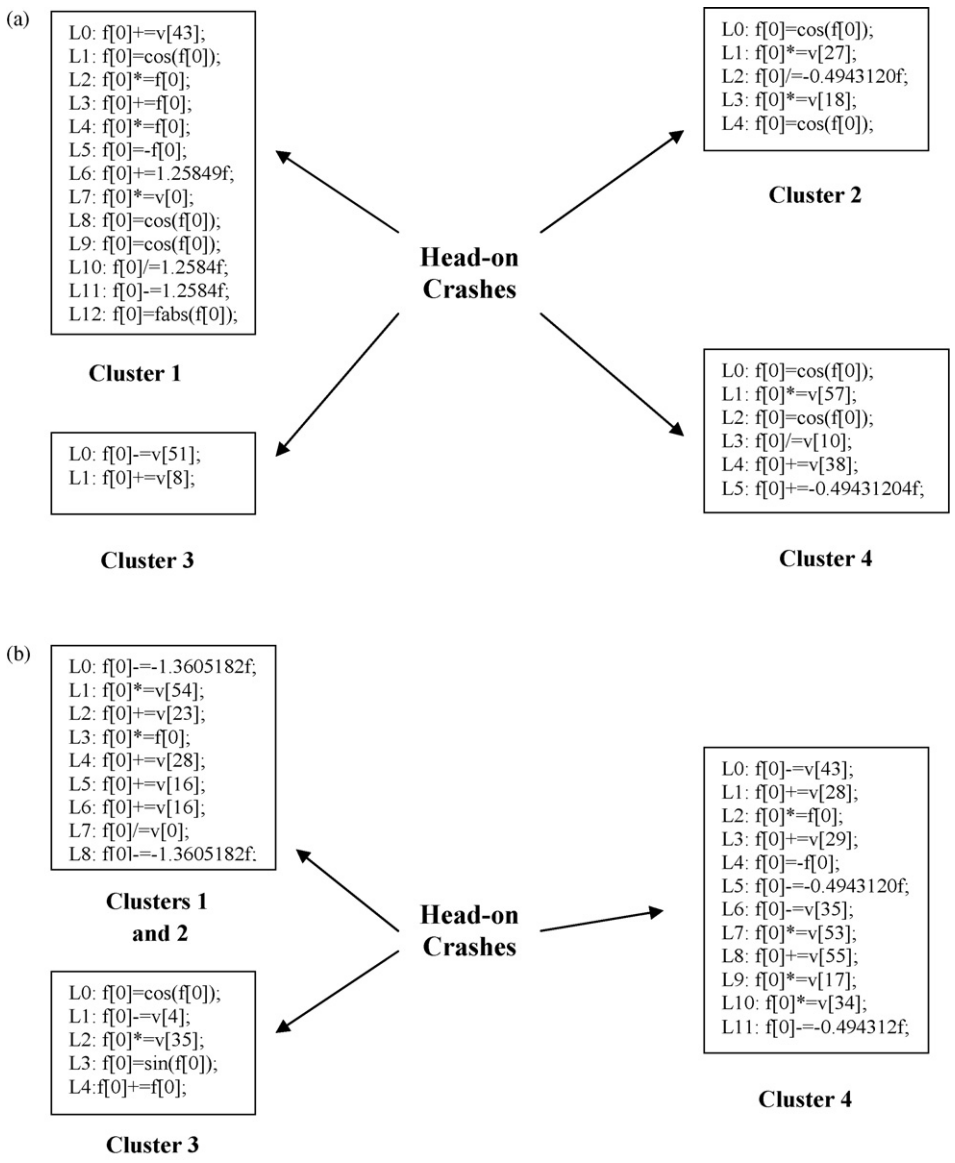

Fig. 6. (a) Injury vs. non-injury analysis for head-on crashes and (b) severe vs. non-severe analysis for head-on crashes.

0.5 then it is classified as a non-injury crash and as a injury crash, otherwise.

As mentioned earlier the corridors in Cluster 1 (1.009–2.89 miles) are the smallest in length. Crashes are most likely to be without injury if the surface width ($V[0]$) is high. Higher surface width gives the driver more maneuvering space and thus more opportunity to take crash avoidance maneuver. Even if the crash does take place, it will mostly likely not to result in an injury. It is interesting to note that even low percentage of trucks on the corridors can result in injuries if a crash occurs. Seriousness of crashes with trucks and other vehicles has been well documented by Bjornstig et al. (2008). Interestingly in Cluster 2 (2.898–5.729 miles) (WHR = 57.8271) corridors higher surface width increases the likelihood of injury in a crash. In Cluster 3 (5.762–10.556 miles) (WHR = 57.7476) corridors, fair pavement condition ($V[40]$) increases the possibility of injury. This indicates that pavement condition has to be good to excellent for safe driving. Deteriorated pavements put the drivers at risk for a crash due to sudden unacceptable changes in the level and also due to poor traction. In Cluster 4 (10.644–78.293 miles) (WHR = 59.514) corridor injuries are more likely to occur when parking is available on higher speed limit segments ($V[9]^*V[1]$). Emphasis on the restrictions of on-street parking has been highlighted in the work of Zegeer et al. (1994).

Fig. 5(b) illustrates the results of the classification between severe and non-severe crashes. For Cluster 1 (WHR = 83.2677) crashes $V[52]$ (variable indicating the presence of a median with curb $\leq 6''$ and lawn) and $V[5]$ (variable indicating dry surface condition) enter the classification rule developed. A careful observation at the entire rule for Cluster 1 indicate that the presence of median with lawn and curb and also dry surface condition decrease the severity of the crash. The cosine function applied on $f(0)$ reduces the value of $f(0)$ when $f(0)$ is higher. Das et al. (2008) also found dry surface conditions to favor less severe crashes probably because of resultant better friction the car is more in control. Hence, even if the crash occurs, the drivers could still be in control. The presence of lawn in the median could help in preventing multi-vehicle which more often results in severe crashes. In Cluster 2 (WHR = 81.944), as the variable $V[4]$ (shoulder plus side walk width) increases the resulting crash tends to be less severe. Fatal crash rates are found to decrease with wider shoulder width (Kweon and Kockelman, 2005). Cluster 4 (WHR = 83.5804) results indicate that with good

pavement condition (*V*[41]) the crash severity will decrease. *V*[24] (curve of roadway between 5° and 8°) also indicate low curvature. The entire rule indicates that with this curvature range the crashes occurring will be less severe. Souleyrette et al. (2001) found that the crash frequency had a direct association with the degree of curvature on horizontal surfaces.

4.2. Head-on crashes

The results for the two types of analysis for the head-on crashes (one for injury and non-injury crashes; the other for severe and non-severe crashes) are illustrated in Fig. 6(a) and (b), respectively. In Cluster 1 (WHR = 70.1923) low skid values (*V*[43]) result in increased likelihood of injury from a crash. Low skid values indicate poor traction control on roads which would increase the chances of loosing control of the vehicle during the event of a crash and thus leading to injury. Reduced friction could also lead to potentially dangerous head injuries on the roadways (Finan et al., 2008). It is interesting to note that in Cluster 2 (WHR = 61.7116) the presence of non-restrictive median at sharper curves (*V*[18]**V*[27]) lead to decreased probability of injuries. In Cluster 3 (WHR = 61.8879) paved median (*V*[51]) is found to decrease the injuries. In Cluster 4 (WHR = 63.2472) crashes occurring during off-peak periods on roadways with surfaces other than blacktop (*V*[38]/*V*[10]) decrease the injury probability. Results in Clusters 2 and 4 of head-on type crashes also indicate the capability of the LGP methodology to discover interaction terms in the injury/no-injury classification.

The results for the severity analysis are illustrated in Fig. 6(b). In this analysis the Clusters 1 and 2 are combined to form one group (for the need of sufficient data). For Clusters 1 and 2 (WHR = 84.5273) variables like *V*[23] (curvature between 4° and 5°) and *V*[28] (ADT ≤ 31,000) increase the chances of severe crashes. Lower ADT means increased possible maneuvers during driving and hence the increased chances of potential conflicts. Lower ADT also indicates higher speeds, given a conflict occurs, would potentially result in severe crashes. Restrictive openings in medians (*V*[16]) also tend to increase the severity of crashes. However the crash severity would decrease with increase in surface width. This is in consistence with findings by Petritsch et al. (2007) who did an evaluation of geometric and operational characteristics for the

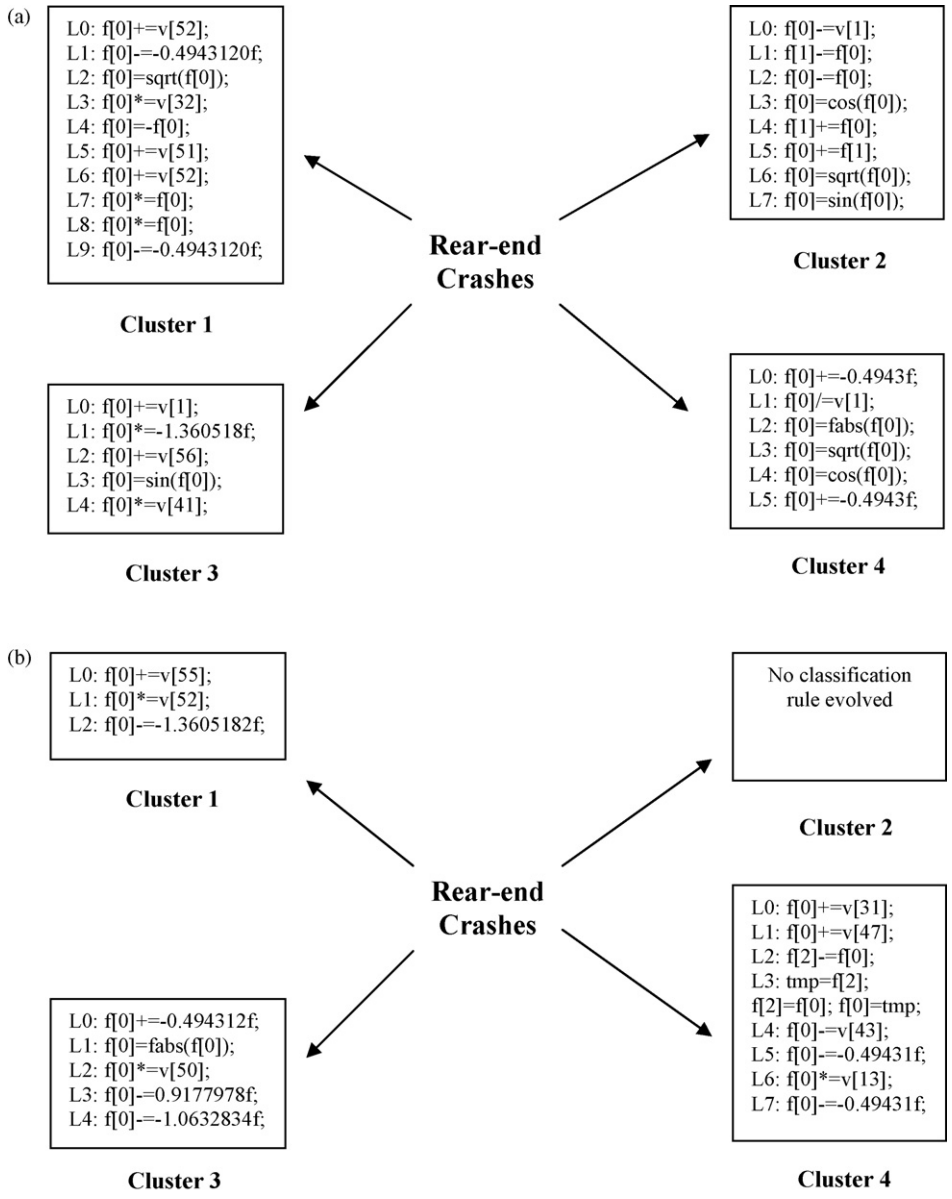

Fig. 7. (a) Injury vs. non-injury analysis for rear-end crashes and (b) severe vs. non-severe analysis for rear-end crashes.

safety of six-lane divided highways for the FDOT. Again in Cluster 3 (WHR = 82.1027), the presence of wide shoulder and side walk (*V*[4]) decrease the severity of crashes. If the crash has occurred during the afternoon peak period (*V*[35]) then the resulting crash would be non-severe. The results are in line with a previous work by the authors (Das et al., 2008). In Cluster 4 (WHR = 81.7513), again ADT less than 40,000 (*V*[28] and *V*[29]) leads to higher severity of injuries. As in Cluster 3, this cluster also has less severe injuries during afternoon peak traffic. Presence of curb and lawn median (*V*[53]) helps avoid crossover head-on crashes or reduce the intensity of it. Hence it would reduce the severity. If a head-on crash occurs on the segment (*V*[55]) then it would be more severe than if it would have occurred at any other roadway element. This could be attributed to higher vehicular speeds on segments than at intersections or access points. Restrictive opening and higher truck factor (*V*[17] and *V*[34]) results in higher severity of crashes. A study by Andreassen (2003) in Australia found that there are areas on corridors which should not have a higher truck percentage. Likewise the corridors with higher truck percentage should be flagged and more administrative measures should be taken to reduce the risk of crash occurrence and imminent severity due to crashes involving trucks.

4.3. Rear-end crashes

The results for the two types of analysis for the rear-end crashes (one for injury and non-injury crashes; the other for severe and non-severe crashes) are illustrated in Fig. 7(a) and (b) respectively. In Cluster 1, the presence of paved and curbed median increase the likelihood of injury, while increase in maximum posted speed limit increase the probability of injury crashes on Cluster 2 corridors. In Cluster 3 rear-end crashes at intersections (*V*[56]) are more injury prone even under good condition of the pavement (*V*[41]). Surprisingly higher posted speed limits tend to be safer in terms of injury occurrence for Cluster 4 corridors. One possible explanation could be that on longer stretches of roadway segments the driver gets used to the speed limit and after a while is more accustomed to the high-speed traffic around it. Hence the injury probability might be reduced as the driver is more aware of the surrounding. The WHRs for the Clusters 1 through 4 are 56.3661, 53.0926, 52.495 and 54.1009 respectively.

In Fig. 7(b) the results for the severity analysis are shown. In Cluster 1 (WHR = 91.8455), crashes related to segment (*V*[55]) and on roadways with curb and lawn median (*V*[52]) give rise to increased severity. In Cluster 3 (WHR = 91.4834) presence of lawn only median (*V*[50]) leads to decreased severity of crashes. Lawn medians are generally wide medians. Wider medians lead to decreased crash rate (Gettis et al., 2005). Even though lawn medians are typically safer for head-on type of crash, yet the very presence of lawn medians can make the drivers make a move towards the lawn in case of imminent rear-end crash situation. This is different from the result obtained in Cluster 1, where curb and lawn median increase the severity. The presence of the curb makes it difficult for the driver to use the median space effectively for the drivers to avoid crashes. This could be a possible explanation as to why the crashes result in higher severity in Cluster 1. In Cluster 4 (WHR = 89.4289) It was observed that *V*[31] (ADT $\geq$ 52,000) causes increased severity of crashes. Thirty-two percent of the crashes have speeds greater than 38 mph and thus indicating that a large number of vehicles were travelling at higher speeds (the number is large as the ADT is high). Thus crashes occurring at higher speeds would more likely lead to a severe crash. This indicates a higher speed variance. For the majority of slower vehicles (<38 mph) severe crashes may occur due to the random aggressive behavior of drivers trying to make their way through a relatively low speed corridor. Nevarez et al. (2009) found ADT per lane to be significantly related to crash severity. A study by Pande and Abdel-Aty (2009) also finds severe rear-end crashes to be significantly related to ADT. A possible explanation to that could be the fact that the rear-end crashes, considered in this study, are occurring on high-speed arterials. In addition to that it must be observed that the severity of rear-end crashes is not entirely dependant on external factors. Rigid seat backs also contribute significantly to severity of injuries in rear-end crashes (Warner and Warner, 2008). Interestingly with the absence of street parking (*V*[13]) the severity is found to diminish.

5. Conclusions

As stressed earlier in the paper, classification is critical to our understanding of the variables of significance and their contribution to the safety problem at hand. In the present study the authors have set up a classification problem for the injury as well as severity of crashes. Typically in a classification problem the algorithm develops a set of rules which when followed leads to a particular category of the target variable. For example, in crash severity analysis when the binary target variable represents severe/non-severe crashes, the classification rule developed would lead to either severe crashes or non-severe crashes.

Classification using trees has been carried out since Breiman et al. (1984) came up with the Classification and Regression Tree (CART) algorithm. Different algorithms have been tried ever since to develop classification models or rules. The advantage or the feature that gives genetic programming the edge over any other existing classification algorithm is the fact that numerous models can be developed for the same dataset. The use of the concept of biological evolution helps the algorithm develop numerous models (by its capacity to perform multiple runs with randomized parameter settings), through the operators like crossover and mutation. A lower crossover frequency and a higher mutation frequency are implemented to prevent genetic drift from taking place. Genetic drift is the accumulation to a suboptimal solution in the search space due to stochastic errors. The process of mutation always brings in novelty to the population of evolved generations. LGP can also assemble teams of models than just individual models which makes it better than most classification algorithms which primarily work on just individual models. The individual models or teams model have been observed to have a lower error rate than other standard classification algorithms. Percent correct classification achieved on the validation dataset for severe/non-severe models were as high as 90% and more as indicated by the WHR values.

As mentioned earlier the two types of analyses carried out in the study includes: (1) injury and non-injury crashes; and (2) severe and non-severe crashes. Some of the results confirm to the traditional well established patterns where as certain other results are not so common and do not confirm to convention. For angle/turning movement crashes presence of parking and higher posted speed limits are responsible for more injury related crashes. Even low percentage of trucks can increase the chance of injury prone crashes. 'Curb and lawn' median and dry surface conditions decrease the severity of crashes where as poor pavement condition result in more severe crashes. Wider shoulders along with sidewalk also tend to make the roads safer from a severity point of view.

In case of head-on crashes low ADT and median openings are the leading operational and geometric factors for severe crashes. Again wide shoulder and sidewalk result in less severe crashes. Crashes occurring on afternoon weekday peak periods also tend to be less severe. Lower skid resistance and the presence of 'curb and lawn' medians are again found to diminish the severity of the crash. Higher truck factor also results in increased severity of head-on crashes. Low skid values increase the injury probability

of a crash while crashes occurring during off-peak periods are less injury prone.

Rear-end crashes at intersections are more likely to be injury prone as well as those at paved and curbed median segments of the roadways. Unlike the angle/turning movement and the head-on crashes, the 'lawn and curb' median causes increased severity in rear-end crashes and similarly for higher ADT values. Absence of street parking also decreases the severity of rear-end crashes.

The results from the Linear Genetic Programming classification are intuitive and their association with severity may be explained. Certain known results about severity of crashes have been confirmed while some new information is discovered about others. The 'lawn and curb' median are found to be safe for angle/turning movement crashes and not so safe for rear-end crashes. Vision obstruction is a leading factor of severe crashes. Dry surface conditions, good pavements also reduce the severity of crashes. On-street parking, higher posted speed limits and lighting conditions do play a role in both injury related crashes and severe crashes.

It can be observed from the results that a lot of interaction terms are discovered in the classification approach for injury/no-injury and severe/non-severe crashes. The heuristic approach that LGP applies has been observed to shed new light on the interaction between variables discussed in this study.

As it can be observed most of the variables of concern relate to geometric and operation factors. Event specific variables have not been included in this study for the sake of interpretability, generalization and the objectives of this study. However, it should be noted that the analysis could be carried with only those variables or by mixing them with geometric and traffic parameters. This could be a part of future investigation. On-street parking has been found to be a hazard for severe injury. Steps should be taken to either remove the facilities for parking or in the case where it is not possible, to restrict the parking hours. Pavement condition should be improved and wherever possible, 'curb and lawn' median should be designed. Higher truck percentage is found to increase severity; hence steps such as lane restriction for trucks or rerouting them from flagged corridors should be taken. Betterment of lighting conditions on the roadways is always desired. Vision obstruction has traditionally been a problem; that however, is not only due to external factors. Nevertheless, transportation authorities should always take design initiatives for the drivers to have a clear view of the surroundings.

For future research the LGP models developed could be compared to traditional classification models like CARTs and Random Forests and a separate comparative analysis could be reported. The categories of variables selected for research could also be enhanced in future work.

References

Abdel-Aty, M., Pande, A., 2006. Comprehensive analysis of relationship between real-time traffic surveillance data and rear-end crashes on freeways. Transportation Research Record 1953, 31–40.

Abdel-Aty, M., Wang, X., 2006. Crash estimation at signalized intersections along corridors: analyzing spatial effect and identifying significant factors. Transportation Research Record 1953, 98–111.

Abdel-Aty, M., Keller, J., 2005. Exploring the overall and specific crash severity levels at signalized intersections. Accident Analysis and Prevention 37 (3), 417–425.

Andreassen, D., 2003. Aspects of road design and trucks from the analysis of crashes. In: Institution of Professional Engineers New Zealand (IPENZ) Transportation Group Technical Conference Papers 2003.

Bjornstig, U., Bjornstig, J., Eriksson, A., 2008. Passenger car collision fatalities—with special emphasis on collisions with heavy vehicles. Accident Analysis and Prevention 40 (1), 158–166.

Brameier, M., Banzhaf, W., 2007. Linear Genetic Programming. Springer, New York.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth International Group, Belmont, CA.

Ceylan, H., Bell, M.G.H., 2004. Traffic signal timing optimisation based on genetic algorithm approach, including drivers' routing. Transportation Research Part B: Methodological 38 (4), 329–342.

Chang, N.B., Chen, W.C., 2000. Prediction of PCDDs/PCDFs emissions from municipal incinerators by genetic programming and neural networking modeling. Waste Management and Research 18, 341–351.

Das, A., Pande, A., Abdel-Aty, M., Santos, J.B., 2008. Urban arterial crash characteristics related with proximity to intersections and injury severity. Transportation Research Record 2083, 137–144.

Deschaine, L.M., Francone, F.D., 2004. White paper: comparison of Discipulus (Linear Genetic Programming software with Support Vector Machines, Classification Trees, Neural Networks and Human Experts). http://www.rmltech.com/Comparison.White.Paper.pdf (accessed 7.02.08.).

Finan, J.D., Nightimgale, R.W., Myers, B.S., 2008. The influence of reduced friction on head injury metrics in helmeted head impacts. Traffic Injury Prevention 9 (5), 483–488.

Gettis, J.L., Balakumar, R., Duncan, L.K., 2005. Effects of rural highway median treatments and access. Transportation Research Record 1931, 99–107.

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Longman Publishing Co., Inc., Boston, Massachusetts.

Holland, J.M., 1975. Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor.

Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

Koza, J.R., 1992. Genetic Programming: On The Programming of Computers by Means of Natural Selection. MIT Press, Cambridge, Massachusetts.

Kweon, Y.J., Kockelman, K.M., 2005. Safety effects of speed limit changes: use of panel models, including speed, use, and design variables. Transportation Research Record 1908, 148–158.

Makkeasorn, A., Chang, N.B., Beaman, M., Wyatt, C., Slater, C., 2006. Soil moisture estimation in a semiarid watershed using RADARSAT-1 satellite imagery and genetic programming. Water Resources Research 42.

National Highway Traffic Safety Administration, 2007. Traffic Safety Facts 2006: A Compilation of Motor Vehicle Crash Data From the Fatality Analysis Reporting System and the General Estimate System. National Highway Traffic Safety Administration, Washington, D.C.

Nevarez, A., Abdel-Aty, M., Wang, X., Santos, J.B., 2009. Large-scale injury severity analysis for arterial roads: modeling scheme and contributing factors. In: Presented at the 88th Annual Meeting of the Transportation Research Board, Washington, DC.

Pande, A., Abdel-Aty, M., 2008. Discovering indirect associations in crash data using probe attributes. Transportation Research Record 2083, 170–179.

Pande, A., Abdel-Aty, M., 2009. Patterns in severe crashes on segments of multilane arterials with partially limited access. In: Presented at the 88th Annual Meeting of the Transportation Research Board, Washington, DC.

Park, B., Messer, C.J., Urbanik II., T., 2000. Enhanced genetic algorithm for signal-timing optimization of oversaturated intersections. Transportation Research Record 1727, 32–41.

Petritsch, T.A., Challa, S., Huang, H., Mussa, R., 2007. Evaluation of Geometric and Operational Characteristics Affecting the Safety of Six-lane Divided Roadways. Sprinkle Consulting, Inc., Florida Department of Transportation.

Souleyrette, R., Kamyab, A., Hans, Z., Knapp, K.K., Khattak, A., Basavaraju, R., Storm, B., 2001. Systematic Identification of High Crash Locations. Center for Transportation Research and Education, Iowa Department of Transportation.

Swets, J.A., 1964. Signal Detection and Recognition by Human Observers. Wiley, New York.

Teklu, F., Sumalee, A., Watling, D., 2007. A genetic algorithm approach for optimizing traffic control signals considering routing. Journal of Computer-Aided Civil and Infrastructure Engineering 22 (1), 31–43.

Wang, X., Abdel-Aty, M., Nevarez, A., Santos, J.B., 2008. Investigation of safety influence area for four-legged signalized intersections: nationwide survey and empirical inquiry. Transportation Research Record 2083, 86–95.

Warner, M.H., Warner, C.Y., 2008. Fatal and Severe Injuries in Rear Impact: Seat Stiffness in Recent Field Accident Data. SAE International.

Zegeer, C.V., Huang, H.F., Stutts, J.C., Rodgman, E., Hummer, J.E., 1994. Commercial bus accidents characteristics and roadway treatments. Transportation Research Record 1467, 14–22.

Yamamoto, T., Hashiji, J., Shankar, V.N., 2008. Underreporting in traffic accident data, bias in parameters and the structure of injury severity models. Accident Analysis and Prevention 40 (4), 1320–1329.