Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/100223

Using Classification and Regression Tree and Dimension Reduction in Analyzing Motor Vehicle Traffic Accidents

Yu-Huei Liu[1,2], *Kuang-Yang Kou*[3,*], *Hsin-Hung Wu*[4] *and Ya-Chi Nian*[4,5]

[1] Graduate Institute of Integrated Medicine, China Medical University, Taichung, Taiwan.
[2] Department of Medical Genetics and Medical Research, China Medical University Hospital, Taichung, Taiwan.
[3] Department of Traffic Science, Central Police University, Taoyuan, Taiwan.
[4] Department of Business Administration, National Changhua University of Education, Changhua, Taiwan.
[5] Department of International Trade, National Hsinchu Commerce and Vocational High School, Hsinchu, Taiwan.

Received: 20 Aug. 2015, Revised: 7 Nov. 2015, Accepted: 8 Nov. 2015
Published online: 1 Mar. 2016

Abstract: This study applies classification and regression tree (CRT) to identify the hidden knowledge in fatal accidents of motor vehicles from Fatal Traffic Accident of National Police Agency, Taiwan. In the beginning, twenty four variables are chosen from Fatal Traffic Accident data set. Later, dimension reduction is used to reduce the number of variables from twenty four to nine variables by principal component analysis. With two different CRT models with twenty four and nine variables to forecast injury severity, a comparison is made in terms of rules generated, model accuracy, type I and type II errors, and evaluation chart generated by IBM SPSS Modeler 14.2. The results show that the CRT model with dimension reduction outperforms the CRT model without dimension reduction almost in every category except for type II error since this model tends to slightly overestimate the injury severity of motor vehicle traffic accidents than the model without dimension reduction.

Keywords: fatal traffic accident, motor vehicle, classification and regression tree, data mining, type I error, type II error, model accuracy

1 Introduction

According to the website information from Ministry of Transportation and Communications, Taiwan, there were 21,374,175 registered vehicles in December 2009, including 14,604,330 motor vehicles, accounted for more than 68 percent. In addition, the motor vehicle accidents ratio has been steadily increased from 34.12% in 2005, 40.45% in 2006, 41.37% in 2007, 44.14% in 2008, and to 43.75% in 2009 and is the highest among all vehicles. Motor vehicles so far play an important means for daily transportation in Taiwan but the data show that people tend to lack the sense of crises in road safety.

When a massive amount of traffic accidents data has been cumulated, the decision maker faces an important but critical issue to convert the data into useful information to make quality decisions [1,2,3]. In the past, statistical methods were typically applied [4]. However, before the use of statistical analysis, data should be collected, re-arranged, coding, and sampling due to the complexity of the data set. Besides, there might be some human errors to possibly remove the causes of the accidents, leading to neglect potentially important factors [5]. Furthermore, statistical analysis tends to focus on the hypotheses and validation with relatively smaller amount of the data, which might limit the effectiveness of traffic safety analysis [4].

Data mining, on the other hand, uses algorithms to actively search the meaningful rules and then to identify unknown and hidden knowledge, which is the biggest difference between the traditional statistical analysis and data mining [4,5]. In this study, Fatal Traffic Accidents data set from 2005-2007 provided by National Police Agency, Ministry of the Interior of Executive Yuan in Taiwan consisting of both continuous and categorical variables is used to analyze the important variables of injury severity in motor vehicle traffic accidents. In addition, classification and regression tree (CRT) will be applied to identify important variables of motor vehicle accidents and generate rules regarding the traffic

* Corresponding author e-mail: peterkou@mail.cpu.edu.tw

accidents. The reason why CRT is chosen in this study is that this approach can be applied to both continuous and categorical variables and can be used to generate the simplicity of the results [6,7]. Furthermore, Rovlias and Kotsou [8] pointed out that CRT is able to make predictions from the data set by incorporating the independent variables that best predict the outcome of the dependent variable.

By classification and regression tree, the important variables of motor vehicle accidents can be identified, and rules regarding the traffic accidents can be established. Further, dimension reduction performed by principal component analysis, one of the very common techniques applied in practice, will be utilized to reduce the number of variables since using too many variables may lead to over fitting and complicate the interpretation of the analysis [4,5,9,10,11,12]. The selected variables by dimension reduction will be used in classification and regression tree. Finally, a comparison between the results of two CRT models before and after dimension reduction will be performed to identify the differences in order to possibly forecast the injury severity in traffic accidents of motor vehicles.

2 Literature review

2.1 Definition of data mining

Frawley et al. [13] stated that data mining can be defined as the nontrivial extraction of implicit and previously unknown and potentially useful information from data set. Chen et al. [14] concluded that data mining can be referred to as knowledge discovery in databases, which is a process to extract knowledge rules, constraints, and regularities from data in databases. Kleissner [15] defined data mining is a process but not a one-time activity for an organization. In contrast, data mining is a commitment of an organization to leverage its business data for an ongoing basis to continuously and iteratively improve the business practices based on a new level of understanding of the data set. Fayyad [16] pointed out that data mining is a centerpiece of an analytics strategy by identifying interesting patterns and developing predictive models from data for an organization to deliver business values. Wu and Chen [17] summarized that data mining is useful in various areas such as market analysis, decision support, fraud detection, and the like, and many approaches have been proposed to extract information from the large amount of the data.

2.2 Classification and regression tree

Classification and regression tree is one of the decision tree algorithms for classification by constructing a flowchart-like structure where each internal node represents a test on an attribute, each branch denotes an outcome of the test, and each external node means a class prediction [9,11,18, 19]. The characteristic of CRT is to use a set of ”if-then” conditions to perform predictions or classification of cases [4,6]. Thus, Razi and Athappilly [6] stated that CRT is very suitable to tackle large problems or smaller data set with both continuous and categorical variables. In fact, the attribute that is not appeared in the tree is assumed to be irrelevant in the analysis. Therefore, the set of attributes appearing in the tree forms the reduced subset of attributes.

The major advantages of CRT are summarized below by Hill and Lewicki [20]. First, the interpretation of the results in a tree is very simple to explain why observations are classified into a particular manner. Second, there is no implicit assumption that the underlying relationships between the predictor variables and the dependent variables are to be linear or follow some specific non-linear link function since CRT inherents non-parametric and non-linear properties. Finally, CRT is very suitable for data mining because little knowledge on any coherent set of theories or predictions regarding which variables are related and how are to be known in advance [21,22].

2.3 Dimension reduction

The database used in data mining typically might have various variables, and it is unlikely that all of the variables are independent without correlation structure among them [23,24]. Data analysts need to guard against multicollinearity, which might lead to instability in the solution space and possible incoherent results. Larose [5] also pointed out that the use of too many predictor variables to model a relationship with a response variable can unnecessarily complicate the interpretation of the analysis and violate the principle of parsimony that one should keep the number of predictors to a size. Bi et al. [25] stated that selecting appropriate variables can enhance the effectiveness and domain interpretability of an inference model. In order to reduce the effects of the correlation structure among the predictor variables, dimension reduction methods are typically applied to reduce the number of predictor variables, to help ensure these components are independent, and to provide a framework for interpretability of the results [5].

Principal component analysis is one of dimension reduction methods and seeks to explain the correlation structure of a set of predictor variables using a smaller set of linear combinations (components) of these variables [5,9,11,19]. To determine the number of components to be extracted, eigenvalue criterion can be used by selecting the component with the eigenvalue greater than one [5].

3 Research method

This study uses IBM SPSS Modeler 14.2 to perform classification and regression tree. The mode of CRT is set to ”Expert”. In addition, the impurity measure for categorical targets is set to ”Gini”, while the values of maximum surrogates, minimum change in impurity, and prune tree use default values in the software. Specifically, maximum surrogates and minimum change in impurity are set to 5 and 0.0001, respectively. The stopping criteria are based on the percentage with minimum records in parents branch (%) of two and minimum records in child branch (%) of one.

The objective of this study is to discuss how the variables would affect injury severity in motor vehicle traffic accidents. In this study, two types of injury are taken into account, namely death and injured. In the beginning, twenty four variables which might be considered as important variables subjectively chosen by the authors from Fatal Traffic Accidents of National Police Agency include age, speed limit, weather, light, road type, road pattern, accident location, road coverage, road condition, road defect, obstacle, sight distance, type of signal, action of signal, traffic lane differentiated facility, accident type and pattern, gender, protection equipment, mobile phone, status of concerned motor vehicles and people, driving qualification, collided part of vehicle, occupation, and traveling purpose. The number of motor vehicle traffic accidents is 6,256. In this study, 80% of the data set is used for training, while the rest of the data set is for testing.

In the second part of the study, dimension reduction is performed by PASW Statistics 18 based on these twenty four variables. With the eigenvalue greater than one as shown in Table 1, only nine variables are left, including weather, road pattern, accident location, road condition, road defect, type of signal, action of signal, status of concerned motor vehicles and people, and driving qualification based on the information of rotated factor matrix as shown in Table 2. The cumulative percentage of total variance explained is 61.447%. The information regarding these nine variables is summarized in Table 3.

Table 1: Total Explained Variance by Dimension Reduction

Component	Initial Eigenvalue		
	Total	% of Variance	Cumulative %
1	3.206	12.825	12.825
2	2.619	10.478	23.302
3	1.751	7.005	30.308
4	1.616	6.464	36.772
5	1.375	5.498	42.270
6	1.298	5.192	47.463
7	1.249	4.997	52.460
8	1.178	4.711	57.170
9	1.069	4.277	61.447
10	.980	3.918	65.365
11	.934	3.735	69.100
12	.887	3.546	72.647
dimension 13	.810	3.240	75.887
14	.785	3.141	79.028
15	.734	2.935	81.963
16	.706	2.823	84.786
17	.652	2.606	87.392
18	.624	2.496	89.888
19	.591	2.366	92.254
20	.540	2.159	94.413
21	.461	1.843	96.256
22	.324	1.297	97.552
23	.308	1.233	98.786
24	.213	.852	99.638
25	.091	.362	100.000

Component	Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %
1	3.206	12.825	12.825
2	2.619	10.478	23.302
3	1.751	7.005	30.308
4	1.616	6.464	36.772
5	1.375	5.498	42.270
6	1.298	5.192	47.463
7	1.249	4.997	52.460
8	1.178	4.711	57.170
9	1.069	4.277	61.447
10			
11			
12			
dimension 13			
14			
15			
16			
17			
18			
19			
20			
21			
22			
23			
24			
25			

Component	Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %
1	3.092	12.368	12.368
2	2.366	9.464	21.832
3	1.695	6.780	28.612
4	1.535	6.139	34.751
5	1.533	6.132	40.883
6	1.388	5.552	46.435
7	1.330	5.319	51.755
8	1.268	5.071	56.826
9	1.155	4.621	61.447
10			
11			
12			
dimension 13			
14			
15			
16			
17			
18			
19			
20			
21			
22			
23			
24			

4 Results

By considering twenty four variables, six variables are chosen by classification and regression tree with weights. These variables are status of concerned motor vehicles and people, mobile phone, occupation, protection equipment, driving qualification, and collided part of vehicle with the respective weights of 0.90, 0.02, 0.02, 0.02, 0.02, and 0.02. In addition, the tree depth is one and Figure 1 depicts the classification tree of the model with twenty four variables, where 1 and 2 in injury severity represent death and injured, respectively. Besides, two rules are generated and summarized in Table 4, where the notations of status of concerned motor vehicles and

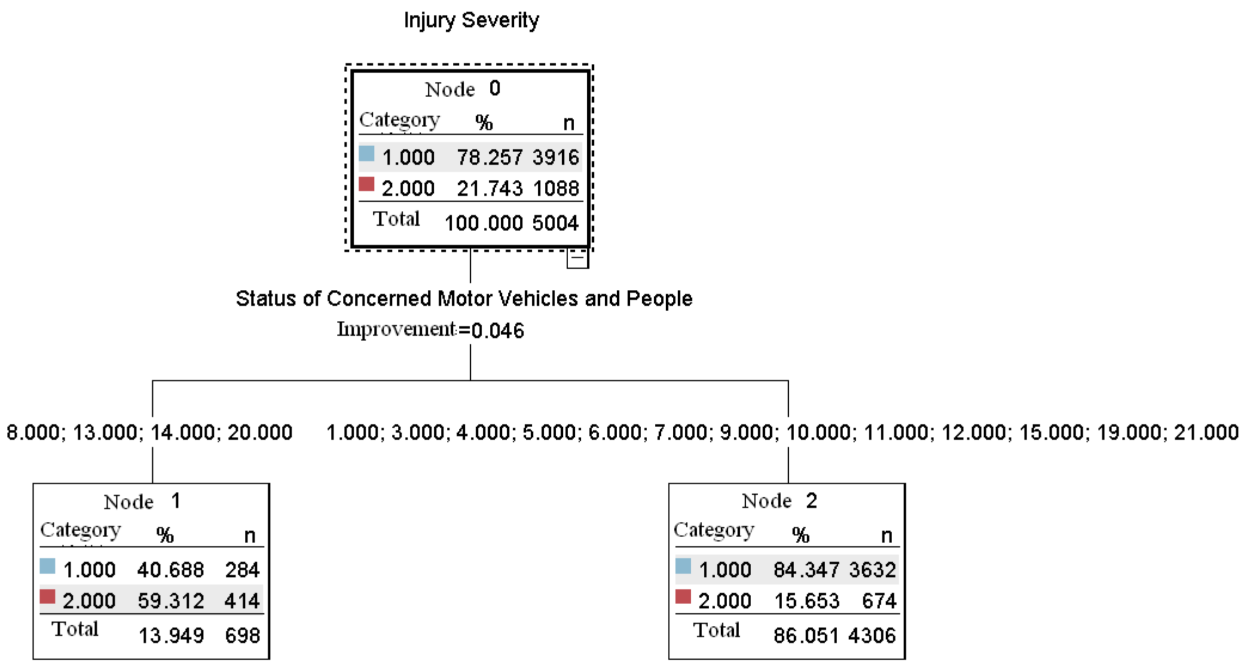


Fig. 1: Tree Plot with Twenty Four Variables

Component	Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %
25			

people are in Table 3. Therefore, the major variable is status of concerned motor vehicles and people.

When a motor vehicle's status is in initial starting, during parking operation, overtaking, turning left, turning right, turning left to change lane, straight on, inserting into the queue, turning over or crossing the road, or still (engine off), the injury tends to be death. When a person's status is getting on and off the motor vehicle, the injury tends to be death, too. In contrast to death, when a motor vehicle's status is in turning right to change lane, still (engine off), and stopping (engine on), the injury tends to be injured.

When nine variables selected by principal component analysis are the input variables for the CRT model, only three variables have weights. These three variables are status of concerned motor vehicles and people, driving qualification, and accident location with the corresponding weights of 0.96, 0.03, and 0.01. The tree depth is two and Figure 2 shows the classification tree of the model with only nine variables. Besides, three rules are generated by CRT and shown in Table 5. The major variable is status of concerned motor vehicles and people.

Two rules for death are (1) a motor vehicle's status is in initial starting, during parking operation, overtaking, turning left, turning right, turning left to change lane, straight on, inserting into the queue, turning over or crossing the road, or still (engine off) and (2) a motor vehicle's status is in turning right to change lane, still (engine off), and stopping (engine on) together with the accident locations of motor vehicle staging area, U turn lane, express way, permissive motor vehicle lane, or pavement. In contrast, one rule is for injured scenario. The only rule for injured is a motor vehicle's status is in turning right to change lane, still (engine off), and stopping (engine on) together with the accident locations of inside the fork, near the fork, traffic island (including channelizing lines), carriage way, ordinary way (not falling into express or carriage way), motor vehicle lane, road shoulder and curb, or near the crosswalk.

By comparing the CRT models before and after dimension reduction, the model with dimension reduction provides more specific rules and much more information to depict both death and injured scenarios. Besides, Table 6 shows that the CRT model with dimension reduction has better performance in terms of forecasting accuracy

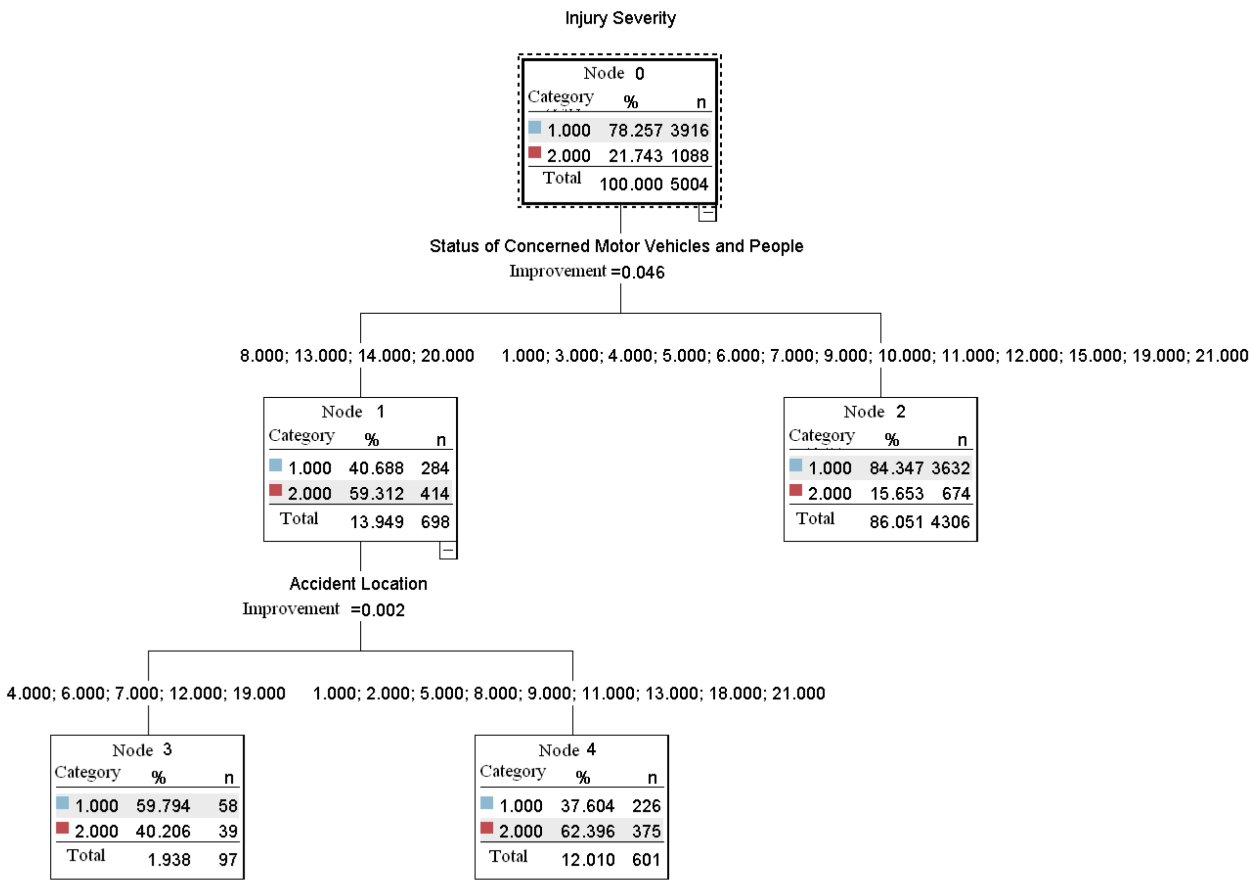


Fig. 2: Tree Plot with Nine Variables by Dimension Reduction

for both training and testing data sets. It is worth noting that type I and type II errors might be another measurement to compare these two CRT models. In this study, type I error is defined as the null hypothesis is true (H_0: death) but rejects it, while type II error is defined as the null hypothesis is not true (H_1: injured) but accepts the null hypothesis. From Table 7, type I errors for both training and testing data sets of CRT model with nine variables outperform those of CRT model with twenty four variables. In contrast to type I error, CRT model with twenty four variables has smaller type II errors in both training and testing data sets. By the overall evaluation in terms of type I and type II errors, CRT model with dimension reduction is a better model to be chosen though its type II errors are relatively larger. This indicates that this model tends to overestimate the outcome of fatal traffic accidents in motor vehicles. From managerial viewpoints, overestimating the outcome of fatal traffic accidents in motor vehicles might be better than underestimating the outcome of fatal traffic accidents in order for people to be aware of traffic safety.

Evaluation chart is also provided in Figure 3 to make a comparison between these two models. The light blue and red lines represent the best and base lines, where the performance for each model falls within these two lines. Besides, if the performance line of a particular model is closer to the best line, it indicates that the model performs better. From Figure 3, the performance lines of CRT ($R-Injury Severity) and CRT with dimension reduction ($R1-Injury Severity) overlap. That is, these two models have the similar performance in terms of evaluation chart. Therefore, by the above discussions and analyses, dimension reduction by principal component analysis helps to reduce the complexity in motor vehicle traffic accidents analysis and even provide better outcomes in

Table 2: Information of Rotated Factor Matrix

	Component		
	1	2	3
Vehicle Type	-.061	.032	.104
Age	-.152	-.224	.196
Speed Limit	-.057	.049	.121
Weather	-.121	.014	**.836**
Light	.247	.040	-.258
Road Type	.001	-.058	-.081
Road Pattern	**.848**	.002	.102
Accident Location	**.802**	-.013	.011
Road Coverage	-.026	-.007	-.121
Road Condition	-.119	.017	**.829**
Road Defect	-.072	-.003	.114
Obstacle	-.091	-.009	.080
Sight Distance	-.129	.009	.105
Type of Signal	**.841**	-.039	.112
Action of Signal	**.858**	-.032	.138
Traffic Lane Differentiated Facility	-.376	-.036	-.145
Accident Type and Pattern	.196	.000	-.064
Gender	-.160	.324	-.003
Protection Equipment	.117	.394	.051
Mobile Phone	.133	.650	-.017
Status of Concerned Motor Vehicles and People	.005	**.835**	-.051
Driving Qualification	-.005	**.840**	-.009
Collided Part of Vehicle	-.076	.661	-.001
Occupation	-.082	.131	.231
Traveling Purpose	-.009	.126	.241

	Component		
	4	5	6
Vehicle Type	.181	.402	.043
Age	.234	.423	.213
Speed Limit	-.569	.016	.198
Weather	.083	-.284	-.160
Light	-.108	-.392	.293
Road Type	.671	.068	-.202
Road Pattern	-.092	.087	-.022
Accident Location	.071	-.011	.008
Road Coverage	.305	-.147	.011
Road Condition	.058	-.300	-.149
Road Defect	-.191	.100	-.033
Obstacle	-.127	.090	-.039
Sight Distance	-.093	.113	.176
Type of Signal	.107	.158	-.140
Action of Signal	.071	.176	-.144
Traffic Lane Differentiated Facility	.562	-.034	-.174
Accident Type and Pattern	.291	-.264	.214
Gender	-.047	.358	-.346
Protection Equipment	.259	-.289	.350
Mobile Phone	.141	-.214	.227
Status of Concerned Motor Vehicles and People	-.041	.009	-.097
Driving Qualification	-.018	.055	-.116
Collided Part of Vehicle	-.028	.178	-.190
Occupation	.043	.389	.439
Traveling Purpose	.231	.275	.584

	Component		
	7	8	9
Vehicle Type	-.142	.398	.200
Age	-.137	.385	.036
Speed Limit	-.142	.243	.053
Weather	-.057	-.050	.035
Light	.104	-.203	.034
Road Type	.227	-.206	-.040
Road Pattern	-.021	-.053	.078
Accident Location	-.023	-.045	.223
Road Coverage	-.253	-.196	.518
Road Condition	-.043	-.058	.017
Road Defect	**.706**	.098	.133
Obstacle	.696	.044	.252
Sight Distance	.044	.020	.473
Type of Signal	.060	.071	-.115
Action of Signal	.047	.059	-.125
Traffic Lane Differentiated Facility	.152	.104	-.121
Accident Type and Pattern	-.015	.131	.485
Gender	-.138	.083	.132
Protection Equipment	.125	.451	-.219
Mobile Phone	.046	.308	-.109
Status of Concerned Motor Vehicles and People	-.014	-.114	.058
Driving Qualification	.013	-.100	.023
Collided Part of Vehicle	-.055	-.094	.090
Occupation	.047	-.450	-.135
Traveling Purpose	.035	-.303	-.067

Table 3: Nine Variables and Notations after Dimension Reduction

Variable	Notations for each Variable
Weather	(1) tempest (2) gale (3) sandy wind (4) fog or smoke (5) snow (6) rainy (7) cloudy (8) sunny
Road pattern	(1) level crossing with remote control (2) level crossing without remote control (3) three-fork road (4) four-fork road (5) multi-fork road (6) tunnel (7) underpass (8) bridge (9) culvert (10) viaduct (11) curved road and its vicinity (12) slope way (13) lane (14) straight road (15) others (16) loop (17) square
Accident location	(1) inside the fork (2) near the fork (3) motorcycle waiting area (4) motorcycle staging area (5) traffic island (including channelizing lines) (6) U turn lane (7) express way (8) carriage way (9) ordinary way (not falling into express or carriage way) (10) bus lane (11) motorcycle lane (12) permissive motorcycle lane (13) road shoulder and curb (14) acceleration lane (15) deceleration lane (16) ring road (17) crosswalk (18) near the crosswalk (19) pavement (20) near the toll station (21) others
Road condition	(1) ice and snow (2) slippery (3) muddy (4) humid (5) dry
Road defect	(1) soft surface (2) rugged surface (3) with pits (4) no defects
Type of signal	(1) traffic control signal (2) pavement control signal (with pedestrian signals)(3) flashing signal (4) no signals
Action of signal	(1) normal (2) abnormal (3) no actions (4) no signals
Status of concerned motor vehicles and people	(I) motorcycle's status: (1) initial starting (2) backing (3) during parking operation (4) overtaking (including surpassing) (5) turning left (6) turning right (7) turning left to change lane (8) turning right to change lane (9) straight on (10) inserting into the queue (11) turning over or crossing the road (12) emergency deceleration or stop (13) still (engine off) (14) stopping (engine on) (15) others (II) person's status: (16) walking (17) still (stopping) (18) running (19) getting on and off the car (20) others (III) uncertain: (21) uncertain
Driving qualification	(1) with proper license (2) without license (under the age for license examination) (3) without license (reaching the age for license examination) (4) over-grade driving (5) license is withheld (6) license is withdrawn (7) uncertain (8) non-car driver

Table 4: Rules with Twenty Four Variables

Rule for 1(Death) - contains 1 rule
Status of concerned motor vehicles and people in [1,4,5,6,7,8,10,11,12,13,16,20,22]
Rule for 2 (Injured) - contains 1 rule
Status of concerned motor vehicles and people in [9,14,15,21]

Table 5: Rules with Nine Variables

Rule for 1(death) - contains 2 rules	
Rule 1	Status of concerned motor vehicles and people in [1,4,5,6,7,8,10,11,12,13,16,20,22]
Rule 2	Status of concerned motor vehicles and people in [9,14,15,21] & Accident location in [5,7,8,13,20]
Rule for 2 (injured) - contains 1 rule	
Rule 1	Status of concerned motor vehicles and people in [9,14,15,21] & Accident location in [1,2,6,9,10,12,14,19,22]

Table 6: A Comparison of Forecasting Accuracy by Two CRT Models

CRT Model with Twenty Four Variables		
Partition	Training	Testing
Correct (%)	4,046 (80.86%)	1,014 (80.99%)
Wrong (%)	958 (19.14%)	238 (19.01%)
Total	5,004	1,252
CRT Model With Nine Variables		
Partition	Training	Testing
Correct (%)	4,065 (81.24%)	1,021 (81.55%)
Wrong (%)	939 (18.76%)	231 (18.45%)
Total	5,004	1,252

Table 7: Types I and II Errors Summary

CRT Model with Twenty Four Variables			
		Predicted death	Predicted injured
Training Data Set	Death in traffic accidents	3,632	284 Type I Error: 7.25%
	Injured in traffic accidents	674 Type II Error: 61.95%	414
Testing Data Set	Death in traffic accidents	924	64 Type I Error: 6.48%
	Injured in traffic accidents	174 Type II Error: 65.91%	90
CRT Model with Nine Variables			
		Predicted death	Predicted injured
Training Data Set	Death in traffic accidents	3,690	226 Type I Error: 5.77%
	Injured in traffic accidents	713 Type II Error: 65.53%	375
Testing Data Set	Death in traffic accidents	939	49 Type I Error: 4.96%
	Injured in traffic accidents	182 Type II Error: 68.93%	82

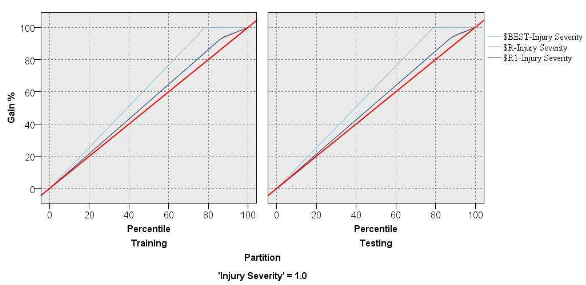


Fig. 3: Evaluation Chart of Two CRT Models

terms of model accuracy, type I and type II errors, and model explanations.

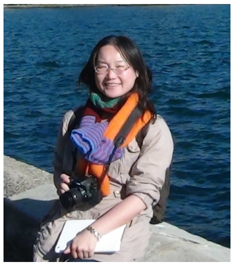
5 Conclusions

When twenty four variables are initially used, six major variables are identified and status of concerned motor vehicles and people has the highest weight among these variables. By applying principal component analysis, only three major variables are chosen and status of concerned motor vehicles and people is the most important variable. To compare the performance of these two CRT models, model accuracy, type I and type II errors, and evaluation chart are used. The results show that CRT model with dimension reduction outperforms the model without dimension reduction. Therefore, using principal component analysis to reduce the number of variables is a better approach when a wide variety of variables might be related to motor vehicle traffic accidents.

References

[1] S.C Huang, E.C Chang, H.H Wu, Expert Systems with Applications **36** (3P2), 5909-5915 (2009).
[2] S. Solomon, H. Nguyen, J. Liebowitz, W. Agresti, Industrial Management & Data Systems **106** (5), 621-643 (2006).
[3] J.T Wei, M.C Lee, H.K Chen, H.H Wu, Expert Systems with Applications, **40** (18), 7513-7518 (2013).
[4] I.H Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Second Edition, Morgan Kaufmann Publishers, San Francisco, CA, 2005.
[5] D.T Larose, Data Mining Methods and Models, Wiley, Canada, 2006.
[6] M.A Razi, K. Athappilly, Expert Systems with Applications **29** (1), 65-74 (2005).
[7] H. Li, J. Sun, J. Wu, Expert Systems with Applications **37** (8), 5895-5904 (2010).
[8] A. Rovlias, S. Kotsou, Journal of Neurotrauma **21** (7), 886-893 (2004).
[9] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Second Edition, Morgan Kaufmann Publishers, San Francisco, CA, 2007.
[10] F.P Garcia Marquez, I.P Garcia-Pardo, Quality and Reliability Engineering International **26** (6), 523-527 (2010).
[11] K.Y Kou, J.T Wei, H.H Wu, Y. H Liu, Applied Mechanics and Materials **145**, 349-353 (2012).
[12] J.T Wei, K.Y Kou, J.S Lin, H.H Wu, Journal of Information & Optimization Sciences **32** (6), 1341-1352 (2011).
[13] W.J Frawley, G. Piatetsky-Shapiro, C.J Matheus, AI Magazine **13** (3), 57-70 (1992).
[14] M.S Chen, J. Han, P.S Yu, IEEE Transactions on Knowledge and Data Engineering **8** (6), 866-883 (1996).
[15] C. Kleissner, Proceedings of the Thirty-First Hawaii International Conference on System Sciences **7**, 295-304 (1998).
[16] U. Fayyad, Intelligent Enterprise **6** (8), 23-33 (2003).
[17] S.Y Wu, Y.L Chen, IEEE Transactions on Knowledge and Data Engineering **19** (6), 742-758 (2007).
[18] S.Y Lin, J.T Wei, C.C Weng, H.H Wu, International Proceedings of Economic Development and Research - Innovation, Management and Service **14**, 131-135 (2011).
[19] IBM Corporation, IBM SPSS Modeler 15 Modeling Nodes, IBM Corporation, 2012.
[20] T. Hill, P. Lewicki, STATISTICS Methods and Applications, StatSoft, Tulsa, OK, 2007.
[21] D. Delen, Expert Systems **26** (1), 100-112 (2009).
[22] StatSoft, Electronic Statistics Textbook, StatSoft, Tulsa, OK, 2010.
[23] D.T. Larose, Discovering Knowledge in Data: An Introduction to Data Mining, Wiley & Sons, New York, 2005.
[24] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, Third Edition, Morgan Kaufmann Publishers, San Francisco, CA, 2012.

[25] J. Bi, K.P Bennett, M. Embrechts, C.M Breneman, M. Song, Journal of Machine Learning Research **3**, 1229-1243 (2003).



Yu-Huei Liu received PhD degree in Biochemistry and Molecular Biology at National Taiwan University. Her primary research interests are in the areas of Biochemistry and Molecular Biology in Cancer, Genetics, Autoimmune diseases and Complementary medicine. Her secondary interests are in the areas of applied statistics including the statistical methods and models for complex systems and numerical methods for kinetic equations.

Kuang-Yang Kou received the PhD degree in Industrial Engineering of University of Texas at Arlington, USA. Now he works in Department of Traffic Science of Central Police University at Taoyuan, Taiwan. His research interests are in the areas of mathematical programming and industrial management. He has published research articles in reputed international journals of mathematical and engineering sciences.

Hsin-Hung Wu received his Ph.D. degree from Department of Industrial & Systems Engineering and Engineering Management at University of Alabama in Huntsville. He is a distinguished professor at National Changhua University of Education as well as an IEDRC Fellow. His major interests include service quality, decision analysis, and data mining.

Ya-Chi Nian received her MBA degree from National Changhua University of Education in Taiwan. Currently, she is a teacher in the Department of International Trade at National Hsinchu Commercial and Vocational High School in Hsinchu, Taiwan.