Effectiveness study of Crash Avoidance technologies by using Clustering and Self Organizing Map

Hitoshi Uno
Yusuke Kageyama
Akira Yamaguchi
Tomosaburo Okabe
Nissan Motor Co. Ltd.
Japan
Paper Number 13-0482

ABSTRACT

Implementation of appropriate safety measures, either from the viewpoint of a vehicle, individual or the infra-structure, it is an important issue to clearly understand the multi-dimension complicated real world accident scenarios. This study proposes a new method to easily capture and to extract the essence of such complicated multi-dimension mutual relationship by visualizing the results of accidents clustering by SOM (Self Organizing Map).

The FARS data from 2010 is used to generate a dataset comprised of 16,180 fatal passenger car drivers and 48 variables. The 16,180 fatal drivers were clustered using hierarchy cluster analysis method and mapped into a two-dimensional square with one dot representing one fatal driver using SOM. From the SOM assessment of the 16,180 fatal drivers, five clusters were created, and they are characterized as follows: Cluster 1 (Interstate highway accidents), Cluster 2 (Drunk speeding), Cluster 3 (Non speeding lane departure), Cluster 4 (Vehicle to vehicle) and Cluster 5 (Intersection).

Three accident scenarios are created to study potential areas of fatal accidents reduction in the SOM map, and the accident scenarios are: [A] Skidding Straight, [B] Lane Departure N.H. (National Highway) and [C] Rear-end. The three accident scenarios mutually had coverage of totally 31% of all the fatal drivers, and the three accident scenarios had high coverage of Cluster 1 (Interstate highway accidents) and some coverage over Clusters 2, 3 & 4. ESC (Electronic Stability Control), LDW (Lane Departure Warning) and FCW (Forward Collision Warning) may be relevant to help reduce the number of fatal accidents in these three accident scenarios.

The remaining areas that the three accident scenarios [A], [B] and [C] did not completely cover were the following accidents:
(1) Young drunk speeding at curves
(2) Speeding on low speed limit roads
(3) Speeding with previous speeding convictions
(4) Drunk driving that are not speeding
(5) Distraction
(6) Elderly
(7) Intersection

1. INTRODUCTION

Today there are many methods and many published values that explain the effectiveness of Crash Avoidance technologies, where the effectiveness is estimated as a value from accident simulations, field data or combination of accident data.
But through these current methods it is difficult to understand a global view to reduce accidents and fatalities in a strategic way and make the priority decision of implementing safety features or social measures.
This study focuses to visually understand the effectiveness of Crash Avoidance technologies in a global view, possible to perceive coverage of the technologies and overlap of the accident factors, which enables intuitive insight in priority decision of measures and remaining areas to be developed and implemented.
The present study focuses on a generalized method, by utilizing SOM to visualize the multi-dimension accident scenario. A Self Organizing map (SOM) is a type of artificial neural network that is trained using unsupervised learning to map mutual relationship into a two-dimensional representation. This can prevent the analyst's arbitrary perspective.
Further, SOM is useful for low-dimensional visualization of high-dimensional input data, by using a neighborhood function to preserve the topological properties of the input space. This can preserve all characteristics of a large dataset (48 variables x 16,180 cases) used in this study and one can visually percept all 48 variables and their mutual relationships across the clusters at a glance.

2. METHODOLOGY

Data Set

This study uses the FARS 2010 data base maintained by NHTSA and UMTRI. Each traffic accident in FARS includes at least one fatality that occurred on a traffic way. Data key of FARS 2010 Occupant (FARS10OC) consists of driving scenarios including road environment, vehicle / driver relating information and occupant characteristics. As the objective of this study is to understand accident scenarios of general cars, the data set conditions are filtered as fatal passenger car and light truck drivers shown in Table 1. From the total 69,124 fatal accident occupants in the FARC10OC data base, 16,180 passenger car fatal drivers can be extracted.

Table 1: Data set used in this study

2010 FARS Occupant cases (**FARS10OC**)	69,124
AUX: VEHICLE BODY TYPE = Passenger Car (1) + Light truck (2-5)	
OCCUPANT TYPE = Driver (1)	
OCCUPANT INJURY SEVERITY = fatal (4)	
Selected cases	16,180

Selection of variables

The first step is to narrow down the whole set of 566 variables to a subset of fewer meaningful variables. Without any loss of information, the fewer variables it will help to perceive the multi-dimension accident scenario more accurately. The 566 variables are narrowed down in the following rule by excluding, those semantically lower order variables, the variables having low frequency and similar variables.

(1) Exclude low-semantic variables in describing the accident. e.g. CASE NUMBER, COUNTY ID, VIN, ACCIDENT DATE-YEAR, etc.

(2) Exclude variables which attribute values having low-frequency occurrence of under 5%. e.g. CRASH RELATED FACTOR has 99.9% of value "0: none"

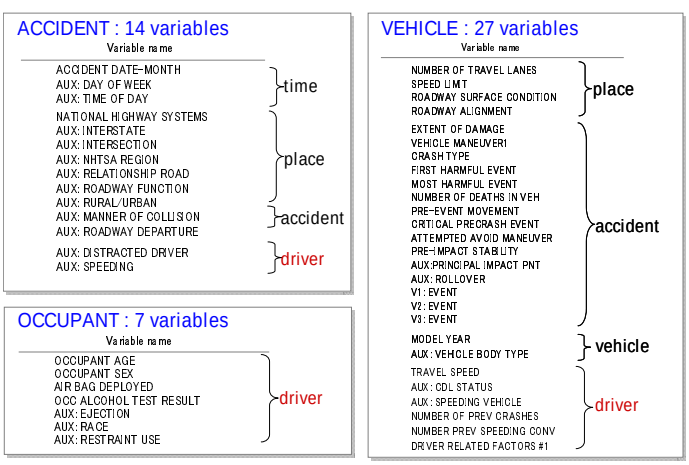
(3) Variables that have similar values with high coherence are grouped, to avoid over represented contribution. e.g. BODY TYPE, VIN BODY TYPE, VIN VEHICLE TYPE, etc. They are grouped by using only one representative variable.

Parameter selection

Based on the above rule, excluding similar variables and unimportant variables, finally a set of 48 variables were selected as shown in Table 2.

Table 2: The selected FARS variables

Total number of FARS10OC variables	566
(1) Low-semantic variables	346
(2) Low-frequency variables	59
(3) Similar variables	113
Selected variables	48



3. SELF ORGANIZING MAP

A Self Organizing map (SOM) is a type of artificial neural network that is trained using unsupervised learning to map mutual relationship into a two-dimensional representation. This can prevent the analyst's arbitrary perspective.

Further, SOM is useful for low-dimensional visualization of high-dimensional input data, by using a neighborhood function to preserve the topological properties of the input space. This can preserve all characteristics of a large dataset (48 variables x 16,180 cases) used in this study and one can visually percept all 48 variables and their mutual relationships across the clusters at a glance.

Vector Quantization

Vector quantization is a classical quantization technique [1], [2]. Thus, this paper will only briefly explain the statistical outline to understand the results.

Vector quantization is to find the discrete approximate of the input vector x of the vector space Rn, by using infinite codebook vector $mi \in Rn$, $i=1,2,3,,,k$. The approximate of x, is to find the most nearest codevector mc to x with Euclidean distance as in equation (1). If the most appropriate value mi is chosen, the square quantization error will be the minimum, as in equation (2).

$$\| x - m_c \| = \min_i \{ \| x - m_i \| \} \qquad (1)$$

$$E = \int \| x - m_c \|^2 \; p(x) dx \qquad (2)$$

Each, codevector is related with a nearest codevector, placed in a nearest neighbor region called a Voronoi region [3]. Voronoi is a space where the codebook vectors are fitted in the two-dimensional space side by side, like a mosaic, and the space is separated into a multiple domain, and each area is divided by a hyper-plane. Each separated area has a vector which has the most nearest vector to the surrounding vectors in the area.

Application

To calculate VQ's for the FARS accident data, this would mean to calculate one VQ for each of all the 16,180 cases. The VQ is calculated by using the normalized 48 variables having a value in between 0 to 1. Similar accidents will have similar VQ values, meaning that the variable values of the 48 variables have a similar distribution.
More details of the method for applying FARS data to SOM can be available in the author's previous study "Method development of multi-dimensional accident analysis using Self Organizing Map" [4].

Mapping

To map the items into a two-dimensional space, a square map of the size of all samples is prepared. From the calculated VQ values using equation (1) and (2) for each item, the items are plotted in the map by plotting each item with the closest VQ value. The mapping steps are shown below.

Step1. Calculate VQ value for each of all of the cases
Step2. Randomly plot one case in the middle of the map
Step3. Randomly select another case and plot beside the first case

Step4. Randomly select another case and plot beside the case with the nearest VQ value
Step5. Repeat Step4, repeat until there are no cases left.

Creating of SOM

As each VQ and accident number have a one-to-one correspondence, for each variable, the accident number is replaced by the variable value with color gradation, from the smallest value as blue to the largest value as red, as shown in Figure 1.

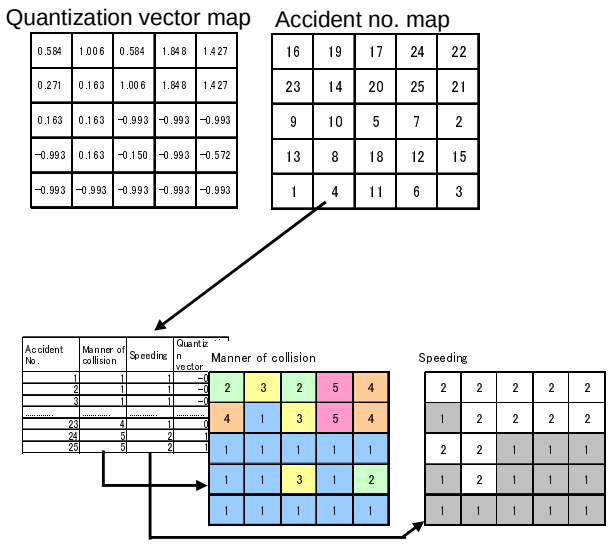


Figure 1: Creating SOM maps for each variable

For each variable, one SOM map was created, resulting in 48 SOM maps.
If there is an unknown value, the value is imputed by calculating the average value of the surrounding cells.

Clustering

Using the hierarchy clustering results, the SOM maps can be divided into clusters, as shown in Figure 2. .

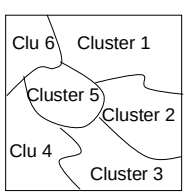


Figure 2: Clustering applied to SOM

The number of clusters can be chosen depending on the level of hierarchy, chosen by the analyst.

3. RESULTS

Clustering

Using hierarchy clustering analysis, FARS 16,180 drivers are clustered into 5 clusters, shown in Figure 3 and Table 3.

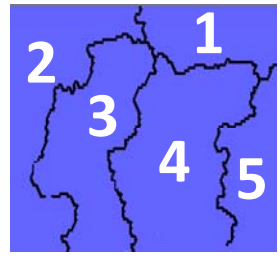


Figure 3: Clustering / SOM results of FARS2010

Table 3: Proportion of each cluster

	Drivers	(%)
Cluster 1	1,931	11.9%
Cluster 2	3,773	23.3%
Cluster 3	3,618	22.4%
Cluster 4	4,773	29.5%
Cluster 5	2,085	12.9%
Total	16,180	100.0%

Self Organizing Map

Each SOM is a representation of 16,180 fatal drivers, with one dot as one fatal driver, and each dot has a color representing the variables value. For example, Figure 4 shows the results of the variable "AUX: MANNER OF COLLISION", and has the following five attribute values.

1. Single (Not collision with motor vehicle): dark blue
2. Rear-end: bright blue
3. Head-on: bright green
4. Angle: bright yellow
5. Other (Sideswipe, Other): orange

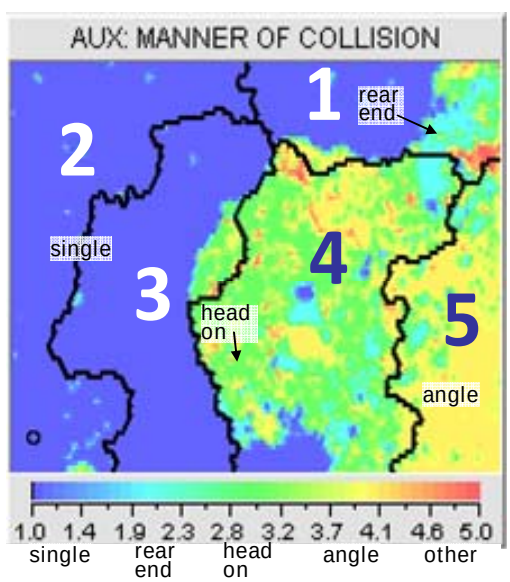


Figure 4: Clustering / SOM results of FARS variable AUX: MANNER OF COLLISION

For example, Cluster 1 has a mix of dark blue “1.Single” accidents and some bright blue “2.Rear-end” accidents. Cluster 2 and 3 are mostly covered with dark blue “1.Single” accident. Cluster 4 has a mix of all types, and Cluster 5 is mainly covered with bright yellow “4.Angle” accidents.
The SOM has a key on the bottom, showing the value and coordinating color. The values for each variable are shown in the FARS codebook in reference [5].
Regarding unknown attribute values, the SOM algorithm imputes a value by calculating the average value of the cells surrounding the unknown value and generates a color which is in between the surrounding cells. For example, for the above AUX: MANNER OF COLLISION, Cluster 2 is almost fully covered in dark blue “1.Single” accidents, thus, if there is an unknown value in this cluster, the value would be most likely imputed as a dark blue “1.Single” accident.

Viewing all 48 SOM maps

For each of the 48 variables, 48 SOM maps are created as shown in Figure 5. In each SOM map, the position of each dot representing a fatal driver is consistent and unique. Thus each SOM map corresponding to each individual variable can be compared. In each of the SOM maps, the cluster boundaries are indicated to clearly identify the characteristics of clusters by viewing several variable SOM maps at a glance.

4. CHARACTERISITICS OF CLUSTERS

This chapter explains the characteristics of each cluster by viewing the major contributing variables. A representative name is given to each of those clusters as shown in Table 4.

Table 4: Representative name of clusters

Cluster 1	Interstate highway accidents	12%
Cluster 2	Drunk speeding	23%
Cluster 3	Non speeding lane departure	22%
Cluster 4	Vehicle to vehicle	30%
Cluster 5	Intersection	13%

Cluster1. Interstate highway accidents

Figure 6 shows the key variables and corresponding SOM maps for Cluster 1. AUX: INTERSTATE, shows that accidents in this cluster occur on interstate highways. Also from AUX: MANNER OF COLLISION and V1: EVENT it can be seen that 62% are single accidents, 13% head-on, 12% rear-end. AUX: SPEEDING VEHICLE shows that half of them are due to Speeding. In this way, the characteristics of each cluster can be understood by viewing the SOM maps and can be quantified as shown in Table 5.

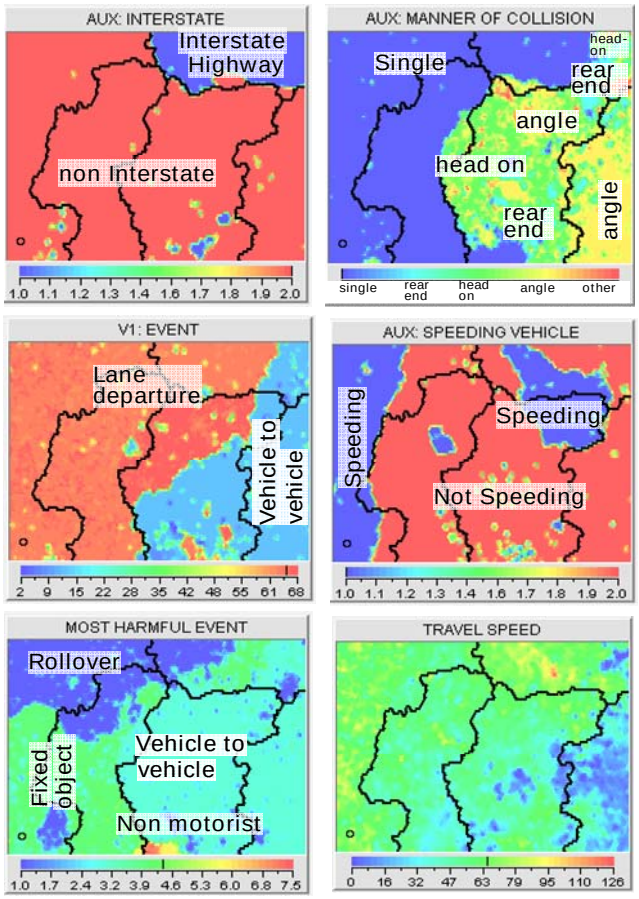


Figure 6: Key variables and corresponding SOM maps for Cluster 1

Table 5: Quantification of variables for each cluster

Variable name	Value	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
V18: NATIONAL HIGHWAY SYSTEMS	This section is on the NHS	100%	11%	14%	28%	25%
V77: AUX: INTERSTATE	Interstate	91%	1%	0%	2%	0%
V78: AUX: INTERSECTION	Intersection	1%	4%	7%	12%	87%
V92: AUX: RURAL/URBAN	Rural	45%	78%	65%	66%	39%
V393: ROADWAY ALIGNMENT	Curve	26%	52%	36%	27%	4%
V26: AUX: MANNER OF COLLISION	Single	62%	98%	91%	6%	0%
	rear end	12%	1%	2%	9%	4%
	head on	13%	0%	5%	47%	15%
	angle	10%	0%	2%	35%	81%
V461: V1 EVENT	Lane departure	62%	95%	93%	38%	0%
V93: AUX: SPEEDING	Speeding	44%	69%	6%	25%	11%
V623: OCC ALCOHOL TEST RESULT	0.08-0.16	13%	24%	11%	7%	4%
	0.16 +	10%	30%	16%	6%	2%
V667: AUX: RESTRAINT USE	unbelted	42%	77%	60%	36%	27%

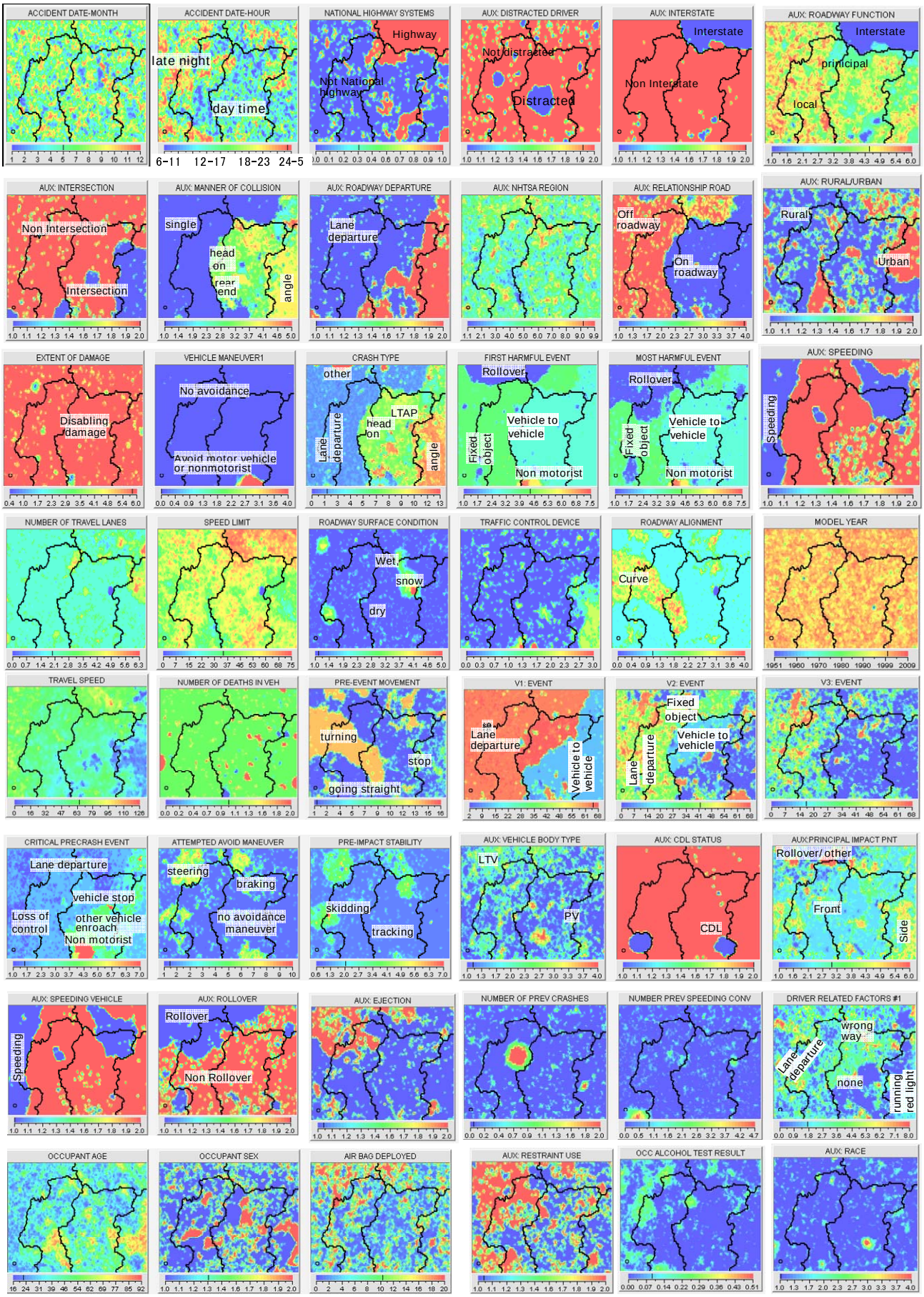


Figure 5: 48 Self Organization Maps

5. MAPPING THREE ACCIDENT SCENARIOS

Three accident scenarios are created to study potential areas of fatal accidents reduction in the SOM map. The three accident scenarios are;

[A] Skidding Straight
[B] Lane Departure N.H. (National Highway)
[C] Rear-end

By mapping these accident scenarios into the SOM map, it is possible to simulate each accident scenario as a virtual CA (Crash Avoidance) technology that can prevent the accidents in that accident scenario. By this simulation, it is possible to understand the coverage areas of these virtual CA technologies, and also clarify uncovered accident scenarios, to study a global approach for reducing traffic accident fatalities.
The definition of the accident scenarios are set by a single or combination of the FARS variables, and are described below.

[A] Skidding Straight
This accident scenario assumes a virtual CA technology to theoretically avoid skidding accidents on straight roads. The accident scenario is defined as PRE-IMPACT STABILITY = "Skidding" and ROADWAY ALIGNMENT = "Straight".
With current available CA technologies, ESC (Electronic Stability Control) may be effective to help reduce the number of fatal accidents in this accident scenario.

Figure 7 shows the areas of PRE-IMPACT STABILITY = "Skidding" and ROADWAY ALIGNMENT = "Straight". The left SOM of Figure 8 shows the SOM of AUX: MANNER OF COLLISION and the black dots represents the mutual assembly of [PRE-IMPACT STABILITY = "Skidding"] AND [ROADWAY ALIGNMENT = "Straight"]. The right SOM of Figure 8 encircles the high density black dots area, to define an easily recognizable area of [A] Skidding Straight.

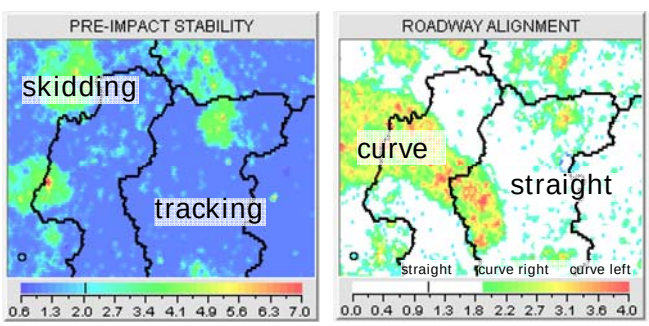


Figure 7: Stability and Road Alignment SOM maps

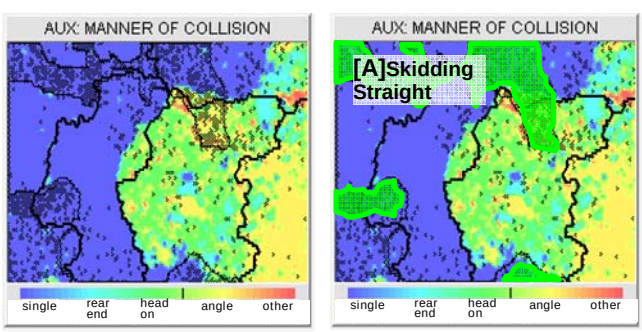


Figure 8: [A] Skidding Straight coverage areas

This [A] Skidding Straight covers 14% of all areas, 27% of Cluster 1(Interstate highway accidents), 25% of Cluster 2(Drunk speeding) and 11% of Cluster 3(Non speeding lane departure).

[B] Lane Departure N.H. (National Highway)
This accident scenario assumes a virtual CA technology to theoretically avoid un-intentional lane departure accidents on national highways. The accident scenario is defined as NATIONAL HIGHWAY SYSTEMS = "National highway", V1: EVENT = "Lane departure (Ran off road, cross median)" and VEHICLE MANUEVER = "No avoidance".
With current CA technologies, LDW (Lane Departure Warning) may be effective to help reduce the number of fatal accidents in this accident scenario.

Figure 9 shows the relevant SOM maps and Figure 10 shows the SOM of AUX: MANNER OF COLLISION with the mutual assembly of NATIONAL HIGHWAY SYSTEMS = "National highway", V1:EVENT = "Lane departure" AND VEHICLE MANUEVER = "No avoidance" with black spots and the [B] Lane Departure N.H. area encircled for simplification.

Figure 9: V1: Event, National Highway Systems and Vehicle Maneuver SOM maps

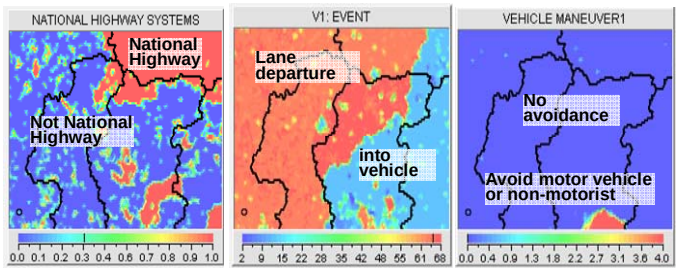
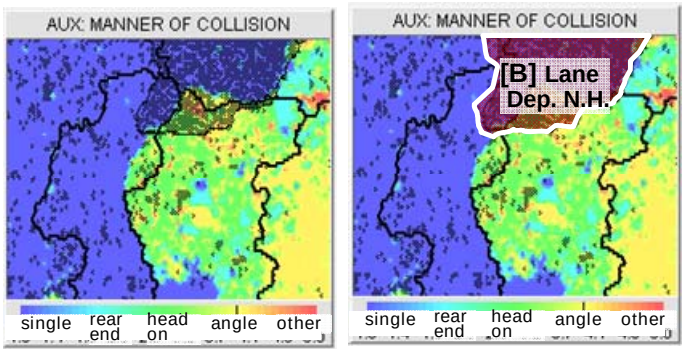


Figure 10: [B] Lane Departure N.H. coverage areas



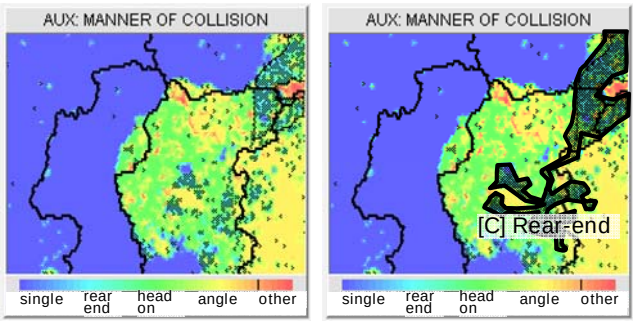
This [B] Lane Departure N.H. covers 16% of all areas, 62% of Cluster 1(Interstate highway accidents), 14% of Cluster 2(Drunk speeding) and 11% of Cluster 3 (Non speeding lane departure) and Cluster 4 (Vehicle to vehicle).

[C] Rear-end

This accident scenario assumes a virtual CA technology to theoretically avoid rear-end accidents, and is defined as AUX: MANNER OF COLLISION = “Rear end”. With current CA technologies, FCW (Forward Collision Warning) may be effective to help reduce the number of fatal accidents in this accident scenario.

Figure 11 shows the SOM of AUX: MANNER OF COLLISION with “Rear end” with black spots and the [C] Rear-end Prevention area encircled for simplification.

Figure 11: [C] Rear-end coverage areas



This [C] Rear-end accident scenario covers 5% of all areas, 12% of Cluster 1(Interstate highway accidents) and 9% of Cluster 4 (Vehicle to vehicle).

Mutual coverage areas of [A], [B] and [C]

Figure 12 and Table 6 show the mutual coverage of [A] Skidding Straight, [B] Lane Departure N.H. and [C] Rear-end. The three accident scenarios have high coverage of Cluster 1 (Interstate highway accidents) and some coverage over Clusters 2, 3 & 4.

Figure12: Coverage areas of [A], [B] and [C]

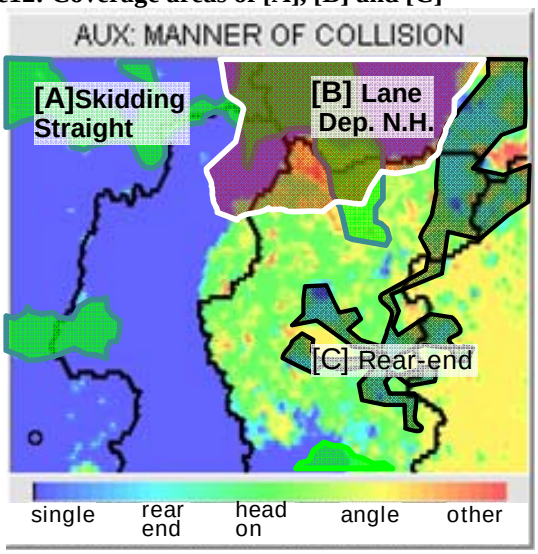


Table 6: Potential coverage of the three accident scenarios

Cluster		[A] Skid	[B] Lane	[C] Rear	Mutual [A][B][C]
1	Interstate highway accidents	27%	62%	12%	**77%**
2	Drunk speeding	25%	11%	1%	**32%**
3	Non speeding lane departure	11%	14%	2%	**23%**
4	Vehicle to vehicle	9%	11%	9%	**27%**
5	Intersection	2%	0%	4%	**5%**
	All areas	14%	16%	5%	**31%**

Characteristics of other areas

Focusing on other areas in the SOM map, the characteristics of the uncovered areas of the three accident scenarios can be understood. Examples of the perception of the remaining areas are shown in Figure 13. The characteristics of the remaining areas can be derived by viewing the relevant major SOM maps in Figure 14.

The summary of the remaining areas are described below, and the implementation of appropriate safety measures, could be either from the viewpoint of a vehicle, individual or the society.

(1) Young drunk speeding at curves
(2) Speeding on low speed limit roads
(3) Speeding with previous speeding convictions
(4) Drunk driving that are not speeding
(5) Distraction
(6) Elderly
(7) Intersection

Figure13: Characteristics of remaining areas

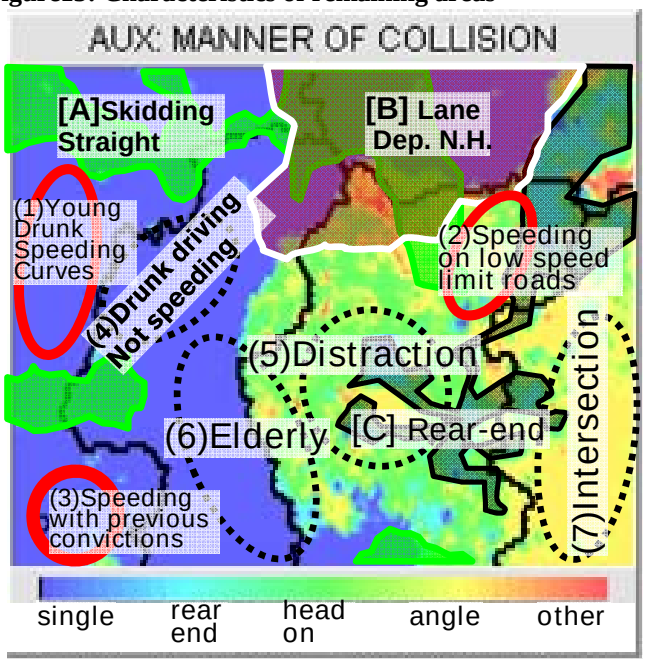
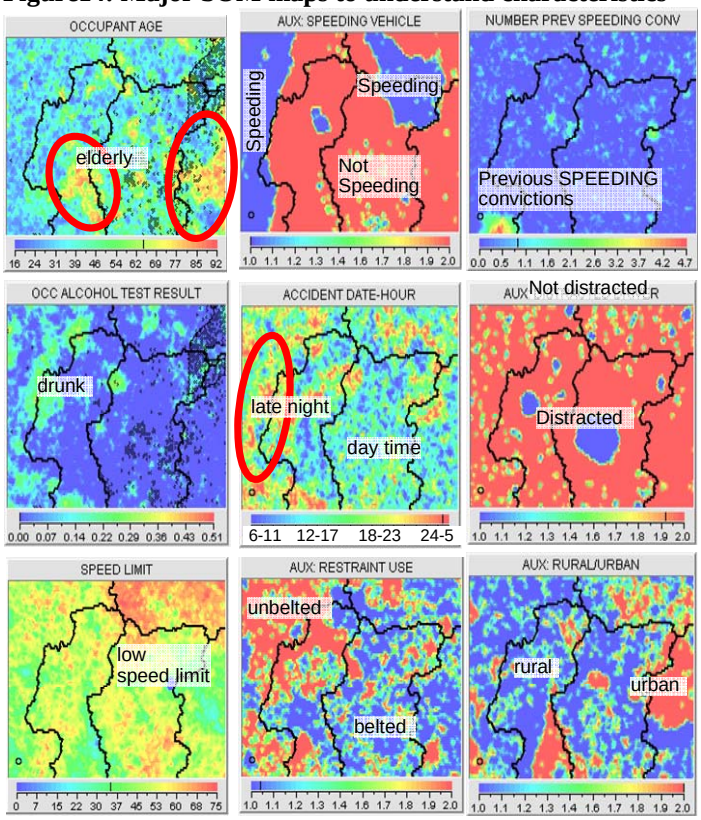


Figure14: Major SOM maps to understand characteristics



4. CONCLUSION

This present paper shows that clustering & SOM analysis can be used to classify FARS accidents into a number of clusters using all necessary factors, without setting any specific evaluation criteria.
Five clusters can be successfully generated from 16,180 fatal drivers from FARS2010 database. They are Cluster 1 (Interstate highway accidents), Cluster 2 (Drunk speeding), Cluster 3 (Non speeding lane departure), Cluster 4 (Vehicle to vehicle) and Cluster 5 (Intersection).

Three accident scenarios are created to study potential areas of fatal accidents reduction in the SOM map, and the accident scenarios are: [A] Skidding Straight, [B] Lane Departure N.H. (National Highway) and [C] Rear-end. The three accident scenarios mutually had coverage of totally 31% of all the fatal drivers, and the three accident scenarios had high coverage of Cluster 1 (Interstate highway accidents) and some coverage over Clusters 2, 3 & 4.
By focusing on the areas that the three accident scenarios did not cover, the characteristics of the uncovered accident scenarios were clarified.

When focusing on the effective areas of CA technologies and other areas which overlap across the adjacent clusters, all important factors such as driver age, drunk driving, seatbelt usage, etc. can be visualized. Thus, by analyzing the characteristics of the clusters using SOM to consider new counter-measures, it may be possible to explore new strategic solutions to reduce US fatalities.

REFERENCES

[1] Teuvo.Kohonen, “Self Organizing Maps”, Helsinki University of Technology Neural Networks Research Centre, 2005
[2] A.Gersho, IEEE Trans Inform Theory IT-25, 1979
[3] G.Voronoi, J.reine angew. Math. 134, 1908
[4] Hitoshi Uno, “Method development of multi-dimensional accident analysis using Self Organizing Map”, 2013
[5] UMTRI, “Data Set Codebook FARS 2010”, Version 03-Nov-11