

## MOTORWAY TRAFFIC CRASH PREDICTION BASED ON THE NEURONAL NETWORK APPROACH: APPLICATION TO THE RING WAY OF PARIS

**BOUHELAL Mediouny**

EPAM  
Rue du Maréchal Tito  
4029 Sousse - Tunisie  
NVAP-Faculté de Médecine  
5000 Monastir-Tunisie  
**bouhelal\_monsef@yahoo.fr**

**ZEGLAOUI Anis**

ENISO  
Avenue 18 Janvier  
4000 Sousse - Tunisie  
NVAP-Faculté de Médecine  
5000 Monastir - Tunisie  
**zeglaoui2001@yahoo.fr**

**HAJ SALEM Habib**

INRETS/GRETIA  
2, Rue de la Butte Verte,  
Le DESCARTES 293166-France  
**haj-salem@inrets.fr**

**ABSTRACT:** The paper focuses on the development of a Risk index model for traffic crash prediction, based on the application of a mixed approach: artificial neural networks, statistical analysis approaches. The inputs of the developed risk index model are the traffic measurements (volume and occupancy rate) and the calculated temporal left gradient. A global database including accidents and traffic measurements are used to validate the risk index model approach. The obtained results are promising while in some traffic conditions, the estimated risk index model is able to detect crash occurrence about 6 to 7 minutes prior to the crash time. This Risk index could be used as off-line safety evaluation index (evaluation process, off-line simulation) or real-time safety index monitoring for user information.

**KEY-WORDS:** *Safety, Risk analysis, ANN, traffic state clustering, fundamental diagram, linear and non linear regression.*

### 1 INTRODUCTION

Safety can be defined in a number of ways, including the official World Health Organisation (WHO) safety definition 'freedom from unacceptable risk of harm'. When we speak about traffic safety we usually think about accidents. Accidents can be defined as (KELLER, 2002): "Any event that due to moving traffic at opened roads and places resulted in fatalities, injuries or/and damages". Safe road traffic is characterised, in an ideal case, by the absence of crashes, injuries and fatalities.

Accidents have a great effect on the safety of responders and on the mobility of the travelling public. Over 40,000 people are killed and 1, 7 million people injured on roads in the EU every year (ETSC, 2001). In the USA studies reported 41,000 people killed and more than 5 million injured. It is estimated that 18% of fatalities on motorways is due to secondary accidents only. Representing human life in numbers is not a favourable practice, however, the following accident cost estimates could be found in literature. In the USA, in addition to the delay costs, there is close to \$200 billion per year of direct economic loss due to accidents and fatalities (FHWA, 2004). In the EU the cost to society has been estimated at 160 billion Euros annually.

In order to increase the traffic safety, control measures are introduced to improve traffic performance in motorway traffic including speed limit control, ramp metering, user information aiming at homogenizing the practical speed along the motorway sections and at minimizing the number and the severity of accidents and consequently increasing safety (Dilmore J., 2005). On the other hand, introduction of electronics and computerization systems in the vehicle technologies have significantly contributed to safety and comfort. However, the prediction of the crash in real time is still in investigation phase and some research efforts are dedicated in this area. During the last five years, there is an increasing focus on the development of real time ("potential crash") prediction algorithm on urban motorway traffic (Abdel-Aty M. & al., 2005; Lee C. & al., 2006; Haj-Salem, & al., 2007).

In the field of safety analysis, the classical traffic evaluation approaches consist in collecting incident/accidents traffic data during the experimented scenarios (traffic control strategies, modification of the infrastructure etc.), and in proceeding to traffic impact and statistical safety analysis of the number of accidents before and after the implementation of these scenarios. Generally, the collection of the accident numbers must get a statistical significance before undertaking an evaluation process. This remark imposes a long time of field data collection (3-5 years), which is the "price to pay" for having a statistical

significance of accidents set and correct safety evaluation.

This paper aims at developing a risk index based on real-data measurements, which can be used either off-line as an evaluation index during the evaluation process leading to the dramatically reduction of the field test periods, or in real-time like: a safety monitoring tool (e.g. safety user warning system), a multi-criterion function to be optimized in real time (safety index combined with a traffic index) within several control strategies such as coordinated ramp metering, speed limit control, route guidance, etc.

The developed index is based on the collection of measured traffic data synchronized with incident/accidents data on two sites in France: the urban motorway A4/A86 and the ring way of Paris. In this paper, the used data concerns the Ring way of Paris only (accidents and traffic data measurements).

## 2 AVAILABLE TRAFFIC DATA DESCRIPTION

### 2.1 Ring way of Paris description

The Corridor Périphérique (CP) consists of two parallel beltways around the city of Paris (see figure 1), each having a length of some 35 km in each direction, and of the connecting radial streets. The outer motorway belt is the Ring way including a total of 70 on-ramps and 70 off-ramps in both directions. Some of these ramps are the beginning or ending points of corresponding motorways that start from or lead to the city of Paris. The inner, signal-controlled arterial belt is the Boulevard des Maréchaux (BM).

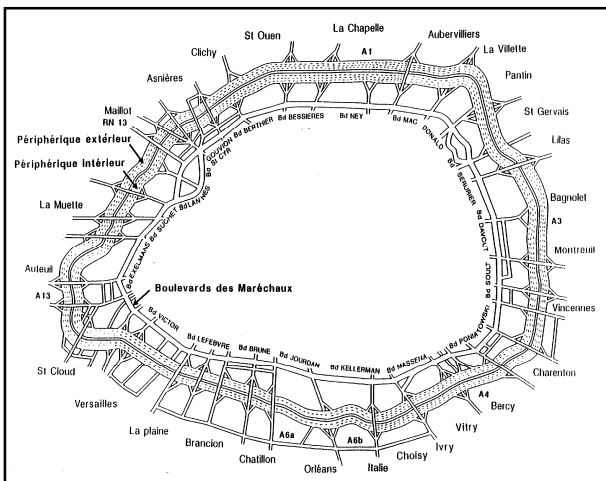


Figure 1: Ring way of Paris.

The Corridor Périphérique is a central highway facility of the extended traffic network of Ile-de-France. It carries a wide variety of traffic types, including daily commuters, holiday traffic and commercial vehicles, and offers connections from Paris to the suburbs and vice

versa, between pairs of Paris locations, and between suburbs. Moreover, the CP is used by a non negligible number of through drivers when changing motorways on their far-distance trip.

CP is managed by the Paris town-hall. Two control centers are dedicated for rapid lanes traffic management and urban traffic management. The first control centre is located in the centre of Paris (LUTECE) and its main function is the urban traffic management (intersections). The second control centre is located at Porte d'Ivry and only manages the traffic on the rapid lanes (Boulevard Périphérique). It is important to note that the Boulevards des Maréchaux management is ensured by the control centre LUTECE and that a data exchange in terms of traffic states is going to be implemented in order to coordinate the control strategies between these two central control rooms.

The Ring way represents 40% of the Parisian traffic for a network surface equal to 2.5% of the overall motorway and urban network of Paris. Today the Boulevard Périphérique is the only complete ring considered as a motorway in Ile-de-France. Linked up to 6 motorways (A1, A3; A6a, A6B, A13, A86), it supports an important national and international traffic; it is the main access to the motorways from Paris and the near suburban area.

The number of vehicles served per day is equal to 1.100.000 vehicles for an average travel distance of 7 kilometers. The tracks represent about 10% of the number of vehicles.

### 2.2 Data base building

Traffic dataset and accident characteristics are collected from historical database stored in the HYPER operating system of the Ring way. The considered sites are fully equipped with real traffic measuring sensors located at around every 500 meters apart. The incidents/accidents data characteristics are automatically and manually collected and include: time of day, location of the accident, weather conditions and severity.

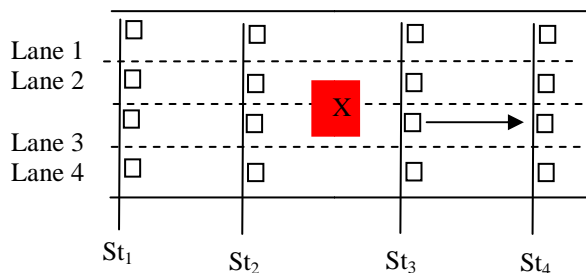


Figure 2: Topology of the considered stretch measurements for each crash

The collected traffic data covers 2 hours (one before and one after the crash) at two upstream and two downstream measurement stations (figure 2), consisting of traffic volume, occupancy rate and speed (if exists). The time

intervals of the traffic measurements are equal to one minute.

The final constituted database includes the overall accidents occurred and traffic data during 4 years (2001-2004) and (2002-2004). The total number of accidents collected is around 900 on the ring way of Paris.

In order to exclude the effect of several factors, the first investigation step is made on a selected number of accidents with the following criterion: same topology (4 lanes), sunny weather conditions and full luminosities (no night-time accidents considered). Among all collected accidents, the available accidents where dramatically reduced, leading to 90 accidents selected on the Ring way of Paris.

### 3 RISK MODEL DEVELOPMENT

The aim is to develop a crash risk function model [Abdel Aty, M., A. Pande, 2005] which is able to decide on line and at the seen of two traffic variable measurements of volume and occupancy rate, whether the traffic state is crash prone or not. For a given traffic situation, more its potential of crash occurrence increases more it's deemed to be as crash prone. The proposed crash risk function model is based on the application of a supervised learning artificial neural networks (ANN) approach. To achieve this research investigation, a labeled data base is required. Unfortunately no human expert is able to decide whether a traffic situation is crash prone or not with a high confidence in view of the partial information at disposal. To overcome this deficiency, some mathematical techniques are used, among them clustering (K-means) and relationship between traffic volume and occupancy rate (Greenberg fundamental diagram) are the most relevant. This investigation is implemented through three steps:

1. The measured traffic data is splitted into two categories: fluid and congested based on the fundamental diagram relationship. Two sub data sets are generated. A temporal left gradient is incorporated in each data base example to enrich the information carried by the input.
2. For each sub data set, K-Means data clustering algorithm is applied. Several experiments using different clusters are achieved to find out the best cluster number. It will be shown later that two clusters is the best data partition.
3. A majority vote was proceeded to select the class which will be labeled as crash prone. The class containing the majority of traffic patterns associated to one minute prior to the crash minute time is the crash prone class and subsequently all its elements are labeled as crash prone.

### 3.1 Basic theoretical risk index model development

In order to alleviate this paper, only the main approach and summary of mathematical results are described in this section.

In this research, let  $X$  refers to a random vector describing the traffic state and taking its values within  $\mathbb{R}^d$ . For each realization  $x$  of  $X$ , let  $R(x)$  denotes the crash occurrence potential of the traffic state  $x$ . A traffic condition situation  $x$  is deemed to be crash prone if the risk value  $R(x)$  exceeds a given level  $\alpha$ .

The following classifier holds:

$$Y(x) = \begin{cases} 1 & \text{if } x \text{ is crash prone} \\ 0 & \text{Otherwise} \end{cases} \\ = \begin{cases} 1 & \text{if } R(x) > \alpha \\ 0 & \text{Otherwise} \end{cases} \quad (1)$$

It's known that the optimal classifier approaching  $Y$  is the Bayes function, say  $g^*$  [Devroye, L. and *al.*, 1996], [Hastie, T. and *al.*, 2001], [Vapnik, V., 2000]. We showed  $g^* = Y$  almost surely. In other words the Bayes loss is zero, that is  $P(g^*(X) \neq Y) = 0$ .

Of course only partial information is available for traffic flow observer. We proved that under this hypothesis, the corresponding Bayes function  $\tilde{g}^*$  is the optimal approximation of  $g^*$  in the  $L^2$  sense. Based on the result established by [Cybenko, G., 1989], [Funahashi, K., 1989], [Hornik, K., and *al.*, 1989] concerning the density of functions implemented by neural networks within the set of classifiers, we showed that multilayer perceptrons approach the conditional probability associated to the Bayes function  $\tilde{g}^*$  and subsequently the risk function  $R(x)$ .

### 4 DATA LABELLING AND CATEGORIZATION

As mentioned above, no human expert can replace an oracle in terms of labeling every traffic pattern as being crash prone or not. To imitate an oracle, each pattern is labeled through the following steps:

- Building accident data series and splitting them into fluid and congested sets.
- Reducing the input dimensions.
- Incorporating the temporal left-gradient in each input.
- Applying K-means clustering on both fluid and congested accident data series.
- A majority vote decides which cluster is crash prone or not in each set (congested and fluid).

#### 4.1 New data base building

In the following the building up of an accident data series is detailed. For each available accident in the data set, measurements of traffic variables (occupancy rate and traffic flow) are extracted for each minute during

one hour prior to the crash minute. These measurements are taken for only two stations: one upstream and one downstream of the crash location. According to the selected accidents which include only four lanes, 16 dimensional traffic patterns are built up (4 lanes \* 2 stations \* 2 variables) at each minute for each considered station. The sequence of these patterns makes up the accident data series.

Further, the reduction of input dimension avoiding the loss of information improves the learning of the artificial neural networks. A data analysis results demonstrated that measurements related to lanes one and two are highly positively correlated and similarly for lanes three and four. Averaging the measurements of two lanes is preferred to select one of them. Then two dummy lanes are obtained and named as lane I and lane II. The previous accident data series is henceforth consisting of 8 dimensional patterns. The data base thus obtained consists of 90 data series each of them contains 60 patterns. Hence the overall examples are 5400.

#### 4.2 Fluid and congested accident data series

Recall that the aim of this section is labeling data after using the clustering statistical function (K-means) which is based on grouping sample examples belonging to an Euclidian state space according to a given distance function (criterion). Generally this distance function is taken as the Euclidian distance. However, crash prone patterns could be very disparate when they correspond to fluid or congested traffic situations, which complicates the task of such clustering methods leading to unsatisfactory results. To palliate this shortcoming, accident data series are categorized into fluid or congested series using the Greenberg fundamental diagram relationship [Greenberg, H., 1959], [Lighthill, M.J., G.B. Whitham, 1955].

For every accident data series, for each underlying station and for each dummy lane, pairs of  $(k, q)$  where  $k$  is the occupancy rate and  $q$  is the traffic volume are extracted. The Greenberg relationship is fitted to the data set of  $(k, q)$  yielding an estimated fundamental diagram. Figures (3, 4, 5, 6) indicate the nonlinear parametric regression related to the accident number 7014. In each figure, squares and asterisks represent respectively accident data and the estimated Greenberg relationship. The diamond and the circle report  $(k, q)$  pairs corresponding respectively to one minute and two minutes prior to the crash time. The  $(k, q)$  pair's corresponding to one minute and two minutes prior the crash time are named as data 59 and data 58. The accident data series stands for fluid, if both data 58 and data 59 related to both dummy lanes I and II at the two stations are in the fluid part of the fundamental diagram. Otherwise, this data series is set to be congested. In figures (3, 4, 5, 6) the data of the accident number 7014 is reported. Using the

previous definition one can conclude that the associated data series is congested.

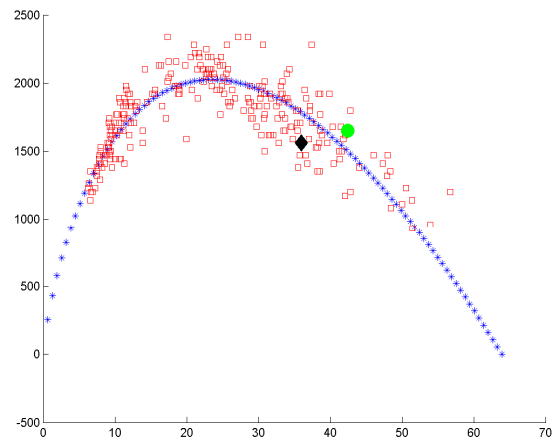


Figure 3: Lane I at Station 2

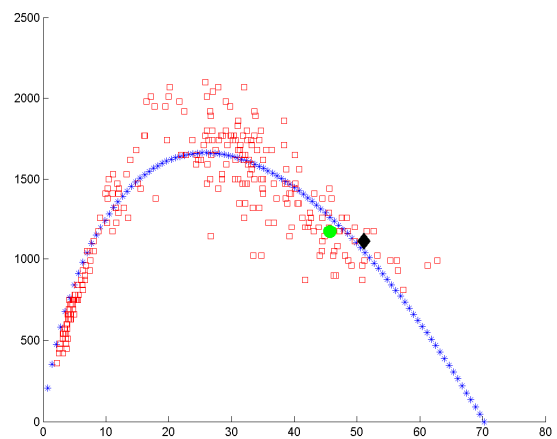


Figure 4: Lane II at Station 2

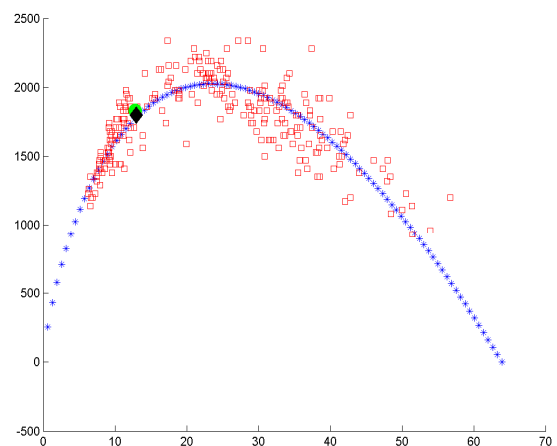


Figure 5: Lane I at Station 3

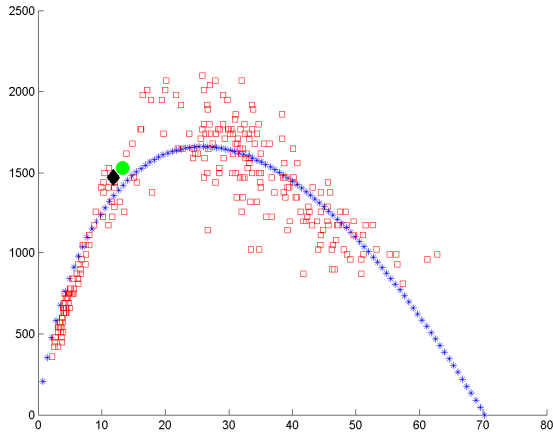


Figure 6: Lane II at Station 3

### 4.3 K-means based data labeling

The accident data series are splitted into fluid and congested series. Each traffic pattern belonging to a series is an eight dimensional input. At this stage, a temporal left gradient is added to every sample example to take into account the time varying of the measurements and hence a 16 dimensional input is obtained. The gradient is computed using the values of  $k$  and  $q$  at minutes  $t$  and  $(t - 1)$  for each dummy lane at each station. Recall that each input contains measurements associated to one upstream station and one downstream station with respect to the location of the crash for each lane. This structure can be considered as a spatial gradient. The injected temporal left gradient enriches the information carried by each input. At the end of this stage, two categories of data are created. The first one contains the whole 16 dimensional examples deriving from fluid accident data series and referred to as fluid category. In the same way, the congested category is created from the congested accident data series.

On each above category, a K-means clustering is applied. Experiments were led using different numbers of clusters  $K$  going from two to five. The best configuration is selected upon the mean and the standard deviation of silhouette patterns. The silhouette displays a measure of how close each point in one cluster is to points in the neighboring clusters. This measure ranges from -1 to +1. The +1 value indicates points that are very distant from neighboring clusters. The 0 value indicates points for which no decision can be taken. The -1 value indicates points that are probably assigned to the wrong cluster. More the mean silhouette clustering is large and its standard deviation is small more the clustering is better. The table 1 reports the summarized results concerning the silhouette mean and its standard deviation of each category (fluid and congested) and for each cluster number. Screening the table 1, one can observe that among the

four cluster characteristics the best one is the number cluster  $K = 2$ . Figures (7, 8) indicate the silhouette plots for both categories (fluid and congested) with  $K = 2$ .

Clusters	2	3	4	5
Mean (fluid)	0.371	0.321	0.368	0.325
Std (fluid)	0.205	0.188	0.189	0.184
Mean (congested)	0.465	0.325	0.335	0.342
Std (congested)	0.213	0.197	0.197	0.191

Table 1: Mean and standard deviation of silhouettes

Recall that the aim of K-means is to be able to label each example as crash prone or not. The main idea of this labeling is: if there are precursors to a crash they would be detected at least through the traffic measurements of one minute prior to the minute of accident occurrence. Therefore the cluster containing the majority of patterns related to 59 minute will be considered as crash prone. All examples belonging to that cluster will be labeled so. The remaining patterns are non crash prone. The next step is merging all crash prone examples deriving from fluid and congested categories in a unique crash prone class. Similarly, non-crash prone class is built up. Approximately two thirds of the overall examples are in the non crash prone class.

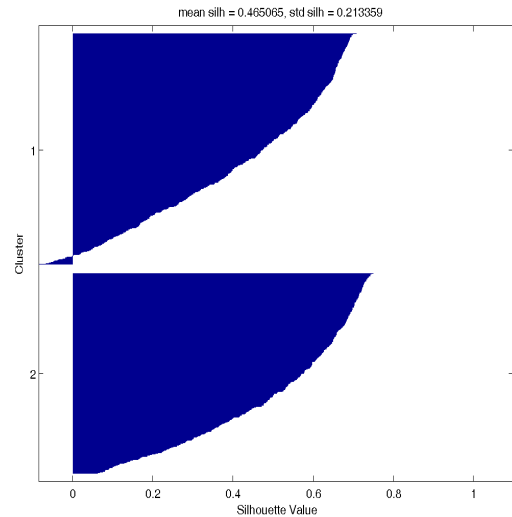


Figure 7: Congested case. Silhouette by cluster with K=2

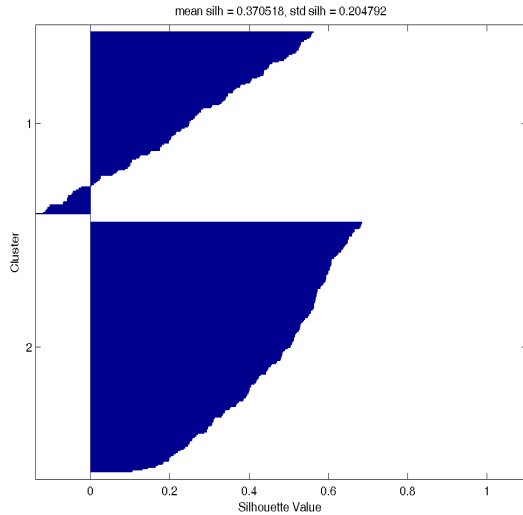


Figure 8: Fluid case. Silhouette by cluster with  $K=2$

## 5 RISK FUNCTION APPROXIMATION USING ARTIFICIAL NEURAL NETWORKS

The used ANN are multilayer perceptrons (MLP) with one hidden layer containing  $k$  neurons having the same sigmoidal transfer function and a linear output layer containing one neuron.

Among the 90 available data series (5400 examples), one third (containing fluid and congested series) is reserved for the needs of the efficiency prediction tests. The training and the validation phases use the whole examples of the remaining data series (3600 examples). The target of each pattern is either +1 or -1 according to whether it is crash prone or not.

### 5.1 ANN architecture investigation

The best architecture (the best number of hidden units  $k$ ) is determined experimentally according to the available training and validation sets. Several configurations were tested with  $k$  going from 5 to 50. To evaluate the efficiency of each MLP, the correlation coefficient  $R$  is computed. This coefficient corresponds to the linear regression of the MLP outputs on their associated targets. Experiments show that the best architectures seem to be those with hidden unit number between 20 and 30. Figures (9, 10, 11) represent this linear regression related to  $k = 5$ ,  $k = 21$  and  $k = 50$ .

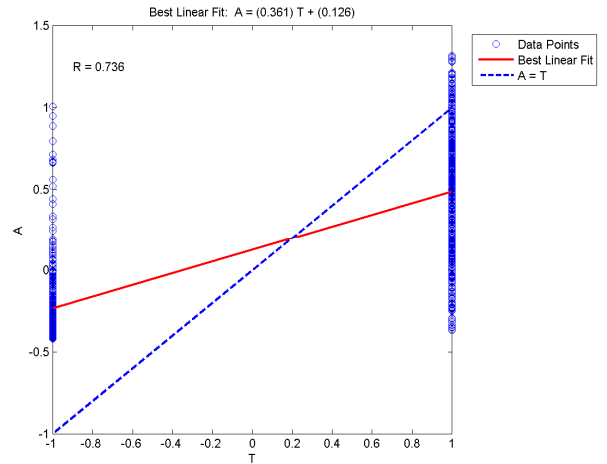


Figure 9: MLP with 5 hidden nodes.  $R = 0.736$

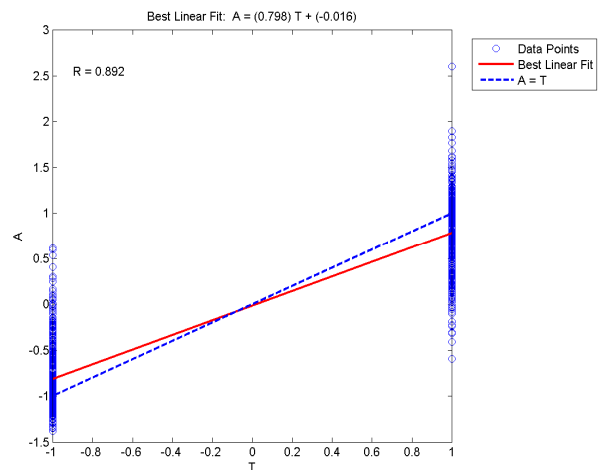


Figure 10: MLP with 21 hidden nodes.  $R = 0.892$

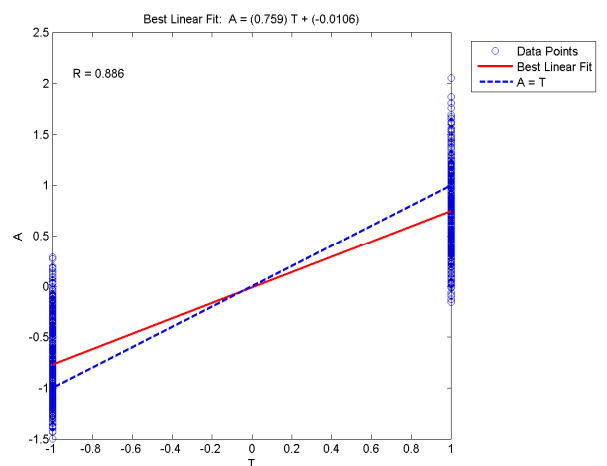


Figure 11: MLP with 50 hidden nodes.  $R = 0.886$

## 5.2 Behavior of the risk index

This subsection is devoted to test the ability of the risk function implemented by the selected MLP to evaluate the crash proneness. Also, the behavior in time of this estimated risk function is studied during a period of ten minutes prior to the crash time. These experiments are achieved using the data series of the test set made up previously.

Due to the previous architecture investigations, an MLP with  $k = 21$  hidden nodes is chosen. Let  $x$  denotes the input of the MLP and  $\varphi(x)$  its output. A traffic situation described by its input  $x$  (measurements and temporal left gradient) is considered to be crash prone if  $\varphi(x) > 0$  and non-crash prone if  $\varphi(x) \leq 0$ . This choice is based on both target labels  $\{-1,1\}$  and the theoretical risk index model development. On the other side, accident data series of each category (fluid and congested) are handled separately. For each data series belonging to the fluid category, the average of MLP outputs for the same minute is evaluated. This process is repeated for the ten minutes prior to the crash time. In the same way an average of MLP outputs is computed for the same period and for the data series belonging to the congested category. Figures (12, 13) plot the obtained results. The first figure corresponds to the congested category while the second one to the fluid case.

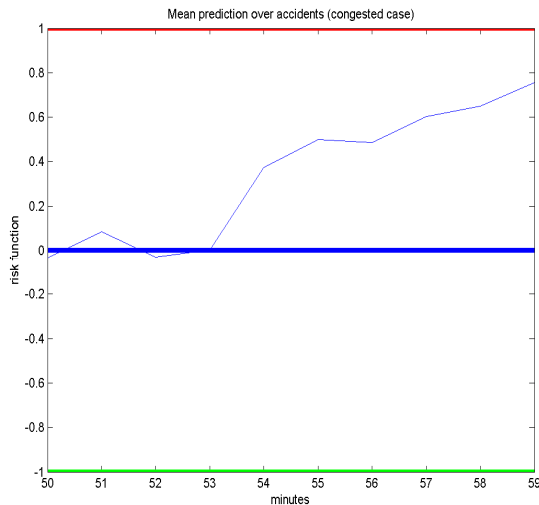


Figure 12: Accident prediction (Congested case)

In the congested case, the estimated mean risk function is positive for about 6 to 7 minutes prior to the crash occurrence. Also, one could see that is an increasing function of time attaining its maximum just before the accident moment. However in the fluid case, the MLP decides that the traffic situations are non-crash prone during all the considered period except the last minute for which the output of the MLP is slightly positive.

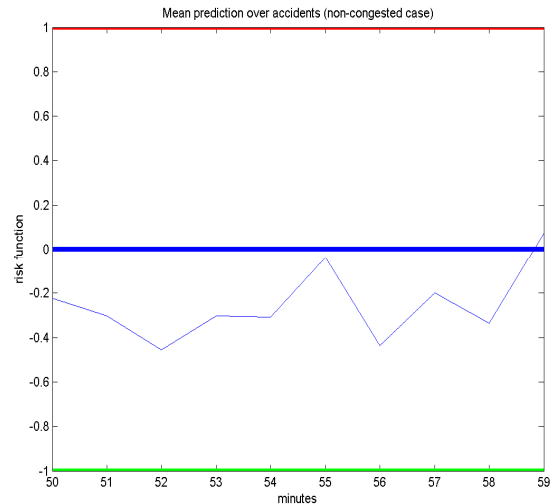


Figure 13: Accident prediction (Fluid case)

## 6 CONCLUSION AND NEXT STEPS

It appears from the above results that in the congested case, the estimated mean risk function developed in this paper is able to predict accident of each traffic situation using the measurements of the traffic volume and occupancy rate. The information carried by these two variables at two successive stations and the computation of the temporal left gradient at each of them contain enough information about the real traffic state. However in the fluid case, the information carried by the input measurements is not sufficiently pertinent. The estimated risk index model predicts the accident occurrence only one minute prior to the crash time. This seems to be acceptable because in the fluid case, traffic flow and occupancy rate measurements are probably not sufficient to explain the crash occurrence. Also the one minute interval time of measurements seems to be too large and the traffic measurements are much smoothed leading to the elimination of the traffic variability. However, additional information such as driver behavior, weather conditions and speed measurements could be useful to improve the Risk index model and the crash prediction mainly in the fluid case of traffic.

## ACKNOWLEDGEMENT

Bouhelal M. and Zéglouai A. gratefully acknowledge Professor M. Masmoudi from University Paul Sabatier of Toulouse (France) for helpful discussions. His critical insights were especially valuable.

## REFERENCES

Abdel-Aty, M., A. Pande, 2005. Identifying crash propensity using specific traffic speed conditions. *Safety Research*, 36:97-108.

- Cybenco, G., 1989. Approximations by superpositions of sigmoidal functions. *Math. Control, Signals, Systems*, 2:303-314.
- Devroye, L., L. Györfi, G. Lugosi, 1996, *A probabilistic theory of pattern recognition*, Springer-Verlag New York.
- Dilmore J. 2005; "Implementation strategies for Real-Time traffic safety Improvement on Urban Freeways". MA Thesis, University of Central Florida, Orlando.
- ETSC: European transportation safety council 2001; Transport accident and incident investigation in the European Union. Brussels.
- FHWA: Federal Highway Administration 2004; [http://ops.fhwa.dot.gov/congestion\\_report/](http://ops.fhwa.dot.gov/congestion_report/)
- Funahashi, K., 1989. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2:183-192.
- Greenberg, H., 1959. An analysis of traffic flow. *Operations Research*, 7 No.1.:79-85.
- Haj Salem, H., Lebacque J.P., 2007. "Risk index modeling for real-time motorway traffic crash prediction" *Traffic and Granular Flow conference*, 55-64.
- Hastie, T., R. Tibshirani, J. Friedman, 2001, *The elements of statistical learning*, Springer-Verlag New York.
- Hornik, K., M. Stinchcombe, H. White, 1989. Multi-layer feedforward networks are universal approximators. *Neural Networks*, 2:359-366.
- Lee C., Hellinga B., Saccomanno F., 2006; "Evaluation of variable speed limits to improve traffic safety". *Transportation Research Part C* 14:213-228.
- Lighthill, M.J., G.B. Whitham, 1955. "On Kinematic waves, II. A theory of traffic flow on long crowded roads. *Proceeding of Roy. Soc. (London)*. 229A, 317-345.
- Keller, H. 2002; *Materialien – Verkehrsleitsysteme im Straßenverkehr – Methodik und Geräte zu Erfassung des Verkehrsablaufs*. Publication at Fachgebiets Verkehrstechnik und Verkehrsplanung, Technical University Munich.
- Vapnik, V., 2000, *The Nature of statistical learning theory*, Springer-Verlag New York.