World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:9, No:1, 2015

Using Data Mining in Automotive Safety

Carine Cridelich, Pablo Juesas Cano, Emmanuel Ramasso, Noureddine Zerhouni, Bernd Weiler

Abstract—Safety is one of the most important considerations when buying a new car. While active safety aims at avoiding accidents, passive safety systems such as airbags and seat belts protect the occupant in case of an accident. In addition to legal regulations, organizations like Euro NCAP provide consumers with an independent assessment of the safety performance of cars and drive the development of safety systems in automobile industry. Those ratings are mainly based on injury assessment reference values derived from physical parameters measured in dummies during a car crash test.
The components and sub-systems of a safety system are designed to achieve the required restraint performance. Sled tests and other types of tests are then carried out by car makers and their suppliers to confirm the protection level of the safety system. A Knowledge Discovery in Databases (KDD) process is proposed in order to minimize the number of tests. The KDD process is based on the data emerging from sled tests according to Euro NCAP specifications. About 30 parameters of the passive safety systems from different data sources (crash data, dummy protocol) are first analysed together with experts opinions. A procedure is proposed to manage missing data and validated on real data sets. Finally, a procedure is developed to estimate a set of rough initial parameters of the passive system before testing aiming at reducing the number of tests.

Keywords—KDD process, passive safety systems, sled test, dummy injury assessment reference values, frontal impact.

I. Introduction

A. Context

WORLDWIDE, road accidents kill 1,2 million persons per year and hurt forty times more [1]. For many years, governments campaign for safe roads and their actions have already proved their efficiency, without being able to influence the inequalities in different countries. Indeed, more than 90% of the deaths on the road take place in low-income or lower middle income countries, which have less developed active and passive safety system compared to Europe or North America. Considering the increasing number of vehicles on the road, safety systems had to be developed during the last years to reduce these accidents. In 2003, 6.613 people died in Germany on the road and, mainly thanks to the development of active and passive safety, this number has been reduced in 2011 by almost 42% (3.648 deaths) [2]. The car buyers understood also the importance of the safety, that is why, when buying a new car, its safety is one of the most important considerations, as shown in a survey from 2013 (Fig. 1). For this reason, the safety systems have to be continually further improved in order to respond to the increasing market demands. Sled tests and other types of tests are carried out by car makers and their suppliers to confirm the protection level of the safety system.

C. Cridelich and B. Weiler are with TRW Automotive GmbH, Alfdorf, Germany (e-mail: carine.cridelich@trw.com).
P. Juesas Cano, E. Ramasso and N. Zerhouni are with FEMTO-ST Institute, UMR CNRS 6174 - UFC / ENSMM / UTBM, 25000, Besançon, France

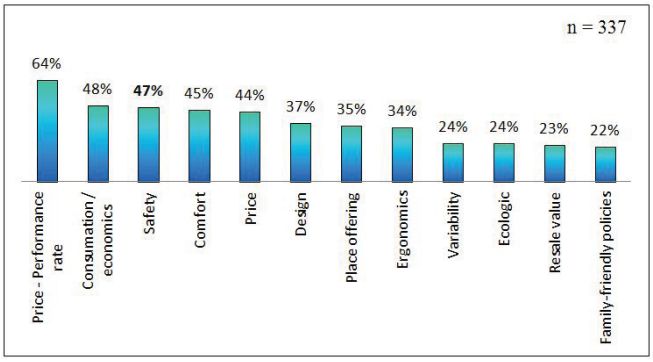

Fig. 1: Reasons for buying a car [3].

Sled tests give a hand in the development of the passive safety system. They are based on the use of specific dummies which simulate a real human body behaviour. The injuries imposed on the human body are estimated through high-tech sensors within the dummies which measure the biomechanical criteria such as the head acceleration and neck movements. A study of German data (Fig. 2) reveals that about 50% of the car collisions are frontal impacts (data collected by GIDAS between July 1999 and June 2013) [4].

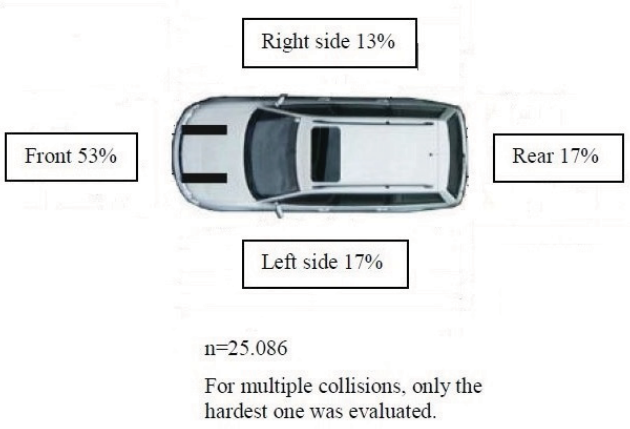

Fig. 2: Car collision's repartition

These frontal impact crashes are simulated and executed by customers and suppliers with the help of real crashes or servo hydraulic HyG-Sled-facilities in order to understand the behaviour of the safety systems during an impact and to reduce the occupants' injuries severity by developing new safety technologies.

B. Objectives

This study is based on thousands of data sets emerging from frontal sled tests. The data sets are mainly biomechanical criteria imposed to the dummies during the impact such as the head acceleration and rib deflection. In this publication, the characteristics of the sled test (airbags, seat belt and dummy position) are defined as input parameters and have been studied to find correlations and patterns between the parameters (inputs) and the data (outputs). Knowledge Discovery in Databases (KDD) is used in addition to numeric simulation to reduce the number of sled tests by learning from historical data. This publication follows the steps of a KDD process, based on the data emerging from sled tests according to Euro NCAP specifications.

II. Type of Data

A. The Injuries through Dummy's Instrumentation

In this study, only the Hybrid III 50th percentile frontal impact dummy used for the Euro NCAP rating was taken into consideration. It corresponds with a Median Adult Male, with 175 cm in height and 78 kg in weight. This dummy is a calibrated device which simulates the physical properties of a human. The high tech sensors measure the injury potential during a sled test such as the velocity of impact, deceleration rates of various body parts, and impact forces. A modern dummy has over 200 sensors. These sensors are able to acquire lots of data and are regrouped into three types of instrumentation: the accelerometers, load sensors and motion sensors. These sensors are used to calculate the criteria for the Euro NCAP rating. One of the most sensitive areas is the dummy neck, that is why a six-Axis Lower Neck Load Cell (Fx, Fy, Fz, Mx, My, Mz) is embedded into the dummy's neck.

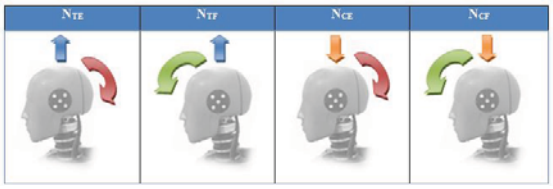

Fig. 3: The four movements of the dummy's neck

B. Euro NCAP Rating

In the last thirty years, lots of new legislations and programmes have been established in order to improve car safety, but also to support the car buyer's choice with the help of the released passive safety results. European New Car Assessment Programme (NCAP) is a European programme which regroups methods of tests, automobile designs and others tests relating to the automobile. Data from sled tests according to Euro NCAP specifications have been considered on the passenger side. These tests represent the collision between two cars (Fig. 4) where the passenger is a Hybrid III 50th dummy percentile. The impact velocity is 64km/h against a deformable barrier with an overlap of 40% [5].

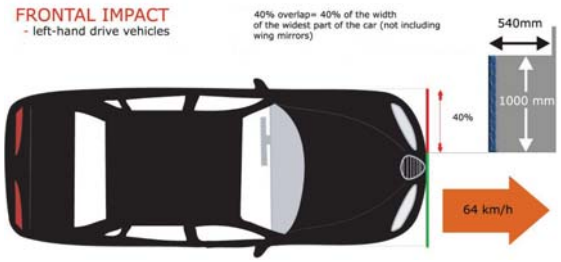

Fig. 4: Frontal impact Euro-NCAP [6]

III. Knowledge Discovery in Databases

The KDD process and data mining methods are being increasingly developed in automobile crash studies in order to find out useful and understandable patterns in data. In order to extract knowledge from data in large databases, the steps of the KDD process are the following: Data selection, pre-processing, transformation, application of a data mining method and evaluation of the model.

A. Data Selection

The passive safety systems were studied in order to determine the parameters of each system which could have an influence on the biomechanical criteria of the dummy. Three components of the safety system have been studied: the passenger airbag, knee bag and seat belt. The dummy position and quality of the pulse have been also analysed. In the analysis, the potential variation of the environment like the buck, instrumental panel or seat, was not taken into consideration. Thanks to an expert elicitation, about 30 key factors have been identified as parameters influencing the occupant's injuries during a crash:

- Passenger airbag: divided panel, volume, vent holes, tethers, inflactor
- Knee bag: volume, tethers, inflactor
- Seat belt: pretensioners (retractor, buckle and anchor), angle shoulder/D-ring, webbing on spool, height adjustment, dynamic locking tongue
- General: dummy temperature, dummy neck constellation, pulse evaluation, car classification

About 50 sled tests were selected and their data have been collected with the aim of creating the largest possible database.

B. Data Pre-Processing

The collected data cover the period 2000-2014. The data had to be prepared and completed from different sources (such as the dummy instrumentation or dummy position). But some data were missing and had to be replaced by an estimated value.

This step is one of the most important challenges of the data pre-processing and is known in literature as "missing data". It means that some values are no more available or unknown. Tseng et al. [9] were confronted to this problem and they developed a new method based on cluster properties. Batista and Monard [8] also studied this problem by using the 10-NNI

method, a k-nearest neighbour imputation algorithm. No one of these methods could be directly applied to the data of this study. Therefore, a new algorithm had to be developed using the k-NN imputation method among other methods.

In the present study, the missing data have been gathered in two cases:

- Case 1: the parameter's value is not given but this parameter shall have to be defined, e.g. the dummy position or temperature. This case will be studied here.
- Case 2: the parameter's value is not given because the restraint system does not have some properties, e.g. if the seat belt system does not have a buckle pretensioner, data like time to fire or length of pretension cannot be recorded. These values are not directly considered as missing values.

There are three possibilities to solve the problem of missing data (from case 1):

1) Repeat the sled test: This solution seems to be the best one but it is also expensive and sometimes impossible (old projects, parts no more available).
2) Average calculation: This method replaces missing values through the average of the values of other sled tests. It was used in the algorithm for some types of missing values.
3) Imputation method by using the k-NN: Some algorithms can be used in order to define the missing value according to the values included in the database. It was used in the algorithm for some types of missing values.

The created algorithm takes 3 types of missing data into consideration:

1) The missing data is independent on the sled test, e.g. the dummy temperature. In this case, the average of all dummy temperature values included in the database is calculated.
2) The missing data is dependent on the project, e.g. the dummy position. Indeed, the dummy position is dependent on the trim parts (seat, instrumental panel) and car body. In this case, the algorithm first selects only the sled tests of the same project and then calculates the average of these sled tests.
3) The missing data is dependent on the project and other parameters, e.g. the webbing on spool of the seat belt. In fact, the webbing on spool is one of the parameters which influences the dummy shoulder force. In this case, the algorithm first selects only the sled tests of the same project, then compares their dummy shoulder force (with a tolerance of +/- 500N) and finally calculates the average of these sled tests.

This algorithm only intervenes if a parameter is missing and has to be defined, in contrary with data which are missing because they do not exist. After running this algorithm, the database has no more missing data.

C. Data Transformation

For each parameter determined during the data section, some input data are transformed in order to recognise nominal descriptions (Fig. 5).

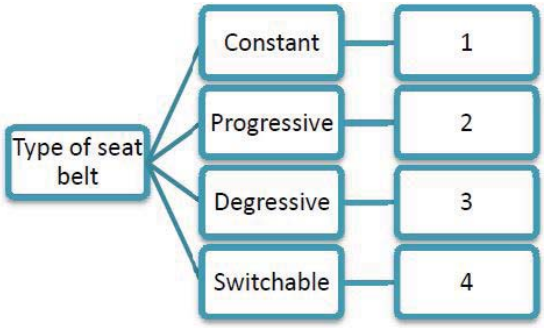

Fig. 5: Example for the numerical transformation of nominal parameters

After this basic transformation, all values were normalised to a range between -1 and 1. Table I shows the real values, Table II their normalized corresponding.

TABLE I: EXAMPLE OF PARAMETER'S VALUES

Parameter	P1	P2	P3	P4	P5	P6	P7	P8
Test 1	13	406	NaN	121	114	0	30	3,5
Test 2	18	NaN	NaN	130	135	0	30	3
Test 3	13	411	300	118	109	15	35	3,5
Test 4	13	411	304	125	118	NaN	35	NaN
New test	18	440	NaN	NaN	133	30	40	2,5

TABLE II: EXAMPLE OF RANGED PARAMETER'S VALUES

Parameter	P1	P2	P3	P4	P5	P6	P7	P8
Test 1	-1	-1	NaN	-0,5	-0,615	-1	-4	1
Test 2	1	NaN	NaN	1	1	-1	-1	0
Test 3	-1	-0,706	-1	-1	-1	0	0	1
Test 4	-1	-0,706	1	0,167	-0,308	NaN	0	NaN
New test	1	1	NaN	NaN	0,846	1	1	-1

D. Application of a Data Mining Method

Wu et al. [10] registered the top 10 algorithms in data mining identified by the IEEE International Conference on Data Mining 2006. The k-NN classification is one of those which has been implemented here.

Parameter weights have been determined because they do not have the same influence on each dummy part. Üçtug et al. [11] applied the Delphi method, i.e. they asked in a survey, expert's opinion in order to determine the weights. Pirdavani et al. [12] also used this technique for finding the weights in a problem of prioritizing accident hotspots on the road without any crash data. The Delphi method has been applied here on the parameters defined in the section "Data selection" in order to quantify their influence on the dummy's biomechanical values.

The calculated weights can be integrated directly in the Euclidean distance used in the k-NN method in order to estimate the similarity between a new test ($\mathbf{x}_{new}$) and the i-th training instance in the database ($\mathbf{x}^i_{train}$) as follows:

$$d(\mathbf{x}_{new}, \mathbf{x}^i_{train}) = \sqrt{\sum_{j=1}^{p} w_j \cdot \left(x_{new}(j) - x^i_{train}(j)\right)^2} \quad (1)$$

where w_j is the parameter weight.

The classification phase then consists in finding the closest training instance by minimizing the aforementioned distance.

Fig. 6: Example of classification

IV. Model Validation: Missing Data

"Leave-one-out cross validation" is a model validation method of machine learning which uses one value as validation and the remaining values as training set. This method has been used here in order to validate the algorithm which should be able to replace a missing value. The model has been validated for the 3 types of missing data described in the subsection "Data transformation", i.e. for the dummy temperature, dummy position (here the left knee, shorter distance ahead) and webbing on spool of the seat belt. The relative error E has been calculated by comparing the exact and approximate value:

$$E = \frac{|\text{exact value} - \text{estimated value}|}{|\text{exact value}|} \quad (2)$$

The considered data set is made of 48 sled tests: 47 values are used as training set and 1 value is the validation value. For each missing value, the algorithm calculated an approximate value and their relative error E. Fig. 7, 8 and 9 present the curves of the exact and approximate values and of the relative error for the different missing parameters.

As shown in Fig. 7, all the dummy's temperature values estimated by the algorithm are similar. In fact, for parameters which are independent on the sled test, the approximate value is determined through the average of all available values. The maximal difference is +1,45°C, that is acceptable.

Fig. 8 shows that the dummy's position values estimated with the help of the algorithm are acceptable. For such parameters, the algorithm took into consideration only the data coming from the same project. The database contains 48 values at the moment, and per project, it is possible that only one sled test was available. That is why some missing values can not be determined (rupture in the curve). The maximal difference between the exact and approximate values is 15 mm. The

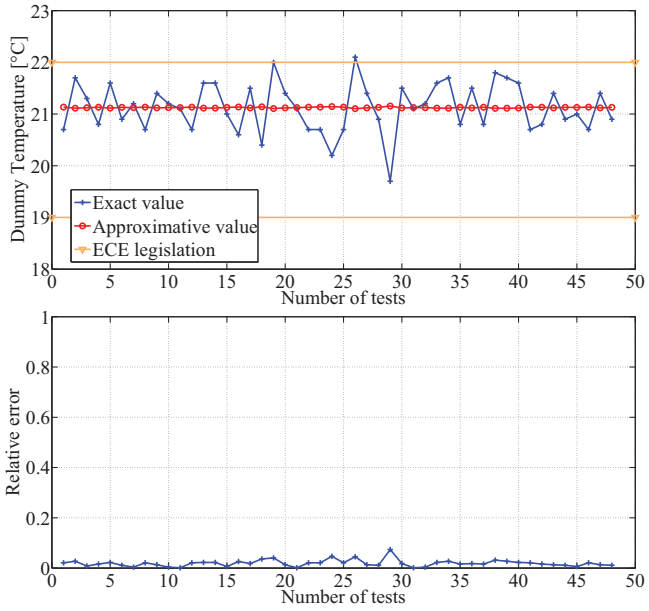

Fig. 7: Estimation's curves for the dummy temperature

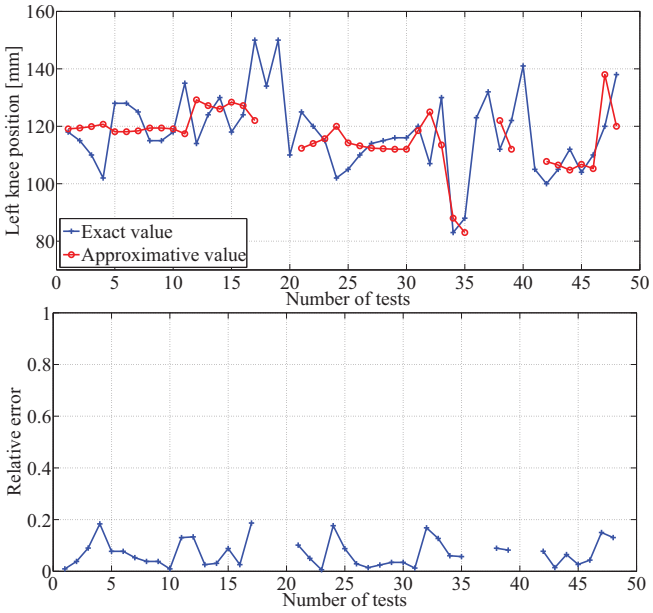

Fig. 8: Estimation's curves for the dummy position

tolerance by dummy position according to the engineers is 10 mm for the shortest distance between the knee and ahead.

As illustrated in Fig. 9, the webbing on spool values estimated with the help of the algorithm are trend similar to the real values. The error curve confirms the validation of the model for missing parameters which are dependent on the project and others parameters. Two approximate values are a little bit far away from the real values. For the first one, only one sled test of the same project was available in the database, that is why its approximate value is the average of other project's values. For the second one, only 3 sled tests of the same project were available. These approximations can be rectified by adding more sled tests. These simulations of

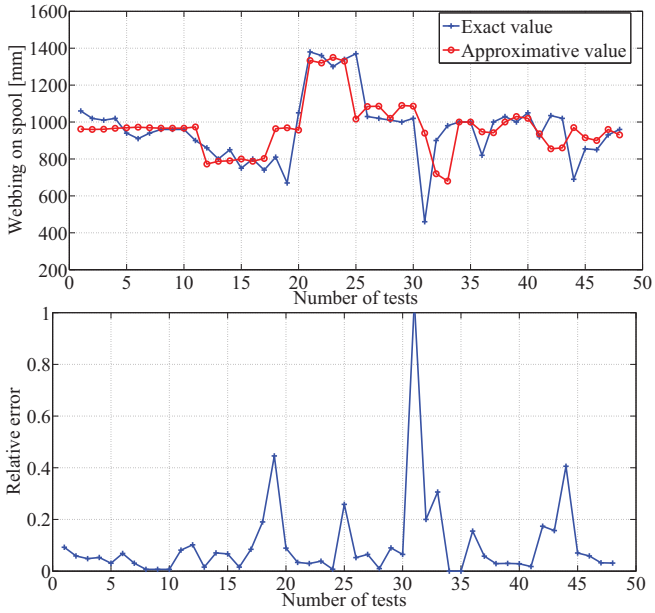

Fig. 9: Estimation's curves for the webbing on spool of the seat belt

missing data confirmed the potential of the algorithm.

In order to collect the information of the sled tests and to analyse it, a graphical user interface (GUI) has been created (Fig. 10). The programme includes sled data of different customers from 2000-2014. The main target of this programme is to classify a frontal sled test, before running, by comparing its configuration (set-up, airbags, dummy specification) with the database (k-NN classification).

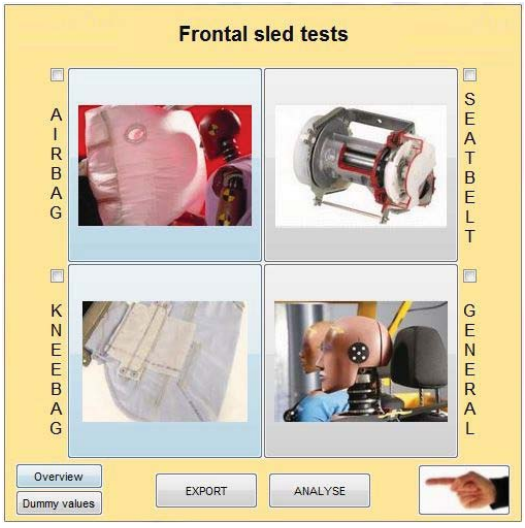

Fig. 10: General panel of the GUI [7]

If desired, the user can adapt the classification by choosing the initial parameters which have to be compared.

V. Conclusion

Safety is one of the growing interest of the customer, that is why automotive suppliers continually develop new passive safety systems which are tested through the execution of sled tests. The present study is focused on the Euro-NCAP rating (passenger side) and aims to classify, before running, a new sled test with the database in order to adapt the parameters of the restraint systems. The biomechanical values of the Hybrid III 50th percentile dummy obtained by the already executed sled tests (database) give a first indication of the new sled test's expected results.

In order to study and classify the sled tests, the KDD process steps have been followed. The first step, the data selection, defined the parameters of each restraint system and the configuration which can influence the dummy biomechanical values. Thanks to an expert elicitation, about 30 parameters have been described as influencing theses values. Other external parameters such as the car pulse and dummy position also have been selected.

The next step, the data pre-processing, was a critical task of the KDD process because the data have to be prepared (data crash, protocol for dummy). The case of "missing data" has also been studied and completed with the help of an algorithm.

The data transformation enabled an homogenisation of all data coming from different sources. This step has been done through a data ranking and basic transformation nominal/numerical value.

In order to classify the sled tests according to the relevant restraint systems, the data mining method, the k-nearest neighbour classification, has been applied: the Euclidean distance has been calculated between two sled tests for each parameter. The parameter weights obtained through expert's elicitation, can be also used in order to quantify the parameter's influence on each body part.

For a better data processing, a graphical user interface (GUI) has been created for entering, saving, analysing and classifying the sled tests. For the future, this method can be used in order to decrease time and costs through the analysis of the historical sled tests before running a new one.

References

[1] *http://www.planetoscope.com/mortalite/1270-mortalite—morts-d-accidentsde-la-route-dans-le-monde.html* 16.09.2014
[2] EuroStat *Persons killed in road accidents by sex (CARE data)* Last update: 21-01-2014
[3] Aral Aktiengesellschaft *Aral Studie Trends beim Autokauf*, Brochure page 12 2013
[4] H. Johannsen *Unfallmechanik und Unfallrekonstruktion* ATZ/MTZ-Fachbuch 2013
[5] Internet Website Euro NCAP, *The official site of the European New Car Assessment Programme*, www.euroncap.com, version October 2014.
[6] Internet Website Wikipedia Euro NCAP, consulted November the 5th, 2014.
[7] Internet Website TRW, consulted November the 30th, 2014.
[8] G. Batista and M.C. Monard, *An Analysis of Four Missing Data Treatment Methods for Supervised Learning* 2003
[9] S. Tseng, K. Howang and C.Lee *A pre-processing method to deal with missing values by integrating clustering and regression techniques* Taylor & Francis, 2003.
[10] X. Wu, V. Kumar, J. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, Z. Zhou, M. Steinbach, D. Hand and D. Steinberg *Top 10 algorithms in data mining* Knowl Inf Syst, 2008.
[11] G.F. Üçtug, N.E.Kabakc, O. Bugu Bekdikhan and B. Akyürek *Multi-Criteria Decision Making-Based Comparison of Power Source Technologies for Utilization in Automobiles* Vol. 3, No.3, May 2015. Journal of Clean Energy Technologies

[12] A. Pirdavani, T. Brijs, G. Wets *A multiple criteria decision making approach for prioritizing accident hotspots in developing countries in the absence of crash data.* 2009.