Available online at www.worldscientificnews.com

World Scientific News

An International Scientific Journal

WSN 103 (2018) 77-93 EISSN 2392-2192

Road accidents frequency control using Bayesian Networks

Francisco Soler-Flores[1,a], **Nicoleta González-Cancelas**[2,b], **Beatriz Molina Serrano**[2,c]

[1]Education Faculty, Universidad Internacional de La Rioja, Calle de Almansa, 101, 28040 Madrid, Spain

[2]Civil Engineering Department, Transport, Universidad Politécnica de Madrid, Avda. Profesor Aranguren s/n 28040 Madrid, Spain

[a-v]E-mail address: francisco.soler@unir.net , nicoleta.gcancelas@upm.es , beatriz.molinas@alumnos.upm.es

ABSTRACT

Theory and application of rare events have been very important in recent years due to its practical importance in very different fields such as insurance, finance, engineering or environmental science. This paper presents a methodology for predicting rare events based on Bayesian Networks which in turn enables the study alternative scenarios to control the frequency of road accidents. This way the model Naive-Poisson and ROCDM is presented in this paper for its validation. The developed model is used to estimate and predict road accidents as rare events and results have been evaluated by using ROCDM curve. Naive-Poisson model and a validation model based on ROC curve is used to study several Spanish roads and the results are here shown.

***Keywords*:** Bayesian Networks, road accidents, estimation, probability

(Received 20 June 2018; Accepted 05 July 2018; Date of Publication 06 July 2018)

1. INTRODUCTION

In recent decades, many techniques have been developed for data analysis and modeling in different areas of both statistics and Artificial Intelligence, and several have been applied to the study of rare events. Data mining techniques, which include those that operate automatically with minimum human intervention, are usually efficient to work with large amounts of information available in the databases of many practical problems. Thus, Artificial Intelligence models such as Bayesian Networks are used in this paper in order to estimate the probability of occurrence of a rare event.

Traffic engineering is a branch of civil engineering that uses engineering techniques to achieve the safe and efficient movement of people and goods on roadways. It focuses mainly on research for safe and efficient traffic flow, such as road geometry, sidewalks and crosswalks, segregated cycle facilities, shared lane marking, traffic signs, road surface markings and traffic lights. Traffic engineering deals with the functional part of transportation system, except the infrastructure provided.

The development of infrastructure, roads and highways in todays world involves social and economic problems caused by road accidents. Road accidents are an important problem for developed countries and are relevant in politics. The study of the relationship between the frequency of road accidents and road characteristics, traffic, the environment and users, is also one of the most important application of statistical analysis in the field of road safety. The mathematical models used to estimate the frequency of accidents throughout history have been different, from the linear regression to models multivariate regression as logistic regression or Poisson regression, or more recently, Cluster Analysis and Classification Trees. Models of data analysis in the field of Artificial Intelligence, such as Neural Networks or Bayesian Networks, are beginning to be used in problems related to the study of the frequency of accidents. However, the Generalized Linear Model is currently the most accepted and, therefore, the more used. Thus, it is widely recognized that the distribution of the frequency of accidents follows a Poisson distribution, but there is not a standard model for the study of this problem in the scientific community. Furthermore, there is no software that enables to work with accident data to predict these frequencies.

An event Et [1] is an observation that happens in an instant t and is described by a set of values. Likewise, an event sequence is a sequence of events ordered temporally, S = {Et1, Et2, … Etn} that includes all events timespan $t1 \leq t \leq tn$. Events are associated with a domain object D, which is the source or generator of events. The target event is the event to predict and specified by a set of variables.

In this century, the theory and applications of rare events and extreme events have been a huge increase in interest. This is due to their practical relevance in different fields such as insurance, finance, engineering, environmental science or hydrology. Treatment of rare events, events that occur with a low probability, is a complex and comprehensive problem whose treatment falls within the scope of modeling uncertainty and decision theory. The ’Law of Rare Events’, demonstrated by Poisson, mathematically based the concept of rare occurrence. This law that bears his name is also called law of rare events [2] or also called ’law of small numbers’ [3].

In recent decades, numerous techniques for data analysis and modeling have been developed in different areas of statistics [4, 5] and Artificial Intelligence [6]. These have been applied to the study of rare events. In this context, Data mining (MD) [7] is a modern interdisciplinary area that encompasses techniques that operate automatically (require minimal human intervention) and also are efficient to work with great amounts of information available in the databases of many practical problems. These techniques can extract useful knowledge (associations between variables, rules, patterns, etc.) from the stored raw data, enabling better analysis and understanding of the problem. In some cases, this knowledge can also be post- processed automatically and it can benefit conclusions, and even make decisions almost automatically in specific practical situations (intelligent systems). The practical application of these disciplines extends too many commercial or research prediction problems in classification or diagnosis fields.

The authors developed the model and a proposed application for prediction of road accidents based on the work done for the development of the doctoral thesis ”Estimating rare events using Bayesian Networks”.

2. PROPOSED METHODOLOGY

2. 1. Bayesian Networks

Among the various techniques available in data mining, Bayesian Networks or probabilistic networks enable to model all the relevant information for a problem and draw conclusions based on the available evidence problem by using probabilistic inference mechanisms. Bayesian Networks have been used in the context of the estimation of rare event occurrence in some jobs [8, 9] without actually signaling a general estimation method.

Bayesian Networks [10, 11] are a compact representation of a multivariate probability distribution. Formally, a Bayesian network is a directed acyclic graph where each node represents a random variable and dependencies between variables are encoded in the structure of the graph (Figure 1) according to the criterion of d-separation [12]. Associated with each node in the network is a conditional distribution of the parents of that node probability, so that the joint distribution factored as the product of conditional distributions associated with the nodes of the network. That is, for a network with n variables X1, X2, …, Xn (equation 1).

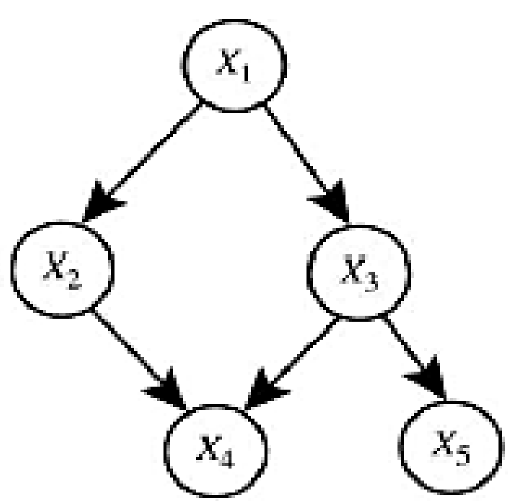

Figure 1. Example of Bayesian network

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i | pa(x_i)) \quad (1)$$

Bayesian Networks usually consider discrete or nominal variables; then these must be discretized before building the model. Although there are models of Bayesian Networks with continuous variables, these variables are limited to Gaussian and linear relationships. Discretization methods are divided into two main types: supervised and unsupervised [13].

The concept of causality [14] in a Bayesian network results in a particular case of these called causal network [15]. Bayesian Networks may have a causal interpretation and although often used to represent causal relationships, the model does not have to represent them in this way, Naive-Bayes is an example of this, relationships are not causal. Bayesian Networks automate the process of probabilistic modeling [16] by using the expressiveness of the graph. The resulting models combine results of graph theory (to represent the relations of dependence and independence of all variables) and the probability (to quantify these relationships). This union enables both efficient machine learning model, through the calculation of parameters [17] which is modeled by a Beta distribution for the case of binary variables and multi-valued variables, the Dirichlet distribution (Table 1) and on the other side hand, inference from the available evidence. The knowledge base of such systems is an estimate of the joint probability function of all variables in the model, while the reasoning module is where the calculation of conditional probabilities is done. The study of this technique provides a good overall view of the problem of statistical learning and data mining.

Table 1. Parameter Estimation

Estimador	**Expresin**
Maximum Likelihood. Multinomial	$\theta_k^* = \frac{N_k}{N}$
Bayesian estimation. Dirichlet	$\theta_k^* = \frac{N_K + a_k}{N + \sum_{i=1}^{r} a_i}$

2. 2. Naive Bayes

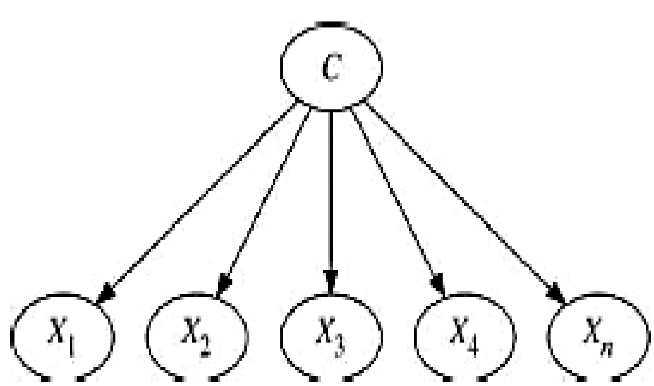

Figure 2. Naive Bayes k

One of the simplest ways that can be devised by considering the structure of a Bayesian network with qualifying goals, is the called Naive-Bayes [18]. Its name comes from the naive assumptions on which it is constructed, which is to consider that all predicting variables are conditionally independent given the variable rated (Figure 2).

The Naive Bayes model is very used because [7]:

- It is simple to build and understand.
- The induction process is fast
- It is very sturdy considering irrelevant attributes.
- It uses many attributes to make the final prediction.

Its predictive power is competitive with other classifiers existing, Nave-Bayes is one of the most effective classifiers. This classifier learns the conditional probability of each attribute Xi given the class C from a training set. The classification process is obtained by applying the Bayes rule, calculating the probability of C, given the instances of X1, X2, …, Xn and taking the highest posterior probability as predicted class. These calculations are based on a strong independence assumption: all attributes Xi are conditionally independent given the value of the C class.

The probability that the j-th instance belonging to the class i-th of the C variable can be applied simply by applying the Bayes theorem, as follows (equation 2)

$$P\left(C=\theta_i \middle| X_1=x_{1j},\dots,X_n=x_{nj}\right) \propto P(C=\theta_i)\cdot P\left(X_1=x_{1j},\dots,X_n=x_{nj} \middle| C=\theta_i\right) \quad (2)$$

Since we assume that the predictor variables are conditionally independent given the C variable, we obtain that (equation 3)

$$P\left(C=\theta_i \middle| X_1=x_{1j},\dots,X_n=x_{nj}\right) \propto P(C=\theta_i)\cdot \prod_{r=1}^{n} P\left(X_r=x_{rj} \middle| C=\theta_i\right) \quad (3)$$

The Naive-Bayes model combined with the Poisson distribution was used for text classification in the paper of [19] with good results. In this paper, the addition of data mining using Bayesian Networks and applying the known probability distribution for the study of rare events is proposed. This form and the known values of the variables used as predictors can study different scenarios and see when the occurrence of a rare event is more likely.

2. 3. Naive-Poisson for rare events

The phases in which the model is divided are:

- Data preprocessing

  1. Selection of variables and obtaining their values.
  2. Discretization of variables

- Building Bayesian Network using Naive Bayes structure
- Naive-Poisson model
- Estimations

2. 3. 1. Data preprocessing

Bayesian Networks usually considered discrete or nominal variables, so the first step is to discretize and then to build the model. Although there are models of Bayesian Networks with continuous variables, these variables are limited to Gaussian and linear relationships. Discretization methods are divided into two main types: supervised and unsupervised [16].

2. 3. 2. Building Bayesian Network

From the previous phase, the construction of the network consists of the following:

- Qualitative part (structure): identify causal relationships; analyze the variables in terms of dependencies and independence
- Quantitative part (probability assessment): quantify relationships and interactions

In this case a specific model, the Naive-Bayes, suitable for the classification process and the variable for which you want to estimate rare events is proposed to be the variable named parent network.

We opted for the development of the methodology for Naive Bayes model. The Naive Bayes model is very used because it has, inter alia, certain advantages:

- It is simple to build and understand.
- The inductions are extremely fast, requiring only a step to do so.
- It is very sturdy considering irrelevant attributes.

Once the structure of the network is given, the probability tables that enable the decomposition of the probability distribution are calculated and then inference is based on evidences and the probability distribution associated with the network is calculated.

This information provides the complete network built by using the Naive-Poisson model described below.

2. 3. 3. Naive-Poisson model

The model or procedure by which, from the Bayesian network built by Naive-Bayes model, we obtain the associated probability distribution with estimating the probability of occurrence of a rare event is called Naive-Poisson [8].

The process of assigning the probability distribution consists of the following sections:

- Poisson Assumption
- Building probability distribution

It is accepted in the scientific community that the distribution of the frequency of rare events is consistent with a Poisson distribution [2]. Thus, it is assumed for the model building, the frequency of rare events follows a Poisson distribution. After obtaining the values of the actual distribution, data were adjusted for this type of distribution [8, 9].

The probability distribution provides Bayesian network constructed to estimate the probability for different values of each of the values of variables that provides the discretization. From the results obtained, the network is adjusted with a Poisson distribution, which as we know is determined by its mean (equation 4).

$$P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!} \quad (4)$$

For the calculation of the parameter that determines the Poisson distribution, the mean, their maximum likelihood estimator is taken, which is given by the equation 5, where xi are the discrete values of the accident and p(xi) the probability values which are provided by the network built through the Naive-Bayes algorithm, being C the discrete variable to classify, the variable whose values considered rare events we want to study.

$$\lambda = \sum_{i=1}^{n} x_i \cdot p(x_i) \quad (5)$$

For each of the values of the strata a1,a2, …, an that provides the discretization of the variable to study rare events, the actual values of the frequency of the rare event to estimate are taken except for the an to which a given value is assigned by the average of the values higher than xi to the higher stratum an-1 (equation 6).

$$a_n = \bar{x}_i \; con \; x_i > a_{n-1} \quad (6)$$

This way we obtain the poisson distribution associated with any of the situations (j = 1 ldots, m) to study (equation 7) with any set of values of different selected variables, which allows to determine the situation where there is a higher probability of occurrence of low probability events once having detected the values of the variable studied, which given their distribution, rare events are of a low probability.

$$P_j(X = x) = e^{-\lambda j} \frac{\lambda_j^x}{x!} \quad (7)$$

3. VALIDATION. ROCDM CURVE

A ROC curve (Receiver Operating Characteristic acronym or Receiver Operating Characteristic) is a graphical plot of the sensitivity vs. (1-specificity) for a binary classifier system as discrimination threshold.

Discrete classifiers such as Decision Trees and Rule Systems return numerical results given as a binary label values. A single point in the ROC space is provided when these classifiers are used with a specific set of instances to classify or predict the performance of the classifier. For other classifiers, such as Naive Bayes classifier, a network or Bayesian artificial neural network, the output values are likely to represent the extent to which one belongs to one of two classes, for instance. In this way a new indicator is needed.

This model generalizes the discreet binary case. The idea is to consider the model prediction for each of the values of the variable and calculate the ROC curve for each individual case and treat it as binary.

The proposal becomes a variable with n discrete values (A = a1; a2; ldots; an) that we call a multivalued binary variable, using function called v(x) so that it can be built for each of the discrete values associated ROC curve, obtaining at the end n ROC curves that will describe the overall prediction accuracy of the qualification (Table 2).

Table 2. ROCDM curve

Curve	Values	Area under the curve
a_0	$v(x) = \begin{Bmatrix} 1 & x = a_1 \\ 0 & x \in \overline{A - a_1} \end{Bmatrix}$	s_1
a_2	$v(x) = \begin{Bmatrix} 1 & x = a_2 \\ 0 & x \in \overline{A - a_2} \end{Bmatrix}$	s_2
…	…	…
…	…	…
a_n	$v(x) = \begin{Bmatrix} 1 & x = a_n \\ 0 & x \in \overline{A - a_n} \end{Bmatrix}$	s_n

Thereby obtaining a ROC curve for each of the n values of the variable and when you combine all the ROC curves obtaining the ROCDM resulting graph (Figure 3).

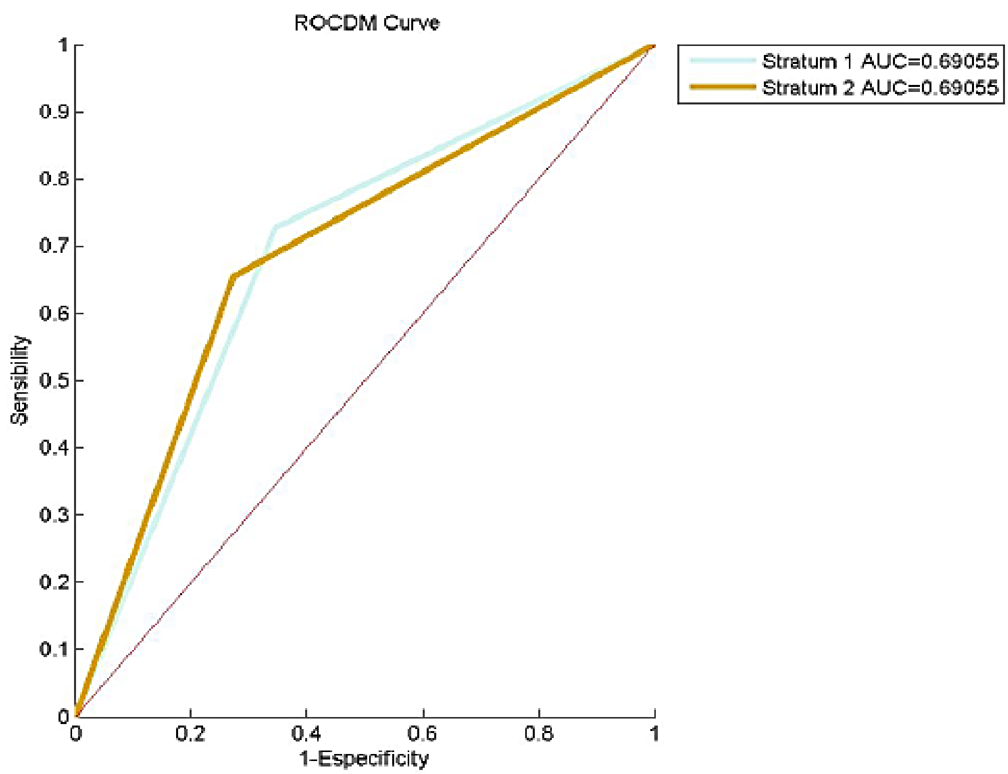

Figure 3. ROCDM curve

The interpretation is the same as the binary model but including all variable values in the same graph. The area under each curve (defined by the diagonal of the first quadrant) is the value of the ROC curve for each value of the variable, so the value of the area of the curve ROCDM as the average of the n is defined areas atextrmROCDM = overlines1, s2; ldots, sn, which takes values between 1 (perfect test) and 0:5 (useless test).

4. IMPLEMENTATION

Based on public data from Spanish roads, for a period of five years, with a selection of sections of 500 meters and the selection of variables given in Table 3 and described in ?? applying the Naive-Poisson model the probability distributions of the frequency of accidents for each section and each variable is obtained, then these variables are discretized in response to the strata given in Table 4 .

To observe rare events represents the case of more than 10 accidents on a stretch in the studied period, as road accidents in traffic engineering represent a rare event, ie an event that occurs with a very low probability.

Table 3. Variable's description

Variable	Description
IMD	Flow (veh/day)
ACC	Road accident frequency
DACIN	Intersection and access density (nt/km)
INME	Medium slope of the section
PPAD	No passing proportion
RV85M	Decrease in specific speed relative to the adjacent sections of 1 km (km/h)

$$[a_0, a_1, \dots, a_n] = \begin{Bmatrix} 1 & if & x \leq a_0 \\ 2 & if & a_0 < x \leq a_1 \\ \cdots & \cdots & \cdots \\ n & if & a_{n-1} < x \leq a_n \\ n+1 & if & x > a_n \end{Bmatrix}$$

Table 4. Discretization

Variable	Discretization
IMD	[524.2, 1055.2,2104.4]
ACC	Model 3
DACIN	[0, 1, 2, 3, 4, 5, 6, 7, 8]
INME	[10, 15, 20, 25, 30, 35, 40, 45]
PPAD	[0, 0.25, 0.75]
RV85M	[5, 15, 20]

After this discretizations an accidents discretization is needed. In this way a preselection of this is studied (Table 5).

Table 5. Discretization of ACC

Analysis	Discretization
1	[10, 20, 30, 40, 50, 60, 70, 80, 90, 100]
2	[100]
3	[10]
4	[10, 100]
5	[50, 100]
6	[50]
7	[20]
8	[20, 40]
9	[30]

To select the best prediction, rare events values predicted is analyzed by studying the different values of probability (Figure 4) and testing their probabilities (Figure 5).

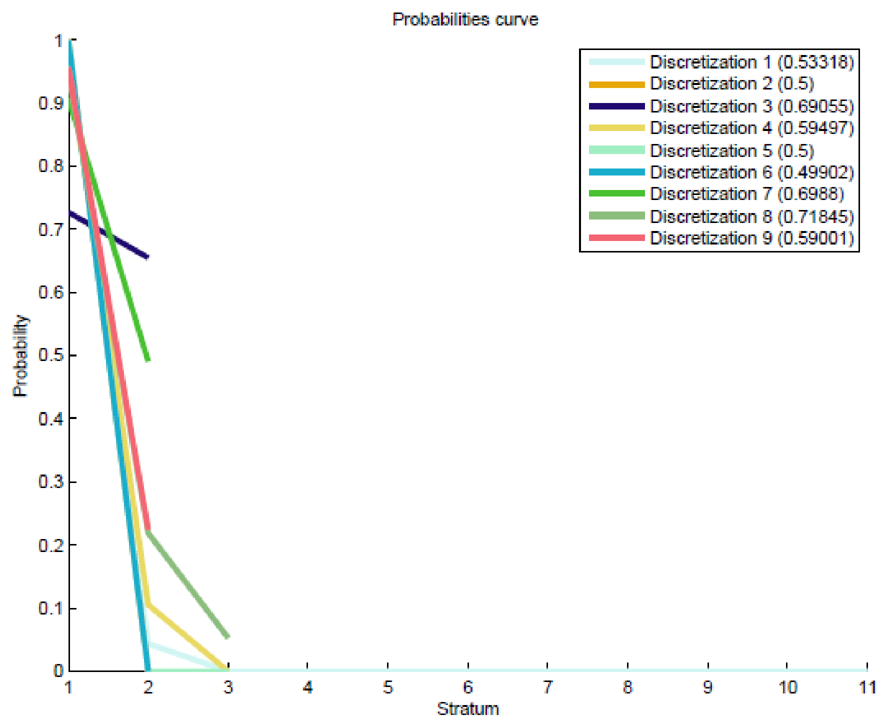

Figure 4. Probability comparison

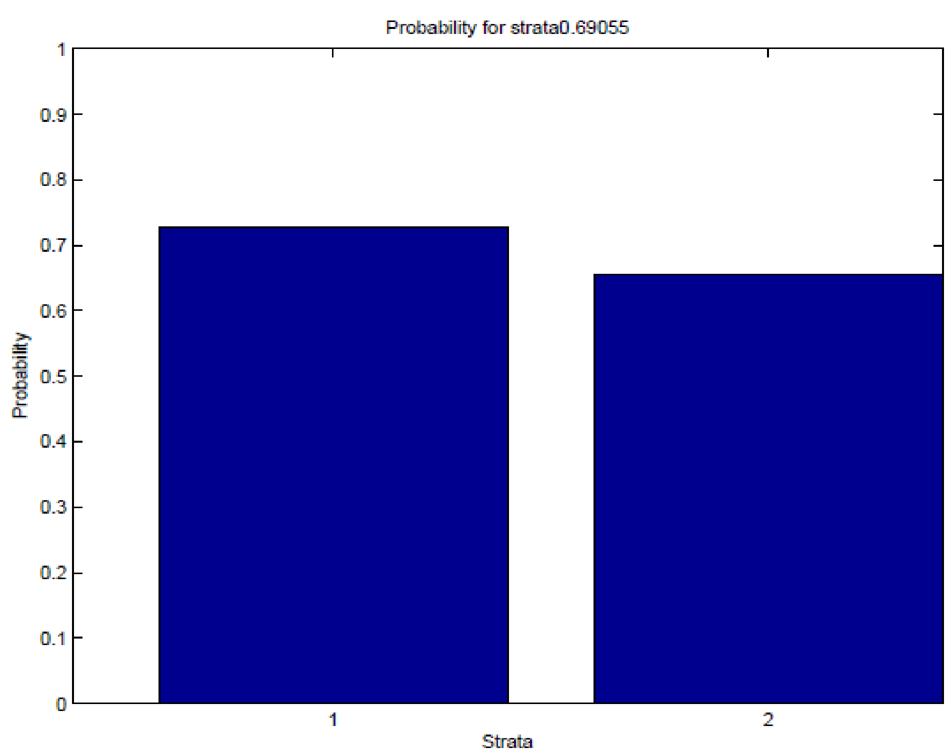

Figure 5. Probabilities

This way, the selected discretization is in Figure 6

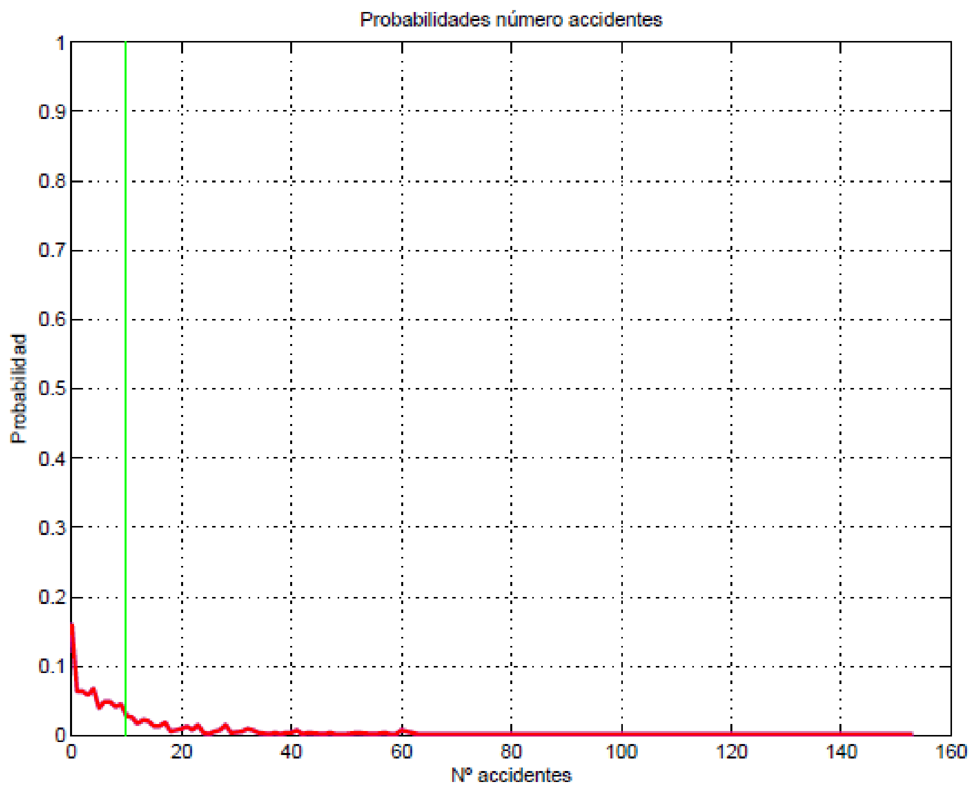

Figure 6. Stratification for accidents

For an example of section selected, the status for each of the variables described. In this selected example section, in each of the figures, the observed distribution of accidents for each of the strata of each variable and in black on the situation study.

4. 1. IMD analysis

(Figure ref fig: IMD) shows that the distribution of accidents for the case 13 makes it more likely for more than 10 accidents but the differences between the other cases are not significant, that is, only AADT in 13 layer has a high probability with respect to the other to be rare events.

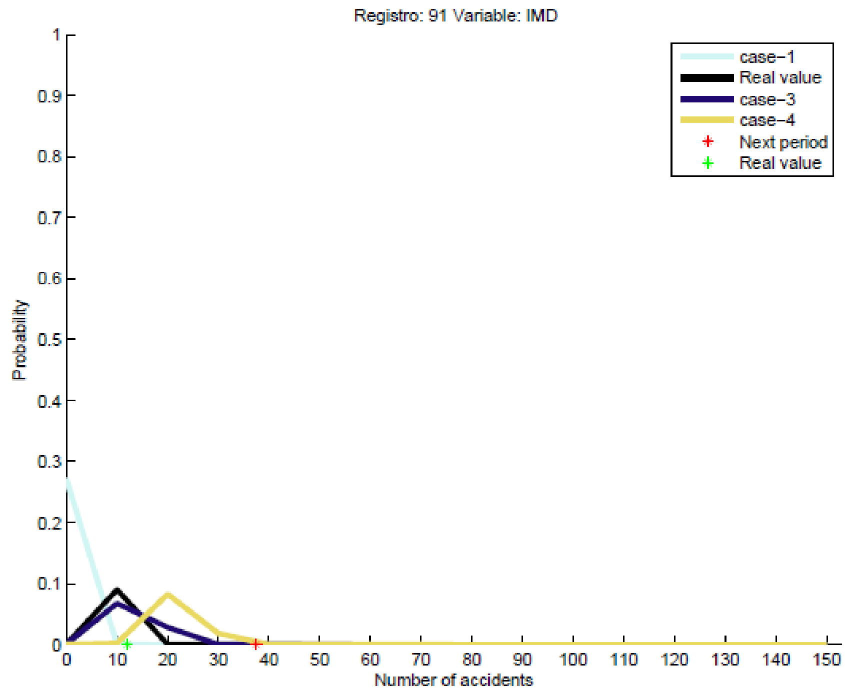

Figure 7. AADT

4. 2. PPAD analysis

In this case (Figure 8) it shows that modifying the section access density stratum 1 would be a lower likelihood of rare events.

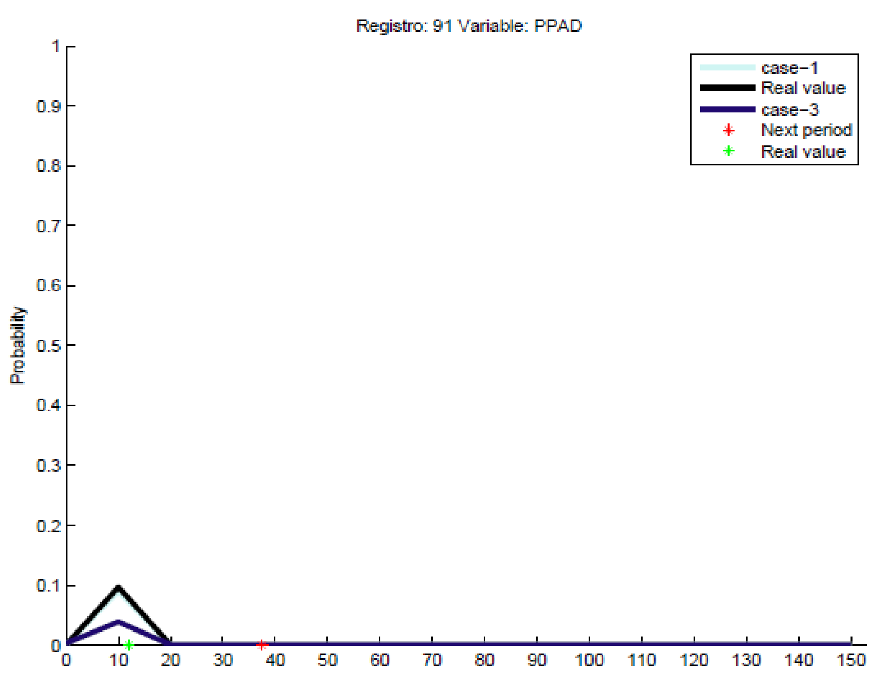

Figure 8. PPAD

4. 3. INME analysis

The IVISI variable in the stratum, in which it is located (Figure 9) for section example, could be modified to improve the situation with rare events.

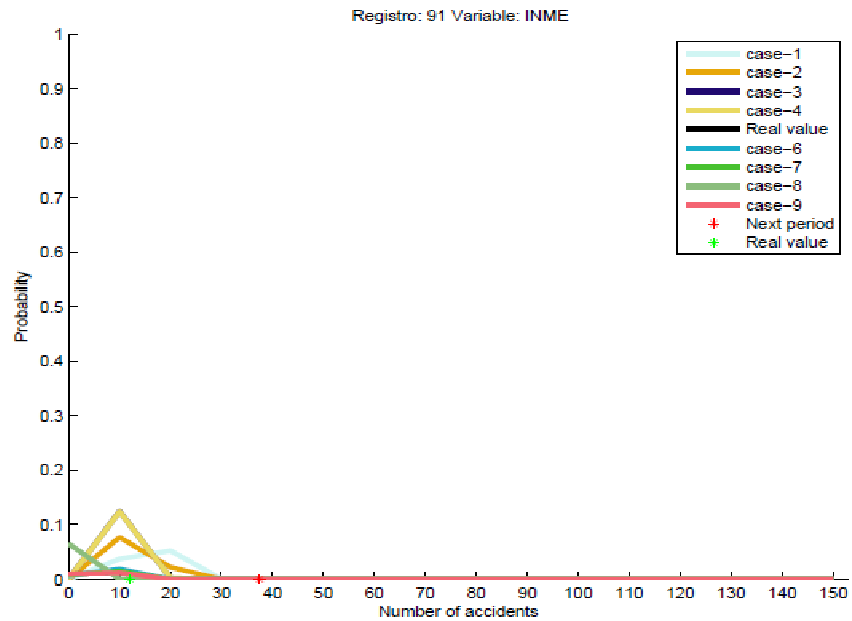

Figure 9. INME

4. 4. DACIN analysis

In this particular section the DCIN variable (Figure 10) can be improved.

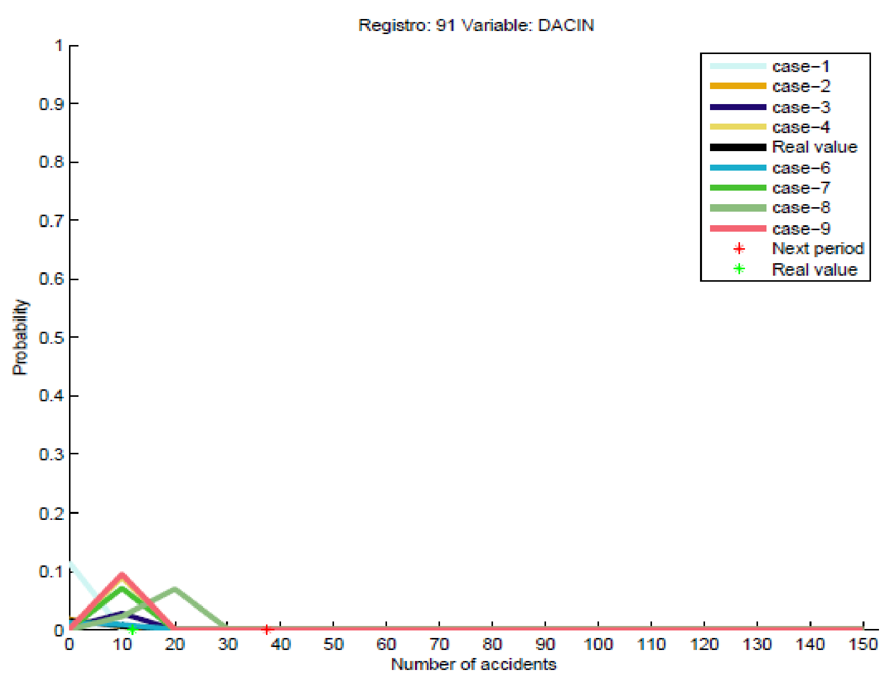

Figure 10. DACIN

4. 5. RV85M analysis

The layer provides 6 for RV85M variable (Figure 11) lower probability for the occurrence of rare events.

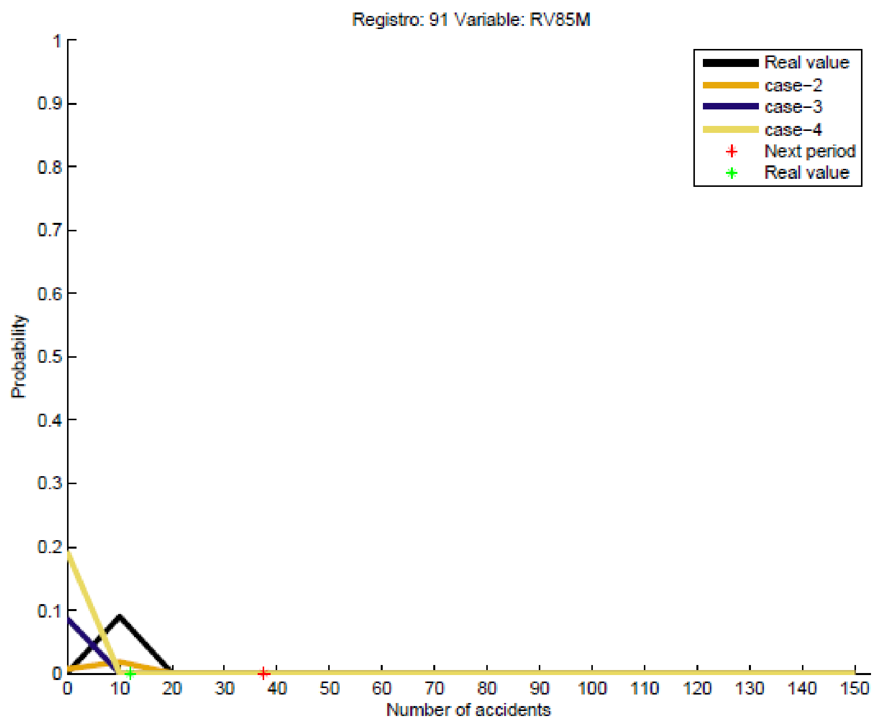

Figure 11. RV85M

5. CONCLUSIONS

This paper presents a methodology and an application of such methodology to determine the frequency of traffic accidents and control it. The main contributions of this paper are summarized as follows:

1 Treatment of rare events is essential in real trouble and that is why the development of a model, and a methodology for the treatment of these allowing the use of systematic way, becomes necessary.

2 From the raw data and by defining the variable from which you want to estimate their low probability events, the developed model defines the probability distribution of the same and compares different alternatives, thus taking appropriate decisions based on the results. The Naive-Poisson model combines the qualifying potential Naive-Bayes model for Bayesian Networks taken with the classic setting of the frequency distribution of so called ”rare events”, the Poisson distribution.

3 The curve called ROC allows draw the probability of correctly classifying a binary variable. The model provided in this paper, ROCDM curve allows extend the original one which can be considered non-binary variables.

4 This paper develops from applying them to real cases, and usage models listed are presented for estimating low probability events and study different alternatives. The developed model in this paper is valid for its application to any estimation problem of occurrence of low probability events.

It has already shown that is possible to model the events of low probability of occurrence by using Bayesian Networks and their possible application to certain problems. It has great impacts and practical implications on a wide range of applications. Since the proposed framework is robust to large within-class variations, it can be used in industry too, in failure analysis. Although the framework proposed work has outperformed, the existing methods are much room for improvement, the data that we used to evaluate the proposal system are relatively simple.

Aspects to be considered in future research to continue the work developed in this paper, can be delimited in the following lines:

- Extension to other learning algorithms such as K2 or DVNSST.
- Generalization of the methodology to problems within the paradigm of Big Data. Map-Reduce algorithms.
- Development of applications implementing the methodology.
- Application of the model to other case studies.

Road accidents have become one of the most serious problems of public health. Factors of physical, technical, weather, deficiency of quality of the road network, the behavior, cognitive and civic / road training has been representing some of the possible causes of accidents that occur today. In this situation the countries have been designing preventive strategies and research which attempts to detect what type of variables can affect the accident rate. In this way we try to decrease the great material cost great material that derives from this fact. The Naive-Poisson model allows the study and analysis of different alternatives for reduce the frequency of accidents.

References

[1] G.M. Weiss & H. Hirsh, H. Learning to Predict Rare Events in Event Sequences. *KDD Journal*, p. 359-363, 1998

[2] H. Aytaç, J. Freitas & S. Vaienti. Laws of rare events for deterministic and random dynamical systems. *Transactions of the American Mathematical Society*, vol. 367, no 11, p. 8229-8278, 2015.

[3] A. Tversky & D. Kahneman. Belief in the law of small numbers. *Psychological bulletin*, vol. 76, no 2, p. 105, 1971.

[4] G. King & L. Zeng. Logistic regression in rare events data. *Political analysis*, vol. 9, no 2, p. 137-163, 2001

[5] N. González-Cancelas, F. Soler-Flores, A. Camarero Orive & I. López Ansorena. Tratamiento de outliers para el estudio de transmisión de vibraciones del ferrocarril. *Ingeniería y Ciencia*, vol. 8, no 16, 2012

[6] C. Seiffert, T.M. Khoshgoftaar, J. Van Hulse & A. Napolitano. Mining data with rare events: a case study. In Tools with Artificial Intelligence, 2007. ICTAI 2007. 19th *IEEE International Conference* on (Vol. 2, pp. 132-139). IEEE, October 2007

[7] E.W.T. Ngai, L. Xiu & D.CK Chau. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, vol. 36, no 2, p. 2592-2602, 2009.

[8] S.P. Cheon, S. Kim, S.Y. Lee & C.B. Lee. Bayesian networks based rare event prediction with sensor data, *Knowledge-Based Systems* 22 (5), 336–343, 2009.

[9] A. Ebrahimi & T. Daemi. Considering the rare events in construction of the bayesian network associated with power systems, *Probabilistic Methods Applied to Power Systems (PMAPS)*, IEEE 11th International Conference on, IEEE, 659–663, 2010.

[10] J. Pearl. Fusion, propagation, and structuring in belief networks. *Artificial intelligence*, vol. 29, no 3, p. 241-288, 1986.

[11] J. Pearl. Probabilistic reasoning in intelligent systems: networks of plausible inference, Morgan Kaufmann, 1988.

[12] E. Castillo, J.M. Gutierrez & A.S. Hadi. Expert systems and probabilistic network models, Springer Verlag, 1997.

[13] J. Dougherty, R. Kohavi & M. Sahami, M. Supervised and unsupervised discretization of continuous features, *ICML*, 194–202, 1995.

[14] C.P. Agueda. Causality in sciencie, *Pensamiento Matemático* (1) 12, 2011.

[15] J. Pearl. Causality: models, reasoning and inference, Vol. 29, Cambridge Univ. Press, 2000.

[16] L. Uusitalo. Advantages and challenges of bayesian networks in environmental modelling, *Ecological Modelling* 203 (3), 312–318, 2007.

[17] O. Y. Al-Jarrah, P.D. Yoo, S. Muhaidat, G. K. Karagiannidis & K. Taha. Efficient machine learning for big data: A review. *Big Data Research*, vol. 2, no 3, p. 87-93, 2015

[18] Y. Ng & M. I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, p. 841-848, 2002.

[19] S.B. Kim, H.C. Seo, H.C. Rim. Poisson naive bayes for text classification with feature weighting, Proceedings of the *sixth international workshop on Information retrieval with Asian languages* Volume 11, Association for Computational Linguistics, 33–40, 2003.