Investigación

Sucesos raros en Ingeniería de Tráfico

Rare events in Traffic Engineering

Francisco Soler Flores, J. Angel Olivas Varela y María Dolores López González

Revista de Investigación

Volumen V, Número 1, pp. 063–074, ISSN 2174-0410
Recepción: 31 Mar'14; Aceptación: 2 Ene'15

1 de abril de 2015

Resumen

La Ingeniería de Tráfico, como rama de la ingeniería del transporte y de la ingeniería civil trata la planificación, diseño y operación de tráfico en las calles, carreteras y autopistas, sus redes, infraestructuras, tierras colindantes y su relación con los diferentes medios de transporte para conseguir una movilidad segura, eficiente y conveniente tanto de personas como de mercancías.

La ingeniería de tráfico estudia los accidentes de tráfico, éstos por ser sucesos de muy baja probabilidad, pueden ser considerados sucesos raros. Así, este trabajo recoge un resumen de resultados de la aplicación del modelo Naive-Poisson a la Ingeniería de Tráfico enfocado a, mediante la estimación de accidentes de tráfico, propuestas de mejora en el diseño de algunas carreteras españolas.
Palabras Clave: Naive-Poisson, Redes Bayesianas, Ingeniería de Tráfico, accidentes de tráfico.

Abstract

Traffic Engineering is a part of transportation engineering and civil engineering and it is the planning, design and operation of traffic on the streets, roads and highways, their networks, infrastructure, adjoining land and its relationship with the different modes of transport to get thi safe, efficient and convenient for people and goods.

Traffic engineering studies the traffic accidents, these are very low probability events and can be considered rare events. Thus, this paper presents a summary of results from the application of Naive-Poisson model to the Traffic Engineering focused, by estimating traffic accidents, suggestions for improvements in the design of some Spanish roads.

Keywords: Naive-Poisson, Bayesian Networks, Traffic Engineering, road accidents.

1. Introducción

Un evento E_t ([1]) es una observación que ocurre en un instante t y es descrita por un conjunto de valores. De la misma forma, una secuencia de eventos es una secuencia de eventos ordenada temporalmente, $S = \{E_{t_1}, E_{t_2}, \ldots E_{t_n}\}$ que incluye todos los eventos de intervalo de

tiempo $t_1 \leq t \leq t_n$. Los eventos están asociados a un dominio objeto D, el cual es la fuente o generador de los eventos. El *Evento objetivo* es el evento a predecir y especificado por un conjunto de variables.

En este siglo, la teoría y aplicaciones de sucesos extremos y eventos raros han recibido un enorme aumento de interés. Esto es debido a su relevancia práctica en diferentes campos, como los seguros, las finanzas, la ingeniería, las ciencias del medio ambiente o la hidrología.

El tratamiento de sucesos raros, sucesos que ocurren con una probabilidad baja, es un problema complejo y amplio cuyo tratamiento se enmarca en el ámbito de la modelización de la incertidumbre, teoría de la decisión y donde el estudio del riesgo, definido en términos de teoría de decisiones como las pérdidas promedio o las pérdidas que se pronostican cuando algo malo sucede, es importante. La 'Ley de Eventos Raros', demostrada por Poisson, fundamenta matemáticamente el concepto de suceso raro. Esta ley que lleva su nombre es denominada también 'ley de los sucesos raros' [2] o también denominada 'ley de los pequeños números' [3].

En las últimas décadas se han desarrollado numerosas técnicas de análisis y modelización de datos en distintas áreas de la estadística ([4],[5]) y la Inteligencia Artificial ([6]) que se han aplicado al estudio de sucesos raros . La Minería de Datos (MD) ([7]) es un área moderna interdisciplinar que engloba a aquellas técnicas que operan de forma automática (requieren de la mínima intervención humana) y, además, son eficientes para trabajar con las grandes cantidades de información disponibles en las bases de datos de numerosos problemas prácticos. Estas técnicas permiten extraer conocimiento útil (asociaciones entre variables, reglas, patrones, etc.) a partir de la información cruda almacenada, permitiendo así un mejor análisis y comprensión del problema. En algunos casos, este conocimiento puede ser también post-procesado de forma automática permitiendo obtener conclusiones, e incluso tomar decisiones de forma casi automática, en situaciones prácticas concretas (sistemas inteligentes). La aplicación práctica de estas disciplinas se extiende a numerosos ámbitos comerciales y de investigación en problemas de predicción, clasificación o diagnosis.

Los autores, desarrollan en [8] y [9] el modelo y una propuesta de aplicación a la predicción de accidentes de tráfico en base a los trabajos realizados para el desarrollo de la tesis doctoral .Estimación de sucesos poco probables mediante Redes Bayesianas".

2. Redes Bayesianas

Entre las diferentes técnicas disponibles en minería de datos, las redes probabilísticas o redes bayesianas permiten modelizar de forma conjunta toda la información relevante para un problema dado y utilizando posteriormente mecanismos de inferencia probabilística para obtener conclusiones en base a la evidencia disponible. Las redes bayesianas han sido utilizados en el contexto de la estimación de sucesos sucesos raros en algunos trabajos([10] y [11]) sin llegar a señalar un método de estimación general.

Las redes bayesianas ([12], [13]) son una representación compacta de una distribución de probabilidad multivariante. Formalmente, una red bayesiana es un grafo dirigido acíclico donde cada nodo representa una variable aleatoria y las dependencias entre las variables quedan codificadas en la propia estructura del grafo (figura 1) según el criterio de d-separación ([14]). Asociada a cada nodo de la red hay una distribución de probabilidad condicionada a los padres de ese nodo, de manera que la distribución conjunta factorizada como el producto de las distribuciones condicionadas asociadas a los nodos de la red. Es decir, para una red con n variables $X_1, X_2, \ldots, X_n$ (ecuacion 1).

$$\prod_{i=1}^{n} p(x_i|pa(x_i)) \tag{1}$$

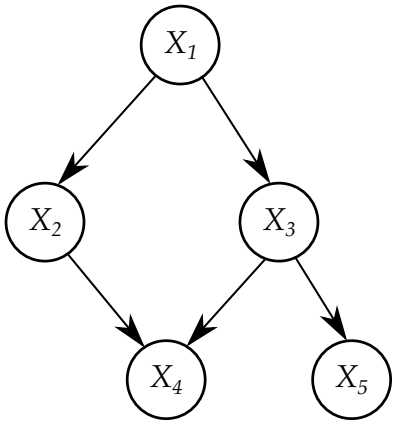

Figura 1. Ejemplo de Red Bayesiana

Normalmente las redes bayesianas consideran variables discretas o nominales, por lo que si no lo son, hay que discretizarlas antes de construir el modelo. Aunque existen modelos de redes bayesianas con variables continuas, éstos están limitados a variables gaussianas y relaciones lineales. Los métodos de discretización se dividen en dos tipos principales: no supervisados y supervisados ([15]).

El concepto de causalidad ([16]) en una red bayesiana se traduce en un caso particular de estas denominado red causal ([17]). Las redes bayesianas pueden tener una interpretación causal y aunque se utilizan a menudo para representar relaciones causales, el modelo no tiene por que representarlas de esta forma, Naive-Bayes es un ejemplo de esto, sus relaciones no son causales.

Las redes probabilísticas automatizan el proceso de modelización probabilística ([18]) utilizando la expresividad de los grafos. Los modelos resultantes combinan resultados de la teoría de grafos (para representar las relaciones de dependencia e independencia del conjunto de variables) y de la probabilidad (para cuantificar estas relaciones). Esta unión permite realizar de forma eficiente tanto el aprendizaje automático del modelo, a través del cálculo de parámetros ([19]) que para el caso de variables binarias se modela con una distribución Beta y para variables multivaluadas mediante su extensión, que es la distrubución Dirichlet (tabla 1),como la inferencia a partir de la evidencia disponible. La base de conocimiento de estos sistemas es una estimación de la función de probabilidad conjunta de todas las variables del modelo, mientras que el módulo de razonamiento es donde se hace el cálculo de probabilidades condicionadas. El estudio de esta técnica proporciona una buena perspectiva global del problema del aprendizaje estadístico y la minería de datos.

Tabla 1. Estimación paramétrica

Estimador	Expresión
Máxima verosimilitud. Multinomial	$\theta_k^* = \frac{N_k}{N}$
Estimación bayesiana. Dirichlet	$\theta_k^* = \frac{N_k + a_k}{N + \sum_{i=1}^{r} a_i}$

2.1. Naive Bayes

Una de las formas más sencillas que se pueden idear al considerar la estructura de una red Bayesiana con objetivos clasificatorios, es el denominado *Naive-Bayes* o *Bayes ingenuo* ([20]). Su denominación proviene de la hipótesis ingenua sobre la que se construye, que consiste en

considerar que todas las variables predictoras son condicionalmente independientes dada la variable a clasificar (figura 2)

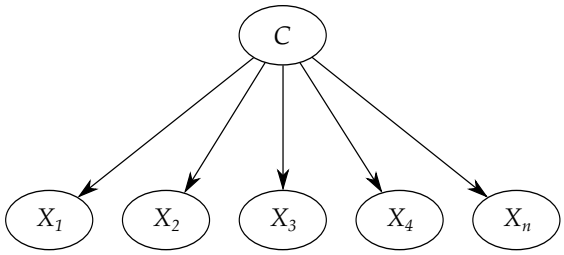

Figura 2. Naive Bayes

El modelo Naive Bayes es muy utilizados debido a que presentan, entre otras, ciertas ventajas ([7]):

- Generalmente, es sencillo de construir y de entender.
- El proceso de inducción es rápido
- Es muy robusto considerando atributos irrelevantes.
- Toma evidencia de muchos atributos para realizar la predicción final.

En el sentido de que su capacidad predictiva es competitiva con el resto de clasificadores existentes, el llamado *Naive-Bayes*, descrito, por ejemplo, por [20] y por [21] es uno de los más efectivos clasificadores. Este clasificador aprende de un conjunto de entrenamiento la probabilidad condicional de cada atributo X_i dada la clase C. La clasificación se hace entonces aplicando la regla de Bayes para calcular la probabilidad de C dadas las instancias de $X_1, X_2, \ldots, X_n$ tomando como clase predicha la de mayor probabilidad a posteriori. Estos cálculos se basan en una fuerte suposición de independencia: todos los atributos X_i son condicionalmente independientes dado el valor de la clase C.

La probabilidad de que el j-ésimo ejemplo pertenezca a la clase i-ésima de la variable C, puede aplicarse sin más que aplicar el teorema de Bayes, de la siguiente manera (ecuación 2)

$$P(C = \theta_i | X_1 = x_{1j}, \ldots, X_n = x_{nj}) \propto P(C = \theta_i) \cdot P(X_1 = x_{1j}, \ldots, X_n = x_{nj} | C = \theta_i) \qquad (2)$$

Dado que suponemos que las variables predictoras son condicionalmente independeientes dada la variable C, se obtiene que (ecuación 3)

$$P(C = \theta_i | X_1 = x_{1j}, \ldots, X_n = x_{nj}) \propto P(C = \theta_i) \cdot \prod_{r=1}^{n} P(X_r = x_{rj} | C = \theta_i) \qquad (3)$$

El modelo Naive-Bayes combinado con la distribución de Poisson es utilizado para la clasificación de texto en el trabajo de [22] con buenos resultados. En este trabajo se propone el añadido de la minería de datos utilizando redes bayesianas y aplicando la distribución de probabilidad conocida para el estudio de sucesos raros. De esta forma y conocidos los valores de las variables que se utilizan como predictoras se pueden estudiar las diferentes situaciones posibles y observar cuándo es más probable la ocurrencia de un suceso raro.

3. Naive-Poisson para tratamiento de sucesos raros

Las fases en las que se divide el modelo son las siguientes:

- Preprocesamiento de los datos
    1. Selección de variables y obtención de sus valores.
    2. Discretización de las variables
- Construcción de la red bayesiana, estructura Naive-Bayes
- Aplicación modelo Naive-Poisson

3.1. Preprocesamiento de los datos

Normalmente las redes bayesianas consideran variables discretas o nominales, por lo que si no lo son, hay que discretizarlas antes de construir el modelo. Aunque existen modelos de redes bayesianas con variables continuas, éstos están limitados a variables gaussianas y relaciones lineales. Los métodos de discretización se dividen en dos tipos principales: no supervisados y supervisados ([15]).

En este apartado de preparación de los datos, éstos son formateados de forma que las herramientas informáticas que se uilicen puedan manipularlos. A su vez, este apartado consiste también en la selección de variables y discretización de los datos en el caso de variables continuas. En este paso se identifica también la variable de la cual se quieren estimar los sucesos raros.

3.2. Construcción de la red bayesiana

A partir de la fase anterior, la construcción de la red consta de los siguientes puntos:

- Identificación de la Estructura: identificar las relaciones causales; analizar las variables en cuanto a dependencias e independencias
- Cálculo de Parámetros (probabilidades): cuantificar relaciones e interacciones

En este caso se propone un modelo concreto, el Naive-Bayes , adecuado para el proceso de clasificación y para el que la variable de la cual se quieren estimar los sucesos raros será la variable denominada padre de la red([14]).

Se va a optar para el desarrollo de la metodología por el modelo Naive Bayes. El modelo Naive Bayes es muy utilizados debido a que presentan, entre otras, ciertas ventajas:

- Generalmente, es sencillo de construir y de entender.
- Las inducciones de son extremadamente rápidas, requiriendo sólo un paso para hacerlo.
- Es muy robusto considerando atributos irrelevantes.
- Toma evidencia de muchos atributos para realizar la predicción final.

Una vez determinada la estructura de la red se calculan las tablas de probabilidades que permiten realizar la descomposición de la distribución de probabilidad y posteriormente realizar inferencia basándose en las evidencias obtenidas y la distribución de probabilidad asociada a la red.

Esta información que proporciona la red construida se completa aplicando el modelo Naive-Poisson descrito a continuación.

3.3. Modelo Naive-Poisson

El modelo o procedimiento por el que, a partir de la red bayesiana construida mediante el modelo Naive-Bayes, obtenemos la distribución de probabilidad asociada que permite estimar la probabilidad de ocurrencia de un suceso raro es denominado Naive-Poisson ([8]).

El procedimiento de asignación de la distribución de probabilidad consta de los siguientes apartados:

- Suposición Poisson
- Construcción de la distribución de probabilidad

Es admitido en la comunidad científica que la distribución de la frecuencia de sucesos raros se ajusta a una distribución de Poisson ([2]). Así, se supondrá para el modelo a construir, la frecuencia de de sucesos raros sigue una distribución de Poisson. Una vez obtenidos los valores de la distribución real se ajustarán los datos por este tipo de distribución ([8],[9]). La distribución de probabilidad que proporciona la red bayesiana construida permite estimar los diferentes valores de probabilidad para cada uno de los valores de las variables que proporciona la discretización. A partir de los resultados obtenidos, la red se ajusta con una distribución de Poisson, que como es sabido viene determinada por su media (ecuacion 4).

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

Para el cálculo del parámetro que determina la distribución e Poisson, la media, tomaremos su estimador máximo verosímil que viene dado por la ecuación 5, donde x_i son los valores discretos de la accidentalidad y $p(x_i)$ los valores de probabilidad que proporciona la red construida mediante el algoritmo Naive-Bayes , siendo la variable discreta a clasificar C, la variable de la que queremos estudiar los valores que consideramos sucesos raros y el resto, las variables utilizadas para tal fin el resto de las variables.

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

Para cada uno de los valores de los estratos $a_1, a_2, \ldots, a_n$ que proporciona la discretización de la variable a estudiar los sucesos raros, se tomarán los valores reales de la frecuencia del suceso raro a estimar excepto para el a_n para el que se asigna un valor dado por la media de los valores x_i superiores al mayor estrato a_{n-1} (ecuación 6).

$$a_n = \bar{x_i} \text{ con } x_i > a_{n-1} \tag{6}$$

De esta forma descrita obtenemos las distribución de poisson asociada a cualquiera de las situaciones ($j = 1 \ldots, m$) que estudiemos (ecuación 7) con cualquier conjunto de valores de las diferentes variables seleccionadas , con lo cual es posible determinar en que situación es más alta la probabilidad de ocurrencia de los sucesos de baja probabilidad una vez detectado cuales son los valores de la variable estudiada que dada su distribución, los sucesos raros son de probabilidad más baja.

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

4. Resultados

A partir de datos públicos de carreteras españolas, para un periodo de cinco años, con una selección de tramos de $500m$ y de la selección de variables dada en la tabla 2 en la que se describen las mimas, a continuación se discretizan dichas variables atendiendo a los estratos dados en la tabla 3 y aplicando el modelo Naive-Poisson se obtienen las distribuciones de probabilidad de la frecuencia de accidentes para cada tramo y cada variable.

Los sucesos raros a observar lo representa el caso de más de 10 accidentes en un tramo en el periodo de tiempo de estudio, dado que los accidentes de tráfico en ingeniería de tráfico suponen un suceso raro o infrecuente, es decir un suceso que ocurre con una probabilidad muy pequeña.

Tabla 2. Descripción de variables

Variable	Descripción
IMD	Intensidad media diaria (veh/día)
ACC	Núumero total de accidentes con víctimas
DACIN	Densidad de accesos e intersecciones (accesos/km)
IVISI	Índice de visibilidad
RMIN	radio mínimo de curvatura en el tramo
LIMV	Valor mínimo del límite de velocidad señalizado (km/h)
INCM	Valor máximo de la inclinación en el tramo de (%)
r4ve	Disminución de velocidad específica respecto de los tramos de 1 km contiguos (km/h)

$$[a_0, a_1, \ldots, a_n] = \begin{cases} 1 & si \quad x \leq a_0 \\ 2 & si \quad a_0 < x \leq a_1 \\ \ldots & \ldots \quad \ldots \\ n & si \quad a_{n-1} < x \leq a_n \\ n+1 & si \quad x > a_n \end{cases}$$

Tabla 3. Discretización de las variables

Variable	Discretización
IMD	[2000,3000,4000,5000,6000,7000,8000,9000,10000,15000,20000,25000]
ACC	[1,2,3,4,5,6,7,8,9,10]
DACIN	[0]
IVISI	Variable discreta
RMIN	[150,300,600,99998]
LIMV	[40,60,80,100]
INCM	[2,3,4,5,6,7]
r4ve	[0,10,20,30,40]

Para un tramo-ejemplo seleccionado, se describe la situación para cada una de las variables. En este tramo ejemplo seleccionado se observa, en cada una de las figuras, la distribución de accidentes de tráfico para cada uno de los estratos de cada variable y en color negro el de la *situación actual* de estudio.

4.1. Análisis IMD

Se observa (figura 3) que la distribución de accidentes para el caso 13 hace más probable el caso de más de 10 accidentes pero las diferencias entre los otros casos no son significativas, es decir sólo IMD en el estrato 13 presenta una probabilidad alta con respecto a las demás de tener sucesos raros.

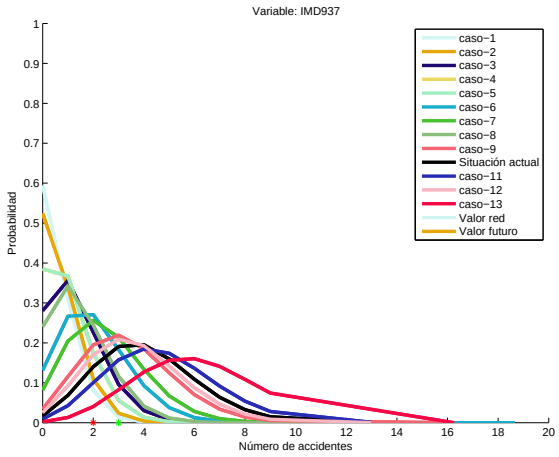

Figura 3. IMD

4.2. Análisis DACIN

En este caso (figura 4) se observa que modificando la *densidad de accesos* en el tramo al estrato 1 habría menos probabilidad de ocurrencia de sucesos raros.

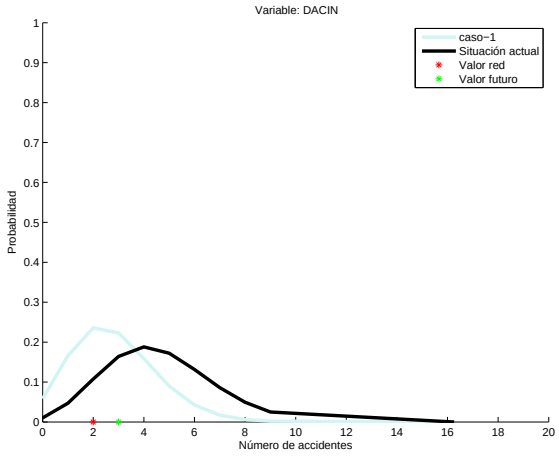

Figura 4. DACIN

4.3. Análisis IVISI

La variable IVISI en el estrato en el que se encuentra (figura 5) para el tramo ejemplo podría ser modificada para mejorar la situación ante eventos raros.

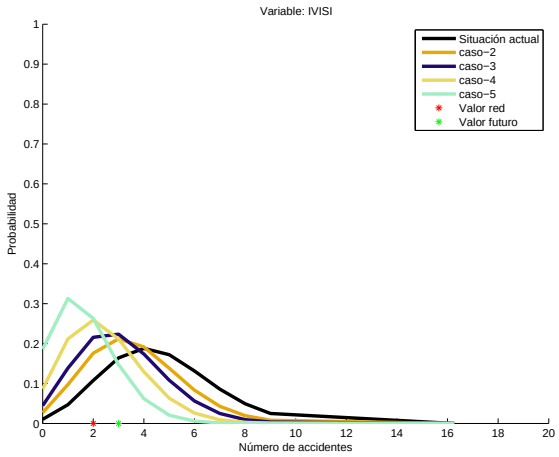

Figura 5. IVISI

4.4. Análisis RMIN

En este tramo concreto la variable RMIN (figura 6) no proporciona diferencias significativas en sus diferentes estratos.

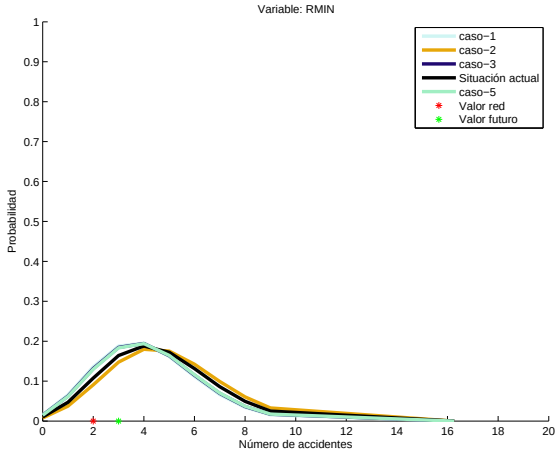

Figura 6. RMIN

4.5. Análisis INCM

El estrato 6 proporciona para la variable INCM (figura 7) una propbabilidad menor para la ocurrencia de sucesos raros.

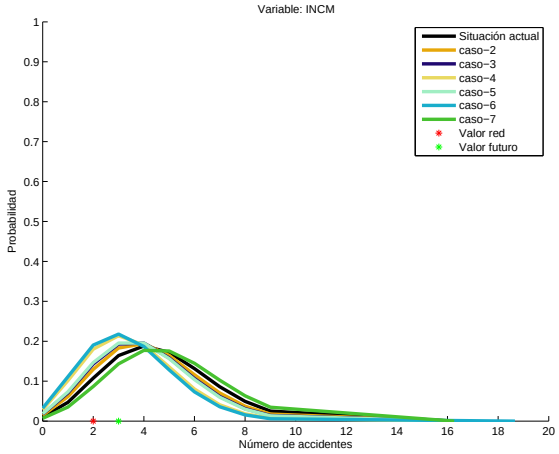

Figura 7. INCM

4.6. Análisis R4VE

Para este caso concreto, la variable R4VE en su estrato número 4 proporciona una distribución de probabilidad en la que los sucesos raros son menos probables.

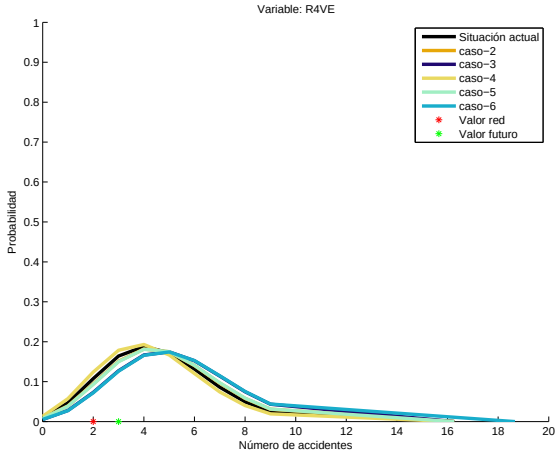

Figura 8. R4VE

5. Conclusiones

Los accidentes de tráfico se han convertido en uno de los problemas de salud pública mas graves. Factores de naturaleza física, técnica, metereológica, deficiencia de la calidad de la red

viaria, aspectos de comportamiento, cognitivos y de formación cívica/vial se ha visto que representan algunas de las posibles causas de accidentes que se registran en la actualidad. Ante esta situación los países han ido diseñando estrategias de tipo preventivo y de investigación donde se intenta detectar qué tipo de variables pueden incidir en el grado de accidentalidad. De esta manera se intenta disminuir el gran coste material que se deriva de este hecho. El modelo Naive-Poisson permite el estudio y análisis de las diferentes alternativas para la disminución de la probabilidad de la frecuencia de accidentes.

Referencias

[1] Gary M. Weiss and Haym Hirsh. Learning to predict rare events in event sequences. In *KDD*, pages 359–363, 1998.

[2] Simon Denis Poisson. *Recherches sur la probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilités*. Bachelier, 1837.

[3] Ladislaus Bortkiewicz. Das gesetz der klienen zahlen (the law of small numbers.). *Leipzig, Germany: Teubner*, 1898.

[4] Micahael Tomz, Gary King, and Langche Zeng. Relogit: Rare events logistic regression. *Journal of statistical software*, 8(i02), 2003.

[5] F. Soler-Flores, J. M. Pardillo Mayora, and R. Jurado Piña. Tratamiento de outliers en los modelos de predicción de accidentes de tráfico. *VIII Congreso de Ingeniería del Transporte, 02/07/2008-04/07/2008, La Coruña, España.*, 2008.

[6] Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. Mining data with rare events: a case study. In *Tools with Artificial Intelligence, 2007. ICTAI 2007. 19th IEEE International Conference on*, volume 2, pages 132–139. IEEE, 2007.

[7] D. E. Holmes, J. Tweedale, and L. C. Jain. Data mining techniques in clustering, association and classification. *Data Mining: Foundations and Intelligent Paradigms*, pages 1–6, 2012.

[8] F. Soler-Flores. Naive-poisson, a mathematical model for road accidents frequency estimation. *Conference of Informatics and Management Sciences*, pages 384–391, 2013.

[9] F. Soler-Flores. Expert system for road accidents frequency estimation based in naive-poisson. *Global Virtual Conference*, page 646, 2013.

[10] Seong-Pyo Cheon, Sungshin Kim, So-Young Lee, and Chong-Bum Lee. Bayesian networks based rare event prediction with sensor data. *Knowledge-Based Systems*, 22(5):336–343, 2009.

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

[12] J. H. Kim and J. Pearl. A computational model for causal and diagnostic reasoning in inference systems. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, pages 190–193. Citeseer, 1983.

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

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

[15] James Dougherty, Ron Kohavi, and Mehran Sahami. Supervised and unsupervised discretization of continuous features. In *ICML*, pages 194–202, 1995.

[16] Cristina Puente Agueda. Causality in sciencie. *Pensamiento Matemático*, (1):12, 2011.

[17] Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Cambridge Univ Press, 2000.

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

[19] Luis Enrique Sucar. Redes bayesianas.

[20] R. O. Duda, P. E. Hart, and D. G. Stork. Pattern classification. 2000. *NY Wiley*, 2000. ID: 129.

[21] P. Langley, W. Iba, and K. Thompson. An analysis of bayesian classifiers. In *Proceedings of the National Conference on Artificial Intelligence*, pages 223–223. JOHN WILEY and SONS LTD, 1992.

[22] Sang-Bum Kim, Hee-Cheol Seo, and Hae-Chang Rim. Poisson naive bayes for text classification with feature weighting. In *Proceedings of the sixth international workshop on Information retrieval with Asian languages-Volume 11*, pages 33–40. Association for Computational Linguistics, 2003.

Sobre los autores:

Nombre: Francisco Soler Flores
Correo Electrónico: f.soler@upm.es
Institución: Departamento de Ingeniería Civil: Ordenación del Territorio, Urbanismo y Medio Ambiente. Universidad Politécnica de Madrid, España.

Nombre: José Ángel Olivas Varela
Correo Electrónico: joseangel.olivas@uclm.es
Institución: Departamento de Tecnologías y Sistemas de Información. Universidad de Castilla La Mancha, España.

Nombre: María Dolores López González
Correo Electrónico: marilo.lopez@upm.es
Institución: Departamento de Matemáticas e Informática Aplicada. Universidad Politécnica de Madrid, España.