Tratamiento de outliers en los modelos de predicción de accidentes de tráfico

Francisco J. Soler Flores
Investigador. Dpto. de Ingeniería Civil:Transportes. Universidad Politécnica de Madrid. España
José María Pardillo Mayora
Profesor Titular. Dpto. de Ingeniería Civil:Transportes. Universidad Politécnica de Madrid. España.
Rafael Jurado Piña
Profesor Titular. (i) Dpto. de Ingeniería Civil:Transportes. Universidad Politécnica de Madrid. España

RESUMEN

El estudio de las relaciones existentes entre la frecuencia de accidentes de tráfico y las características de la carretera, del tráfico, del entorno y de los usuarios, constituye una de las aplicaciones más importantes del análisis estadístico en el campo de la seguridad vial. En el Departamento de Ingeniería Civil: Transportes de la Universidad Politécnica de Madrid en el marco del proyecto de investigación DISCAM, subvencionado por el Centro de Estudios y Experimentación de Obras Publicas en la convocatoria para el año 2006 de ayudas para la realización de proyectos de I+D+i ligados aldesarrollo del PEIT, se está calibrando y depurando un conjunto de estos modelos a partir de muestras obtenidas en las carreteras españolas. En el proceso de ajuste de los modelos se detecta en muchos casos la existencia outliers o valores extremos, lo que afecta a la precisión de los modelos de predicción. En esta comunicación se presentan algunos de los métodos estadísticos existentes para el tratamiento de outliers y se analizan los resultados obtenidos mediante estas metodologías a los modelos de predicción de accidentes calibrados con muestras de tramos de carreteras de la red española.

1.INTRODUCCIÓN

En el marco del proyecto de investigación D.I.S.C.A.M.: *Herramientas para un DIseño Seguro de CArreteras y sus Márgenes*, un proyecto subvencionado por el *Centro de Estudios y Experimentación de Obras Publicas* en la convocatoria para el año 2006 de ayudas para la realización de proyectos de I+D+i ligados al desarrollo del Plan Estratégico de Infraestructuras y Transporte en el marco del Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica 2004-2007, que se está desarrollando en el Departamento de Ingeniería Civil: Transportes, en la Universidad Politécnica de Madrid, se está haciendo recapitulación sobre los modelos utilizados hasta ahora en la predicción de frecuencias de accidentes de tráfico, dentro del Modelo Lineal Generalizado, para intentar mejorar los resultados obtenidos hasta hoy, volviendo a utilizarlos, depurando por completo el estudio estadístico e intentando aplicar modelos que, hasta ahora, no se habían aplicado a datos de vías españolas. Una de las mejoras que se ha obtenido es el mejor ajuste de los modelos mediante el tratamiento de outliers, objeto de esta comunicación.

2.ANTECEDENTES

El estudio de las relaciones existentes entre la frecuencia de accidentes de tráfico y las características de la carretera, del tráfico, el entorno e indefectiblemente el usuario, constituye una de las aplicaciones del análisis estadístico más importantes en el terreno de la seguridad vial y el desarrollo de modelos de predicción de accidentes (Lord y Persaud ,2000).

Los modelos de predicción de accidentes fueron, durante mucho tiempo, desarrollados aplicando técnicas de regresión lineal simple y múltiple pero a finales de los ochenta, principios de los noventa estos fueron se replantearon los modelos. Para el desarrollo de un modelo de predicción, Miaou (1994) sugirió utilizar un modelo de regresión de Poisson y en el caso de que existiese sobredispersión (varianza superior a la media) moderada o alta en los datos se explorase la aplicación de un modelo de regresión binomial negativo o uno de Poisson con ceros aumentados. Además, sugirió que un modelo de regresión de Poisson con exceso de ceros pudiera aplicarse en el caso de que en la frecuencia observada se apreciara una alta cantidad de ceros. La utilización de los modelos de ceros aumentados ha sido recientemente desestimada (Lord, Washington e Ivan, 2007). Por tanto y, según el trabajo realizado por Pardillo J.M. y Llamas R. (2003) y Pardillo J.M., Bojórquez R. y Camarero A. (2006) los métodos de regresión lineal o aquellos que se basan en el concepto de mínimos cuadrados, son poco apropiados para tratar de modelar la relación entre la ocurrencia de accidentes con los factores geométricos del trazado y el tráfico. Nuevos estudios confirman que un modelo de regresión de Poisson o, en el caso de sobredispersión de los datos, el modelo de regresión binomial negativo resultan más apropiados (Soler, Pardillo y Jurado (2007).

3. MODELOS PARA LA PREDICCIÓN DE ACCIDENTES DE TRÁFICO

3.1 El Modelo Lineal Generalizado

Una de las mayores contribuciones en el ámbito de la estadística en los últimos años, ha sido la introducción por parte de Nelder y Wedderburn (1972) de los modelos lineales generalizados. Siendo las dos grandes aportaciones del MLG:

- La introducción del concepto de modelado estadístico como procedimiento general para el análisis de datos.
- La integración del modelado de datos categóricos y cuantitativos en un mismo entorno.

3.1.1 Construcción del modelo

En la construcción del modelo se determinan los siguientes pasos:

- Se evalúan los supuestos del componente aleatorio.
- Se establece la función del componente sistemático.
- Se determina cómo los dos componentes son combinados en el modelo mediante la función de enlace.
- Se estiman los coeficientes del modelo mediante el principio de máxima verosimilitud: maximizando la función de verosimilitud del modelo.
- Se contrasta la bondad de ajuste del modelo.

3.1.2 Modelo de regresión Binomial Negativo

Como hemos citado anteriormente es el modelo de regresión de Poisson el que debería utilizarse para ajustar los datos provenientes del conteo de accidentes de tráfico, pero es conocido que estos suelen presentar sobredispersión . En el departamento de Ingeniería Civil: Transportes se están realizando diversas aplicaciones de estos modelos para la predicción de la frecuencia de accidentes de tráfico a partir de datos de carreteras españolas. Este modelo permite analizar datos de recuento en presencia de sobredispersión (Cameron y Trivedi, 1998).

3.1.3 Bondad de ajuste del modelo

La valoración del poder explicativo del modelo de predicción de accidentes no es ajena a los procesos metodológicos desarrollados en la implantación de los modelos de regresión, por lo que este poder se valora a partir de contrastes de bondad de ajuste.

Tratando de proponer alternativas a los contrastes tradicionales de *R2* en los modelos de predicción de accidentes, Fridstrom (1995) propuso un conjunto de contrastes de bondad de ajuste para tal propósito. De este conjunto en el trabajo de Vogt (1999), destacan la Desviación. Así mismo en diversos trabajos también se ha aplicado el criterio de información de Akaike *(AIC)*, que se hace destacar por la oportunidad de su aplicación independientemente de las hipótesis en la estructura del modelo de regresión, así como el contraste χ^2 a nivel de análisis de residuos con datos puntuales como también a nivel de distribuciones de frecuencias en datos agrupados.

Desviación (*D*): Medida de la bondad del ajuste del modelo respecto a los datos. Cuanto menor sea el valor, mejor será el ajuste.

R^2 **de máxima verosimilitud:** También es una medida análoga al R^2 de los modelos lineales y fue propuesta por Maddala (1983). Cuánto más próxima a 1 mejor será el modelo.

Cuando se tiene una serie de modelos M1, M2,... con parámetros K1, K2,…, respectivamente, una metodología para compararlos corresponde a la función de máxima verosimilitud. La máxima verosimilitud permite seleccionar el modelo que realiza el mejor ajuste de los datos pero no penaliza su complejidad, lo que si sucede cuando se emplean medidas de contraste como el AIC y el BIC.

Criterio de información de Akaike (AIC): Los mejores modelos son aquellos que presentaron el menor valor de AIC. Cuando los valores de AIC están muy cercanos, la elección del mejor modelo se puede realizar con base en el cálculo los pesos de Akaike (Burnham y Anderson, 1998), que se calculan a partir de los coeficientes de Akaike.
Criterio de información bayesiano (BIC): El BIC (Schwarz, 1978) es calculado para los diferentes modelos como una función de la bondad de ajuste del logaritmo de verosimilitud, el número de parámetros ajustados y el número total de datos. El modelo con el más bajo valor de BIC es considerado el mejor en explicar los datos con el mínimo número de parámetros.

Gráfico de residuos acumulados:. En resumen el procedimiento descrito por de Vogt (1999) y Hauer (2004) se valora de la siguiente forma:

- Sí el ajuste es adecuado el diagrama de residuos acumulados deberá “oscilar” alrededor del eje de las abscisas y con una escasa amplitud.

- Sí el ajuste es adecuado el diagrama de residuos cerrará muy próximo a cero (en el punto del valor mayor de la variable explicativa).

3.2 Outliers

Un *Outlier* se define como “aquella observación (o conjunto de observaciones) inconsistentes con el resto del conjunto de datos” (Barnet y Lewis ,1994) . Es decir, aquella observación que siendo atípica y/o errónea, tiene un comportamiento muy diferente con respecto al resto de los datos frente al análisis que se desea realizar sobre las observaciones experimentales .

En estas definiciones conviene observar las siguientes peculiaridades:

- Las observaciones atípicas y erróneas exigen que los errores o variabilidades sean grandes
- Los outliers no consideran todas las observaciones atípicas o erróneas, sino aquellas que tienen un comportamiento muy diferente respecto al resto de los datos.

Ello viene motivado porque las técnicas o procedimientos para determinar o corregir este tipo de observaciones sólo tienen sentido en estas situaciones, ya que aquellas observaciones que no tienen un gran error o que se comportan como la mayoría, no van a afectar de forma determinante a las conclusiones que se realicen a partir de las mismas.

3.2.1 Detección de Outliers

Algunos de los métodos clásicos utilizados para la identificación de outliers son:

-Método basado en el recorrido intercuartílico.
-Diagrama Box&Whisker (Cajas y bigotes).
-Método basado en la mediana de las desviaciones absolutas (MEDA).

En esta comunicación hemos basado nuestro procedimiento de detección de outliers en el utilizado habitualmente para la detección de tramos de concentración de accidentes, basado en la desigualdad de Tchebychev con un tratamiento estratificado descrito a continuación.

3.2.2 Método basado en la desviación típica.

Este método se basa en la desigualdad de Tchebychev ($P(|X-\mu|>k\sigma)\leq\frac{1}{k^2}$) y viene dado por los registros pertenecientes al complementario del siguiente conjunto: $\left\{x_i :|x_i-\bar{x}|<kS\right\}$ que verifiquen que $f\left(x_i\right)\geq 1-\frac{1}{k^2}$ de donde se deduce que en el intervalo $(\bar{x}-kS,\bar{x}+kS)$ se encuentran al menos el $100(1-\frac{1}{k^2})\%$ de las observaciones y si k es tal que $1-\frac{1}{k^2}$ es próximo

a 1, las observaciones fuera de $(\overline{x} - kS, \overline{x} + kS)$ pueden ser considerados como outliers.

4.METODOLOGÍA

Partiendo del modelo resultante construido en investigaciones anteriores (Pardillo, 2008) (2):

$$ACC = e^{(-0.3937136 + 0.8700814 \cdot LNEXPO + 0.265021 \cdot NINTER - 0.1050077 \cdot VISMIN - 0.005216 \cdot VEL\_MIN - 0.0249549 \cdot INCMAX + 0.0069947 \cdot SPDVEL)} \quad (2)$$

Donde la variable dependiente a predecir es la f*recuencia estimada de accidentes con víctimas=ACC* y las variables independientes con las que se relaciona son la *exposición al riesgo=expo*, la *visibilidad mínima=vismin*, la *velocidad mínima=vel_min, el número de intersecciones en el tramo de 1km=ninter, la inclinación máxima=incmax* y *la desviación de la velocidad específica=spdvel.*

El primero de los pasos ha sido el obtener los outliers de la muestra para la variable IP (Índice de peligrosidad ($IP = \frac{N^o\ Accidentes\_con\_víctimas}{10^8\ veh - km}$, Pardillo (2004)) que es la variable tenida en cuenta habitualmente en los modelos de detección de tramos de concentración de accidentes, utilizando el método basado en la desviación diferenciado por intervalos de intensidades y para los outliers en la variable IP.

- Se han estratificado las muestras por intensidades (0-1000, 1000-2000 ,2000-3000, 3000-4000, 4000-5000, 6000-7000, 7000-15000 y 15000-20000).

- Se han eliminado los outliers para la variable índice de peligrosidad a diferentes niveles de significación (85, 90, 95 y 99) para cada estrato.

- Se estudia el modelo utilizando el modelo de regresión binomial negativo

En nuestro caso seleccionaremos como mejor modelo aquel que sea más parsimonioso (es decir, el modelo que mejor explica la variación de los datos y que requiere el menor número de parámetros) o lo que es lo mismo, el modelo que presente menor valor del criterio de información de Akaike (AIC).

5. RESULTADOS

Aplicado el método basado en la desviación típica para obtener los outliers y eliminados para su posterior tratamiento obtenemos los siguientes resultados y este gráfico, que refleja los outliers detectados en la muestra a un nivel de significación del 85% por el procedimiento estratificado. Figura 1:

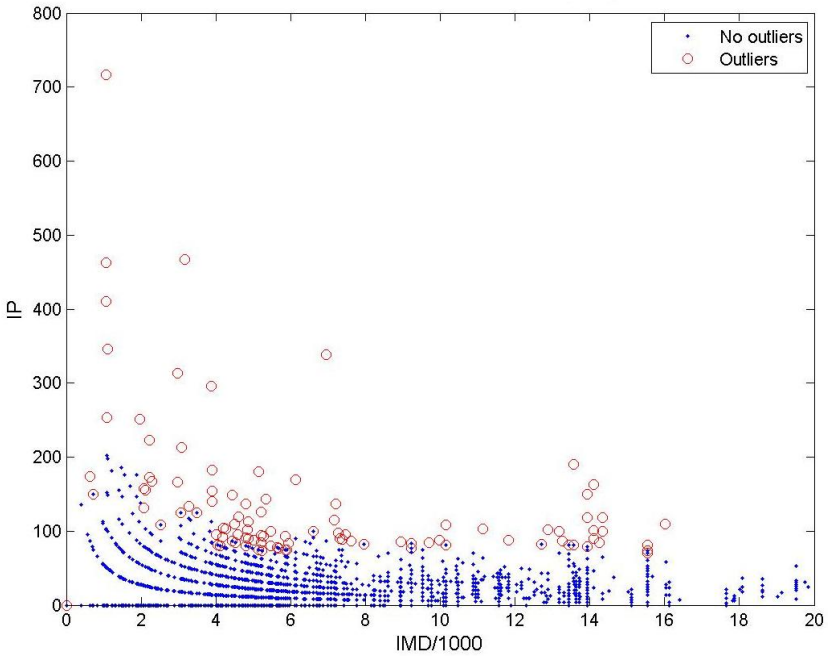


Figura 1: Gráfico outliers a NC: 85%

Aplicado el modelo resultante de estudios anteriores (Pardillo, 2008), partimos de las variables recogidas en el primer registro de la siguiente tabla que también incluye los parámetros de bondad de ajuste para los mejores modelos construidos, a un nivel de confianza del 95% para los diferentes conjuntos de datos. Tabla 1.

MODELO	NOUT	OUT	DESVIACIÓN	R2 M.L.	BIC	AIC	PESO AIC
Modelo incial	3297.00	0.00	12848.14	0.32	-13795.29	3.90	0.16
Modelo_80%	3186.00	111.00	11496.74	0.33	-14146.73	3.61	0.18
Modelo_85%	3216.00	81.00	11794.73	0.34	-14120.82	3.67	0.17
Modelo_90%	3247.00	50.00	12115.58	0.33	-14081.39	3.74	0.17
Modelo_95%	3276.00	21.00	12471.34	0.33	-13989.19	3.81	0.16
Modelo_99%	3295.00	2.00	12797.29	0.32	-13827.94	3.89	0.16

Tabla 1: Ajustes de los modelos con todos los criterios de bondad de ajuste y nº outliers

Así vemos que con todos los criterios de bondad de ajuste el modelo mejora (incluyendo incluso, en algún caso, una variable menos) al observar que las cuatro medidas de bondad de ajuste mejoran en todos los casos al modelo que trata todos los datos. Además, debido a que la diferencia entre los diferentes coeficientes de Akaike es baja, al estudiar el peso de Akaike confirmamos la elección, por ser el mayor de todos.

Por otro lado, el gráfico de residuos acumulado mejora el modelo seleccionado previamente sin tratamiento de outliers. Figura 2:

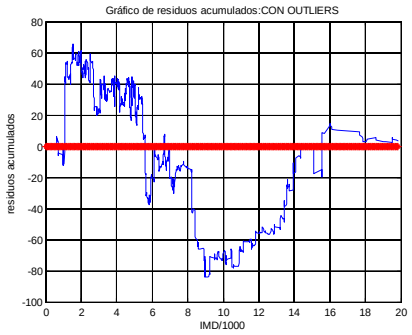
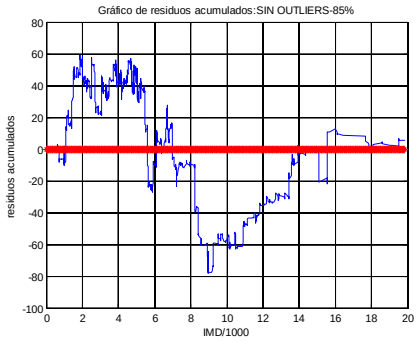
MODELO INICIAL	MODELO SIN OUTLIERS AL 85%
	

Figura 2: Gráficos Residuos Acumulados de los modelos inicial y seleccionado.

El modelo seleccionado por tener los mejores coeficientes de bondad de ajuste descritos es el Modelo_85% (modelo construido eliminando los outliers a un nivel de confianza del 85%) (3):

$$ACC = e^{(-0.9361179 \; + 0.9154474 \cdot LNEXPO + 0.1910163 \cdot NINTER - 0.0836082 \cdot VISMIN - 0.0025976 \cdot VEL\_MIN + 0.0051495 \cdot SPDVEL)} \quad (3)$$

6. CONCLUSIÓN

Al obtener una mejora en los coeficientes de bondad de ajuste y por lo tanto en los modelos, podemos concluir que el modelo mejora en los casos en los que se eliminan los ouliers, por lo que los TCAs podrían considerarse casos excepcionales que no recoge el modelo de regresión de partida utilizado para la predicción de accidentes de tráfico.

7. BIBLIOGRAFÍA

BARNETT, V. y Lewis, T. (1994). *Outliers in Statical Data*. John Wiley & Sons. New York.

BURNHAM, K. P. y ANDERSON D. R. (1998). *Model selection and inference: a practical information-theoretic approach*. Springer-Verlag, New York.

CAMERON, A. C. y TRIVEDI, P. K. (1998). *Regression* Analysis of Count Data. *Econometric Society Monographs, 30. Cambridge: Cambridge University Press*

DOMINIQUE, L., WASHINGTON, S. e IVAN, J. (2007). Further notes on the application of zero-inflated models in highway Safety. *Accident Analysis and Prevention, pp. 53-482,Volume 57.*

FRIDSTROM, L., IFVER, J., INGEBRIGTSEN, S., KULMALA, R. y THOMSEN, L. (1995). Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts. *Accident Analysis and Prevention*, pág. 1-20, Volume 27, Issue 1.

HAUER, E. (2004) *Statistical Road Safety Modeling*. *83rd Annual Meeting of the Transportation Research Board, Washington D.C., EEUU.*

LORD, D. y PERSAUD, B. (2000). Accident prediction models with and without trend: application of the generalized estimating equations procedure. *Transportation Research Record 1717* , Washington, D.C.

MADDALA, G. S. (1983): Limited-dependent and qualitative variables in econometrics. *Cambridge University Press*

MIAOU (1994).The relationship between truck accidents and geometric design of road sections:Poisson versus negative binomial regressions. *Accident Analysis and Prevention, pp. 471-482,Volume 26, Issue 4*

NELDER , J.A. y WEDDERBURN, W.M. (1972) Generalized linear models. *Journal of the Royal Statistical Society-Series A*, pp. 370-384.

PARDILLO, J.M. y LLAMAS, R. (2003). *Relevant Variables for Crash Rate Prediction in Spain´s Two Lane Rural Roads*. 82nd Annual Meeting of the Transportation Research Board, Washington

PARDILLO, J.M. (2004). Procedimientos de estudio, diseño y gestión de medidas de seguridad vial en las infraestructuras. Fundación Agustin de Betancourt. Madrid.

PARDILLO, J.M., BOJÓRQUEZ, R. y CAMARERO, A. . (2006): Refinement of Accident Prediction Models for Spanish National Network. *Transportation Research Record Journal of the Transportation Research Board nº 1950, pp. 65-72*. Washington D.C.

PARDILLO, J.M. (2008) Calibración de modelos para el análisis de la seguridad de la carretera y sus márgenes en el proyecto DISCAM. *I Jornadas de Presentación de Proyectos I+D de Transportes del Plan Nacional 2004-2007 ligados al PEIT*. Centro de Estudios y Experimentación de Obras Públicas. Madrid.

SCHWARZ, G. (1978) Estimating the Dimension of a Model. *The Annals of Statistics*, Vol. 6, No. 2. pp. 461-464.

SOLER F.J., PARDILLO J.M. y JURADO R. (2007). El Modelo Lineal Generalizado aplicado a la predicción de accidentes de tráfico. *I Congreso Internacional de Matemáticas en Ingeniería y Arquitectura.30 de Mayo-1 de Junio 2007.* Madrid.

VOGT, A. (1999) *Crash models for rural intersections: four-lane by two-lane stop-controlled and two-lane by two-lane signalized*. Federal Highway Administration, Report No. FHWA-RD-99-128. McLean,Virginia, EEUU.