¿CÓMO MODELIZAR DATOS CON EXCESO DE CEROS? MÉTODOS Y APLICACIÓN A LA INVESTIGACIÓN FORESTAL

Rafael Calama Sainz [1], Rubén Manso González [1] y Jose A. Tomé [2]

[1] Dpto. Selvicultura y Gestión Forestal. CIFOR-INIA. Ctra. A Coruña, km 7,5. 28040 – MADRID (España)
[2] Dpto. de Engenharia Florestal. Instituto Superior de Agronomia. Tapada da Ajuda. 1349– 017 LISBOA (Portugal)

Resumen

Los datos con exceso de ceros son muy frecuentes en el ámbito de la ecología y la ciencia forestal. Este tipo de datos presentan una serie de particularidades que impiden la aplicación de las técnicas estadísticas clásicas basadas en la asunción de normalidad, como es la regresión por mínimos cuadrados. En el trabajo se describen los problemas asociados a este tipo de datos, se muestran las posibles alternativas para la modelización de los mismos y el desarrollo y aplicación de las técnicas propuestas a través de su programación en el paquete estadístico SAS®, y se presenta un caso de estudio en el ámbito de la modelización forestal.

Palabras clave: *Modelos cero-inflados, Binomial negativa, Poisson, Modelos hurdle*

INTRODUCCIÓN: PLANTEAMIENTO DEL PROBLEMA

En el ámbito de la modelización en ecología, agronomía y ciencia forestal es habitual trabajar con variables truncadas en los valores negativos, entendiendo por éstas aquellas variables que sólo pueden adoptar valores cero y positivos. Este tipo de variables pueden ser de tipo discreto o continuo. En el caso de variables discretas, el caso más común es el de los conteos. Ejemplo de trabajos de modelización ecológica o forestal con variables discretas de conteo son el estudio del número de pies muertos en un rodal (Affleck, 2006), el número de agentes patógenos detectado por cm^2 de hoja (Hall et *al.*, 1997), el número de frutos cosechados por planta y año (Calama & Montero, 2007) o el número de plántulas emergidas en una superficie dada (Fortin & Deblois, 2007). Dentro de variables continuas, las más habituales en modelización forestal son aquéllas que sólo adoptan valores positivos, no estando definidas o careciendo de sentido para el valor cero. Es el caso de todas las referidas a dimensiones del árbol o del rodal, tales como diámetro normal, altura, área basimétrica, volumen, índices de espesura, crecimientos... Sin embargo, existen variables de tipo continuo que pueden presentar observaciones con valor cero, como serían las hectáreas afectadas por incendio en distintas unidades de gestión, el total de biomasa reproductora producida por una planta en un año determinado, el porcentaje de acículas de la planta afectadas por clorosis…

En las variables truncadas en valores negativos, un problema habitual es la presencia de un exceso o sobreabundancia de observaciones con valor cero (Hall et *al.*, 1997). Esta situación conlleva una serie de limitaciones a la aplicación de técnicas estadísticas tradicionales, como puede

ISSN: 1575-2410

ser la regresión lineal por mínimos cuadrados ordinarios o generalizados. La presencia de un exceso de ceros implica un severo alejamiento de la distribución normal, presentando habitualmente las observaciones una distribución de frecuencias con una única moda en cero y una larga cola hacia la derecha, o una moda en cero y una segunda moda en un valor positivo (LAMBERT, 1992; TOOZE et *al.*, 2002; AFFLECK, 2006).

La Figura 1 muestra un ejemplo simulado de variable de conteo con una distribución con exceso de ceros, donde sobre un total de 1.080 observaciones 669 (53,09% del total) tienen valor cero. En esta muestra –al contrario que en la distribución normal– el valor medio (4,71) está desplazado respecto de la moda, presentando además una larga cola hacia la derecha, lo que implica que en las distribuciones con abundancia de ceros, un mismo valor de la varianza implica la presencia de observaciones con valores extremos. La aplicación de métodos basados en asunción de normalidad – como la regresión por mínimos cuadrados ordinarios – sobre esta muestra conllevaría serios problemas de sesgo en la estimación de los parámetros y sus errores estándar, asociados al incumplimiento de los supuestos de normalidad y homocedasticidad en la varianza de los residuos. Asimismo implicaría una subestimación de la incidencia de los ceros, la posible predicción de valores negativos, una sobreestimación de los valores positivos cercanos a cero y la no estimación de valores positivos extremos (LI et *al.*, 2008), lo que supondría que la distribución de valores esperados no se asemejaría a la de los valores observados. En este sentido, la Figura 2a muestra el histograma de frecuencias esperado para una distribución normal con la misma media y varianza que la distribución simulada de estudio.

La transformación de los datos mediante funciones logarítmicas u otras se ha propuesto como método de corrección habitual para trabajar en situaciones de desviación de la normalidad o heterocedasticidad (PEÑA, 1989). Sin embargo, la aplicación de esta técnica a estos datos con abundancia de ceros no resulta adecuada (Figura 2b), puesto que en primer lugar requiere de la adición de una constante k en el caso de la transformación logarítmica, para evitar problemas de definición, lo que implica que la distribución de frecuencias presentará una moda en el valor de $\log(k)$, nuevamente desplazada respecto de la media (LI et *al.*, 2008). En cualquier caso, sí que se consigue corregir en cierta medida la larga cola hacia la derecha y presencia de valores extremos, pudiéndose recomendar este método para aquellos datos muestrales en los que la presencia de ceros se cifre en un 20% como máximo (p. ej. CALAMA & MONTERO, 2007). Otra alternativa propuesta en el caso de distribuciones alejadas de la normalidad ha sido el empleo de métodos no-paramétricos de rangos, no válidos para el caso de exceso de ceros, debido al elevado número de empates en las diferencias de rango entre las observaciones cero y a la imposibilidad de obtener predicciones de la variable de respuesta (TOOZE et *al.*, 2002)

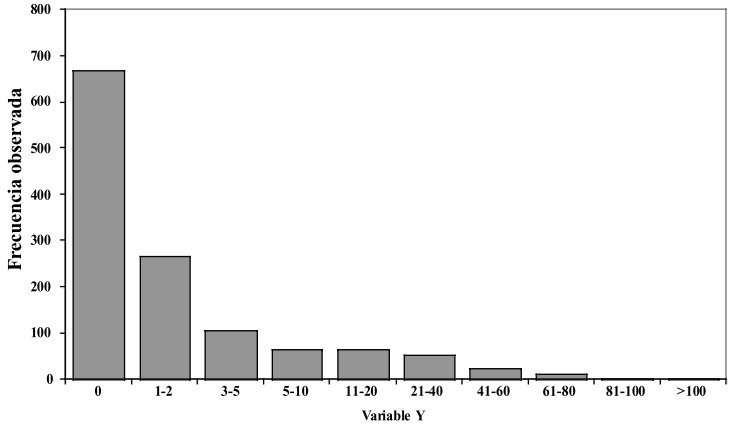

***Figura 1**. Histograma de frecuencias observadas del caso de estudio simulado: media (4,71), varianza (152,6) y 53,1% de observaciones con valor cero*

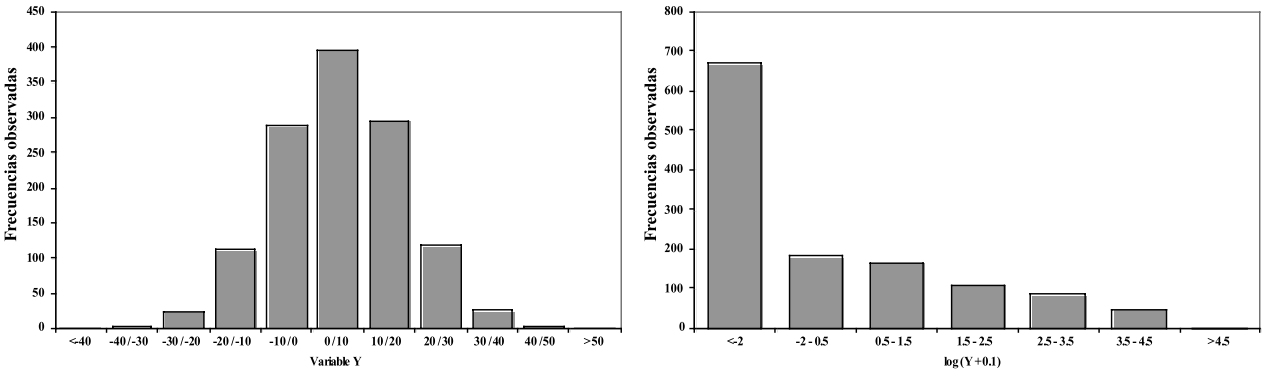

***Figura 2**. (a) Histograma de frecuencias esperadas para una distribución normal de media (4,71), varianza (152,6) (b) Histograma de frecuencias observadas tras aplicar a la distribución original la transformación log (y+0,1)*

En el presente trabajo se hace una revisión de los distintos métodos existentes para la modelización de datos inflados en cero, presentando el desarrollo teórico de la metodología asociada a la resolución de este tipo de modelos, y desarrollando el ajuste y resolución práctica de los mismos mediante el procedimiento NLMIXED del paquete estadístico SAS-STAT ®. El trabajo se complementa con la presentación de un caso práctico de modelización de una variable de interés forestal que presenta exceso de ceros y una distribución continua para la abundancia, como es el peso anual de fruto producido en árboles de *Pinus pinea* L.

MÉTODOS ESTADÍSTICOS PARA LA RESOLUCIÓN DEL PROBLEMA

El análisis y desarrollo de modelos que utilicen como variables de respuesta variables infladas en cero se ha realizado identificando en primer lugar a qué modelo de distribución teórica se asemejan los datos, procediendo a continuación a establecer una relación lineal mediante una función *link* entre alguno de los parámetros caracterizadores de la distribución (habitualmente la media) y una o varias variables explicativas de tipo continuo o categórico, a través de una serie de parámetros que serán estimados mediante métodos de máxima verosimilitud (Affleck, 2006). En este sentido, la resolución de este tipo de modelos puede considerarse un caso particular de los modelos lineales generalizados (McCullagh & Nelder, 1989; Tu, 2002).

Un aspecto adicional a considerar como paso previo a la modelización es la naturaleza de los ceros en los datos observados (Ridout et *al.*, 1998; Kuhnert et *al.*, 2005). La presencia de un cero en un sujeto muestreado puede deberse a que en la población existe una probabilidad de ocurrencia del valor cero, y que ese sujeto tiene asignado ese valor cero al azar (dando lugar al denominado *cero aleatorio*), o bien a que ese sujeto pertenece a una subpoblación cuyos individuos siempre tienen valor cero (denominados *ceros estructurales*). Un caso particular de cero aleatorio lo constituyen los *falsos ceros*, cuando el valor cero está asociado a un error del muestreo o del observador. Un ejemplo sería la presencia de frutos en plantas de una determinada especie. Si los frutos han sido comidos, o el muestreo se realiza en una época no adecuada, podríamos encontrar *falsos ceros*. Desde el punto de vista estadístico, su tratamiento sería similar al de aquellos ceros que debido a la variabilidad natural del proceso, podríamos encontrar en una muestra de plantas cuyas características no difieran de la del resto de individuos de la muestra, y que consideraríamos como *cero aleatorio*. Por otra parte, podría haber un grupo o subconjunto de plantas que, por alguna característica determinada (p. ej. edad o tamaño), no produjesen fruto, correspondiendo este caso a un *cero estructural*. Según las variables de estudio es posible identificar la presencia dentro de una muestra de uno o varios tipos de ceros, lo que condicionará el tipo o familia de distribución a seleccionar.

La distribución más utilizada en el caso de datos de conteo es la de Poisson (Johnson et *al.*,

2005). La distribución de Poisson es una distribución univariante para datos discretos caracterizada por un único parámetro λ, que define tanto la media y la varianza de la distribución, y cuya función de densidad para un valor y = Y_i es:

$$f(Y_i \mid \lambda) = \frac{e^{-\lambda}\lambda^{Y_i}}{Y_i!} \qquad [1]$$

Donde λ puede expresarse como un modelo lineal sobre distintas variables explicativas xi a través de una función link logarítmica, definiendo la regresión de Poisson:

$$\log(\lambda) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots = \mathbf{\beta x} \qquad [2]$$

El hecho de que en la distribución de Poisson la media y la varianza coincidan hace que no sea útil para caracterizar la distribuciones con un alto porcentaje de ceros y valores positivos extremos, que presentan habitualmente sobredispersión de la varianza (RIDOUT et *al.*, 1998, FORTIN & DEBLOIS, 2007. La Figura 3a muestra el histograma de frecuencias esperadas para una distribución de Poisson de parámetro λ = 4,71 junto al histograma de los datos observados en el ejemplo propuesto, observándose como la distribución Poisson subestima tanto el porcentaje observado de ceros como los valores de la larga cola hacia la derecha. En este sentido, la distribución de Poisson podrá considerarse válida únicamente para aquellos conjuntos de datos en los que la media y la varianza tengan valores próximos, y el número de ceros no sea mucho mayor que es esperado de acuerdo a una Poisson.

La primera alternativa a la distribución de Poisson lo constituyen aquellas distribuciones que permiten una mayor dispersión de las observaciones, como la distribución binomial negativa (JOHNSON et *al.*, 2005). Ésta es una distribución univariante discreta obtenida a partir de la distribución de Poisson en la que el parámetro λ se distribuye de acuerdo a una distribución gamma, quedando la distribución final caracterizada por dos parámetros μ y k, lo que la dota de mayor flexibilidad que a la distribución Poisson. La función de densidad asociada a la binomial negativa es:

$$f(Y_i \mid \mu, k) = \frac{\Gamma\left(Y_i + k^{-1}\right)}{\Gamma\left(k^{-1}\right) Y_i!}\left(\frac{k\mu}{k\mu+1}\right)^{Y_i}\left(\frac{1}{k\mu+1}\right)^{1/k} \qquad [3]$$

Donde Γ representa la función gamma, Ì la media de la distribución, y k el parámetro de sobredispersión de la varianza, que queda definida como $var(y) = \mu + k\mu^2$. La parametrización de μ en función de variables explicativas xi mediante una función link logarítmica, análoga a la ecuación 2, daría lugar al modelo de *regresión binomial negativa*¸ en el que los parámetros β_i se estimaría de forma conjunta al parámetro k. La distribución binomial negativa permite considerar con bastante precisión elevados porcentajes de cero en la muestra y la sobredispersión de la varianza (ver Figura 3b), habiéndose utilizado con asiduidad en modelización ecológica (BLISS & FISHER, 1953; HALL et *al.*, 1997). Las limitaciones al uso de esta distribución en modelización de muestras con exceso de ceros son la necesidad de aplicarse a datos discretos (tipo conteo), la imposibilidad de considerar ceros estructurales en la población y la subestimación en el número de ceros predichos en el caso de muestras con un porcentaje de ceros muy elevado (> 70%).

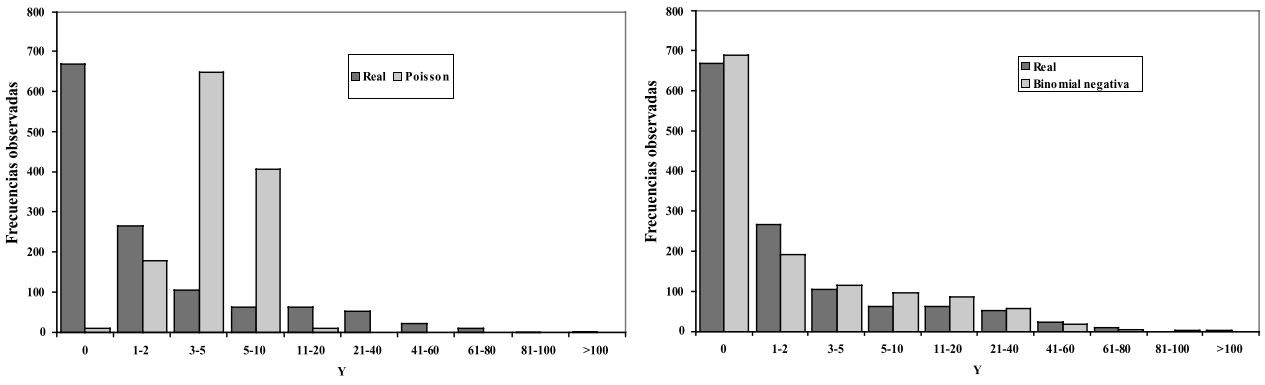

***Figura 3**. Comparación entre la distribución de frecuencias real observada y (a) distribución de frecuencias esperadas para una distribución Poisson de media y varianza (4,71) (b) Distribución binomial negativa de media (4,71) y parámetro de sobredipsersión k (5,44)*

Las distribuciones de Poisson y binomial negativa, y otras derivadas como la Poisson generalizada o la Neyman tipo A, entrarían dentro de la categoría de distribuciones univariantes discretas, considerándose las observaciones de la muestra realizaciones dentro de estas distribuciones. Una alternativa para el manejo de datos con abundancia de ceros se fundamenta en considerar el proceso en estudio como el resultado de dos fenómenos diferenciados: uno definiendo la ocurrencia o no del evento, y otro definiendo la extensión del mismo condicional a la ocurrencia. A partir de esta consideración es posible desarrollar modelos independientes para el ajuste de los fenómenos de ocurrencia-no ocurrencia (asumiendo una distribución binomial y regresión logística) y de abundancia (a través de un modelo estándar de regresión sobre los datos estrictamente mayores que cero), dando lugar a los *modelos condicionales en dos partes* (WELSH et *al.*, 1996; WOOLLONS, 1998).

Una opción más adecuada se basa en asumir que los datos provienen de una distribución conjunta, mezcla de una distribución binomial para describir la ocurrencia o no del evento en estudio, y cualquier otra distribución discreta o continua (Poisson, Poisson truncada, Binomial Negativa, Log-normal…) para caracterizar la abundancia del evento condicional a la ocurrencia del mismo. Este tipo de distribuciones conjuntas se conocen como distribuciones cero-infladas (*zero-inflated*, LAMBERT, 1992) asociadas a la distribución condicional seleccionada (p. ej. *zero-inflated Poisson*: distribución cero-inflada Poisson), y permiten considerar los dos procesos de ocurrencia/abundancia a través de una única distribución. Una ventaja adicional es que permite asumir distribuciones bimodales, con modas en cero y otro valor intermedio en la distribución. Por último, su formulación permite además estimar de forma simultánea, mediante técnicas de máxima verosimilitud, los parámetros que permita predecir la probabilidad de ocurrencia a partir de una regresión logística, y la abundancia del evento, condicional a su ocurrencia, a partir de un modelo de regresión definido por el modelo de distribución condicional.

En el caso de que la distribución condicional no sea truncada en cero (ej. Poisson), este tipo de modelos permitiría considerar la presencia conjunta de ceros estructurales y aleatorios (HALL, 2000). La función de densidad asociada sería:

$$g\left(Y_i \mid \pi, f\right)=\begin{cases}\pi+(1-\pi)\,f(Y_i=0) & \text{si } Y_i=0\\ (1-\pi)\,f(Y_i>0) & \text{si } Y_i>0\end{cases} \quad [4]$$

Donde f representa la función de densidad de una distribución conocida no truncada en cero (ecuación 1 para el caso de una distribución condicional Poisson, ecuación 3 para una binomial negativa); π es el parámetro caracterizador de la distribución binomial (0,1), que define la probabilidad de no ocurrencia del evento (ó 1 – probabilidad de ocurrencia) asociada a ceros de tipo estructural, mientras que (1 - π) $f(Y_i=0)$ indiciaría la probabilidad de ocurrencia de ceros aleatorios. La regresión cero-inflada se basa en relacionar tanto el parámetro π como el parámetro o parámetros caracterizadores de la distribución condicional asociada con una serie de variables explicativas mediante una función link lineal. Para el caso del parámetro de probabilidad, esta función será la logit:

$$\text{logit}\left(\pi\right)=\log\left(\frac{\pi}{1-\pi}\right)=\mathbf{x}\beta \quad [5]$$

Mientras que para el caso de la distribución condicional el parámetro a relacionar será habitualmente la media, que se expandirá de forma lineal mediante una función link logarítmica (en el caso más habitual de los modelos cero-inflado Poisson ZIP y cero-inflado Binomial Negativa ZINB, sería equivalente a la ecuación 2). Las variables explicativas a incluir en el modelo de regresión logística para ocurrencia y el modelo para abundancia pueden ser o no las mismas (FORTIN & DEBLOIS, 2007).

La Figura 4 presenta el resultado del ajuste de una distribución cero-inflada Poisson a los datos de estudio. Aunque el parámetro adicional π permite caracterizar la probabilidad de ocurrencia de ceros, el uso de la distribución Poisson para definir la abundancia adolece de los mismos problemas identificados anteriormente, al subestimar la ocurrencia de valores extremos.

La distribución cero-inflada binomial negativa (ecuación 6) representa una alternativa al caso anterior, ya que al contar con tres parámetros a estimar dotaría de mayor flexibilidad al modelo.

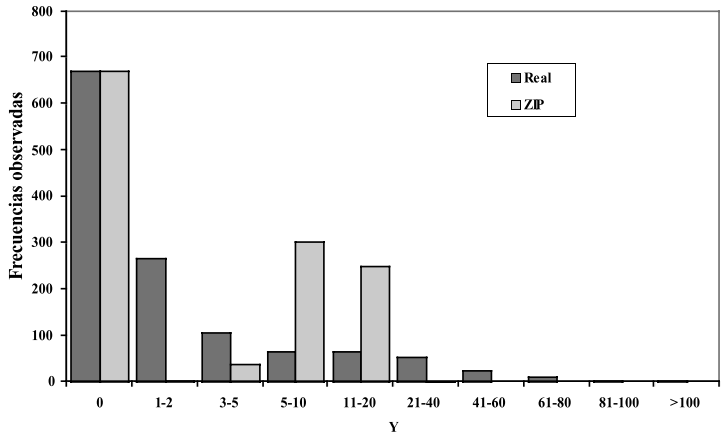

Figura 4. *Comparación entre la distribución de frecuencias real observada y distribución de frecuencias esperadas para una distribución Cero inflada - Poisson de parámetros π = 0,5309 y media y varianza μ (4,71)*

$$g(Y_i/\pi,k,\mu)=\begin{cases}\pi+(1-\pi)\left(\dfrac{1}{k\mu+1}\right)^{1/k} & \text{si } Y_i=0\\ (1-\pi)\dfrac{\Gamma\left(Y_i+k^{-1}\right)}{\Gamma\left(k^{-1}\right)Y_i!}\left(\dfrac{k\mu}{k\mu+1}\right)^{Y_i}\left(\dfrac{1}{k\mu+1}\right)^{1/k} & \text{si } Y_i>0\end{cases} \quad [6]$$

A pesar de las ventajas del modelo ZINB, en el caso particular de estudio el ajuste no supone una mejora respecto al ajuste de la distribución binomial negativa (resultado no presentado), con un valor estimado para el parámetro π = $8{,}089\cdot10^{-7}$, lo que indica que la mayor parte de los ceros pueden quedar explicados mediante la distribución binomial negativa estándar. Esto puede deberse al hecho de que el exceso de cero no es extremo, siendo inferior al 70%.

Una alternativa a los modelos cero inflados son los modelos de salto o modelos *hurdle* (MULLAHY, 1986). Cuando no hay diferenciación entre ceros aleatorios y estructurales, el fenómeno quedaría definido mediante dos procesos estadísticos diferenciados, uno que genera los ceros (definido por una binomial) y otro independiente que genera los valores de abundancia, caracterizado por una distribución truncada en cero (p. ej. Poisson o binomial negativa truncadas). La denominación de *modelos de salto* pretende captar la idea de que cualquiera que sea el mecanismo que condiciona la respuesta positiva, debe superarse un determinado valor frontera antes de obtener resultados no nulos. En estos casos, la función de densidad para una distribución conjunta entre una binomial y una distribución truncada *h* es:

$$g(Y_i\mid\pi,h)=\begin{cases}\pi & \text{si } Y_i=0\\ (1-\pi)\dfrac{h(Y_i>0)}{h(Y_i=0)} & \text{si } Y_i>0\end{cases} \quad [7]$$

La diferencia entre los modelos *cero-inflados* y los modelos hurdle no se detecta en la fase inicial de definición de la distribución teórica, sino en el ajuste de los modelos lineales que permitan definir los parámetros caracterizadores en función de covariables explicativas, por lo que no se presentan resultados de comparación sobre la distribución simulada de estudio.

Las distribuciones analizadas hasta el momento tienen validez para muestras de variables de tipo discreto. En el caso de variables de tipo continuo, el número de alternativas es mucho menor, siendo la distribución conjunto más ampliamente utilizada la *distribución delta* de AITCHINSON (1955), también conocida como cero-inflada log-normal (ZILN, TOOZE et *al.*, 2002). En esta distribución la probabilidad de ocurrencia o no del evento viene definida mediante una distribución binomial de parámetro π, mientras que la abundancia condicional al evento se caracteriza mediante una distribución log-normal, que es truncada en cero por naturaleza (definiendo un modelo de salto). La función de densidad asociada a esta distribución es:

$$g(Y_i\mid\pi,\mu,\sigma)\begin{cases}\pi\\ (1-\pi)\dfrac{1}{Y_i\sigma\sqrt{2\pi}}e^{-(\ln Y_i-\pi)^2/2\sigma^2}\end{cases} \quad [8]$$

Esta distribución queda caracterizada por tres parámetros: π, el caracterizador de la bino-

mial (que indica probabilidad de ocurrencia de cero), y los parámetros media (μ) y desviación típica (σ) caracterizadores de la distribución log-normal; lo que dota a esta función de gran flexibilidad. Sin embargo, su uso ha quedado muy restringido, habiendo sido aplicada en estudio médicos (Tooze et *al.*, 2002; Li et *al.*, 2008) y recientemente en el trabajo de Calama et *al.* (2011) sobre producción de fruto de *Pinus pinea*, que se desarrollará a continuación.

AJUSTE DE LOS MODELOS, ESTIMACIÓN DE LOS PARÁMETROS Y ESTADÍSTICOS DE COMPARACIÓN

Tanto el ajuste de los modelos de distribución básicos (sin expansión de los parámetros caracterizadores) como de los modelos expandidos incluyendo covariables se realiza mediante métodos de máxima verosimilitud. Este método se base en identificar aquellos parámetros que maximizan la probabilidad de ocurrencia de las observaciones para una distribución dada. Habitualmente se maximiza el logaritmo de la función de verosimilitud. La Tabla 1 recoge la expresión del logaritmo de la función de verosimilitud para los distintos modelos presentados. Al no existir una solución analítica dada (como en el caso de los mínimos cuadrados ordinarios o generalizados), los parámetros deben estimarse mediante procedimientos iterativos de optimización.

El paquete estadístico SAS® tiene distintos procedimientos que permiten el ajuste de modelos basados en las distribuciones anteriores (Liu & Cela, 2008). El procedimiento GENMOD permite ajustar distribuciones de tipo Poisson y Binomial Negativa, mientras que COUNTREG permite modelizar distribuciones cero-infladas Poisson. Sin embargo, es el procedimiento NLMIXED el que da mayor flexibilidad, al permitir optimizar la función de verosimilitud asociada a cualquier distribución utilizando técnicas de *cuadratura gaussiana adaptativa* (Pinheiro & Bates, 2000). Asimismo, este procedimiento posibilita la inclusión de parámetros aleatorios. En el apéndice 1 se presentan los programas SAS® para la resolución de este tipo de modelos. Otros paquetes estadísticos también permiten resolver modelos mixtos lineales generalizados, como es el caso de R (Zuur et *al.*, 2009, cap. 11). El contraste entre los modelos anteriores aplicados sobre una misma base de datos puede hacerse mediante la comparación del logaritmo de la verosimilitud (*ll*), y de los criterios de información de Akaike (AIC) y Bayesiano (BIC). La comparación entre dos modelos anidados – uno completo y otro reducido con una submuestras de las variables incluidas – se puede realizar aplicando el test de la razón de verosimilitud, de forma tal que el estadístico $-2(ll_{reducido} - ll_{completo})$ se distribuya de acuerdo a una χ^2. En este tipo de modelo deben de compararse además los histogramas de frecuencias observadas y predichas, aplicando el test de Kolmogorov o un test χ^2. Debe de evaluarse además la capacidad del modelo para predecir los ceros, tanto en frecuencia como en especificidad (porcentaje de observaciones correctamente clasificadas como cero). Para ello, en los modelos cero-inflados y en los modelos *hurdle* es necesario definir un punto de corte o *cutoff*, valor límite del parámetro de probabilidad π por encima del cual se considera que el evento no va a ocurrir. Este valor puede ser fijo, o variar de forma aleatoria en cada observación. Por último, el conocimiento biológico del proceso a modelizar debe de ser determinante a la hora de elegir entre las distintas alternativas de modelización.

CASO DE ESTUDIO

Como caso de estudio y aplicación de las técnicas de modelización sobre bases de datos con exceso de ceros se presenta el estudio de la producción anual de fruto en masas regulares de *Pinus pinea* L. en la Meseta Norte de España (para más detalles sobre el caso de estudio consultar Calama et *al.*, 2011). Se ha utilizado una muestra de 6.998 observaciones procedentes de más 700 árboles localizados en 140 parcelas y en los que se ha cosechado, contado y pesado la producción de piña sana durante 10 años. Como variable dependiente se ha utilizado el peso (kg) de piña cosechado en un árbol un año determinado, variable de tipo continuo. Se ha identificado que el 54,55% de observaciones de la muestra se corresponden con valores ceros (ver Figura 1), proponiéndose entonces el ajuste de un modelo cero-inflado log-normal mediante la maximización de la función correspondiente de la Tabla 1.

Modelo	ll = Logaritmo de la función de verosimilitud (sobre la muestra)	Parámetros
Poisson	$\sum_i \left[-\lambda + Y_i \log(\lambda) - \log(Y_i!)\right]$	λ
NB	$\sum_i \left[\log\left(\frac{\Gamma\left(Y_i + k^{-1}\right)}{\Gamma\left(k^{-1}\right) Y_i!}\right) - \left(Y_i + k^{-1}\right)\log(k\mu + 1) + Y_i \log(k\mu)\right]$	μ, κ
ZIP	$\sum_i (1-I)\left[\log\left(\pi + (1-\pi)e^{-\lambda}\right)\right] + (I)\left[\log(1-\pi) - \lambda + Y_i \log(\lambda) - \log(Y_i!)\right]$	π, λ
ZINB	$\sum_i (1-I)\log\left(\pi + (1-\pi)\left(\frac{1}{k\mu+1}\right)^k\right) +$ $(I)\left[\log(1-\pi) + \log\left(\frac{\Gamma\left(Y_i + k^{-1}\right)}{\Gamma\left(k^{-1}\right) Y_i!}\right) - \left(Y_i + k^{-1}\right)\log(k\mu+1) + Y_i \log(k\mu)\right]$	π, μ, κ
Hurdle ZIP	$\sum_i (1-I)\left[\log(\pi)\right] + (I)\left[\log(1-\pi) - \lambda + Y_i \log(\lambda) - \log\left(1-e^{-\lambda}\right) - \log(Y_i!)\right]$	π, λ
ZILN	$\sum_i (1-I)\log(\pi) + (I)\left[\log(1-\pi)\ \frac{1}{Y_i\sigma\sqrt{2\pi}}\ \exp\left(-\frac{1}{2\sigma^2}(\ln(Y_i) - \mu)^2\right)\right]$	π, μ, σ

Tabla 1. *Logaritmo de la función de verosimilitud para los principales modelos presentados y parámetros a estimar. Donde I es una variable ficticia que adopta el valor 0 si Y=0 y 1 si Y>0*

Se ha propuesto la expansión de los parámetros π (probabilidad de no ocurrencia de fructificación) y μ (valor medio esperado de producción de fruto en peso en kg arbol^{-1}·año^{-1}, condicional a la ocurrencia de la fructificación) utilizando variables de árbol individual (diámetro, alturas, dimensiones de copa), rodal (densidad, área basímetrica, edad, altura dominante), estación (tipología general de suelo, índice de calidad de estación) y climáticas (eventos de precipitación y temperatura acaecidos durante los tres años anteriores a la maduración de los conos). El parámetro σ^2 de varianza del error de la componente lognormal se ha estimado sin expansión. Como los árboles se encontraban localizados en parcelas, se aproximó esta posible falta de independencia mediante la inclusión de dos componentes aleatorios a nivel de parcela que afectaban tanto a la regresión logit para definir la ocurrencia de fructificación como al modelo de regresión log-normal que define la abundancia:

$$\log\left(\frac{\pi}{1-\pi}\right) = 1{,}7062 - 0{,}0131\,pp_{my\_jn\_3} - 0{,}0111\,pp_{oc\_nv\_3} + 0{,}1003\,nhel + 0{,}5664\ln(N) - 0{,}0754\,SI + 2{,}7243\,T_{20} + 0{,}3298\,T_{50} - 2{,}2421\,d/dg + UN_1 + u \quad [9]$$

$$\ln(\mu) = -2{,}8930 + 0{,}0089\,pp_{my\_jn\_3} + 0{,}0055\,pp_{oc\_nv\_3} + 0{,}0030\,pp_{summ\_2} + 0{,}0036\,pp_{fb\_my\_0} - 0{,}0425\,nhel - 0{,}2673\ln(N) + 0{,}0421\,SI + 0{,}0454\,d + 0{,}5895\,d/dg + UN_2 + v \quad [10]$$

Donde $pp_{my\_jn\_3}$ es el valor de la precipitación (mm) en los meses de mayo y junio del año anterior a la floración; $pp_{oc\_nv\_3}$ es el valor de la precipitación (mm) en los meses de octubre y noviembre del año anterior a la floración; pp_{summ_2} es el valor de la precipitación (mm) del verano (julio–septiembre) posterior a la floración; $pp_{fb\_my\_0}$ es el valor de la precipitación (mm) en los meses de febrero a mayo del año de maduración del fruto; nhel es el número de días de helada severa (mínima inferior a -5ºC) durante el primer invierno tras la floración; N es la densidad de masa (pies.ha); SI el índice de sitio (m); T_{20} y T_{50} son dos variables categóricas que valen uno si la edad del rodal es menor que 20 ó 50 años, respectivamente, y 0 en el resto de casos; d es el diámetro normal (cm); d/dg el cociente entre el diámetro normal y el diámetro medio cuadrático; UN_1 y UN_2 los valores correspondientes a la estratificación edáfica y climáti-

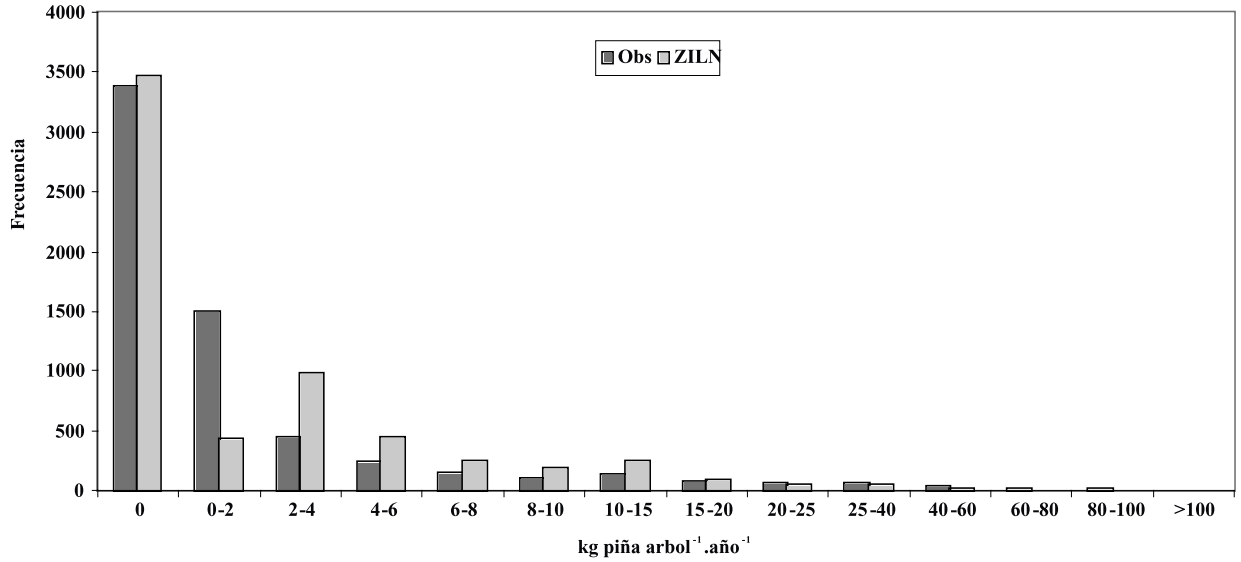

***Figura 5**. Comparación entre la distribución de frecuencias real observada (obs) y distribución de frecuencias esperadas para una distribución Cero inflada – log normal (ZILN), para el caso real de estudio de producción de piña*

ca del territorio; u y v son dos parámetros aleatorios de parcela, distribuidos según una normal bivariante de media (0,0) y varianzas σ^2_u = 0,6490 y σ^2_v = 0,2142 (coef correlación = 0,6364). El término de varianza del error σ^2 del componente log-normal es 1,0608.

Del modelo anterior se desprende que la probabilidad de fructificación viene condicionada por prácticamente las mismas variables que la abundancia de frutos, aumentando en ambos casos con la precipitación de la primavera y otoño del año anterior a la floración (coincidiendo con los momentos de formación y diferenciación de las yemas), viéndose afectadas de forma negativa por fenómenos acaecidos durante el proceso, y en general, favoreciéndose la fructificación en masas poco densas, maduras, y en los pies dominantes y de mayor tamaño. El modelo permite obtener estimaciones insesgadas de la producción de fruto (0,062 kg·arbol^{-1}·año^{-1}, p-valor=0,4254), alcanzando valores de eficiencia del modelo de 35,27% en la predicción a escala de árbol y año y del 51% a escala de parcela y año. En cuanto a la frecuencia de ceros, el modelo predice un porcentaje de ceros de 56,05% de las observaciones (frente al 54,55% observado), con una especificidad del 74,96%. Por último, los histogramas presentados (Figura 5) muestran una identificación entre las distribuciones de valores observados y predichos, indicador de un buen ajuste del modelo.

CONCLUSIONES

La sobreabundancia de ceros muestrales es un fenómeno habitual en los datos de estudio en el ámbito de la ecología y la ciencia forestal. La aplicación de métodos estadísticos clásicos, basados en la asunción de normalidad muestral, da lugar a resultados de interpretación errónea. En el presente trabajo se han presentado, comparado y discutido distintas alternativas como son la regresión Poisson, la regresión binomial negativa y la aplicación de modelos basados en distribuciones cero-infladas y distribuciones hurdle. Determinados aspectos como la naturaleza de la variable (continua o discreta), la presencia de ceros estructurales en la población, la abundancia de ceros, o la presencia de bimodalidad en la muestra condicionará la selección de uno u otro método. El procedimiento NLMIXED del paquete estadístico SAS® permite el ajuste mediante técnicas de máxima verosimilitud tanto de los modelos de distribución teóricos anteriores como de modelos expandidos sobre variables explicativos.

Agradecimientos

El presente trabajo se ha desarrollado en el marco financiero y funcional del proyecto AGL-15521 "*Dinámica y gestión en masas heterogé-*

neas de Pinus pinea: de la respuesta fisiológica a la modelización a escala regional en un escenario de cambio global".

BIBLIOGRAFÍA

AFFLECK, D.L.R.; 2006. Poisson mixture models for regression analysis of stand level mortality. *Can. J. For. Res.* 36: 2994-3006.

AITCHINSON, J.; 1955. On the distribution of a positive random variable having a discrete probability mass at the origin. *Journal of the American Statistical Association* 50: 901-908.

BLISS, C.I. & FISHER, R.A.; 1953. Fitting the negative binomial distribution to biological data and note on the efficient fitting of the negative binomial. *Biometrics* 9: 176-200.

CALAMA, R. & MONTERO, G.; 2007. Cone and seed production of Pinus pinea L. *Eur. J. For. Res.* 126: 23-35.

CALAMA, R.; MUTKE, S.; TOMÉ, J.A.; GORDO, J.; MONTERO, G. & TOMÉ, M.; 2011. Modelling spatial and temporal variability in a zero-inflated variable: The case of stone pine (Pinus pinea L.) cone production. *Ecol. Model.* 222: 606-618.

FORTIN, M. & DEBLOIS, J.; 2007. Modelling tree recruitment with zero-inflated models: the example of hardwood stands in Southern Québec, Canada. *For. Sci.* 53(4): 529-539.

HALL, DB.; 2000. Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics* 56: 1030-1039.

HALL, D.G.; CHILDERS, C.C.; EGER, J.E. & ALLEN, J.C.; 1997. Citrus rust mite (Acari: Eriophyidae) counts on fruit and the negative binomial distribution. Florida *Entomologist* 80: 1-10.

JOHNSON, N.L.; KEMP, A.W. & KOTZ, S.; 2005. *Univariate Discrete Distributions* (3rd Edition). Wiley Series in Statistics and Probability. John Wiley & Sons Inc.

KUHNERT, P.M.; MARTIN, T.G.; MENGERSEN, K. & POSSINGHAM, H.P.; 2005. Assessing the impacts of grazing levels on bird density in woodland habitats: a Bayesian approach using expert opinion. *Environmetrics* 16: 717-747.

LAMBERT, D.; 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 34:1-14.

LI, N.; ELASHOFF, D.A.; ROBBINS, W.A. & XUN, L.; 2008 A hierarchical zero-inflated log-normal model for skewed responses. *Stat. Meth. Med. Res.* DOI: 10.1177/0962280208097372.

LIU, WS. & CELA, J. 2008. *Count data models in SAS®*. Paper 371. SAS Global Forum.

MCCULLAGH, P. & NELDER, J.; 1989. *Generalized Linear Models.* Second Edition. Chapman and Hall/CRC. Boca Ratón.

MULLAHY, J.; 1986. Specification and testing of some modified count data models. *Journal of Econometrics* 33: 341-365.

PEÑA, D.; 1989. *Estadística: modelos y métodos. Vol 2: Modelos lineales y series temporales*. Alianza Universidad Textos. Madrid.

PINHEIRO, J.C. & BATES, D.M.; 2000. *Mixed-Effects Models in S and S-PLUS.* Springer Series in Statistics and Computing.

RIDOUT, M.; DEMETRIO, C.G.B. & HINDE, J.; 1998. Models for count data with many zeros. *In*: *Proceedings of the XIXth International Biometrics Conference*. 179-192. Cape Town.

TOOZE, J.; GRUNWALD, G.K. & JONES, R.H.; 2002. Analysis of repeated measurements data with clumping at zero. *Stat. Meth. Med. Res.* 11: 341-355.

TU, W.; 2002. Zero-inflated data. *In*: El- AH. Shaarawi & W.W. Piegorsch (eds.), *Encyclopedia of Environmetrics*: 2387-2391. John Wiley and Sons. Chichester.

WELSH, A.H; CUNNINGHAM, R.B.; DONNELLY, C.F. & LINDENMAYER, D.B.; 1996. Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecol. Model.* 88: 297-308.

WOOLLONS, R.C.; 1998. Even-aged stand mortality estimation through a two-step regression process. *Forest Ecol. Manage.* 105: 189-195.

ZUUR, A.F.; IENO, E.N.; WALKER, N.J.; SAVELIEV, A.A. & SMITH, G.M.; 2009. *Mixed Effects Models and extensions in Ecology.* R. Springer. NY.

Apéndice 1. Resolución de modelos con exceso de ceros mediante el paquete SAS®

```
/* Los siguientes modelos se refieren al ajuste de los datos a la distribución correspondiente, sin incluir covariables explicativas*/ /* Notas: p1=variable de interes; ll= logaritmo de la funcion de verosimilitud asociada; a0, b0, alpha, mu, varerror = parametros a estimar*/
/*Poisson*/
proc nlmixed data=prueba1 tech = dbldog;
parms b0=0 ; etap = b0 ; exp_etap = exp(etap);
ll = p1 * etap - exp_etap - lgamma(p1 + 1);
model p1 ~ general(ll);
run;
/*Binomial Negativa*/
proc nlmixed data=prueba1;
parms b0=2; mu=exp(b0);
ll = lgamma(p1 + 1 / alpha) - lgamma(p1 + 1) - lgamma(1 / alpha) +
p1 * log(alpha * mu) -(p1 + 1 / alpha) * log(1 + alpha * mu);
model p1 ~ general (ll);
run;
/*ZIP (Cero Inflado-Poisson)*/
proc nlmixed data=prueba1 tech = dbldog;
parms a0=0 b0=0 ; eta0 = a0; exp_eta0 = exp(eta0); p0 = exp_eta0 / (1 + exp_eta0); etap = b0; exp_etap = exp(etap);
if p1= 0 then ll = log(p0 + (1 - p0) * exp(-exp_etap));
else ll = log(1 - p0) + p1 * etap - exp_etap - lgamma(p1 + 1);
model p1 ~ general(ll);
run;
/*Hurdle ZIP*/
proc nlmixed data=prueba1 tech = dbldog;
parms a0=0 b0=0 ; eta0 = a0; exp_eta0 = exp(eta0); p0 = exp_eta0 / (1 + exp_eta0); etap = b0; exp_etap = exp(etap);
if p1= 0 then ll = log(p0) ;
else ll = log(1 - p0) + p1 * etap - exp_etap - lgamma(p1 + 1)-log(1-exp(-exp_etap));
model p1 ~ general(ll);
run;
/*ZINB*/
proc nlmixed data=prueba1 tech = dbldog maxiter=1000;
parms a0=0 b0=0 ; eta0 = a0; exp_eta0 = exp(eta0); p0 = exp_eta0 / (1 + exp_eta0); etap = b0; mu= exp(etap);
if p1=0 then ll = log(p0 + (1 - p0) * ((1/(1+alpha*mu))**(1/alpha)));
else ll = log(1 - p0) + lgamma(p1 + 1 / alpha) - lgamma(p1 + 1) - lgamma(1/alpha) + p1 * log(alpha * mu) -(p1 + 1 / alpha) * log(1 + alpha * mu) ;
model p1~ general(ll);
run;
/*Hurdle ZINB*/
proc nlmixed data=prueba1 tech = dbldog maxiter=1000;
parms a0=0 b0=0 ; eta0 = a0; exp_eta0 = exp(eta0); p0 = exp_eta0 / (1 + exp_eta0); etap = b0; mu= exp(etap);
if p1=0 then ll = log(p0);
else ll = log(1 - p0) + lgamma(p1 + 1 / alpha) - lgamma(p1 + 1) - lgamma(1/alpha) + p1 * log(alpha * mu) -(p1 + 1 / alpha) * log(1 + alpha * mu) ;
model p1~ general(ll);
run;
/*ZILN*/
proc nlmixed data=prueba1 tech=dbldog;
parms a0=0 b0=0 varerror=1;
bounds varerror>0; eta0 = a0; exp_eta0 = exp(eta0); p0 = exp_eta0/(1 + exp_eta0); etap = b0; mu= exp(etap);
parte1=(-1/(2*varerror)); parte2= log(1/sqrt(2*4*atan(1)*varerror));
if y = 0 then ll = log(p0);
else ll = log(1 - p0) + (parte1*((y-log(mu))**2))+ parte2;
model y ~ general(ll);
run;
```