UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Noviembre 2020, Vol. 13, No. 2, pp. 9–32

Comparación del modelo COM-Poisson y el modelo Poisson

Comparison of the COM-Poisson model and the Poisson model shorttitle

Álvaro Arley Castaño Colorado[a]
aacastan@unal.edu.co

Juan Carlos Correa Morales[b]
correo electronico

Resumen

La modelación de datos de conteo se hace típicamente usando el modelo Poisson, en el cual se asume equidispersión (ED), en donde la media y la varianza son iguales. Cuando esta condición no es fácil de justificar, han surgido diferentes alternativas, unas más flexibles que otras, en cuanto a la capacidad de manejar tanto sobredispersión (OD) como subdispersión (UD). Una de ellas es el modelo COM-Poisson el cual fue propuesto recientemente y ha sido evaluado en términos inferenciales (Sellers & Shmueli 2010*a*). Esta investigación quiere cuantificar la calidad predictiva del modelo COM-Poisson con respecto al modelo Poisson, y así establecer la pérdida en la eficiencia que se tiene al ajustar el modelo inadecuado cuando la propiedad de equidispersión no es satisfactoria. El estudio de simulación efectuado determinó que al ajustar el modelo inadecuado, ya sea en sobre o subdispersión, no representa, en la mayoría de los casos, ni una ganancia o pérdida en cuanto a la calidad predictiva de los valores ajustados. Dos estudios de caso aplicados a la ecología ilustran los resultados obtenidos.

***Palabras clave*:** Datos de Conteo, Modelos Lineales Generalizados, Eficiencia Relativa, Regresión Poisson, Regresión Conway-Maxwell-Poisson, Capacidad Predictiva, Dispersión.

Abstract

When modeling count data, the poisson model is typically used, in which the equidispersion (ED) assumption is assumed, where the mean and variance are equal. When this condition is not easy to justify, different alternatives have been proposed, some more flexible than others in terms of accounting for both overdispersion (OD) and underdispersion (UN). One of them is the COM-Poisson model which

[a]Universidad Nacionall
[b]Afiliación institucional

was recently proposed and has been evaluated in inferential terms. The investigation presented here aims to compare the COM-Poisson model predictive quality with respect to the Poisson model and establish the loss in efficiency that occurs when the inadequate model is fitted when the property of equidispersion is not satisfactory. A simulation study determined that adjusting the inappropriate model either over or underdispersion does not represent in most cases, a gain or loss of the predictive quality. Two case studies illustrate ours findings obtained here.

***Keywords*:** Count Data, Generalized Linear Models, Relative Efficiency, Poisson regression, Conway-Maxwell-Poisson regression, Predictive Power, Dispersion.s.

1. Introducción

Los datos de conteo se refieren al número de veces que se da un evento en un período de tiempo o espacio definido, por ejemplo, el número de accidentes aéreos, el número de días de permanencia en un hospital, la cantidad de frutos en un árbol. Este tipo de datos toman valores enteros no negativos y se asume que los eventos en un intervalo de tiempo o espacio determinado son independientes e idénticamente distribuidos (Cameron & Trivedi 1998).

El problema que han tenido los modelos para datos de conteo son los niveles de dispersión que estos pueden tomar. Según Hilbe (2011), en la mayoría de los casos no es frecuente que los datos de conteo en la realidad tengan ED lo cual siempre se asume en la distribución Poisson. El modelo típico sobre el cual parte el análisis de este tipo de datos es el modelo Poisson. Éste se caracteriza por el supuesto de ED, en donde la media y la varianza son iguales, lo cual puede ser causante de un ajuste inadecuado cuando no se cumpla dicha condición, es decir, que este modelo no explica bien conjuntos de datos que presentan casos de OD o UD (Sellers & Shmueli 2010*a*). Es más común encontrar datos con OD o UD, aunque este último con menos frecuencia. Cuando se habla de OD en los conteos por unidad de tiempo o espacio, se refiere a que la varianza excede su media y cuando la varianza es menor que la media se denomina UD. Según Dobson (2002), hay una forma de determinar estos niveles de dispersión hallando un término de dispersión de acuerdo con la ecuación (1):

$$Var(Y) = \phi E(Y) = \phi\mu, \tag{1}$$

donde Y es la variable de conteo, μ es la media poblacional y ϕ es la constante de variación la cual indica que si $\phi > 1$, hay OD, y si $\phi < 1$, hay UD.

Con el tiempo se han desarrollado diversas alternativas para modelar conteos bajo la violación de este supuesto, entre ellas están la regresión Binomial Negativa (BN) (Hilbe 2011), la regresión Poisson Generalizada Restringida (PGR) (Famoye 1993), la regresión hyper-Poisson (Sáez-Castillo & Conde-Sánchez 2013), entre otras. Según Sellers & Shmueli (2010*a*) la regresión BN, a pesar de que ex-

plica correctamente datos con OD, no es adecuada para la modelación cuando la varianza es inferior a la media. En cuanto a la regresión PGR, los autores enuncian que dicho modelo puede ajustar tanto datos con OD como UD, pero su capacidad de estimación es limitada en este último caso presentando fallas en la convergencia del modelo.

Recientemente ha surgido una propuesta más flexible la cual se adapta bien a los diferentes niveles de dispersión en los datos de conteo, es denominado el modelo Conway-Maxwell Poisson (CMP). El establecimiento de esta alternativa dentro del contexto inferencial está en proceso de estudio y evaluación. Hasta ahora se han evaluado sus propiedades inferenciales e incluso se han hecho modificaciones a la propuesta original. Una de ellas es el planteamiento de Guikema & Goffelt (2008) quienes reparametrizaron el modelo CMP original y lo adaptaron dentro del marco de un modelo lineal generalizado. Desde una perspectiva Bayesiana, Lord et al. (2008) presentaron una formulación con una función de enlace $\log \lambda^{1/\nu}$ logrando buenos ajustes del modelo CMP a pesar de no poseer características de la familia exponencial. Sin embargo, esta propuesta fue muy demandante a nivel computacional a la hora de usar la estimación por el método de Markov Chain Monte Carlo. Se han desarrollado estudios con el fin de establecer la calidad de las estimaciones del modelo CMP, evaluando el comportamiento de los estimadores en diversos escenarios (Geedipally et al. 2008, Jowaheer & Mamode 2009, Lord et al. 2010, Sellers & Shmueli 2010*b*, Francis et al. 2012). Los métodos de estimación de parámetros, los niveles de dispersión, las medias y tamaños muestrales han sido los componentes para establecer dichos escenarios y hacer la comparación y evaluación respectiva.

En vista de que se tiene un modelo tradicional y parsimonioso como lo es el Poisson, surge la necesidad de compararlo con un modelo más flexible especificado para ajustar diferentes niveles de dispersión. Sellers et al. (2012) señalan que aunque las investigaciones han arrojado un buen comportamiento en la bondad de ajuste del modelo CMP, su capacidad predictiva aún es una incógnita. En esta investigación se logró determinar vía simulación la eficiencia relativa entre el modelo Poisson y el modelo CMP, camparándolos por medio de medidas de calidad de las predicciones, en diversos escenarios que tendrán como factores a controlar la variación de la dispersión y del intercepto. También se efectuó un contraste entre dos métodos de predicción propuestos por Sellers & Shmueli (2010*a*), ya que según Minka et al. (2003) la aproximación para obtener predicciones (ecuación (6)) no es adecuada en ciertos casos. Con esta investigación se quiere aportar información que permita determinar el comportamiento inferencial de las estimaciones del modelo en cuanto al desempeño predictivo, y al ser contrastado con estudios similares permitirá establecer una base teórica y aplicada más robusta en cuanto a su flexibilidad para ajustar conteos carentes de ED.

2. El Modelo Poisson

Según Cameron & Trivedi (1998), la distribución Poisson, que lleva el apellido de su formulador, se estableció a partir de un caso límite de la distribución binomial. Su propiedad fundamental es la equidispersión donde la varianza es igual a la media, y a partir de esta relación se derivan los condicionamientos para la formulación de otros modelos para datos de conteo. La ecuación (2) muestra su función de masa de probabilidad:

$$P(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!}, \quad y = 0, 1, 2, \ldots, \tag{2}$$

donde λ es la media del número de eventos en un intervalo de longitud de espacio o tiempo. El modelo de regresión Poisson pertenece a la familia de los modelos lineales generalizados, ya que su función de distribución pertenece a la familia exponencial, su predictor es lineal ($\eta = X\beta$) y tiene una función de enlace g tal que $E(Y) = \mu = g^{-1}(\eta)$ (McCullagh & Nelder 1972). Este modelo es expresado por la función dada en la ecuación (3):

$$E(y_i|x_i) = \mu_i = \exp\left\{x_i'\beta\right\}, \quad i = 1, \ldots, n, \tag{3}$$

la cual sigue una distribución condicional de y_i (variable dependiente) en función de un vector de covariables x_i y de parámetros β (Cameron & Trivedi 1998). Esta es la forma multiplicativa del modelo y se expresa así ya que de esta manera asegura que μ tendrá valores enteros no negativos. Si se plantea una forma aditiva hay un riesgo de que ciertas combinaciones no cumplan con esta restricción (Cameron & Trivedi 1998).

La estimación por máxima verosimilitud se configura según las dos ecuaciones anteriores y teniendo en cuenta que las observaciones son independientes. La ecuación (4) es la función de log-verosimilitud obtenida para esta distribución:

$$\log L(\beta) = \sum_{i=1}^{n} \left\{y_i x_i'\beta - \exp(x_i'\beta) - \log y_i!\right\}, \tag{4}$$

la cual es globalmente concava y los parámetros estimados pueden ser obtenidos mediante algoritmos iterativos tales como el de Gauss-Newton o Newton-Raphson (Cameron & Trivedi 2003).

La mayor desventaja del modelo Poisson es que no explica correctamente conjuntos de datos en los que existe sobredispersión o subdispersión dada su propiedad de equidispersión (Sellers & Shmueli 2010*a*), lo que puede generar errores estándar de los coeficientes subestimados generando coeficientes significativos cuando en realidad estos no lo son (Cameron & Trivedi 2003). Por ello, se han diseñado nuevas propuestas que pretenden ser más flexibles y que abarcan los diferentes niveles de dispersión que puede tomar este tipo de datos.

3. El Modelo COM-Poisson

La distribución COM-Poisson fue propuesta por Conway y Maxwell en 1962, pero sus propiedades probabilísticas y de regresión fueron estudiadas por Shmueli et al. (2005). La función de masa de probabilidad está dada por la ecuación (5):

$$P(Y = y) = \frac{\lambda^y}{(y!)^v \, Z(\lambda, v)}, \quad y = 0, 1, 2, \cdots, \lambda > 0, \nu \geq 0, \tag{5}$$

donde $Z(\lambda, \nu) = \sum_{s=0}^{\infty} \frac{\lambda^s}{(s!)^\nu}$ es la constante de normalización y $\nu \geq 0$ es el parámetro de forma o de dispersión, indicando OD si $\nu < 1$ y UD si $\nu > 1$ (Sellers & Shmueli 2010*a*). Esta distribución pertenece a la familia exponencial y generaliza a la distribución Poisson (cuando $\nu = 1$), la distribución geométrica (cuando $\nu = 0$ y $\lambda < 1$) y la distribución Bernoulli (cuando $\nu \to \infty$ y con probabilidad $\frac{\lambda}{1+\lambda}$) (Shmueli et al. 2005).

El valor esperado y la varianza están dados por las ecuaciones (6) y (7):

$$E(Y) = \frac{\partial \log Z(\lambda, \nu)}{\partial \log \lambda} \approx \lambda^{1/\nu} - \frac{\nu - 1}{2\nu}, \tag{6}$$

$$Var(Y) = \frac{\partial E(Y)}{\partial \log \lambda} \approx \frac{1}{\nu} \lambda^{1/\nu}, \tag{7}$$

como se puede observar estas funciones no tienen una forma cerrada y se relacionan entre sí mediante expresiones aproximadas (Sellers & Shmueli 2010*a*).

La construcción del modelo se da a partir de un caso log-lineal de la regresión Poisson. De allí se deduce la función de log-verosimilitud representada en la ecuación (8):

$$\log L(\lambda_i, \nu) = \sum_{i=1}^{n} y_i \log \lambda_i - \nu \sum_{i=1}^{n} \log y_i! - \sum_{i=1}^{n} \log Z(\lambda_i, \nu), \tag{8}$$

que al ser maximizada mediante optmización lineal no restringida y teniendo en cuenta que $\nu \geq 0$ se obtienen los coeficientes estimados de máxima verosimilitud. Esta función también puede ser maximizada usando mínimos cuadrados ponderados a través de soluciones iterativas (Sellers & Shmueli 2010*a*).

Los valores ajustados se pueden obtener por medio de medias o medianas estimadas, ya que según Minka et al. (2003) la aproximación de la ecuación (6) es adecuada cuando $\nu \leq 1$ o $\lambda > 10^\nu$ (Sellers & Shmueli 2010*a*).

Este modelo ajusta bien datos con diferentes niveles de dispersión (Sellers & Shmueli 2010*a*), pero tiene una restricción análoga al supuesto de homocedasticidad en el caso de la regresión lineal, en donde se asume que el modelo tiene en cuenta un nivel de dispersión constante a través de todas las observaciones.

4. Estudio de simulación

Para evaluar la eficiencia entre las predicciones de los modelos que son objeto de comparación en este estudio, se diseñó un procedimiento de simulación, el cual se basa en el algoritmo descrito a continuación. Una de las características específicas de tal procedimiento es que se utilizó un tamaño muestral (n) constante para todos los conjuntos de datos generados. De acuerdo con Sellers & Shmueli (2010*a*) y Miller (2007), la normalidad asintótica de los parámetros estimados no se puede asegurar en tamaños muestrales pequeños. Teniendo en cuenta la anterior afirmación, además de los problemas de convergencia en el ajuste de los modelos y la demanda computacional al variar n, se definió un nivel constante para este factor con el fin de diagnosticar el comportamiento predictivo al lograr la normalidad asintótica de los parámetros del modelo CMP (consultar apéndice A). Para comparar los modelos en términos de calidad de las predicciones se determinó un tamaño muestral de 1000 observaciones, el cual es el tamaño usado en los trabajos de Francis et al. (2012) y Winkelmann (2008).

4.1. Algoritmo de simulación

1. Generar covariables fijas y ortogonales x_1 y x_2 con $n = 1000$ a partir de una distribución uniforme de 0 a 1:

   $x_1 \sim U(0,1)$ y $x_2 \sim U(0,1)$

2. Generar un conjunto de datos de conteo con un tamaño $n = 1000$ de una distribución Poisson para ED o de una distribución CMP para OD y UD.

   $Y_i \sim Poisson(\lambda_i)$, para ED con la función `rpois`
   $Y_i \sim CMP(\lambda_i, \nu)$, para OD y UD con la función `rcomp` de la librería CompGLM

3. Ajustar modelos Poisson, CMP y BN en OD. Y ajustar modelos Poisson, CMP y PGR en ED y UD.

4. Almacenar predicciones de los modelos ajustados y calcular medidas de calidad predictiva.

5. Repetir los pasos del 1 al 4 hasta 1000 simulaciones.

En total, se generaron 1000 conjuntos de datos para cada uno de los escenarios conformados por varias intensidades de dispersión (ν) y los modelos asumidos (m). Las intensidades de dispersión abarcan una rango amplio tanto de OD como UD, dado que se definieron tres niveles para OD ($\nu = 0.25, 0.5, 0.75$) y tres para UD ($\nu = 1.5, 2.5, 5$) y $\nu = 1$ para el caso de ED. Cuatro clases de los modelos asumidos se definieron según el valor verdadero adoptado para el parámetro β_0 dejando constantes los coeficientes asumidos asociados a las variables predictoras, indicando que el menor valor asumido de β_0 corresponde al modelo asumido de más baja denominación y así respectivamente hasta el modelo de mayor denominación.

En las tablas 1, 2, 3, se muestran los coeficientes asumidos para generar los datos para cada uno de los escenarios.

A los conjuntos de datos generados se les ajustaron los modelos CMP y Poisson, para ser contrastados con un modelo BN el cual tiene buen desempeño bajo OD. Y en ED y UD con el modelo PGR de buen comportamiento en UD. El modelo Poisson fue ajustado usando la función `glm` especificando la familia Poisson con función de enlace log. Esta función utiliza el método de Mínimos Cuadrados Reponderados Iterativamente (MCRI) para obtener las estimaciones de los parámetros del modelo. Para ajustar los modelos CMP se usó la función `glm.comp` del pauqete CompGLM que provee el algoritmo de optimización con el método de quasi-Newton.

Luego se almacenaron las predicciones de los modelos ajustados y se calcularon medidas de calidad predictiva.

Tabla 1: *Coeficientes asumidos para el estudio de simulación de eficiencia entre el modelo CMP y el modelo Poisson en, OD.*

	ν =**0.25**				ν =**0.50**				ν =**0.75**			
	m1	**m2**	**m3**	**m4**	**m1**	**m2**	**m3**	**m4**	**m1**	**m2**	**m3**	**m4**
β_0	-0.50	0.30	0.50	0.70	-0.30	0.70	1.10	1.50	-0.10	1.20	1.70	2.20
β_1	-0.50	-0.50	-0.50	-0.50	-0.50	-0.50	-0.50	-0.50	-0.50	-0.50	-0.50	-0.50
β_2	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50

Tabla 2: *Coeficientes asumidos para el estudio de simulación de eficiencia entre el modelo CMP y el modelo Poisson, en ED.*

	ν =**1.00**			
	m1	**m2**	**m3**	**m4**
β_0	0.10	1.60	2.30	3.00
β_1	-0.50	-0.50	-0.50	-0.50
β_2	0.50	0.50	0.50	0.50

Tabla 3: *Coeficientes asumidos para el estudio de simulación de eficiencia entre el modelo CMP y el modelo Poisson, en UD.*

	ν=**1.50**				ν =**2.50**				ν =**5.00**			
	m1	**m2**	**m3**	**m4**	**m1**	**m2**	**m3**	**m4**	**m1**	**m2**	**m3**	**m4**
β_0	0.50	2.50	3.50	4.50	1.00	4.20	6.00	7.50	2.00	8.50	12.0	15.0
β_1	-0.50	-0.50	-0.50	-0.50	-0.50	-0.50	-0.50	-0.50	-0.50	-0.50	-0.50	-0.50
β_2	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50

Tal como se ve en las tablas de coeficientes asumidos y con base en la metodología propuesta por Francis et al. (2012), el rango de niveles de dispersión es amplio y en los diferentes modelos asumidos (m1, m2, m3, m4) el coeficiente verdadero para β_0 es diferente mientras que para β_1 y β_2 fueron constantes.

4.2. Comparación de métodos de predicción

En el trabajo de Sellers & Shmueli (2010*a*) se proponen dos métodos de predicción para obtener valores ajustados. El método de predicción de medias que se basa en el cálculo de la media condicional,

$$\hat{y}_i|x_i = \hat{\lambda}_i^{1/\hat{\nu}} - \frac{\hat{\nu}-1}{2\hat{\nu}}, \tag{9}$$

la cual está en función de $\hat{\lambda}$ y $\hat{\nu}$. El método de predicción de medianas consiste en el cálculo de probabilidades consecutivas por medio de la ecuación (10):

$$P(Y_i = y_i) = \left(\frac{\lambda_i}{y_i}\right)^{\nu} P(Y_i = y_i - 1), \tag{10}$$

cuando la suma supera el valor de 0.5, es decir, el valor correspondiente a la mediana se obtiene tal predicción (Sellers & Shmueli 2010*b*).

Aunque Sellers & Shmueli (2010*b*) señalan que el método de predicción de medianas tiene ventajas en cuanto a que predice valores enteros y que la mediana es una medida de tendencia central más robusta en distribuciones sesgadas, no es claro si es más adecuada o no en términos del comportamiento predictivo, especialmente en el escenario de subdispersión donde de acuerdo con Minka et al. (2003) la aproximación a la media no es tan exacta. Para evaluar cuál de los métodos es más adecuado se compararon las predicciones obtenidas en los diferentes escenarios configurados en términos de la calidad predictiva.

4.3. Medidas de calidad predictiva

Luego de obtener los valores ajustados para cada modelo y en cada conjunto de datos generado se calcularon medidas estadísticas para caracterizar el comportamiento predictivo en los diferentes escenarios planteados. Las métricas que se tomaron en cuenta son presentadas a continuación.

4.3.1. Error Cuadrático Medio de Predicción (ECMP)

Ésta es una métrica implementada por Lord et al. (2008) y Sellers & Shmueli (2010*b*) para evaluar el comportamiento de las predicciones de los modelos comparados. Esta medida de calidad predictiva se obtuvo mediante la ecuación (11):

$$ECMP = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2, \tag{11}$$

donde, y es la respuesta observada, $\hat{y}$ es el valor ajustado o predicho, n es el número de observaciones o tamaño muestral.

4.3.2. Eficiencia Relativa (ER)

Las eficiencias relativas respectivas se obtuvieron con la ecuación (12):

$$ER_{(\hat{Y}_1,\hat{Y}_2)} = \frac{ECMP_{\hat{Y}_2}}{ECMP_{\hat{Y}_1}}, \tag{12}$$

donde ECMP es el error cuadrático medio de predicción. A manera de interpretación, si ER > 1, entonces, las predicciones del modelo 1 ($\hat{Y}_1$) son más eficientes que las del modelo 2 ($\hat{Y}_2$). Con esta medida se establece un criterio para evaluar si se pierde o no calidad en las predicciones al ajustar un modelo equivocado respecto al modelo adecuado o alternativo.

Se utilizó R (R Core Team 2017), un paquete computacional con enfoque estadístico de carácter libre y gratuito, para implementar la simulación y obtener los resultados estadísticos que serán objeto de análisis dentro de la investigación. Todas las simulaciones se realizaron en un computador con procesador Intel® Core™ i5-2430M con velocidad de 2.4 Ghz y con el sistema operativo Microsoft® Windows™ 7 Ultimate de arquitectura de 64 bits.

5. Resultados de la simulación

En esta sección dan a conocer los resultados que arrojó el estudio de simulación. En cada escenario de dispersión se obtuvieron los comportamientos de la calidad predictiva y la ER de los diferentes modelos que son objeto de comparación. También se presentan las tablas de proporciones de ER, las cuales indican el número de veces que un modelo fue más eficiente con respecto a otro durante la simulación.

5.1. Eficiencia Relativa bajo OD

La evaluación de la ER indica que la diferencias más notorias se presentan en el nivel de OD más fuerte. En el modelo asumido m1 fue poco eficiente el modelo CMP respecto a los modelos Poisson y BN. Lo contrario se dio en los modelos asumidos m2 y m3, en donde el modelo CMP es más eficiente que los otros dos modelos. Y en el modelo asumido m4, el modelo CMP solo es eficiente respecto al modelo BN. También se alcanza a notar que el modelo Poisson es ligeramente más eficiente que el modelo BN. En los niveles de OD menos severos las diferencias fueron menos perceptibles entre las diversas distribuciones contrastadas.

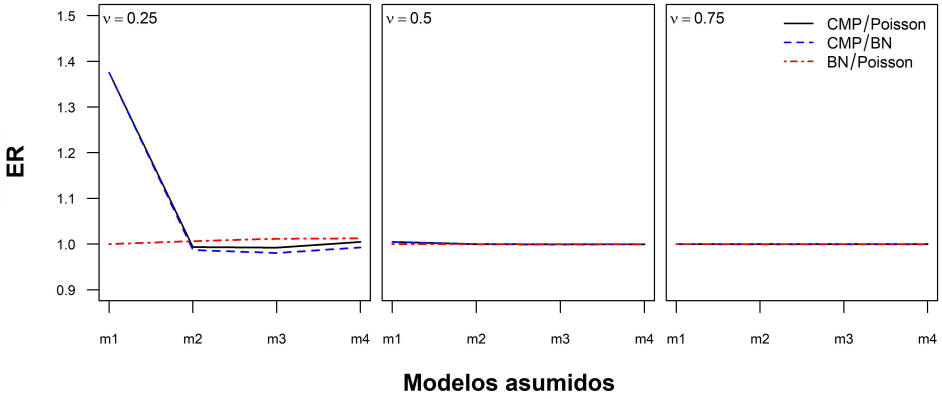

Figura 1: *Desempeño predictivo bajo OD.*

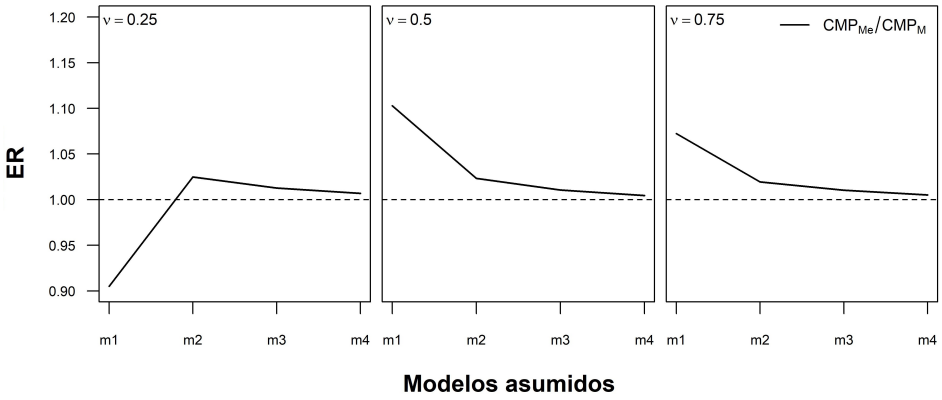

Figura 2: *Comparación del método de predicción de medias (M) y de medianas (Me) en el modelo CMP, bajo OD.*

La ER señala que el único caso donde las predicciones de mediana fueron más eficientes que las de la aproximación de la media condicional, fue en el modelo asumido de menor denominación m1, en el nivel de OD más fuerte ($\nu = 0.25$). En el resto de modelos asumidos las predicciones de media fueron más eficientes, aunque hay una tendencia en la ER a ser más cercana a 1 a medida que aumenta el coeficiente asumido para β_0 (Figura 2),.

La tabla 4 corrobora los resultados anteriores. Muestra que la comparación entre el modelo CMP y Poisson, el modelo CMP obtuvo el mayor número de casos de eficiencia en los modelos asumidos m2 y m3, especialmente en el nivel de OD más fuerte. La comparación entre el modelo CMP y el BN indicó que el modelo CMP obtuvo mayores casos de eficiencia entre los modelos asumidos m2 y m4, incluso en niveles de OD menos severos. Y característicamente, el contraste entre los modelos

Tabla 4: *Proporción de ER en un escenario de OD con* $n = 1000$.

OD		**Comparaciones**			
ν	**m**	CMP vs Poisson	CMP vs BN	BN vs Poisson	$\hat{Y}_{Me}$ vs $\hat{Y}_M$
ν =**0.25**	**m1**	1.000	1.000	0.996	0.353
	m2	0.226	0.159	0.996	1.000
	m3	0.173	0.095	0.995	1.000
	m4	0.659	0.351	0.992	0.656
ν =**0.50**	**m1**	0.945	0.931	0.999	1.000
	m2	0.481	0.448	0.996	1.000
	m3	0.407	0.371	0.996	0.994
	m4	0.432	0.319	0.992	0.944
ν =**0.75**	**m1**	0.755	0.730	1.000	1.000
	m2	0.494	0.471	1.000	1.000
	m3	0.466	0.407	0.998	0.975
	m4	0.468	0.315	1.000	0.931

BN y el Poisson, arrojó una mayor tasa de casos de eficiencia a favor del modelo Poisson a través de los diferentes niveles de OD.

En cuanto a la comparación de las propuestas de predicción, se nota la ventaja en casos de eficiencia de la predicción de medianas en el nivel de OD más fuerte y en m1, ya que en el resto de escenarios fue predominante los casos de eficiencia de la predicción de medias (Figura 2).

5.2. Eficiencia Relativa bajo ED

Al evaluar la ER entre ellos, es casi imperceptible algún rasgo que de un indicio de eficiencia en las predicciones a favor de uno u otro modelo. La figura 3 izquierda muestra que con número de tres cifras decimales ninguna de las propuestas para análisis de datos de conteo es más eficiente una de la otra.

La comparación de los métodos de predicción a pesar de ser muy similares entre sí en cuanto a su calidad predictiva, muestra según la ER, que fueron más eficientes las predicciones de medias en todos los casos, respecto a las de medianas (Figura 3 derecha inferior). Sin embargo, se nota una tendencia a disminuir la brecha en el desempeño predictivo a medida que aumenta el valor asumido para β_0, es decir, la denominación del modelo asumido.

En cuanto a la proporción de casos de ER, la mayoría de los escenarios muestran que hubieron proporciones equilibradas de eficiencia de un modelo respecto al otro. Las comparaciones del modelo CMP respecto al modelo Poisson y el modelo PGR muestran que la proporción de casos de eficiencia entre estos tres modelos fue cercana al 50 %, con una leve ventaja de estos dos últimos en m1. En la comparación del modelo PGR y el Poisson hay una proporción de ventaja a favor

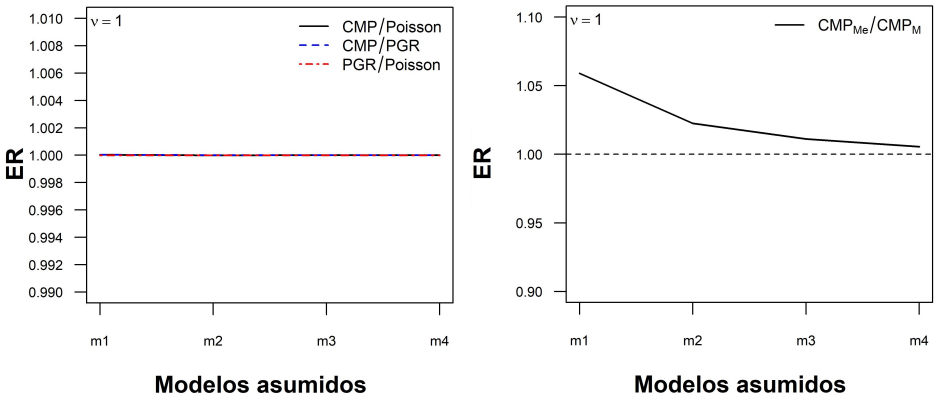

Figura 3: *Desempeño predictivo (izquierda) y comparación del método de predicción de medias (M) y de medianas (Me) en el modelo CMP (derecha), bajo ED.*

de la eficiencia del modelo PGR.

Tabla 5: *Proporción de ER en un escenario de ED con* $n = 1000$.

ED		**Comparaciones**			
ν	**m**	CMP vs Poisson	CMP vs PGR	PGR vs Poisson	$\hat{Y}_{Me}$ vs $\hat{Y}_M$
ν =**1.00**	**m1**	0.664	0.666	0.445	1.000
	m2	0.519	0.521	0.447	0.996
	m3	0.495	0.530	0.408	0.974
	m4	0.491	0.545	0.400	0.914

El método de predicción de medias en el modelo CMP demostró mejor calidad predictiva en ED ya que la proporción de casos de eficiencia de este método estuvo siempre por encima del 90 % respecto a las predicciones de mediana en todos los modelos asumidos (Tabla 5).

5.3. Eficiencia Relativa bajo UD

Bajo UD, la ER demostró que las predicciones del modelo CMP fueron menos eficientes respecto a los demás modelos cuando se asumió el valor más bajo de β_0. Mientras, que en la comparación del modelo PGR y el Poisson no se logró detectar una eficiencia de un modelo respecto al otro dado que los valores de ER son muy cercanos a 1.

La figura 5 muestra los comportamientos de la calidad predictiva entre los dos métodos de predicción planteados para obtener valores ajustados en el modelo CMP. Al determinar la ER se pudo notar que de forma generalizada el procedimiento de obtener valores ajustados por medio de la aproximación a la media fue

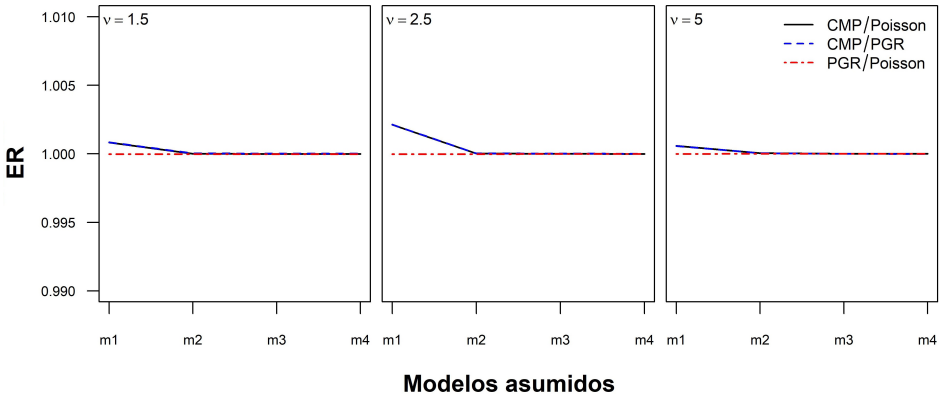

Figura 4: *Desempeño predictivo bajo UD.*

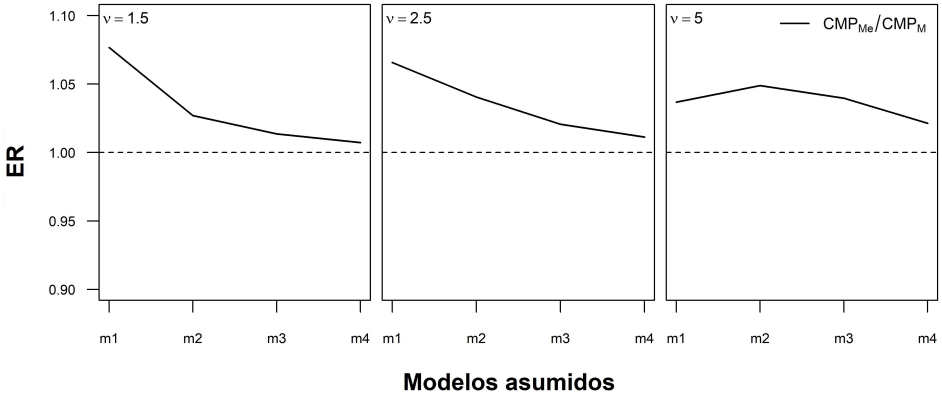

Figura 5: *Comparación del método de predicción de medias (M) y de medianas (Me) en el modelo CMP, bajo UD.*

más eficiente que el método de predicción de medianas.

La tabla 6 complementa la información gráfica descrita anteriormente. La comparación en el comportamiento de las predicciones entre el modelo CMP y las dos propuestas muestran proporciones similares en los casos más leves de UD (ν =1.5 y ν =2.5); mientras que el en el caso más fuerte de UD, además de presentar casos totales de menor eficiencia en m1, en m2 esta proporción fue superior al 95 %, indicando que los modelos Poisson y PGR obtuvieron mayores casos eficiencia respecto al modelo CMP cuando se asumieron valores bajos de β_0. La comparación entre el modelo Poisson y el PGR muestra que este último obtuvo los mayores casos de eficiencia en ν =1.5 y ν =2.5 y en el caso más severo de UD, las proporciones muestran ventajas para el modelo Poisson excepto en m1.

En cuanto a la comparación de los métodos de predicción, es claro que los valores

Tabla 6: *Proporción de ER en un escenario de UD con* $n = 1000$.

UD		**Comparaciones**			
ν	**m**	CMP vs Poisson	CMP vs PGR	PGR vs Poisson	$\hat{Y}_{Me}$ vs $\hat{Y}_M$
ν =**1.50**	**m1**	1.000	1.000	0.000	1.000
	m2	0.557	0.598	0.002	0.997
	m3	0.528	0.584	0.001	0.978
	m4	0.522	0.657	0.000	0.931
ν =**2.50**	**m1**	1.000	1.000	0.003	1.000
	m2	0.665	0.693	0.005	0.999
	m3	0.546	0.572	0.012	0.990
	m4	0.524	0.572	0.012	0.951
ν =**5.00**	**m1**	1.000	1.000	0.001	1.000
	m2	0.957	0.955	0.616	1.000
	m3	0.641	0.628	0.920	0.999
	m4	0.546	0.523	0.944	0.992

ajustados por medio de la media condicional propuesta para el modelo CMP superan en mayor número de casos de eficiencia a aquellos obtenidos por predicción de medianas.

6. Discusión

Los resultados presentados en la sección anterior mostraron varios componentes de análisis que son discutidos en esta sección. Aspectos tanto de la calidad predictiva como de la evaluación del método de predicción más adecuado para el modelo CMP, son analizados en esta sección dentro del marco de la normalidad asintótica de las estimaciones.

La comparación entre las propuestas de análisis de datos de conteo marcaron algunos resultados relevantes. Uno de ellos, se relaciona con la comparación que es objeto de esta investigación. La comparación del desempeño predictivo entre el modelo CMP y el modelo Poisson en el marco de la normalidad asintótica de las estimaciones determinó que únicamente en el caso de OD más severo y cuando se definieron los modelos asumidos m2 y m3, las predicciones de media del modelo CMP fueron ligeramente más eficientes que las del modelo Poisson. El modelo CMP presentó un desempeño deficiente de sus predicciones especialmente cuando se asumió el valor más bajo para el intercepto, siendo el nivel de OD más fuerte la evidencia más clara ya que allí fueron mucho más eficientes las predicciones de los modelos Poisson y BN. En UD también se detectó un bajo desempeño de las predicciones del modelo CMP en valores bajos asumidos para β_0 pero en una escala menor a la presentada en OD. También este resultado concuerda con lo expresado por Francis et al. (2012), en donde se concluye que el modelo reparametrizado

CMP (adaptado a un modelo lineal generalizado) tiene un desempeño deficiente o limitado en OD cuando se asumieron valores bajos para β_0.

Respecto a la comparación del modelo BN con el modelo CMP, se destaca la eficiencia en OD extrema que hay a favor de la calidad predictiva de este último, excepto cuando se asumió el valor más bajo para el intercepto. En los escenarios de OD más cercanos a la ED ya no se marcan eficiencias a favor de uno u otro modelo. Lo mismo sucedió al establecer la ER entre los modelos BN y Poisson, donde solo en el caso más fuerte de OD las predicciones logradas por el modelo BN obtuvieron mejor desempeño predictivo. El modelo CMP comparado con el modelo PGR en UD, solo logró ser más eficiente justo donde las predicciones de media tuvieron limitaciones, es decir, cuando se asumieron valores bajos para β_0. Tanto en ED como en UD, no se logró determinar una eficiencia marcada entre los modelos PGR y Poisson, ya que su ER fue muy cercana a 1 en todos los escenarios configurados, a pesar de que la proporción de casos de eficiencia fue a favor del modelo PGR en los niveles de UD menos fuertes, lo cual no se dio cuando se asumió un valor para ν de 5. Esto puede evidenciar la limitación que señala Sellers & Shmueli (2010*a*), respecto al modelo PGR, en donde dicha propuesta es una alternativa para ajustar datos de conteo en UD pero en un rango no tan flexible como lo hace el modelo CMP.

En cuanto a la comparación de los métodos de predicción propuestos para el modelo CMP, el único caso donde las predicciones de mediana fueron más eficientes que las logradas por la aproximación a la media, se dio en el escenario de OD más extrema en m1. En el resto de los casos incluso en UD, los valores ajustados obtenidos por medio de la media condicional fueron más eficientes, con una ligera tendencia a reducir la brecha en la calidad predictiva al aumentar el valor asumidos de β_0. Sellers & Shmueli (2010*a*) proponen el método de predicción de medianas como una alternativa generalizada para obtener valores ajustados, sin embargo, al parecer esta propuesta no logra ser más eficiente incluso cuando las predicciones de media son de baja calidad, es decir, cuando $\nu > 1$, por lo tanto, aún en UD en donde se presumía que las predicciones de mediana lograrían cierta ventaja, éstas no obtuvieron un desempeño predictivo suficiente para minimizar la baja calidad de la aproximación de la media condicional en este escenario de dispersión.

7. Aplicación

Por medio de dos casos de estudio reales aplicados dentro del campo de la ecología, se evaluó la calidad tanto de las predicciones de media como de mediana en las distribuciones que han sido comparadas. Como medidas de calidad predictiva se tuvieron en cuenta la raíz cuadrada del ECMP (RECMP) y la Mediana del Error Porcentual Absoluto (EPAMe), la cual es una medida utilizada por Sellers & Shmueli (2010*b*) ya que evita posibles indeterminaciones en el caso de la existencia de ceros (Armstrong & Collopy 1992).

El primer caso de estudio consiste en la predicción de la abundancia de una espe-

cie de interés en términos algunas variables ambientales, lo cual es un proceso de análisis esencial a la hora de tomar decisiones de manejo y conservación silvestre lugares donde no se ha hecho un muestreo previo. El segundo caso se desarrolla dentro del contexto del estudio de fauna silvestre ya que se trata de establecer la relación entre el tamaño del nido en aves (asociado con el número de huevos por nido) y las características morfológicas de las especies observadas y así configurar estrategias de producción sostenible y/o de conservación basadas en la caracterización del potencial de crecimiento de un conjunto de aves con características taxonómicas similares.

7.1. Estudio de abundancia en mango

La información tenida en cuenta para este estudio de caso está contenida en un conjunto de datos en donde se determinó la abundancia de especies de plantas leñosas en un total de 96 parcelas de muestreo distribuidas a través de un área que tiene una superficie de unos 22.000 km^2 y localizada al suroeste de la India.

El estudio de caso planteado consistió en estimar las existencias por hectárea que hay del árbol de mango (*Mangifera indica*) de forma silvestre dadas unas condiciones de ambientales definidas por el número de estratos del bosque y la duración en meses de la época de sequía. Dado esto, el conjunto de datos para este análisis contiene 96 observaciones correspondientes las mediciones en cada una de las parcelas de muestreo para cada una de las variables consideradas.

La información que presenta la tabla 7 es el resumen de las estimaciones de los parámetros de los diferentes modelos ajustados. En primer lugar se logró diagnosticar que la variable de conteo tiene una varianza mayor que la media y por lo tanto, el parámetro de dispersión estimado por el modelo CMP cae en el rango de OD. Al comparar las estimaciones se nota la gran diferencia que obtuvieron los coeficientes estimados y sus respectivos errores estándar en el modelo CMP luego de ser escalados. También se logró identificar que los errores estándar obtenidos por el modelo Poisson fueron los más bajos en contraste con los de las demás distribuciones.

La tabla 8 presenta las diferentes medidas que se adoptaron para evaluar la calidad

Tabla 7: *Resumen de las estimaciones en los modelos comparados en el estudio de abundancia.*

Modelo	β_0		β_1		β_2		Dispersión	
	$\hat{\beta}_0$	$\hat{\sigma}_{\hat{\beta}_0}$	$\hat{\beta}_1$	$\hat{\sigma}_{\hat{\beta}_1}$	$\hat{\beta}_2$	$\hat{\sigma}_{\hat{\beta}_2}$	$\hat{\phi}$	$\hat{\sigma}_{\hat{\phi}}$
Poisson	-5.9071	1.0324	0.8820	0.0735	0.7693	0.1642	–	–
CMP[1]	-13636.9261$\hat{\nu}$	4178.7752$\hat{\nu}$	1337.6726$\hat{\nu}$	331.7317$\hat{\nu}$	1283.5193$\hat{\nu}$	677.7827$\hat{\nu}$	0.0001	11.8296
BN	-5.8599	2.1009	0.9914	0.1865	0.7069	0.3340	0.5072	0.2142
PGR	-5.6810	2.0868	1.0419	0.2089	0.6540	0.3278	0.5844	0.1071

[1] Los coeficientes y sus errores estándar están divididos por $\hat{\nu}$ (excepto los de dispersión) ya que según Sellers & Shmueli (2010*a*) deben ser escalados para ser comparados con los de la regresión Poisson.

de las predicciones en el estudio de la abundancia de la especie *Mangifera indica*. Se nota que el modelo Poisson fue el que obtuvo el mejor desempeño tanto en la calidad de las predicciones de media como de mediana respecto a las demás distribuciones. También se resalta la calidad deficiente de las predicciones de media del modelo CMP debido a sus valores altos de RECMP, lo cual no sucede con las predicciones de mediana en donde si bien no presentaron un buen desempeño es considerable la diferencia que marca este método respecto al de los valores ajustados obtenidos mediante la aproximación a la media condicional propuesta para esta distribución.

Tabla 8: *Calidad de las predicciones en los modelos comparados en el estudio de abundancia.*

Medida	**Poisson**		**CMP**		**BN**		**PGR**	
	M	**Me**	**M**	**Me**	**M**	**Me**	**M**	**Me**
RECMP	4.227	4.180	3447.570	4.460	4.285	4.596	4.341	4.775
EPAMe	5.850	3.000	3449.608	1.000	6.808	0.955	7.465	1.000

Los resultados que se presentaron en el estudio de las predicciones de abundancia de la especie *Mangifera indica* reflejan las consecuencias de un caso de OD extrema ya que el parámetro de dispersión estimado por el modelo CMP al parecer tuvo problemas en su estimación. Este comportamiento fue muy común en el escenario más severo de OD asumido en las simulaciones, en donde con tamaños muestrales pequeños y en valores muy bajos asumidos para el intercepto dicho parámetro presentó problemas en su estimación obteniendo valores muy cercanos a cero. Al dividir los coeficientes y sus respectivos errores estándar por el valor de $\hat{\nu}$ tal como lo propone Sellers & Shmueli (2010*a*), se obtuvieron coeficientes sobrestimados en comparación con los obtenidos por las demás distribuciones.

Otro aspecto que se evidencia en los resultados es la subestimación de los errores estándar asociados a los coeficientes del vector de β en el modelo Poisson. Tal como lo señala Cameron & Trivedi (2003), una de las consecuencias de ajustar un modelo Poisson en OD es que los errores estándar tienden a ser subestimados generando coeficientes significativos cuando en realidad estos no lo son.

En cuanto a la calidad de las predicciones, es claro que el modelo Poisson es el que mejor desempeño logró. Debido a las estimaciones deficientes que produjo el modelo CMP, especialmente del parámetro de dispersión, las predicciones obtenidas a través de la aproximación a la media condicional presentaron valores muy bajos de desempeño predictivo, lo que contrastó con lo obtenido por el método de estimación de medianas. Esto corrobora los resultados de las simulaciones previas en donde definitivamente no es conveniente usar la aproximación de la media en casos de OD extrema y con valores muy bajos para el intercepto.

7.2. Estudio del tamaño del nido

En este estudio de caso se planteó estudiar la relación del tamaño del nido que es una medida asociada con el número de huevos por nido, respecto a la masa del huevo y al peso de la hembra en gramos. Para ello, se consideró realizar el estudio en especies de aves del orden de los Passeriformes. Este conjunto de datos tiene un total de 2061 observaciones para cada una de las tres variables evaluadas en este estudio.

Para este conjunto de datos se diagnosticó UD, dado que el parámetro de forma estimado por el modelo CMP fue de 1.815. La comparación de las estimaciones y sus respectivos errores estándar marca una similaridad entre aquellas obtenidas por el modelo Poisson y el modelo BN.

Tabla 9: *Resumen de las estimaciones en los modelos comparados en el estudio del tamaño del nido.*

Modelo	β_0		β_1		β_2		Dispersión	
	$\hat{\beta}_0$	$\hat{\sigma}_{\hat{\beta}_0}$	$\hat{\beta}_1$	$\hat{\sigma}_{\hat{\beta}_1}$	$\hat{\beta}_2$	$\hat{\sigma}_{\hat{\beta}_2}$	$\hat{\phi}$	$\hat{\sigma}_{\hat{\phi}}$
Poisson	1.2283	0.0185	-0.0312	0.0067	0.0013	0.0003	–	–
CMP[1]	$1.2925\hat{\nu}$	$0.0471\hat{\nu}$	$-0.0279\hat{\nu}$	$0.0049\hat{\nu}$	$0.0011\hat{\nu}$	$0.0002\hat{\nu}$	1.8151	0.0020
BN	1.2283	0.0185	-0.0312	0.0067	0.0013	0.0003	13903.3396	0.0585
PGR	1.2317	0.0181	-0.0329	0.0066	0.0013	0.0003	0.9895	0.0046

[1] Los coeficientes y sus errores estándar están divididos por $\hat{\nu}$ (excepto los de dispersión) ya que según Sellers & Shmueli (2010*a*) deben ser escalados para ser comparados con los de la regresión Poisson.

Tabla 10: *Calidad de las predicciones en los modelos comparados en el estudio del tamaño del nido.*

Medida	**Poisson**		**CMP**		**BN**		**PGR**	
	M	**Me**	**M**	**Me**	**M**	**Me**	**M**	**Me**
RECMP	1.602	1.638	1.603	1.639	1.602	1.638	1.601	1.639
EPAMe	0.328	0.400	0.327	0.400	0.328	0.400	0.326	0.400

También se logró detectar ciertas diferencias en los errores estándar del modelo CMP respecto a las otras distribuciones, especialmente en las estimaciones de β_0 y β_1. En la estimación del parámetro de dispersión del modelo BN se presentó el error estándar más elevado, mientras que el del modelo CMP fue el más bajo.

Las predicciones de media del número de huevos de aves de la orden Passeriformes evaluadas mediante la RECMP indican que la distribución PGR obtuvo el mejor desempeño mientras que en las predicciones de mediana los modelos Poisson y BN obtuvieron la mejor calidad predictiva en cuanto a la RECMP (Tabla 10).

El estudio del tamaño del nido en especies de aves del orden los Passeriformes, demostró ser coherente con los resultados presentados en los estudios de simulación. En primera instancia, el nivel de dispersión estimado por el modelo CMP indicó que los conteos del número de huevos tuvo una media mayor que la varianza

aunque no tan contrastantes ya que $\hat{\nu}$ fue muy cercano a 1. Considerando el alto número de observaciones, con el cual se puede asegurar la normalidad asintótica de las estimaciones; y que los interceptos estimados son bajos, se puede deducir que el desempeño del modelo CMP no fue tan bueno como el del modelo Poisson e incluso del modelo BN que logró una calidad en las predicciones de media similar.

A pesar del buen desempeño del modelo Poisson respecto al modelo CMP, para el conjunto de datos considerado en este estudio, la distribución PGR fue la que mejor calidad de predicción de medias obtuvo. Lo anterior, se puede explicar ya que esta distribución captura correctamente un rango parcial de UD (Sellers & Shmueli 2010*a*), por lo tanto, en un nivel bajo de UD se esperaría un buen desempeño en su calidad predictiva. De igual forma, vale la pena indicar que a partir de la tercera cifra decimal se empezaron a notar las diferencias entre las distribuciones contrastadas. Lo mismo sucedió con la estimación de medianas, pero ya las diferencias fueron a favor de los modelos BN y Poisson.

8. Conclusiones

La comparación entre la calidad predictiva del modelo CMP y el modelo Poisson fue evaluada mediante un estudio de simulación en el cual se tuvieron en cuenta factores como la intensidad de la dispersión y la variación del intercepto expresada mediante los modelos asumidos, en un marco donde el tamaño muestral fue lo suficientemente grande para así asegurar la normalidad asintótica de las estimaciones logradas por las diferentes propuestas para ajustar datos de conteo. Los resultados que fueron objeto de análisis en la anterior discusión arrojaron las conclusiones descritas a continuación.

Debido al desempeño limitado que tuvo el modelo CMP en cuanto a sus predicciones especialmente en el escenario más severo de OD cuando se asumió el valor más bajo de β_0, la ER entre este modelo y el modelo Poisson fue a favor de éste último. En esa misma intensidad de dispersión, en los únicos casos que el modelo CMP logró ser más eficiente fue en m2 y m3. Mientras que en las intensidades más cercanas a la ED, no se logró establecer una eficiencia marcada de un modelo respecto al otro. En UD, se detectó una eficiencia leve a favor del modelo Poisson a través de las diferentes intensidades de dispersión y específicamente en m1. En este sentido, se concluye que ya sea en OD o UD, el ajustar un modelo inadecuado, en este caso el modelo Poisson, en la mayoría de los casos no se incurre en una pérdida de la calidad predictiva. Si bien en el nivel de OD más fuerte, hubo dos escenarios que representaron una ligera eficiencia a favor de las predicciones del modelo CMP, esto no justifica su uso generalizado, ya que cuando se asumieron valores bajos para el intercepto fue claramente ineficiente, incluso comparado con el modelo BN.

La propuesta de Sellers & Shmueli (2010*a*) logró en algunos casos ser más eficiente y en otros no tanto, respecto a los modelos BN y PGR. El modelo CMP obtuvo una mayor eficiencia de las predicciones sobre las del modelo BN en el caso más

fuerte de OD, pero cuando la intensidad de OD fue más cercana a 1, y en la misma ED, las diferencias en desempeño predictivo ya no fueron tan notorias. Mientras que en UD, se lograron percibir eficiencias a favor del modelo PGR en m1.

Incluso en UD, donde se esperaba un mejor desempeño de las predicciones de mediana, (ya que según Minka et al. (2003) citado por Sellers & Shmueli (2010*a*) la aproximación de la ecuación (6) no es adecuada cuando $\nu > 1$) éstas no lograron la suficiente calidad predictiva para superar a la de las predicciones logradas por la aproximación de la media condicional de la distribución CMP. Particularmente, en el caso de OD extrema cuando se asumió el valor más bajo para el intercepto el comportamiento de las predicciones logrado por el método de predicción de medianas superó a la de las predicciones de media, por lo tanto, solo sería recomendable utilizar esta propuesta en este escenario.

Complementando la comparación de los modelos en cuanto a su desempeño predictivo, se evaluó la bondad de ajuste por medio del Criterio de Información de Akaike (AIC). Los resultados se pueden consultar en el anexo B. Allí se puede observar la capacidad que tiene el modelo CMP para explicar la relación funcional entre una respuesta de conteo y las variables predictoras, respecto a las otras alternativas contrastadas. Únicamente en el nivel de ED, el modelo Poisson presentó ventajas en cuanto a la bondad de ajuste, de resto en OD y UD, el modelo con mejor desempeño incluso en los niveles de dispersión más fuertes fue la CMP. Por lo tanto, se concluye que al parecer el modelo CMP no representa una mayor eficiencia a la hora de hacer predicciones en comparación con el modelo Poisson, sin embargo, éste posee ventajas en cuanto a su capacidad de explicar una relación funcional.

Recibido:
Aceptado:

Referencias

Armstrong, B. J. S. & Collopy, F. (1992), 'Error Measures For Generalizing About Forecasting Methods: Empirical Comparisons', **8**(1), 69–80.
*http://dx.doi.org/10.1016/0169-2070(92)90008-W

Cameron, A. C. & Trivedi, P. K. (2003), Essentials of Count Data Regression, *in* B. H. Baltagi, ed., 'A Companion to Theoretical Econometrics', Blackwell Publishing Ltd, pp. 331–348.
*http://dx.doi.org/10.1002/9780470996249.ch16

Cameron, A. & Trivedi, P. K. (1998), *Regression Analysis of Count Data*, Cambridge University Press, New York.

Dobson, A. J. (2002), *An introduction to generalized linear models*, 2nd ed. edn, Chapman & Hall/CRC.
*http://dx.doi.org/10.1002/sim.1493

Famoye, F. (1993), 'Restricted generalized poisson regression model', *Communications in Statistics - Theory and Methods* **22**(5), 1335–1354.
*https://doi.org/10.1080/03610929308831089

Francis, R., Geedipally, S. R., Guikema, S. D., Dhavala, S. S., Lord, D. & Larocca, S. (2012), 'Characterizing the Performance of the Conway-Maxwell Poisson Generalized Linear Model', *Risk Analysis* **32**(1), 167–183.
*https://doi.org/10.1111/j.1539-6924.2011.01659.x

Geedipally, S. R., Guikema, S. D., Dhavala, S. S. & Lord, D. (2008), Characterizing the Performance of the Bayesian Conway-Maxwell Poisson Generalized Linear Model, *in* A. S. Association, ed., 'Joint Statistical Meetings', p. 22.

Guikema, S. D. & Goffelt, J. P. (2008), 'A Flexible Count Data Regression Model for Risk Analysis', *Risk Analysis* **28**(1), 213–223.
*http://doi.wiley.com/10.1111/j.1539-6924.2008.01014.x

Hilbe, J. (2011), *Negative Binomial Regression*, 2nd ed. edn, Cambridge University Press.
*https://doi.org/10.1017/CBO9780511973420

Jowaheer, V. & Mamode, N. (2009), 'Estimating Regression Effects in Com Poisson Generalized Linear Model', *World Academy of Science, Engineering and Technology* **29**(1), 1040–1044.

Lord, D., Geedipally, S. R. & Guikema, S. D. (2010), 'Extension of the Application of Conway-Maxwell-Poisson Models: Analyzing Traffic Crash Data Exhibiting Underdispersion', *Risk Analysis* **30**(8), 1268–1276.
*http://dx.doi.org/10.1111/j.1539-6924.2010.01417.x

Lord, D., Guikema, S. D. & Geedipally, S. R. (2008), 'Application of the Conway-Maxwell-Poisson generalized linear model for analyzing motor vehicle crashes', *Accident Analysis and Prevention* **40**(3), 1123–1134.
*https://doi.org/10.1016/j.aap.2007.12.003

McCullagh, P. & Nelder, J. (1972), *Generalized linear models*, 2nd ed. edn, Chapman & Hall/CRC, New York.

Miller, J. (2007), Comparing Poisson, Hurdle and ZIP model fit under varying degrees of Skew and Zero-Inflation, Ph.d. thesis, University of Florida.

Minka, T. P., Shmueli, G., Kadane, J. B., Borle, S. & Boatwright, P. (2003), Computing with the COM-Poisson distribution, Technical report, Carnegie Mellon University, Pittsburgh, PA.
*http://repository.cmu.edu/statistics/170/

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
*https://www.R-project.org/

Sáez-Castillo, A. & Conde-Sánchez, A. (2013), 'A hyper-Poisson regression model for overdispersed and underdispersed count data', *Computational Statistics & Data Analysis* **61**, 148–157.
*http://dx.doi.org/10.1016/j.csda.2012.12.009

Sellers, K. F., Borle, S. & Shmueli, G. (2012), 'The COM-Poisson model for count data: A survey of methods and applications', *Applied Stochastic Models in Business and Industry* **28**(2), 104–116.
*http://dx.doi.org/10.1002/asmb.918

Sellers, K. F. & Shmueli, G. (2010*a*), 'A flexible regression model for count data', *Annals of Applied Statistics* **4**(2), 943–961.
*http://www.jstor.org/stable/29765537

Sellers, K. F. & Shmueli, G. (2010*b*), 'Predicting Censored Count Data with COM-Poisson Regression', *SSRN Electronic Journal* p. 18.
*http://dx.doi.org/10.2139/ssrn.1702845

Shmueli, G., Minka, T., Kadane, J., Borle, S. & Boatwright, P. (2005), 'A Useful Distribution for Fitting Discrete Data: Revival of the Conway-Maxwell-Poisson Distribution', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **54**(1), 127–142.
*https://doi.org/10.1111/j.1467-9876.2005.00474.x

Winkelmann, R. (2008), *Econometric Analysis of Count Data*, 5th ed. edn, Springer-Verlag, Berlin.

Anexos

A. Distribución empírica de β_0 en un modelo CMP

A manera de simulación previa se determinó la distribución empírica del coeficiente β_0 en un modelo CMP. Se puede notar en la figura 6, que en pequeños tamaños muestrales la distribución del coeficiente evaluado tiene una forma asimétrica y que a medida que va incrementando n la asimetría tiende a centralizarse, es decir, que los coeficientes estimados son menos sesgados, en especial cuando el tamaño muestral es de 1000 (Figura 7).

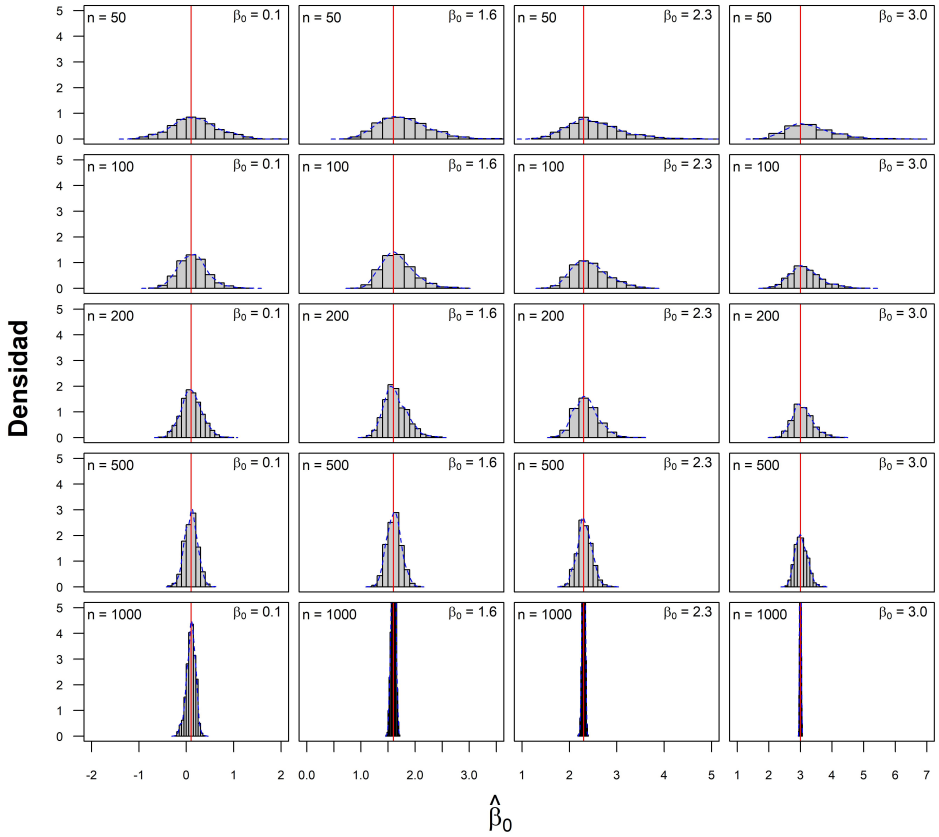

Figura 6: *Gráfico de distribución empírica de β_0 en un modelo de regresión COM-Poisson en diferentes tamaños muestrales.*

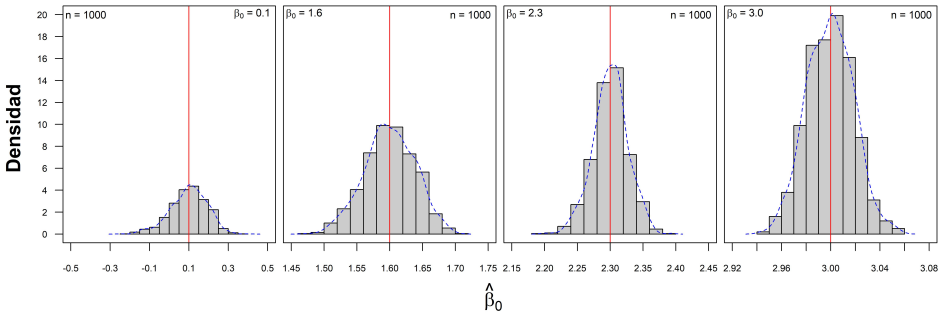

Figura 7: *Gráfico de distribución empírica de β_0 en un modelo de regresión COM-Poisson con $n = 1000$.*

B. Evaluación de la bondad de ajuste

Tabla 11: Criterio de Información de Akaike promedio con $n = 1000$.

ν	Modelo	*AIC*			
		m1	m2	m3	m4
	Poisson	2862.57	5938.47	7063.43	8151.06
0.25	**CMP**	2754.79	5205.56	6042.69	6879.13
	BN	2756.01	5239.09	6111.37	7011.29
	Poisson	2780.35	4963.18	5897.52	6752.51
0.50	**CMP**	2741.41	4780.93	5655.68	6485.71
	BN	2742.42	4791.97	5674.22	6507.11
	Poisson	2742.20	4683.20	5388.81	6067.10
0.75	**CMP**	2734.63	4651.69	5352.39	6029.63
	BN	2735.09	4654.46	5355.55	6032.63
	Poisson	2722.06	4393.98	5115.94	5828.02
1.00	**CMP**	2723.03	4395.03	5116.97	5829.06
	PGR	2723.05	4395.03	5116.98	5829.07
	Poisson	2692.08	4139.74	4819.80	5492.22
1.50	**CMP**	2662.67	4074.89	4749.92	5419.55
	PGR	2664.84	4079.37	4753.16	5421.90
	Poisson	2500.21	3883.10	4621.72	5228.84
2.50	**CMP**	2322.12	3589.59	4314.13	4914.71
	PGR	2344.98	3608.64	4325.66	4922.80
	Poisson	2250.09	3694.62	4418.88	5029.17
5.00	**CMP**	1579.36	2924.71	3626.86	4227.10
	PGR	1751.59	2969.73	3650.32	4241.97