La unión hace la fuerza: modelos de distribución de especies integrando diferentes fuentes de datos

Javier Fernández-López[1,2], Pelayo Acevedo[3], Olivier Gimenez[1]

(1) CEFE, Université Montpellier, CNRS, EPHE, IRD, Montpellier (Francia)

(2) Universidad Complutense de Madrid, Madrid (España)

(3) IREC, CSIC-UCLM-JCCM, Ciudad Real (España)

Autor para correspondencia: Javier Fernández-López (javfer05@ucm.es)

Palabras clave

Proceso de Puntos de Poisson; modelos jerárquicos; verosimilitud conjunta; NIMBLE

Keywords

Poisson Point Process; hierarchical models; joint-likelihood; NIMBLE

La distribución y la abundancia de las especies son dos métricas básicas implicadas directamente en la gestión y la conservación de las poblaciones silvestres (Tellería, 2012). En las últimas décadas, los modelos de distribución (SDM por sus siglas en inglés) se han convertido en una de las herramientas más utilizadas para el estudio de la ocupación (presencia/ausencia) y abundancia (número de individuos) de las especies (Elith y Leathwick, 2009). Estos modelos relacionan la presencia o abundancia de individuos con una serie de covariables predictoras, permitiendo i) conocer la relación de esas covariables con las poblaciones y ii) realizar predicciones espacialmente explícitas sobre la ocupación o abundancia de las especies en función de esas covariables. Entre las causas que han motivado el auge de estas herramientas se podrían mencionar el aumento de la disponibilidad de cartografía ambiental (Hijmans et al., 2005), el desarrollo de nuevos algoritmos para ajustar este tipo de modelos (Phillips et al., 2006) o la aparición de numerosas plataformas con bases de datos accesibles sobre la presencia o abundancia de las especies (p.ej. GBIF). Recientemente, el crecimiento y diversidad de estas plataformas ha motivado la investigación de las aproximaciones analíticas para la integración de diferentes fuentes de datos en un único modelo que refleje la distribución o abundancia de las especies (*integrated species distribution models*, en adelante iSDM). Estas metodologías pueden ser de gran utilidad en programas de monitoreo de poblaciones a gran escala o en especies para las que no se dispone de mucha información, pues permiten combinar unos pocos muestreos sistemáticos de gran calidad y llevados a cabo por especialistas con bases de

datos más oportunistas, como las procedentes de la ciencia ciudadana (Miller et al., 2019). El uso de la verosimilitud conjunta (en adelante *joint-likelihood*) es una de las metodologías más utilizadas para ajustar este tipo de modelos y se asienta principalmente en dos marcos teórico-prácticos: el Proceso de Puntos de Poisson y la modelización jerárquica.

Proceso de Puntos de Poisson

El Proceso de Puntos de Poisson es un modelo matemático que estudia la distribución de elementos (puntos) en un espacio continuo, asumiendo que siguen una distribución de Poisson cuyo parámetro λ_i significaría el número esperado de puntos por unidad de área en la localidad i. Además de resultar una forma natural de representar individuos en el territorio (puntos en el espacio), este proceso tiene dos propiedades que lo hacen muy atractivo a la hora de utilizarlo como modelo subyacente en el estudio de la distribución y abundancia de las especies. La primera es que partiendo de una distribución de individuos (puntos en un espacio continuo), es muy fácil e intuitivo *discretizar* ese espacio (colocando encima una cuadrícula, por ejemplo) y obtener dos métricas de gran interés: la abundancia (conteo del número de puntos por cuadrícula) y la ocupación (cuadrículas con al menos un punto) de la especie (Figura 1). Esto significa que la ocupación y la abundancia pueden ser entendidas como realizaciones discretizadas del mismo proceso de puntos que ocurre en un espacio continuo y que, por tanto, están relacionadas entre sí. Esta propiedad permite combinar distintos tipos de datos (presencias/ausencias, solo presencias y conteos) en un mismo marco estadístico. La segunda propiedad tiene que ver con la resolución. El proceso de puntos ocurre, tal y como sucede con la distribución de individuos en la naturaleza, en un espacio continuo y por lo tanto es independiente de la resolución. Son las métricas *abundancia* y *ocupación* las que son dependientes de la resolución, ya que previamente hemos discretizado el espacio en cuadrículas cuyo tamaño va a condicionar el número o la presencia/ausencia de puntos en su interior (Figura 1). Por ello, utilizar el Proceso de Puntos de Poisson como modelo subyacente nos permite reconciliar diferentes bases de datos a diferentes resoluciones, ya que asumimos que todas provienen de la misma distribución de individuos en un espacio continuo.

Modelos jerárquicos

La modelización jerárquica es un marco estadístico que se caracteriza por diferenciar entre una variable latente (o variable de estado), que suele ser el objeto de nuestro interés pero que no puede ser estudiada directamente, y unas observaciones, que son muestreos de esa variable latente y que aunque dependen de ella, emergen a través de un proceso observacional y por lo tanto pueden estar distorsionadas (Kéry y Royle, 2015). En nuestro caso, la variable latente es la abundancia o la ocupación de la especie, mientras que las observaciones pueden ser muestreos de conteos directos, muestreos de detecciones/no-detecciones, registros de presencias oportunistas o cualquier otra fuente de datos (Fernández-López et al., 2022). Todas estas observaciones estarán relacionadas con

la variable latente, pero además se verán afectadas por un proceso observacional en el que influirán otros factores. Uno de los factores que se intenta controlar más a menudo es la detectabilidad imperfecta (falsos negativos: no detectar un individuo/especie cuando en realidad sí está presente en un determinado sitio), aunque en el marco de la modelización jerárquica se pueden acomodar igualmente otras situaciones, como el sesgo de muestreo o los falsos positivos (Louvrier et al., 2018). La ventaja de los modelos jerárquicos es que permiten incluir aquellas covariables que afectan a los dos procesos (ecológico para la variable latente y observacional para la fuente de datos) de forma independiente y mediante ajustes habituales de los modelos lineales generalizados (Figura 2).

Modelos de distribución de especies integrados: *joint-likelihood*

La aproximación de *joint-likelihood* para el ajuste de iSDM utiliza como modelo subyacente el proceso de puntos de Poisson, siendo la variable latente el número de individuos por unidad de área (Miller et al., 2019). Partiendo de este modelo unificador, se superponen los diferentes juegos de datos con sus respectivos procesos (submodelos) observacionales, pero compartiendo una misma variable latente de la cual provienen (Figura 2). Al compartir la misma variable latente, las fuentes de datos comparten información sobre el proceso ecológico que las origina, teniendo en cuenta a su vez los diferentes procesos observacionales específicos de los cuales emanan cada una de ellas. Es importante destacar que para ajustar un iSDM mediante *joint-likelihood* se requiere una comprensión profunda de cada uno de los procesos observacionales que originan las diferentes fuentes de datos. Si alguno de los submodelos observacionales no es adecuado por cualquier motivo, la información compartida por esa fuente de datos será errónea, por lo que el resultado del modelo integrado estará así mismo sesgado (Ahmad Suhaimi et al., 2021).

La Figura 3 representa de forma esquemática los componentes y resultados de un ejemplo de iSDM a partir de datos simulados (ver Apéndice 1 para el código completo). En este modelo se combinaron dos fuentes de datos: i) una resultante de un muestreo en el que se visitaron 4 veces un total de 10 localidades, contabilizando todos los individuos que fueron capaces de detectarse (“conteos repetidos”); y ii) una base de datos de ciencia ciudadana que contenía un total de 223 registros de presencia oportunistas (“registros oportunistas”). A partir de estos datos, se ajustaron tres modelos, uno con cada una de las fuentes de datos por separado y un tercer modelo integrando ambas fuentes. Para todos los modelos se siguió una estrategia de modelización jerárquica similar, en la que se asumió un Proceso de Puntos de Poisson como modelo subyacente para la variable latente (abundancia de individuos), y se ajustó un submodelo observacional para cada una de las fuentes de datos siguiendo el esquema de la Figura 2. Todos los modelos se ajustaron mediante inferencia Bayesiana en NIMBLE (Valpine et al., 2017). En la Figura 3 se muestran las fuentes de datos, las covariables predictoras y los principales resultados de los análisis realizados. Los patrones de abundancia obtenidos a partir de los diferentes modelos fueron similares a la abundancia simulada (Figura 3, segunda fila), mostrando una mayor linealidad para el modelo a partir de registros oportunistas y el iSDM. Sin embargo, el iSDM obtuvo estimas de abundancia totales más cercanas a

la abundancia simulada que el resto de modelos (N abundancia simulada = 25,801; N Conteos SDM = 14,656.; N Registros oportunistas = 17,107; N iSDM = 18,003). La precisión de los parámetros estimados fue mayor en el caso del iSDM (Apéndice 1).

Los modelos integrados permiten combinar la información de calidad obtenida a partir de muestreos sistemáticos, pero que normalmente son escasos y de ámbito local por su elevado coste, con otro tipo de datos oportunistas, como los derivados de la ciencia ciudadana, que suelen ser muy abundantes y ampliamente distribuidos pero que a menudo se ven gravemente afectados por sesgos de diferente tipo (Dorazio, 2014; Isaac et al., 2020; Merow et al., 2022). El desarrollo de los modelos de distribución integrados permitirá el aprovechamiento de las diferentes bases de datos de una forma más eficiente, cobrando especial importancia en programas de monitorización a gran escala y en aquellas especies menos estudiadas o con mayores problemas de conservación (Tellería, 2012).

Contribución de los autores

Javier Fernández-López: Conceptualización, análisis formal y redacción - borrador original. Pelayo Acevedo: Conceptualización y redacción - revisión y edición. Olivier Gimenez: Conceptualización, validación y redacción - revisión y edición.

Agradecimientos

Queremos agradecer a Verónica Cruz, Julen Astigarraga, Elena Quintero y en general al Grupo de Ecoinformática de la Asociación Española de Ecología Terrestre por organizar y promover los “Seminarios Ecoinformáticos”, a partir de uno de los cuales se originó esta nota. Las conversaciones con Valentin Lauret motivaron reflexiones que ayudaron a una mejor evaluación y comunicación de los modelos de esta nota. JF-L está actualmente financiado por un contrato postdoctoral Margarita Salas (fondos de la Unión Europea – NextGenerationEU) a través de la Universidad Complutense de Madrid.

REFERENCIAS

Ahmad Suhaimi, S.S., Blair, G.S., Jarvis, S.G. 2021. Integrated species distribution models: A comparison of approaches under different data quality scenarios. *Diversity and Distributions* 27: 1066-1075.

Dorazio, R.M. 2014. Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography* 23: 1472-1484.

Elith, J., Leathwick, J.R. 2009. Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution and Systematics* 40: 677-697.

Fernández-López, J., Blanco-Aguiar, J.A., Vicente, J., Acevedo, P. 2022. Can we model distribution of population abundance from wildlife–vehicles collision data? *Ecography* 2022: e06113.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology: A Journal of the Royal Meteorological Society* 25: 1965-1978.

Isaac, N.J., Jarzyna, M.A., Keil, P., Dambly, L.I., Boersch-Supan, P.H., Browning, E., Freeman, S.N. et al. 2020. Data integration for large-scale models of species distributions. *Trends in ecology & evolution* 35: 56-67.

Kéry, M., Royle, J.A. 2015. *Applied Hierarchical Modeling in Ecology: Analysis of Distribution, Abundance and Species Richness in R and BUGS*. Elsevier, Londres, Reino Unido.

Louvrier, J., Chambert, T., Marboutin, E., Gimenez, O. 2018. Accounting for misidentification and heterogeneity in occupancy studies using hidden Markov models. *Ecological modelling* 387: 61-69.

Merow, C., Galante, P.J., Kass, J.M., Aiello-Lammens, M.E., Babich Morrow, C., Gerstner, B.E., Grisales Betancur, V. et al. 2022. Operationalizing expert knowledge in species' range estimates using diverse data types. *Frontiers of Biogeography*.

Miller, D.A., Pacifici, K., Sanderlin, J.S., Reich, B.J. 2019. The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution* 10: 22-37.

Phillips, S.J., Anderson, R.P., Schapire, R.E. 2006. Maximum entropy modeling of species geographic distributions. *Ecological modelling* 190: 231-259.

Tellería, J.L. 2012. *Introducción a la conservación de las especies*. Hernández, V. (ed.),. Tundra Ediciones, Valencia, España.

Valpine, P. de, Turek, D., Paciorek, C.J., Anderson-Bergman, C., Lang, D.T., Bodik, R. 2017. Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics* 26: 403-413.

PIES DE FIGURA

Figura 1. El mismo proceso de puntos (A y D) y sus métricas derivadas a diferentes resoluciones: B y C a 1x1 km; E y F a 4x4 km. Los gráficos B y E muestran el número de puntos por cuadrícula (abundancia). Los gráficos C y F muestran la presencia de al menos un punto en cada una de las cuadrículas (ocupación). Mientras que los patrones observados a partir de las métricas derivadas (abundancia B y E; ocupación C y F) varían dependiendo de la resolución de la malla, el patrón de puntos permanece constante (A y D).

Figure 1. The same point process (A and D) and their derived metrics at different resolutions: B and C at 1x1 km; E and F at 4x4 km. Graphs B and E show the number of points per grid (abundance). Graphs C and F show the presence of at least one point in each of the grids (occupancy). While the patterns observed from the derived metrics (abundance B and E; occupancy C and F) vary depending on the resolution of the grid, the pattern of points remains constant (A and D).

Figura 2. Esquema de un modelo jerárquico integrando dos fuentes de datos diferentes (ver Apéndice 1). A) La variable latente o de estado (abundancia de individuos) puede ser modelizada como un Proceso de Puntos de Poisson discretizado, cuya función de intensidad (λ) vendrá determinada por dos covariables (cobertura de bosque y altitud). B) La primera fuente de datos, conteos sistemáticos en 15 localidades repetidos cuatro veces por cada localidad, puede ser modelizada con una distribución Binomial (detección - no detección de cada uno de los individuos en cada conteo y localidad). C) La segunda fuente de datos, registros oportunistas a partir de ciencia ciudadana, puede modelizarse como un Proceso de Puntos de Poisson adelgazado con una distribución de Bernoulli (para cada celda, detección - no detección de al menos un individuo).

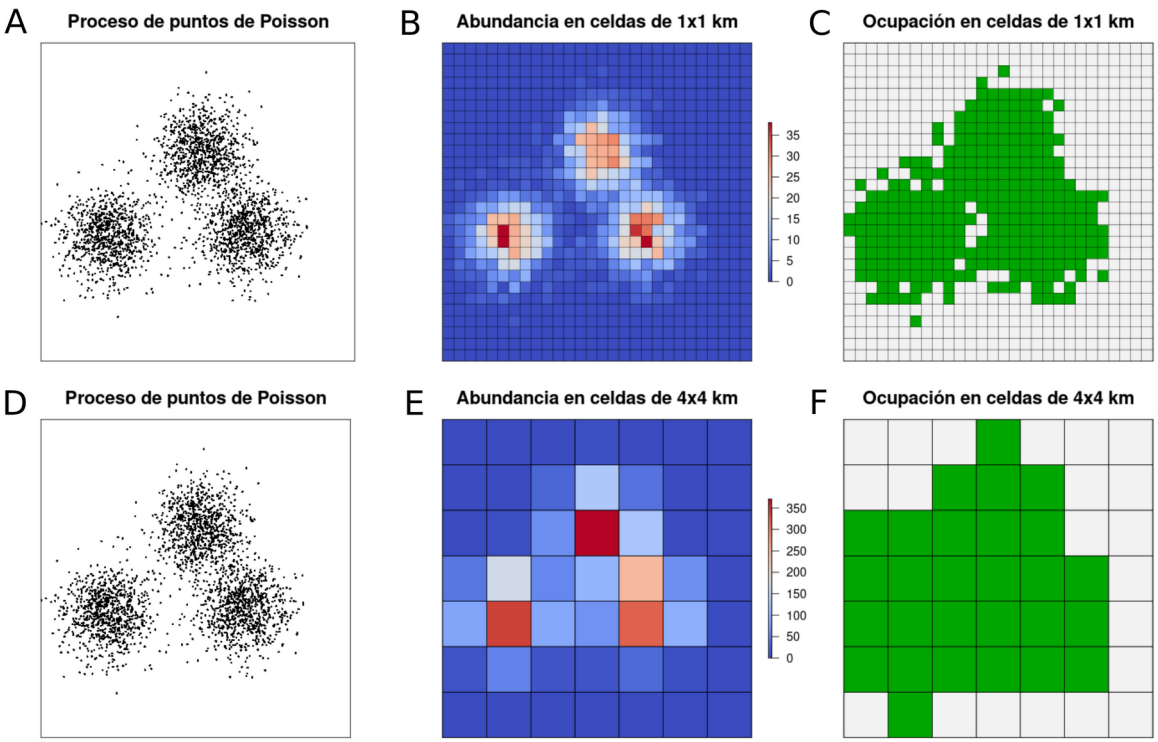
Figure 2. Diagram of a hierarchical model integrating two different data sources (see Appendix 1). A) The latent or state variable (abundance of individuals) can be modeled as a discretized Poisson Point Process, whose intensity function (λ) will be determined by two covariates (forest cover and altitude). B) The first source of data, systematic counts in 15 locations repeated four times at each location, can be modeled with a Binomial distribution (detection - non-detection of each individual in each count and location). C) The second data source, opportunistic records from citizen science, can be modeled as a thinned Poisson Point Process with a Bernoulli distribution (for each cell, detection - non-detection of at least one individual).

Figura 3. Fuentes de datos, covariables predictoras y resultados del ejemplo desarrollado en el Apéndice 1. Se muestra el patrón de distribución de la abundancia simulada (Abundancia), y las predicciones para cada uno de los modelos: a partir de conteos repetidos (Conteos SDM), a partir de registros oportunistas (Registros SDM) y el modelo integrando ambas bases de datos (iSDM).

Figure 3. Data sources, predictor covariates and results of the example developed in Appendix 1. We show the simulated abundance distribution pattern (Abundancia), and the predictions for each of the

models: from repeated counts (Conteos SDM), from opportunistic records (Registros SDM) and the model integrating both databases (iSDM).

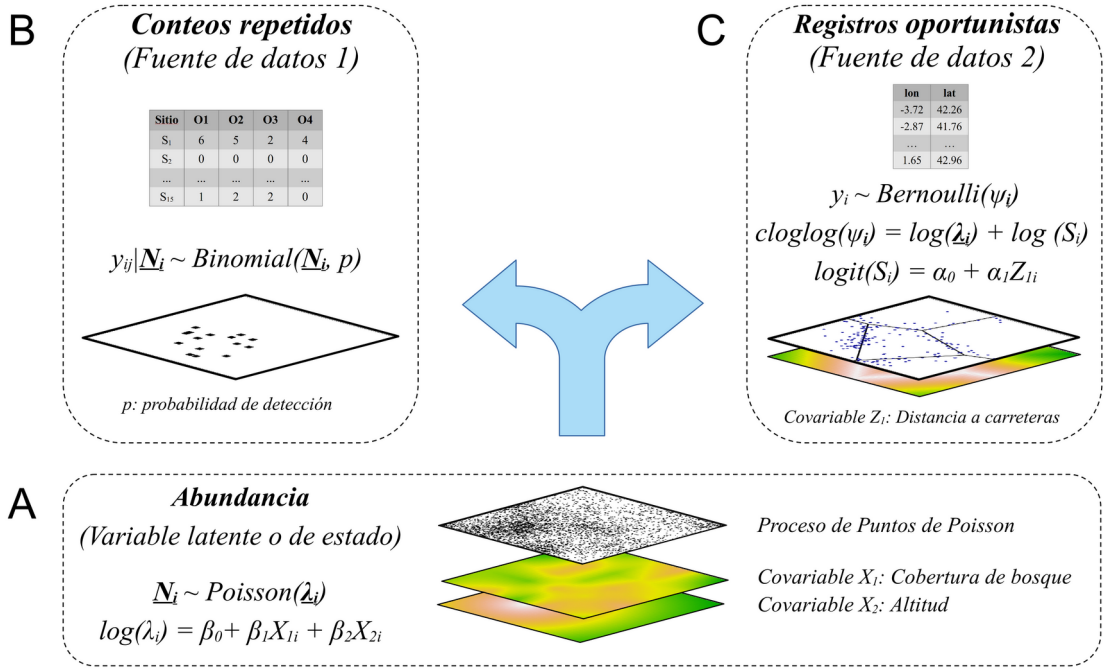
196 FIGURA 1



197

198 *Figura 1. El mismo proceso de puntos (A y D) y sus métricas derivadas a diferentes resoluciones: B y*
199 *C a 1x1 km; E y F a 4x4 km. Los gráficos B y E muestran el número de puntos por cuadrícula*
200 *(abundancia). Los gráficos C y F muestran la presencia de al menos un punto en cada una de las*
201 *cuadrículas (ocupación). Mientras que los patrones observados a partir de las métricas derivadas*
202 *(abundancia B y E; ocupación C y F) varían dependiendo de la resolución de la malla, el patrón de*
203 *puntos permanece constante (A y D).*

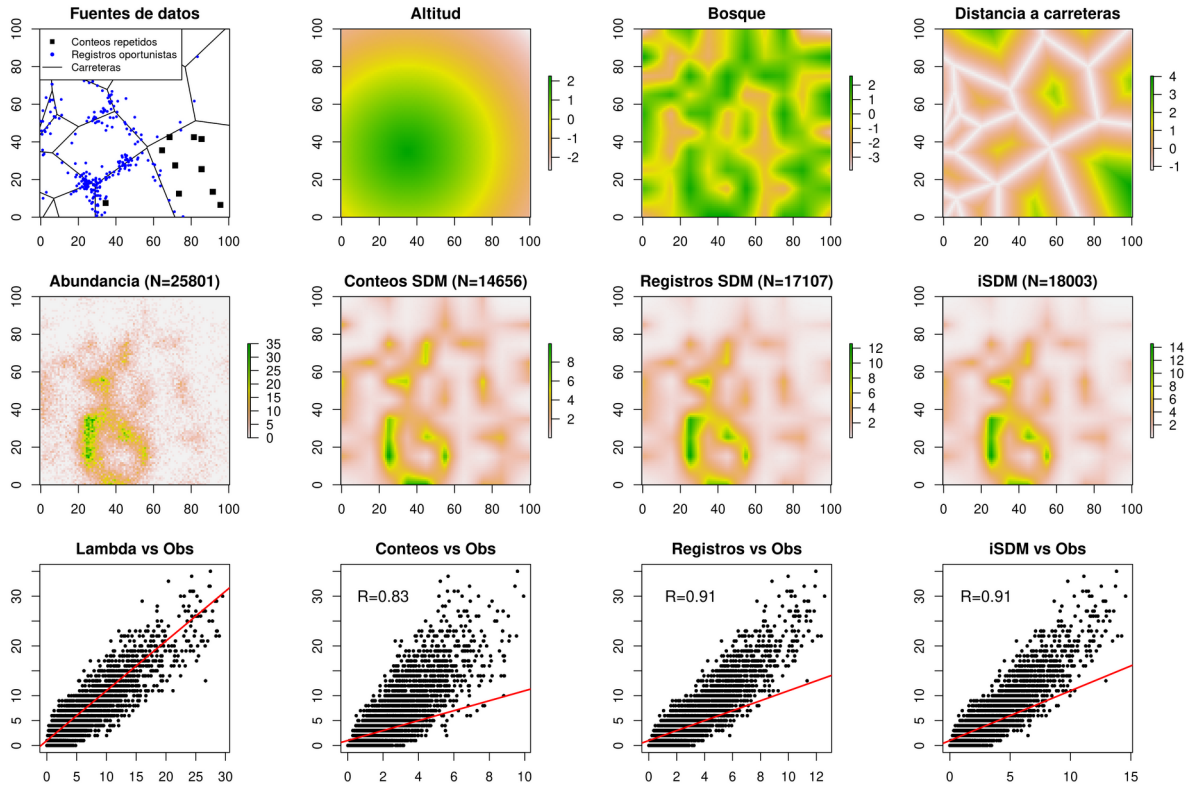
204 FIGURA 2



205

206 *Figura 2. Esquema de un proceso de modelado jerárquico integrando dos fuentes de datos diferentes*
207 *(ver Apéndice 1). A) La variable latente o de estado (abundancia de individuos) puede ser*
208 *modelizada como un Proceso de Puntos de Poisson discreto, cuya función de intensidad (lambda)*
209 *vendrá determinada por una serie de covariables (cobertura de bosque y altitud). B) La primera*
210 *fuente de datos, conteos sistemáticos de 15 localidades repetidos cuatro veces en cada localidad,*
211 *puede ser modelizada con una distribución Binomial (detección - no detección de cada uno de los*
212 *individuos en cada conteo y localidad). C) La segunda fuente de datos, registros oportunistas a partir*
213 *de ciencia ciudadana, puede modelizarse como un Proceso de Puntos de Poisson adelgazado con*
214 *una distribución de Bernoulli (para cada celda, detecto - no detecto al menos un individuo).*

215 FIGURA 3



216

217 *Figura 3. Fuentes de datos, covariables predictoras y resultados del ejemplo desarrollado en el*
218 *Apéndice 1. Se muestra el patrón de distribución de la abundancia simulada (Abundancia), y las*
219 *predicciones para cada uno de los modelos: a partir de conteos repetidos (Conteos SDM), a partir de*
220 *registros oportunistas (Registros SDM) y el modelo integrando ambas bases de datos (iSDM).*