Hombre, Ciencia y Tecnología ISSN: 1028-0871 Vol. **25, No. 2, abr-jun. pp. 67-74, 2021**

Análisis estadístico de la especie de abejas *Megachile armaticeps*.
Statistical analysis of the solitary bee's species *Megachile armaticeps*

Autores: Lic. Mairelis Videaux-Aguilar[1], Ing. José Rolando Dupuy-Parra[2], Lic. Lidysse Thaureaux-Jhones[1].

Organismo: Universidad de Guantánamo, Guantánamo, Cuba[1], Centro de Aplicaciones Tecnológicas para el Desarrollo Sostenible (CATEDES), Guantánamo, Cuba[2].

E-mail: denise@cug.co.cu ; mairelis.videaux@nauta.cu

Resumen

Las abejas son fundamentales en el funcionamiento de los ecosistemas tropicales y son reconocidas por su importancia económica. En Cuba, la apifauna está compuesta en su mayoría por abejas solitarias, lo cual le confiere a este grupo un valor especial desde el punto de vista de su conservación. El objetivo de este trabajo es crear un modelo de regresión logística que permita predecir la presencia o ausencia de la especie *Megachile armaticeps*, a través de diferentes variables relacionadas con las características ambientales de su hábitat. Los resultados se obtuvieron aplicando el modelo de regresión logística binaria en el software SPSS en su versión 21. Se obtuvo que la estacionalidad de la temperatura, la precipitación del mes más húmedo, la precipitación del mes más seco y la distancia a la costa son las propiedades climáticas que más influyen en su distribución.

Palabras clave: abejas; correlación; distribución; hábitat; regresión

Abstract

Bees are fundamental at functioning of tropical ecosystems and are recognized for their economic importance. In Cuba, the apifauna is composed mostly of solitary bees, which gives this group a special value from the point of view of its conservation. The objective of this work is to create a logistic regression model that allows predicting the presence or absence of the *Megachile armaticeps* species, through different variables related to the environmental characteristics of its habitat. The results were obtained by applying the binary logistic regression model in the SPSS software version 21. It was obtained that temperature seasonality, precipitation of the wettest month, precipitation of the driest month and distance to the coast are the climatic properties that most influence its distribution.

Keywords: bees; correlation; distribution; habitat; regression

Introducción

El 20 de mayo del 2018 se realizó la primera celebración del Día Mundial de las Abejas. Esta celebración nace de una propuesta realizada por la República de Eslovenia en el 2016, la cual fue aprobada en la 40ª reunión de la conferencia de la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO), y proclamada por la Organización de las Naciones Unidas (ONU) en el 2017 (FAO, 2019). Podemos decir también que el análisis de los patrones de distribución de las especies a nivel mundial ha aportado información valiosa sobre el impacto potencial del cambio climático en la distribución de las especies, específicamente en la capacidad que presentan las poblaciones de migrar para soportar estos cambios previstos.

En Cuba las abejas son consideradas de importancia capital para el crecimiento y la calidad de la producción desde el inicio de los programas de investigación científica para la apicultura, actividad dedicada a la crianza de las abejas y a prestarles los cuidados necesarios con el objetivo de obtener y consumir los productos que son capaces de elaborar y recolectar. Varios son los factores que amenazan a las abejas: la pérdida del hábitat, las prácticas de la agricultura industrializada, el uso de plaguicidas, y los impactos del cambio climático (Videaux, 2015).

Por todo lo antes mencionado, el Instituto de Ecología y Sistemática de Cuba realiza una investigación de 18 especies de abejas solitarias, entre las que se encuentra la especie *Megachile armaticeps*, a la cual se le miden variables de tipo climático en su hábitat, es por ello que surge la necesidad de encontrar un modelo que permita predecir la presencia-ausencia de la especie de abejas *Megachile armaticeps*, así como, los factores ambientales que más afectan a esta especie, lo cual constituye el objetivo de la presente investigación.

Materiales y métodos

Recolección de los datos originales:

La información primaria se tomó de las colecciones del Museo Nacional de Historia Natural de Cuba y el Instituto de Ecología y Sistemática, ambos pertenecientes al Ministerio de Ciencia Tecnología y Medioambiente de Cuba. La base de datos aportó registros de 18 especies de abejas solitarias que representa la distribución de éstas en Cuba y las características ambientales del territorio.

Figura 1: Especie de abeja *Megachile armaticeps*.

La base de datos de las abejas solitarias *Megachile armaticeps*, representa la distribución de esta especie y contiene 500 observaciones y 24 variables, una de las cuales es binaria. La variable binaria se denomina mapa binario, la cual, representa la variable dependiente dicotómica y sólo toma dos valores, 0 cuando no se encuentra dicha especie en el área y 1 cuando si se encuentra. Las 23 variables restantes denominadas variables independientes o covariables son continuas, las cuales miden en general las características ambientales del área. Estas variables son: temperatura media anual (Bio1), variación diurna promedio (Bio2), isotermalidad (Bio3), estacionalidad de la temperatura (Bio4), temperatura máxima del mes más cálido (Bio5), temperatura mínima del mes más frío (Bio6), variación anual de temperatura (Bio7), temperatura media del trimestre más húmedo (Bio8), temperatura media del trimestre más seco (Bio9), temperatura media del trimestre más cálido (Bio10), temperatura media del trimestre más frío (Bio11), precipitación anual (Bio12), precipitación del mes más seco (Bio13), precipitación del mes más húmedo (Bio14), estacionalidad de las precipitaciones (Bio15), precipitación del trimestre más húmedo (Bio16), precipitación del trimestre más seco (Bio17), precipitación del trimestre más cálido (Bio18), precipitación del trimestre más frío (Bio19), modelo digital de elevación (MDE), pendiente (PEN), distancia a la costa (DISTC) y el índice topográfico (TOPO).

Modelo de regresión logística binaria:

Los modelos de Regresión Logística Binaria son fórmulas estadísticas en las cuales se desea conocer la relación entre una variable dependiente cualitativa, dicotómica y una o más variables explicativas independientes ya sean cualitativas o cuantitativas. El objetivo fundamental es encontrar el mejor ajuste del modelo con el menor número de parámetros y describir la relación entre la variable respuesta y un conjunto de variables explicatorias independientes.

El modelo de Regresión Logística Binaria se define como:

$Y = E\ (Y/X) + e$

Donde:

$E(Y/X) = \pi(X) = \frac{\exp\ (\alpha+\beta X)}{1+\exp\ (\alpha+\beta X)}$.

e tiene distribución Bernoulli con media cero y varianza $\pi(X)[1 - \pi(X)]$.

La transformación de $\pi(X)$, conocida como transformación logit o logito, se define mediante la función $g(X) = ln\left[\frac{\pi(X)}{1-\pi(X)}\right] = \alpha + \beta X$ Hosmer y Lemeshow en el 2013 plantearon que es una función lineal en los parámetros, continua y sus valores se encuentran en toda la recta numérica $-\infty < g(X) < +\infty$.

Clásicamente, la exactitud de una prueba diagnóstica se ha evaluado en función de dos características: la sensibilidad y la especificidad. Sin embargo, éstas varían en función del criterio elegido como punto de corte entre la población. Una forma más global de conocer la calidad de la prueba en el espectro completo de puntos de corte es mediante el uso de curvas ROC (*Receiver Operating Characteristic*). Siempre que el problema y el resultado de la prueba diagnóstica puedan plantearse en términos de dicotomía (presencia o ausencia, positivo o negativo), la exactitud de la prueba puede definirse en función de su sensibilidad y especificidad.

El punto de corte es un valor límite que permite resumir los resultados en dos categorías: positivo y negativo, presencia o ausencia de cierta característica.

La sensibilidad (*S*) de una prueba diagnóstica es la probabilidad de que la prueba indique como positivo a aquel que realmente lo es.

La especificidad (*E*) de una prueba indica la probabilidad de que la prueba clasifique como negativo a aquel que realmente lo es. No existe ninguna manera teórica de medición de estas dos características de una prueba. El único procedimiento es el experimental, sometiendo a un grupo clasificado mediante un método diagnóstico de referencia exacto e independiente a la prueba que se quiere estudiar.
Una curva ROC es una representación gráfica para una prueba de clasificación binaria según varía el umbral de discriminación, es decir, es el resultado de representar en un eje de coordenadas los puntos (x, y) dados por (1-E, S) para cada punto de corte.
El criterio de información de Akaike (AIC) elige dado un conjunto de modelos candidatos para los datos, el modelo que tiene el valor mínimo.
El criterio de información bayesiano (BIC) es similar al AIC excepto que el término de penalidad es más grande. En estas circunstancias, BIC penaliza modelos complejos más fuerte que AIC, favoreciendo la selección de modelos más simples.

Resultados y Discusión

Ajuste de los datos originales al modelo

El estudio descriptivo permitió constatar de forma general que la base de datos tiene más observaciones de ausencia (444) que de presencia (56), por lo que se espera un modelo que posea mayor predicción en la ausencia. Las variables presentan un alto rango excepto distancia a la costa (DISTC) que posee un máximo de 0,57499999 debido a que los valores tomados por esta variable son muy pequeños, además, distancia a la costa y la pendiente (PEN) poseen un mínimo de 0. La desviación típica es alta, distinta de la unidad, evidenciando la alta dispersión en las observaciones, con excepción de DISTC que tiene una desviación menor a 1, la media en todas las variables es distinta de 0. Debido a los problemas identificados a través del análisis descriptivo es necesario estandarizar las variables con el propósito de generar un modelo con los datos de mejor calidad posible.

Tabla 1: Análisis descriptivo de las covariables.

Covariables	Rango	Mínimo	Máximo	Media	Desv. típ.	Varianza
Bio1	78	186	264	249,51	10,100	102,010
Bio2	48	76	124	103,03	8,885	78,939
Bio3	12	58	70	64,11	2,449	5,996
Bio4	777	1444	2221	1878,03	193,493	37439,362
Bio5	89	253	342	325,13	12,310	151,543
Bio6	80	116	196	166,02	9,769	95,426
Bio7	52	126	178	158,68	10,462	109,458
Bio8	76	201	277	265,90	10,486	109,957
Bio9	84	165	249	223,68	10,512	110,503
Bio10	80	205	285	270,93	10,454	109,294
Bio11	76	165	241	222,76	10,085	101,712
Bio12	1250	825	2075	1343,05	176,957	31313,689
Bio13	190	120	310	212,60	29,423	865,732
Bio14	95	9	104	29,56	14,921	222,643

Bio15	43	30	73	59,19	8,624	74,379
Bio16	409	327	736	547,19	78,639	6184,060
Bio17	286	51	337	110,28	45,852	2102,437
Bio18	523	179	702	491,00	103,415	10694,651
Bio19	423	51	474	113,77	54,723	2994,615
DISTC	,57499999	,00000	,57499999	,18804046	,12799849	,016
MDE	1393,000	1,0000	1394,0000	110,1760	164,17247	26952,602
PEN	1233	0	1233	108,78	188,799	35645,047
TOPO	1226	171	1397	626,61	211,147	44582,923

Las variables explicatorias independientes están relacionadas con las características ambientales del hábitat de esta especie, por ello, se realiza un análisis de correlación de las variables antes de aplicar la regresión logística. Las variables analizadas presentan alta correlación con un nivel de significación de 0.05 donde, Bio5 y Bio15 se correlacionan con todas. A continuación, se muestra un resumen de las variables que no están correlacionadas:

- Bio1 con Bio3, Bio4 y DISTC.
- Bio3 con Bio1, Bio8, Bio10, MDE, PEN.
- Bio4 con Bio1, Bio2 y DISTC.
- Bio6 con Bio14 y Bio19.
- Bio7 con Bio13.
- Bio16 con Bio2, Bio14 y PEN.
- Bio18 con Bio8, Bio17, MDE, PEN y TOPO.
- DISTC con Bio1, Bio4, Bio8, Bio9, Bio10, Bio11, Bio12, MDE y TOPO.

Al aplicar la técnica de regresión logística mediante el software SPSS versión 21 a la base de datos estudiada, teniendo en cuenta que $\pi(X)$ es la probabilidad de que la especie se encuentre en dicha área y $1-\pi(X)$ la probabilidad de que no se encuentre en el área, se obtuvo el siguiente modelo:

Tabla 2: Estimaciones de los parámetros del Modelo.

Modelo								
Mapa Binario[a]	B	Error típ.	Wald	Gl	Sig.	Exp(B)	Intervalo de confianza al 95% para Exp(B)	
							Límite inferior	Límite superior
Intersección	-5.794	.746	60.338	1	.000			
TOPO	-.721	.286	6.353	1	.012	.486	.277	.852
Bio13	2.257	.430	27.523	1	.000	9.552	4.111	22.194
Bio4	-3.779	.597	40.108	1	.000	.023	.007	.074
Bio14	-1.419	.312	20.626	1	.000	.242	.131	.446
DISTC	-2.591	.508	26.010	1	.000	.075	.028	.203
La categoría de referencia es: 0.								

En la tabla de clasificación podemos comprobar que nuestro modelo tiene una especificidad del 98.2% y una sensibilidad del 83.9%, siendo su capacidad predictiva del 96.6%. Se corrobora que el modelo posee mayor predicción de ausencia.

Tabla 3: Tabla de clasificación del Modelo.

Observado	Pronosticado		
	0	1	Porcentaje correcto
0	436	8	98.2%
1	9	47	83.9%
Porcentaje global	89.0%	11.0%	96.6%

En el ajuste del modelo se obtuvo que AIC=107.969 y BIC=133.257 por lo que se puede decir que el modelo es adecuado, y con el contraste de la razón de verosimilitud se verifica que no podemos rechazar la hipótesis de que los datos se ajustan al modelo supuesto.
Debido a la existencia de correlación entre algunas variables independientes del modelo obtenido, se incluyeron en el logit las interacciones con el propósito de mejorar la capacidad predictiva del mismo, y el modelo resultante fue:

Tabla 4: Estimaciones de los parámetros del Modelo con interacciones.

Modelo con interacciones								
Mapa Binario[a]	B	Error típ.	Wald	Gl	Sig.	Exp(B)	Intervalo de confianza al 95% para Exp(B)	
							Límite inferior	Límite superior
Intersección	-7.645	1.116	46.968	1	.000			
Bio13	4.144	.690	36.073	1	.000	.016	.004	.061
Bio4	-5.013	.789	40.412	1	.000	150.288	32.044	704.860
Bio14	-1.233	.360	11.704	1	.001	3.432	1.693	6.956
DISTC	-2.878	.582	24.414	1	.000	17.775	5.676	55.662
Bio4 * Bio13	1.316	.290	20.612	1	.000	.268	.152	.473
La categoría de referencia es: 0.								

En la tabla de clasificación podemos comprobar que nuestro modelo tiene una especificidad del 98.4% y una sensibilidad del 85.7%, siendo su capacidad predictiva del 97%.

Tabla 5: Tabla de clasificación del Modelo con interacciones.

Observado	Pronosticado		
	0	1	Porcentaje correcto
0	437	7	98.4%
1	8	48	85.7%
Porcentaje global	89.0%	11.0%	97.0%

En el ajuste del modelo se obtuvo que AIC=88.764 y BIC=114.052 por lo que se puede decir que el modelo es adecuado, y con el contraste de la razón de verosimilitud se verifica que no podemos rechazar la hipótesis de que los datos se ajustan al modelo supuesto.

De manera general, los modelos poseen un índice de precisión global y un grado de acuerdo con la observación real muy buenos como se muestra en el gráfico:

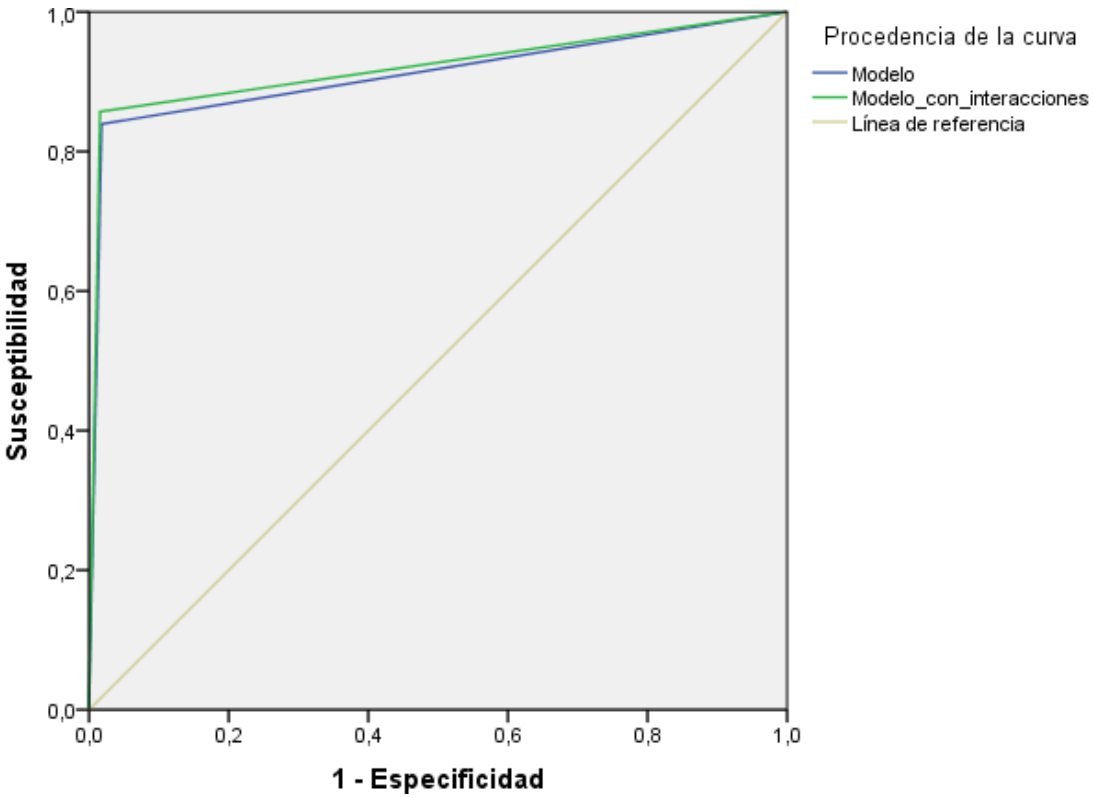

Figura 2: Curva ROC de los modelos.

El modelo con mayor capacidad predictiva es el Modelo con interacciones como se esperaba, ya que en dicho modelo se analiza la significación e influencia de todas las covariables en el pronóstico de la presencia o ausencia de la especie, los estadísticos informan que el mismo, es el mejor modelo en cuanto a la bondad de ajuste, mayor área bajo la curva ROC y el que más explica la proporción de varianza de la variable dicotómica. Este modelo plantea que las covariables más significativas en la predicción son: estacionalidad de la temperatura (Bio4), precipitación del mes más seco (Bio13), precipitación del mes más húmedo (Bio14) y distancia a la costa (DISTC).

De la información anterior se puede corroborar que la distribución de esta especie depende tanto de la variabilidad ambiental como de su tolerancia. Algunas de las abejas solitarias de esta especie suelen ser muy selectivas en cuanto al hábitat que utilizan y tienden a concentrarse en puntos donde las condiciones son especialmente favorables por la incidencia de la estacionalidad de las temperaturas en el área, así como, la distancia de esta región a la costa, precipitación del mes más seco, precipitación del mes más húmedo y la topografía si se excluye la interacción medio-ambiental. Esto puede ocurrir a lo largo del año o en épocas específicas. Cuanto más se aproximan las condiciones ambientales a las tolerancias mínima y máxima de un organismo, menor será el número de individuos. La estacionalidad de las temperaturas, la distancia a la costa, precipitación del mes más húmedo y la topografía influyen negativamente en su presencia, por lo que son parámetros a tener en cuenta en la elección del área, ya que, restringen su distribución por la baja tolerancia de la misma a estos factores ambientales. La precipitación del mes más seco influye positivamente.

Conclusiones

A partir de los resultados obtenidos mediante la aplicación de las técnicas estadísticas aplicadas a la base de datos de las abejas solitarias *Megachile armaticeps*, se determinó que los modelos obtenidos tienen mayor capacidad para pronosticar la ausencia de la especie. Existe alta correlación entre las covariables. El modelo de mayor capacidad predictiva es el que posee las interacciones, con un 97%, el cual tiene una especificidad del 98.4% y una sensibilidad del 85.7%. Las variables que influyen en el estudio de esta especie de abejas son: estacionalidad de la temperatura, precipitación del mes más húmedo, precipitación del mes más seco y distancia a la costa. La interacción más fuerte se establece entre estacionalidad de la temperatura y precipitación del mes más seco.

Referencias bibliográficas

Cruz, D. (2015). Distribución y evaluación de los grados de amenaza de abejas solitarias (Hymenoptera: Apoidea). Memoria para optar al Título de Master en Ciencias, Facultad de Biología, Universidad de La Habana, Cuba.

García, C. (2012). Estimación del modelo logístico mixto: revisión y nueva propuesta. Memoria para optar al Título de Master en Ciencias, Escuela de Ciencias y Humanidades, Universidad EAFIT Medellín, Colombia.

Genaro, J. (2008). Origins, composition and distribution of the bees of Cuba (Hymenoptera: Apoidea: Anthophila). Insecta Mundi. 583. Recuperado de https://digitalcommons.unl.edu/insectamundi/583

Hosmer, D. W., Lemeshow, S. & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.). New Jersey: John Wiley and Sons.

Minerva, A. B., Martínez, A. (2011). Análisis Multinivel de movimientos migratorios: consideraciones y estrategia. Investigación Operacional, vol. 32, núm. 1, pp. 20-29.

Organización de las Naciones Unidas para la Alimentación y la Agricultura. (2019). Acción mundial de la FAO sobre servicios de polinización para una agricultura sostenible. Recuperado de http://www.fao.org/pollination/world-bee-day/es/

Pérez, V. (2012). Los modelos multinivel en el análisis de factores de riesgo de sibilancias recurrentes en lactantes. Memoria para optar al Título de Doctor en Ciencias, Universidad de Murcia.

Videaux, M. (2015). Modelo estadístico para el pronóstico de la presencia o ausencia de la especie Agapostemon viridulus en Cuba. Memoria para optar al Título de Licenciado en Matemáticas, Facultad de Matemática-Computación, Universidad de Oriente, Cuba.

Fecha de recibido: 29 oct. 2020
Fecha de aprobado: 15 ene. 2021