



Estimación de la disposición a pagar en el Brunas usando modelos Logit aplicados a datos balanceados con el algoritmo Rose

Estimation of the willingness to pay in the Brunas using Logit models applied to balanced data with the Rose algorithm

Carlos López de Castilla Vásquez^{1*}, Jaime Porras Cerrón¹, Gesabel Villar Morales¹

¹ Universidad Nacional Agraria La Molina, Lima, Perú.

* Autor de correspondencia: clopez@lamolina.edu.pe
* <https://orcid.org/0000-0001-7724-5522>

Recepción: 05/06/2021; Aceptación: 15/12/2021; Publicación: 30/06/2023

Resumen

Uno de los métodos de valoración económica más conocidos es denominado valoración contingente. En su aplicación, los encuestados responden una pregunta sobre su disposición a pagar (DAP) por el uso y preservación de un bien público. Sin embargo, este método tiene una seria limitación relacionada con la posibilidad de obtener sesgo hipotético, debido a la falta de honestidad en las respuestas, lo que produce un desbalance en el tamaño de los grupos de respuestas negativas (no dispuestas a pagar) y positivas (dispuestas a pagar). Utilizando datos de valoración contingente del Bosque Reservado de la Universidad Nacional Agraria de la Selva (BRUNAS), se calculó la DAP utilizando diferentes modelos de regresión logística binaria. El primer modelo se obtuvo considerando el conjunto original de datos con grupos no balanceados. El segundo modelo se estimó luego de aplicar el algoritmo ROSE, que permitió balancear los grupos a través de la generación de datos sintéticos. Después de aplicar ROSE, el modelo obtenido logró una estimación más realista de la DAP y de su error estándar lo que resultó en intervalos de confianza con menor amplitud en comparación con el modelo inicial.

Palabras clave: modelo de regresión logística binaria; bootstrap; algoritmo ROSE; valoración contingente; disposición a pagar (DAP), Bosque Reservado de la Universidad Nacional Agraria de la Selva (BRUNAS).

Forma de citar el artículo: López, C., Porras, J., & Villar, G. (2023). Estimación de la disposición a pagar en el Brunas usando modelos Logit aplicados a datos balanceados con el algoritmo Rose. *Natura@economía*, 8(1), 1-10. <http://dx.doi.org/10.21704/ne.v8i1.1697>

DOI: <http://dx.doi.org/10.21704/ne.v8i1.1697>

© El autor. Este artículo es publicado por la revista *Natura@economía* de la Universidad Nacional Agraria La Molina. Este es un artículo de acceso abierto, distribuido bajo los términos de la Licencia Creative Commons Atribución 4.0 Internacional. (<https://creativecommons.org/licenses/by/4.0/>) que permite Compartir (copiar y redistribuir el material en cualquier medio o formato), Adaptar (remezclar, transformar y construir a partir del material) para cualquier propósito, incluso comercialmente.

Abstract

One of the best-known methods of economic valuation is called contingent valuation. In their application, respondents answer a question about their willingness to pay (WTP) for the use and preservation of a public good. However, this method has a serious limitation related to the possibility of hypothetical bias, due to the lack of honesty in the answers, which produces an imbalance in the size of the groups of negative (unwilling to pay) and positive (willing to pay) responses. Using contingent valuation data from the Reserved Forest of the National Agrarian University of the Jungle (BRUNAS), the DAP was calculated using different binary logistic regression models. The first model was obtained considering the original set of data with unbalanced groups. The second model was estimated after applying the ROSE algorithm, that allows to balance the groups through the generation of synthetic data. After applying the ROSE, the obtained model achieved a more realistic estimate of the DAP and its standard error which resulted in confidence intervals with less amplitude compared to the initial model.

Keywords: logistic regression model, bootstrap, ROSE algorithm, contingent valuation, willingness to pay (WTP), Reserved Forest of the National Agrarian University of the Jungle (BRUNAS).

1. Introducción

El método de valoración contingente, en su conceptualización, intenta concretar el vínculo de los seres humanos y los ecosistemas (Bouwma *et al.*, 2018) y permite la descomposición del valor total del bien en sus atributos (Melo *et al.*, 2020) brindando un enfoque útil para la valoración económica de bienes públicos carentes de un mercado, como los bienes ambientales, lo que resulta en mayores ventajas para los tomadores de decisiones y gestores en política ambiental. Es un método directo basado en preferencias que las propias personas declaran (McFadden & Train, 2017). El trabajo pionero de Bishop & Heberlein (1979), luego Hanemann (1991), más recientemente MINAM (2021) y Vilela *et al.* (2022) señalaron que bajo este método se somete a los consumidores a experimentos controlados aplicados mediante un cuestionario que permite observar sus preferencias haciendo uso de un mercado hipotético, que es un escenario donde se provee el bien o servicio ecosistémico a valorar y donde se podrá observar las preferencias en las respuestas de las personas.

La pregunta se realiza consultando al entrevistado cuanto está dispuesto a pagar como máximo por el bien que se está valorando. Melo *et al.* (2020) señalaron que dado que el ejercicio

de valoración se basa en un escenario simulado, es factible que se presente “sesgo hipotético” (Rakotonarivo *et al.*, 2016); es decir, que las preferencias expresadas por los encuestados podrían diferir ante circunstancias económicas reales o que no respondan con total honestidad sobre el valor para su disposición a pagar.

Hanemann (1991) indicó que el método de valoración contingente considera una elección binaria, donde se escoge entre la posibilidad de pagar o no pagar por la preservación de un bien, reflejando así sus verdaderas preferencias. Según el argumento de Hanemann los individuos son clasificados como:

- Los que están dispuestos a pagar el monto propuesto
- Los que no están dispuestos a pagar el monto propuesto

Esta clasificación conlleva a tener grupos altamente desbalanceados debido al sesgo hipotético, que ocurre cuando el entrevistado no asimila el mercado simulado o presume que el pago propuesto no se efectúa como un desembolso real. Ogrodowczyk (2003), indicó que el sesgo hipotético genera una sobre o subvaloración en la estimación de la DAP.

Tudela (2017) indicó que, si hay dos alternativas en el proceso de valoración contingente, generalmente se aplica el modelo de regresión logística. La variable respuesta binaria es:

$$y_i = \begin{cases} 0 & \text{si el encuestado } i \text{ no acepta pagar} \\ 1 & \text{si el encuestado } i \text{ acepta pagar} \end{cases}$$

tal que $\Pr(y_i=1) = \pi_1$ y $\Pr(y_i=0)=1-\pi_1$, donde π_1 es la probabilidad que el encuestado i acepta pagar por el uso y preservación del bien. Menardi & Torelli (2014) indicaron que un desbalance en los grupos definidos por la variable respuesta, debido al sesgo hipotético, puede dar lugar a inconvenientes en el proceso de estimación y evaluación de los modelos ya que los resultados se obtendrían básicamente con los datos del grupo mayoritario (los encuestados que no aceptan pagar), mientras que los datos del grupo minoritario (los encuestados que aceptan pagar) tendrían muy poca participación en el proceso de estimación de los parámetros del modelo. Labandeira (2007) recomendó el uso de escenarios mucho más realistas para los encuestados, dejando de lado las ambigüedades o generalizaciones.

Uno de los enfoques populares para resolver el problema de desequilibrio en los grupos es sobremuestrear el grupo minoritario o submuestrear el grupo mayoritario. Sin embargo, estos enfoques tienen sus debilidades. En el método básico de sobremuestreo, busca duplicar algunas observaciones al azar del grupo minoritario por lo que esta técnica no agrega ninguna información nueva. Por otro lado, el método de submuestreo se lleva a cabo eliminando algunas observaciones al azar del grupo mayoritario, asumiendo el costo de eliminar parte de la información de los datos originales.

Lunardon *et al.* (2014) proporcionaron un marco adecuado para tratar simultáneamente con los problemas de estimación y evaluación de la precisión en modelos de regresión logística binaria en presencia de grupos desequilibrados en la distribución de

la variable respuesta. Su algoritmo, conocido como ROSE (Random Over-Sampling Examples), combina el sobremuestreo y submuestreo para generar observaciones artificiales usando el método Bootstrap. Se considera un conjunto de entrenamiento T_n , de tamaño n , cuya observación es (x_i, y_i) para $i = 1, \dots, n$. Los grupos definidos por la variable respuesta son $\{c_0, c_1\}$ y se asume que los valores observados para las variables predictoras corresponden a un vector aleatorio cuya función de densidad desconocida es $f(x)$. Además, $n_j < n$ donde n_j es el tamaño del grupo c_j , para $j = 0, 1$. El procedimiento para generar una observación artificial usando ROSE se presenta a continuación:

1. Seleccionar $y^* = c_j$ con probabilidad π_j .
2. Seleccionar $(x_i, y_i) \in T_n$ al que (y_i, y^*) .
3. Seleccionar yx^* de $k_{H_j}(\cdot, x_i)$ siendo k_{H_j} una distribución de probabilidad centrada en x_i y cuya matriz de covarianzas es H_j .

Esencialmente, se elige desde el conjunto de entrenamiento una observación que pertenece a uno de los dos grupos y se genera una nueva observación dentro de su vecindario cuya ancho estará determinado por H_j . Se considera, además, que k_{H_j} corresponde a una distribución de probabilidad simétrica y unimodal.

Una vez seleccionado $y^* = c_j$, se tiene que:

$$\Pr(x|y = c_j) = \sum_{i=1}^{n_j} \frac{1}{n_j} \Pr(x|x_i) = \sum_{i=1}^{n_j} \frac{1}{n_j} K_{H_j}(x - x_i)$$

de tal forma que la generación de nuevas observaciones del grupo corresponde a la generación de datos a partir de la estimación de la densidad de kernel de $f(x|c_j)$.

Dentro de todas las posibles alternativas, se considera el kernel Gaussiano con matriz de suavización diagonal $H_j = \text{diag}(h_1^{(j)}, h_2^{(j)}, h_d^{(j)})$, que minimiza la media integrada del error cuadrado asintótica (AMISE) bajo el supuesto que la verdadera densidad condicional corresponde a la distribución normal. Bajo esta consideración, se tiene:

$$h_q^{(j)} = \left(\frac{4}{(d+2)n} \right)^{\frac{1}{d+4}} \hat{\sigma}_q^{(j)}$$

para $q=1, \dots, d$ y $j=1, 2$. Además, $\hat{\sigma}_q^{(j)}$ es la desviación estándar muestral de las observaciones de la q -ésima variable, que pertenecen al grupo c_j .

En el Perú se han realizado recientemente estudios de aplicación del método de valoración contingente, usando modelos de regresión logística binaria: Albarracín-Valdivia *et al.* (2021), Medalla (2020), Lavado (2021), Quispe (2020), Ramírez (2022), etc. En el contexto internacional, Garzon (2013) presentó una revisión de literatura de aplicaciones de valoración contingente en América Latina y el Caribe y adicionalmente se han consultado a Hernández *et al.* (2019) en México y Gualli Guamán (2022) en Ecuador. En todos los estudios mencionados, se realizó la estimación directa de la DAP usando modelos de regresión con respuesta binaria, sin mencionar los riesgos que supone el trabajar con grupos de respuesta no balanceados.

El proceso de evaluación de los modelos de regresión logística binaria se realiza a través de la construcción de una matriz de confusión, presentada en la Tabla 1, que permite comparar las predicciones del modelo con las respuestas obtenidas por los encuestados. A partir de esta comparación se pueden observar los errores y aciertos en las predicciones que pueden ser utilizados en la construcción de indicadores. En los modelos de respuesta binaria se pueden obtener cuatro posibles resultados en esta matriz. Cuando el modelo predice correctamente que una persona aceptará el pago propuesto se dice que se ha obtenido un Verdadero Positivo (VP); sin embargo, si el modelo predice incorrectamente que una persona aceptará el pago se dice que se ha obtenido un Falso Positivo (FP). De igual modo, se tiene un Verdadero Negativo (VN) si el modelo predice correctamente que una persona no aceptará el pago, y será Falso Negativo (FN) si el modelo predice incorrectamente que una persona no aceptará el pago.

Tabla 1. Matriz de confusión

Predicción según el modelo	Valores observados		Total
	$y_i=0$	$y_i=1$	
$\hat{y}_i=0$	VN	FN	VN+FN
$\hat{y}_i=1$	FP	VP	FP+VP
Total	VN+FP	FN+VP	

La Precisión se define como la proporción de individuos correctamente clasificados que hay en el grupo de grupo de encuestados que el modelo predice como dispuestos a realizar el pago.

$$\text{Precisión} = \text{VP} / (\text{FP} + \text{VP})$$

La Sensitividad o Recall se define como la proporción de elementos correctamente clasificados en el grupo de encuestados que realmente están dispuestos a realizar el pago.

$$\text{PSensitividad} = \text{VP} / (\text{VP} + \text{FN})$$

Este indicador se define como la media armónica de Precisión y Recall. Se recomienda utilizar cuando se trabaja con grupos no balanceados.

$$\text{F score} = 2[(\text{Precisión} \times \text{Recall}) / (\text{Precisión} + \text{Recall})]$$

La curva característica operativa (ROC) es una representación gráfica para diferentes valores de la tasa de verdaderos positivos en función de la tasa de falsos positivos del modelo de clasificación. El área debajo de esta curva (AUC) es un indicador de la capacidad predictiva del modelo de regresión logística estimado.

El objetivo del presente trabajo de investigación fue estimar la DAP por el servicio de ecoturismo del Bosque Reservado de la Universidad Nacional Agraria de la Selva (BRUNAS) en Tingo María, usando los datos obtenidos por Ruiz (2007), considerando un modelo de regresión logística binaria estimado luego de realizar un proceso de balance de grupos a través del algoritmo ROSE (Random Over-Sampling Examples).

2. Materiales y métodos

Se utilizaron los datos del estudio desarrollado por Ruiz (2007), obtenidos mediante la aplicación de encuestas a un total de 250 personas mayores de 18 años que radican en el lugar y que conocían el bosque, se les preguntó por su DAP para acceder al servicio de ecoturismo. Los entrevistados respondieron a la siguiente pregunta:

¿Usted estaría dispuesto a pagar la cantidad (2, 10, 20 o 50) PEN, como único pago incluido en su recibo de luz del mes siguiente, que le dará derecho a ingresar cuatro veces durante los 3 meses siguientes a fin de disfrutar del servicio de ecoturismo que brinda el BRUNAS?

Las variables consideradas en el estudio se presentan en la Tabla 2.

La población estaba formada por todos los visitantes del BRUNAS de Tingo María. El estudio se realizó en el año 2007 considerando una muestra de 250 visitantes, obteniéndose 175 respuestas positivas, que aceptaron pagar el monto propuesto, y 75 respuesta negativas, que rechazaron pagar el monto propuesto.

Para la estimación de la DAP se considera un modelo de regresión logística múltiple:

$$\log \frac{\pi_i}{1-\pi_i} = \mathbf{x}_i^T \boldsymbol{\beta} \quad (1)$$

donde \mathbf{x}_i^T es un vector de valores observados para las variables predictoras y $\boldsymbol{\beta}$ es el vector de coeficientes de regresión. El modelo estimado con las respuestas obtenidas por los encuestados es:

siendo $X_1, X_2, \dots, X_{(p-1)}$ variables predictoras y A es el monto propuesto de pago por el uso y conservación del bien público. Una estimación de la DAP, a partir del modelo anterior, es:

$$DAP = \int_0^{\infty} \frac{1}{1 + \exp\{\mathbf{x}_{p-1}^T \hat{\boldsymbol{\beta}}_{p-1} + \hat{\beta}_p A\}} dA \quad (2)$$

Tabla 2. Identificación de las variables

Variable respuesta	Descripción
DISPAGAR	Indica el rechazo o la aceptación a pagar el monto propuesto del encuestado. <ul style="list-style-type: none"> • 0 = No, rechazo pagar el monto propuesto • 1 = Sí, acepto pagar el monto propuesto
MONPROP	Indica el monto propuesto por el encuestador, el cual será aceptado o no como pago (en PEN) para el acceso al parque 2, 10, 20 o 50 soles.
MFLO	Indica si el motivo de la visita es observar la flora del parque. <ul style="list-style-type: none"> • 0 = No • 1 = Sí
VISIT	Indica si el entrevistado visitaría otra vez el parque. <ul style="list-style-type: none"> • 0 = No • 1 = Sí
LUGN	Indica el lugar de procedencia del entrevistado. <ul style="list-style-type: none"> • 0 = Foráneo • 1 = Local
EDAD	Indica la edad del entrevistado en años.
SEXO	Indica el sexo del entrevistado. <ul style="list-style-type: none"> • 0 = Hombre • 1 = Mujer
EDUC	Indica en grado de instrucción del entrevistado. <ul style="list-style-type: none"> • 0 = No universidad • 1 = Universidad
SITL	Indica la situación laboral del entrevistado. <ul style="list-style-type: none"> • 0 = Desempleado • 1 = Empleado
TRABH	Indica el número de personas que trabajan en el hogar del entrevistado.
ING	Indica el ingreso mensual del entrevistado en soles.

donde $\bar{\mathbf{x}}_{(p-1)}$ es el vector correspondiente a la media de cada variable predictorada obtenida a partir de las respuestas brindadas por los encuestados.

La estimación por intervalo para la DAP se obtuvo usando el método Bootstrap (Efron & Tibshirani, 1994). Se trata de una técnica de remuestreo que permite obtener muchas muestras con reemplazo, a partir de una muestra inicial, calculando en cada una de ellas la cantidad de interés y obteniendo

Tabla 3. Proporción de respuestas de aceptación o rechazo pagar el monto propuesto

Disposición a pagar	Cantidad	Proporción
Sí = 1	175	0.7
No =	75	0.3
Total	250	1.0

$$\log \frac{\hat{\pi}_i}{1 - \hat{\pi}_i} = \hat{\beta}_0 + \hat{\beta}_1 X_{1,t} + \hat{\beta}_2 X_{2,t} + \dots + \hat{\beta}_{p-1} X_{p-1,t} + \hat{\beta}_p A_t - \mathbf{X}_{p-1,t} \hat{\beta}_{p-1} + \hat{\beta}_p A_t$$

información adicional para el proceso de inferencia. Se pueden obtener intervalos de confianza para el parámetro de interés θ usando los resultados obtenidos con el método Bootstrap. El intervalo de confianza del $(1-\alpha)\%$ para el parámetro $\theta = g(F)$ por el método de percentiles, está dado por:

$$g_{\alpha/2}^{-1}(\hat{F}) \leq \theta \leq g_{1-\alpha/2}^{-1}(\hat{F}) \tag{3}$$

donde, $g_{\alpha}^{-1}(\hat{F})$ es el percentil que corresponde a una probabilidad α , obtenido desde la distribución del estimador Bootstrap $\hat{\theta} = g(\hat{F})$.

El proceso de generación de muestras sintéticas a través del algoritmo ROSE considera un conjunto de entrenamiento T_n , de tamaño n , cuya observación es (x_i, y_i) para $i = 1, \dots, n$. Los grupos definidos por la variable respuesta son $\{c_0, c_1\}$ y se asume que los valores observados para las variables predictoras corresponden a un vector aleatorio cuya función de densidad desconocida es $f(x)$. Además, $n_j < n$ donde n_j es el tamaño del grupo c_j , para $j = 0, 1$. El procedimiento para generar una observación artificial usando ROSE se presenta a continuación:

- Seleccionar $y^* = c_j$ con probabilidad π_j .
- Seleccionar $(x_j, y_j) \in T_n$ tal que $y_j = y^*$.
- Generar una observación sintética x^* desde $K_{H_j}(\cdot, x_j)$, siendo K_{H_j} una distribución de probabilidad centrada en x_j y cuya matriz de covarianzas es H_j .

El procedimiento de análisis de los datos y estimación de la DAP se desarrolló bajo la siguiente secuencia:

1. Estimación del modelo de regresión logística binaria usando todas las variables predictoras disponibles de la Tabla 2 y el conjunto original de datos con grupos desbalanceados (Modelo 1). Identificación de las variables predictoras significativas al 5% de nivel de significación.
2. Estimación del modelo de regresión logística binaria considerando las variables predictoras identificadas en el paso anterior (Modelo 2).
3. División del conjunto original de datos en el conjunto de entrenamiento y el conjunto de prueba (80% y 20% respectivamente).
4. Balanceo del conjunto de entrenamiento usando las muestras sintéticas obtenidas mediante el algoritmo ROSE.
5. Estimación del modelo de regresión logística binaria usando el conjunto de entrenamiento con grupos balanceados y las variables predictoras significativas identificadas en el paso 2 (Modelo 3).
6. Estimación de la DAP, su error estándar y el intervalo de confianza Bootstrap a partir de los Modelos 2 y 3.
7. Cálculo de los indicadores de la capacidad predictiva de los Modelos 2 y 3.

El procedimiento de análisis se realizó usando el programa estadístico R, en específico los paquetes DCchoice y ROSE. El paquete DCchoice permite obtener modelos de elección discreta para la estimación del valor económico de un bien público y el paquete ROSE permite balancear los grupos a través de la generación de muestras sintéticas.

3. Resultados y discusión

Se estimó el modelo de regresión logística binaria considerando todas las variables

predictoras disponibles (Modelo 1) y el conjunto original de datos. A continuación, se eliminaron las variables predictoras no significativas usando un nivel de significación del 5%. Luego, el modelo de regresión logística binaria final considera las variables predictoras VISIT, LUGN e ING (Modelo 2). Los coeficientes de regresión estimados y sus errores estándar correspondientes para ambos modelos se presentan en la Tabla 4.

Tabla 4. Modelos 1 y 2 usando el conjunto de datos con grupos no balanceados

	Modelo 1	Modelo 2
(Intercepto)	-5,009*** (1,427)	-4,433*** (1,027)
MFLO	-0,694 (0,542)	
VISIT	4,975*** (0,879)	4,884*** (0,837)
MONPROP	-0,087*** (0,014)	-0,084*** (0,014)
LUGN	1,019* (0,471)	1,137** (0,440)
EDAD	-0,010 (0,024)	
SEXO	0,034 (0,466)	
EDUC	-0,312 (0,504)	
SITL	0,228 (0,495)	
TRABH	0,664 (0,370)	
ING	0,003*** (0,001)	0,002*** (0,001)

Nota: Significancia: ***0,001**0,01*0,05.
Desviación o error estándar del coeficiente estimado en paréntesis

La Tabla 5 muestra la distribución de los conjuntos de entrenamiento y prueba. Para este fin se eligen aleatoriamente el 80% de las observaciones para constituir el conjunto de entrenamiento y el 20% restante conforma el conjunto de prueba.

Tabla 5. Distribución del tamaño del conjunto de entrenamiento y de prueba

Datos	Respuesta		Total	Porcentaje
	No	Si		
Entrenamiento	62	138	200	80%
Prueba	13	37	50	20%
Datos originales	75	175	250	100%

El proceso de generación de muestras sintéticas se realizó utilizando el algoritmo ROSE. Este método busca equilibrar los grupos definidos por la variable de respuesta y se realizó solo en el conjunto de entrenamiento, tal como se muestra en la Tabla 6. El conjunto de prueba mantiene su estructura inicial debido a que se trata del conjunto de datos usado para evaluar el modelo estimado.

Tabla 6. Distribución del tamaño del conjunto de entrenamiento sin y con balanceo

Datos del conjunto de entrenamiento no balanceada (antes de aplicar Rose)		
Respuesta	Cantidad	Proporción
Si	140	0.70
No	60	0.30
Total	200	1.00
Datos del conjunto de entrenamiento balanceada (después de aplicar Rose)		
Respuesta	Cantidad	Proporción
Si	92	0.46
No	108	0.54
Total	200	1.00

Se estimó el modelo de regresión logística binaria, considerando todas las variables predictoras presentes en el Modelo 2, usando el conjunto de entrenamiento con grupos balanceados con el algoritmo ROSE (Modelo 3). Los coeficientes de regresión estimados y errores estándar correspondientes a los Modelos 2 y 3 se presentan en la Tabla 7.

Tabla 7. Modelos 2 y 3 (antes y después de ROSE)

Variables	Modelo 2 (antes de ROSE)	Modelo 3 (después de ROSE)
(Intercepto)	-4,433*** (1,027)	-5,321 ** (1,870)
VISIT	4,884*** (0,837)	5,712 ** (1,852)
MONPROP	-0,084*** (0,014)	-0,064*** (0,012)
LUGN	1,137** (0,440)	0,610 *** (0,395)
ING	0,002*** (0,001)	0,001** (0,0004)

Nota: Significancia: ***0,001**0,01*0,05.
Desviación o error estándar del coeficiente
estimado en paréntesis

La Tabla 8 muestra la estimación de la DAP, usando la ecuación (2), su error estándar y el intervalo de confianza correspondiente, obtenidos a partir de los Modelos 2 y 3.

Tabla 8. Valor estimado de la DAP (PEN) en los Modelos 2 y 3

	Modelo 2 (antes de ROSE)	Modelo 3 (después de ROSE)
DAP estimada	35,6	14,4
Error estándar	8,605	3,884
Intervalo de confianza	(8,1505; 42,510)	(4,7115; 20,536)
Número de datos	250	200
Grupo Balanceados	No	Si

Los resultados obtenidos en la Tabla 8 muestran que el valor estimado de la DAP a partir del Modelo 3 es de aproximadamente 14,4 PEN, mucho menor en comparación con los 35,6 PEN obtenidos a partir del Modelo 2. Al utilizar el Modelo 2 es posible obtener una sobreestimación del valor la DAP. Por otro lado, con el Modelo 3 se obtuvo una reducción en el valor del error estándar correspondiente, lo que implica una mayor precisión en la estimación obtenida.

Los indicadores de la capacidad predictiva de los Modelos 2 y 3 no presentaron diferencias importantes según los valores de la Tabla 9: la Precisión aumenta de 0,896 a 0,967, el Recall disminuye de 0,937 a 0,829, el valor F-score disminuye ligeramente de 0,916 a 0,892 y el área bajo la curva ROC aumenta de 0,923 a 0,971 al pasar del modelo estimado con grupos desbalanceados al modelo estimado luego de aplicar el algoritmo ROSE.

Tabla 9. Indicadores de la capacidad predictiva para los modelos 2 y 3

	Modelo 2 (antes de ROSE)	Modelo 3 (después de ROSE)
Precisión	0,896	0,967
Recall	0,937	0,829
F-score	0,916	0,892
AUC	0,923	0,971

4. Conclusión

Sobre la base de los resultados obtenidos se puede concluir que:

1. El Modelo 2 (grupos no balanceados) y el Modelo 3 (grupos balanceados) no presentan diferencias significativas en los valores obtenidos para sus indicadores de capacidad predictiva.
2. La estimación de la DAP obtenida con el Modelo 3 presenta mayor eficiencia estadística en comparación con la estimación obtenida por el Modelo 2, debido a que se reduce el error estándar y la amplitud del intervalo de confianza correspondiente asociados a las estimaciones.
3. Sin embargo, a partir del Modelo 3 se obtuvo una estimación mucho más realista para la DAP (14,2 PEN), en comparación con la estimación obtenida con el Modelo 2 (35,6 PEN). De esta forma, utilizar el algoritmo ROSE para el balanceo de grupos permite una mejor estimación del modelo logit evitándose una sobreestimación de la DAP.

Conflictos de intereses

Los autores firmante del presente trabajo de investigación declara no tener ningún potencial conflicto de interés personal o económico con otras personas u organizaciones que puedan influir indebidamente con el presente manuscrito.

Rol del autor

CLDCV, JPC y CVM: Conceptualización, Investigación, Escritura-Preparación del borrador original, Redacción-revisión y edición.

Fuentes de financiamiento

Esta investigación no recibió ninguna subvención específica de ninguna agencia de financiación, sector gubernamental ni comercial o sin fines de lucro.

Aspectos éticos / legales:

Los autores declaran no haber incurrido en aspectos antiéticos ni haber omitido normas legales.

ORCID y correo electrónico

Carlos López De Castilla Vásquez	 clopez@lamolina.edu.pe
	 https://orcid.org/0000-0001-7724-5522
Jaime Porras Cerrón	 jaimpec@lamolina.edu.pe
	 https://orcid.org/0000-0001-9805-8785
Gesabel Villar Morales	 gesabel.villar@gmail.com
	

5. Referencias

Albarracin-Valdivia, A., & Alarcón-Novoa, J. (2021). Valoración económica de los ecosistemas del Área Natural Protegida “Vilacota Maure”, Tacna- Perú. *Revista*

Nicolaitade Estudios Económicos, 16(1), 23-51.

Bishop, R., & Heberlein, T. (1979) Mesuaring the market goods: are indirect measures biased. *American Journal of Agricultural Economics*, 61(5),1-15.

Bouwma, I., Schleyer, C., Primmer, E., Winkler, K. J., Berry, P., Young, J., & Vadineanu, A. (2018). Adoption of the ecosystem services concept in EU policies. *Ecosystem Services*, 29, 213-222. <https://doi.org/10.1016/j.ecoser.2017.02.014>

Efron, B., & Tibshirani, R.J. (1994). An introduction to the bootstrap. CRC press.

Garzón, L. P. (2013). Revisión del método de valoración contingente: experiencias de la aplicación en áreas protegidas de América Latina y el Caribe. *Espacio y Desarrollo*, (25), 65-78.

Gualli Guamán, L.E. (2022). Valoración económica y ambiental de las Microcuencas de los Ríos Alao y Maguazo de la Parroquia Pungalá, Cantón Riobamba, Provincia de Chimborazo. Tesis de Bachiller. Universidad Nacional de Chimborazo. Ecuador.

Hanemann, M., Loomis, J., & Kanninen, B. (1991). Statistical efficiency of double-bounded dichotomous choice contingent valuation. *American Journal of Agricultural Economics*, 73(4), 1255-1263. <https://doi.org/10.2307/1242453>

Hernández Valdivia, M.S., Valdivia Alcalá, R., & Hernández Ortíz, J. (2019). Valoración de servicios ambientales y recreativos del Bosque San Juan de Aragón, Ciudad de México. *Revista Mexicana de Ciencias Forestales*, 10(54), 100-117. <https://doi.org/10.29298/rmcf.v10i54.557>

Labandeira, X., León, C.J., & Vázquez, M.X. (2007). *Economía ambiental*. Pearson Educación.

Lavado Solís, K.N. (2021). Valoración económica y disposición a pagar por la conservación: aplicación de indicadores de la biodiversidad. Tesis para optar por el grado de PhD en Economía de los Recursos Naturales y el Desarrollo Sustentable, Universidad Nacional Agraria La Molina, Lima. Perú.

- Lunardon, N., Menardi, G., & Torelli, N. (2014). ROSE: a package for binary imbalanced learning. *R journal*, 6(1). <https://journal.r-project.org/archive/2014-1/menardi-lunardon-torelli.pdf>
- McFadden, D., & Train, K. (2017). *Contingent Valuation of Environmental Goods. A Comprehensive Critique* Cheltenham, UK: Edward Elgar Publishing. <https://doi.org/10.4337/9781786434692>
- Medalla Macedo, J.K. (2020). Valoración económica del servicio ecosistémico de los toboganes del Encanto de la Novia del distrito de Padre Abad provincia de Padre Abad-Ucayali.
- Melo Guerrero, E., Rodríguez Laguna, R., Martínez Damián, M.Á., Hernández Ortíz, J., & Zárate, R. (2020). Consideraciones básicas para la aplicación de experimentos de elección discreta: una revisión. *Revista Mexicana de Ciencias Forestales*, 11(59), 4-30. <https://doi.org/10.29298/rmcf.v11i59.676>
- Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, 28, 92-122. <https://doi.org/10.1007/s10618-012-0295-5>
- Ministerio del Ambiente [MINAM] (2021). “Guía de Valoración Económica de Impactos Ambientales en el marco del Sistema Nacional de Evaluación del Impacto Ambiental”. Lima: Ministerio del Ambiente. Resolución Ministerial N° 047-2022-MINAM <https://cdn.www.gob.pe/uploads/document/file/2841548/RM.%20047-2022-MINAM.pdf.pdf?v=1645360244>
- Ogrodowczyk, J.D. (2003). A theoretical and statistical exploration into the effects of morals, personality and uncertainty on hypothetical bias in contingent valuation. Thesis submitted for the degree of Doctor of Philosophy. University of Massachusetts Amherst.
- Quispe Aguilar, M.A. (2020). Valoración económica del servicio de ecoturismo en los humedales de Pisco, a partir del método de valorización contingente. Tesis para optar el Grado de Magister Scientiae en Ciencias Ambientales. Universidad Nacional Agraria La Molina, Lima. Perú.
- Rakotonarivo, O.S., Schaafsma, M., & Hockley, N. (2016). A systematic review of the reliability and validity of discrete choice experiments in valuing non-market environmental goods. *Journal of Environmental Management*, 183,98-109. <https://doi.org/10.1016/j.jenvman.2016.08.032>
- Ramírez Melgarejo, E. L. (2022). Valoración económica de la belleza paisajística de la bella durmiente del pntm por la población de la ciudad de Tingo María, Huánuco. Tesis para optar por el grado de Ingeniero Ambiental. Universidad Nacional Agraria de la Selva.
- Ruiz Aguilar, M. (2007). Valor económico del servicio de ecoturismo, usando el método de valoración contingente: el caso del bosque reservado de la Universidad Nacional Agraria de la Selva-Tingo María. Tesis para optar el grado de Magister en Economía de los Recursos Naturales y del Ambiente. Universidad Nacional Agraria La Molina, Lima. Perú.
- Tudela Mamani, J.W., & Leos Rodríguez, J.A. (2017). Herramientas metodológicas para aplicaciones del método de valoración contingente.
- Vilela, T., Harb, A. M., & Vergara, C. M. (2022). Chileans’ willingness to pay for protected areas. *Ecological Economics*, 201, 107557. <https://doi.org/10.1016/j.ecolecon.2022.107557>