

Modelado ecológico con Epi-Chagas

Una plataforma de minería de datos espaciales

<https://epichagas.c3.unam.mx/>



Este documento es introducción básica para el uso de la plataforma web Epi-Chagas, una herramienta interactiva para el modelado de distribución de los agentes causales (patógeno, vectores, hospederos) de la enfermedad de Chagas, así como la distribución de casos reportados de la enfermedad. Epi-Chagas ha sido desarrollado con el apoyo de la Fundación Río Arronte I.A.P, Fundación Carlos Slim y la Universidad Nacional Autónoma de México



RÍO ARRONTE
FUNDACIÓN

FUNDACIÓN
Carlos Slim

Epi-Chagas: herramienta interactiva para el análisis del nicho epidemiológico de la enfermedad de Chagas

MARCO TEÓRICO

La enfermedad de Chagas está muy extendida en todo el continente americano y representa un gran problema de salud pública. La Organización Mundial de la Salud (OMS) estima que aproximadamente 25 millones de personas están en riesgo de infección y entre 6 y 7 millones de personas están infectadas. Particularmente para México, se ha estimado alrededor de 5.5 millones de personas potencialmente afectadas por la enfermedad de Chagas. El agente etiológico de la enfermedad de Chagas es el protozoo *Trypanosoma cruzi*, que se transmite por heces infectadas de triatomíneos que ingresan al torrente sanguíneo humano. En condiciones naturales, el ciclo de vida de *T. cruzi* alterna entre insectos vectores (triatomíneos de la familia Reduviidae) y huéspedes vertebrados (principalmente especies de mamíferos). Actualmente, se han reportado 40 especies de chinches infectadas naturalmente por *T. cruzi* en América del Norte. Estas especies de chinches pertenecen a los géneros *Rhodinus*, *Triatoma* y *Panstrongylus*. En México, se han confirmado 21 especies positivas para *T. cruzi*, y hasta el momento se han confirmado 52 especies de mamíferos silvestres positivas a *T. cruzi*, pertenecientes a las Clases Marsupialia, Edentata, Chiroptera, Carnivora, Arthiodactyla, Rodentia y Primates. La diversidad de vectores y huéspedes vertebrados refleja la complejidad de los ciclos de transmisión de *T. cruzi*.

Caracterizar las condiciones donde puede suceder la transmisión de *T. cruzi* a las poblaciones humanas deber ser una meta en el corto plazo. Conocer estas características ambientales es un primer paso hacia la comprensión e interpretación de la interrelación entre el ambiente y la presencia de *T. cruzi*. Sin embargo, lograr estas metas presenta varios desafíos, entre ellos contar con información de los agentes causales involucrados en los ciclos de transmisión de la enfermedad, por otro lado, contar con métodos analíticos y herramientas informáticas que nos permitan evaluar las relaciones entre patógenos, vectores, hospederos y variables socioambientales considerando la complejidad asociada a la enfermedad.

Ante el riesgo que representa la enfermedad de Chagas a la salud pública en México, aquí se presenta una herramienta informática, la cual permite construir modelos de riesgo para la presencia de *T. cruzi* en México. La plataforma implementa una metodología de minería de datos espaciales que permite integrar, explorar y analizar variables de diferente origen y tipo (biológicas, sociales, climáticas). Esta propuesta metodológica utiliza un marco Bayesiano para crear modelos de la probabilidad condicional $P(C|X_{(t)})$, donde C es a clase de interés, por ejemplo, un patógeno, vector o caso de una enfermedad; y $X_{(t)} = (X_{(t1)}, X_{(t2)}, \dots, X_{(tm)})$ es un vector de variables predictoras, por ejemplo, temperatura. Para evaluar la confianza estadística de esta asociación se aplicará la prueba binomial $\epsilon(C|X_{(t)}) = [N_X(P(C|X_{(t)}) - P(C))]/[(N_X * (P(C) * (1 - P(C))))]^{1/2}$. La cual nos permite determinar si uno o más factores $X_{(t)}$ están correlacionados con C en una manera inconsistente con la hipótesis nula $P(C)$, donde valores de $|\epsilon| > 2$ corresponden a una asociación estadísticamente significativa ($P < 0.05$). De esta forma $\epsilon(C|X_{(t)})$ permite identificar la importancia relativa de los factores asociados a la presencia de C ; tal que, si $P(C|X_{(t)}) > P(C)$, entonces X puede ser interpretado como factor de “nicho” para la presencia de C . Mientras combinaciones tal que $P(C|X_{(t)}) < P(C)$, entonces X puede ser considerado un factor que desfavorece la presencia de C (i.e. “anti-nicho”).

Este marco de análisis también permite construir modelos espacialmente explícitos con lo cual se generarán mapas de riesgo para la presencia del patógeno, vector, hospedero o casos clínicos. Para generar los modelos de distribución potencial se calculará una función de *score*: $S(X) = \sum_{i=1}^n \ln P(X|C)/P(X|\underline{C})$, donde X es el conjunto de variables potencialmente predictoras (biológicas, ambientales, sociales) para la presencia/no presencia de C . Así, asignando el score correspondiente de cada variable [$S(X) = S(X_1) + S(X_2) + \dots + S(X_n)$] a una región determinada, se determina el perfil

socioambiental de diferentes regiones geográficas. Los valores altos/bajos de $S(X)$ indican si las condiciones son o no favorables para la presencia de *C*. Los análisis en la plataforma se realizan a nivel municipal, por lo tanto, la función de *score* nos indicara los municipios con alto/bajo riesgo de presencia del agente modelado.

IMPLEMENTACIÓN

La plataforma Epi-Chagas implementa la metodología descrita anteriormente para caracterizar el nicho ecológico de las especies potencialmente involucradas en la transmisión de *T. cruzi*. Epi-Chagas contiene una base de datos geográficos con los registros de triatomos y mamíferos que han sido confirmados positivos a *T. cruzi*, así como casos clínicos de la enfermedad. Para caracterizar el nicho de estas especies o de los casos clínicos, Epi-Chagas contiene la base de datos del Sistema Nacional de Biodiversidad de CONABIO, 19 variables bioclimáticas obtenidas del portal WorldClim (<http://www.worldclim.org/>) y las variables sociodemográficas del Censo de población y vivienda de INEGI (INEGI, 2020). De esta forma, se puede identificar qué variables bióticas, climáticas y/o sociodemográficas pueden favorecer o limitar la distribución de las especies positivas al patógeno. Adicionalmente, Epi-Chagas permite construir mapas de riesgo a nivel municipal.

A continuación, se presenta la implementación de esto métodos de minería de datos en la plataforma web, <https://epichagas.c3.unam.mx/>. Esta plataforma se puede acceder desde los navegadores de internet Google Chrome, Microsoft Edge, Mozilla Firefox o Safari. Para acceder se requiere tener un *usuario* y *contraseña*. Para ejemplificar el flujo de trabajo en Epi-Chagas se presentan casos de estudios para los diferentes agentes causales y tipos de covariables predictivas.

Página de inicio Epi-Chagas

Como se menciono anteriormente se requiere el nombre de usuario y contraseña, para ingresar al sistema.



The screenshot shows the Epi-Chagas web platform interface. At the top, there is a navigation bar with logos for UNAM, Fundación Carlos Slim, and CHILAM. Below the navigation bar is a large blue banner with the text "Epi-Chagas" and a network diagram. The main content area is divided into two columns. The left column contains the text "Bienvenido a Epi-Chagas" and "Plataforma de Eco-epidemiología Espacial". Below this, it states: "Epi-Chagas es una herramienta interactiva para el análisis del nicho epidemiológico de la enfermedad de Chagas y la creación de modelos predictivos para analizar su dinámica espacial en México." and "Análisis de nicho epidemiológico". It then says: "Si deseas realizar un análisis de nicho epidemiológico basta con seguir los siguientes pasos:". The right column contains the "Iniciar Sesión" section, which includes input fields for "Usuario:" and "Contraseña:", and an "Iniciar Sesión" button.

Módulos de Epi-Chagas

Modulo para selección de nuestra clase, C, de interés.

Epi-Chagas realiza análisis a nivel de país (Región: México) con una resolución a nivel municipal (Resolución: Municipios). Estos parámetros son fijos en la plataforma.

Epi-Chagas
Plataforma de simulación de transmisión de la enfermedad de Chagas en México

Región: MEXICO Resolución: Municipios

1 Agente: Seleccione un agente 2 Raíz taxonómica: Seleccione Grupo

Enfermedad:

+ -

Ver Grupo de interés

Reportar error

Para iniciar un análisis se debe elegir el agente con el que se quiere trabajar, **patógeno, vectores u hospederos**:

Epi-Chagas
Plataforma de simulación de transmisión de la enfermedad de Chagas en México

Región: MEXICO Resolución: Municipios

1 Agente: Seleccione un agente 2 Raíz taxonómica: Seleccione Grupo

Patógeno
Hospedero
Vector

+ -

Ver Grupo de interés



Reportar error

Una vez seleccionado el agente de interés, se selecciona la raíz taxonómica que se quiere utilizar. Debido a que se pueden tener diferentes grupos taxonómicos, la selección de la raíz nos permite desplegar un árbol taxonómico que nos permite seleccionar especies individuales o conjuntos de especies por Genero o Familia. A continuación, se muestra las diferentes visualizaciones de los tres agentes con la raíz taxonómica Familia:

The three screenshots show the Epi-Chagas interface with the following configurations:

- Screenshot 1:** 'Agente' is set to 'Patógeno'. The 'Raíz taxonómica' is 'Familia'. The list on the right includes Trypanosomatidae, Trypanosoma, and Trypanosoma cruzi.
- Screenshot 2:** 'Agente' is set to 'Hospedero'. The 'Raíz taxonómica' is 'Familia'. The list on the right includes Homínidae, Bovidae, Muridae, Procyonidae, Felidae, and Phyllostomidae.
- Screenshot 3:** 'Agente' is set to 'Vector'. The 'Raíz taxonómica' is 'Familia'. The list on the right includes Reduviidae, Triatoma, Rhodnius, and Dipetalogaster.

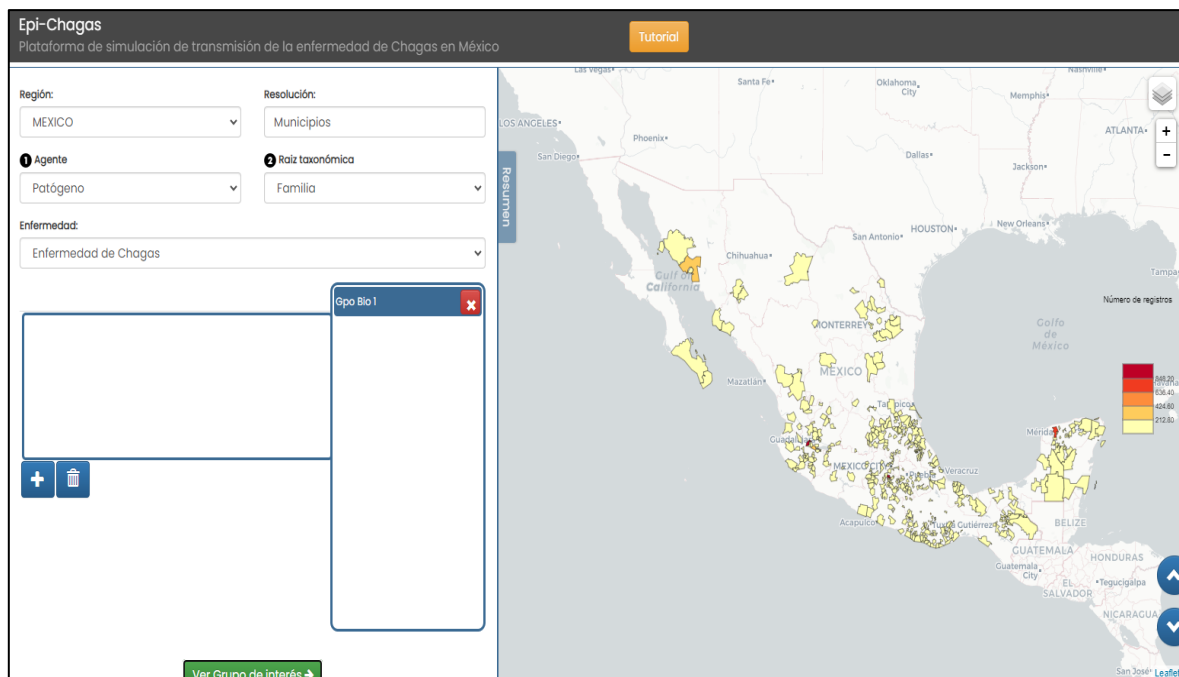
Si el usuario tiene interés en evaluar la presencia de caso clínicos, en la sección de *Hospedero* se debe seleccionar la familia Hominidae, si usa Familia, *Homo*, si usa Genero u *Homo sapiens* si usa Especie. En este caso el usuario estará construyendo un modelo para la enfermedad de Chagas. Si se seleccionan vectores u hospederos no humanos, se estarán construyendo modelos para la presencia de especies positivas al patógeno.

Para ejemplificar el uso de Epi-Chagas, se ira construyendo un modelo para *T. cruzi*. Una vez que seleccionamos *Agente*/*Patógeno*, *Raíz taxonómica*/*Familia*, los siguientes pasos son: 1) selecciona la especie con el botón , para agregar en la siguiente ventana. 2) asignar la especie al mapa de municipios, , en este paso se está agregando las presencias de la especie a los municipios correspondientes.

The two screenshots show the Epi-Chagas interface with the following configurations:

- Screenshot 1:** 'Agente' is set to 'Patógeno'. The 'Raíz taxonómica' is 'Familia'. The list on the right includes Trypanosomatidae and Trypanosoma. An arrow labeled '1' points to the 'Trypanosoma' entry.
- Screenshot 2:** 'Agente' is set to 'Patógeno'. The 'Raíz taxonómica' is 'Familia'. The list on the right includes Trypanosomatidae and Trypanosoma. An arrow labeled '2' points to the 'Ver Grupo de Interés' button.

Una vez que usamos el comando “ver grupo de interés” podremos ver en el mapa los municipios donde ha sido confirmada la presencia del grupo con el que estemos trabajando, en este caso es la distribución de *T. cruzi*.



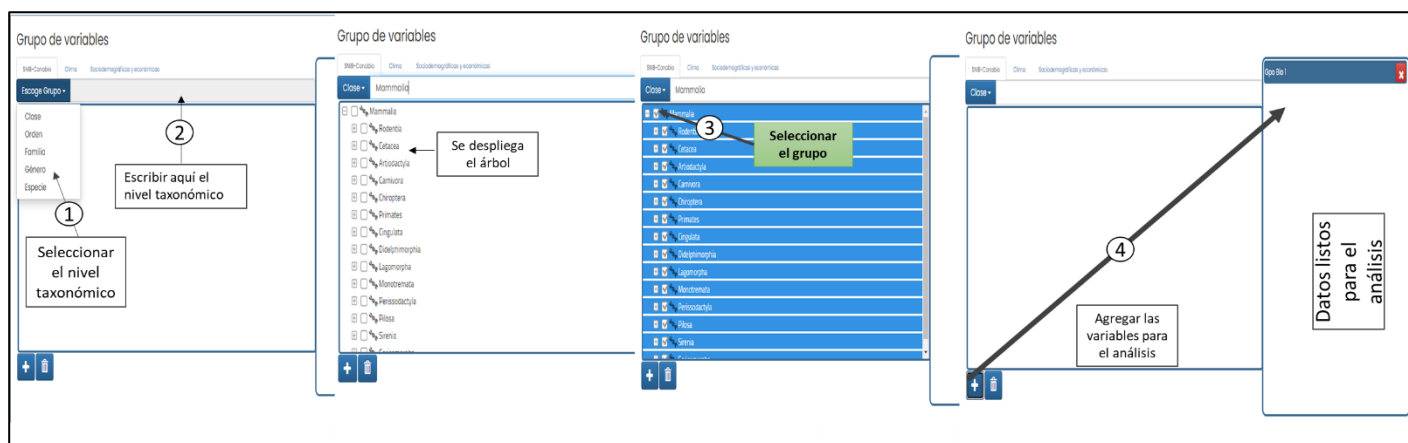
Modulo para selección de variables, X.

El siguiente modulo podemos seleccionar las variables que usaremos como predictoras, la plataforma tiene tres conjuntos de variables: biológicas, climáticas y sociodemográficas.

The screenshot shows the 'Grupo de variables' selection module. It features tabs for SNIB-Conabio, Clima, and Sociodemográficas y económicas. A 'Escoge Grupo' button is present. The right panel shows 'Parámetros' with options for 'Validación espacial' (No) and 'Min. Celdas con ocurrencia (n):' (1).

A continuación, se muestra como seleccionar las variables.

Las biológicas corresponden a los registros de especies de flora y fauna del Sistema Nacional de Información sobre Biodiversidad de México de CONABIO. Para estas variables se debe seleccionar el grupo taxonómico que se quiere utilizar y escribir el nivel taxonómico seleccionado, por ejemplo, si uso Clase, puede ser Mammalia, si uso Orden puedo escoger Rodentia. Una vez que se escribe el nivel taxonómico se desplegará el árbol correspondiente, este debe seleccionarse y agregarse a la siguiente ventana con el botón **+**. De esta forma tendremos listas nuestras variables para hacer el modelo. Como ejemplo use Clase: Mammalia.

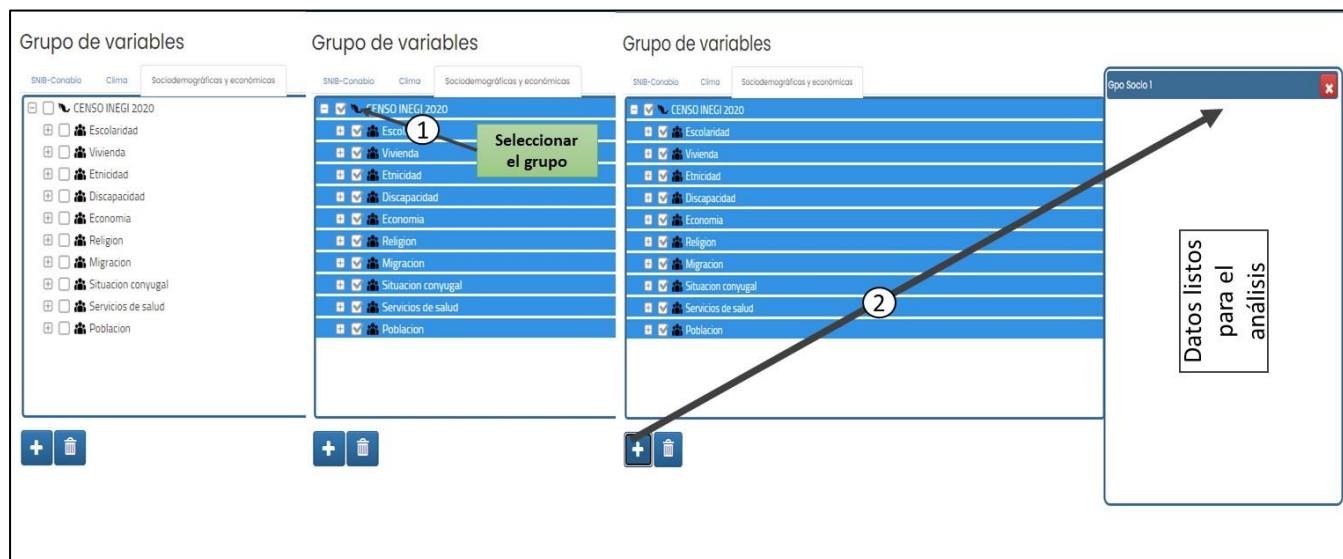


Si el interés es utilizar variables climáticas, se ingresa en la siguiente pestaña. Estas variables ya aparecen visibles. Corresponden a 19 variables bioclimáticas, que son derivadas de temperatura y precipitación (<https://www.worldclim.org/data/bioclim.html>). Aquí solo deben seleccionar del árbol las variables de interés, por ejemplo, solo temperaturas o solo precipitaciones o las 19 variables. Una vez seleccionadas se agregan para el análisis con el botón **+**.

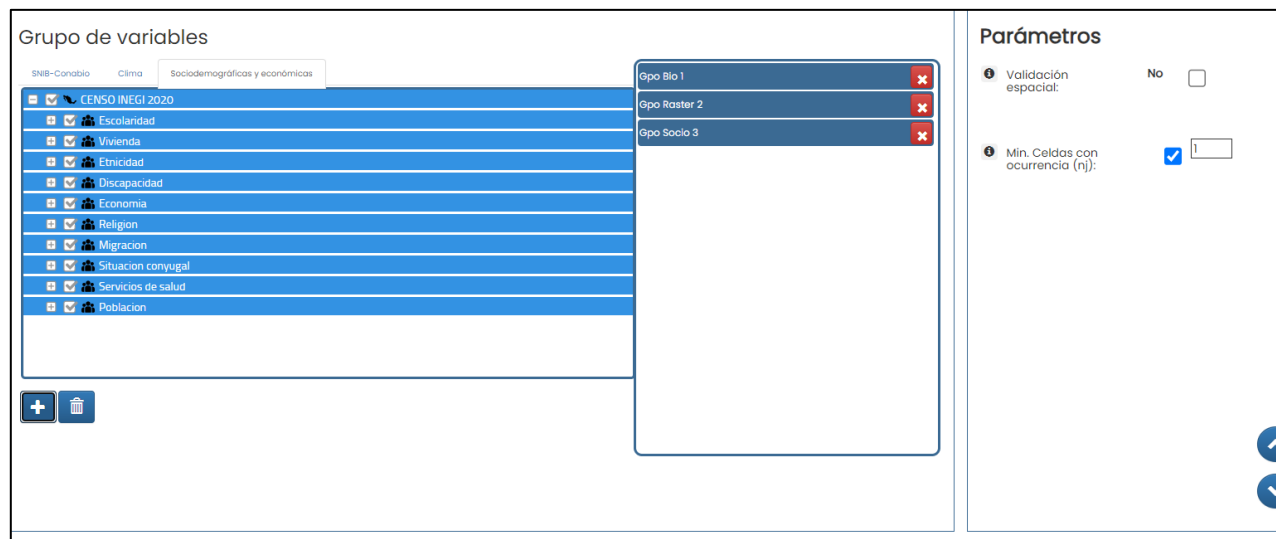


Finalmente, si el interés es utilizar variables sociodemográficas, se puede ingresar en la última pestaña. En esta sección se encuentran las variables del censo poblacional de INEGI, que están agrupadas de acuerdo al tipo de información que proporcionan. Al igual que con clima, las variables de INEGI ya

aparecen visibles y solo se debe seleccionar las que sean de interés o si se desea, se seleccionan todas y se agregan para el análisis.



Como se ha visto hasta aquí, se puede seleccionar algún tipo de variable (biológica, climática, sociodemográfica) que sea de interés para el usuario. Sin embargo, una de las ventajas de este sistema es que podemos seleccionar todos los tipos de variables para construir un modelo. Para esto solo debe ir repitiendo los pasos anteriores seleccionando uno a una los grupos de variables para tener los tres tipos en la ventana de variables para análisis. Nota: Cuando se agregan las variables para análisis, el sistema les da un nombre, si es biológica es “Bio”, clima es “Raster”, y sociodemográficas es “Socio” y les agrega un número. Si agrega dos grupos taxonómicos, los distinguirá con un número. Si agregan climáticas o sociodemográficas por separado también identificar con un número los subconjuntos.



En la sección de “**Parámetros**” se tiene la opción de seleccionar “validación espacial”, con esta opción el sistema divide los datos en entrenamiento y validación (70%:30%). Y corre 5 repeticiones para calcular una curva de $Recall = VP / (VP + FP)$.


Modulo resultados.

Para correr el modelo se estilízale botón “Ejecutar análisis

Ejecutar análisis

A continuación, se presentan los tipos de resultados que ofrece la plataforma.

El primer resultado que se observa es el modelo de distribución. A continuación, se muestran las salidas utilizando un tipo de variable predictor: mamíferos, clima y sociodemográficos. Además, se muestra una salida utilizando de forma conjunto los tres tipos de variables. Nota: la plataforma solo presenta un mapa a la vez, por lo cual, si se realiza un modelo con diferentes variables, por ejemplo: clima + sociodemográficos, el mapa que se despliega el modelo completo y no se presentan los individuales. Por lo tanto, si se quieren tener los modelos individuales deben correrse primero por separado descargarlos y al final un modelo completo.

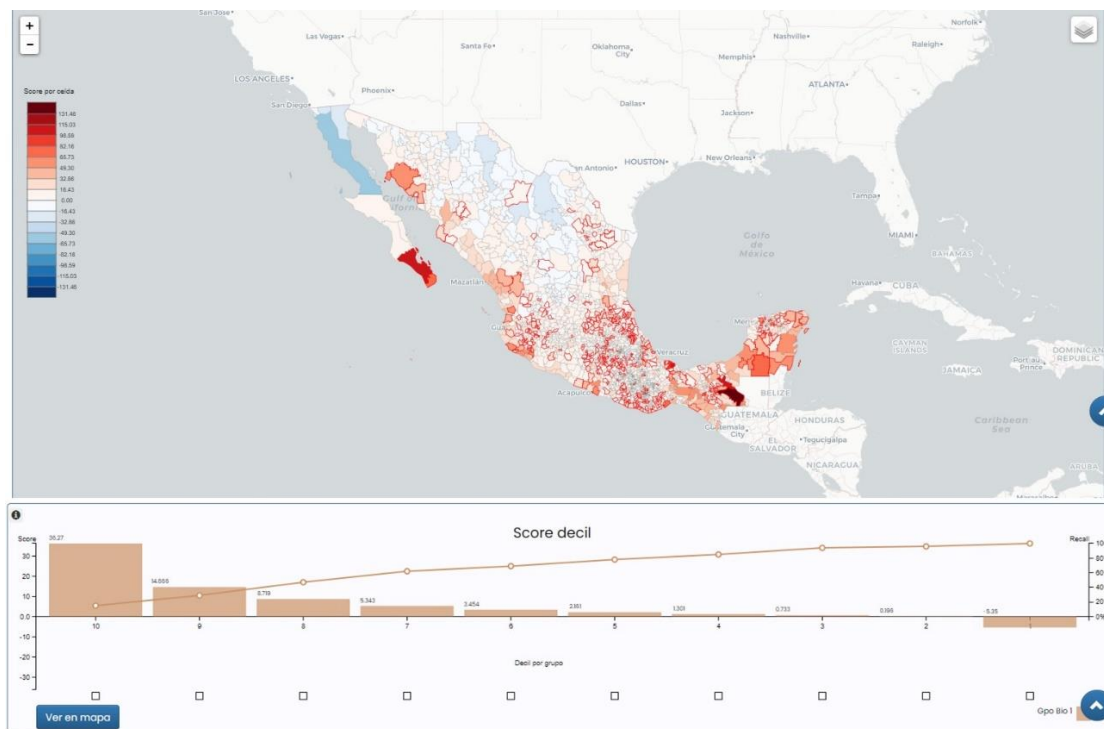
En el mapa se despliegan tanto el resultado como la distribución de la clase (municipios resaltados en rojo). Para ver solo el mapa, se puede utilizar el botón superior derecho  y desactivar la clase,

deseleccionando la opción “target”:

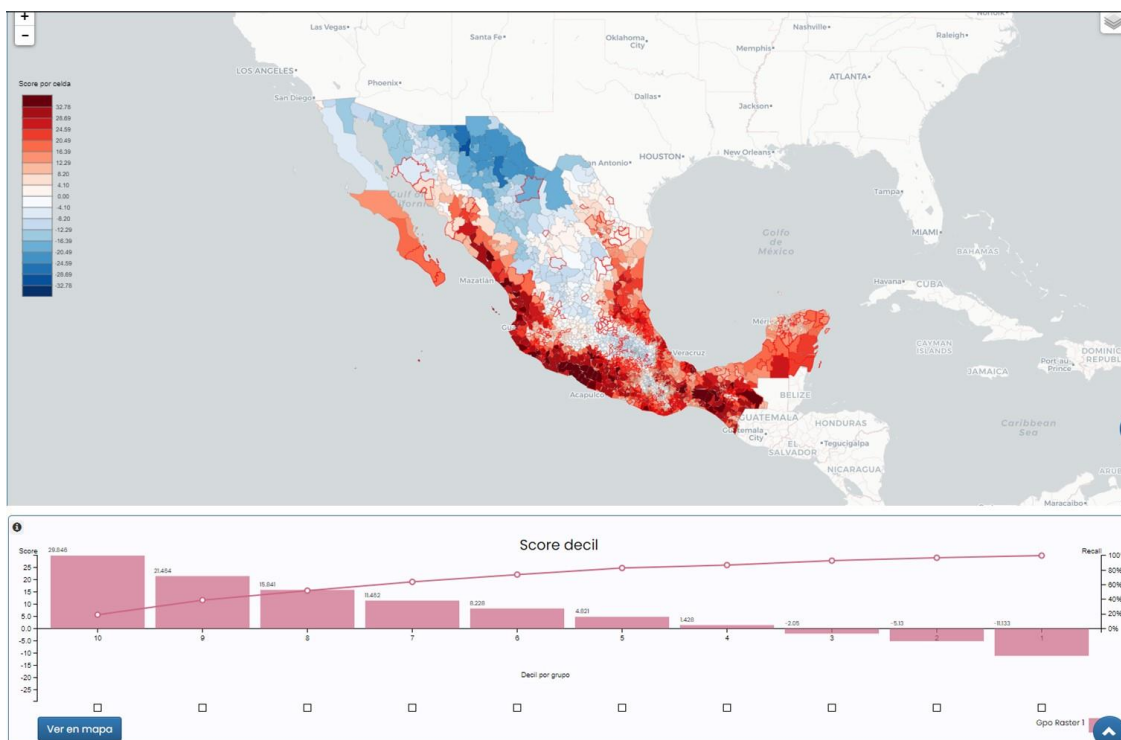


. A continuación, se presentan diferentes modelos de distribución.

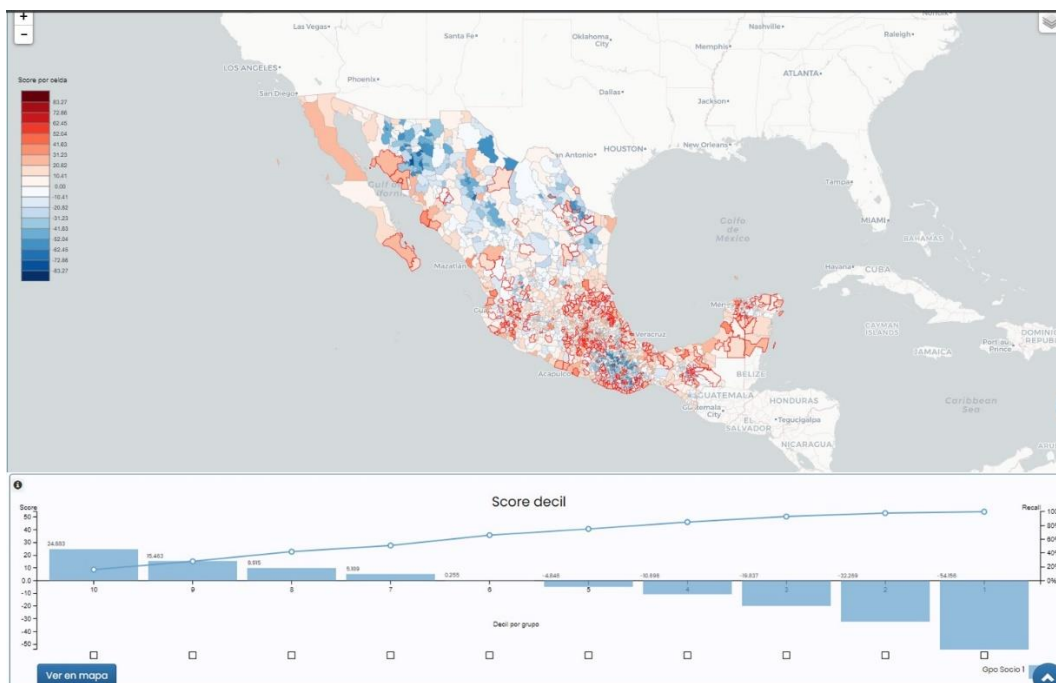
Modelo de distribución para *T. cruzi*. utilizando mamíferos.



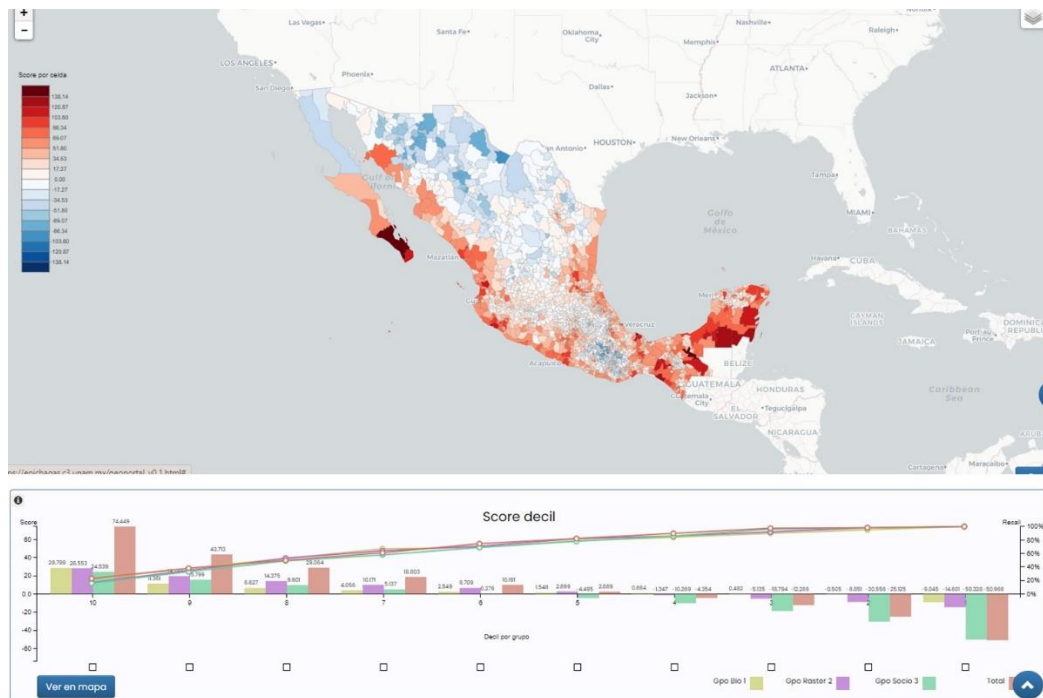
Modelo de distribución para *T. cruzi*. utilizando clima.



Modelo de distribución para *T. cruzi*. utilizando sociodemográficos.



Modelo de distribución para *T. cruzi*. utilizando mamíferos + clima + sociodemográficos.



Después del mapa se presenta una gráfica del desempeño de los modelos, para esto se ordenan los municipios de mayor a menor *score* y se dividen en deciles, es decir, se generan 10 grupos donde cada grupo corresponde a 10% de municipios. Para cada decil se calcula el promedio de *score*, que es el valor que se presenta en la gráfica. Los deciles con el mayor *score* representan los sitios donde las variables de nicho son favorables para la presencia de la especie y conforme disminuye el valor de *score* por decil nos indica que las condiciones no son favorables para la presencia de la especie. Una vez obtenidos los deciles se calcula el porcentaje de datos de validación correctamente predicho para cada *score*-decil donde se espera que el mayor porcentaje de datos de validación se encuentren en los deciles con mayor *score* y este porcentaje disminuya en los deciles con menor *score*. Con esto se construye la curva de *Recall* que nos permite observar el desempeño del modelo. Si se utilizan varios grupos de variables, la grafica mostrará el *score*-decil para cada modelo individual y en conjunto, y la curva de *Recall* para cada modelo individual y en conjunto.

Finalmente, se presentan dos tablas. En la primera tabla se presentan las variables presentes en para cada decil, es decir, que variables se observan para cada grupo del 10% de municipios. La segunda tabla es la lista completa de variables usadas en el modelo con sus valores de *epsilon* y *score* de acuerdo a su asociación con la especie de interés. De esta forma podemos cuantificar la importancia relativa de cada variable para nuestra especie, esta tabla puede exportarse en formato CSV, Excel.

Decil	Variable	Epsilon	Score	Porcentaje especie	Porcentaje decil
10	Población masculina de 18 años y más con educación postbásico 0.89%-4.02%	-3.121	-0.893	0.407	0.407
10	Población masculina de 18 años y más con educación postbásico 12.6%-35.4%	-0.908	-0.807	12.581	12.582
10	Población masculina de 18 años y más con educación postbásico 4.02%-5.98%	-0.103	-0.021	2.834	2.846
10	Población masculina de 18 años y más con educación postbásico 5.39%-9.44%	-1.924	-0.466	3.239	3.262
10	Población masculina de 18 años				

Mostrando 1 a 3,089 de 3,089 entradas

Especie/huésped	Hj	Hj	Hj	Hj	Epsilon	Score	Reino	Filum	Clase	Orden	Familia
[Grato promedio de escolaridad de la población femenina] (13.05.85)	17	247	275	2469	-2.058	-0.928					
[Grato promedio de escolaridad de la población	23	248	275	2469	0.578	0.015					

Vista general de la plataforma

