# **EpI-PUMA 1.0**

El propósito de EpI-PUMA 1.0 es predecir la evolución de la pandemia de SARS-Cov-2 en México en espacio y tiempo a nivel municipal. Lo hace al construir modelos predictivos de Aprendizaje de Máquina de la forma: P(C(t)|X(m)), que representa la probabilidad que un municipio "m" esté dentro de una "clase de interés" C en un mes t basado en sus características X(m), que representa un conjunto de factores (variables) socio-económicos, socio-demográficos, climáticos, de contaminación y de movilidad.

La clase C representa preguntas epidemiológicas de interés como: ¿Cuáles municipios presentarán el mayor número de casos de COVID19 en el siguiente mes? ¿En cuáles municipios habrá mayor mortalidad en el próximo mes? o ¿En cuáles de los municipios sin casos previos de COVID19 aparecerán nuevos casos? Generalmente los C representan el 10% de municipios con mayor riesgo para la característica epidemiológica seleccionada. Se puede construir tu propia clase de interés en el sistema en la parte "Criterios del Modelo".

Los datos para C vienen del reporte oficial de la Dirección General de Epidemiología de la Secretaría de Salud, y los otros datos de diferentes fuentes oficiales, como el Instituto Nacional de Estadística Geografía e Informática (INEGI). Más información de los datos utilizados se encuentra en: <a href="https://chilam.c3.unam.mx/proy-epipuma/datos-epipuma">https://chilam.c3.unam.mx/proy-epipuma/datos-epipuma</a>

#### Criterios del Modelo

Primero, se selecciona el "Tipo" de modelo: "Predictivo" significa que se pronosticará C para el próximo mes (t) con base en los factores X(t-1) en el mes previo, mientras "Perfilado" significa que nos interesa saber cuáles factores X(t) caracterizan los municipios asociados con C. Por ejemplo, ¿cuáles factores son más indicativos de un municipio que actualmente tiene muchos casos de COVID19? Es importante poder establecer perfiles de riesgo para entender cuáles factores llevan a más riesgo como, por ejemplo, altos niveles de movilidad.

"Variable objetiva" refiere a que característica "clínica" es de interés: contagios o disfunciones. Además, se puede seleccionar "Pruebas" para ver donde se han realizado más o menos pruebas. Los "Modificadores" determinan diferentes aspectos de interés de una "Variable objetiva" dada. Por ejemplo, para contagios, preguntarse en dónde hay o habrá casos, es decir "presencia" de casos independiente de su número, es distinto a analizar la incidencia, en donde hay más casos de forma relativa a la población del municipio. "Sin modificador" representa simplemente la presencia del "Variable objetiva" en un municipio independientemente de su grado.

El rubro de "Enfoque" es otro modificador para la pregunta de interés. Por ejemplo, un enfoque es "empeoramiento" y otro es "mejoramiento". "Empeoramiento" como clase C representa aquellos municipios que no estaban en el 10% de mayor riesgo según los criterios previos seleccionados, pero si son en el siguiente mes. De la misma manera "Mejoramiento" representa aquellos municipios que estaban en el 10% de mayor riesgo según los criterios previos seleccionados y en el siguiente mes está en el 90% de menor riesgo. Es decir que su perfil de riesgo mejoró. Finalmente, el enfoque "Estrella" da un modelo que predice cuáles municipios estarán en el top 10% de riesgo según la "Variable

objetiva" y "Modificador" seleccionados; el próximo mes en el caso de "Tipo" Predictivo y en el mes actual en el caso de "Tipo" Perfilado.

Los Criterios del Modelo permiten al usuario analizar más de 50 clases de interés C distintas para investigar la evolución de la pandemia.

#### Periodo de Validación

Después de seleccionar un tipo de clase de interés hay que seleccionar un "Periodo de Validación". Esto significa en el caso de "Tipo" Predictivo el mes que se quiere predecir. Por ejemplo, el seleccionado 2020-11 dará un modelo para predecir lo que pasaría en noviembre 2020 basado en datos disponibles en octubre de 2020. Asimismo, para "Tipo" Perfilado la selección 2020-11 dará el perfil de la clase de interés para ese mes. Se puede seleccionar cualquier mes entre enero 2020 y marzo 2023.

Es importante enfatizar que dado la dinámica compleja y adaptativa de la pandemia, ningún modelo predictivo generado al inicio de la pandemia podría mantener su validez por toda la pandemia donde factores de riesgo pueden ser más importantes en una etapa versus otra. Por lo tanto, la capacidad de crear diferentes modelos en diferentes etapas de la pandemia es una importante ventaja de EpI-PUMA 1.0 donde se puede analizar si los factores de riesgo (predictores) más importantes cambian en tiempo (Stephens et al. 2023, Salinas et al. 2022).

Después de seleccionar la clase de interés y periodo de validación es importante activar el botón "Confirmar modelo". Luego aparecerá un mapa en la ventana a la derecha que ilustra aquellos municipios asociados con la clase de interés (Figuras 1.1 y 1.2). El próximo paso es seleccionar los factores de riesgo/predictores  $\mathbf{X}$  que se usarán para construir el clasificador  $P(C|\mathbf{X})$ . Hay variables de múltiples tipos, que potencialmente pueden ser causas directas o indirectas en la evolución de la pandemia. Se puede seleccionar un predictor, o cientos, y crear un modelo correspondiente al accionar el botón "Ejecutar análisis".

La selección de variables es a la discreción del usuario y permite investigar un sinfín de hipótesis respecto a la importancia relativa de diferentes factores de riesgo. El sistema da resultados a través de distintas visualizaciones interactivas complementarias. Primero, se muestra un mapa de riesgo, donde el grado de riesgo (predictibilidad) está dado por un "score" de Bayes calculado para cada municipio (Figura 2). El "score" es un proxy para P(C|X) tal que un alto score significa una alta probabilidad y viceversa. Debajo del mapa se muestra una visualización del desempeño del modelo a través de curvas de "recall" (ver abajo para más descripción) como función del "score" donde se divide el conjunto de municipios ranqueados por el score en deciles (Figura 3). Así, el decil 10/1 representa el 10% de los municipios con más alto/bajo score y consecuentemente los municipios de más alto/bajo riesgo.

Abajo de la grafica de Recall se presenta el análisis de los factores de riesgo en varios cuadros. Primero, los factores de riesgo asociados con un decil de score y asociado con un cierto modelo. Por ejemplo, por seleccionar los grupos de variable "Pobreza" (datos de CONEVAL) y "Demográficos" (datos de INEGI) como "Factores de Riesgo" se crea tres modelos predictivos: uno basado únicamente en datos de Pobreza, otro basado en datos Demográficos y otro "Total" que combina los dos conjuntos. En las curvas de Recall se puede medir el desempeño relativo de cada modelo para descubrir cuál modelo tiene mejor desempeño y si por sumar diferentes tipos de datos se puede crear un modelo de

mejor desempeño. Por ejemplo, se puede contestar: "¿Cuáles tipos de variable (factores) predicen mejor aquellos municipios con el mayor número de casos? Al seleccionar un tipo de variable entre los cuadrantes al fondo y a la derecha en la imagen (Covariable 1, Covariable 2,..., Total) y luego un decil (10, 9,...,1) se puede ver en el mapa el conjunto de municipios que corresponden a ese decil de score para el modelo seleccionado. Por ejemplo, se puede ver aquellos municipios que están de más alto/bajo riesgo (decil 10/decil 1) para el modelo seleccionado. Asimismo, en la cuadro siguiente a las curvas de Recall se puede ver el conjunto de factores de riesgo asociados con ese modelo y decil (Ver los Ejemplos de uso).

Próximo hay una tabla que da el perfil de riesgo para cada municipio según el modelo seleccionado en "Criterios de modelo". El cuadro muestra los 5 factores de más alto y más bajo riesgo asociado con cada municipio; y se puede descargar el cuadro completo con la lista completa de la predictibilidad asociada con cada variable (predictor) para poder identificar cuáles son los principales factores de riesgo para la clase de interés seleccionada. Esto en particular es donde entra la Inteligencia Humana para ayudar en la interpretación de los resultados y si o no los resultados llevan a cualquier observación que puede afectar la toma de decisión.

Una descripción más detallada de la metodología asociada con EpI-PUMA 1.0 puede ser encontrada en el artículo "Does a Respiratory Virus Have an Ecological Niche, and If So, Can It Be Mapped?" Yes and Yes" (https://www.mdpi.com/2414-6366/8/3/178).

## Ejemplos de uso

En este ejemplo, construiremos dos modelos con los mismos predictores (Demográficos, Movilidad y Clima) y cambiaremos la clase de interés. Para tal efecto cambiaremos los Modificadores en la sección Criterios del Modelo. En el modelo 1 analizaremos la aparición de nuevos casos de COVID19 en municipios que no presentaban ninguno; y en el modelo 2 analizaremos a los municipios que presentan un aumento en el número de casos confirmados, sin importar si previamente se confirmaron casos o no. Así el modelo 1 representa la expansión de la enfermedad en los municipios libres de enfermedad, y el modelo 2 representa el crecimiento epidemiológico en la población.

#### Modelo 1

#### Selecciona la clase de interés

Tipo: Predictivo

Variable Objetivo: Covid-19 Confirmado

Modificadores: Sin modificadores

Enfoque: Empeoramiento Selección de Periodo: 2021-01

Construye y compara las predicciones de tres tipos de factores para explicar la clase de interés (Presencia de casos confirmados de COVID-19).

Selecciona os criterios del modelo y los factores de riesgo: variables demográficas, de movilidad y climáticas.



**Figura 1.1** Panel de selección de la clase de interés y los factores predictivos. En este caso seleccionamos las factores de riesgo sociales (variable 1: Demográficos, variable 2: Movilidad) y Factores de Riesgo Ambientales (World Clim).

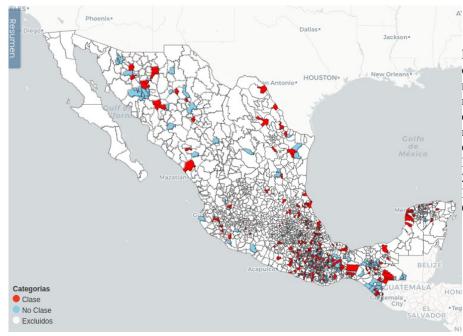
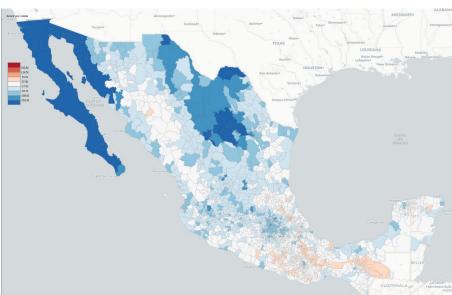


Figura 1.2. Mapa de la clase y No clase para el entrenamiento. El resumen muestra el numero de municipios de la clase y la no clase que se utilizan para entrenar al modelo y hacer las predicciones. En este caso la clase son 314 Municipios Donde Surgieron Confirmados en 2020-12, y la No clase son 447 Municipios Donde no Surgieron Confirmados en 2020-12.

#### Resultados del Modelo 1

El modelo 1 permite identificar a los municipios en mayor riesgo de presentar casos de COVID19 y los factores que lo colocan en esa categoría de riesgo.



**Figura 2.** Mapa de Municipios que no han tenido casos confirmados de Covid-19 y tienen mayor riesgo de presentar casos confirmados en el siguiente mes (color rojo).



**Figura 3**. Deciles de la población de municipios ordenados con base en su score. La gráfica de barras representa el score promedio de cada decil, para cada tipo de factor y el total. Las curvas de recall muestran el desempeño del modelo.

Con base en la suma del Score para cada municipio (barras de color purpura), se ordenan de mayor a menor y se agrega a los municipios en deciles. Para cada decil de municipios se analiza la contribución de cada tipo de factor. En la Figura 3, las barras representan el score promedio de los municipios en ese decil y el color indica el tipo de factor: Demográficos (covariable 1), Movilidad (covariable 2) y Climáticos (covariable 3).

Las líneas en la Figura 3 representan las curvas de "recall" que nos indican el desempeño del modelo de predicción. La escala de Recall da el total de municipios del grupo de interés identificado correctamente por el modelo como función del score (riesgo). P.ej. 40% de recall en el top 10% significa que el modelo predice 400% (4X) mejor que el azar.

Profundizando nuestro análisis en los factores que contribuyen al riesgo de que se observen casos de COVID19 confirmados en los municipios, podemos utilizar los siguientes cuadros, ordenándolos de mayor a menor Epsilon.

| 6 Decll | <b>∂</b> Variable  | <b>⊕</b> Epsilon 1 | 6 Score | 1 Porcentaje especie | Porcentaje decil |
|---------|--|--------------------|---------|----------------------|------------------|
| 0       | Viviendas particulares habitadas que disponen de automóvil o camioneta 12.2%;17.4% | 6.25               | 0.89    | 11.67%               | 11.07%           |
| 0       | Viviendas particulares habitadas con dos cuartos 29.1%:32.9%                       | 5.48               | 0.80    | 16.67%               | 15.81%           |
| 0       | Población de 15 años y más con secundaria completa 1.5% 8.5%                       | 5.44               | 0.79    | 58.51%               | 55.73%           |
| 0       | Población masculina no económicamente activa 27.3%:63.2%                           | 5.32               | 0.78    | 21.76%               | 20.55%           |
| 0       | Precipitación del mes mas seco (38.000 mm : 62.000 mm)                             | 5.14               | 0.60    | 34.92%               | 58.10%           |
| 0       | Precipitación del trimestre mas seco (137.000 mm : 216.000 mm)                     | 5.12               | 0.63    | 36.27%               | 53.75%           |
| 0       | Población con limitación para poner atención o aprender cosas sencillas 0%0.1%     | 5.09               | 0.75    | 38.75%               | 36.76%           |
| 0       | Viviendas particulares habitadas con dos dormitorios y más 45.8%.50.3%             | 5.02               | 0.74    | 15.70%               | 15.02%           |
| 0       | Flujo laboral que recibe el municipio desde otros municipios. 52:99                | 5.02               | 0.74    | 18.60%               | 17.79%           |
| 0       | Población económicamente activa 4.4%:27.4%   | 4.87               | 0.73    | 49.79%               | 47.43%           |

Figura 4. Cuadro de deciles. Muestra los factores que contribuyen al score de cada decil.

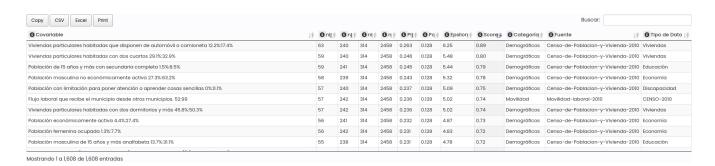
Las variables en el decil con mayor importancia son demográficas, climáticas y de movilidad. Las variables demográficas muestran carencias. Por ejemplo, municipios con pocas viviendas de dueños de un automóvil o camioneta (12 -17 %) y pocas viviendas con dos cuartos (29-32%), baja escolaridad (1-8 % de la población mayor de 15 años con secundaria completa), y desempleo en la población masculina (27-63%).

Las variables climáticas también contribuyen a predecir los municipios con mayor riesgo. EpI-PUMA 1.0 identifica a municipios con alta precipitación, en el mes más seco o en el trimestre más seco. Estas son características climáticas de los municipios del sur-oeste del territorio nacional (Mapa en la Figura 2).



**Figura 5.** Cuadro de riesgo calculado para cada municipio de México. Como un indicador de riesgo se muestra el score, que es la diferencia entre el score positivo y el score negativo. El score positivo es la suma del score de la variables con score positivo, y el score negativo es la suma del score de las variables con score negativo.

Ordenando el cuadro de la Figura 5, de mayor a menor score, podemos detectar el riesgo de que ocurra un evento epidemiológico entre los municipios sin casos confirmados de COVID19. En enero del 2021, los municipios en mayor riesgo estaba en los estados de Chiapas y Oaxaca, Veracruz, Yucatán y Quintana Roo.



**Figura 6**. Cuadro de la contribución de factores de riesgo. En este cuadro se muestran los factores predictores (j) de la clase de interés (i), y sus estadísticos (epsilon y score). El número de observaciones con las clases i y j (nij), las observaciones de la clase i (ni), las observaciones de la clase j (nj), el número total de observaciones (n); la probabilidad condicional de i dado j (Pij), y la probabilidad de observar la clase i (Pc).

#### Modelo 2

#### Selecciona la clase de interés

Tipo: Predictivo

Variable Objetivo: Covid-19 Confirmado

Modificadores: Número de casos

Enfoque: Empeoramiento Selección de Periodo: 2021-01

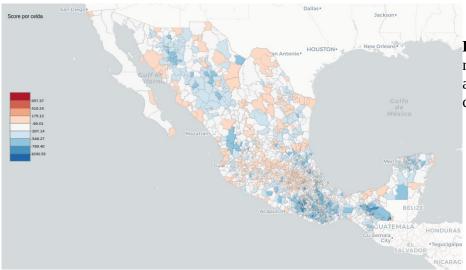
En este segundo ejemplo, probaremos los predictores del modelo 1 para analizar otra clase de interés: el aumento en el número de casos confirmados de Covid-19 en los municipios. Para ello, utilizamos el modificador Número de casos en la sección Modificadores del panel de control.

#### I. Criterios del Modelo (Todos los campos son obligatorios)



**Figura 7**. Criterios del modelo.

Seleccionamos los mismos factores que en el modelo 1: Demográficos, Movilidad y Clima. Ejecutamos el modelo y los resultados son los siguientes.



**Figura 8.** Municipios con mayor riesgo de presentar un aumento en el número de casos confirmados de Covid-19.



**Figura 9**. Deciles de la población de municipios ordenados con base en su score. La gráfica de barras representa el score promedio de cada decil, para cada tipo de factor y el total. Las curvas de recall muestran el desempeño del modelo.

El Recall del primer decil es cerca del 40%, es decir que el modelo predice 4 veces mejor que el azar a los municipios en el decil 10 de riesgo de que aumente el número de casos en el siguiente mes. Las barras en la figura representan el score promedio de los municipios en ese decil. El promedio es en general bajo e incluso negativo, lo que indica anti correlaciones entre la clase de interés y los factores predictores.

En el siguiente cuadro, se muestran los estadísticos Epsilon y Score para cada uno de los factores de riesgo analizados (Figura 10). Al ordenar el cuadro con base en Epsilon podemos observar los factores que estadísticamente contribuyen positivamente a la probabilidad de observar la clase de interés C.

| Copy CSV Excel Print Buscar: |  |                    |                |                      |                    |  |  |  |  |  |
|------------------------------|--|--------------------|----------------|----------------------|--------------------|--|--|--|--|--|
| 6 Decll                      | <b>♦</b> Variable  | <b>⊕</b> Epsllon ↓ | <b>⊕</b> Score | 1 Porcentaje especie | 1 Porcentaje decil |  |  |  |  |  |
| 0                            | Flujo laboral interno al municipio. 11548:25696                                      | 4.71               | 1.40           | 21.67%               | 20.55%             |  |  |  |  |  |
| 0                            | Población masculina de 18 años y más con educación pos-básica 18.3%:22.9%            | 4.69               | 1.40           | 31.54%               | 30.04%             |  |  |  |  |  |
| 0                            | Viviendas particulares habitadas que disponen de computadora 19.2%:26.4%             | 4.66               | 1.39           | 30.45%               | 29.25%             |  |  |  |  |  |
| 0                            | Población de 3 años y más que habla alguna lengua indígena 1.4%:3.9%                 | 4.08               | 1.29           | 21.67%               | 20.55%             |  |  |  |  |  |
| 0                            | Viviendas particulares habitadas que disponen de internet 10.8%:17.1%                | 4.06               | 1.29           | 30.71%               | 29.25%             |  |  |  |  |  |
| 0                            | Viviendas particulares habitadas con dos cuartos 18.2%:20.4%                         | 4.06               | 1.29           | 16.60%               | 15.81%             |  |  |  |  |  |
| 0                            | Población de 3 años y más que habla alguna lengua indígena y habla español 1.2%:3.5% | 4.06               | 1.29           | 23.24%               | 22.13%             |  |  |  |  |  |
| 0                            | Viviendas particulares habitadas que disponen de drenaje 96.9%:100%                  | 4.05               | 1.28           | 29.34%               | 28.06%             |  |  |  |  |  |
| 0                            | Población no económicamente activa 20.3%:33.8%                                       | 4.05               | 1.28           | 28.93%               | 27.67%             |  |  |  |  |  |
| 0                            | Población de 18 años y más con educación pos-básica 18.1%:22.7%                      | 4.05               | 1.28           | 31.82%               | 30.43%             |  |  |  |  |  |

Figura 10. Estadísticas sobre la contribución de cada factor a la probabilidad de observar la clase de interés.

El flujo laboral hacia el municipio es un factor de movilidad importante en determinar el aumento en el número de casos. El virus se transmite por el contacto entre las personas, por lo que la movilidad es un predictor natural en la transmisión de un agente infeccioso entre la población.

Los siguientes factores con mayor significado estadístico son variables asociadas a la educación y carencias de la población. Municipios con un 20% de jóvenes con educación posbásica y con poco acceso a internet y a recursos tecnológicos como una computadora. Poco acceso a una computadora, es característica de un estilo de vida menos proclive a quedarse en casa durante la pandemia.

El cuadro de la Figura 11 revela a los municipios en donde aumentará el número de casos en la población. Los diez primeros en los estados de Tlaxcala, Estado de México, Baja California Sur, Morelos, Baja California, Sonora y Puebla.

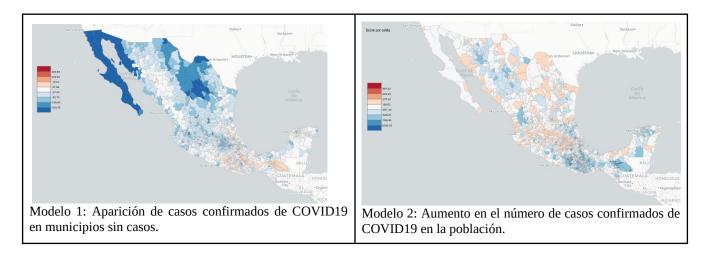
| Сору     | CSV Excel Print |          |        |                   |           |   |                          |   |                      |   |                      |  | Buscar            | :   |
|----------|-----------------|----------|--------|-------------------|-----------|---|--------------------------|---|----------------------|---|----------------------|--|-------------------|---|
| Estado   | ↓↑              | Score \$ |        | Score<br>Negativo | rlesgo \$ | Mejor PI<br>nombre  | Mejor †<br>P1 ¢<br>valor | Mejor P2<br>nombre  | Mejor<br>P2<br>valor | Mejor P3 nombre   | Mejor<br>P3<br>valor | Mejor P4 nombre  | Mejor<br>P4 valor | Mejor P5 nombre   |
| Tlaxcala | Chlautempan     | 116.03   | 136.08 | -20.05            | Muy alto  | Flujo laboral<br>interno al<br>municipio.<br>1.15e+4:2.57e+4                                  | 1.4                      | Población<br>masculina de 18<br>años y más con<br>educación pos-<br>básica<br>1.83e+1:2.29e+1 | 1.4                  | Viviendas particulares habitadas que<br>disponen de computadora 19.2%-26.4% | 1.39                 | Población de 3 años y más que habla<br>alguna lengua indígena 1.4%3.9% | 1.29              | Viviendas particulares habitadas que disponen de internet 10.8%:17.1%     |
| México   | Nextialpan      | -36.63   | 122.05 | -158.68           | Muy alto  | Población<br>masculina de 18<br>años y más con<br>educación pos-<br>básica<br>1.83e+1:2.29e+1 | 1.4                      | Población de 3<br>años y más que<br>habla alguna<br>lengua Indígena<br>1.40e+0:3.90e+0        | 1.29                 | Población no económicamente activa 20.3%:33.8%                              | 1.28                 | Viviendas particulares habitadas que<br>disponen de drenaje 96.9%:100% | 1.28              | Viviendas particulares habitadas que<br>disponen de luz eléctrica 0%.0.7% |
|          |                 |          |        |                   |           | Flujo laboral   |                          | Población<br>masculina de 18  |                      |   |                      |  |                   |   |

Figura 11. Estadísticas de riesgo de los municipios.

### Comparación entre los ejemplos 1 y 2

Los dos modelos 1 y 2 representan dos clases de interés que complementariamente nos muestran una imagen de lo que ocurrió en enero del 2021. La infección en municipios que no habían tenido ningún caso confirmado (modelo 1), y el aumento de casos observado en los municipios (modelo 2).

El modelo 1 nos muestra a los municipios que no presentaban casos de Covid-19 y son más susceptibles de presentar un caso confirmado de covid-19 en el siguiente mes. Este modelo sugiere que las poblaciones más vulnerables se encuentran en los estados de Chiapas y Oaxaca, Veracruz, Yucatán y Quintana Roo, con carencias económicas. En cambio, el modelo 2 muestra a los municipios en los que puede aumentar el número de casos confirmados, que generalmente tienen flujo laboral hacia el municipio, baja escolaridad y poco acceso a tecnologías de información. Estos municipios se encuentran distribuidos en todo el territorio nacional, a diferencia de lo que predice el modelo 1.



#### Referencias

C. R. Stephens, C. González-Salazar, and P. Romero-Martínez, "'Does a Respiratory Virus Have an Ecological Niche, and If So, Can It Be Mapped?' Yes and Yes," Trop. Med. Infect. Dis., vol. 8, no. 3, p. 178, Mar. 2023, doi: 10.3390/tropicalmed8030178.

A. Salinas et al., "EpI-PUMA: Plataforma Universitaria de Inteligencia Epidemiológica de SARS-CoV-2, Versión 1.0," TIES Rev. Tecnol. E Innov. En Educ. Super., no. 7, Mar. 2023, [Online]. Available: <a href="https://ties.unam.mx">https://ties.unam.mx</a>