TIES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

# Marco de Trabajo Tecnológico y Computacional para la Modelación de Sistemas Complejos Adaptativos

**José Luis Gordillo Ruiz**
Universidad Nacional Autónoma de México, Centro de Ciencias de la Complejidad Ciudad Universitaria, Ciudad de México, México.
ORCID: 0009-0000-1267-6647

**Christopher Rhodes Stephens**
Universidad Nacional Autónoma de México, Centro de Ciencias de la Complejidad / Instituto de Ciencias Nucleares. Ciudad Universitaria, Ciudad de México, México.
ORCID: 0000-0002-2491-606X

Recepción: 02 de octubre de 2024.
Aceptación: 14 de noviembre de 2024.

Diciembre 2024 • número de revista 11 • https://doi.org/10.22201/dgtic.26832968e.2024.11.43

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

# Marco de Trabajo Tecnológico y Computacional para la Modelación de Sistemas Complejos Adaptativos

## Resumen

Todos los grandes problemas, tanto nacionales como internacionales, tales como el cambio climático, las enfermedades emergentes, las enfermedades crónico-degenerativas, la pérdida de biodiversidad y muchos más, son reconocidos universalmente como problemas "complejos". Aunque no hay un consenso sobre la definición de un Sistema Complejo Adaptativo, su multifactorialidad/multicausalidad es un factor crucial que dificulta enormemente su comprensión y predicción. En este artículo, analizamos los retos informáticos, tecnológicos, científicos y políticos que representan, presentando un marco tecnológico y computacional para su modelado en el contexto de algunas de las experiencias generadas en el desarrollo de las plataformas para el Laboratorio para la Simulación de Sistemas Complejos Adaptativos —CHILAM— del Centro de Ciencias de la Complejidad.

**Palabras Clave:** multifactorialidad, modelos explicativos, inteligencia híbrida, arquitecturas de sistemas.

# *A Technological and Computational Framework for Modelling Complex Adaptive Systems*

## *Abstract*

*All major problems, both national and international, such as climate change, emerging diseases, chronic-degenerative diseases, biodiversity loss, and many more, are universally recognized as "complex" problems. Although there is no consensus on the definition of a Complex Adaptive System, its multifactoriality/multicausality is a crucial factor that makes it extremely difficult to understand and predict. In this article, we analyze the computational, technological, scientific and*

*political challenges they represent, presenting a technological and computational framework for their modeling in the context of some of the experiences generated in the development of the platforms for the Laboratory for the Simulation of Complex Adaptive Systems —CHILAM— of the Center for Complexity Sciences.*

***Keywords:*** *multifactoriality, explainable models, hybrid intelligence, system architectures.*

## 1. Introducción

Si preguntamos: "¿Por qué hay una pandemia de obesidad?" o "¿Por qué el impacto de SARS-Cov-2 era tan diferente entre un país y otro?", aunque podríamos mencionar algunos factores que se nos ocurren, la verdad es que hay un sinfín de factores involucrados, desde la genética y epigenética hasta la economía y las políticas públicas. De hecho, sería difícil identificar una disciplina que no fuera relevante para la comprensión y predicción de estos fenómenos. En corto, son fenómenos sumamente multifactoriales y multicausales.

Aunque uno puede aceptar en lo abstracto la multifactorialidad de estos fenómenos y de otros Sistemas Complejos Adaptativos (SCA), únicamente veremos la posibilidad de comprenderlos y controlarlos mejor si logramos operacionalizar este hecho en nuestros análisis y modelos de predicción de tales sistemas. Fundamentalmente, esta operacionalización requiere superar retos no simplemente científicos, sino también de la política, administración y práctica de la ciencia y la medicina. Además, ofrece retos grandes en las tecnologías de la información y en la Inteligencia Artificial. Finalmente, requiere cambios en cómo individuos y sociedades ponemos en marcha potenciales soluciones de estos problemas complejos para incorporarlos a nuestra toma de decisiones. En este artículo, usaremos las experiencias ganadas en el desarrollo de los proyectos y plataformas del metaproyecto CHILAM del Centro de Ciencias de la Complejidad (chilam.c3.unam.mx) para ilustrar varios de los obstáculos al modelado de los SCA y el marco de trabajo utilizado para intentar superarlos.

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

## 2. Multifactorialidad

Aunque existe aún un debate sobre cuáles son todas las características que distinguen si un sistema es complejo o no [1], e incluso podemos decir que no existe una definición concluida y universalmente aceptada de la complejidad en sistemas, desde nuestro punto de vista, uno de sus aspectos más fundamentales es la multifactorialidad, que expresa que en un SCA interactúan una enorme cantidad de factores, sumamente heterogéneos en cuanto a su tipo y origen, asociados con distintas escalas espaciales, temporales y orgánicas, y que todos ellos pueden jugar un papel importante en el comportamiento del mismo.

En realidad, la multifactorialidad no es difícil de captar. Muchos la aceptamos como una propiedad de los problemas complejos, donde su existencia es evidente en una gran cantidad de fenómenos. Por ejemplo, en el área de la obesidad y las enfermedades metabólicas, en [2] se enfatiza la caracterización del fenómeno, es decir, su naturaleza multifactorial. Es importante preguntarnos qué hacer con ella. Una estrategia casi universal, tanto entre los científicos como los no-científicos, es "dividir y conquistar" —reduccionismo asociado con tratar de enfocarse en algunas variables particulares como si fueran de carácter general. Desafortunadamente, muchas veces este reduccionismo no surge de un análisis extenso de los distintos factores, buscando de forma cuantitativa y objetiva cuáles son más importantes que otros, sino, más bien, está basado en sesgos disciplinarios, donde, por ejemplo, los geneticistas se concentran en las variables genéticas, los psicólogos en las variables psicológicas, los sociólogos en las variables sociológicas, etc.

Los modelos de tipo Susceptible-Infectado-Recuperado (SIR) son otro ejemplo de la tendencia a tratar de simplificar las cosas en el contexto de las enfermedades transmisibles, donde el número pronosticado de casos de la enfermedad a tiempo $t$ depende únicamente de su número a tiempo $t-1$ y un parámetro ($R_0$). Sin embargo, es inútil tratar de entender, explicar, predecir y modificar un SCA a través de la interacción de solamente unos cuantos de los factores que lo influyen. Como ya hemos mencionado, el punto clave es operacionalizar la multifactorialidad, es decir, proporcionar métodos y herramientas que permitan su inclusión cognitiva en el intento de comprensión, análisis y predicción de un SCA.

Pongamos por ejemplo la salud pública. Ésta se compone de la salud de cada uno de los individuos de un grupo social (podemos considerar un cierto sector de la población o la

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

de un lugar en particular). La salud de un individuo se determina a través de su estado físico, estudios de laboratorio, etc., pero ¿Cuáles son las causas de su estado de salud? Podemos considerar un sinfín de factores: su alimentación, su estilo de vida, su perfil psicológico, su perfil familiar, sus particularidades genéticas, su microbiota. Éstos, a su vez, están influidos por su entorno: acceso a servicios de salud, de educación, de cultura, entretenimiento, deporte, entre muchos otros. Podemos ver entonces que se involucran factores sociales, geográficos, económicos, etc., específicos de los lugares en donde éste vive.

En el contexto de la comprensión y predicción de un SCA y su uso en la realidad, es importante considerar los requisitos necesarios. Imaginemos que se quiere predecir la probabilidad de que alguien padece de diabetes o que un municipio tendrá un aumento en el número de difuntos por Covid. En ambos casos, se puede representar una variable dependiente C e intentar relacionarla con un conjunto de predictores **X** como variables independientes. Si imaginamos que la relación entre C y **X** puede ser modelada por una probabilidad P(C|**X**) —la probabilidad de ver el valor de la variable dependiente C dado los valores de las variables independientes—, podemos preguntarnos: ¿basta tener una predicción precisa de esta cantidad? Para contestar esto, se puede imaginar un escenario en donde se informe a alguien que tiene una alta probabilidad de ser diabético en 20 años. Como individuo o médico, de igual o más importancia es la causa asociada con la predicción. ¿La probabilidad es alta por razones genéticas o por estilo de vida? Y, si es el último, ¿es más por falta de actividad física o de malnutrición? Las acciones necesarias para reducir el riesgo dependen completamente de la posibilidad de analizar las relaciones entre cada factor $X_i$ y su relación con C.

Dado este requisito, no basta el uso de un algoritmo de Inteligencia Artificial (IA), pues es una "caja negra". Al contrario, es vital que la IA sea explicable. Independientemente del algoritmo usado, hay tres propiedades de la relación entre C y Xi que deben ser consideradas: i) la fuerza de la correlación entre C y Xi, así como su grado de significado estadístico; ii) el grado de causalidad de la relación entre C y Xi —por ejemplo, si es directo, indirecto o una pura correlación; iii) el grado de accionabilidad de la variable Xi, es decir, si es una variable que puede ser sujeta a un cambio por una intervención externa. Aunque varios algoritmos de IA son capaces de ayudar con i), es la Inteligencia Humana (IH) la que es un componente esencial para ii) y iii). La combinación óptima entre la IA y la IH —la Inteligencia Híbrida— es

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

aquella que es vital para un mejor entendimiento y predicción de los SCA y que cualquier sistema de ayuda en la toma de decisiones debería tomar en cuenta.

## 3. Enfrentando la Multifactorialidad: El problema de los Datos

¿Cómo operacionalizamos la multifactorialidad? En primer lugar, en vez de tratar de producir modelos simples, debemos encontrar la forma de incluir todos los factores que sean posibles. Para cualquier tipo de modelo —simple o multifactorial—, se necesitan datos. Podemos imaginar, desde un enfoque interdisciplinario, qué información necesitamos para representar cada uno de los factores generales mencionados en los ejemplos anteriores, tanto para enfermedades metabólicas como para enfermedades transmisibles. Justamente, buscamos que sean estos datos, heterogéneos y complejos, la materia prima para crear modelos que incluyan dicha multifactorialidad. Sin embargo, involucrar estos datos en los modelos implica el enfrentar una serie de retos tecnológicos y computacionales, no solamente por su número, sino también por la multidisciplinariedad y multi-institucionalidad de su origen, la heterogeneidad de sus formatos y sus aspectos tanto sintácticos como semánticos.

### a) Multidisciplinariedad: fragmentación de la información.

Empezamos entonces por determinar qué datos necesitamos para construir nuestros modelos. Esto está influido por las preguntas que estemos planteando sobre el SCA en cuestión. Por ejemplo, si queremos averiguar si hay una correlación entre un cierto polimorfismo y obesidad, quizá solo necesitemos datos genéticos. Por el contrario, si nuestro interés es encontrar todos los factores que configuran un riesgo para la obesidad, entonces necesitaremos una gran multiplicidad de datos, tal como ejemplificamos anteriormente. Aquí encontramos una primera dificultad: la fragmentación de la información. Vivimos en un mundo que es sumamente disciplinar, en donde las organizaciones que recolectan, curan y ofrecen información se dedican exclusivamente a los datos de una disciplina particular, o en una área de responsabilidad particular en el caso de gobiernos y otras organizaciones. Existen pues bancos de datos “ómicos” (genómicos, proteómicos, etc.), bancos de datos

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

geográficos, bancos de datos socioeconómicos, pero no existe un sitio o servicio que ofrezca datos de varias disciplinas con criterios unificados de recolección y exposición. Además, no todos los datos están disponibles públicamente.

Otro elemento crucial es la cuestión del grado de conmensurabilidad de los datos. Para entender esto, primero hay que recordar que la gran mayoría de la información está indizada en una base de datos por un identificador de una persona o un lugar como elementos atómicos, por ejemplo, en los expedientes clínicos de los miembros del IMSS o los datos censales de la INEGI sobre los AGEBs o municipios de México. Puede ser, en el caso de individuos, que exista una base de datos de una población asociada con un estudio transversal llevado a cabo por el Instituto Nacional de Medicina Genómica sobre varios factores de riesgo genético para el diabetes mellitus tipo 2. Igual, puede existir una base de datos de un estudio longitudinal en el Hospital General de México asociada con una población de la CDMX con factores fisiológicos —estudios de bioquímica sanguínea y de antropometría— de una población de pacientes de ese hospital. Igualmente, existe la base de datos de la ENSANUT 2020 [3], donde hay información en extenso del consumo de alimentos de la población que participa en esta encuesta. El punto importante es que no existe una base de datos sobre una población que contenga la información detallada de estos tres estudios independientes, aunque puede ser que los tres diferentes estudios tengan una meta en común: entender los factores de riesgo de la diabetes mellitus tipo 2 en México. Es por eso que estos estudios no son conmensurables, pues cada estudio ofrece nada más que una versión muy simplificada de la realidad compleja de la enfermedad.

Es un reto enorme de las políticas públicas, tanto dentro como fuera de la ciencia, tomar una perspectiva multifactorial que rete y rebase las fronteras disciplinarias y rompa las paredes entre instituciones. Primero, requiere una coordinación entre diferentes actores e instituciones que simplemente no existe actualmente. Segundo, requiere que los investigadores identifiquen y reduzcan sus sesgos tanto disciplinarios como personales, y que subordinen sus propios intereses al beneficio del todo. Desafortunadamente, los incentivos a la comunidad científica y médica no propician un enfoque así.

En el caso de datos espacio-temporales, el fenómeno de inconmensurabilidad también existe. Por ejemplo, una base de datos asociada con un estudio de los potenciales huéspedes o vectores de la enfermedad de Chagas en Tamaulipas no es conmensurable con una base

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

asociada a un estudio de los factores ambientales que fomentan la presencia de triatominos en Chiapas.

Afortunadamente, en el caso de datos espacio-temporales, existen varias bases de datos públicas con cobertura nacional. Algunas de las organizaciones que han servido como fuentes de los datos (ver Tabla 1) que hemos incorporado a las diferentes plataformas del Laboratorio Chilam son:

- CONABIO. Sistema SNIB, que contiene los registros de observaciones de diferentes especies a lo largo de todo el territorio de México [3].
- Worldclim. Bases de datos de información sobre el clima (temperaturas, niveles de precipitación, entre otros) a nivel planetario y con una muy alta resolución [4].
- INEGI. En particular, se han utilizado los Censos 2010 y 2020 [5].
- CONEVAL. Datos de medición de pobreza [6].
- Secretaría de Salud. Datos abiertos de la pandemia de Covid19 [7].
- Secretaría del Bienestar. Datos sobre programas sociales de apoyo al campo [8].

Vale la pena enfatizar la pertinencia de que exista información pública y procesable de fenómenos emergentes. Por ejemplo, la información disponible acerca de los casos de SARS-CoV2 en México permitió la creación de las plataformas EPI-PUMA 1.0 [9] y EPI-PUMA 2.0 [10], en las cuales se pudieron construir modelos predictivos de diversos fenómenos relacionados con la epidemia, por ejemplo, predicción de su evolución territorial, o modelos de riesgos para personas de acuerdo a sus perfiles de edad, presencia de comorbilidades, lugar de residencia y otros factores sociales.

Es importante enfatizar que ninguna de estas bases de datos fue armada por una organización con un propósito que tomara en cuenta los propósitos de las demás organizaciones. El hecho de que tengan traslapes significativos en espacio y tiempo permite que puedan ser potencialmente integradas para dar una perspectiva mucho más multifactorial de un lugar como un municipio.

Cuando no existen los datos requeridos, constituye un reto colosal generarlos. Como casos específicos, podemos citar el proyecto “Project 42” [11] del Laboratorio Chilam y el Atlas de Enfermedades Infecciosas del Instituto de Ciencias de la Atmósfera y Cambio Climático

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

[12]. En el primero, se ha conseguido reunir información sumamente amplia de distintos grupos de individuos. La información recabada incluye datos antropométricos, familiares, de hábitos, clínicos, genéticos y psicológicos, entre varios otros. En el segundo, se ha recabado información de casos de enfermedades infecciosas que incluye al patógeno, los hospederos y los vectores. Entre las enfermedades incluidas están Chagas, Lyme, Chikunguya y leishmaniasis, entre varias más.

**Tabla 1.** Información de algunas fuentes de datos usadas en las plataformas del Laboratorio Chilam.

| Datos | Origen | Número de variables | Tipo de valores | Número de registros |
|---|---|---|---|---|
| Encuesta Trabajadores/UNAM | C3/UNAM | 855 | Diversos tipos | 1075 |
| Encuesta UNAM/Ibero/UGto/UniSalud | C3/UNAM | 523 | Diversos tipos | 721 |
| SNIB | CONABIO | >30 | Categórico (+60 mil valores) | más de 15 millones |
| WorldClim | WorldClim | 19 | Numéricos / espaciales | 238324 |
| Censo2020, Censo2010 | INEGI | 221 | Diversos tipos / espaciales | 545650 |
| COVID DGE | Secretaría de Salud | 30 | Diversos tipos | más de 18 millones |

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

| Datos | Origen | Número de variables | Tipo de valores | Número de registros |
|---|---|---|---|---|
| Atlas de Enfermedades Infecciosas | ICACC/UNAM | > 20 | Categórico (136 valores) | 72580 |

### b) Heterogeneidad: La Transformación de Datos.

Después de conseguir los datos, el reto es encontrar la manera de relacionar aquellos que son de distinta naturaleza en varios sentidos. Los datos multifactoriales no solamente son de distinto tipo (categóricos, numéricos, series de tiempo, etc.), sino también de distintas escalas en alcance, espacio y tiempo (individuales o poblaciones, de una localidad particular, o regionales, semanales, históricos, etc.). Algunos datos son medidas de extensión y otros de intensidad. ¿Cómo podemos involucrar en un mismo modelo el número de camas de hospital de un municipio y el porcentaje de casas con agua potable? En nuestra estrategia, que está basada en las metodologías de los clasificadores Bayesianos [13], la clave es construir modelos clasificadores, para los que, en primer lugar, hay que definir un ensamble *o* conjunto de entidades. Este ensamble es primordial para la construcción de los modelos, porque constituye la forma en que transformamos datos de múltiples dominios (en un sentido matemático) a datos en un único dominio, que es el ensamble. De esta forma, todos los datos se vuelven conmensurables. Esta conmensurabilidad consiste en co-ocurrencias, es decir, dos variables ocurriendo en el mismo elemento del ensamble.

Los ensambles, en sí mismos, son un conjunto abstracto, pero en Chilam hemos encontrado que es útil para la semántica de cada aplicación conceptualizarlos en dos géneros bien diferenciados: de objetos (típicamente personas) y de lugares (espaciales). Los primeros se utilizan en el estudio de problemas en donde la pregunta principal es "¿a quién?". Los ensambles mismos surgen a partir de los datos, cuya naturaleza es discreta, de modo que su construcción es muy directa ya que el ensamble suele ser el mismo conjunto de entidades que constituyen el conjunto de datos. Por ejemplo, en modelos para Covid, el

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

ensamble es el grupo de personas cuyas pruebas y resultados están registrados en los datos que proporciona la Secretaría de Salud.

Los ensambles de lugares se utilizan en el estudio de problemas ecológicos y/o geográficos, en los que la pregunta principal es "¿en dónde?". En este caso, el ensamble no surge directamente de los datos, sino de la región espacial sobre la que se hace la investigación, por ejemplo, el territorio de México. La forma en que se construye el ensamble es partiendo el territorio en un conjunto de celdas regulares (en resoluciones de algunos $km^2$) o irregulares (por divisiones políticas, como AGEBs o municipios) [14]. Para poder convertir un conjunto de datos en un ensamble de lugares, es necesario que cada dato en el conjunto contenga una referencia espacial (coordenadas, municipios, etc.). Una aclaración pertinente es por qué no hacemos que las propias coordenadas contenidas en los datos sean los elementos del ensamble. Las razones son: a) no todos los conjuntos de datos con referencias espaciales están a una resolución tal que nos permitan hacer eso; b) más importante todavía, aunque todos los datos estuvieran a una resolución máxima, ¿cuántas veces podríamos observar la co-ocurrencia de dos variables exactamente en el mismo punto (medido al menos en grados y minutos)? La respuesta es que casi ninguna. Así pues, en los casos de ensambles de lugares, es necesario transformar todos los datos que queremos involucrar en una resolución común. Esto se consigue mediante alguna función de agrupamiento (conteos, promedios, máximos, etc.). A este proceso lo denominamos "*coarse graining* espacial".

Generalmente, es recomendable modelar una resolución asociada con los datos de menor resolución. Por ejemplo, si se quieren combinar los datos del clima usando WorldClim, que están al nivel de un pixel de un ráster, con datos socio-económicos y demográficos del censo, que están al nivel de AGEB o municipio, es necesario convertir los datos de mayor resolución, en este caso pixeles, en los de menor resolución. Si se toma una variable en particular, como temperatura promedio anual, hay múltiples maneras en que se puede mapear esta variable en un municipio. Por ejemplo, se puede tomar el valor promedio de esa variable en el municipio, su varianza o cualquier otra propiedad que se pueda derivar de la distribución de valores en el grupo de pixeles que lo conforman. Es importante notar que no es posible mapear de mayor hasta menor resolución sin hacer supuestos. Por ejemplo, sabiendo el municipio en que un individuo reside, es posible asociar una variable del censo

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

al nivel municipal a esa persona. Sin embargo, uno está asumiendo que cada persona en el mismo municipio cuenta con el mismo valor de esa variable.


**Figura 1.** Visualización de la conversión de registros de observaciones de la especie *lynx rufus* a un ensamble espacial [4].

Por otro lado, hay que hacer otra consideración: la temperatura promedio de un elemento de un ensamble espacial puede ser, digamos, 14.453 °C. La estatura de una persona puede ser, digamos, 1.654 metros. ¿Cuántos lugares y cuántas personas puede haber en el ensamble que tengan esa misma configuración? Lo más probable es que, a esa resolución de la medición, ninguna más. Esto implica que el número de co-ocurrencias con otra variable sería 0 ó 1, y eso es inservible para el método del clasificador. Para este tipo de datos, es necesario hacer una nueva transformación, que denominamos "coarse graining de dominio", que puede ser vista como la conversión de una variable no-categórica a una categórica. Existen varios métodos, por ejemplo, crear deciles ('n'-iles, en realidad), o grupos utilizados en la disciplina de donde provienen los datos (p.ej, infante, niño/a, adolescente, adulto/a, adulto/a mayor). En el caso de una variable ordinal, como temperatura o estatura, la pregunta surge: ¿cuántas categorías son mejores? Con más categorías uno puede tener una mejor representación de la relación entre una variable dependiente —una clase— y una independiente categorizada. Sin embargo, para un número fijo de datos, uno terminará con menos datos dentro de una categoría con importantes errores de muestra correspondientes. Además, puede ser que haya mucha heterogeneidad en el número de datos en diferentes

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

categorías y, por lo tanto, cualquier estimado de la significancia estadística entre una variable dependiente y una categoría de una variable independiente será dependiente de ese número. Una manera de evitar ese problema es seleccionar categorías tales que el número de datos en cada categoría sea similar. En el caso de categorías predefinidas, esto no es posible.

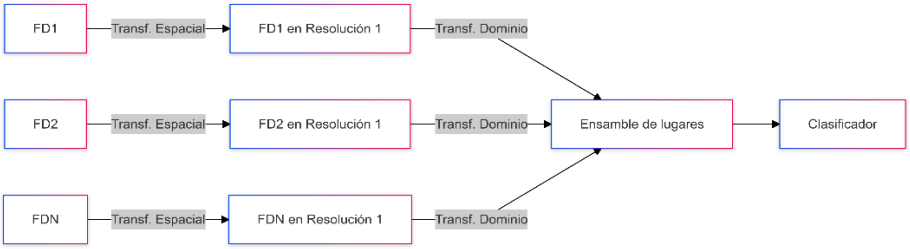


**Figura 2.** Transformación de múltiples fuentes de datos a un ensamble de lugares.

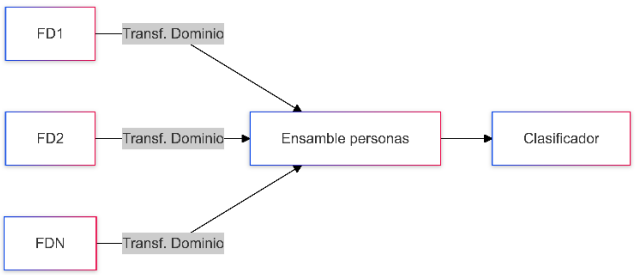


**Figura 3.** Transformación de fuentes de datos a ensambles de “personas”.

### c) Calidad de los datos.

La calidad de los datos se refiere a la cantidad relativa de errores que surgen de sus procesos de recolección y transformación. Es imposible que no existan errores en la recolección de datos. En la práctica, es casi imposible y definitivamente vano tratar de enmendar o filtrar estos errores para poder usar fuentes perfectas, pues eso requeriría una inversión muy grande en recursos humanos, lo que nunca está disponible y, de estarlo, sería mejor utilizarlos en procesos más productivos.

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

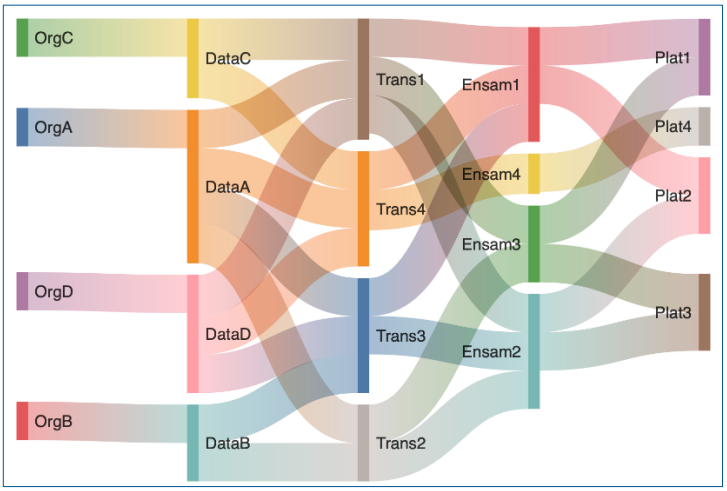


**Figura 4.** Esquema del flujo de información desde las fuentes originales hasta las plataformas.

La transformación (los procesos de *coarse graining)* requiere una buena cantidad de recursos humanos. En la Figura 4, podemos apreciar de forma muy esquemática la cantidad de procesos que surgen de transformar varias fuentes de datos de diferentes maneras y su uso en diferentes aplicaciones. Si bien los métodos y algoritmos para hacer las transformaciones no son complicados, es necesario determinar para cada variable (cada medición contenida en las fuentes de datos) su naturaleza y la forma en que ésta debe ser tratada en la transformación, así como la manera en que se deben manejar casos límite. Por ejemplo, no es lo mismo transformar las mediciones "grado de estudios", "grado promedio de estudios de la población" y "porcentaje de la población con estudios de primer nivel", aunque las tres se refieren a una misma materia. Otro ejemplo son los deciles: en varios casos, en los que aparentemente la forma más natural de utilizar una variable es su transformación a deciles, simplemente no hay suficientes valores distintos para generarlos o, peor aún, en variables que tiene valores en el tiempo, habrá periodos en donde sí se puedan generar los deciles y en otros no. Otras dificultades surgen debido a las estructuras subyacentes de la información, como pueden ser entidades organizacionales a las que están referidas. Por ejemplo, los municipios en los censos 2010 y 2020 no son los mismos. Todo lo anterior constituye un reto no tanto por la dificultad de manejar cada una de las situaciones, sino porque no existe una forma algorítmica para decidir el proceso para todas las variables,

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

de modo que se convierte en un proceso semi-manual que consume una gran cantidad de tiempo, ya que el número de variables es bastante grande, como se puede ver en la Tabla 1.

## 4. Modelado de Datos Multifactoriales Conmensurables.

Hemos descrito de forma muy simplificada todo un proceso de selección, recolección y transformación de datos para llegar a conjuntos de datos conmensurables. Esta conmensurabilidad tiene dos importantes componentes: i) que las fuentes originales son conmensurables —es decir, que corresponden a la misma población en el ensamble de personas o la misma región geográfica y en el ensamble de lugares; ii) que las resoluciones de los datos son conmensurables —en otros términos, que cada variable tiene un valor que se puede asignar a cada elemento atómico del ensamble. En el caso de una población, donde el elemento atómico es una persona, significa que cada variable es asignable a una persona. Si el nivel de agregación es a nivel de una familia, requiere que cada variable pueda estar asignada a ésta. Si los datos originales son de individuos, esto requiere un *coarse graining*. En el caso de una región geográfica, requiere que cada variable sea asignable al nivel de un elemento del ensamble, sea AGEB, municipio o una celda regular.

Habiendo establecido los requisitos respecto a los datos mismos, hace falta aún hablar del paso final en la operacionalidad de la multifactorialidad: ¿cómo usar todos estos datos? Esto requiere que se tome un conjunto de variables independientes, representado, por ejemplo, como un vector $\mathbf{X} = (X_1, X_2..., XN)$ y se use para predecir una variable dependiente C. Hemos enfatizado que la perspectiva de los clasificadores Bayesianos ofrece un marco universal para hacer esto a través del cálculo de $P(C|\mathbf{X})$, la probabilidad condicional para ver un valor (categoría) de la variable dependiente dado un vector de valores (categorías) de las variables independientes. Hay varios algoritmos de Aprendizaje de Máquina que pueden ser usados para el cálculo de $P(C|\mathbf{X})$. En CHILAM, se usa la aproximación de Bayes Ingenuo donde se usa el teorema de Bayes para relacionar el posterior $P(C|\mathbf{X})$ a la verosimilitud $P(\mathbf{X}|C)$, haciendo la suposición de que los componentes $X_i$ del vector son independientes respecto a la clase C. Este algoritmo tiene la gran virtud de facilitar el cumplimiento de las tres características mencionadas en la sección 2. En primer lugar, asigna un *score* a cada valor (categoría) de cada

variable y su significado estadístico. Segundo, por ser transparente, variable por variable permite al modelador analizar el potencial de causalidad y accionabilidad de cada una.

## 5. Modelado de Sistemas Complejos Adaptativos como Plataforma-como-un-Servicio

Uno de los objetivos del Laboratorio Chilam es construir plataformas en donde las personas puedan tanto aprovechar los datos como los algoritmos correspondientes para hacer investigación de SCA sin que sean expertos en IA.

La primera consideración es acerca de cómo conectamos las diferentes fuentes de datos con las diferentes plataformas. Para esto, lo primero que debemos considerar es cómo organizamos los datos, para lo cual es necesario tener en cuenta al menos cuatro aspectos:

1. Flexibilidad para acceder a datos que son completamente heterogéneos.
2. Eficiencia en la lectura y procesamiento de los datos.
3. Simplicidad para la modificación y adición de nuevos datos.

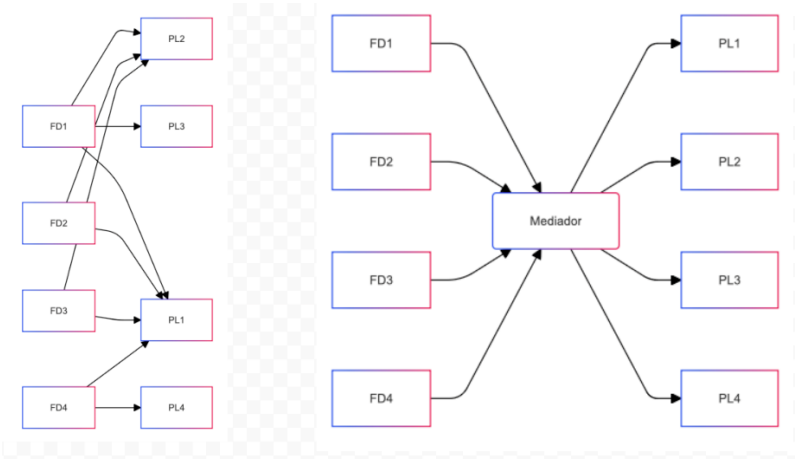


**Figura 5.** Integración de datos a plataformas con y sin mediador.

Estos aspectos configuran una gama de opciones, en cuyos extremos podríamos mencionar, por un lado, el agrupar a todos los datos en una única base de datos e, incluso,

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

en una única relación (tabla) con algunos mecanismos para manejar su heterogeneidad; por el otro, mantener cada elemento de información en su propia base de datos. Durante la experimentación que hemos realizado a través del desarrollo de distintas plataformas, hemos encontrado que un elemento crucial para reducir la complejidad tecnológica en estos procesos es no usar los datos directamente, sino a través de un mediador, siendo GraphQL [15] una de las herramientas que hemos adoptado en algunos de nuestros desarrollos para este fin.

Por otro lado, ¿qué podemos esperar si ofrecemos a una persona la posibilidad de hacer una selección de entre cinco mil opciones y después le presentamos los resultados de éstas con múltiples combinaciones? Si hiciéramos una preselección, ¿no estaríamos dirigiendo al usuario hacia una dirección en particular, eliminando así la riqueza que proporciona la multifactorialidad? En cambio, dejar todos los factores posibles conduce al problema de sobrecarga de opciones, un sesgo cognitivo que ocurre cuando alguien está enfrentado con demasiadas opciones [18]. Así, un elemento muy importante, y de difícil solución, es proporcionar a los usuarios interfaces intuitivas y adecuadas para manejar la multiplicidad de datos y resultados involucrados en los modelos. Una estrategia útil es agrupar factores y permitir la selección/despliegue por grupos según una ontología particular.

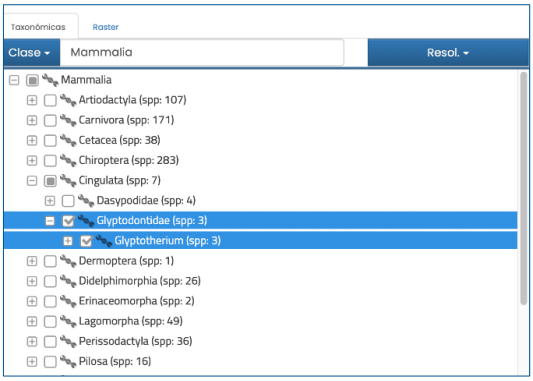


**Figura 6.** Interfaz de selección por grupos (taxonómicos). Permite la selección a diferentes niveles. También se muestra la selección por tipos ("Taxonómicas", "Ráster") [4].

Para el caso de la selección, una estrategia recurrente es agrupar factores y permitir la selección por grupos. Estos grupos se pueden formar bajo diferentes criterios: las diferentes fuentes de los datos (p.ej WorldClim, INEGI, etc.), su semántica (datos de clima, datos

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

demográficos, datos socioeconómicos, etc.), o información intrínseca de los datos (p.ej, taxonomías de especies). Al mismo tiempo, los grupos son útiles para la presentación de los modelos. Los modelos para ensambles espaciales se pueden beneficiar fácilmente de elementos visuales, principalmente mapas interactivos aumentados con escalas de colores y elementos desplegables, como el mostrado en la Figura 7 que está tomado de la plataforma SPECIES (species.conabio.gob.mx).

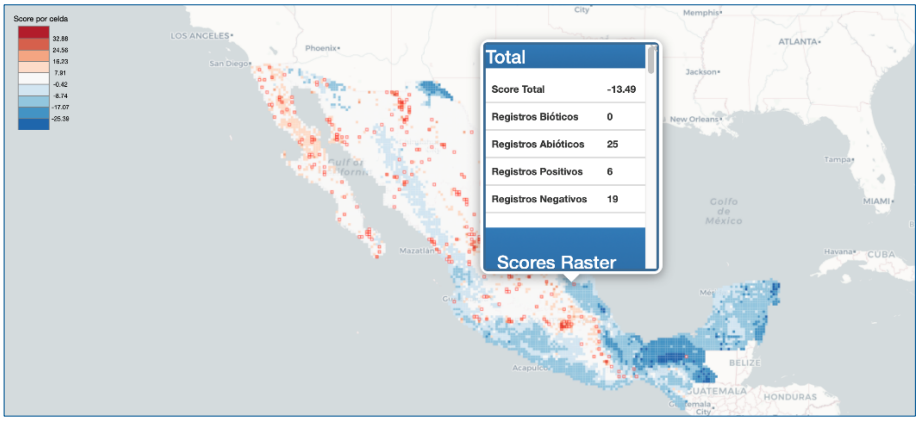


**Figura 7.** Representación gráfica de un modelo para la especie "Lynx Rufus" de la plataforma SPECIES [4].

## Ejemplo de uso del marco de trabajo: la plataforma EpISpecies.

EpISpecies es una plataforma para la modelación de enfermedades zoonóticas en México. Es útil para responder preguntas del tipo de "¿En qué zonas del país se pueden presentar brotes de alguna enfermedad zoonótica?". Las fuentes de datos que se han integrado a EpISpecies son: Atlas de enfermedades infecciosas, SNIB, WorldClim, INEGI y Clima Futuro. A partir de los datos de la Tabla 1, podemos decir que en esta plataforma se integran cerca de 250 variables numéricas, observaciones de más de 60 mil especies y varios millones de registros.

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

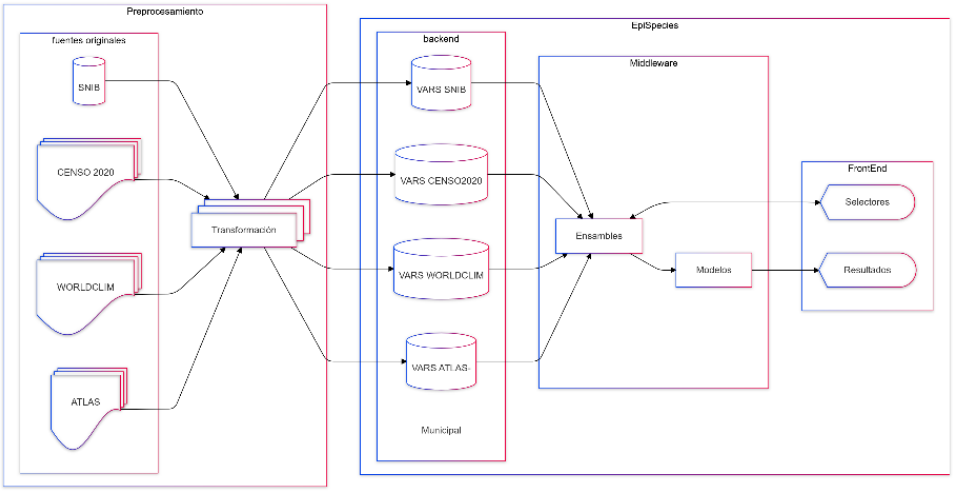


**Figura 8.** Diagrama de datos y procesos de la plataforma Epi-Species.

En la Figura 8, podemos observar de forma esquemática la aplicación del marco de trabajo en la construcción de esta plataforma. La sección de "preprocesamiento" consiste en hacer conmensurables las diferentes fuentes y tipos de datos, tomando en consideración lo discutido en la Sección 3. El resultado es un conjunto de bases de datos (sección *backend*) que contiene las representaciones de los datos en un ensamble de municipios de México. La sección *middleware* contiene dos procesos: "Ensambles" y "Modelos". El primero se encarga de tres cosas: a) descubrir las variables almacenadas en el *backend*, b) presentar estas variables al *frontend* y c) enviar la selección de variables (la clase y los variables independientes) al proceso "Modelos". Este último se encarga de construir el modelo de acuerdo a los parámetros recibidos y enviar esta información al *frontend*.

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

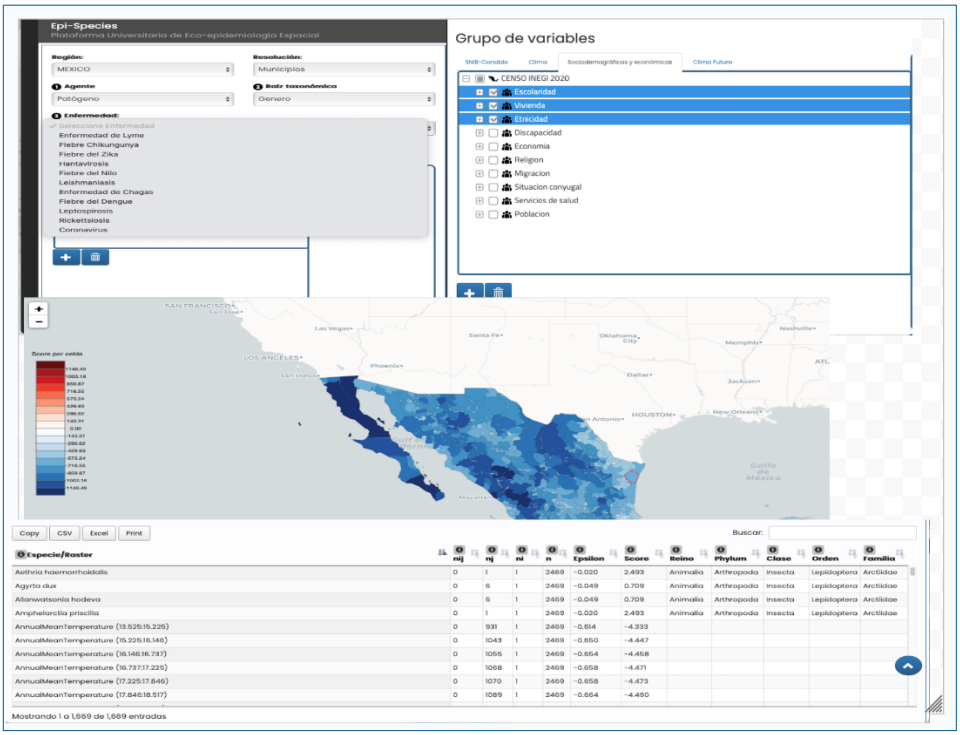


**Figura 9.** Diferentes elementos del *frontend* de EpISpecies. Arriba, se pueden observar los elementos para seleccionar la clase y las variables dependientes. Abajo, partes de los elementos que muestran los resultados del modelo.

En la Figura 9, se muestran algunos elementos del *frontend* o "interfaz de usuario" de la plataforma EpISpecies. Arriba, se pueden observar los elementos para seleccionar la clase y las variables dependientes. La clase —que en este caso es la presencia de una enfermedad— se puede escoger a partir de sus patógenos, hospederos o vectores. Las variables dependientes se pueden escoger por grupos. Por ejemplo, en la figura vemos que las variables del censo están organizadas con una semántica particular. Abajo, se muestra parte de un mapa que representa el riesgo para cada municipio, que es una de los resultados del modelo, mientras que, en la tabla, se muestra la contribución de cada variable.

Es importante señalar que, en el contexto de la creación de una multitud de distintos modelos predictivos en plataformas que contienen hasta miles de variables como predictores, hay tres principales dimensiones para la validación y evaluación de las

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

plataformas y los modelos individuales: i) en términos de las plataformas: ofrecen la flexibilidad y rapidez para poder crear un clasificador P(C|X) para un gran variedad de clases de interés C, y con un número grande de predictores X, de diferentes tipos en segundos; ii) en términos de la predictibilidad de un modelo, permiten validar el desempeño de los modelos usando métricas estándares de la clasificación, como la matriz de confusión y la curva de ROC; iii) en términos de la explicabilidad de los resultados, permiten ver la contribución predictiva explícita de cada variable Xi al clasificador, su significancia estadística y, usando la Inteligencia Humana, su interpretación y posible accionabilidad. Por ejemplo, en EpISpecies, se puede crear un modelo con target siendo casos de una enfermedad, incluyendo datos climáticos y factores bióticos, como mamíferos, que pueden ser potenciales hospederos del patógeno, como predictores. Al analizar el modelo, se puede evaluar la contribución de los factores bióticos y abióticos a su desempeño, notando que los factores bióticos son más predictivos y que un modelo que incluye los dos conjuntos de factores es aún más predictivo que un modelo con cada conjunto por separado. El modelo luego puede ser interpretado, por ejemplo, hipotetizando que los mamíferos más predictivos pueden ser hospederos del patógeno.

## 6. Conclusiones

Los grandes problemas nacionales y globales son tanto multifactoriales como multidisciplinarios. Es necesario que estas dos características deban no sólo ser tomadas en cuenta, sino ser operacionalizadas en modelos si queremos realmente entender y predecir estos problemas. Operacionalizarlas implica afrontar varios retos. En primer lugar, la fragmentación de visiones y de información que es el resultado de un contexto que por mucho fortalece más a la unidisciplinariedad. En segundo lugar, las dificultades computacionales y tecnológicas que representan la integración y utilización conjunta de una multiplicidad de datos, en cuanto a fuentes, intenciones, sintaxis y semántica. Las estrategias de IA que se adopten deben ofrecer la posibilidad de hacer conmensurable dicha multiplicidad, además de proveer modelos que no solamente sean predictivos, sino también explicativos, de modo que sea posible entender la intensidad de las relaciones causa-efecto de los diferentes factores. Los modelos clasificadores Bayesianos cumplen con ambos requisitos. A su vez, el uso de arquitecturas de sistemas adecuadas es una clave para afrontar

las dificultades tecnológicas de la incorporación de datos a diversas plataformas y aplicaciones. Por último, es necesario construir interfaces que permitan a los investigadores interactuar de forma práctica con miles de factores, al mismo tiempo que lo invitan a dejar de lado sus sesgos cognitivos en la selección de algunos de ellos.

## Agradecimientos

Estamos agradecidos por el apoyo financiero a través del proyecto DGAPA-PAPIIT IV100520.

El metaproyecto CHILAM es un resultado de los esfuerzos de más de 80 investigadores y estudiantes. Estamos muy agradecidos a todos los miembros del proyecto por su compromiso y esfuerzos.

## Referencias

[1] C. R. Stephens, "What isn't complexity?", *arXiv*, vol. 1502.03199v1, [nlin.AO], 2015. doi: 10.48550/arXiv.1502.03199.

[2] J. A. Rivera Dommarco, C. A. Aguilar Salinas, y M. Hernández Ávila, eds., *Obesidad en México: recomendaciones para una política de Estado*. Dirección General de Publicaciones y Fomento Editorial, Instituto Nacional de Salud Pública, 2015.

[3] Instituto Nacional de Salud Pública, Secretaría de Salud, *Encuesta Nacional de Salud y Nutrición (ENSANUT)*. [En línea]. Disponible en: https://ensanut.insp.mx/

[4] CONABIO, *Sistema Nacional de Información sobre Biodiversidad en México (SNIB)*. [En línea]. Disponible en: https://www.snib.mx/

[5] *WorldClim: Global climate and weather data*. [En línea]. Disponible en: https://worldclim.org

[6] INEGI, *Censo de Población y Vivienda 2020*. [En línea]. Disponible en: https://www.inegi.org.mx/programas/ccpv/2020/

[7] Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL). [En línea]. Disponible en: https://www.coneval.org.mx/

[8] Secretaría de Salud, *Datos Abiertos de la Dirección General de Epidemiología*. [En línea]. Disponible en: https://www.gob.mx/salud/documentos/datos-abiertos-152127

[9] Plataforma Universitaria de Inteligencia Epidemiológica (*EpIPUMA 1.0*). [En línea]. Disponible en: https://epipuma10.c3.unam.mx/

[10] Plataforma para el análisis epi-ecológico y epidemiológico de SARS-CoV-2. [En línea]. Disponible en: https://epipuma20.c3.unam.mx/

[11] Proyecto 42: Causas del síndrome metabólico y obesidad en México. [En línea]. Disponible en: https://project42.c3.unam.mx/landing

[12] C. Gonzalez, *Atlas de enfermedades infecciosas: Una herramienta de eco-epidemiología espacial*. [En línea]. Disponible en: https://www.atmosfera.unam.mx/eventos/atlas-de-enfermedades-infecciosas-una-herramienta-de-eco-epidemiologia-espacial/

[13] C. R. Stephens et al., *Epi-PUMA: Plataforma Universitaria de Inteligencia Epidemiológica de SARS-CoV-2, Versión 1.0*, *Revista de Tecnología e Innovación en Educación Superior*, núm. 7, marzo 2023.

[14] R. Sierra and C. R. Stephens, "Exploratory analysis of the interrelations between co-located boolean spatial features using network graphs", *International Journal of Geographical Information Science*, vol. 26, no. 3, pp. 441–468, 2012, doi: 10.1080/13658816.2011.594799.

[15] *GraphQL: A query language and execution engine*. [En línea]. Disponible en: https://spec.graphql.org/

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

[16] M. C. Hansen et al., "High-resolution global maps of 21st-century forest cover change", *Science*, vol. 342, pp. 850–853, 2013, doi: 10.1126/science.1244693.

[17] C. R. Stephens, V. Sánchez-Cordero, y C. González Salazar, "Bayesian inference of ecological interactions from spatial data", *Entropy*, vol. 19, no. 12, p. 547, 2017, doi: 10.3390/e19120547.

[18] A. Chernev, U. Böckenholt, y J. Goodman, "Choice overload: A conceptual review and meta-analysis", *J. Consum. Psychol.*, vol. 25, no. 2, pp. 333–358, 2015.