

Programa de información MAD-Mex (Sistema de Monitoreo de Datos de Actividad del Programa REDD+ México, CONABIO-CONAFOR)

Contenido

1. Antecedentes
2. Objetivo del MAD-Mex
3. Marco institucional
4. Documentación conforme al Modelo de Proceso Estadístico y Geográfico (MPEG-INEGI)
 1. Fase 1. Documentación de necesidades
 2. Fase 2. Diseño del programa de información
 1. Diseño conceptual y de los productos de información
 2. Diseño de la Construcción
 3. Diseño de la Captación
 4. Diseño del Procesamiento
 5. Diseño del Análisis de la producción
 6. Diseño de la Difusión (completar)
 3. Fase 6. Análisis de la producción. Resultados para al menos un ciclo del Programa
5. Conclusiones (completar)

Referencias bibliográficas

1. Antecedentes

Las series de INEGI documentan a alto grado de detalle temático la superficie terrestre de México por lo que han sido la base cartográfica para diversas decisiones sobre infraestructura, inversión y manejo del territorio durante varias décadas. Sin embargo, su tiempo de producción varía entre tres y cinco años y esta temporalidad no siempre se corresponde con las necesidades de política para reducir las emisiones de carbono del sector Uso de Suelo, Cambio en el Uso de Suelo y Silvicultura (USCUSS), además de que no en todos los casos de manejo del territorio se requiere un detalle de 220 clases de usos de suelo y vegetación.

Por estas dos razones en 2011 se conformó un grupo de trabajo entre el INEGI, la CONAFOR y la CONABIO. La intención fue diseñar un sistema de producción cartográfica que aunque no obtuviera un alto detalle temático sí proporcionara un mayor detalle espacial, con tiempos de producción adecuados para apoyar las decisiones públicas relacionadas como el manejo territorial, el suministro de agua para sistemas de riego, el crecimiento urbano y las políticas públicas referentes al sector forestal y agrícola.

Durante los últimos cuatro años el grupo diseñó un sistema eficiente que combina métodos automáticos de tecnología de punta con la experiencia de expertos y puede generar hasta cuatro mapas nuevos sobre el territorio en el marco de doce meses. Estos mapas tendrán la misma alta calidad de los mapas de las series INEGI, con un nivel de detalle de cincuenta a doscientas veces más alto y con un tiempo de actualización mucho menor a estos productos.

El proyecto del *Sistema de Monitoreo de Datos de Actividad del Programa REDD+ México (MAD-Mex)* inicia con la firma de un acuerdo internacional en la que México y otros países se adhieren al mecanismo REDD (*Reducing Emissions from Deforestation and Forest Degradation*) en su modalidad de conservación e incremento de los acervos de carbono forestal (REDD+ *plus*), que implica entre otras cosas generar datos de emisiones de carbono a partir de la detección de cobertura de suelo, cambio de cobertura y uso de suelo.

El presente documento describe el proceso de producción, los actores y las actividades por cubrir en este proyecto, así como los presupuestos necesarios para llevar a cabo los trabajos previstos.

2. Objetivo del MAD-Mex

El sistema MAD-Mex tiene como objetivo generar los insumos de información pertinentes, oportunos y confiables *sobre cobertura de suelo* que coadyuven a la producción de cartografía nacional. Parte esencial de este objetivo es refrendar y cumplir con los compromisos internacionales celebrados por México, en particular, el Acuerdo de París (2015), el mecanismo REDD+ y los Objetivos del Desarrollo Sostenible (ODS) referentes al sector forestal y de uso de suelo.

De manera particular el sistema MAD-Mex produce información anual sobre cobertura de suelo a nivel nacional a partir del procesamiento automático de imágenes satelitales de media (30 m) y alta resolución (de 10 a 5 m) altamente disponibles durante todo el año que son recopiladas por el área geoespacial de la CONABIO. Tal información se presenta con una alta precisión geométrica, posicional y temática y con un sistema de referencia explícito.

3. Marco institucional

El programa de información MAD-Mex cuenta con un *Grupo de trabajo* (GT) conformado por el Instituto Nacional de Estadística y Geografía (INEGI), la Comisión Nacional Forestal (CONAFOR) y la

Comisión Nacional para el uso y Conocimiento de la Biodiversidad (CONABIO). Cada institución tiene una serie de tareas previstas en la implementación nacional del sistema REDD+ – Monitoreo, Reporte y Verificación (MRV), programa que requiere de la cooperación interinstitucional de varias instancias del gobierno federal, así como del sector académico y de los representantes de las entidades federativas.

Al INEGI le corresponde la normatividad de los productos cartográficos, de los datos geoespaciales; además revisa y valida los aspectos metodológicos, técnicos y científicos de la información generada, así como su contenido, formato y presentación. Tales validaciones se enmarcan con apego a las fases del Modelo de Proceso Estadístico y Geográfico (MPEG) de este instituto. El INEGI asimismo propone elementos de control de calidad respecto al proceso de producción, tanto en los componentes automáticos como en los analógicos. También revisará que sean aplicados los estándares nacionales e internacionales pertinentes respecto a la cantidad y tipo de clases de cobertura de suelo y respecto a las características técnicas de la representación de información geoespacial, tales como la Unidad mínima de mapeo (UMM), escalas, entre otras.

La CONAFOR, por su parte, provee los elementos necesarios para elaborar el documento técnico sobre las necesidades de información geoespacial para elaborar los reportes de uso de suelo y cambio de uso de suelo y silvicultura (USCUSS), en el marco de los trabajos con el Grupo Intergubernamental de Expertos sobre el Cambio Climático (IPCC) sobre emisiones de CO₂. Asimismo, la CONAFOR contribuye con información geoespacial sobre las zonas de manejo forestal, datos sobre incendios e información similar para la generación de reportes. La CONAFOR también apoya con validación de campo e información puntual sobre lugares con cambio de uso de suelo, ya sea por deforestación, degradación, zonas de tala o cosechas de productos maderables. Es importante destacar que esta comisión contribuye con la disposición de fondos financieros para la adquisición de imágenes de satélites con fines de procesamiento o para la validación con datos de muy alta resolución donde el Grupo de trabajo lo considere necesario.

La CONABIO provee el mantenimiento técnico del sistema de procesamiento automatizado MAD-Mex y con la implementación de nuevas tecnologías y algoritmos para procesar de manera eficiente los insumos de información que reciba. Esto incluye los sensores de satélite más actualizados y disponibles, así como la tecnología de segmentación y de clasificación de los datos satelitales. Apoya a los cuerpos técnicos del INEGI y la CONAFOR en la operación del sistema de producción y aporta una serie de insumos tales como mapas de cobertura nacional o regional, datos de campo, fotos aéreas para elaborar la información geoespacial. Asimismo, la CONABIO aporta un sistema de manejo de datos satélites, la implementación del sistema de producción en la nube y el control de calidad de las imágenes satélites adquiridas en el marco de los contratos de servicios externos con la empresa *PlanetLabs* o con cualquier otro proveedor similar.

Otros actores claves para asegurar el funcionamiento y la calidad de los mapas generados por el sistema MAD-Mex – MRV son la Comisión Nacional de Áreas Naturales Protegidas (CONANP) y el Instituto Nacional de Ecología y Cambio Climático (INECC). En el mismo sentido, el Grupo de trabajo ha colaborado con cinco entidades federativas (entidades de acción temprana), con las que desarrolló entre enero de 2011 y diciembre de 2014 los estándares y soluciones técnicas que permiten el monitoreo de deforestación y degradación de forma anual a nivel nacional.

Para dar cumplimiento a los preceptos de la Ley del Sistema Nacional de Información Estadística y Geográfica (LSNIEG), el INEGI ha constituido un Comité Técnico Especializado (CTE) para revisar y validar la información de uso de suelo y vegetación (CTEUSV). En este comité se discuten y acuerdan

los temas y estándares nacionales para la elaboración de la información geoespacial de interés nacional.

En el marco de este Comité, desde 2011, se ha desarrollado la tecnología y el *Sistema Operacional de Cartografía de Alta Resolución (SOCRATES)*, que es la cara operativa del MAD-Mex.

La normatividad aplicable al proyecto MAD-Mex abarca desde tratados, convenios internacionales y leyes hasta las políticas, planes y programas de distintos niveles de gobierno.

4. Documentación conforme al Modelo de Proceso Estadístico y Geográfico (MPEG-INEGI)

4.1 Fase 1. Documentación de necesidades

La información sobre cobertura de suelo se inserta dentro de un programa de información más amplio que tiene la finalidad de contribuir al manejo integrado del territorio y cumplir con los compromisos internacionales como el Acuerdo de París, el mecanismo REDD+ y los Objetivos del Desarrollo Sostenible. Además, se busca preservar la biodiversidad y los servicios ambientales que ofrecen estos recursos, así como contribuir a la reducción de emisiones de gases de efecto invernadero generados a partir de la deforestación y degradación de los bosques. Para el caso concreto de MAD-Mex los usuarios principales de los productos del sistema son el gobierno federal y sus agencias, los gobiernos estatales y las organizaciones de la sociedad civil. Otro usuario importante será la academia nacional e internacional.

Las necesidades identificadas son las siguientes:

- Proveer información sobre cobertura de suelo definida como la capa visible para el sensor de materiales: vegetación leñosa, herbácea, agua, concreto, suelo que se extiende sobre la superficie de una unidad espacial a observar, por ejemplo: zonas forestales, zonas urbanas, cuerpos de agua, entre otras, con alta precisión temática a partir de datos satelitales de media y alta resolución, bajo un sistema de referencia explícito que ayude a la información anual disponible sobre todo el territorio nacional.
- Descripción de una localidad a partir de objetos geométricos que comparten características en común generados con medidas espectrales obtenidas de imágenes satelitales.
- La generación de información sobre cobertura debe partir de un sistema automatizado que permita procesar la información de forma estandarizada, confiable, reproducible y automática en un tiempo factible; todo esto debido a la extensión y complejidad de un país como México.
- Se requiere un sistema que procese coberturas completas del territorio con tecnología de punta a partir de una metodología como los algoritmos de aprendizaje de máquina (por ejemplo, máquinas de soporte vectorial y árboles de decisión) y algoritmos de preprocesamiento para eliminar efectos de la atmósfera sobre los valores de reflectancia, generalmente adversos, así como otros algoritmos que permitan lograr un alto grado de co-registro de las imágenes satelitales y algoritmos de segmentación para la obtención de objetos geométricos.

4.2 Fase 2. Diseño del programa de información

4.2.1 Diseño conceptual y de los productos de difusión

El diseño conceptual del MAD-Mex aborda las siguientes consideraciones:

- Es un sistema de producción de información geoespacial sobre cobertura de suelo que procesa la información de manera automatizada a través de algoritmos de alto desarrollo tecnológico con la finalidad de trabajar con grandes cantidades de información.
- La producción de coberturas de suelo se realiza a partir de modelos, metodologías y algoritmos de la comunidad de aprendizaje de máquina y percepción remota aplicados a los objetos geométricos generados. Adicionalmente el sistema tiene algoritmos de detección de cambio de cobertura que trabajan sobre las imágenes de los satélites en diferentes fechas para la obtención de objetos geométricos que se consideran potenciales polígonos de cambio y se utilizan las coberturas de suelo para determinar la naturaleza de los cambios potenciales.

Metodológicamente, el MAD-Mex presenta varios aspectos relevantes:

- Los niveles de precisión y número de clases. Este proyecto establece el número de clases con base en la Detección de necesidades. Al inicio del proyecto se trabajó con las clases de las Series del INEGI sobre Uso de Suelo y Vegetación. Las 70 clases resultantes generaba errores en el proceso automatizado de clasificación (por las especificidades de cada región), por lo que se sometieron a un proceso de agregación y de pruebas de precisión que respondieran a las necesidades de información de los reportes de la CONAFOR. El resultado fue un número de clases cercano a 35.
- La capacidad de procesamiento de imágenes. El sistema MAD-Mex tiene actualmente la capacidad de procesar imágenes de los satélites RapidEye, Sentinel2 y de toda la serie Landsat con resoluciones de 5 m, 10 m y 30 m, respectivamente. Es necesario indicar que pueden añadirse más sensores con poco esfuerzo. Las coberturas históricas disponibles de Landsat, por ejemplo, de imágenes de buena calidad para México datan del año 1996 hasta la fecha.
- El sistema genera de forma automática -a partir de las firmas espectrales medidas por los distintos canales de los sensores-, objetos geométricos que describen de forma única y definitiva una localidad que comparte características en común (unidad mínima mapeable)

4.2.2 Diseño de la construcción de los sistemas

A fin de contar con un sistema robusto capaz de procesar grandes cantidades de datos es necesario construirlo con tecnologías especializadas y actuales. Existen dos aspectos críticos en la arquitectura del sistema: el procesamiento y el almacenamiento. Una solución estratégica se encontró en los

servicios de la nube de *Amazon*¹, tanto para almacenar los insumos del sistema como el poder de cómputo para generar productos.

El sistema MAD-Mex utiliza un software de procesamiento de imágenes basado en la biblioteca de GDAL para la lectura y escritura de formatos de datos geoespaciales² y por simplicidad en el uso y capacidad de desarrollo se usan paqueterías y extensiones del lenguaje *Python*³ para el uso de tal librería. *Python* es un lenguaje de alto nivel que permite los paradigmas de programación orientada a objetos y estructurada. Fue elegido para el desarrollo del sistema MAD-Mex debido a su creciente popularidad en la comunidad científica y de desarrollo de software. Cuenta con un gran número de bibliotecas ampliamente documentadas, con soporte y actualizaciones constantes que mantienen al sistema desarrollado siempre trabajando con las tecnologías de vanguardia. Dentro de las paqueterías y extensiones que el sistema utiliza se encuentran *Rasterio*⁴, *Fiona*⁵, *Shapely*⁶ y *el cubo de datos de Australia*⁷, entre otros.

Asimismo, tanto en el trabajo como en el desarrollo de aplicaciones se requiere de un ambiente homogéneo que permita realizarlas con la menor cantidad de cambios posibles y de una forma limpia, aislada y reproducible.

- Limpia, debido a que se requieren versiones específicas de las paqueterías y extensiones para realizar el procesamiento
- Aislada, porque el sistema MAD-Mex ensambla diferente software cuando parte del mismo se encuentra ya instalado en el sistema operativo de una computadora, por lo que realizar un aislamiento de los procesos que el sistema ejecuta permite un mejor control de éstos
- Reproducible, porque en el desarrollo y producción de software es indispensable la búsqueda de errores y soluciones a los mismos de una forma iterativa lo más rápido posible.

Por las razones anteriores el sistema MAD-Mex utiliza Docker⁸, una herramienta ampliamente utilizada por la comunidad de desarrollo de software y Sun Grid Engine⁹, Slurm¹⁰, Dask¹¹ y Kubernetes¹² para el despliegue de los contenedores en una arquitectura de clúster de máquinas (la elección del distribuidor de trabajo depende del clúster o máquinas que se utilicen en el procesamiento).

Además, la información y datos que el sistema maneja requiere una forma de consultarla y alojarla, por esta razón el sistema utiliza una base de datos PostgreSQL¹³ con extensión postGIS¹⁴ para realizar consultas espaciales.

¹ Amazon Web Services, 2019. <https://aws.amazon.com/>

² Geospatial Data Abstraction Library, GDAL, 2019. <https://www.gdal.org/>

³ Python, 2019. <https://www.python.org/>

⁴ Rasterio, 2019. <https://rasterio.readthedocs.io/en/stable/>

⁵ Fiona, 2019. <https://fiona.readthedocs.io/en/latest/README.html>

⁶ Shapely, 2019. <https://shapely.readthedocs.io/en/latest/>

⁷ Datacube, 2019 <https://datacube-core.readthedocs.io/en/latest/>

⁸ Docker, 2019. <https://www.docker.com/>

⁹ Sun Grid Engine, 2019. <http://gridscheduler.sourceforge.net/>

¹⁰ Slurm, 2019. <https://slurm.schedmd.com/documentation.html>

¹¹ Dask, 2019. <https://docs.dask.org/en/latest/>

¹² Kubernetes, 2019. <https://kubernetes.io/>

¹³ PostgreSQL, 2019. <https://www.postgresql.org/>

¹⁴ PostGis, 2019. <https://postgis.net/>

La construcción del sistema es colaborativa. Para tener acceso a la última versión estable del código y versiones de prueba, se eligió GitHub¹⁵ como plataforma de administración y control de versiones. Todos los desarrolladores tienen acceso y permisos de edición al repositorio de MAD-Mex para la integración continua de algoritmos y corrección de “bugs” del sistema. Con plataformas como GitHub la colaboración internacional es particularmente fácil y los procesos de depuración, ampliación y entrega de versiones estables se establecen como metas a alcanzar en un tiempo determinado.

En el desarrollo de sistemas existe la llamada Integración Continua. Se trata de un modelo informático que consiste en hacer integraciones automáticas de un proyecto lo más a menudo posible con la finalidad de detectar fallos cuanto antes. Se trata de una parte importante del desarrollo de software y tiene varias ventajas como un ciclo de desarrollo más rápido y mayor calidad de software. Travis¹⁶ es un servicio que proporciona integración continua y probará el proyecto de software para cada cambio. Su servicio proporciona entornos de prueba para muchos lenguajes de programación entre ellos Python y es utilizado por el sistema MAD-Mex.

Para conocer las capacidades de procesamiento, ver el **Anexo 1**.

4.2.3 Diseño de la captación de la información

Los modelos y métodos que el sistema MAD-Mex utiliza requieren de insumos como imágenes de satélite, datos de entrenamiento y regiones definidas para el procesamiento. Se definen cada uno de estos insumos:

- Imágenes de satélite¹⁷

El desarrollo de los sensores remotos ha permitido la obtención automática de datos de una manera cada vez más rápida. Los sensores se han afinado y cada vez se obtienen datos con mayor detalle. La multitemporalidad de esta información representa la primera oportunidad para comenzar a llevar a cabo acciones de monitoreo ecosistémico a gran escala.

En junio del año 2013, la NASA publicó el primer mapa mundial de cobertura de vegetación generado a partir de información recolectada a lo largo de un año por el nuevo satélite *Suomi NPP*. Ese mismo mes la NASA misma lanzó el satélite *Aquarius* cuyo objetivo primordial es medir la salinidad de los cuerpos de agua en nuestro planeta. La cantidad de satélites para investigación científica en órbita es sumamente elevada, así como su diversidad. Las diferencias técnicas (láser, radar, por ejemplo) entre los sensores representan la captación de una alta gama de señales.

México ha estado involucrado durante varios años en esfuerzos hacia la generación de conocimiento a partir de sensores remotos. Es común que la gente utilice productos públicos, como MODIS y LANDSAT, para sus investigaciones. Los modelos digitales de elevación más actuales y más finos del país se generaron utilizando láser (LIDAR). El gobierno mexicano también ha comenzado a adquirir imágenes de alta resolución, por ejemplo, RapidEye.

En un principio, el objetivo de estas imágenes fue generar mapas de uso de clases de suelo a esta resolución. Esta tarea continúa en proceso pues los investigadores a cargo de este proyecto han

¹⁵ GitHub, 2019. <https://github.com/>

¹⁶ Travis, 2019. <https://travis-ci.org/>

¹⁷ Equihua-Benítez, 2014, Monitoreo Ecosistémico, Capítulo 1 Tesis de licenciatura, ITAM.

encontrado imprevistos a la hora de procesar esta información tan detallada. Una mayor resolución implica una mayor cantidad de información que procesar. Cada imagen RapidEye consta de 5 bandas de 25,000,000 de píxeles respectivamente para un total de 125,000,000 de píxeles. Se necesitan alrededor de 4000 imágenes para cubrir al país.

Adicionalmente, la resolución temporal de estas imágenes es aún bastante pobre. Se tienen imágenes del año 2011 hasta 2016, pero con un máximo de cuatro y un mínimo de dos fechas de adquisición. Los mapas de uso de suelo anteriores basados en otras imágenes satelitales (MODIS y LANDSAT) se generaron por medio de clasificaciones supervisadas que dependieron en gran medida de variables generadas a partir del estudio del comportamiento del país a lo largo de grandes intervalos de tiempo. Esta información aún no existe a una resolución más alta. A mediados del año 2011 comenzó el intento de generar un producto nacional de cambios con base en imágenes RapidEye. Debido a que la resolución temporal es aún sumamente baja, se optó por explorar métodos bitemporales de comparación de las imágenes.

La detección de cambios es un tema de mucho interés para la gente encargada de la toma de decisiones en el mundo. Es una de las únicas herramientas que permitirán medir de manera precisa urbanización, estudiar la deforestación, la reforestación, la degradación y los cambios en uso de suelo en general. Además de ayudar a estudiar dinámicas de carbono, fotosíntesis y producción primaria.

- Datos de entrenamiento:

Los modelos en el aprendizaje de máquina que generan la clasificación automática de imágenes se pueden dividir en supervisados y no supervisados. Mientras que los modelos no supervisados requieren la definición de clases de cobertura terrestre después de agrupar la similitud o disimilitud de los datos en la imagen, los enfoques supervisados necesitan la definición de todas las clases posibles antes de la clasificación¹⁸, lo que significa que el conjunto de clases debe ser definido de forma previa y exhaustiva¹⁹. La definición inapropiada de clases o la falta de alguna o algunas de ellas inevitablemente conducen a resultados imprecisos^{20 21 22}. El uso correcto de cualquier modelo supervisado requiere que todas las clases que ocurren en el área de estudio se incluyan en la etapa de entrenamiento²¹. La definición y selección de clases de cobertura terrestre han demostrado ser cruciales, sin embargo, estas no deben ser simples adaptaciones de los esquemas existentes y el

¹⁸ Di Gregorio, A., 2005. Land cover classification system: classification concepts and user manual: LCCS. Food & Agriculture Org.

¹⁹ Congalton, R.G., Green, K., 2008. Assessing the accuracy of remotely sensed data: principles and practices. CRC press.

²⁰ Foody, G.M., 2002. Hard and soft classifications by a neural network with a non-exhaustively defined set of classes. *Int. J. Remote Sens.* 23, 3853–3864.

²¹ Foody, G.M., Mathur, A., Sanchez-Hernandez, C., Boyd, D.S., 2006. Training set size requirements for the classification of a specific class. *Remote Sens. Environ.* 104, 1–14. doi:10.1016/j.rse.2006.03.004

²² Heinel, M., Walde, J., Tappeiner, G., Tappeiner, U., 2009. Classifiers vs. input variables - The drivers in image classification for land cover mapping. *Int. J. Appl. Earth Obs. Geoinformation* 11, 423–430. doi:10.1016/j.jag.2009.08.002

esquema adoptado se debe enfocar más en discriminar las clases por sus propiedades espectrales, así como por sus relaciones con la topografía o por su estacionalidad, o ambas²².

Los datos de entrenamiento se proveen al sistema en un formato ráster o *shapefile* para su registro en la base de datos. Enseguida se detalla el proceso de obtención de clases de entrenamiento.

La incorporación de un análisis discriminante, basado en los datos de imágenes reales, a usarse en el proceso de clasificación requiere del conocimiento de todas las clases presentes en el área de estudio y de la disponibilidad de datos de entrenamiento y validación independientes. Mientras que para una identificación manual de clases sobre una sola escena o un mosaico de pocas escenas, lo anterior es factible, tal proceso está completamente fuera de alcance para esfuerzos de mapeo de áreas muy grandes. Por ejemplo, una clasificación de cobertura terrestre de pared a pared para México o países de tamaño similar requiere alrededor de 130 escenas Landsat distintas o alrededor de 4000 escenas RapidEye individuales. Resulta obvio que un análisis de separabilidad de clases basado en escenas que parte de datos de entrenamiento separados manualmente no es una opción. Lo mismo resulta cierto para la generación de datos de validación. Todo lo anterior bajo el supuesto de que la clasificación debe tener un tiempo de producción, razonablemente corto.

El sistema MAD-Mex, entre otros atributos, permite la clasificación de la cobertura de la tierra y su cambio para todo México, pared a pared, a través de la clasificación automática de series de tiempo de imágenes satelitales de media a alta resolución (p.ej. Landsat, RapidEye, Sentinel).

Para ello se requirió generar una clasificación inicial de la cobertura terrestre del país partiendo de la mejor información temática y espacial disponible para la totalidad de territorio, misma que se encuentra presente en las distintas series de Uso de Suelo y Vegetación generadas por el INEGI (series II a V). Gebhardt y otros²³ generaron una agregación partiendo de las clases de Uso de Suelo y Vegetación del INEGI mediante el uso de un método iterativo, con el cual llegan a la propuesta de un esquema de clasificación de la cubierta de la tierra para todo el país de utilidad en el análisis discriminante, incorporado al procesamiento automatizado MAD-Mex.

Los mismos autores partieron de 57 clases diferentes de cobertura para la validación del producto. Un algoritmo iterativo de agregación de clases se aplicó sobre cada matriz de confusión resultante. En cada iteración, el algoritmo identificó la clase con la precisión de productor más baja y buscó dentro del mismo *tipo de covariante* la clase con los mayores errores de omisión y comisión. Con estas dos clases identificadas se realizó una agregación a una nueva clase compuesta.

Todas las muestras correspondientes a la nueva clase se re-etiquetaban y la matriz de confusión era calculada de nuevo. El procedimiento se detuvo cuando se cumplió un cierto umbral de precisión (definido como 60% en este estudio). Con cada iteración, la información a partir de la cual las clases se combinaron, con sus respectivas las exactitudes, se registraron en el protocolo para su posterior análisis. El enfoque aplicado depende en gran medida de la representación de todas las clases en ambos procedimientos: entrenamiento y validación. Sin embargo, este no fue siempre el caso ya que clases reales existentes como cobertura de la tierra en el país, pudieron ser omitidos al no existir

²³ S. Gebhardt, P. Maeda, T. Wehrmann, J. Argumedo Espinoza, and M. Schmidt, 2015. A proper Land Cover and Forest Type Classification Scheme for Mexico, *Int Arch Photogramm Remote Sens Spat. Inf Sci*, vol. XL-7/W3, pp. 383–390.

polígonos de referencia (no se registró un solo polígono persistente de dicha clase a lo largo de las series de Uso de Suelo y Vegetación del INEGI). Tal fue el caso de la *selva mediana perennifolia* (SMP).

Por otra parte, la distribución desigual del número de muestras disponibles por clase (que es función del área real ocupada) pudo influir en la agregación automatizada aplicada. Esto inevitablemente resulta en algunos errores de agrupación que no hacen sentido en términos ecológicos. Los resultados del protocolo de agregación de clase, por tanto, fueron analizados en términos de compatibilidad de clases agregadas. Por último, se definió un esquema jerárquico de clases de cobertura de la tierra (MAD-Mex), estableciendo en él la relación entre las clases resultantes y sus distintos niveles de agrupación con las clases originales de Uso de Suelo y Vegetación del INEGI. De igual manera se generaron tablas que relacionan las categorías superiores (I a III de MAD-Mex) con aquellas del IPCC y la Evaluación de los Recursos Forestales Mundiales (FRA).

- Regiones de procesamiento:

El sistema MAD-Mex ajusta los modelos de aprendizaje de máquina por regiones que se definen a partir de un *shapefile* y se registran en la base de datos para su uso por el sistema.

Para conocer los datos de entrenamiento del MAD-Mex, así como el tipo de imágenes utilizadas, ver **Anexo 2**.

4.2.4 Diseño del procesamiento de la información

Preprocesamiento

Los píxeles resultantes de la toma de imágenes por satélites requieren un paso de preprocesamiento en el que se corrijan a valores de reflectancia de superficie y puedan ser utilizadas para un análisis cuantitativo que permita el cálculo de características. En esta fase se generan máscaras de nubes.

Respecto a las herramientas utilizadas en esta etapa, el sistema MAD-Mex utiliza algoritmos desarrollados y proveídos de fuentes oficiales e internacionales: *United States Geological Survey* (USGS, <https://www.usgs.gov/>) para el caso de Landsat, para Sentinel de ESA/copernicus (<https://scihub.copernicus.eu/>) y para Rapideye se tienen metodologías desarrolladas por personal de CONABIO.

Procesamiento

- Procesamiento de series de tiempo. La fase de preprocesamiento genera máscaras de nubes que, en combinación con la corrección a reflectancia de superficie, permite utilizar la información y calcular estadísticas e índices a partir de la disponibilidad de imágenes en un año y para una determinada región únicamente para aquellos píxeles libres de nubes. Se utilizan también conjuntos de datos auxiliares como modelos digitales de elevación o de terreno (DEM/DTM).
- Procesamiento basado en objetos. Con base en las medidas de similitud se generan objetos a partir de respuestas espectrales calculadas directamente de las distintas bandas que componen una escena y/o a través de índices como el NDVI y el NDMI. El sistema MAD-Mex

cuenta con algoritmos propietarios como BIS²⁴ y otros de código abierto de la biblioteca OpenCV²⁵. Ambos segmentadores admiten parámetros de configuración que darán origen a las distintas geometrías que se usarán en las etapas de ajuste de modelos y clasificación.

- Segmentación. El sistema MAD-Mex tiene implementados dos algoritmos:
- Segmentación de Berkley para imágenes (BIS). Tiene un parámetro de *threshold* que determina qué tan largo(s) son los segmentos. Otros parámetros son *shaperate* y *compactness* que son pesos que determinan la forma de los segmentos. *Shaperate* pondera el atributo de la forma del objeto vs. el atributo del color. Al calcular los segmentos la tasa de *compactness* pondera el cálculo del mismo sobre el cálculo del *smoothness*. Documentación en la liga: <http://www.imageseg.com/faq>
- Segmentación de *OpenCV* para imágenes. Posee diferentes algoritmos de segmentación para imágenes ráster, imágenes de satélite de varios formatos soportados por GDAL y varios *layouts* como multiespectral o *hyperspectral*. Utiliza *OpenCV* para sus principales algoritmos y GDAL para el correspondiente manejo de los formatos. Documentación en la liga: <https://github.com/cbalint13/gdal-segment>

I. Clasificación

Aprendizaje de máquina en MAD-MEX

- Ajuste de un modelo. Los modelos de aprendizaje de máquina que están implementados en MAD-MEX son árboles de decisión y bosques aleatorios. Para este punto es importante contar con un conjunto de entrenamiento, que se detalló en cuanto su obtención para el caso de México en la sección anterior. Para el proceso de ajuste se utiliza un aprendizaje supervisado. Este paso se le conoce dentro de la comunidad de aprendizaje de máquina con el nombre de entrenamiento del modelo. El conjunto de entrenamiento en MAD-MEX se encuentra registrado en la base de datos y contiene las geometrías y atributos de cada objeto perteneciente a éste.

Cada uno de los modelos en el aprendizaje supervisado requieren que se defina el porcentaje de muestra del conjunto de entrenamiento que se utilizará para su ajuste y también se requiere que la estructura de datos se coloque en un formato específico (por ejemplo, tabular).

El proceso de ajuste del modelo utiliza un porcentaje de objetos seleccionados del conjunto de entrenamiento para obtener las etiquetas de cada uno de los objetos que se generaron con el proceso de segmentación de acuerdo a una intersección geoespacial entre ellos y la región de interés (p.ej., una entidad federativa). Una vez etiquetados los segmentos se realizan agregaciones espaciales para cada objeto (promedio, desviación estándar, mínimo, máximo). Se transforma a una estructura tabular con renglones que representan cada objeto y con columnas que representan las agregaciones espaciales realizadas sobre los canales espectrales considerados. El paquete de *scikit-learn* de *Python*

²⁴ BIS, 2019. <http://www.imageseg.com>

²⁵ OpenCV, 2019. <https://opencv.org/>

contiene una cantidad diversa de modelos de aprendizaje de máquina y es utilizado en MAD-MEX. Otra implementación que se utiliza de los árboles de decisión es con el *algoritmo C5*²⁶

- Predicción de la clase a la que cada objeto pertenece. En este paso el modelo ajustado se utiliza para etiquetar a los objetos resultado del proceso de segmentación que no fueron utilizados para el ajuste del modelo. En los pasos intermedios se realiza la conversión a formato específico de cada modelo (por ejemplo, tabular; éste fue explicado en el punto anterior) y se utilizan las funciones de predicción para un modelo seleccionado. Los resultados, es decir, los objetos etiquetados, son registrados en la base de datos. En el caso del algoritmo C5 se generan intervalos de confianza para cada una de las clases que un objeto tendría y se le asigna la clase de mayor confianza.

Para consultar los diagramas de Clasificación, ver **Anexo 3**.

Para una breve introducción al aprendizaje de máquina, ver **Anexo 4**.

Para algunas especificaciones de la base de datos geoespacial, ver **Anexo 5**.

4.2.5 Diseño del análisis de la producción

En esta fase se puede validar la clasificación realizada. Para ello es necesario un conjunto de validación independiente del conjunto de entrenamiento que fue utilizado para el ajuste del modelo; para así medir su desempeño (¿qué exactitud/precisión tiene el modelo ajustado?). Tal conjunto de validación debe tener el mismo esquema de clasificación para generar métricas utilizadas en el reporte del desempeño del modelo.

En la comunidad del aprendizaje de máquina lo importante es que un modelo generalice de tal manera que el error en el conjunto de prueba (el conjunto de prueba en el caso de MAD-MEX son los objetos por predecir su clase) y de validación sea pequeño (este último término siendo dependiente de la medida utilizada).

Por ejemplo, una medida del error del modelo es la tasa de clasificación incorrecta. Se debe tomar en cuenta que diferentes aplicaciones otorgan importancia distinta en cuanto a qué se equivoca el modelo, por ejemplo, determinar si se realizará una política para conservación en un área natural. La distinción en la importancia en este caso surge si se considera que el área natural es protegida o no lo es. Es común que para el problema de clasificación se reporte la probabilidad de que un caso pertenezca a una clase.

Otra medida típica en la medición del desempeño del modelo es el análisis de los tipos de errores en los que se incurre, siendo los anteriores nombrados como: verdaderos negativos, falsos negativos (también llamados errores de omisión), verdaderos positivos, falsos positivos (también llamados errores de comisión). Y también es común construir una matriz de confusión en la que cada renglón

²⁶ C5, 2019. <https://www.rulequest.com/see5-info.html> Consúltense en: <http://topepo.github.io/c5.0/>

tiene el número de instancias de cada clase (según los ejemplos de entrenamiento, la clase real) y cada columna tiene el número de instancias de cada clase según el clasificador.

Con esta matriz de confusión es posible obtener las tasas de falsos positivos y falsos negativos. También el desempeño de un modelo se mide de acuerdo a su sensibilidad (¿qué tan bien se detecta a los casos positivos?) y su especificidad (¿qué tan bien se descartan a los negativos? o ¿qué tan bien se detectan a los negativos?, dependiendo de la perspectiva utilizada). En este caso es útil generar una curva *Receiver Operating Characteristic* (ROC). Puede verse la documentación para las métricas del desempeño de un modelo en esta liga²⁷

Los mapas estarán sujetos a un proceso de evaluación de exactitud siguiendo un protocolo para estimar, a distintos niveles, el porcentaje de área correctamente clasificada por el sistema. Para esta validación se requiere que un conjunto de intérpretes revise las clasificaciones del mapa y evalúe si son correctas.

Siguiendo la literatura de evaluación de mapas en percepción remota^{28 29}, se selecciona una muestra con base a la cual se estiman las exactitudes del mapa, el diseño de la muestra se caracteriza por ser probabilístico y en el caso del sistema MAD-Mex es estratificado.

Los objetivos de la validación son:

1. Tener estimaciones de exactitud a total: ¿qué porcentaje del área del país fue correctamente clasificada?
2. Tener estimaciones de exactitud por clase, para cada clase se calculan las exactitudes de usuario y productor:
 - a. Exactitud de usuario: para la *i*-ésima clase ¿qué proporción del área clasificada como clase *i*, en realidad es clase *i* ?
 - b. Exactitud de productor: para la *j*-ésima clase ¿qué proporción del área que en realidad es clase *j*, fue clasificada como clase *j* ?
3. Tener estimaciones de exactitud para otros dominios espaciales, por ejemplo, estado: ¿para cada estado que porcentaje del área fue correctamente clasificada?

4.2.6 Diseño de la difusión de la información

[Completar]

La difusión de la información geoespacial que se genere estará disponible a través de servidores web desde los que se podrán descargar productos de cobertura de suelo generados por el sistema MAD-Mex. La página web es:

²⁷ <https://hal.archives-ouvertes.fr/hal-0611319/document>

²⁸ Stehman, S. V., & Foody, G. M., 2009. Accuracy assessment. In T. A. Warner, M.D. Nellis, & G. M. Foody (Eds.), *The SAGE handbook of remote sensing*. London: Sage Publications.

²⁹ Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E. and Wulder, M.A., 2014. Good Practices for Estimating Area and Assessing Accuracy of Land Change. *Remote Sensing of Environment*, 148, 42-57. <http://dx.doi.org/10.1016/j.rse.2014.02.015>

http://monitoreo.conabio.gob.mx/snmb_charts/dashboard_mapabase.html

Esta página está a disposición para los años 2017 y 2018 en el caso del satélite Landsat y Sentinel2 y para 2015 en el caso de *RapidEye* los productos de cobertura en formato ráster.

Fase 6. Fase de análisis de la producción. Resultados de al menos un ciclo del Programa

Mapa de referencia 2015

Los detalles del diseño para el mapa de referencia 2015 son los siguientes:

- Marco muestral: conjunto de polígonos homogeneizados que conforman el mapa 2015, correspondiente al procesamiento de imágenes *RapidEye*
- Unidad de observación: polígono
- Estratos: se definieron de acuerdo a las variables clase MADMEX (31 clases), entidad y el tamaño del polígono (categorizado), dando lugar a 1,680 estratos
- Tamaño y asignación de muestra: la muestra se asignó a lo largo de los estratos de tal manera que se minimizó el error estándar de la estimación de la proporción de área correctamente clasificada a total

La siguiente tabla muestra los resultados y las exactitudes a nivel nacional para distintas agregaciones de las clases. El número en el paréntesis indica el número de clases en cada caso. Las exactitudes indican el porcentaje de área clasificada correcta a nivel nacional.

	% Área	Error est.	% Pol.	Error est.
MADMEX (31)	79.3	0.7	75.6	0.8
CONABIO- SEMARNAT-INEGI (17)	80.7	0.6	77.3	0.8
CONABIO- SEMARNAT-INEGI (14)	84.6	0.6	82	0.7
CONABIO- SEMARNAT-INEGI (4)	88.1	0.6	85.9	0.7
Boscosa / No boscosa	93.4	4.5	92.3	0.5

Reportamos también estimaciones de exactitud para distintas desagregaciones, en el caso de las 31 clases MAD-Mex las estimaciones de exactitud de usuario y productor para el porcentaje de área correctamente clasificada son:

	Usuario		Productor	
	% Área	Error est.	% Área	Error est.
Bosque de Coníferas de	87.2	3.7	50.3	12.1

Oyamel Ayarin Cedro					
Bosque de Coníferas de Pino y Tascate	79.4	2.2		88.6	1.7
Bosque de Encino y Bosque de Galeria	70.8	2.8		75.7	2.7
Chaparral	84.4	5.5		84.6	4.2
Mezquital y Submontano	66.4	6.6		47.2	4.4
Selva Baja Perennifolia y Bosque Mesofilo	77.2	8.4		74.5	5.2
Selva Baja y Mediana Subperennifolia a Galeria y Palmar Natural	92.6	2.4		86.5	3
Manglar y Peten	90.2	4.2		78.6	8.9
Selva Mediana y Alta Perennifolia	81.4	4.1		77.4	5.2
Selva Alta Subperennifolia	83.1	5.9		72.5	18.6
Selva Baja Caducifolia Subcaducifolia y Matorral Subtropical	85	1.9		84.7	2.2
Selva Mediana Caducifolia y Subcaducifolia	92.5	2.2		93	2.4
Mezquital Xerofilo Galeria y Desertico Microfilo	87.6	1.4		83.6	1.5
Matorral Crasicaule	71.8	5.8		59.7	7.5
Matorral Espinoso Tamaulipeco	75.3	5.9		69.1	5.5
Matorral Sarco-Crasicaule	88.2	4.5		85.2	4.5
Matorral Sarcocaulo	91.9	2.4		92.1	2.4
Matorral Sarco-Crasicaule de Neblina	96.1	1.6		85.3	8.5

Matorral Rosetofilo Costero	86.4	8.2		64.9	15.6
Matorral Desertico Rosetofilo	84.6	2.3		72.2	2.9
Popal	58	8.8		81.2	8.3
Tular	93.8	2.2		91.4	3.5
Vegetacion de Dunas Costeras	90.6	5.8		58.5	15
Vegetacion de Desiertos Arenosos	85.6	4.9		80.2	5.9
Vegetacion Halofila Hidrofila	57.8	7.4		90.7	4.1
Vegetacion Halofila Xerofila y Gipsofila	85.6	4.2		65.1	5.1
Pastizales	53.3	2.5		77.4	2.6
Tierras Agricolas	84.5	1.8		77.7	1.7
Urbano y Construido	83	4.2		54.8	6.2
Suelo Desnudo	74	4.4		50.2	5.1
Agua	88.6	3.4		95.2	2

La siguiente tabla muestra las estimaciones de exactitud para el porcentaje de polígonos correctamente clasificados. Nuevamente reportamos estimaciones para exactitud de usuario y productor.

	Usuario		Productor	
	% poli.	error est.	% poli.	error est.
Bosque de Coniferas de Oyamel Ayarin Cedro	87.6	3	52	11.8
Bosque de Coniferas de Pino y Tascate	75.4	2.7	83.9	2.5
Bosque de Encino y Bosque de Galeria	69	3	75.9	2.6
Chaparral	84.7	5.9	79.8	5.1
Mezquital y Submontano	72.3	6.2	49.1	5.1
Selva Baja Perennifolia y Bosque Mesofilo	82	6.5	68.7	6.6

Selva Baja y Mediana Subperennifolia a Galeria y Palmar Natural	89.7	2.9		82.2	4.5
Manglar y Peten	85.6	6.8		86.6	4.4
Selva Mediana y Alta Perennifolia	73.7	5.9		72.9	6.5
Selva Alta Subperennifolia	90	6.4		91	8.1
Selva Baja Caducifolia Subcaducifolia y Matorral Subtropical	81.8	2.3		82.6	2.6
Selva Mediana Caducifolia y Subcaducifolia	90.1	2.7		91.4	3.2
Mezquital Xerofilo Galeria y Desertico Microfilo	83.3	1.8		79	1.9
Matorral Crasicaule	71.5	5.5		54.5	8.3
Matorral Espinoso Tamaulipeco	75.8	6.4		70.8	5.9
Matorral Sarco-Crasicaule	88.9	5.6		77.7	7
Matorral Sarcocaulo	87	3.3		90.4	2.6
Matorral Sarco-Crasicaule de Neblina	95.6	2.2		78.7	11.3
Matorral Rosetofilo Costero	88.6	8.5		59.4	14.2
Matorral Desertico Rosetofilo	85.3	2.3		69.1	3.2
Popal	51.5	12.6		84.8	8.4
Tular	91.2	3.7		90.7	3.2
Vegetacion de Dunas Costeras	85.3	8		58.2	15.3
Vegetacion de Desiertos Arenosos	80.7	5.2		70.7	9.1
Vegetacion Halofila Hidrofila	53.8	8.6		84.4	7.7

Vegetacion Halofila Xerofila y Gipsofila	80.3	5.9		61.1	5.6
Pastizales	52.7	2.7		73.4	2.9
Tierras Agricolas	77.9	2.3		76.1	1.9
Urbano y Construido	73.8	5.5		53	5.4
Suelo Desnudo	66.9	6		38	4.9
Agua	80.3	4.5		92.5	2.2

Incluimos las estimaciones de porcentaje de área correctamente clasificada para las 17 clases CONABIO-SEMARNAT-INEGI:

	Usuario		Productor	
	% area	error est.	% area	error est.
Bosques de aciculifolias y escuamifolias	79.8	2.2	88.3	1.7
Bosques latifoliados	70.8	2.8	75.7	2.7
Bosque Húmedo de Montaña	77.2	8.4	74.5	5.2
Manglar y petén	90.2	4.2	78.6	8.9
Selvas húmedas	88.3	2.1	83	2.7
Selvas secas	87	1.5	86.9	1.8
Matorral alto denso	76.9	4.1	68.5	4.1
Matorral méxico	73.6	4.9	58	3.7
Matorral bajo abierto	93.2	0.8	87	1.1
Vegetación acuática menor	92.8	2.1	92.8	3.2
Vegetación de suelos arenosos	85.9	4.6	78.5	5.4
Vegetación halófila	81.1	3.5	72.8	4.6
Pastizales y otra vegetación herbácea	53.3	2.5	77.4	2.6

Tierras agrícolas	84.5	1.8		77.7	1.7
Urbano y Construido	83	4.2		54.8	6.2
Suelo desnudo	74	4.4		50.2	5.1
Agua	88.6	3.4		95.2	2

Y las estimaciones de exactitud para el porcentaje de polígonos correctamente clasificados son:

	Usuario		Productor	
	% poli.	error est.	% poli.	error est.
Bosques de aciculifolias y escuamifolias	75.7	2.6	83.7	2.5
Bosques latifoliados	69	3	75.9	2.6
Bosque Húmedo de Montaña	82	6.5	68.7	6.6
Manglar y petén	85.6	6.8	86.6	4.4
Selvas húmedas	83.1	3	78.8	3.7
Selvas secas	84.1	1.8	85	2.1
Matorral alto denso	76.2	4.4	66.4	4.6
Matorral méxico	78.4	4.5	59.2	4.1
Matorral bajo abierto	91.6	1	83.9	1.3
Vegetación acuática menor	90.6	3.4	94.4	2.4
Vegetación de suelos arenosos	81	4.9	69.5	8.3
Vegetación halófila	74.4	4.6	68.2	5.1
Pastizales y otra vegetación herbácea	52.7	2.7	73.4	2.9
Tierras agrícolas	77.9	2.3	76.1	1.9
Urbano y Construido	73.8	5.5	53	5.4
Suelo desnudo	66.9	6	38	4.9
Agua	80.3	4.5	92.5	2.2

Estimaciones para desagregación en 4 clases:

	Usuario		Productor	
	% area	error est.	% area	error est.

bosques	82.7	1.6		89.4	1.3
matorrales	93.9	0.7		85.8	1
selvas	88.8	1.2		86.8	1.4
otros	86.2	1.1		90.4	0.9

	Usuario			Productor	
	% poli.	error est.		% poli.	error est.
bosques	80.9	1.7		87.4	1.5
matorrales	92.6	0.9		82.6	1.2
selvas	85.6	1.5		84.9	1.8
otros	83.7	1.2		88.6	1

Boscosa/no boscosa (clase boscosa sin incluir matorral), tenemos nuevamente estimaciones de área correcta,

	Usuario			Productor	
	% area	error est.		% area	error est.
No boscosa	95.6	0.5		94.1	0.5
Boscosa	89.7	0.9		92.2	0.8

Y para porcentaje de polígonos correctamente clasificados.

	Usuario			Productor	
	% poli.	error est.		% poli.	error est.
No boscosa	95.1	0.6		93.3	0.6
Boscosa	87.3	1		90.6	1

Finalmente, presentamos exactitudes por estado para la desagregación en 17 clases CONABIO-SEMARNAT-INEGI:

	% área	error est.	% pol.	error est.
Aguascalientes	75.5	4.5	74	4.4
Baja_California	89.5	2.3	87.5	2.6
Baja_California_Su r	90.7	2	87.7	2.4
Campeche	87.1	3.2	83	4.9
Chiapas	81.7	3.1	79.7	3.7
Chihuahua	84.6	1.7	80	2.1

Coahuila_de_Zaragoza	83.2	2.2	82.2	2.3
Colima	83.2	4.3	79.9	4.3
Distrito_Federal	80.3	3.8	69.2	5.8
Durango	75.9	3	74.3	3.7
Guanajuato	70.4	6.8	69.9	7.1
Guerrero	82.3	3.6	76.4	5
Hidalgo	70.6	4.1	66.1	5.1
Jalisco	87.4	3.3	87.2	3.6
Mexico	82.2	4.2	69	6.7
Michoacan_de_Ocampo	73.6	5.4	68.4	5.6
Morelos	90.8	2.8	79.2	6.5
Nayarit	79.6	4.4	76.5	4.3
Nuevo_Leon	74.2	4.4	72.4	4.7
Oaxaca	82.7	3.7	81.8	4
Puebla	67.1	4.8	58.4	6.5
Queretaro	67.3	4.8	62.4	5.3
Quintana_Roo	89	3.5	85.5	4.4
San_Luis_Potosi	81.5	2.9	78.2	3.6
Sinaloa	85.3	3.8	75.5	5.7
Sonora	83.1	2.3	78.7	2.7
Tabasco	74.8	7	75.7	5.9
Tamaulipas	74.1	3.9	70.4	4.8
Tlaxcala	76.5	8.9	74.1	8.9
Veracruz_de_Ignacio_de_la_Llave	53.9	4.2	55.9	4.5
Yucatan	88	3.1	86.4	2.8
Zacatecas	77.2	3.4	71.9	4.5

5. Conclusiones

[Sin información]

6. Referencias bibliográficas

[¹] Amazon Web Services, 2019. <https://aws.amazon.com/>

[²] Geospatial Data Abstraction Library, GDAL, 2019. <https://www.gdal.org/>

[³] Python, 2019. <https://www.python.org/>

- [4] Rasterio, 2019. <https://rasterio.readthedocs.io/en/stable/>
- [5] Fiona, 2019. <https://fiona.readthedocs.io/en/latest/README.html>
- [6] Shapely, 2019. <https://shapely.readthedocs.io/en/latest/>
- [7] Datacube, 2019 <https://datacube-core.readthedocs.io/en/latest/>
- [8] Docker, 2019. <https://www.docker.com/>
- [9] Sun Grid Engine, 2019. <http://gridscheduler.sourceforge.net/>
- [10] Slurm, 2019. <https://slurm.schedmd.com/documentation.html>
- [11] Dask, 2019. <https://docs.dask.org/en/latest/>
- [12] Kubernetes, 2019. <https://kubernetes.io/>
- [13] PostgreSQL, 2019. <https://www.postgresql.org/>
- [14] PostGis, 2019. <https://postgis.net/>
- [15] GitHub, 2019. <https://github.com/>
- [16] Travis, 2019. <https://travis-ci.org/>
- [17] Equihua-Benítez, 2014. Monitoreo Ecosistémico, Capítulo 1 Tesis de licenciatura, ITAM.
- [18] Di Gregorio, A., 2005. Land cover classification system: classification concepts and user manual: LCCS. Food & Agriculture Org.
- [19] Congalton, R.G., Green, K., 2008. Assessing the accuracy of remotely sensed data: principles and practices. CRC press.
- [20] Foody, G.M., 2002. Hard and soft classifications by a neural network with a non-exhaustively defined set of classes. *Int. J. Remote Sens.* 23, 3853–3864.
- [21] Foody, G.M., Mathur, A., Sanchez-Hernandez, C., Boyd, D.S., 2006. Training set size requirements for the classification of a specific class. *Remote Sens. Environ.* 104, 1–14. doi:10.1016/j.rse.2006.03.004
- [22] Heisl, M., Walde, J., Tappeiner, G., Tappeiner, U., 2009. Classifiers vs. input variables - The drivers in image classification for land cover mapping. *Int. J. Appl. Earth Obs. Geoinformation* 11, 423–430. doi:10.1016/j.jag.2009.08.002
- [23] S. Gebhardt, P. Maeda, T. Wehrmann, J. Argumedo Espinoza, and M. Schmidt, A proper Land Cover and Forest Type Classification Scheme for Mexico, 2015. *Int Arch Photogramm Remote Sens Spat. Inf Sci*, vol. XL-7/W3, pp. 383–390.
- [24] BIS, 2019. <http://www.imageseg.com>
- [25] OpenCV, 2019. <https://opencv.org/>
- [26] C5, 2019. <https://www.rulequest.com/see5-info.html> Consúltense en: <http://topepo.github.io/c5.0/>
- [27] <https://hal.archives-ouvertes.fr/hal-0611319/document>

[28] Stehman, S. V., & Foody, G. M., 2009. Accuracy assessment. In T. A. Warner, M.D. Nellis, & G. M. Foody (Eds.), *The SAGE handbook of remote sensing*. London: Sage Publications.

[29] Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E. and Wulder, M.A., 2014. Good Practices for Estimating Area and Assessing Accuracy of Land Change. *Remote Sensing of Environment*, 148, 42-57. <http://dx.doi.org/10.1016/j.rse.2014.02.015>

Anexo 1. Capacidad de procesamiento

Referente a la capacidad de procesamiento la CONABIO cuenta con el siguiente clúster de máquinas para procesamiento.

Se cuenta con cuatro nodos de las siguientes capacidades:

- **Procesador:** 2 procesadores Intel® Xeon® E5-2670 (2.50 GHz), 10 cores, 20 hilos de ejecución en total.
- **Memoria:** 128GB
- **Discos duros:** 2 discos de 1TB, RAID 1
- **Sistema Operativo:** CentOS Linux release 7.5.1804
- **GPUs:** 2 NVIDIA® Tesla K20Xm, arquitectura Kepler, 6GB de memoria GDDR5, 2688 CUDA cores

Adicionalmente, se tienen dos nodos con más capacidad en términos de memoria:

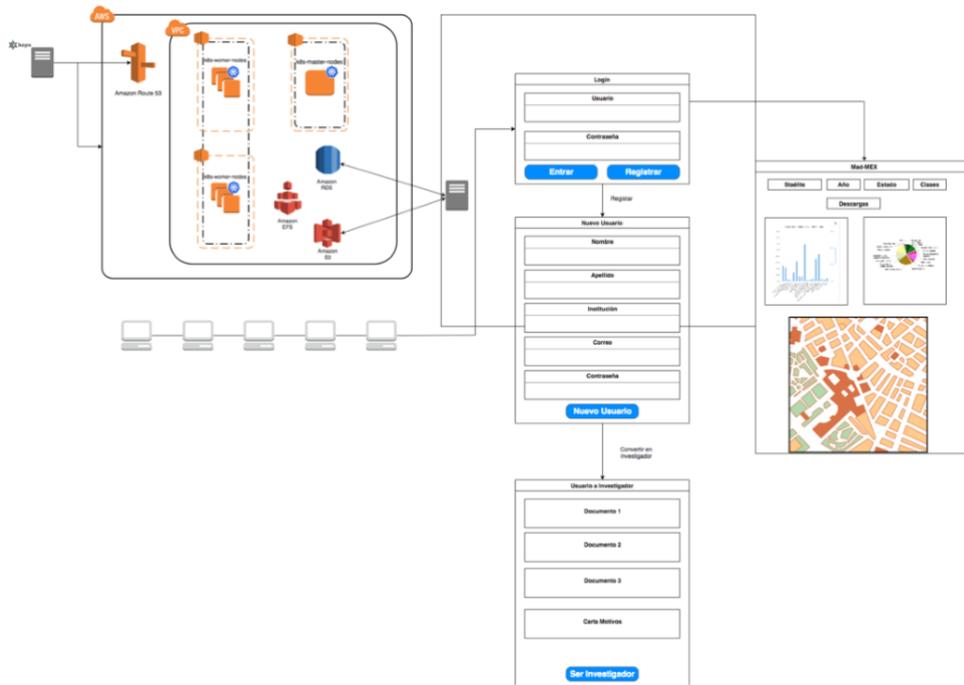
Primer nodo:

- **Procesador:** 2 procesadores Intel® Xeon® E5-2670 (2.50 GHz), 10 cores, 20 hilos de ejecución en total.
- **Memoria:** 256GB
- **Discos duros:** 2 discos de 1TB, RAID 1
- **Sistema Operativo:** CentOS Linux release 7.5.1804

Segundo nodo:

- **Procesador:** 4 procesadores Intel® Xeon® E7-4809 V3 (20M Cache, 2.00 GHz), 8 cores, 32 hilos de ejecución en total.
- **Memoria:** 512GB
- **Discos duros:** 5 discos de 1.2TB, RAID 5
- **Sistema Operativo:** CentOS Linux release 7.5.1804

En la nube de Amazon Web Services se tiene la siguiente infraestructura construida:



En ésta se utilizan servicios tales como:

- *Amazon Route 53* para hospedar el dominio: conabio-route53.net y tener en un subdominio acceso vía web browsers a ventanas de monitoreo de procesos y de uso del sistema MAD-Mex.
- *Amazon Relational Database System Engine: Aurora* database. En específico se utiliza la funcionalidad de postgresSQL con extensión postGIS para almacenar información de los productos generados e información de clases, por ejemplo, geometrías para localizar una región y esquemas de clasificación.
- *Amazon Elastic File System*: para compartir archivos entre las instancias de contenedores de docker y almacenar resultados de procesamiento del sistema.
- *Amazon S3*: para almacenamiento de diferentes niveles de procesamiento de los productos derivados de los datos de satélite, por ejemplo, los segmentos generados con los algoritmos implementados en el sistema MAD-Mex.

Anexo 2. Datos de referencia de entrenamiento e imágenes satelitales.

- De las series INEGI se toman las áreas persistentes entre todos los productos de la serie II hasta serie IV. Se ha creado un proceso que armoniza las series INEGI y crea una nueva columna llamada MAD-MEX con las clases de cobertura usando el código MAD-MEX.
- El mapa de referencia 2015, entrenamiento a partir del producto 2018, CONABIO 2017. El producto ha sido rasterizado y proyectado a una ventana con la misma extensión que las series de INEGI.
- Inventario Nacional de Manglares, CONABIO 2016. El producto se descargó del geoportal CONABIO, se proyectó a Cónica Conforme de Lambert y se proyectó a la ventana de la misma extensión de los productos INEGI.
- € Producto de Cuerpos de Agua de la superficie, Global Surface Water, *Centro Común de Investigación de la Unión Europea*. <https://global-surface-water.appspot.com>. Los píxeles con arriba de 80% de ocurrencia de agua se han filtrado y se ha creado un formato shape la cual está integrada en la base de datos. Este shape fue convertido en formato raster para integrarlo en el producto global de entrenamiento.
- € Global Urban Footprint – registro global de asentamientos (https://www.dlr.de/eoc/en/desktopdefault.aspx/tabid-11725/20508_read-47944/) con base en datos TerraSar-X el centro alemán para la observación de la tierra (Earth Observation Center) ha creado este producto con resolución de 30m para aplicaciones libres de licencias. El producto original es comercial y tiene resolución de 10m.

Todos estos productos de referencia para el entrenamiento se combinaron en un ráster único. Se han filtrado todos los píxeles cuya clase no ha cambiado. La información de agua y de manglares se mantuvo de forma fija en el ráster final, por lo que se sobre-escribieron las etiquetas de los demás productos (INEGI, MAD-Mex).

Esquema de clases (tabla)

Las etiquetas usadas para describir cada objeto y la clase a la que pertenecen son las siguientes, conforme a la tabla a las clases y claves de INEGI.

Tabla 1. Esquema jerárquico de cuatros niveles, basado en discusiones entre el proyecto, INEGI, CONABIO y CONAFOR.

NIVEL 1	NIVEL 2	Cobertura de Suelo - IIN CONABIO- SEMARNAT- INEGI	MAD-MEX ~ (31 clases)	Claves Básicas de los Tipos de Vegetación (INEGI)		
BOSQUE	BOSQUES	1	Bosques de aciculifolias y escumifolias	1	Bosque de Coníferas: Oyamel, Ayarín y Cedro	BA, BB, BS
				2	Bosque de Coníferas: de Pino y Tásate	BJ, BP, BPQ, MJ
		2	Bosques latifoliados	3	Bosque de Encino y Bosque de Galería	BG, BQ, BQP

		3	Bosque húmedo de montaña	6	Bosque Mesófilo y Selva Baja Perennifolia	BM, SBP
	SELVAS	4	Manglar y petén	8	Manglar y Petén	PT, VM
		5	Selvas húmedas	7	Selva Baja y Mediana Subperennifolia, Bosque de Galería y Palmar Natural	SBQ, SBQP, SG, SMQ, VPN
				9	Selva Mediana y Alta Perennifolia	SAP, SMP
				10	Selva Alta Subperennifolia	SAQ
		6	Selvas secas	11	Selva Baja Caducifolia y Subcaducifolia, Matorral Subtropical	MTS, SBC, SBK, SBS, VPI
				12	Selva Mediana Caducifolia y Subcaducifolia	SMC, SMS
	MATORRALES	7	Matorral alto denso	14	Matorral Crasicaule	MC
				18	Matorral Sarco-Crasicaule de Neblina	MSN
				15	Matorral Espinoso Tamaulipeco	MET
		8	Matorral méxico	4	Chaparral	ML
				5	Mezquital y Matorral Submontano	MK, MKE, MSM
		9	Matorral bajo abierto	13	Matorral Desértico Micrófilo, Mezquital Xerófilo, Vegetación de Galería	MDM, MKX, VG
				16	Matorral Sarco-Crasicaule	MSCC
				17	Matorral Sarcocaula	MSC
				19	Matorral Rosetófilo Costero	MRC
				20	Matorral Desértico Rosetófilo	MDR
	VEGETACIÓN MENOR O TIERRAS FORESTALES MENORES	10	Vegetación acuática menor	21	Popal	VA
				22	Tular	VT

		11	Vegetación de suelos arenosos	23	Vegetación de Dunas Costeras	VU
				24	Vegetación de Desiertos Arenosos	VD
		12	Vegetación halófila	25	Vegetación Halófila Hidrofila	VHH
				26	Vegetación Halófila Xerófila y Gipsófila	VH, VY
NO BOSQUE	NO BOSQUE Y OTRAS TIERRAS	13	Pastizales y otra vegetación herbácea	27	Pastizales	PC, PH, PI, PN, PY, VS, VSI, VW
		14	Tierras agrícolas	28	Tierras Agrícolas	temporal, riego, humedad
		15	Urbano y construido	29	Urbano y Construido	AH, ZU
		16	Suelo desnudo	30	Suelo Desnudo	ADV, DV
	AGUA	17	Agua	31	Agua	ACUI, H2O
2	6	17		31		

Insumos del MAD-Mex

El principal insumo del MAD-Mex es un conjunto de imágenes ópticas calibradas con una resolución espectral que abarca al menos cinco canales visibles e infrarrojos y que permiten distinguir objetos de un tamaño mínimo de 400 m². A continuación, se detallan algunos ejemplos.

Imágenes de la constelación RapidEye

Características de la misión	Información
Año de lanzamiento	2008
Cantidad de satélites	5
Altitud de la órbita	630 km solar sincrónico
Tiempo de cruce del ecuador	11:00 a.m. (tiempo local solar)
Tipo de sensor	Multi-spectral push-broom imager
Bandas espectrales	Azul 440 - 510 nm Verde 520 - 590 nm Rojo 630 - 685 nm Red Edge 690 - 730 nm NIR – 760 -850 (Infrarrojo cercano)
Tasa de muestreo en la tierra (GSD)	6.5 m
Tamaño del pixel (orto-rectificado)	5 m
Ancho de barrido	77 km
Rango dinámico de la cámara	12 bit
Productos usados para MAD-MEX: 3 A	RapidEye Ortho Product – Radiometric, sensor and geometric corrections applied to the data. The product accuracy depends on the quality of the ground control and DEMs used. Product is processed as an individual 25 km by 25 km tile.

	Para México: Uso de un MDT/DEM INEGI de 30 m, exactitud de la georeferenciación de 30 m, exactitud del corregistro entre fechas: 1 pixel = 5 m
--	---

Características del producto Orto de RapidEye

Product Attribute	Description
Product Components and Format	RapidEye Ortho image product consists of the following file components: <ul style="list-style-type: none"> · Image File – GeoTIFF file that contains image data and geolocation information · Metadata File – XML format metadata file · Browse Image File – GeoTIFF format · Unusable Data Mask (UDM) file – GeoTIFF format
Product Orientation	Map North up
Product Framing	Image Tile (image tiles are based on a worldwide, 24km by 24km fixed grid system (see Appendix B for full tile grid definition). To each 24km by 24km grid square, a 500m overlap is added to produce a 25km by 25km image tile. Image tiles are black-filled 1km beyond the order polygon used during order placement. Tiles only partially covered by an image take will be also be black-filled in areas containing no valid image data.
Pixel spacing	5 m
Bit Depth	16-bit unsigned integers
Product Size	Tile size is 25km (5000 lines) by 25km (5000 columns). 250 Mbytes per Tile for 5 bands at 5m pixel spacing.
Geometric Corrections	Sensor-related effects are corrected using sensor telemetry and a sensor model, bands are co-registered, and spacecraft-related effects are corrected using attitude telemetry and best available ephemeris data. Orthorectified using GCPs and fine DEMs (30m to 90m posting)
Horizontal Datum	WGS84
Map Projection	Universal Transverse Mercator (UTM)
Resampling Kernel	Cubic Convolution (default), MTF, or Nearest Neighbor
Acceso a la documentación técnica de Planet	https://www.planet.com/products/satellite-imagery/files/160625-RapidEye%20Image-Product-Specifications.pdf

Imágenes de la constelación Sentinel-2

Características de la misión	Información
Acceso a la información	http://esamultimedia.esa.int/docs/S2-Data_Sheet.pdf
Año de lanzamiento	S-2A: 23/06/2015; S-2B: 07/03/2017
Cantidad de satélites	Sentinel-2A y Sentinel-2B
Altitud de la órbita	786 km solar sincronico
Tiempo de cruce del ecuador	10:30 a.m. (tiempo local solar, descending node)
Tipo de sensor	Multi-spectral push-broom imager

Bandas espectrales - # - denominación - longitud de onda central - resolución espacial	Banda 1 - Coastal Aerosol - 0.443 (µm) - 60m Banda 2 - Azul - 0.490 (µm) - 10m Banda 3 - Verde - 0.560 (µm) - 10m Banda 4 - Rojo - 0.665 (µm) - 10m Banda 5 - Red Edge Vegetación - 0.705 (µm) - 20m Banda 6 - Red Edge Vegetación - 0.740 (µm) - 20m Banda 7 - Red Edge Vegetación - 0.783 (µm) - 20m Banda 8 - NIR - 0.842 (µm) - 10m Banda 8A - Red Edge Vegetación - 0.865 (µm) - 20m Banda 9 - Vapor de Agua - 0.945 (µm) - 60m Banda 10 - SWIR Cirrus - 1.375 (µm) - 60m Banda 11 - SWIR - 1.610 (µm) - 20m Banda 12 - SWIR - 2.190 (µm) - 20m
Tasa de muestreo en la tierra (GSD)	10 m (4 bandas), 20 m (6 bandas), 60 m (3 bandas)
Tamaño del pixel (orto-rectificado)	10 m, 20 m, 60 m
Ancho de barrido	290 km
Rango dinámico de la cámara	12 bit
Productos usados para MAD-MEX:	Sentinel-2 Level 2A - Orthorectified product providing Bottom-Of-Atmosphere reflectances, and basic pixel classification (including classes for different types of cloud)

Características del producto Sentinel-2A

Product Attribute	Description
Product Data Items (PDI)	<ol style="list-style-type: none"> 1. Image data (MSI data); 2. Image metadata; 3. Image quality reports including quality data indicators and quality checks; 4. Auxiliary data; 5. Satellite Ancillary data; 6. Preview Image data.
Format / compliance	<ol style="list-style-type: none"> 1. SENTINEL - SAFE (Standard Archive Format for Europe) 2. ISO 14721:2003 OAIS (Open Archival Information System) 3. compliant to the INSPIRE (Infrastructure for Spatial Information in the European Community) Metadata regulation (cfr. [EC- INSPIRE-CR] and [EC-INSPIRE-DIR])
Product Framing / Tiles	<ol style="list-style-type: none"> 1. Tile = 100kmx100km squared ortho-images in cartographic reference frame UTM/WGS84 (Universal Transverse Mercator / World Geodetic System 1984) projection.
Bit Depth	16-bit unsigned integers
Geometric Corrections	<ol style="list-style-type: none"> 2. Exactitud de la posición / geo-localación: 20 m (2 σ) sin punto de control en el suelo (GCPs) 3. Ortho-rectified products are systematically projected on UTM-UPS/WGS84 projection and tiled. A UTM tiling following the US-MGRS (US-Military Grid Reference System) grid approach is proposed:

	<ol style="list-style-type: none"> 4. The vertical UTM boundaries and horizontal latitudinal band boundaries define 6° X 8° Grid Zones. 5. The MGRS is derived from the UTM grid system and the UPS (Universal Polar Stereographic) grid system but uses a different labelling convention.
Horizontal Datum	WGS84
Map Projection	Universal Transverse Mercator (UTM)
Acceso a la documentación técnica de la ESA:	https://sentinels.copernicus.eu/documents/247904/685211/Sentinel-2-Products-Specification-Document

Imágenes de la constelación LandsatThematic Mapper (TM)

Bands	Wavelength (micrometers)	Resolution (meters)
Band 1 - Blue	0.45-0.52	30
Band 2 - Green	0.52-0.60	30
Band 3 - Red	0.63-0.69	30
Band 4 - Near Infrared (NIR)	0.76-0.90	30
Band 5 - Shortwave Infrared (SWIR) 1	1.55-1.75	30
Band 6 - Thermal	10.40-12.50	120* (30)
Band 7 - Shortwave Infrared (SWIR) 2	2.08-2.35	30

Landsat-7: Enhanced Thematic Mapper Plus (ETM+)

Bands	Wavelength (micrometers)	Resolution (meters)
Band 1 - Blue	0.45-0.52	30
Band 2 - Green	0.52-0.60	30
Band 3 - Red	0.63-0.69	30
Band 4 - Near Infrared (NIR)	0.77-0.90	30
Band 5 - Shortwave Infrared (SWIR) 1	1.55-1.75	30
Band 6 - Thermal	10.40-12.50	60 * (30)
Band 7 - Shortwave Infrared (SWIR) 2	2.09-2.35	30
Band 8 - Panchromatic	.52-.90	15

Landsat-8: Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS)

Bands	Wavelength (micrometers)	Resolution (meters)
Band 1 - Ultra Blue (coastal/aerosol)	0.435 - 0.451	30
Band 2 - Blue	0.452 - 0.512	30
Band 3 - Green	0.533 - 0.590	30
Band 4 - Red	0.636 - 0.673	30
Band 5 - Near Infrared (NIR)	0.851 - 0.879	30
Band 6 - Shortwave Infrared (SWIR) 1	1.566 - 1.651	30
Band 7 - Shortwave Infrared (SWIR) 2	2.107 - 2.294	30
Band 8 - Panchromatic	0.503 - 0.676	15
Band 9 - Cirrus	1.363 - 1.384	30
Band 10 - Thermal Infrared (TIRS) 1	10.60 - 11.19	100 * (30)

Band 11 - Thermal Infrared (TIRS) 2	11.50 - 12.51	100 * (30)
--	---------------	------------

Características del producto Reflectancia de Superficie Landsat

Parameter	Parameter Landsat 4–5, 7 (LEDAPS)	Landsat 8 OLI (LaSRC)
(Original) research grant	NASA GSFC, MEaSUREs (Masek)	NASA GSFC
Global coverage	Yes	Yes
TOA	Visible (1–5,7) + Brightness temp (6) bands	Visible (1–7, 9) + Thermal (10–11) bands
SR	Visible (1–5, 7) bands	Visible (1–7) bands (OLI/TIRS only)
Radiative transfer model	6S	Internal algorithm
Thermal correction level	TOA only	TOA only
Thermal band units	Kelvin	Kelvin
Pressure	NCEP Grid	Surface pressure is calculated internally based on the elevation
Water vapor	NCEP Grid	MODIS CMA
Air temperature	NCEP Grid	MODIS CMA
DEM	Global Climate Model DEM	Global Climate Model DEM
Ozone	OMI/TOMS	MODIS CMG Coarse resolution ozone
AOT	Correlation between chlorophyll absorption and bound water absorption of scene	MODIS CMA
Sun angle	Scene center from input metadata	Scene center from input metadata
View zenith angle	From input metadata	Hard-coded to 0
Undesirable zenith angle correction	SR not processed when solar zenith angle > 76 degrees	SR not processed when solar zenith angle > 76 degrees
Pan band processed?	No	No
XML metadata?	Yes	Yes
Brightness temperature calculated	Yes (Band 6 TM/ETM+)	Yes (Bands 10 & 11 TIRS)
Cloud mask	Internal algorithm; CFmask	Internal algorithm; CFmask
Data format	INT16	INT16
Fill values	-9999	-9999
QA bands	Cloud Adjacent cloud Cloud shadow DDV Fill Land water Snow Atmospheric opacity	Cloud Adjacent cloud Cloud shadow Aerosols Cirrus Aerosol Interpolation Flag
Acceso a la documentación técnica:	https://landsat.usgs.gov/sites/default/files/documents/ledaps_product_guide.pdf	https://landsat.usgs.gov/sites/default/files/documents/lasrc_product_guide.pdf

Anexo 3. Diagramas de clasificación

En los cuadros siguientes un diagrama se presenta en forma general y otros tres para el caso de LANDSAT, RapidEye y Sentinel.

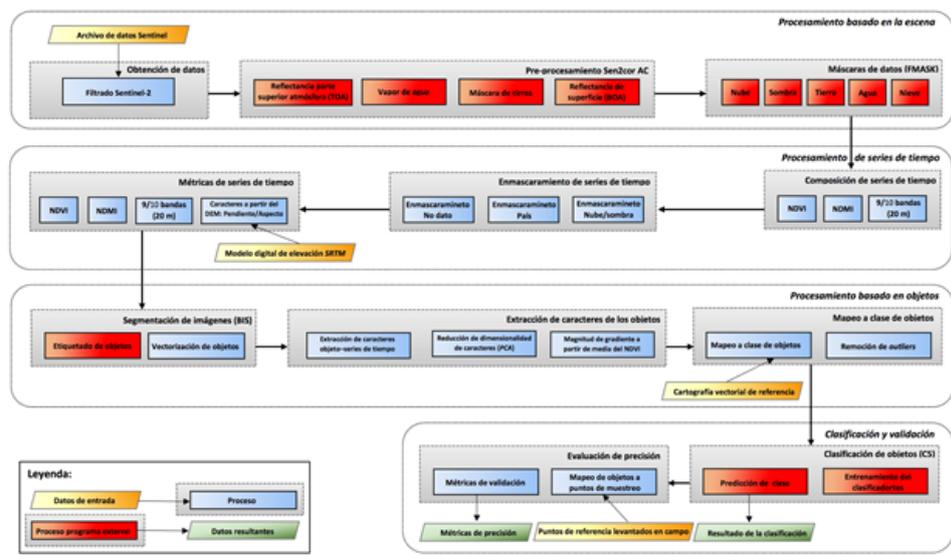


Tabla: Características y complejidad de los productos

ID	Tipo	UMM	# objetos distintos
CS 30	Cobertura de suelo	0.01 ha = 100 m ²	< 2,000,000,000 polígonos

Anexo 4. Introducción al aprendizaje de máquina

Breve introducción al aprendizaje máquina

Se dice que una máquina aprende de su experiencia E con respecto a una tarea T y función de evaluación F si su desempeño en la tarea T (con respecto a la evaluación F) mejora con la experiencia E.

- Función de evaluación: la medida de desempeño de nuestro modelo
- Función objetivo: la función que se desea aprender y el modelo es la que la define

En términos sencillos, aprender es ajustar el modelo (función objetivo) para maximizar su desempeño (función de evaluación).

En general un problema de aprendizaje debe definir:

1. Clase de tareas que se pretenden realizar
2. Función de evaluación
3. Fuente de experiencia

Ejemplos:

Predecir si un día es bueno para jugar tenis

4. Tarea: clasificar días como buenos o malos para jugar tenis.
 1. Función de evaluación: Porcentaje de días bien clasificados y mal clasificados
 2. Experiencia: historia de un mes.

Segmentar un grupo de clientes de un supermercado

5. Tarea: crear n categorías de clientes. Clasificar a cada cliente como miembro de una categoría
 1. Función de evaluación: que haya sentido estratégico en los grupos. Incrementar ventas
 2. Experiencia: transacciones en los supermercados

Estimación de la cantidad de población en una región dentro de 5 años

6. Tarea: predecir número de habitantes en la región en el futuro
7. Función de evaluación: mejoramiento de condiciones de vida de la población medido en por ejemplo abastecimiento de agua
8. Experiencia: números históricos de habitantes en la región

Experiencia

Es la información (datos) con la que se realizará el aprendizaje.

Puede sesgar el resultado y lo ideal es que los ejemplos que se utilizarán para aprender sigan la misma distribución de los ejemplos y casos reales. Por ejemplo, si se recolecta sólo información de un año será difícil estimar el número de habitantes en 5 años.

Función objetivo

Se desea aprender una función que dado un ejemplo (un dato) nos entregue un valor.

En los ejemplos anteriores se tienen:

9. Función objetivo para jugar tenis: sí es un buen día / no es un buen día (variable categórica)
10. Función objetivo para segmentar a un grupo de clientes: x cantidad de grupos y las personas menores a $z1$ años van al grupo $i1$, mayores a $z2$ años van al grupo $i2$, ... (variables categóricas)
11. Función objetivo para estimación de la cantidad de población: x cantidad de personas vivirán en la región dentro de 5 años (variable numérica)

La función objetivo determina en general:

12. Cantidad y calidad de datos de entrenamiento que deben usarse
13. Tiempo de cómputo para el aprendizaje
14. Método numérico por utilizar

¿Qué es aprendizaje de máquina?

Métodos computacionales ^(a) para aprender ^(b) datos (experiencia) con el fin de mejorar el desempeño en alguna tarea ^(c) o toma de decisión.

(a) Se habla del término computacional porque se pretende que las máquinas realicen tal objetivo. Los métodos computacionales tienen detrás un modelo o técnica y algoritmo ideado o implementado por personas.

(b) Aprender por ejemplo en el contexto de un método paramétrico consiste en estimar sus parámetros.

(c) Típicamente la tarea es predecir datos no observados (sea por el costo de obtención o se observan sólo en el futuro) y en general es modelar patrones o estructuras interesantes en los datos.

Anexo 5. Base de datos geoespacial

El registro básico en la base de datos geoespacial es un objeto geométrico que está generado con base en las firmas espectrales mediadas por los sensores usados en el año específico al cual se refiere la base de datos geoespacial. Los objetos describen de forma única y definitiva una localidad con base en su firma espectral en los canales de los sensores usados para la medición. Los objetos tienen información explícita de:

1. Extensión espacial en m^2
2. Localización con coordenadas en cada punto del polígono según el sistema geográfica (proyección y datum) aplicado

3. Etiqueta de clases (según catálogo de las 17 clases de cobertura de suelo)
4. Fecha (año)

BD geoespacial nacional	# objetos registrados	OTs en Yucatán, est.	% de reducción
# OO.	> 2,000,000,000	n/a	
# OT.	> 1,500,000,000	> 3,000,000	25%
# OT UMM5	> 400,000,000	> 2,000,000	80%

OO= Objetos Originales (distinguidos por el algoritmo de segmentación);

OT= Objetos técnicos (después de aplicar un post-procesamiento);

OT UMM5= Objetos técnicos con UMM de 0.5ha;

Otras características de la base de datos geoespacial CS30:

Título	Base de datos geoespacial con objetos que describen cobertura del suelo
presentación espacial	Formato ráster o vector
Resolución espacial	100 m ² (2017, 2018) y 5000 m ² (2015)
Resolución temporal	<ul style="list-style-type: none"> o tri-anual (años calendarios) o Multi-año: rango de años calendarios (mapas de 2, 3 o 5 años)
Exactitud geométrica	<ul style="list-style-type: none"> ● Según la mejor obtenible exactitud del producto de imágenes satelitales usada para generarla (Landsat 5, 7, SPOT-5, 6 o 7, RapidEye) Para RapidEye (2015) = 1 pixel, 6,5 m del GSD Para Sentinel: (2017 2018): 1 pixel, 10 m del GSD
Cobertura de tiempo	<p>A. 1º de enero a 31º de diciembre del año calendario respectivo</p> <p>B. resumen de datos de rango definido de años por el usuario para incrementar cobertura temporal y calidad general del producto</p>
Cobertura espacial	Territorio nacional de los Estados Unidos Mexicanos
Exactitud temática (área correctamente clasificada, exactitud de usuario)	<p>Bosque = 89.7% (0.9% error estándar) no-bosque = 95.6%, (0.5% error estándar)</p> <p>IPCC (6 clases) = agrícolas (84,5%, ee 1.8%) asentamientos (83%, ee 4.2%), forestales (95.6%, ee 0.4%) húmedales (88.6%, ee 3.4%), otras (74%, ee 4.4%), praderas (53.3%, ee 2.5%)</p>
Exactitud temática (area clasificada correctamente) por clase (exactitud del usuario / productor)	<p>Bosques de aciculifolias y escuamifolias = 79.8% / 88.3%</p> <p>Bosques latifoliados = 70.8% / 75.7%</p> <p>Bosque Húmedo de Montaña = 77.2% / 74.5%</p> <p>Manglar y petén = 90.2% / 78.6%</p>

Selvas húmedas	= 88.3% /
83%	
Selvas secas	= 87.0% /
86.9%	
Matorral alto denso	= 76.9% /
78.5%	
Matorral bajo abierto	= 93.2% /
87%	
Vegetación acuática menor	= 92.8% /
92.8%	
Vegetación de suelos arenosos	= 85.9% /
78.5%	
Vegetación halófila	= 81.1% /
72.8%	
Pastizales y otra vegetación herbácea	= 53.3% /
77.4%	
Tierras agrícolas	= 84.5% /
77.7%	
Urbano y Construido	= 83.0% /
54.8%	
Suelo desnudo	= 74.0% /
50.2%	
Agua	= 88.6% /
95.2%	
