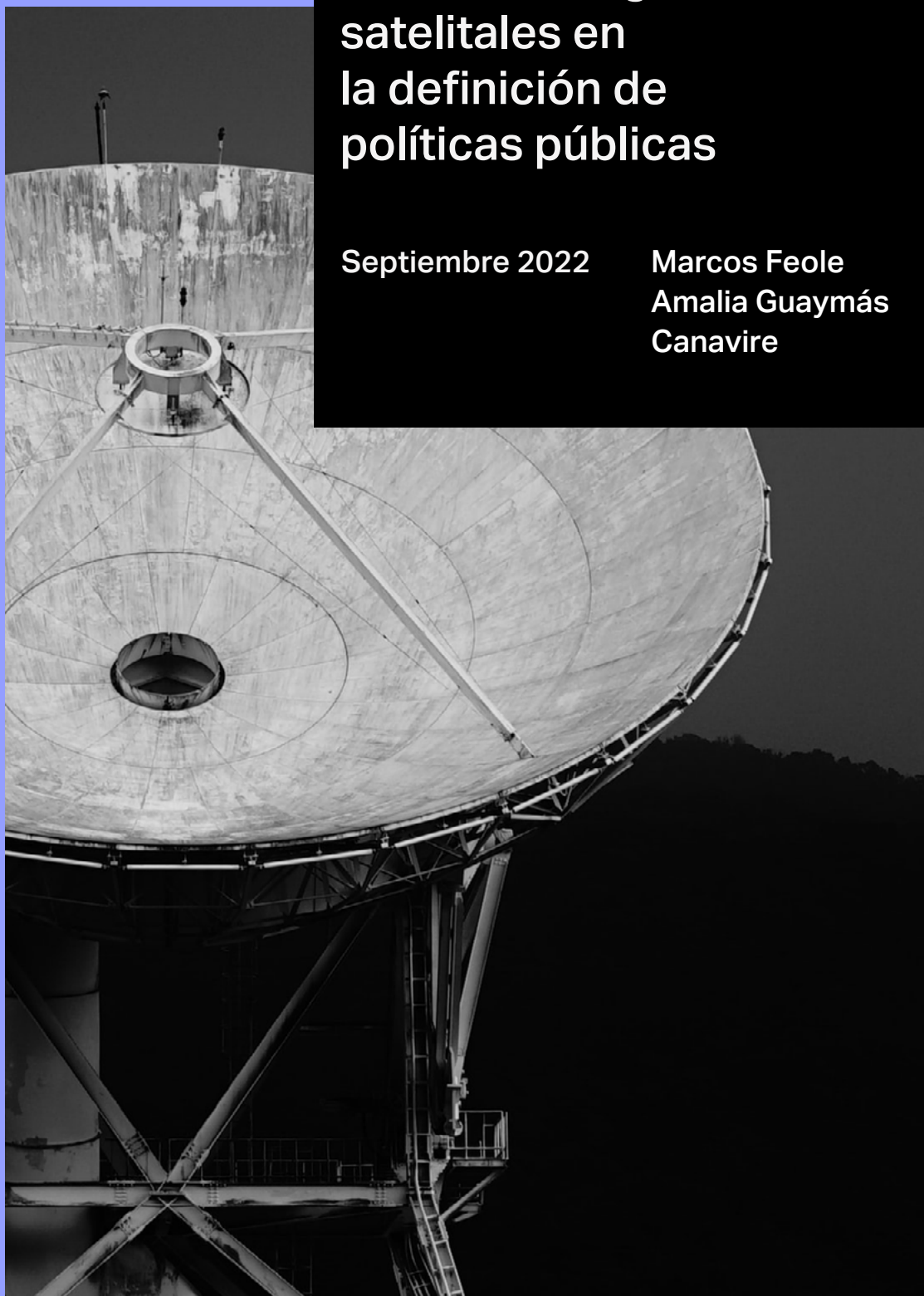


Área de
Datos

Guía práctica para el uso de imágenes satelitales en la definición de políticas públicas

Septiembre 2022

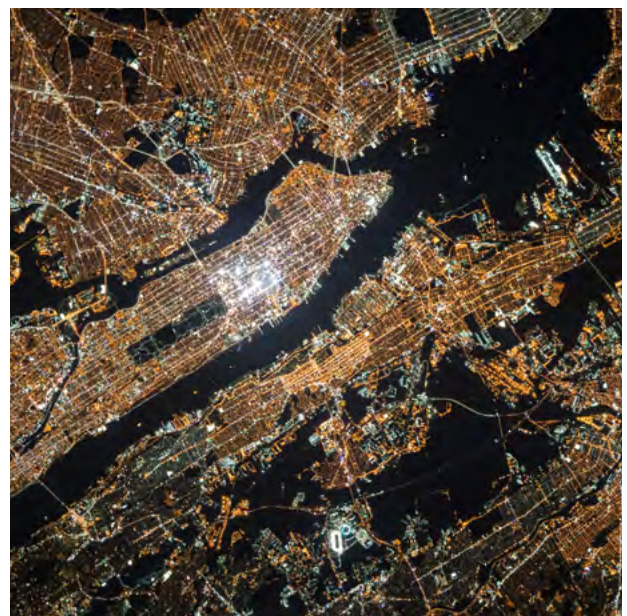
Marcos Feole
Amalia Guaymás
Canavire



Guía práctica para el uso de imágenes satelitales en la definición de políticas públicas

En la actualidad, contamos con abundantes datos de diverso origen, y en particular datos geoespaciales o georeferenciados. Esta creciente masa crítica, acompañada por el desarrollo tecnológico y de algoritmos de aprendizaje automático, permiten realizar análisis originales que pueden proveer información valiosa para el diseño de políticas públicas. Esta información puede ser inalcanzable a través de otros métodos y resulta útil para reducir incertidumbres y valiosa en la toma de decisiones. El presente documento está orientado a guiar este tipo de proyectos de manera general.

También, describe una experiencia de trabajo concreta realizada por el Área de Datos de FUNDAR en colaboración con un organismo público, cuyo objetivo fue desarrollar un modelo de detección automática de distintos tipos de suelos sobre imágenes satelitales.





¿Qué detectan los algoritmos en imágenes satelitales?

El uso de imágenes satelitales resulta particularmente conveniente a la hora de precisar información geográfica detallada, desde la escala de kilómetros hasta la escala de metros. Además, el sensoramiento remoto provee datos en tiempos pasados y actuales, lo cual permite monitorear la evolución temporal de la información geográfica. Existe ya una amplia gama de aplicaciones concretas de estos análisis, asociadas a diversas problemáticas de carácter geográfico: resultan útiles y efectivas en la detección de basurales a cielo abierto, crecimiento urbano y penetración de construcciones, asentamientos informales, calidad de la infraestructura, mapeo de cultivos y minería, estimación de la producción agropecuaria, monitoreo de precipitaciones, sequías o catástrofes naturales, análisis demográficos varios, estimación de tiempos y distancias de viaje hacia centros de salud u otros destinos de interés, entre otras aplicaciones. En otras palabras, son un insumo invaluable para el diseño de políticas públicas y la toma de decisiones.

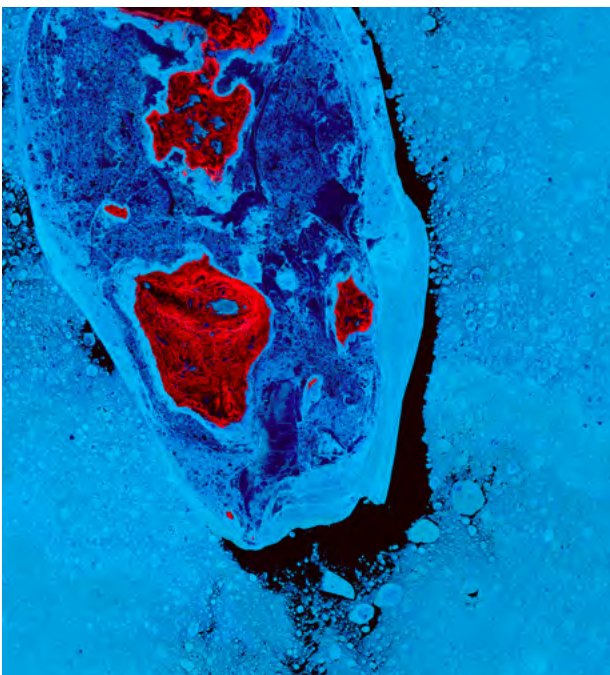


Resultan útiles y efectivos en la detección de basurales a cielo abierto, crecimiento urbano y penetración de construcciones, asentamientos informales, calidad de la infraestructura, mapeo de cultivos y minería, estimación de la producción agropecuaria, monitoreo de precipitaciones, sequías o catástrofes naturales, análisis demográficos varios, estimación de tiempos y distancias de viaje hacia centros de salud u otros destinos de interés, entre otras aplicaciones.

Trabajar en colaboración

Durante el año 2021, desde el Área de Datos de FUNDAR, se ejecutó un proyecto en colaboración con un organismo público para la detección temprana de cambios en el uso del suelo sobre imágenes satelitales de la Provincia de Buenos Aires. Se llevó a cabo una recopilación de los recursos necesarios para su realización, y se desarrolló una prueba de concepto de un modelo mínimo viable de clasificación de usos del suelo en imágenes satelitales de radar, que permitió diferenciar con una precisión razonable píxeles construidos, vegetación y masas de agua. El paso a paso de este proyecto en colaboración nos acompañará en nuestro recorrido como un ejemplo concreto del proceso del uso de imágenes satelitales. Cuando hablamos de “pasos” en este documento, tengamos en cuenta que es a fines didácticos: un uso correcto de las imágenes satelitales debe contemplar la posibilidad de estar integrando en un proceso continuo todos los pasos que vamos a describir.

Guía paso a paso

**1** →

PASO 1

Definir el problema

Como en toda actividad del conocimiento, es fundamental definir el problema que queremos resolver antes de ir a buscar el insumo de datos que necesitamos. No obstante, es muy común que en el proceso que va hasta el desarrollo del modelo de aprendizaje automático se revelen nuevos problemas que no nos habíamos planteado originalmente. O incluso potenciales soluciones a problemas que no estaban en nuestra agenda. Tengamos la mirada sutil para lidiar con esos hallazgos.



2 →

PASO 2

Recopilar la información

El siguiente paso para ejecutar un análisis efectivo basado en datos y algoritmos consiste justamente en **recopilar la información necesaria**. En el caso de la detección de distintos tipos de eventos en imágenes satelitales, se precisan en primer lugar las imágenes satelitales sobre las cuales se desean detectar estos elementos. Para ello existe una buena cantidad de fuentes públicas y privadas de imágenes de diversa resolución, calidad y frecuencia de actualización: cada una resulta más o menos conveniente de acuerdo con el tipo de aplicación buscada. En particular, en el proyecto aquí descrito se utilizó la fuente pública de imágenes satelitales de radar provistas por el [satélite Sentinel-1 de la Agencia Espacial Europea, ESA](#), que cuenta con una resolución espacial promedio

de 20 metros al nivel del suelo. Otras fuentes públicas de datos abiertos disponibles y consideradas en este trabajo fueron también de la ESA, las imágenes ópticas del [satélite Sentinel-2](#), y también de la NASA, de los satélites [MODIS](#) y [LANDSAT](#). De fuentes públicas nacionales se puede mencionar las imágenes del [satélite SAOCOM de la CONAE](#). De fuentes de datos privadas existen las empresas [Satellogic](#), [Planet](#), [Airbus](#), [Maxar](#), [ICEYE](#) o [Capella Space](#), que en general cuentan con mayor grado de resolución y frecuencia; pueden alcanzar hasta menos de 1 metro de resolución espacial sobre el suelo. Resulta fundamental en cada caso contar con la documentación y la guía de usuario de cada una de estas fuentes, disponer de los metadatos de las imágenes recolectadas, e investigar cuáles son los distintos tipos de aplicaciones para las cuales cada una de estas fuentes resultan apropiadas y eficaces.

PASO 3

Buscar las “etiquetas”

Luego de la primera recolección, es necesario avanzar en la búsqueda de un conjunto de datos indispensables para el desarrollo de modelos automáticos de detección sobre imágenes satelitales —y para la mayoría de modelos de aprendizaje automático a partir de datos en general—: los comúnmente denominados “etiquetas”. En este caso deben corresponder a los elementos del suelo que se deseen detectar en las imágenes a través de un modelo automático. Es decir, las etiquetas están estrechamente relacionadas con la definición del problema que se quiere resolver; y, más aún, la definición del problema a resolver por parte del modelo de detección automática estará determinado por los datos de etiquetas que se puedan conseguir, recolectar o generar en cada organización. Son las etiquetas las que definen entonces lo que el modelo podrá y no podrá detectar durante su aplicación en imágenes nuevas: pueden ser obtenidas a través de bases de datos abiertas de la web, o bases mantenidas por instituciones, e inclusive pueden ser etiquetadas a mano por personas con el conocimiento experto o capacitadas específicamente para detectar visualmente los elementos deseados sobre las imágenes satelitales, usualmente denominados “anotadores”. Es de suma importancia que los anotadores cuenten con lineamientos claros y coherentes a la hora de etiquetar un conjunto de datos; también, que estos lineamientos sean debidamente actualizados ante posibles dudas que aparezcan en el proceso. La consistencia interna de las etiquetas es la clave para determinar la calidad máxima de detección de un algoritmo de aprendizaje automático.



En la colaboración entre FUNDAR y el organismo público, se utilizaron etiquetas de distintos usos del suelo provistas de manera abierta por la ESA. Estas etiquetas contienen información georeferenciada sobre píxeles construidos, distintos tipos de vegetación y masas de agua, con una resolución espacial de 300 metros sobre el suelo. La información, para nuestra experiencia, resultó efectiva como punto de partida para el desarrollo de un modelo de detección de distintos tipos de suelo, y que es pasible de ser extendido a detectar elementos de mayor resolución una vez que se cuenten con las etiquetas pertinentes. Para conseguir los datos de etiquetas es necesario entonces identificar primero el problema a resolver, y luego conseguir o generar fuentes de datos que se correspondan a la problemática propuesta.

Para conseguir los datos de etiquetas es necesario entonces identificar primero claramente el problema a resolver, y luego tratar de conseguir o generar fuentes de datos que se correspondan o acerquen de la manera más aproximada posible a la problemática propuesta.

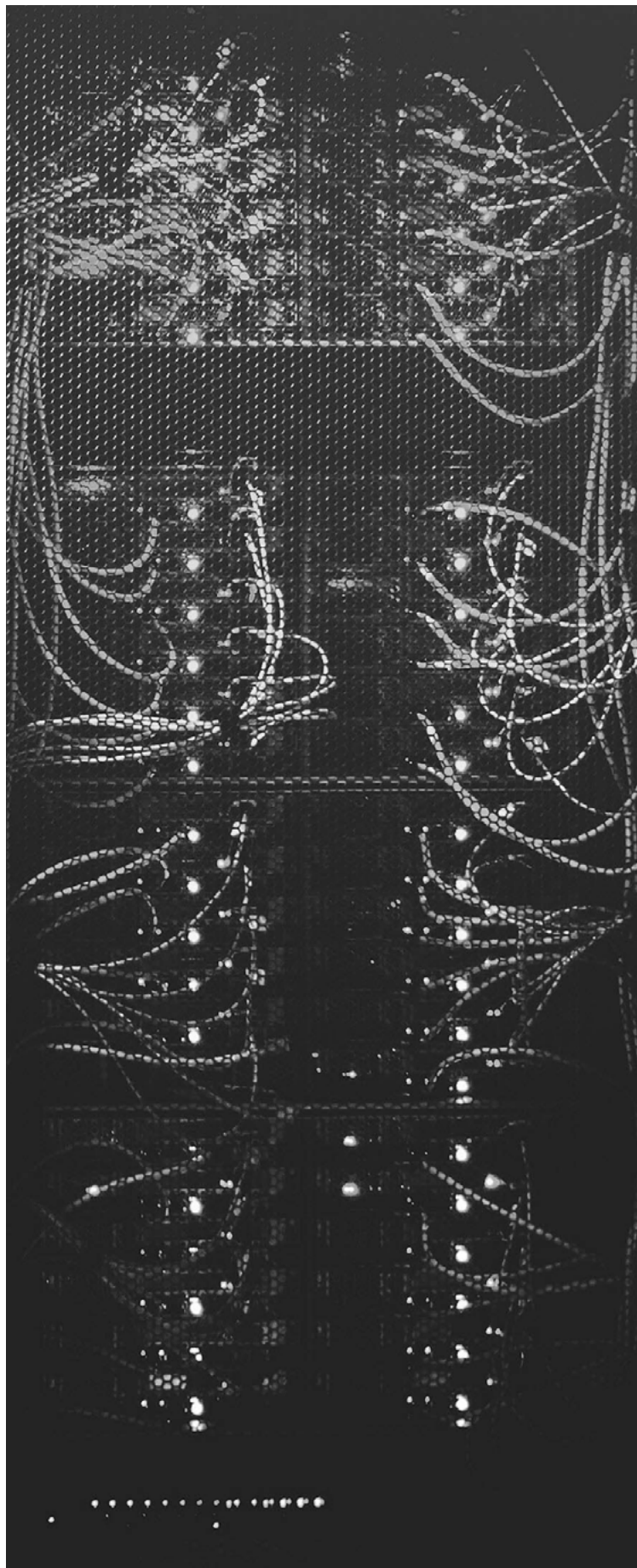
4 →

PASO 4

Asegurar la calidad de datos y el mantenimiento de bases

La recopilación de datos es uno de los primeros pasos en un proceso continuo: los datos cambian, no atender a ese cambio es reducir su valor. Por eso, es fundamental no solo el uso y la recopilación, sino también el mantenimiento actualizado de bases de datos que contengan la información correspondiente a las distintas problemáticas y eventos de interés, catalogados según ubicación geográfica, fecha y hora, y categoría del suceso, además de toda otra información relevante sobre cada fenómeno que cada organización en particular tenga previsto detectar. **La efectividad de las soluciones obtenidas mediante el desarrollo de modelos de aprendizaje automático está supeditada a la existencia, provisión, mantenimiento y calidad de estos datos.**

La efectividad de las soluciones obtenidas mediante el desarrollo de modelos de aprendizaje automático está supeditada y determinada por la existencia, provisión, mantenimiento y calidad de los datos.



Houston, we have a problem

El correcto mantenimiento de estas bases suele enfrentarse con datos faltantes o incompletos: es importante que ciertos campos de interés como la ubicación geográfica, la fecha y hora, y la categoría del suceso se encuentren presentes en todas las instancias de la base. Además, y sobre todo si el etiquetado se debe realizar por anotadores, estos deben mantener una clara consistencia interna en la forma de etiquetar las imágenes. Para esto se deben generar documentos con lineamientos que guíen claramente el etiquetado, y que contengan ejemplos de todos los elementos que el algoritmo necesite detectar, incluyendo la categoría específica de cada elemento etiquetado. Los problemas más comunes a la hora de desarrollar modelos de aprendizaje automático tienen que ver con la falta de consistencia interna entre las etiquetas, y también con la poca cantidad de etiquetas que usualmente pueden conseguirse de esta manera; las aplicaciones más comunes requieren al menos del orden de cientos o incluso miles de etiquetas para cada tipo de elemento que se desee detectar sobre las imágenes. La forma de reducir los errores de etiquetado consiste en trabajar de manera iterativa: las inconsistencias deben ser debidamente documentadas y clarificadas.

Otros problemas comunes consisten en la necesidad de integrar distintas fuentes de datos, ya que estos pueden encontrarse distribuidos en distintos servidores o directamente provenir de orígenes diversos. La consistencia interna es clave también aquí. Para ello es necesario realizar un trabajo de homogeneización, haciendo uso del conocimiento experto de cada tema en particular. De lo contrario se limitará considerablemente la calidad y el desempeño de los modelos de detección automática desarrollados a partir de estos datos.



Los problemas más comunes a la hora de desarrollar modelos de aprendizaje automático tienen que ver con la falta de consistencia interna entre las etiquetas, y también con la poca cantidad de etiquetas que usualmente pueden conseguirse de esta manera; las aplicaciones más comunes requieren al menos del orden de cientos o incluso miles de etiquetas para cada tipo de elemento que se desee detectar sobre las imágenes.

5



PASO 5

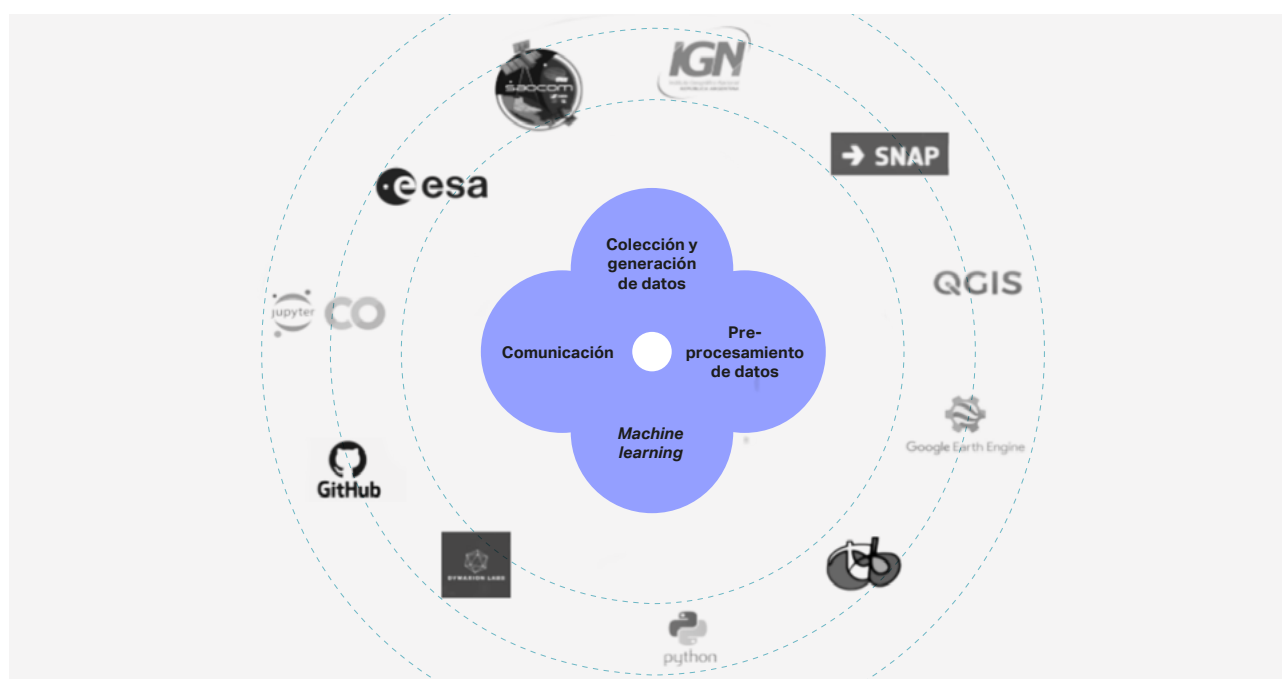
Desarrollar modelos de aprendizaje automático



Los modelos de aprendizaje automático son algoritmos que, una vez programados para tal fin, logran detectar automáticamente los tipos de etiquetas anteriormente descritas sobre las imágenes satelitales (u otros datos de interés). **En caso de éxito en su desarrollo, logran generalizar la detección efectiva de los elementos etiquetados sobre imágenes satelitales nuevas, es decir imágenes nunca antes vistas por el algoritmo durante su desarrollo.** El desarrollo de los modelos de aprendizaje automático requiere de distintas capacidades y recursos, desde los servidores o computadoras que permitan entrenar estos modelos, hasta las capacidades técnicas de los programadores o científicos de datos que los desarrollan. Además, existen desafíos específicos en cada área de aplicación de estos modelos. En este sentido, es necesario establecer una colaboración dinámica entre los científicos de datos que programan los modelos y los expertos en la materia de su aplicación. En el caso de la detección en imágenes satelitales es indispensable la colaboración con geógrafos o expertos en el conocimiento de campo asociados a la temática de trabajo, ya que el desarrollo de un modelo efectivo debe diseñarse artesanalmente considerando los distintos casos de uso y sus limitaciones. La alta variabilidad espacio-temporal de los ambientes naturales requiere del conocimiento específico de los fenómenos de los componentes del sistema terrestre en cada región de estudio. Por ejemplo, es importante tener en cuenta la dinámica de climas o los cambios en la vegetación y los suelos a lo largo de las áreas de interés específicas para cada problema.

Una vez que un modelo mínimo viable fue desarrollado, el paso final consiste en realizar una evaluación de su rendimiento en imágenes nuevas que resulten de particular interés para la problemática en cuestión. En caso de ser necesario, iterar para perfeccionarlo y resolver cada uno de los errores que puedan haberse acarreado en etapas tempranas de su desarrollo.

Herramientas para el desarrollo de modelos de aprendizaje automático



Existen una gran cantidad de herramientas de acceso libre y código abierto para el diseño y la implementación de modelos de aprendizaje automático. A continuación se muestra un diagrama general de las herramientas utilizadas en el desarrollo del modelo de prueba de concepto.

Para el trabajo en colaboración entre Fundar y un organismo público, se utilizaron los programas de acceso libre [QGIS](#) y [SNAP](#), dos sistemas de información geográfica que permiten realizar un preprocesamiento exhaustivo de las imágenes satelitales y homogeneizar las etiquetas de la ESA. Para el desarrollo del modelo de detección automático se utilizó el software [Orfeo Toolbox \(OT\)](#), una herramienta también provista por la ESA y que simplifica enormemente el desarrollo de modelos de aprendizaje

automático de clasificación de píxeles en imágenes satelitales. OT utiliza una serie de funciones pre-programadas, pudiendo procesar imágenes ópticas, multispectrales y de radar de alta resolución a una escala de tamaño de terabytes. El resto de las tecnologías, como el lenguaje de programación Python y la herramienta Github para el alojamiento del código y el control de versiones, son ubicuas en los proyectos de desarrollo de software y de modelos de aprendizaje automático en general, por lo que resultan útiles e incluso necesarias más allá de las aplicaciones concretas de la detección de elementos sobre imágenes satelitales.

Enfoque ético de la recopilación y la utilización de datos

En cuanto a la visión ética de la recopilación y utilización de datos satelitales, existen [riesgos potenciales](#) en el uso de estas tecnologías, en su gran mayoría asociados a cuestiones de privacidad, seguridad y vigilancia, ya que existe la capacidad de observar directa o inadvertidamente propiedad privada o capturar información personal confidencial. La tendencia actual en el uso de imágenes satelitales implica un gran incremento en la disponibilidad de estos datos con alta resolución tanto temporal como espacial (y que pueden llegar hasta menos de 1 metro al nivel del suelo). Esta tendencia, en conjunto con los enormes avances recientes en el desarrollo de modelos de aprendizaje automático (reconocimiento facial por ejemplo), permiten realizar análisis que conllevan grandes riesgos potenciales hacia la privacidad y la seguridad de individuos e instituciones.

Si bien los desafíos éticos que se deben afrontar son importantes, los datos satelitales son irremplazables a la hora de asistir a ciertos propósitos políticos, humanitarios o ecológicos, por ejemplo para coordinar respuestas ante desastres naturales, o marcar y controlar la trayectoria de distintas enfermedades, como la malaria. Los datos satelitales son un valioso activo para el bien público y permiten tomar decisiones en contextos de alta incertidumbre, por lo que es el deber de la gestión pública garantizar el bienestar individual y colectivo durante su utilización. Para ello, se requiere de un alto grado de compromiso con la generación y puesta en práctica de [normas de seguridad, privacidad, y protección de datos, además de las normativas que atañen a su acceso, control, y propiedad.](#)

Si bien los desafíos éticos que se deben afrontar son importantes, los datos satelitales son irremplazables a la hora de asistir a ciertos propósitos políticos, humanitarios o ecológicos, por ejemplo para coordinar respuestas ante desastres naturales, o marcar y controlar la trayectoria de distintas enfermedades, como la malaria. Los datos satelitales son un valioso activo para el bien público y permiten tomar decisiones en contextos de alta incertidumbre, por lo que es el deber de la gestión pública garantizar el bienestar individual y colectivo durante su utilización.

Síntesis y conclusiones principales



1 →

El uso de imágenes satelitales tiene el potencial de producir un alto valor público. Actualmente existen aplicaciones de diversa índole para la detección de elementos y sucesos sobre estas imágenes.

2 →

Existen fuentes públicas y privadas de imágenes de diversa resolución, calidad y frecuencia de actualización, y cada una resulta más o menos conveniente para el tipo de aplicación que se desee perseguir. Actualmente cientos de misiones satelitales orbitan la Tierra y llevan a bordo una amplia variedad de sensores con el fin de observar las modificaciones sobre la superficie del planeta. Para asociar cada fuente de datos a su aplicación apropiada se recomienda contar con la documentación y la guía de uso de cada fuente.

3 →

Los problemas atendidos por soluciones basadas en datos y algoritmos requieren de la disponibilidad de los datos y la comprensión del dominio de los mismos. La definición del problema a resolver por parte del modelo de detección automática estará unívocamente determinado por los datos de etiquetas que se puedan conseguir, recolectar o generar.

4 →

La calidad de los datos es la clave para el desarrollo de modelos de detección automática que resulten útiles para la toma de decisiones. Los datos faltantes, incompletos o inconsistentes son algunos de los problemas usuales que se deben evitar en el mantenimiento de bases de datos.



5 →

Para reducir los errores de consistencia de datos es necesario trabajar de manera iterativa o recursiva sobre los lineamientos que determinan su generación. Además, integrar distintas fuentes de datos requiere de un trabajo de homogeneización de esas fuentes.

6 →

El desarrollo de modelos de aprendizaje automático requiere servidores que permitan entrenar estos modelos, las capacidades técnicas de los científicos de datos que los desarrollan, y la colaboración dinámica entre éstos y los expertos en el campo de aplicación de los mismos, por ejemplo, geógrafos.

7 →

Existe una gran cantidad y variedad de herramientas de acceso libre y código abierto para el diseño y la implementación de modelos de aprendizaje automático. Para la prueba de concepto aquí descrita se utilizaron los sistemas de información geográfica QGIS, SNAP, y los lenguajes de Orfeo Toolbox y Python, entre otros programas.

8 →

En cuanto al enfoque ético de la recopilación y utilización de los datos satelitales, los enormes avances recientes en el nivel de resolución de estos datos y el desarrollo de modelos de detección automática permiten realizar análisis que conllevan grandes riesgos potenciales hacia la privacidad y la seguridad de los individuos y las instituciones. Por ello, se recomienda estudiar y aplicar los estándares y normativas de acceso, control y propiedad de los datos, para poder proteger efectivamente la privacidad y disminuir los riesgos asociados a su utilización.

Acerca del autor

Marcos Feole

Científico de datos del Área de Datos

Licenciado y magíster en Física por el Instituto Balseiro, y magíster en Estadística por la Universidad de Illinois en Urbana-Champaign. Trabajó en investigación, desarrollo de software, ciencia de datos y modelado matemático para el sector financiero.

Amalia Guaymás Canavire

Licenciada en Análisis de Sistemas de la Universidad de Salta. Especialista en Exploración de Datos y Descubrimiento de Conocimiento de la UBA. Es profesional en el área de ciencia de datos, con experiencia en el sector privado, público y en proyectos de investigación. En el momento de la elaboración de este documento, se desempeñaba como científica de datos del Área de Datos de Fundar

Dirección ejecutiva: Martín Reydó

Coordinación editorial: Gonzalo Fernández Rozas

Diseño: Jimena Zeitune

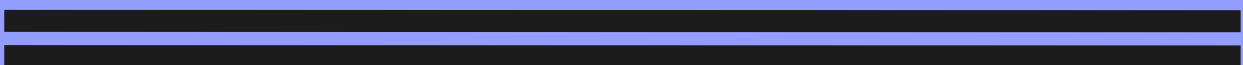
Fundar es un centro de estudios y diseño de políticas públicas que promueve una agenda de desarrollo sustentable e inclusivo para la Argentina. Para enriquecer el debate público es necesario tener un debate interno: por ello lo promovemos en el proceso de elaboración de cualquiera de nuestros documentos. Confiamos en que cada trabajo que publicamos expresa algo de lo que deseamos proyectar y construir para nuestro país. Fundar no es un logo: es una firma.

Esta obra se encuentra sujeta a una [licencia Creative Commons 4.0 Atribución-NoComercial-SinDerivadas Licencia Pública Internacional \(CC-BY-NC-ND 4.0\)](#). Queremos que nuestros trabajos lleguen a la mayor cantidad de personas en cualquier medio o formato, por eso celebramos su uso y difusión sin fines comerciales.

En Fundar creemos que el lenguaje es un territorio de disputa política y cultural. Por ello, sugerimos que se tengan en cuenta algunos recursos para evitar sesgos excluyentes en el discurso. No imponemos ningún uso en particular ni establecemos ninguna actitud normativa. Entendemos que el lenguaje inclusivo es una forma de ampliar el repertorio lingüístico, es decir una herramienta para que cada persona encuentre la forma más adecuada de expresar sus ideas.

Modo de citar

Feole, Marcos y Guaymás Canavire, Amalia (2022). Guía práctica para el uso de imágenes satelitales en la definición de políticas públicas. Disponible en <https://www.fund.ar>



fundar