Informe final - Beca ABJA

Asociación de Becarios de Japón en Argentina
ABJA

EL BIG DATA Y LOS ODS

Dr. Nazareno Castillo Marin[1]

ncmarin@untref.edu.ar

Este informe fue escrito para un público no especialista en la materia interesado en la temática del big data y su aplicación a los ODS. A lo largo del mismo, se describen de manera simple las principales metodologías de trabajo que se utilizan para generar conocimiento a partir de los datos de la telefonía móvil, las imágenes satelitales, las redes sociales y las búsquedas en internet. Para cada una de estas fuentes de datos se mencionan líneas de trabajo vinculadas al cumplimiento y monitoreo de los ODS.

Contenido

[1] Doctor en Ciencias Biológicas de la Universidad de Buenos Aires (UBA), tiene una Carrera de Especialización en “Ciencias Químicas y Ambientales” y otra en “Explotación de Datos y Descubrimiento de Conocimiento” de la misma universidad. Es autor de "17 Objetivos para un mundo mejor: una guía para entender los ODS" (Amazon, 2019). Trabaja en el MAyDS de la nación desde el 2002. Es profesor e investigador en la Universidad Nacional de Tres de Febrero (UNTREF).

Este trabajo se desarrolló con el apoyo financiero de la Asociación de Becarios de Japón en la Argentina (ABJA) que es una asociación civil sin fines de lucro que fue fundada con el objetivo difundir la Cooperación Japonesa en Argentina y crear una comunidad entre los becarios de JICA, para favorecer la transferencia tecnológica y potenciar los conocimientos adquiridos en Japón, a través del fortalecimiento de la red profesional entre sus miembros. Más de 2.700 becarios de JICA Argentina son miembros de ABJA, quienes se especializaron en becas técnicas en Japón en diversos ámbitos de aplicación (http://abja.com.ar).

Resumen ejecutivo

En 2015 las Naciones Unidas adoptaron la agenda 2030 para el desarrollo sostenible, incluyendo 17 Objetivos con 169 metas de carácter integrado e indivisible que abarcaba las esferas económica, social y ambiental. Este proceso en marcha pretende intensificar las acciones para poner fin a la pobreza, promover el crecimiento económico y abordar necesidades sociales y ambientales. Para llevar adelante estas acciones se requerirá contar con información confiable, desagregada, actualizada y a un costo accesible. Información que complemente apropiadamente a la que ya generamos a partir de los censos, encuestas y registros administrativos.

En este contexto, la utilización de nuevas fuentes de información, englobadas dentro de lo que se conoce como el "big data", puede complementar y mejorar las estimaciones de los métodos actualmente utilizados, suministrando estadísticas más rápidas y oportunas, y en muchos casos, reduciendo la carga sobre el encuestado.

El término big data, habitualmente traducido como datos masivos o grandes datos, proviene originalmente del ámbito de las ciencias de la computación, y se refiere a un conjunto de datos cuyo tamaño excede al que puede manejar el software y hardware estándar.

Para el procesamiento y análisis de los datos masivos ha surgido una disciplina denominada Ciencia de Datos. Esta combina un conjunto amplio de técnicas provenientes de las Ciencias de la Computación y Estadística, entre otras.

Las fuentes de big data principales incluyen: sensores de satélites, teléfonos móviles, datos de escáner, datos de tarjetas de crédito, posteos de internet, redes sociales y tendencias de búsquedas, entre otras.

El análisis de estos datos se convierte en un elemento esencial para la mejor comprensión de los requerimientos de las sociedades y los ciudadanos, y por lo tanto para la formulación de mejores políticas basadas en la evidencia.

Los registros de uso de telefonía móvil permiten inferir atributos socioeconómicos y demográficos de los usuarios y contribuyen a entender los patrones de movilidad de las poblaciones en el territorio. La cantidad y duración de las llamadas y la frecuencia y monto de carga de tarjetas prepagas han sido utilizadas en investigaciones para determinar niveles socioeconómicos, género y otros atributos de los usuarios de telefonía móvil. Por otro lado, los datos de localización contenidos en los registros de telefonía móvil se están utilizando para el diseño de mejores sistemas de transporte urbano, para la elaboración de mapas de riesgo a enfermedades y para analizar patrones de movilidad poblacional asociados al turismo o a la ocurrencia de catástrofes naturales como los temblores o inundaciones.

El análisis de las conversaciones en redes sociales o de las búsquedas en Internet contiene información en tiempo real acerca temas diversos, incluyendo, el costo de los alimentos, el desempleo, la inflación, y testimonios sobre catástrofes naturales. A partir del monitoreo de las redes resulta posible identificar cambios en procesos, como la inflación o el desempleo, antes de que sean registrables en las estadísticas oficiales.

Las mejoras en la accesibilidad a datos satelitales en la última década han hecho que la observación de la Tierra y la información geoespacial sean más atractivas que nunca para

abordar diversos aspectos vinculados con el desarrollo. Las imágenes satelitales están siendo utilizadas para identificar bolsones de pobreza, patrones de deforestación, disponibilidad de reservas de agua subterránea, concentración de gases efecto invernadero en la atmósfera, la dinámica de crecimiento de las ciudades e incluso para monitorear las actividad y existencias de industrias específicas. Por otro lado, la medición de la luminosidad nocturna a través de satélites se está utilizando para el monitoreo del acceso a la energía eléctrica y como un proxy de desarrollo socio-económico.

La información que surge de todas estas fuentes se puede transformar en conocimiento aplicable al monitoreo y al logro de las metas de los ODS. Para ello, es necesario contar con algoritmos computacionales que permitan identificar patrones y conocimiento a partir del análisis de millones de datos. Esos algoritmos, forman parte de una rama de la inteligencia artificial conocida como “machine learning” o “aprendizaje automático” que se basa en la idea de que los sistemas pueden aprender de datos, identificar patrones y tomar decisiones con mínima intervención humana.

La aplicación exitosa del big data requiere superar diversos desafíos. El acceso a los datos, que muchas veces está en manos del sector privado, puede no resultar simple, sobre todo cuanto estos consideran que puede haber algún rédito económico asociado a su uso. Incluso, cuando los datos pertenecen a organismos públicos pueden surgir complicaciones para el acceso, entre otras, vinculadas a la confidencialidad de los datos personales. Vale mencionar, que el manejo de los datos masivos, ha generado una razonable preocupación con respecto a su capacidad de vulnerar el derecho a la privacidad de las personas. A pesar esto, está claro que el big data vino para quedarse y puede ser de gran ayuda alcanzar las metas de los ODS.

Nuestro país, tiene la ventaja de contar con una sociedad altamente conectada, incluyendo millones de usuarios con acceso a la web y a la telefonía móvil. No obstante, las aplicaciones de big data no se darán de forma automática. Para ello se requerirá, entre otras, el establecimiento de canales apropiados para el acceso a los datos, garantizando al mismo tiempo la privacidad de las personas.

En la Argentina, la materia prima, los datos y los recursos humanos están, resta una política pública que los articule de manera inteligente para lograr las metas de desarrollo sostenible.

Hipótesis y objetivos

La hipótesis principal del presente trabajo es que el big data puede contribuir al monitoreo de los ODS en Argentina. Para abordarla se buscará dar respuesta a las siguientes preguntas:

1- ¿Cuáles son las metas en las que el big data puede hacer aportes más significativos?

2- ¿Quiénes son los proveedores principales de datos para el monitoreo de los ODS en Argentina?

3- ¿Cuáles son las técnicas/algoritmos que se están utilizando para la preparación y utilización de los datos?

4- ¿Cuáles son las ventajas y desafíos de estos nuevos datos en comparación con los obtenidos a través de las metodologías tradicionales?

Que es el big data?

El término big data (habitualmente traducido como datos masivos o grandes datos) proviene originalmente del ámbito de las ciencias de la computación, y se refiere a un conjunto de datos cuyo tamaño excede al que puede manejar el software y hardware estándar. De esta manera, hablamos de big data cuando el tamaño se vuelve parte del problema.

Para el procesamiento y análisis de los datos masivos ha surgido una disciplina denominada Ciencia de Datos. Esta combina un conjunto amplio de técnicas provenientes de las Ciencias de la Computación y la Estadística, entre otras.

El big data, además de incluir a los datos tradicionales, textos y números, aborda también a los datos no estructurados, como fotografías, audios y videos que se pueden almacenar en una diversidad de formatos.

Muchas veces para analizar un conjunto masivo de datos, lo primero que tendremos que hacer es organizarlos y ponerlos en un formato estructurado (por ejemplo, tabulados en filas y columnas).

Se puede clasificar el big data en tres grupos según el tipo de fuente que los genera.

- En primer lugar, se encuentran los datos en línea generados por humanos en forma consciente y, a menudo, voluntaria (ej. los tuits escritos por las personas)

- En segundo lugar, están los datos que son recopilados de manera invisible y pasiva, en general como subproducto de otro servicio (ej. los datos que registran automáticamente las compañías telefónicas cada vez que un móvil está activo).

- Finalmente, más allá de la actividad humana, también son fuentes generadoras de datos a gran escala distintos tipos de artefactos y sensores que capturan el comportamiento de entidades no humanas (ej. la información generada por los satélites que orbitan alrededor de nuestro planeta).

Las fuentes de big data principales incluyen: sensores de satélites, teléfonos móviles, datos de escáner, datos de tarjetas de crédito, posteos de internet, redes sociales y tendencias de búsquedas, entre otras.

De manera general las aplicaciones del big data se pueden encuadrar en tres categorías:

- Descriptivas: se utilizan los datos para estudiar y caracterizar el pasado y presente.

- Predictivas: como su nombre lo indica hacen posible la creación de modelos que permiten vaticinar lo que va a ocurrir con antelación. Buscan extraer conocimiento en forma de patrones, modelos o tendencias que ayuden predecir situaciones futuras.

- Prescriptivas: se dan en el caso de que lo que estamos caracterizando es un proceso que gobernamos por completo, y por tanto no sólo podemos observar y almacenar los datos que caracterizan la ejecución de dicho proceso, sino también modificar los valores de dichos indicadores de manera que el resultado previsto mejore respecto a una predicción inicial.

La Agenda 2030 y los ODS

Nuestro mundo globalizado se caracteriza por contar con avances extraordinarios en términos de comunicación e interdependencia a la par de niveles inaceptables e insostenibles de miseria, discriminación, injusticia y un comportamiento cuando menos irresponsable con el ambiente.

Es un mundo que transita una nueva era geológica en la historia del planeta, el Antropoceno, en la cual el hombre emerge como una nueva fuerza capaz de afectar los procesos fundamentales de la biosfera a través de la conjunción de dos fenómenos relacionados: el rápido crecimiento poblacional y el incremento del consumo de recursos per cápita.

En el plano social, el hambre, la pobreza y la desigualdad persisten en muchas regiones, adonde el progreso ha pasado por alto a una gran cantidad personas, sobre todo a aquellos que se encontraban en los escalones económicos más bajos o estaban en desventaja debido a su género, edad, discapacidad o etnia.

En este contexto de crisis ambiental y social, la conferencia Rio+20 (Conferencia de las Naciones Unidas sobre el Desarrollo Sostenible, 2012) estableció un proceso[2] para la adopción de un nuevo cuerpo de Objetivos de Desarrollo Sostenible (ODS) que permitieran dar continuidad y potenciar el alcance de los Objetivos de Desarrollo del Milenio (ODM, 2000) cuyas metas terminaban en 2015 (*Dodds et al. 2017*).

Tres años después de RIO+20 ciento noventa y tres países reunidos en la Asamblea General de Naciones Unidas adoptan los Objetivos de Desarrollo Sostenible (ODS) con la premisa fundamental de “no dejar a nadie atrás”

Los 17 Objetivos y 169 metas de la agenda 2030 intensificarán los esfuerzos para poner fin a la pobreza en todas sus formas y reducir la desigualdad, de la mano de estrategias que favorezcan el crecimiento económico y aborden una serie de necesidades sociales, a la vez que promueven la protección del medio ambiente *(Castillo Marin, 2019).*

El seguimiento y examen a nivel mundial de la Agenda 2030 y de sus 17 Objetivos de Desarrollo Sostenible (ODS) está a cargo del Foro Político de Alto Nivel de las Naciones Unidas sobre el Desarrollo Sostenible (HLPF, por sus siglas en inglés) integrado por los Estados Miembros de la ONU. El Foro, reemplaza a la Comisión de Naciones Unidas sobre el Desarrollo Sostenible y proporciona el liderazgo político, la orientación y las recomendaciones para implementación, seguimiento y monitoreo de esta agenda.

Desde su primera sesión en 2013, el HLPF convoca cada año a ministros de Estados bajo los auspicios del Consejo Económico y Social de las Naciones Unidas (ECOSOC) y reúne, cada cuatro años, a Jefes de Estado bajo los auspicios de la Asamblea General para impulsar el desarrollo sostenible. Cada año incluye una serie de eventos paralelos y de intercambios que motivan las alianzas de múltiples partes interesadas en pos de crear soluciones innovadoras. Este año el foro se hizo de manera virtual entre el 7 y el 16 de julio y los debates se organizaron en torno a las áreas de trabajo principales señaladas en el Informe Global sobre desarrollo sostenible 2020[3].

[2] El documento final de esta Conferencia “El futuro que queremos” está disponible aquí: https://rio20.un.org/sites/rio20.un.org/files/a-conf.216-l-1_spanish.pdf.pdf

[3] Disponible en: https://unstats.un.org/sdgs/report/2020/

Agenda 2030 (septiembre,2015)

En relación con la Agenda 2030 para el Desarrollo Sostenible, la Comisión de Estadística de las Naciones Unidas impulsó la creación del Grupo Interagencial y de Expertos (IAEG, por sus siglas en inglés), que tiene el mandato de definir el marco de indicadores para el desarrollo sostenible a nivel mundial y apoyar su aplicación. Cada uno de los 17 objetivos tiene una serie de metas a cumplir y sobre cada una de ellas se han definido uno o más indicadores de seguimiento. La diapositiva de abajo muestra un ejemplo de metas e indicadores para el Objetivo 1 de poner fin a la pobreza.

Anexo

Marco de indicadores mundiales para los Objetivos de Desarrollo Sostenible y metas de la Agenda 2030 para el Desarrollo Sostenible

Objetivo 1. Poner fin a la pobreza en todas sus formas y en todo el mundo

METAS	INDICADORES
1.1 De aquí a 2030, erradicar para todas las personas y en todo el mundo la pobreza extrema (actualmente se considera que sufren pobreza extrema las personas que viven con menos de 1,25 dólares de los Estados Unidos al día)	1.1.1 Proporción de la población que vive por debajo del umbral internacional de pobreza, desglosada por sexo, edad, situación laboral y ubicación geográfica (urbana o rural)
1.2 De aquí a 2030, reducir al menos a la mitad la proporción de hombres, mujeres y niños de todas las edades que viven en la pobreza en todas sus dimensiones con arreglo a las definiciones nacionales	1.4.1 Proporción de la población que vive en hogares con acceso a los servicios básicos

El Grupo Interinstitucional de Expertos se ha encargado de compilar y publicar los metadatos de los indicadores definidos por las respectivas agencias, fondos y programas con el objeto de mejorar la calidad y la cantidad de información estadística disponible y asegurar la comparabilidad internacional de los indicadores que reportan los países.

Los indicadores se clasifican en tres niveles, de acuerdo a su desarrollo metodológico y la disponibilidad de los datos:

- Nivel I: la metodología y las normas están disponibles y los datos se producen periódicamente por los países.

- Nivel II: la metodología y las normas están disponibles, pero los datos no se producen periódicamente por los países.

- Nivel III: no se dispone de metodología o normas establecidas.

La clasificación de los indicadores en los distintos niveles se va actualizando periódicamente[4]. Para diciembre de 2019, el marco mundial abarcaba 232 indicadores con la siguiente clasificación: Nivel I (116 indicadores); Nivel II (92 indicadores); Nivel III (20 indicadores); y Multinivel (4 indicadores).

Los ODS en Argentina

Argentina suscribió la Agenda 2030 en septiembre de 2015 y designó al Consejo Nacional de Coordinación de Políticas Sociales como organismo encargado de coordinar la aplicación y el seguimiento de la Agenda 2030.

En 2016, conformó la Comisión Nacional Interinstitucional de Implementación y Seguimiento de los ODS. Una de las principales funciones de esta Comisión es llevar adelante el proceso de adaptación de las metas de ODS y la selección de indicadores pertinentes y factibles para su monitoreo.

A comienzos de 2019 se publicó un informe con las fichas técnicas para el cálculo de los 242 indicadores de seguimiento (incluyendo sus líneas de base, metas intermedias y finales) que conforman en la actualidad la Agenda Nacional *(CNCPS, 2019).* En el Metadata, no se incluyen los indicadores de niveles II y III según la clasificación del Grupo Interagencial de Expertos en Indicadores de ODS. A los fines de mantener una vinculación con los indicadores del marco de monitoreo internacional, se les agregó el signo asterisco (*) para distinguir los indicadores de la Argentina relacionados pero, que no son estrictamente los mismos, que son complementarios o que implican desagregaciones de los internacionales.

[4] La revisión más reciente (2020) se puede consultar aquí: https://unstats.un.org/sdgs/iaeg-sdgs/2020-comprev/UNSC-proposal/

Por otro lado, como parte de sus mecanismos de seguimiento y revisión, la Agenda 2030 alienta a los Estados miembros a realizar revisiones periódicas del progreso a nivel nacional que contribuyan a facilitar el intercambio de experiencias entre pares, incluidos los éxitos, los desafíos y las lecciones aprendidas.

Argentina presentó un primer informe de revisión en 2017 y uno segundo *(CNCPS, 2020)* en el marco de la reunión del HLPF de julio de 2020.

Big data y ODS

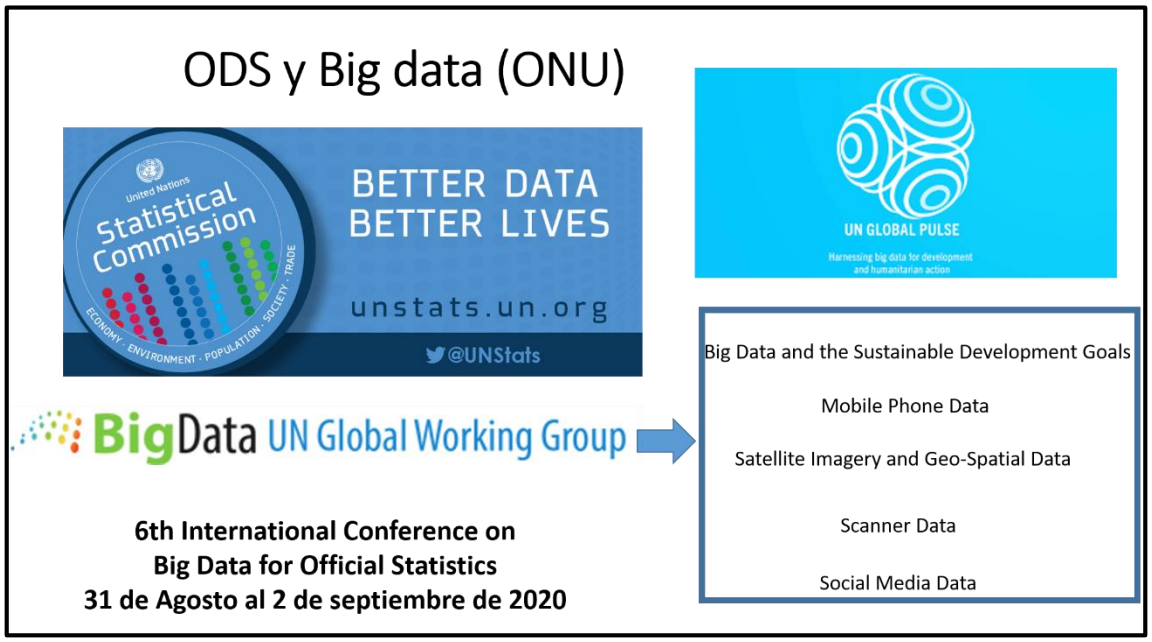
Un componente esencial para lograr el éxito de la agenda 2030 es el del monitoreo de sus metas. Para ello, los países deberán fortalecer las capacidades de sus sistemas estadísticos para dar respuesta a las nuevas necesidades de información que serán requeridas en el seguimiento de los ODS. Históricamente esta información ha dependido casi exclusivamente de la información recolectada a partir de censos, encuestas y registros administrativos. Pero esto está cambiando. Grandes volúmenes de datos (big data) están siendo generados a través de los satélites, la telefonía móvil y la internet. Esta nueva información viene a complementar la que ya generábamos a partir de las metodologías tradicionales.

En la esfera internacional, la Organización de las Naciones Unidas a través de su Comisión de Estadística ha creado el Grupo de Trabajo Mundial sobre Big Data en el año 2014. El mismo propuso una estrategia para un programa mundial de utilización de big data en estadísticas oficiales[5]. Se establecieron distintos equipos de trabajo asignados a tareas específicas que incluían: telefonía celular, imágenes satelitales y datos de las redes sociales con el fin de desarrollar actividades prácticas a través de proyectos piloto.

Otra iniciativa innovadora en el marco de la ONU es el "Global Pulse[6]" que tiene como objetivo principal acelerar el descubrimiento, desarrollo y adopción del análisis del big data como bien público. Esta iniciativa funciona a través de una red de laboratorios de innovación en Nueva

[5] https://unstats.un.org/bigdata/bureau/
[6] https://www.unglobalpulse.org/

York, Kampala y Yakarta que colaboran con las agencias de la ONU y los gobiernos y experimenta con nuevas fuentes de big data.



En la escala doméstica, agencias nacionales de estadística de distintos países han implementado divisiones dedicadas específicamente al tema. En Argentina, el Instituto Nacional de Estadística y Censos (INDEC) ha dado un primer paso a través de la firma del acuerdo sobre cooperación en temáticas de innovación estadística con el Central Bureau of Statistics (CBS) de los Países Bajos[7] en el año 2017.

De los datos al conocimiento para tomar mejores decisiones

Los datos son esenciales para tomar decisiones y la materia prima para exigir responsabilidades. Sin datos, no podemos conocer el nivel de pobreza en una sociedad o cuántas mujeres han muerto víctimas de la violencia machista. Sin embargo, una gran parte de los países que adoptaron la Agenda 2030 carecen de estadísticas fiables, incluso sobre cuestiones tan elementales como la cantidad de nacimientos y muertes. Por otra parte, muchos de los indicadores de desarrollo existentes provienen de trabajosas encuestas domésticas, que insumen un tiempo considerable, con lo que a menudo las políticas públicas se basan en datos desactualizados.

En este contexto, las nuevas fuentes de información, que se engloban bajo el concepto del big data, representan un desafío y una oportunidad extremadamente interesante. Durante los últimos 20 años hemos presenciado una verdadera revolución digital, primero gracias a la expansión de Internet, y luego a través de la masificación de los teléfonos inteligentes. Se estima que un 42% de la población mundial está conectada a Internet y que alrededor de un tercio tiene un teléfono inteligente. Y ambas cifras siguen creciendo a una tasa cercana al 10% anual.

En el sector privado, el análisis del big data es habitual desde hace tiempo, lo que permite al sector comercial crear perfiles de clientes, servicios personalizados y análisis de previsiones, que después son usados para optimizar las ventas. Técnicas similares podrían adoptarse en el sector público para conseguir en tiempo real y a un menor costo que por los métodos tradicionales, un

[7] https://www.indec.gob.ar/indec/web/Institucional-GacetillaCompleta-143

nuevo conocimiento sobre el bienestar de las personas, el estado de los recursos naturales, acciones frente a eventos de desastres naturales, epidemiologia, diseño de mejores sistemas de transporte urbano, etc.

El uso del big data puede complementar y mejorar las estimaciones de los métodos actualmente utilizados, suministrando estadísticas más rápidas y oportunas, y en muchos casos, reduciendo la carga sobre el encuestado.

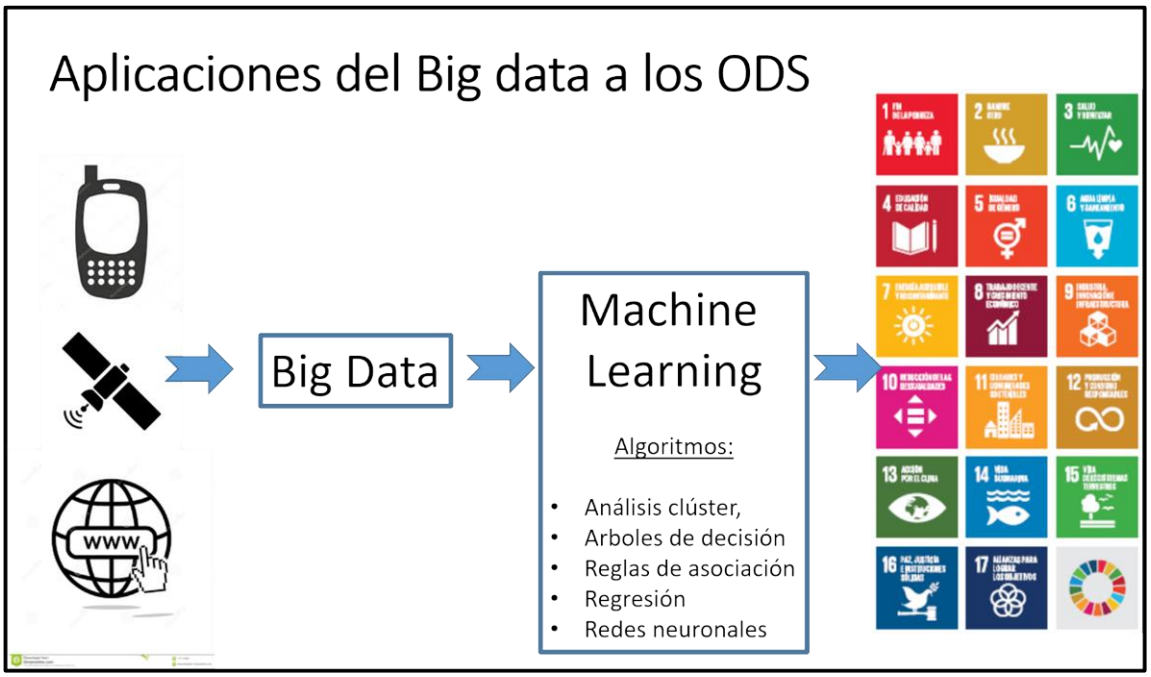
El análisis de estos datos se convierte en un elemento esencial para la mejor comprensión de los requerimientos de las sociedades y los ciudadanos, y por lo tanto para la formulación de mejores políticas basadas en la evidencia.

Los registros de uso de telefonía móvil permiten inferir atributos socioeconómicos y demográficos de los usuarios y contribuyen a entender los patrones de movilidad de las poblaciones en el territorio.

El análisis de las conversaciones en redes sociales (*Russell et al., 2019*) contiene información en tiempo real acerca temas diversos, incluyendo, el costo de los alimentos, la accesibilidad a puestos de trabajo, el acceso a la atención sanitaria, la calidad de la educación y testimonios sobre catástrofes naturales.

Las mejoras en la accesibilidad a datos satelitales en la última década han hecho que la observación de la Tierra y la información geoespacial sean más atractivas que nunca para abordar la pobreza, el desarrollo económico, la deforestación y las inundaciones, entre otras.

Los datos que surgen de estas fuentes se pueden transformar en conocimiento aplicable al monitoreo y al logro de las metas de los ODS *(LIRNEasia, 2017; MacFeely, 2019)*. Para ello, es necesario contar con algoritmos computacionales que permitan identificar patrones y conocimiento a partir del análisis de millones de datos. Esos algoritmos, de los cuales nos ocuparemos más adelante, forman parte de una rama de la inteligencia artificial conocida como "*machine learning*" o "*aprendizaje automático*". La misma se basa en la idea de que los sistemas pueden aprender de datos, identificar patrones y tomar decisiones con mínima intervención humana.



Datos tradicionales vs big data

El big data se suma a las fuentes tradicionales de datos como las encuestas y los registros administrativos. Por ejemplo, el empleo se puede medir a través de encuestas a empresas donde se les solicita informar mensualmente sobre su personal contratado. Complementariamente, en algunas otras operaciones estadísticas, se trabaja con las cuentas de la Seguridad Social de las empresas, que provienen del correspondiente registro administrativo generado y gestionado por organismos públicos (ej. ANSES o Ministerio de Trabajo). Por último, es posible rastrear y descargar datos de empleo de portales específicos y webs de empresas en Internet.

Muchas veces las aplicaciones de big data se basan en "proxies" que son aproximaciones o indicadores cuantitativos de datos reales. Por ejemplo, las búsquedas en Google en base a términos como "bolsa de trabajo" algunas veces funciona como un proxi de la tasa de desempleo en un lugar. De manera análoga, la actividad que generan los teléfonos móviles en nuestras antenas es un proxy de la movilidad de las personas en una ciudad.

En este contexto, es importante remarcar que la utilización del big data en estadísticas oficiales, en comparación con los datos obtenidos por métodos tradicionales (*Sosa, 2019; Gonzalez, 2018; Data-Pop Alliance, 2016)* presenta una serie de oportunidades y también de retos que se describen en una sección posterior.

Etapas de un proyecto big data

Un proyecto big data involucra al menos tres etapas: (1) adquisición y almacenamiento de los datos, (2) preparación de los datos y, (3) análisis y descubrimiento de conocimiento.

Adquisición y almacenamiento de datos

Lo primero que hay que considerar es que no siempre resulta simple acceder a los datos. A veces porque están en manos del sector privado y no los quiere compartir porque los considera información estratégica para su negocio o porque tiene la obligación de proteger la privacidad de los datos de sus clientes. La necesidad de proteger la privacidad de las personas también puede ser un impedimento para acceder a los datos en el sector público.

Adicionalmente, como fue mencionado previamente, los grandes volúmenes de datos que se manejan en proyectos de big data muchas veces requieren el uso de software[8] y hardware especial.

Preparación de los datos

Los datos pueden venir en formatos heterogéneos, tanto de forma estructurada como desestructurada, por lo que será necesario compatibilizarlos a un formato común que sea apropiado al tipo de análisis a desarrollar. La preparación también incluye la limpieza y depuración de los datos por ejemplo caracterizando y/o eliminando los "outliers" y los vacíos de información.

Analítica

[8] En lo referente al software requerido para administrar los recursos de una plataforma de Big Data, hoy en día el principal entorno de trabajo utilizado es "Hadoop" cuyo desarrollo pertenece a: "The Apache Software Foundation".

La analítica de datos se propone dar respuesta a preguntas y/o hipótesis de trabajo. Para ello se utilizan algoritmos y técnicas que se encuadran dentro de lo que conocemos como "machine learning" o "aprendizaje automático".

El objetivo del aprendizaje automático es crear un modelo que nos permita resolver una tarea dada. Para ello, primero se entrena un modelo usando una gran cantidad de datos (ejemplos). El modelo aprende a partir de ellos y después es capaz de hacer predicciones. Este aprendizaje puede ser "supervisado", cuando conocemos cuales son los resultados deseados a partir de los datos de entrenamiento o "no supervisado", cuando los desconocemos[9].

En una tarea supervisada, se busca capturar la relación entre algún ejemplo (entrada) y alguna variable objetivo (salida). Para ello, resulta necesario contar tanto con los ejemplos de entrada, así como con los resultados esperados (salidas) de modo que el modelo "aprenda" la relación entre ambos. Las entradas pueden ser, por ejemplo, datos sobre edad, nivel educativo, género o ingreso; y la salida, el partido político que apoya el individuo.

En una tarea no supervisada, en cambio, la variable objetivo es desconocida, por lo que no es posible ajustar el modelo a salidas conocidas, dadas las características de las entradas. El análisis de componentes principales y el clustering son ejemplos de aprendizaje no supervisado.

Clustering

Se trata de una técnica estadística multivariada para el agrupamiento automático de objetos de forma que objetos similares se pongan en el mismo grupo o cluster, mientras que objetos disimilares terminan en clusters tan diferentes como sea posible.

La mayoría de los algoritmos de clustering computan la proximidad o similaridad entre cada par de objetos, a partir de sus atributos. La proximidad o similaridad de dos objetos se define por una fórmula objetivo que considera las propiedades conocidas de cada uno de ellos. Por ejemplo, si los objetos son documentos, la similaridad se puede medir al considerar la cantidad de palabras que cada par de objetos tiene en común. En caso de que haya muchas palabras compartidas, es plausible suponer que ambos documentos discuten la misma temática y entonces son categorizados en el grupo que responde a esa temática.

Resumiendo, dado un conjunto de atributos conocidos (dimensiones), la meta de un algoritmo de clustering es dividir automáticamente los objetos en grupos, sobre la base de la proximidad o similaridad entre los objetos, de modo que los grupos resultantes sean tan homogéneos como sea posible. Los objetos con atributos similares deberían situarse en el mismo grupo y la disimilaridad entre grupos debería ser tan alta como sea posible.

Se puede utilizar el clustering para la exploración de datos o bien como instancia de su pre-procesamiento, previa a la aplicación de otros algoritmos.

Un ejemplo de algoritmo clustering es el algoritmo de K-medias que intenta encontrar una partición de las muestras en K agrupaciones, de forma que cada ejemplo pertenezca a una de ellas, concretamente a aquella cuyo centroide esté más cerca. El mejor valor de K para que la clasificación separe lo mejor posible los ejemplos no se conoce a priori, y depende completamente de los datos con los que trabajemos. Aunque se puede parecer al

[9] Existe una tercera variante, el "aprendizaje por refuerzo". De manera general, en este caso se le dan a la maquina una serie de reglas a cumplir y el algoritmo aprende a buscar el resultado óptimo en función de esas reglas.

funcionamiento del algoritmo k Nearest Neighbour que veremos más adelante, aquí se ve claramente la diferencia con un algoritmo supervisado: en este caso, no tenemos un conocimiento a priori que nos indique cómo deben agruparse ninguno de los datos de los que disponemos, es decir, no hay un protocolo externo que nos indique lo bien o mal que vamos a realizar la tarea.

Ejemplos de aprendizaje supervisado

Dentro del aprendizaje supervisado, dependiendo del tipo de variable objetivo, existen dos clases principales: regresión y clasificación. En los casos de clasificación, la variable objetivo es de tipo categórico (discreta), mientras que, en los casos de regresión, es de tipo numérico (continua).

Algunos de los algoritmos habitualmente se aplican para el aprendizaje supervisado son:

- ***Árboles de decisión:*** es una estructura similar a un diagrama de flujo que utiliza un método de bifurcación para ilustrar cada resultado posible de una decisión. Cada nodo dentro del árbol representa una prueba en una variable específica, y cada rama es el resultado de esa prueba. Los algoritmos de árbol de decisión desglosan el conjunto de datos mediante la formulación de preguntas hasta conseguir el fragmento de datos adecuado para hacer una predicción. Basado en las características de los datos de entrenamiento, el árbol de decisión "aprende" una serie de factores para inferir las etiquetas de clase de los ejemplos. El nodo de comienzo es la raíz del árbol, y el algoritmo dividirá de forma iterativa el conjunto de datos en la característica que contenga la máxima ganancia de información, hasta que los nodos finales (hojas) sean puros.

- ***k Nearest Neighbour:*** el algoritmo de los k vecinos más cercanos (k-NN, o k Nearest Neighbour) es un algoritmo de clasificación supervisado basado en criterios de vecindad. En particular, k-NN se basa en la idea de que los nuevos ejemplos serán clasificados con la misma clase que tengan la mayor cantidad de vecinos más parecidos a ellos del conjunto de entrenamiento.

- ***Clasificación de Naïve Bayes:*** este tipo de algoritmos por clasificación están basados en el teorema de Bayes y permiten predecir una clase o categoría en función de un conjunto dado de características, utilizando la probabilidad.

- ***Regresión:*** en las tareas de regresión, el programa de aprendizaje automático debe estimar y comprender las relaciones entre las variables. El análisis de regresión se enfoca en una variable dependiente y una serie de otras variables cambiantes, lo que lo hace particularmente útil para la predicción y el pronóstico. Los valores de salida son continuos.

- ***Support Vector Machines (SVM).*** es un algoritmo de aprendizaje de máquina que puede ser usado para clasificación y problemas de regresión. En SVM se traza cada observación como un punto en un espacio dimensional donde cada dimensión corresponde a una característica. El valor de cada característica es el valor de coordenadas particulares. Luego se intenta encontrar un hiper-plano que separe las clases.

- ***Redes neuronales[10]:*** Las redes neuronales tratan de imitar la conexión de las neuronas en el cerebro a partir de nodos que en función de los valores de entrada que reciben producen salidas específicas. Hay una gran variedad de algoritmos de redes neuronales tanto para aprendizaje supervisado como para no-supervisado.

- ***Métodos "Ensemble" (Conjuntos de clasificadores):*** los métodos combinados (métodos de ensemble) utilizan múltiples algoritmos de aprendizaje para obtener un rendimiento predictivo que mejore el que podría obtenerse por medio de cualquiera de los algoritmos de aprendizaje individuales que lo constituyen.

Big data y el problema de sobreajuste de los modelos

Se denomina sobreajuste al hecho de hacer un modelo tan ajustado a los datos de entrenamiento que haga que no generalice bien a los datos de test.

Hay que recordar que el objetivo de los modelos de aprendizaje automático es el de obtener patrones de los datos de entrenamiento disponibles, de cara a predecir o inferir correctamente datos nuevos. Es decir, el concepto clave es el de entrenar y obtener patrones generales que sean extrapolables a nuevos datos. Algo similar ocurre en el aprendizaje de los seres humanos, el sobreajuste se ocurriría cuando aprendemos las cosas de memoria, sin entender el concepto.

El sobreajuste se puede evitar de varias formas, las más claras son las siguientes:

- Incorporando mayor cantidad de datos: al tener más cantidad de datos es más probable que el algoritmo generalice mejor[11].
- Cambiando los parámetros de ciertos algoritmos, haciendo los algoritmos más simples. Por ejemplo, reduciendo la profundidad de un árbol de decisión se ajusta menos al hacer el modelo más simple.

Telefonía móvil

Los datos que surgen de la telefonía móvil pueden ser usados para el monitoreo de diversos indicadores de los ODS. Por otro lado, el teléfono móvil también se está transformando en una herramienta clave para el desarrollo a través de brindar oportunidades de inclusión financiera a los más pobres[12], permitiéndoles cobrar el sueldo, pagar cuentas, enviar dinero, etc.

Los patrones de uso del teléfono varían según los atributos demográficos y socioeconómicos de los usuarios. La cantidad y duración de las llamadas y la frecuencia y monto de carga de tarjetas prepagas han sido utilizadas en investigaciones para determinar niveles socioeconómicos, género y otros atributos de los usuarios de telefonía móvil. Los datos de los celulares también

[10] En la actualidad se está usando cada vez más los algoritmos de aprendizaje profundo (Deep learning) que integran capas de algoritmos de redes neuronales, las cuales pasan una representación simplificada de los datos de una a otra capa. El aprendizaje profundo se puede usar en ambos acercamientos, supervisado y no supervisado.

[11] Una técnica para evitar el sobreajuste es la "validación cruzada" que se basa en dividir los datos de entrenamiento en partes, entrenando al modelo en una subserie de los datos (75%) y luego probándolo en los datos restantes (25%). Esto permite que los valores diagnosticados de la serie de prueba provean una muestra más precisa de cómo el modelo rendiría en un escenario fuera de la muestra.

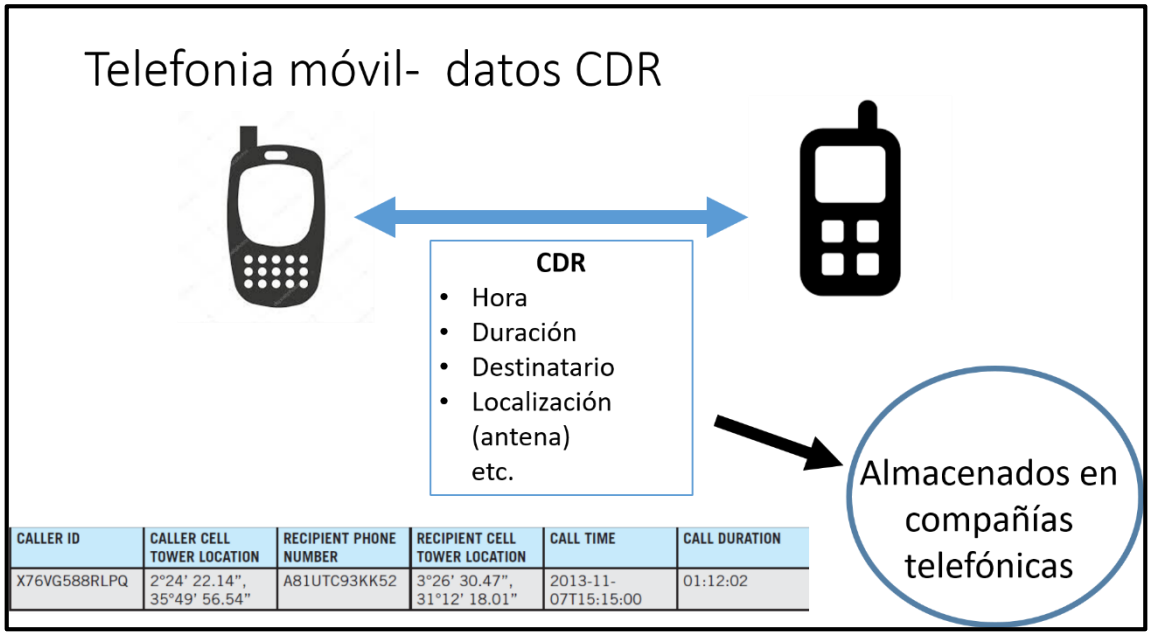
[12] Dos tercios de los 1.700 millones de adultos no bancarizados poseen un celular y 480 millones tienen acceso a internet (fuente: Global Findex).

incluyen la localización espacial de los usuarios a lo largo del tiempo, lo cual está siendo utilizado para el diseño de mejores sistemas de transporte urbano y para entender patrones de movilidad en poblaciones sujetas a epidemias o eventos de desastres naturales.

Los datos de telefonía móvil incluyen los CDR, los GPS, la actividad en internet y los datos de crédito de tarjetas prepagas.

La mayor cantidad de trabajos de investigación en base a datos de telefonía móvil se han hecho utilizando los registros CDR.

Los teléfonos móviles generan datos en forma de registros (que en inglés se denominan CDR[13]) y que incluyen, entre otros, el originante y destinatario de la llamada (o mensaje), su duración, el horario en que ocurre la comunicación y la localización de la antena de telefonía móvil que se activa. Las operadoras de redes móviles registran y almacenan los CDR de sus clientes, primariamente para propósitos de cobro.



Un proyecto de big data que se propone estudiar la pobreza a partir de datos de CDR incluye las siguientes etapas:

1- ***Obtención de los datos:*** la privacidad de datos es un problema sensible y usualmente controvertido, por lo que las operadoras telefónicas suelen tomar ciertos recaudos antes de proveer datos de CDR. Los identificadores deben ser oscurecidos antes que los registros sean exportados desde los sistemas de las telefónicas, para que los números de teléfonos celulares o campos similares no permanezcan en los datos salientes. Este proceso es referido como "anonimización".

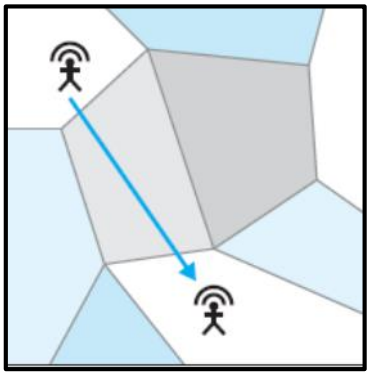
2- ***Pre-procesamiento:*** los datos crudos de CDR son invariablemente ruidosos y requieren de un pre- procesamiento antes de poder ser analizados. Las fuentes primarias de ruido en datos son: (1) huecos o inconsistencias causadas por decisiones de operación de las compañías telefónicas, incluyendo cambios tecnológicos que afectan la comparabilidad de los CDR; y (2) la presencia de líneas de negocios, números para recargar textos u otras

[13] Call Detail Registry

conexiones que no reflejan comunicaciones entre clientes individuales. El preprocesamiento comienza al identificar los CDR inconsistentes o irrelevantes y eliminarlos de la serie de datos.

3- ***Identificación de hogares:*** cada cliente en la serie de datos de CDR asignado a una ubicación de hogar, la cual es generalmente inferida a partir de la antena en la cual ese cliente se encuentra más activo después de las 6pm.

4- ***Agregado espacial de individuos:*** a través del análisis de los CDR no es posible identificar la ubicación precisa del usuario, pero si tener una aproximación del área geográfica o polígono de Voronoi[14] donde se encuentra a través de determinar las antenas de celular con las que se conecta[15]. Los CDR correspondientes a cada polígono con combinados para formar agregados estadísticos (ej. media, mediana, etc.). Son estos agregados geográficos por polígono los que se comparan con los datos de encuesta y censos.



5- ***Armonización espacial:*** asegurar que todas las series de datos compartan una escala espacial común es un paso importante en la preparación de datos. Mientras que la red celular (polígonos de Voronoi) es la unidad espacial natural para los datos de CDR, los índices de pobreza son generalmente calculados usando límites administrativos, como los de los municipios. Estas estructuras espaciales deben ser reconciliadas para poder analizar los datos. Una alternativa consiste en asignar a cada polígono de Voronoi un promedio de área pesada de los índices de pobreza en los municipios que cubre.

6- ***Análisis de los datos***: se aplican diferentes algoritmos de aprendizaje automático para extraer conocimiento de los datos. Por ejemplo, se puede entrenar un modelo usando como datos de entrada los CDR y como salida índices de pobreza obtenidos a partir de encuestas. De esta manera el modelo aprende a asociar determinados atributos contenidos en los CDR (ej: duración de la llamada, cantidad de llamadas, etc.) con un mayor o nivel de pobreza. Una vez que el modelo aprendió puede generar estimaciones de pobreza frente a nuevos CDR que no vio en el entrenamiento.

Existen al menos tres diferentes maneras en las cuales los datos de CDR pueden complementar a las estadísticas de pobreza elaboradas en base a encuestas:

[14] Los polígonos de Voronoi también conocidos como polígonos de Thiessen son una construcción geométrica que permite construir una partición del plano euclídeo

[15] Esto suponiendo que el móvil siempre se conecta a la antena celular más cercana.

- Relleno espacial: generando estadísticas de áreas pequeñas. Dada la relativamente alta resolución espacial de CDR, el relleno espacial posiblemente ofrece el mayor valor agregado como un método de investigación socioeconómica. Una encuesta de hogar particular con un limitado tamaño de muestra solamente puede apoyar los estimados en una resolución espacial relativamente áspera, tal como al nivel del departamento. Las señales de comportamiento de alta resolución en los CDR pueden mejorar la fortaleza estadística de los datos de encuestas, permitiendo estimados precisos y más detallados. Idealmente, el periodo de recolección de CDR debe coincidir con el periodo en el que la encuesta fuese conducida.

- Interpolación/extrapolación de tiempo. La alta frecuencia potencial de los CDR puede también complementar las fuentes convencionales de datos. Las encuestas de hogares son actualizadas generalmente en intervalos largos. Los CDR pueden ser usados para actualizar estos estimados, proveyendo a los oficiales con información actual sobre temas específicos de políticas. Un modelo predictivo puede ser construido para un año de encuesta usando datos CDR contemporáneos. Este modelo puede entonces ser aplicados a datos CDR más recientes para los cuales los datos de encuestas correspondientes no están disponibles. Existe un riesgo, sin embargo, de que la relación entre señales de comportamiento CDR y el objetivo variable, en este caso el índice de pobreza, pudiera cambiar con el tiempo.

- Extrapolación espacial. Esta es la más ambiciosa aplicación potencial de los datos de CDR. En países o regiones en los cuales no hay datos recientes de encuestas disponibles, tal como áreas afectadas por conflictos o países que han experimentado inestabilidad política severa, los estimados pueden ser generados al usar datos de encuestas y CDR para una ubicación similar y luego aplicar este modelo a los datos CDR de la ubicación objetivo. Este enfoque requiere asunciones significativas, pero podría ser útil en casos donde no existe una fuente fuerte de datos.

Determinación de atributos socioeconómicos y demográficos (ODS 1)

El análisis de los CDR de teléfonos móviles se está utilizando para estimar atributos socio-económicos de la población.

- *Thomas et al. (2009)* encontraron que los usuarios de niveles socioeconómicos más bajos en Alemania utilizan el teléfono durante un periodo de tiempo más prolongado a lo largo del día.

- *Blumenstock et al. (2010*) descubrieron que los hombres, en Rwanda, hacen más llamadas y reciben menos llamadas que las mujeres. Además, los usuarios que poseen Tv y heladera (indicadores de un status socioeconómico superior) tienen llamadas de duración más larga en comparación con otros.

- *Frias-Martinez et al. (2012)* concluyeron que poblaciones con altos niveles socioeconómicos tienen rangos de movilidad territorial más grandes que los de poblaciones de niveles socioeconomicos bajos.

- *Gutierrez et al. (2013)* en base a datos de carga de tarjetas prepagas y de actividad del móvil de usuarios infirieron su nivel socioeconómico. Los que hacen menos recargas pero de mayor monto en general tienen un nivel socioeconómico más alto que los que hacen muchas recargas de bajo monto.

- *Blumenstock et al. (2015)* desarrollaron un índice compuesto de bienestar a partir de resultados de una encuesta telefónica (se les preguntaba por características de su vivienda, posesiones, etc.) a individuos elegidos al azar combinados con la actividad registrada en sus teléfonos móviles (bajo su previo consentimiento). A partir de estos datos entrenaron un modelo para predecir el nivel económico de clientes de telefonía móvil a partir de la actividad de sus teléfonos.

- *Steele et al. (2017)* usaron datos de CDR (métricas de movilidad, patrones de recarga y de uso del teléfono) combinados con datos de imágenes satelitales (como acceso a caminos y luminosidad nocturna) para construir un modelo que predecía aceptablemente índices de pobreza.

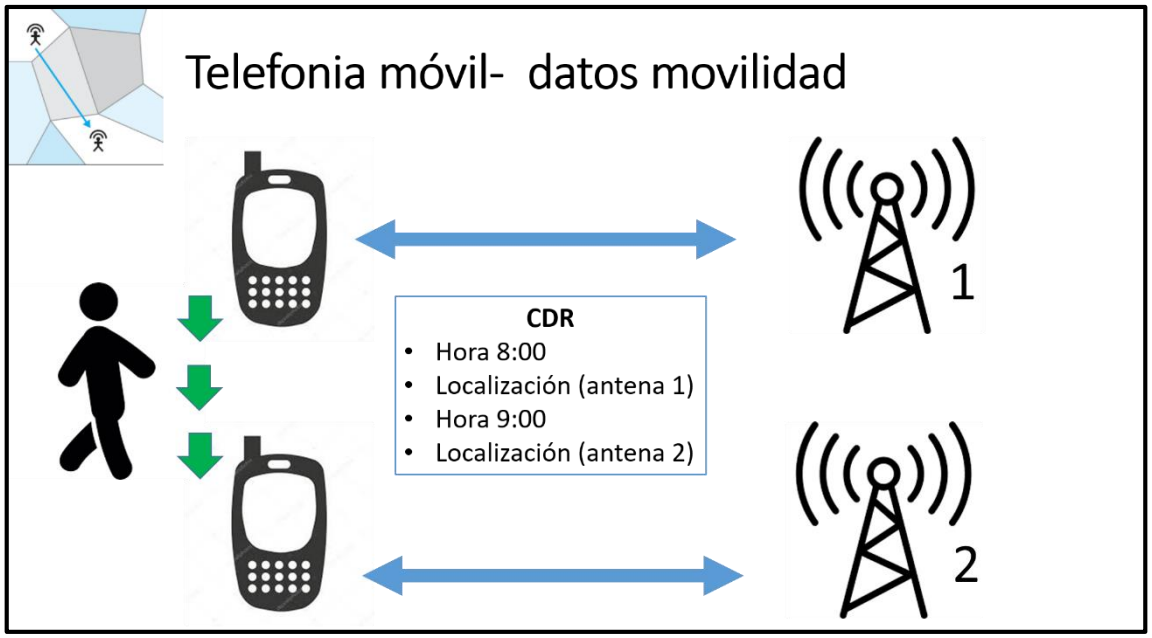
- *Decuyper et al. (2014)* compararon el uso del teléfono movil (actividad y compras de tarjeta prepaga) de un país del este de Africa con una encuesta de nacional conducida por el Programa Alimentario Mundial. A partir de ello encontraron una buena correlación entre las compras de tarjetas prepagas y los resultados de la encuesta para diversos productos alimenticios, indicando que este dato puede servir como un proxy de gasto en alimento de los hogares.

Datos de movilidad

Como vimos previamente, la actividad de cada teléfono móvil (llamadas, mensajes) es captada por la antena[16] más próxima al dispositivo dejando asentada la localización del usuario en cada registro CDR[17]. De esta manera analizando los registros se puede determinar la localización (en términos de polígonos de Voronoi) en el espacio y tiempo de los usuarios, diferenciando entre los periodos en que permanece en un mismo lugar (que registran la misma antena) y los traslados (que registran cambios en las antenas que se activan).

[16] Las compañías telefónicas mantienen una lista de las coordenadas de cada antena, haciendo posible determinar la ubicación general de un teléfono cada vez que es usado.

[17] Si no se usa el teléfono, no se registra CDR y consecuentemente no puede registrar el movimiento



Estos datos también se pueden obtener a partir de teléfono inteligentes con GPS incorporado. De hecho, tienen la ventaja de poder registrar los movimientos aun cuando el teléfono no esté en uso. Además el margen de error en la localización de los GPS es de 10 a 30 m, un orden de magnitud menor que al que se logra con los CDR. Una desventaja es que la penetración del teléfono inteligente con GPS en la sociedad es menor que el convencional, sobre todo en las regiones de menores recursos que es en donde usualmente se requiere mayor fortalecimiento de las estadísticas.

Como veremos a continuación los datos de localización contenidos en los CDRs se están utilizando para el diseño de mejores sistemas de transporte urbano, para la elaboración de mapas de riesgo a enfermedades y para analizar patrones de movilidad poblacional asociados al turismo o a la ocurrencia de catástrofes naturales como los temblores o inundaciones.

Planificación del transporte del transporte urbano (ODS 11)

Hasta ahora, los estudios de movilidad urbana y la planificación del transporte se han basado sobre todo en las encuestas domiciliarias de movilidad, que ofrecen información muy rica pero restringida a un momento en el tiempo, por lo que suele quedar rápidamente desactualizada. En cambio los datos de localización de los CDR se están generando automáticamente de manera continua, por lo que su uso puede contribuir a mantener actualizada la información de las encuestas.

El análisis de los registros de CDR de usuarios de telefonía movil permite conocer su localización en el espacio y en el tiempo. Consecuentemente, a partir de este tipo de datos es posible conocer los lugares en los que el usuario realiza sus actividades (domicilio, trabajo, otras actividades), analizar sus patrones de movilidad e incluso se puede inferir qué tipo de transporte (ej. carretera, tren, etc.) están utilizando en sus desplazamientos.

11.2 De aquí a 2030, proporcionar acceso a sistemas de transporte seguros, asequibles, accesibles y sostenibles…

11 CIUDADES Y COMUNIDADES SOSTENIBLES

- Hasta ahora, los estudios de movilidad urbana y la planificación del transporte se han basado sobre todo en las encuestas domiciliarias de movilidad, pero esa información se refiere a un momento en el tiempo y queda pronto desactualizada
- El análisis de los registros de actividad de usuarios de teléfonos móviles permite conocer su localización en el espacio y en el tiempo

- *Calabrese et al. (2011); Toole at al. (2014); Samarajiva et al. (2015)* utilizaron datos de localización de usuarios de telefonía movil contenidos en los CDRs para identificar patrones de movilidad (flujos de origen-destino) en ciudades, incluyendo la identificación de las horas pico, congestiones de tránsito y los lugares donde se producen aglomeraciones de personas.

- El *BID (2020)* construyó un modelo de transporte de San Salvador empleando utilizando la información de CDRs y registros de localización de usuarios de teléfonos con sistema operativo Android y de aplicaciones de Google Maps, entre otros tipos de tecnología. Con el Modelo de Transporte es posible simular el impacto de diferentes proyectos de infraestructura como creación de nuevas vías, ampliación de carriles, cambios de sentido vial, así como de modificaciones en la oferta de transporte público, un reordenamiento de las rutas de transporte público, etc.

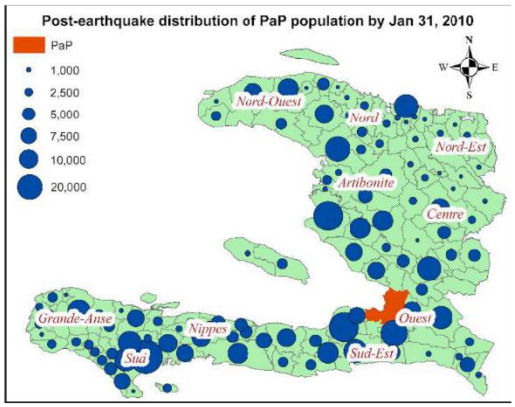
Movilidad en poblaciones afectadas por desastres naturales (ODS 11)

Para poder proveer asistencia humanitaria y restaurar servicios básicos en las áreas afectadas por el desastre, los gobiernos y las organizaciones requieren de datos actualizados que puedan ser recolectados rápidamente y a un costo modesto.

En este contexto, los datos de telefonía móvil pueden contribuir a entender patrones de movilidad en poblaciones afectadas por desastres naturales como los huracanes y los temblores Este fue el caso de Haití después del terremoto de 2010, el cual inspiró a varios investigadores a examinar el potencial de datos de CDR para producir rápidamente información de rastreo de alta frecuencia sobre el desplazamiento de la población a corto plazo. Por ejemplo, *Lu et al., (2012)* analizando millones de CDR identificaron los destinos de las personas que dejaron la capital de Haití después del terremoto. En el caso de un nuevo temblor, esta información puede ser de utilidad para diseñar las estrategias de ayuda post-evento orientadas a los lugares donde se espera recibir personas afectadas.

11.5.1 Número de personas muertas, desaparecidas y afectadas directamente atribuido a desastres por cada 100.000 personas

Lu et al., 2012 – analizando millones de CDR identificaron los destinos de las personas que dejaron la capital de Haití despúes del terremoto del 2010. En el caso de un nuevo temblor, esta información pueder ser de utilidad para diseñar las estrategias de ayuda post-evento.



Análisis de los destinos y patrones de movilidad de los turistas (ODS 8)

Las estadísticas tradicionales sobre turismo incluyen la recogida de datos en los pasos de aduana, el transporte, las pernoctaciones, cuestionarios y diferentes conjuntos de datos modelizados. Sin embargo, las estadísticas existentes son limitadas en el espacio y en el tiempo y no permiten el análisis de cuestiones más complejas como la elección de destinos, la valoración de lugares de interés o los puntos de atracción turística visitados. El análisis de los registros de CDR puede contribuir a generar conocimiento en estos temas. . Además, hay que resaltar la mayor precisión espacio - temporal de los datos de los teléfonos móviles, respecto de lo que ofrecen hoy las estadísticas oficiales existentes.

En el municipio de Girona en España la Unidad da datos (LUCA) de Telefónica analizo el movimiento de los visitantes a lo largo de todo el municipio a través del seguimiento de las señales de celulares. A partir del análisis de los CDR pudieron además identificar el número de turistas y su procedencia, el tiempo que permanecen en el municipio y generar mapas de calor identificando las áreas de mayor interés para los turistas.

Movilidad de las personas para estudios epidemiológicos (ODS 3)

Los datos de la telefonía móvil se están utilizando para monitorear patrones de movilidad de las personas en el estudio de la propagación de enfermedades transmitidas por mosquitos como el dengue y la malaria (*Tatem et al. 2009; Ruktanonchai et al. 2016*). La teoría subyacente es que la interacción social asociada a la movilidad de las personas facilita la propagación de las enfermedades.

Wesolowski, et al. (2015) desarrollaron un modelo epidemiológico para la transmisión del dengue a partir de información climática y datos de telefonía móvil de 40 millones de usuarios para predecir la propagación del dengue en Pakistán.

Amy et al. (2012) a partir del rastreo de la información detallada sobre la ubicación de cada uno de los 15 millones de propietarios de teléfonos celulares en Kenia, en conjunto con mapas de la incidencia de la malaria por región, crearon un mapa del movimiento del parásito causante de esta enfermedad.

Para estimar el potencial de propagación de la malaria es importante conocer no solo la ubicación de los mosquitos que transmiten el parásito sino también la movilidad de las personas que estando infectadas no presentan síntomas. Vale aclarar que si un mosquito libre de parásito pica a una persona enferma, el insecto se infecta y puede continuar con el ciclo de transmisión de la enfermedad.

3.3 De aquí a 2030, poner fin a las epidemias del SIDA, la tuberculosis, la malaria y …

Mapeo de malaria en Kenia (Global Pulse,2013): analizaron los patrones de movilidad regional de millones de usuarios de telefonía móvil para mapear las localidades donde existen mayores probabilidades de dispersión de la malaria.

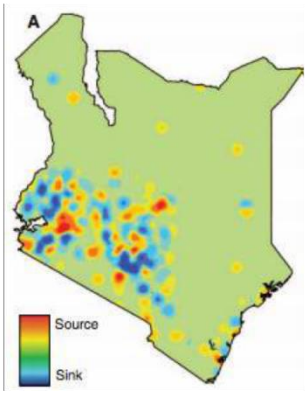


Figure shows sources and sinks of human travel and parasites. Kernel density mapsshow ranked sources (red) and sinks (blue) ofhuman travel

Entre junio de 2008 y junio de 2009 los investigadores localizaron cada llamada o texto realizados por cada uno de los 14.816.521 abonados de telefonía móvil de Kenia en una de las 11.920 torres de telefonía móvil situadas en 692 asentamientos diferentes. Cada vez que un individuo dejaba su asentamiento inicial, se calculaba el destino y la duración de cada recorrido.

Mediante el uso de datos de prevalencia de la enfermedad, los investigadores estimaron la probabilidad de que cada persona estuviera llevando parásitos de la malaria y construyeron un mapa de los movimientos de los parásitos entre las áreas "fuente" (áreas que principalmente propagan la enfermedad) y las áreas "sumidero" (áreas que principalmente reciben la enfermedad). Encontraron que una fracción sorprendentemente grande de las infecciones, son importadas, es decir, son llevadas por personas que se desplazan de un lugar a otro.

Satélites

La abrupta caída del costo de los datos satelitales en la última década ha hecho que la observación de la Tierra y la información geoespacial sean más atractivas que nunca para abordar la pobreza, monitorear los cambios ambientales y estimar el crecimiento económico.

Los satélites tienen sensores que captan la luz reflejada o emitida por la superficie terrestre. Existen de dos tipos:

- Pasivos: son aquellos que pueden registrar información a través de la energía emitida o reflejada por los distintos objetos, siendo el origen de esta energía una fuente ajena al sensor.

- Activos: son aquellos que pueden registrar información a través de la energía reflejada por los distintos objetos, siendo el origen de esta energía el propio sensor. Es decir, que es capaz de emitir su propio haz de energía (Radar[18] y LIDAR[19]).

Los sensores poseen diferente capacidad para registrar información de detalle en función de su resolución espacial, espectral, temporal y radiométrica (Naciones Unidas, 2017).

- Resolución espacial: se refiere al objeto más pequeño que puede ser distinguido en una imagen. Por ejemplo, el sensor MSS de los satélites Landsat tiene una resolución de 80 m. Es decir, que todo objeto cuyas dimensiones sean menores a esa medida, no va a ser registrado en la imagen.

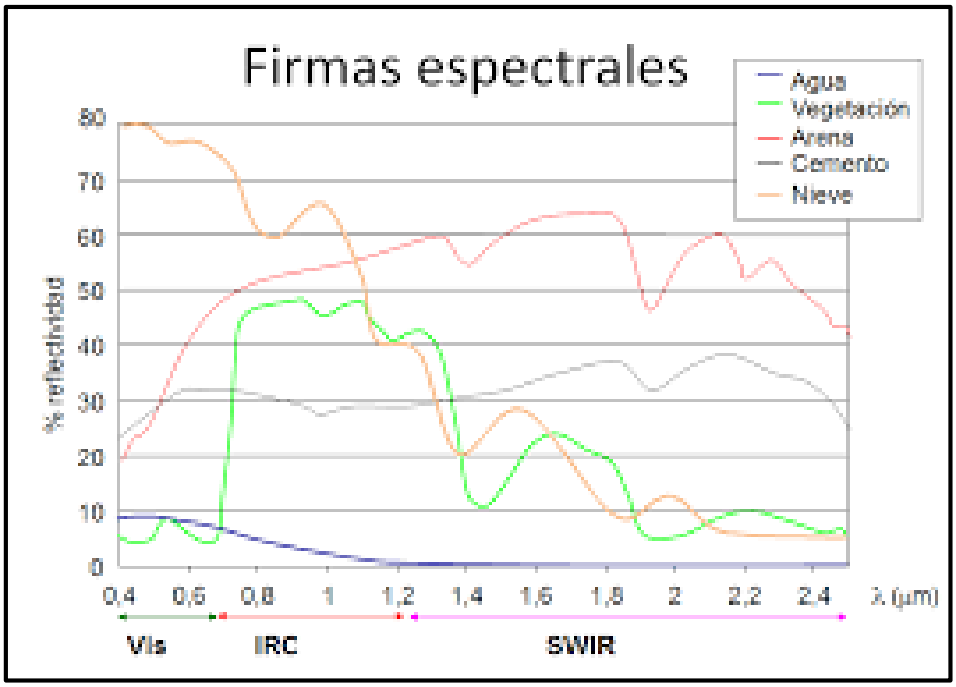
- Resolución espectral: es la cantidad de bandas[20] espectrales en las cuales el sensor puede registrar información. La elección de la cantidad de bandas en las que trabaje un sensor está estrechamente relacionada con los objetivos de su diseño. Por ejemplo, un sensor destinado a fines meteorológicos tiene que contar con una banda en el visible porque no existen diferencias cromáticas en las nubes, dos en el térmico que le permitan conocer la temperatura de dichas nubes y otra en el infrarrojo medio para observar el contenido de humedad en la atmósfera.

- Resolución radiométrica: es la cantidad de energía que puede captar un sensor. En el caso de los sistemas fotográficos se la puede definir como la cantidad de tonos de grises que pueden registrar los mismos. Hay que destacar que cuanto mayor sea la precisión radiométrica, mejor será la interpretación de una imagen.

- Resolución temporal: es la frecuencia con la que el sensor adquiere imágenes de la misma porción de la superficie terrestre, pudiendo conformar la evolución del área registrada. El ciclo de cobertura está en función de las características orbitales de la plataforma (altura, velocidad, inclinación), así como del diseño del sensor. Por ejemplo, los satélites meteorológicos están obligados a facilitar una información muy frecuente, ya que se dedican a observar un fenómeno de gran dinamismo.

De la interacción de la energía electromagnética con los cuerpos que se encuentran sobre la superficie terrestre surgen las denominadas curvas o firmas espectrales. Las firmas espectrales se obtienen del reflejo o emisión de energía a distintas longitudes de onda de una determinada cubierta. En principio, se asume que cada cuerpo u objeto va a poseer una única firma que lo caracterizará. El agua, la vegetación, el cemento y los diferentes cuerpos sobre la superficie de la tierra absorben y reflejan de manera diferente las distintas bandas espectrales generando un patrón único que permite diferenciarlos.

[18] Este dispositivo emite una onda electromagnética, que al rebotar en el objeto regresa al emisor, dando además de las formas, la distancia al mismo. Trabaja en la región de las micro-ondas (0,1 cm y 1 m)
[19] Significa Light-wave Detection and Ranging que en castellano quiere decir Detección y Medición por Ondas Luminosas. Emite pulsos de luz entre el ultravioleta y el infrarrojo cercano.
[20] Una banda espectral designa un rango de longitudes de onda del espectro electromagnético.

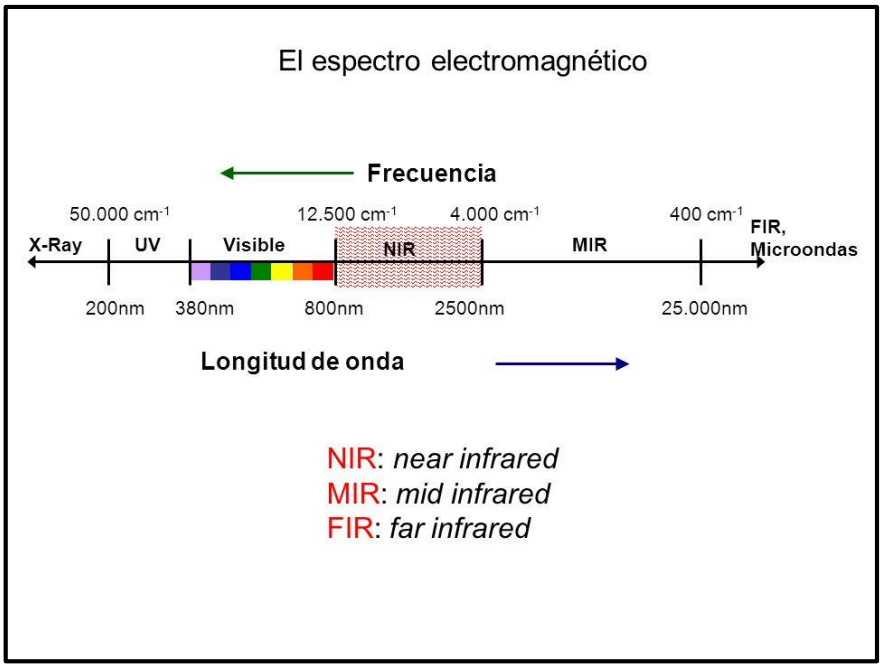


Las bandas espectrales más utilizadas incluyen:

- Espectro visible –VIR - (0,4 a 0,7 micrones): lleva ese nombre porque es la única radiación electromagnética que puede percibir la visión humana. Se distinguen tres bandas elementales, la azul (0,4 a 0,5 micrones), verde (0,5 a 0,6 micrones) y la roja (0,6 a 0,7 micrones).

- Infrarrojo próximo o cercano – NIR y SWIR - (0,7 a 1,3 micrones): esta banda resulta importante por su capacidad de detectar masas vegetales y concentraciones de humedad.
- Infrarrojo medio -MIR- (1,3 a 8 micrones): en esta banda se entremezclan los procesos de reflexión de la luz solar y de emisión de la superficie terrestre.

- Infrarrojo lejano o térmico –TIR- (8 a 14 micrones): incluye la porción emisiva del espectro terrestre. La TIR es la única de estas bandas que se utiliza para imágenes nocturnas.



Existen diversos factores que afectan la firma espectral de un objeto que registra el sensor: a) Angulo de iluminación solar, dependiente de la fecha del año y del momento de paso del satélite, b) Modificaciones que el relieve introduce en el ángulo de iluminación, como las pendiente de las laderas; c) Influencia de la atmósfera, especialmente en lo que se refiere a la dispersión selectiva en distintas longitudes de onda; d) Angulo de observación, relacionado con la órbita del satélite y con las características del sensor.

Estos efectos deben ser corregidos antes de analizar una imagen satelital.

Metodologías de trabajo con imágenes satelitales

Un proyecto de big data en base a imágenes satelitales típicamente incluye las siguientes etapas:

1- ***Obtención de las imágenes:*** las imágenes satelitales de acceso gratuito[21] en general tienen una resolución espacial y temporal menor que las pagas[22].

2- ***Preprocesamiento:*** las imágenes satelitales deben ser corregidas antes de usarse. Por ejemplo, hay que eliminar los efectos introducidos por las partículas de la atmósfera y los que se deben a los ángulos de la radiación solar y la visión del sensor. También hay que georreferenciarlas[23]. Las imágenes que ya han sido corregidas se suelen denominar ARD (*Analysis Ready Data*).

3- ***Tipo de análisis a realizar:*** fundamentalmente existen tres variantes de análisis de imágenes satelitales: i) análisis visual; ii) análisis digital; iii) análisis basado en objetos

[21] Repositorios de imagenes satelitales de acceso gratuito incluyen entre otros: USGS Earth Explorer; ESA Sentinel Mission; NOAA CLASS; NASA Reverb; Earth Observation Link(EOLi); National Institute for Space Research (INPE), specific to South America and Africa; NOAA Data Access Viewer Discover Authoritative Datasets; VITO Vision Coarse Vegetation Data; NOAA Digital Coast Snorkel the Seashore; Global Land Cover Facility Derived Satellite Data; DigitalGlobe Free Product Samples.

[22] Entre otras empresas que se dedican a comercializar imágenes satelitales se encuentran: Digital Globe; AIRBUS Industrie; Planet; Google Terra Bella.

[23] Paso de un sistema de filas y columnas a un sistema de coordenadas estandard.

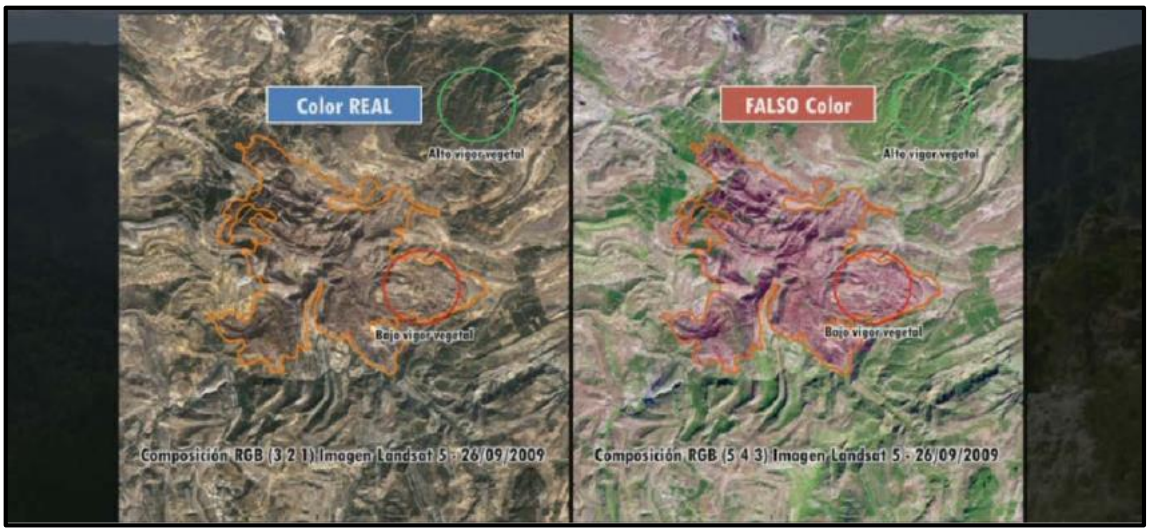
3.1 Análisis visual y composiciones de color

El análisis visual de imágenes es similar en muchos aspectos a la fotointerpretación clásica con las ventajas que aporta la fotografía digital en cuanto a las posibilidades de retocar y realzar las imágenes. Sin embargo una imagen de satélite en bruto presenta un aspecto bastante apagado, por lo que el análisis visual no resulta sencillo. En teledetección se han desarrollado diferentes técnicas que permiten resaltar determinados aspectos para facilitar este análisis.

Aunque el espectro electromagnético abarca un amplio número de regiones y el ojo humano tiene una gran capacidad de discriminación de estos colores, podemos descomponer cualquier color en tres componentes (azul, verde y rojo) que corresponden a tres regiones del espectro visible. Los dispositivos de visualización de imágenes (monitores, televisiones, etc) forman sus imágenes mediante la combinación de diferentes niveles de intensidad en estos tres colores.

Una imagen de satélite tiene varias bandas, algunas de ellas responden a estos colores y otras a regiones fuera del espectro visible. Para visualizarlas podremos pasar las diferentes bandas por cada uno de los cañones o por todos a la vez (imagen en blanco y negro).

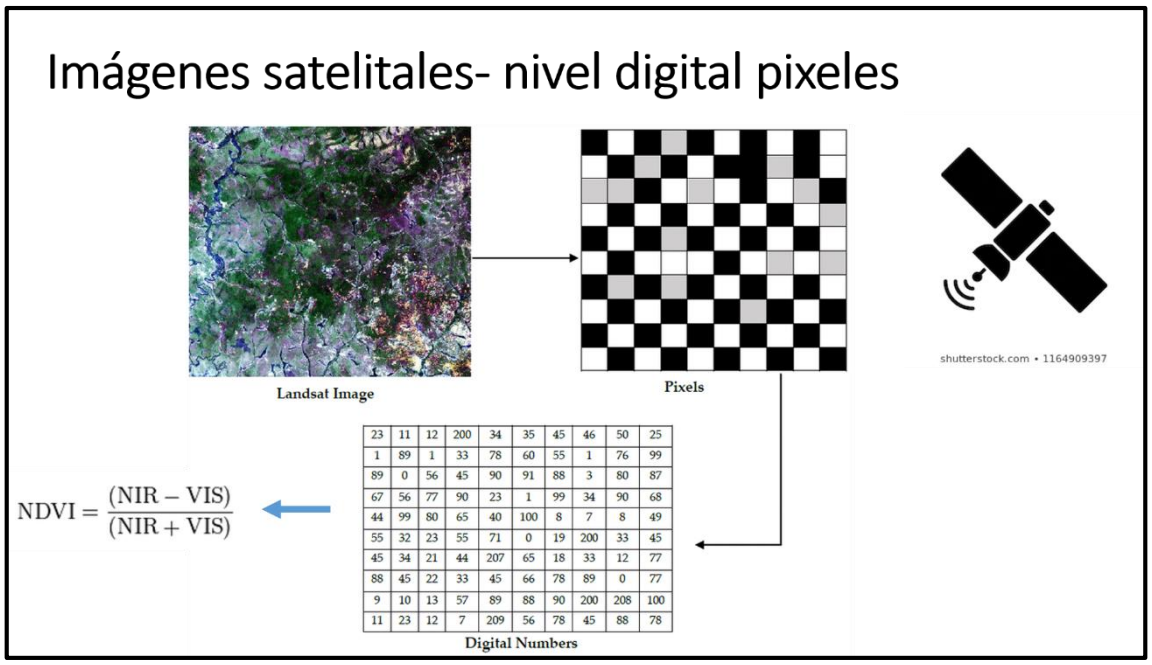
Puesto que la imagen de cada banda representa niveles de intensidad de un color (azul, verde, rojo, etc.) y los monitores y tarjetas de video disponen de 3 canales: R rojo G verde B azul para representar los 3 colores básicos; puede utilizarse cada canal para representar la intensidad de una banda y obtener así una composición de color, la más obvia seria simular el color real. Pero como se dispone de más bandas, nada impide utilizarlas para generar visualizaciones en “falso color”. Estas composiciones servirán para resaltar los elementos que mayor reflectividad presentan en las bandas utilizadas, además de obtener visualizaciones más o menos estéticas. Por ejemplo, si se pasa la banda 4 de landsat (con alta reflectividad por parte de la vegetación) por el canal verde, la vegetación se verá mucho más claramente.



En general, se trata de aprovechar que podemos visualizar tres canales a la vez para introducir las tres bandas que más nos van a ayudar a discriminar visualmente los elementos que nos interesan.

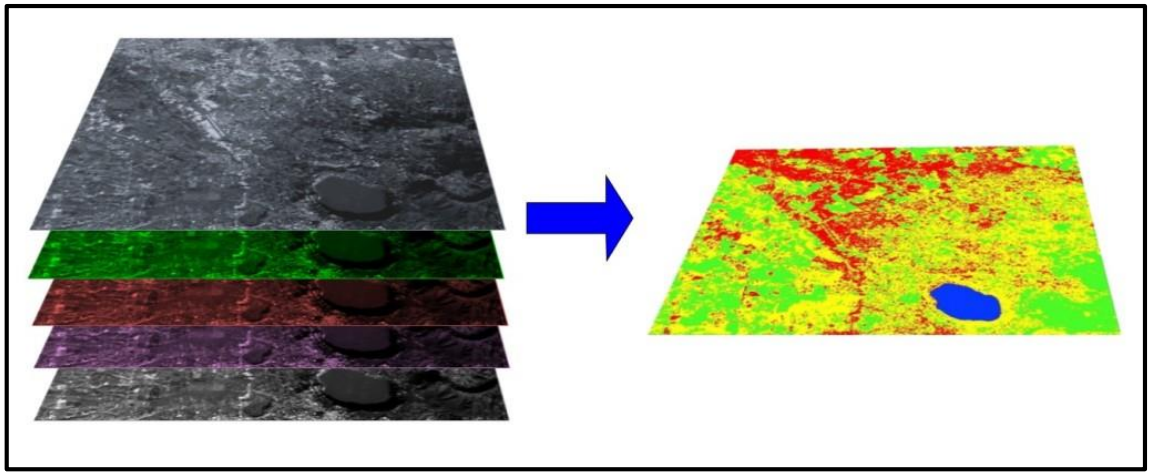
3.2 Análisis digital de imágenes

Una imagen de satélite en bruto, tal como normalmente llega al usuario final, consiste en un conjunto de matrices, una por cada canal del sensor, en la que en cada celda aparecen números del 0 al 255. El cero índica que no llega nada de radiación desde ese punto y el 255 que llega el valor más alto de radiación. Estos valores se denominan “niveles digitales”.

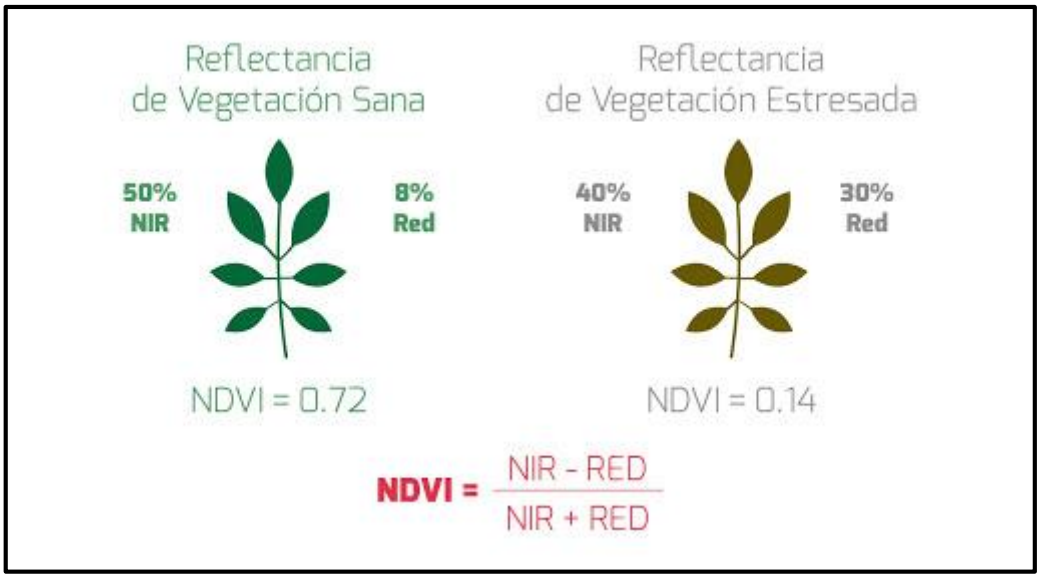


23	11	12	200	34	35	45	46	50	25
1	89	1	33	78	60	55	1	76	99
89	0	56	45	90	91	88	3	80	87
67	56	77	90	23	1	99	34	90	68
44	99	80	65	40	100	8	7	8	49
55	32	23	55	71	0	19	200	33	45
45	34	21	44	207	65	18	33	12	77
88	45	22	33	45	66	78	89	0	77
9	10	13	57	89	88	90	200	208	100
11	23	12	7	209	56	78	45	88	78

Cuando interesa detectar algún aspecto específico de la superficie terrestre, se construyen índices en base a fórmulas específicas que combinan diferentes bandas. .



Uno de los índices más utilizados es el NDVI (Indice Normalizado de Vegetación). El mismo, se basa en que la vegetación tiene una reflectividad muy alta en la banda 4 (infrarrojo cercano) del sensor del landsat y muy baja en la banda 3 (visible – rojo). Por tanto cuanto mayor sea la diferencia entre ambas bandas mayor es el porcentaje de cobertura vegetal y más sana es esta.



Las imágenes satelitales se utilizan para estimar rendimientos de cultivos y a través de ello impuestos al agro[24].

La Agencia de Recaudación de la Provincia de Buenos Aires viene realizando en forma sistemática un proceso de Monitoreo Estratégico Satelital Integrado (MESI), basado en la utilización de tecnología satelital, a los efectos de fiscalizar el cumplimiento de las obligaciones tributarias y detectar maniobras de evasión.

Las variables agronómicas más importantes de los cultivos que pueden ser monitoreadas a partir de imágenes satelitales incluyen: el contenido de agua de los cultivos, el índice de área foliar y la biomasa. Estas variables son habitualmente utilizadas como insumos para modelos de estimación de rendimiento de cultivos.

A partir de la utilización de tecnología satelital se puede estimar el rendimiento de los cultivos de grano, para luego contrastarlos con su situación de inscripción e información expuesta en sus declaraciones juradas del Impuesto.

Estimación de la concentración de geis en la atmósfera (ODS 13)

Los satélites también están siendo utilizados para el estudio de la atmosfera (*Matsunaga et al., 2018*). El ENVISAT, fue el primer satélite equipado para cuantificar la concentración de metano atmosférico, y funcionó durante un breve período (2003-2006); a partir de 2009 se cuenta con los datos del GOSAT que tienen mayor frecuencia temporal y una menor frecuencia espacial. Ambos equipos cuentan con detectores tipo radiómetro de onda corta (SWIR). Los datos son posteriormente sometidos a una selección importante, ya que se descartan aquellos tomados en momentos de alta nubosidad o alta concentración de aerosoles en la atmósfera entre otras razones.

Para aplicar los datos de estos satélites al monitoreo de las concentraciones de gases efecto invernadero[25] como el CO2 y el CH4 es necesario tener medidas en tierra para calibrarlos, ya que

[24] https://www.arba.gov.ar/Informacion/InfoGeneral/Catastro/PropiedadesInfraccion.asp
[25] Las nuevas guías del IPCC incorporan la posibilidad de ajustar inventarios nacionales de Geis a través de mediciones atmosféricas incluyendo aquellas que llevan adelante satélites

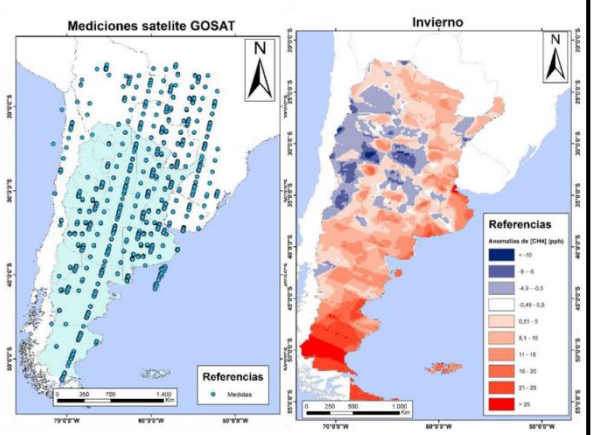
la concentración medida es un promedio a través de la columna de aire en la que se realiza la detección. Para validar las abundancias de las columnas de CO2 y CH4, el equipo de validación de GOSAT utiliza información de espectrómetros de Fourier de alta resolución ubicados en la superficie terrestre e instrumentos de observación instalados en aeronaves

En Argentina se han hecho estudios vinculados a la identificación de fuentes y sumideros de CH4 a partir de las series de datos 2003 y 2004. *Marinone (2016)* pudo construir un mapa de anomalías para la Argentina (zonas con concentraciones desviadas del promedio) diferenciándose regiones de fuerte emisión, como son los lagos, zonas de cría de rumiantes y zonas urbanas, de otras con alto secuestro, como por ejemplo zonas áridas

13 ACCIÓN POR EL CLIMA

13.2 Incorporar medidas relativas al cambio climático …

- Satélite japonés de Observación de Gases de Efecto Invernadero (GOSAT, por sus siglas en inglés) para monitorear los niveles globales de dióxido de carbono (CO2) y metano (CH4) desde el espacio



Identificación de fuentes y sumideros de metano dentro del territorio nacional a partir de mediciones satelitales.
Marinone Esteban, FCE, UNCPBA (2016)

Luminosidad nocturna como proxy de la actividad económica (ODS 8)

Durante casi 30 años, los científicos han utilizado imágenes satelitales nocturnas de la Tierra para estudiar la actividad humana.

Cada píxel de una imagen satelital representa una de la Tierra asociada con un número digital que mide la luminosidad durante la noche. Cuanto más luminoso es el lugar, más alto es el número para ese píxel. La agregación de estos números para todos los píxeles en un país se convierte en un indicador del nivel de actividad de ese país por la noche. Cuando un indicador de este tipo se compara entre países y a lo largo del tiempo, se transforma en un instrumento de medición del desarrollo y las fluctuaciones económicas.

Las imágenes satelitales de luminosidad que se utilizan provienen de dos fuentes de información: el Departamento de Defensa de Estados Unidos para el periodo 1993-2013; y la Administración Nacional de la Aeronáutica y del Espacio[26] para el periodo 2014 - 2017.

https://public.wmo.int/en/resources/bulletin/from-atmospheric-observations-and-analysis-of-greenhouse-gases-emission-estimates

[26] Dentro de cada pixel la luminosidad es medida en una escala de 0 (ausencia total de luz) a 63 (grado máximo de luminosidad) en el caso del DoD, mientras que las imágenes generadas por la NASA no tienen un límite preestablecido.

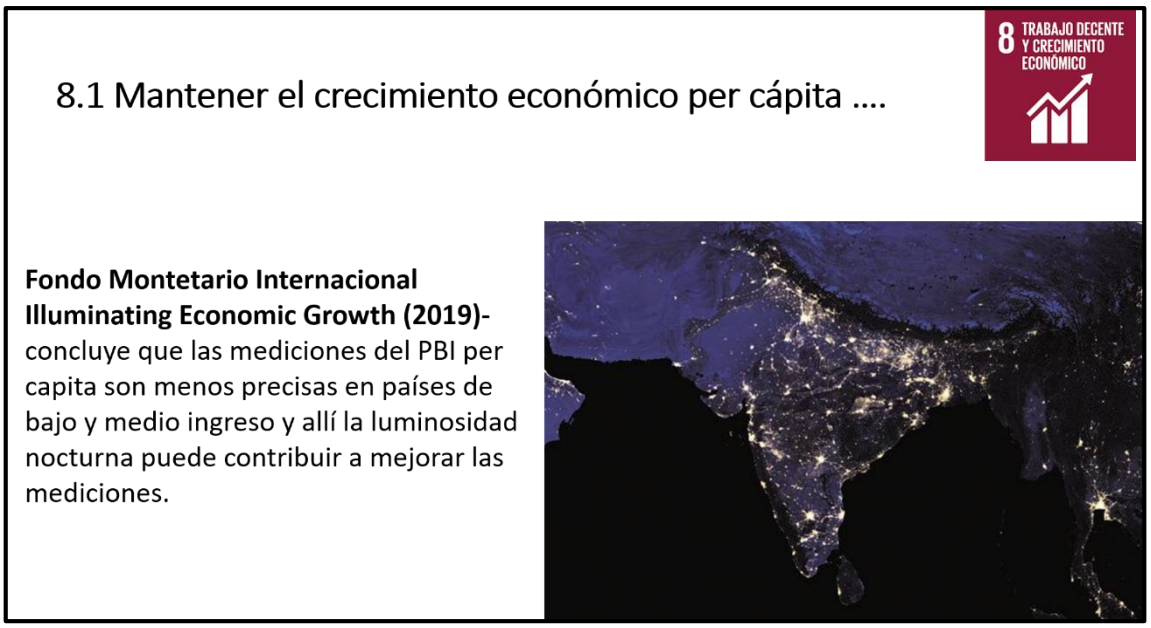
La importancia de las luces nocturnas para la economía está dada por su fuerte correlación con actividades económicas, aunque gran parte de estas ocurren durante el día. En las economías en crecimiento, con el transcurso del tiempo se iluminan más áreas y un mayor número de píxeles comienzan a registrar luz. En cambio, en regiones envueltas en conflictos, un mayor número de parcelas se oscurecen, y un mayor número de píxeles comienzan a perder luminosidad. El otro aspecto es la intensificación. A medida que se urbanizan las zonas rurales, se aglomeran las ciudades y se moderniza la infraestructura, el mismo cielo nocturno se ilumina y los sensores satelitales captan luz de mayor intensidad.

Los países en etapas rudimentarias de desarrollo se concentran en gran medida en la infraestructura: construir carreteras y puentes, edificar estaciones ferroviarias y aeropuertos y renovar el tendido eléctrico y las telecomunicaciones, todo lo cual emite luz por la noche. En consecuencia, con el crecimiento de la economía el cielo nocturno se presenta cada vez más luminoso en las imágenes satelitales.

Las economías avanzadas, por otro lado, potencian su economía con innovación científica y tecnológica, y el consiguiente aumento de la productividad suele tener menos que ver con las luces nocturnas que con la infraestructura que subyace a esa innovación.

Un informe reciente del Fondo Monetario Internacional (*IMF ,2019)* recomienda el uso de la tecnología satelital para ajustar las estimaciones del PBI en países de renta media y baja, donde las estimaciones por las metodologías tradicionales tienen márgenes de error grandes.

Desde el sector privado también existen iniciativas en marcha. Por ejemplo, la compañía japonesa Nowcast[27] brinda un servicio de pronóstico del PIB utilizando, entre otras, la iluminación nocturna medida desde satélites.



Las imágenes nocturnas también están permitiendo estimar, a partir de los cambios de luminosidad en el tiempo, las regiones que presentan mejoras en el acceso a la energía eléctrica[28].

[27] https://www.nowcast.co.jp/en
[28] https://phys.org/news/2017-08-electricity-access-sixth-world-people.html

Dinámica de crecimiento de las ciudades (ODS 11)

Las imágenes satelitales nocturnas también se utilizan para entender la dinámica de crecimiento de las ciudades, que es fundamental para lograr una planificación territorial sostenible.

Buzai et al. (2020) a través del procesamiento digital de imágenes satelitales nocturnas y el uso de Sistemas de Información Geográfica analizaron la expansión y conurbación de la megaciudad Buenos Aires desde 1992 hasta el 2012.

Como el objetivo de la investigación era determinar la extensión de la superficie urbana a partir de la superficie iluminada y, como no todos los píxeles con valores mayores que cero corresponden a las áreas de cobertura urbana, se estimó un valor mínimo (intensidad lumínica) de referencia ligado a diversas construcciones materiales humanas (que en el presente trabajo se denominan de manera amplia infraestructura gris). Para concretar este paso, fue necesario contar con una capa de cobertura urbana que permitiera realizar una superposicion sobre las imágenes nocturnas y poder aproximar un valor mínimo como umbral de luminosidad urbana. A partir de ello se estableció como valor umbral de luz urbana 44.41. A partir de este valor de referencia se delimitó la superficie de luminosidad urbana en las imágenes de 1992, 2002 y 2012, tomándose en consideración aquellos píxeles que tuvieran valores ≥ 44.41.

Detección de cambios en el agua subterránea almacenada (ODS 6)

Los satélites además de medir radiación electromagnética tienen la capacidad de medir variaciones en el campo gravitatorio de la superficie del planeta. Estas variaciones han mostrado ser de utilidad para detectar cambios en el agua subterránea almacenada.

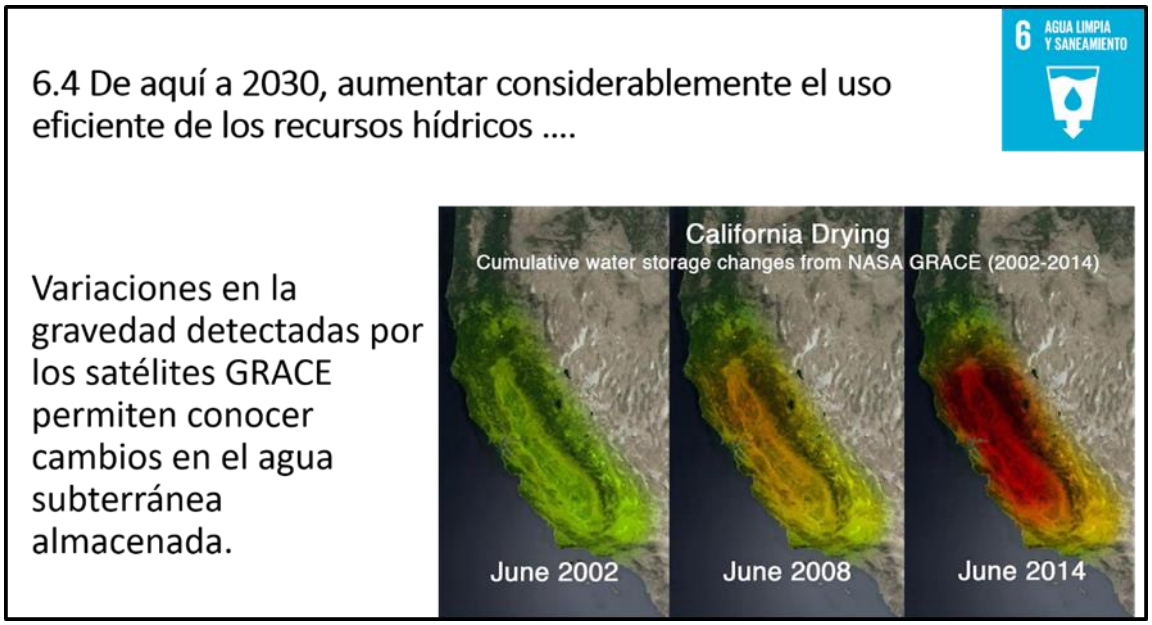
Los satélites GRACE llevan a bordo instrumentos ultrasensibles para detectar pequeñas variaciones en la atracción gravitacional del planeta.

Las mediciones precisas de la gravedad tienen aplicaciones importantes en una variedad de campos. Los datos suministrados por los satélites son utilizados para estudiar las corrientes

marinas, los cambios del nivel del mar, la altura de la superficie terrestre y los niveles de reserva de agua subterránea.

Richey et al. (2015) utilizando datos de los satélites climáticos GRACE encontraron que 13 de los 37 acuíferos más grandes del planeta estudiados entre 2003 y 2013 se estaban agotando, ya que reciben poca o ninguna recarga. Ocho fueron clasificados como "estresados", casi sin reposición natural para compensar el uso. Otros cinco resultaron ser "extremadamente" o "muy estresados", dependiendo del nivel de reposición de cada uno, según informa la NASA. Es decir una parte significativa de la Humanidad está consumiendo agua subterránea rápidamente sin saber cuándo podría agotarse.

Los acuíferos más sobrecargados están en las zonas más secas del mundo, donde las poblaciones usan en gran medida de las aguas subterráneas. Se espera que el cambio climático y el crecimiento de la población intensifiquen el problema.



El equipo de investigación encontró que el Sistema Acuífero de Arabia, una importante fuente de agua para más de 60 millones de personas, es el que padece la tensión más excesiva en el mundo. El acuífero de la Cuenca del Indo en el noroeste de la India y Pakistán es el segundo más estresado, y la Cuenca del Murzuk-Djado en el norte de África es el tercero. El Valle Central de California, que se utiliza en gran medida para la agricultura y sufre un rápido agotamiento, va un poco mejor, pero todavía se le considera altamente estresado.

Monitoreo de la deforestación (ODS 15)

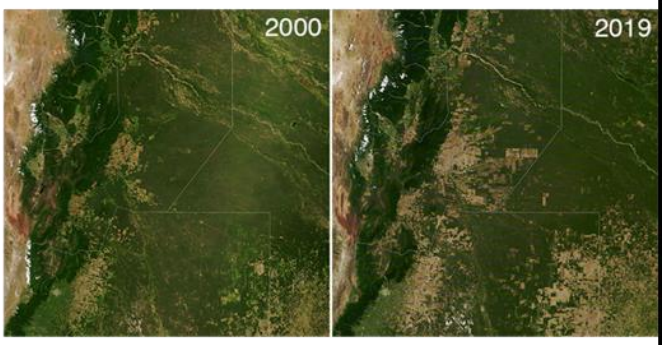
Las imágenes ofrecidas por los satélites de observación de la Tierra se han convertido en la principal herramienta para la determinación del estado de la masa forestal y el avance de fenómenos como la deforestación y desertificación en diferentes regiones del globo.

En Argentina como en otros países del mundo hace años que las estadísticas forestales se basan en la utilización de imágenes satelitales. Además desde el 2018 también contamos con el SAT, una herramienta que monitorea la perdida de bosque nativo de forma continua y mediante procesos automatizados basados en imágenes satelitales.

15.1 De aquí a 2020, asegurar la conservación, el restablecimiento y el uso sostenible de los ecosistemas terrestres y los ecosistemas interiores de agua dulce y sus servicios, en particular los bosques …

El monitoreo de la superficie de bosque nativo se realiza utilizando técnicas de teledetección y un sistema de información geográfica.

Desde el 2018 contamos con el Sistema de Alerta temprana de deforestación (SAT)



El SAT es una herramienta que monitorea la pérdida de bosque nativo de forma continua, mediante procesos automatizados basados en imágenes satelitales. Comenzó a funcionar en forma operativa en nuestra cartera en noviembre de 2018 para las subregiones del Chaco húmedo y Chaco semiárido, y tiene como objetivo fortalecer las acciones de control y vigilancia sobre los bosques nativos de las autoridades locales de aplicación de la Ley n.° 26331.

Desde el mes de enero de este año, se están implementando nuevos algoritmos para el procesamiento de las imágenes satelitales para la generación de las alertas. Los mismos se realizan completamente por medio del Google Earth Engine.

Como resultado, las alertas son generadas en menos tiempo y con resultados más precisos, permitiendo emitir los reportes a las provincias con una mayor frecuencia.

El sistema procesa automáticamente cada 16 días imágenes satelitales MODIS y Landsat 8, aplicando 3 algoritmos que analizan con diversas técnicas, series de tiempo y patrones espaciales.

Finalmente, se envía a cada provincia un reporte con el detalle de las alertas y el requerimiento de información sobre la legalidad de cada evento de deforestación (si estaba autorizado o no, instrumento que autoriza el desmonte, número de expediente y medidas a tomar en caso de los eventos ilegales, entre otros datos).

3.3 Análisis de imágenes basado en objetos

Hasta ahora venimos hablando de como analizar las imágenes satelitales en función del valor de los píxeles individualmente sin considerar su contexto. Pero las imágenes satelitales pueden ser abordadas a partir de técnicas que permiten identificar objetos como podría ser el techo de una vivienda. Este nuevo paradigma se denomina “análisis de imágenes basado en objetos" (OBIA por sus siglas en inglés).

Muchas veces la información necesaria para el análisis de las imágenes de satélite no se encuentra en los píxeles de la imagen, sino en los objetos significativos de la misma y en sus relaciones mutuas.

La clasificación basada en los objetos de una imagen tiene en cuenta, entre otros aspectos, las formas, las texturas y la información espectral presentes en la imagen. Para llevarla adelante se utilizan algoritmos de segmentación. Los mismos se pueden clasificar en: a) Basados en el histograma: agrupan píxeles que tienen las mismas propiedades; b) Basados en la detección de bordes: los objetos destacan de su entorno y tienen bordes definidos; y c) Basados en regiones: combinan información de ubicación espacial y similaridad de los píxeles.

Material de los techos como proxy de pobreza en Africa (ODS 1)

En muchos países de Africa el material del techo de una vivienda es un buen indicador del nivel de pobreza del hogar. En general las familias que viven bajo un techo de metal tienen mayores ingresos que las que lo hace bajo un techo de paja. A partir de este hecho y utilizando imágenes satelitales de distintos años, un proyecto de Global Pulse[29] se propuso identificar regiones con procesos de reducción de pobreza a partir de contabilizar la cantidad de techos de metal con relación a los de paja.

1.2 De aquí a 2030, reducir al menos a la mitad la proporción de hombres, mujeres y niños de todas las edades que viven en la pobreza en todas sus dimensiones …

Global Pulse-
Contabilidad de techos de paja xs techos de metal como proxy de la pobreza en ciertas regiones de Africa.
.

Por diversas razones[30] el algoritmo utilizado para contabilizar los techos no funcionó del todo bien, pero sirvió para inspirar el desarrollo de otros trabajos como el que sigue a continuación.

Rasgos en imágenes asociados con la pobreza (ODS 1)

Una pregunta clave para el monitoreo de la pobreza a partir de imágenes satelitales es determinar los rasgos a buscar en las imágenes. Eso lo puede definir manualmente un humano (como en el caso de los techos referido previamente) o de manera automática una máquina.

¿Qué datos recogidos por satélite ayudan a detectar la pobreza? Las carreteras sin asfaltar, extensiones agrícolas o instalaciones ganaderas, los recursos hídricos, los materiales que

[29] https://www.unglobalpulse.org/project/measuring-poverty-with-machine-roof-counting
[30] Que se explican aquí https://www.datakind.org/projects/using-the-simple-to-be-radical/

componen los tejados de las viviendas en las zonas urbanas o rurales, entre otros, son factores que ayudan a discernir una zona próspera de otra donde hay pobreza.

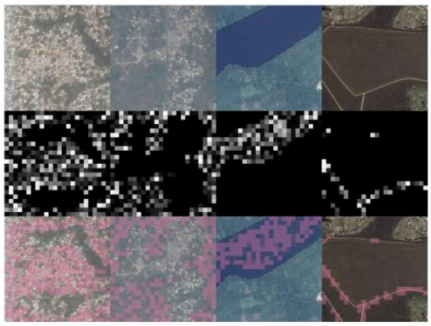
Jean et al (2016) cruzan datos entre imágenes nocturnas y diurnas como una novedad con respecto a estudios anteriores que también pretendían medir el grado de pobreza de una zona.

En este trabajo se entrenó un algoritmo mostrándoles imágenes satelitales diurnas y nocturnas del mismo lugar. En base a este entrenamiento el algoritmo aprendió a diferenciar rasgos (como caminos pavimentados, granjas, techos de metal, etc.) en las imágenes diurnas que se correspondían con mayores o menores niveles de pobreza (evaluados a partir de la mayor o menor luminosidad nocturna).

Usando esta metodología los investigadores lograron trazar los niveles de pobreza (estimando el consumo promedio per cápita en dólares) en los países de prueba: Nigeria, Uganda, Tanzania y Malawi.

1.2 De aquí a 2030, reducir al menos a la mitad la proporción de hombres, mujeres y niños de todas las edades que viven en la pobreza en todas sus dimensiones …

Jean et al (2016) - a partir de imágenes satelitales diurnas y nocturnas estimaron el consumo promedio per cápita en dólares de países en Africa.



Learned features corresponding to buildings, undeveloped areas, water, and roads

1- Imágenes de día y noche
2- Entrenaron algoritmo usando:
   - Input: imágenes de día
   - Output: imágenes de noche
3- El algoritmo aprendió a identificar rasgos en las imágenes de día que implican mayor o menor grado de pobreza (luminosidad nocturna)

Detección de villas y asentamientos urbanos (ODS 11)

Las villas y asentamientos en general se caracterizan por el trazado irregular de las manzanas y calles con una alta densidad de techos por manzana y escases de vegetación y espacios abiertos. Estas características se pueden analizar a partir de imágenes satelitales.

Bayle (2018) trabajó en la detección de villas y asentamientos en el Partido de La Matanza utilizando, entre otras fuentes de información, imágenes satelitales.

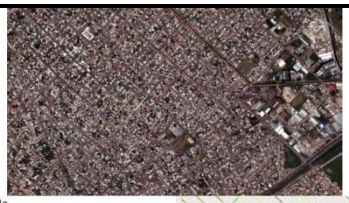
El objetivo de este trabajo fue elaborar una metodología que identificara potenciales villas y asentamientos permitiendo reducir las zonas donde realizar trabajo de campo para relevarlas.

Las imágenes utilizadas en este trabajo fueron tomadas por el satélite comercial WorldView-2 (WV2). Las imágenes en el canal pancromático tienen una resolución de 46cm mientras que las de las bandas multiespectrales tienen 1.85m.

Los atributos más importantes en la clasificación de las imágenes dependieron del algoritmo utilizado, pero los estadísticos calculados sobre el NDVI, cercanía a distinto tipo de calles, vías de tren y cursos de agua fueron en general los más importantes.

Considerando solamente imágenes satelitales, la metodología logra reducir en un 30% del área total, el territorio a estudiar.

11.1 De aquí a 2030, asegurar el acceso de todas las personas a viviendas y servicios básicos adecuados …



Atributos en las imágenes:

- Forma de manzanas
- Trazados de calles
- Densidad de techos y vegetación
- Materia de techos y tamaño

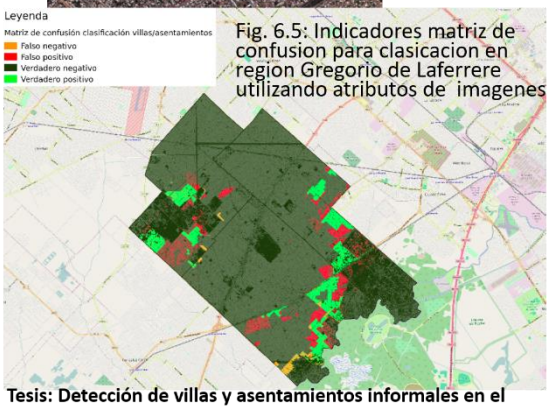


Fig. 6.5: Indicadores matriz de confusion para clasicacion en region Gregorio de Laferrere utilizando atributos de imagenes

Tesis: Detección de villas y asentamientos informales en el partido de La Matanza (Baylé, 2016)

Actividad industrial y monitoreo de existencias (ODS 9)

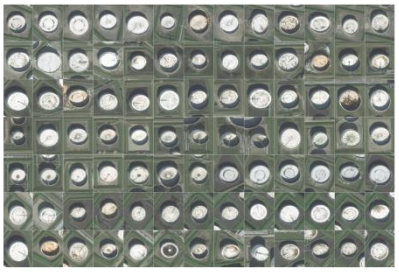
Las imágenes también pueden dar información sobre las existencias y nivel de actividad de las industrias. Por ejemplo, la contabilización en imágenes satelitales de los camiones en los estacionamientos de las fábricas funciona como un proxy de la producción industrial.

Spaceknow, una empresa norteamericana, generó a partir de imágenes satelitales infrarrojas, un índice de actividad de las fábricas en la región de Cantón, China. El índice identifica a partir de un algoritmo específico signos de actividad económica, como los vehículos de transporte en los estacionamientos.

Orbital Insight, otra compañía norteamericana, monitorea actividades en más de 260.000 estacionamientos de minoristas en todo el país. Por otro lado, también desarrollo una metodología para estimar satelitalmente las existencias en tanques de compañías petroleras.

9.2 Promover una industrialización inclusiva y sostenible …

SpaceKnow- Contabilizando camiones y autos través de imágenes satelitales en estacionamientos de comercios y fábricas es posible predecir su nivel de actividad.



Google- la cantidad de crudo almacenado en tanques puede estimarse a partir de las sombras que proyectan los mismos observadas desde satélites.

El crudo se almacena en tanques cuya capacidad puede estimarse a partir de las sombras que proyectan. Cuando un tanque está lleno, la sombra es mínima a diferencia de cuando está vacío. Midiendo el tamaño de estas sombras, que tienen una forma similar a la de la fase lunar creciente, se obtiene una estimación casi en tiempo real de las reservas de petróleo de las compañías[31].

Twitter

Obtener información a partir de las publicaciones que realizan las personas día a día en Twitter es una de las principales formas de construir una base de datos para ser analizada posteriormente con big data. Con estos datos, es posible conocer los gustos, preferencias e interacciones de las personas a través de Internet.

La base de datos de Twitter contiene información en tiempo real sobre muchos temas, incluyendo el costo de los alimentos, la disponibilidad de puestos de trabajo, el acceso a la asistencia sanitaria, la calidad de la educación, y los informes de los desastres naturales.

En este contexto, Twitter y la iniciativa Global Pulse Naciones Unidas establecieron en 2016 una asociación que proporcionará a las Naciones Unidas el acceso a las herramientas de datos de Twitter para apoyar los esfuerzos para alcanzar los Objetivos de Desarrollo Sostenible.

Metodología de trabajo con tweets

Para generar conocimiento a partir de tweets es necesario considerar al menos los siguientes pasos:

1- ***Extracción de datos de Twitter:*** un tweet es un objeto con información diversa (geolocalización, usuario y otros metadatos) pero en general lo que nos interesa es el

[31]https://www.researchgate.net/publication/332193936_Estimating_the_Volume_of_Oil_Tanks_Based_on_High-Resolution_Remote_Sensing_Images

mensaje que el usuario escribió en forma de texto. Una alternativa para extraer estos datos es utilizar la API de Twitter.

2- ***Transformacion de tweets a texto***: la colección de tweets es transformada en un "corpus" extrayendo solo el texto de cada publicación. El producto de esta etapa es un listado de documentos (oraciones) llamado corpus.

3- ***Tokenizacion:*** cada documento de nuestro corpus es transformado en un listado de términos llamados tokens. Esta representación de datos también es conocida como bolsa de palabras (bag of words). Los tokens son cadenas de caracteres entre espacios en blanco o puntuación. El conjunto total de palabras utilizadas, distintas y únicas, es el vocabulario del corpus. En este paso, también se filtran las palabras que no tienen una semántica referencial clara, como artículos, pronombres, preposiciones, etc. stop words que son muy frecuentes en el lenguaje.

4- ***Stemming.*** Muchas veces diferentes tokens pueden hacer referencia al mismo concepto, ya que pueden ser representados por variantes morfológicas de una misma ´ familia de palabras. Por ejemplo, podemos representar "escribo", "escribíamos" y "escribimos", en su raíz como "escrib" ya que tienen un significado similar y derivan del mismo verbo. Esto nos permite relacionar aquellos tweets que contienen palabras con similitud semántica gracias a su raíz, lo que facilitará los procesos de análisis posteriores. A este proceso de reducción de tweets se lo denomina "stemming".

5- ***Vectorizacion.*** el próximo paso consiste en generar a partir de la lista de palabras de cada tweet un vector donde cada columna es una característica. De esta forma, podemos considerar cada palabra del vocabulario como una columna, o podemos obtener representaciones más complejas si consideramos como posibles características secuencias de palabras, llamadas n-gramas, que han ocurrido en el texto. Los n-gramas pueden ser secuencias de una palabra (unigramas), de dos palabras (bigramas), de tres palabras (trigramas), y así sucesivamente.

6- ***Construcción de una matriz binaria:*** la suma de todos los vectores (cada uno correspondiente a un tweet) forma una matriz que abarca al conjunto de tweets. Cada tweet se representa en una fila, y los valores que toma en cada columna están determinados por la ocurrencia de las palabras de cada columna en el tweet. La forma más simple de asignar valores a las celdas es con valores binarios: 1 si el n-grama representado por la columna i ocurre en el tweet j, y 0 si no ocurre. Teniendo un vocabulario grande, es de esperar que la mayor parte de n-gramas no ocurra en un tweet. Por lo tanto las matrices serán ralas, es decir, con gran cantidad de ceros. A esta forma de representar los tweets la identificaremos como binarizacion.

7- ***Filtrado en función de la frecuencia de las características***: una palabra o n-grama que aparece en la mayor parte de los tweets no brinda mucha información para diferenciar un tweet de otro. Para esto es útil definir un umbral de frecuencia para determinar que características deben ser incluidas en la matriz, de tal forma que aquellas características que son muy frecuentes sean ignoradas, ya que pueden ser consideradas stop words propias del corpus. Tampoco debemos incluir aquellas características que son poco frecuentes ya que no proveen generalizaciones sobre tendencias en los textos.

Consecuentemente resulta habitual filtrar características en la matriz que estén por arriba o debajo de umbrales definidos.

8- ***Analítica***. el análisis del tweets implica técnicas diferentes según el objetivo. Una primera alternativa consiste en utilizar el algoritmo K-Means de clustering para identificar grupos de tweets que hablan sobre temas comunes en función de la frecuencia con que aparecen determinadas palabras. Otra posibilidad es clasificar tweets sobre un tema específico a través de técnicas de "análisis de sentimiento" que se basan en determinar si la valoración del tema por un usuario es positiva, neutra o negativa.

Monitoreo de la inflación de los alimentos (ODS 2)

Los resultados de un estudio conducido por *Global Pulse (2014)* indicaron la existencia de una relación entre las estadísticas de inflación de la comida y el volumen de tweets vinculados al incremento de precios. Los investigadores analizaron tweets relacionados con los precios de los alimentos en Indonesia durante dos años a través del desarrollo de taxonomías relacionadas al término y un algoritmo de clasificación para categorizar los resultados. Posteriormente a partir de un análisis de series temporales identificaron una correlación entre las estadísticas de inflación del alimento y los tweets relacionados con el precio de los alimentos. De esta manera los investigadores esperan que a partir del monitoreo de las redes se puedan identificar procesos inflacionarios antes de que sean registrables en las estadísticas oficiales.

Google trends

Otra fuente importante de big data es Google Trends (tendencias en español) un servicio gratuito de Google que cuantifica las búsquedas realizadas en este motor.

Teniendo en cuenta que, en nuestro entorno, más del 85% de las búsquedas se realizan en Google, los datos de este buscador constituyen una fuente muy completa para conocer los intereses de la población.

El servicio permite conocer la frecuencia con que se realizan búsquedas de términos específicos con la posibilidad de acotarla por países y regiones del mundo y en un período de tiempo determinado.

Para cada palabra o frase clave de búsqueda lo que se obtiene es un índice de intensidad relativa de búsquedas. Este índice, se construye dividiendo el número de búsquedas de una palabra o frase clave entre el número total de búsquedas en Google[32]. Posteriormente se normaliza el índice de 0 a 100, siendo 100 el número de búsquedas más alto para el periodo analizado

Un proyecto big data basado en búsquedas en google trends incluye al menos los siguientes pasos:

1- Elegir los términos clave de búsqueda para el fenómeno que quiero abordar[33]. Por ejemplo, si el objetivo es analizar el desempleo se podrían buscar las tendencias de un término como "bolsa de trabajo".

2- Extraer los valores del índice relativo de búsqueda durante un periodo de tiempo dado para el término elegido en google trends.

3- Comparar estadísticamente (por ejemplo a través de la determinación de coeficientes de correlación) la serie temporal del índice de búsqueda para el término con una serie de datos reales, por ejemplo la tasa de desempleo medida por la Oficina Estadística Nacional.

Monitoreando el desempleo (ODS 8)

Considerando que internet es cada vez un medio más importante de búsquedas de empleo, los investigadores están usando Google Trends para estudiar diversos aspectos vinculados con esta temática (*Pavlicek et al.,2015;* Marinescu et al.,2016)

Vázquez et al. (2020) buscaron predecir la tasa de desempleo de México antes de que el Instituto Nacional de Estadística y Geografía (INEGI) publique la información oficial.

Los investigadores encontraron una correlación significativa entre las series temporales de la tasa de desempleo medida por encuestas y el índice relativo de búsquedas en google trends para los términos "empleo + 'bolsa de trabajo'. Además, utilizaron diferentes modelos para predecir su valor en función del índice de búsquedas.

[32] De esta manera se independizan de los efectos asociados con el crecimiento en la cantidad de usuarios de google. Por otro lado, hay que considerar que una disminución en el valor del índice de un para un determinado término de búsqueda, puede explicarse tanto por un menor número de búsquedas asociadas a este término como por un aumento en el resto de las búsquedas de Google.

[33] Para ello, se puede utilizar "Google Correlate", una herramienta que permite encontrar asociaciones entre diferentes palabras para todas las búsquedas realizadas en Google.

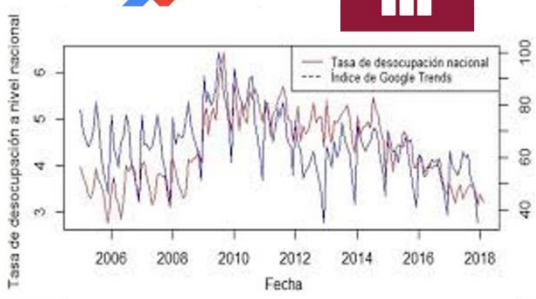
8.5 De aquí a 2030, lograr el empleo pleno y productivo …

1) Seleccionar una o mas palabras clave correlacionadas con el fenómeno que se desea medir. Ej: búsquedas con la palabra “bolsa de trabajo” para el monitoreo del desempleo.

2- Analizan correlación entre series temporales de tasa de desempleo y el índice de intensidad relativa de búsqueda de los términos en google.

Google Trends

8 TRABAJO DECENTE Y CRECIMIENTO ECONÓMICO



GRANDES DATOS, GOOGLE Y DESEMPLEO

BIG DATA, GOOGLE AND UNEMPLOYMENT

Raymundo M. Campos Vázquez

Sergio E. López-Araiza B.

El Colegio de México, A. C.

Detección temprana de epidemias (ODS 3)

La detección temprana de epidemias es un objetivo importante de la vigilancia para permitir una intervención oportuna. No obstante, los datos de vigilancia tradicionales pueden no estar disponibles en el plazo requerido para el control epidémico agudo.

Millones de personas buscan información en línea sobre enfermedades específicas o síntomas médicos. Esto hace que las consultas de búsqueda en internet sean una fuente de información valiosa sobre las tendencias de salud y puedan utilizarse con fines de vigilancia.

Existen estudios (*Sudhakar et al., 2014*) que han demostrado que las búsquedas sobre palabras asociadas a síntomas o medicación se vinculan con tendencias estacionales de diversas enfermedades. Estos estudios se basan en el análisis de google trends y otros datos (ej. Climáticos) para buscar correlaciones. Estos análisis han demostrado ser de utilidad para mapear la propagación espacial de enfermedades lo que facilita el manejo del riesgo.

Quizás la aplicación más conocida de google trends haya sido su algoritmo para predecir picos de gripe.

En 2009 la revista “Nature”[34] publicó un estudio que mostraba cómo las consultas en el motor de búsqueda de Google se habían traducido en una predicción casi exacta de la incidencia de la gripe en cada región de EEUU.

A partir del análisis de millones de búsquedas relacionadas con la gripe —"síntomas gripe", "virus gripe", se podría decir casi al instante si habría colas en las urgencias en determinado punto del país. Mientras, los sistemas predictivos de los Centros para el Control y la Prevención de Enfermedades de EE.UU. necesitaban entre una o dos semanas para recoger la información necesaria.

Por diversas razones el algoritmo de google flue no funciono y de hecho google dejo de publicar sus datos. Las razones de las fallas fueron descritas por un grupo de expertos, en una publicación

[34] https://www.nature.com/articles/nature07634

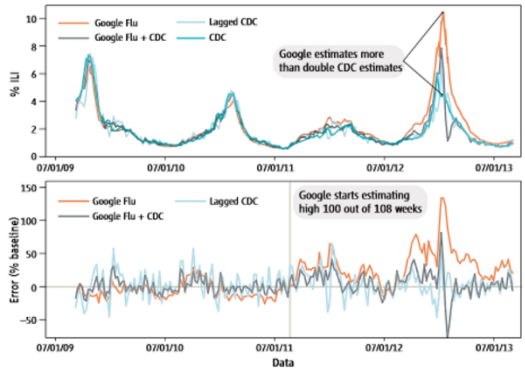
en "Science[35]", donde entre otras destacaron las dificultades que tuvieron para poder analizar el desempeño teniendo en cuenta que google nunca libero por completo la información sobre el algoritmo. Este hecho, derivó en una discusión sobre la necesidad de que las empresas privadas hagan públicos los algoritmos con que generan los datos para que puedan ser revisados por otros científicos.

Las tendencias de búsqueda con algoritmos diferentes han sido utilizados por otros investigadores (incluyendo en Argentina) para analizar la incidencia de enfermedades tipo influenza (ETI).

Orellano et al. (2015) generaron un modelo para estimar los casos de enfermedades tipo influenza (ETI), a partir de los términos de búsquedas[36] en Internet provistos por google trends, y validaron los resultados comparándolos con los casos de ETI informados por el Sistema Nacional de Vigilancia de la Salud de Argentina.

Google flu

- El algoritmo en base a millones de búsquedas relacionadas con la gripe —"síntomas gripe", "virus gripe", etcétera— podía decir al instante si habría colas en las urgencias en determinado punto del país. Mientras, los sistemas predictivos de los Centros para el Control y la Prevención de Enfermedades de EE.UU. (CDC) , necesitaban entre una o dos semanas para recoger la información necesaria.



Uso de la herramienta Google Trends para estimar la incidencia de enfermedades tipo influenza en Argentina
(Orellano, Pablo Wenceslao; Reynoso, Julieta Itatí; Antman, Julián; Argibay, Osvaldo, 2015)

Ventas de supermercados minoristas (ODS 8)

Uno de los elementos más importantes para estudiar el comportamiento del ciclo económico es el flujo de las ventas minoristas. *Camusso et al. (2019)* analizaron dicha variable por medio de un proxy, la serie sobre las ventas nominales de supermercados de la provincia de Santa Fe publicada por el INDEC. El indicador refleja las ventas correspondientes a 68 bocas de expendio de grandes superficies localizadas en el territorio sub-nacional.

El flujo de la serie de ventas de supermercados se contrastó con el flujo de búsquedas de Google Correlate, obteniendo como resultado una lista de palabras cuyo movimiento presente un alto poder predictivo de la variable objetivo.

[35] https://science.sciencemag.org/content/343/6176/1203.full
[36] Las búsquedas de Internet se obtuvieron de la base de datos del google trends, usando 6 términos: gripe, fiebre, tos, dolor de garganta, paracetamol e ibuprofeno.

Del total de palabras encontradas, se opta por tomar en cuenta cinco con alto nivel de correlación y, al mismo tiempo, razonabilidad desde el punto de vista económico. A saber: “Carrefour”, “El Entrerriano”, “Falabella”, “microcomponente“ y “nuevo”.

Las palabras seleccionadas en Google Correlate se ingresaron a Google Trends dando como resultados las series temporales del índice de búsqueda para estos términos. Luego a través de la construcción de un modelo logran predecir de manera aceptable las ventas minoristas de supermercados en Santa Fe a partir de los índices de búsqueda de los términos clave en gogole trends.

Web scraping

Además de las redes sociales las páginas web son una fuente inagotable de datos que se está utilizando, entre otras, para obtener información en tiempo real sobre la dinámica de los precios o de los mercado laborales (Cárdenas et al., 2015). Para lograr esto, se usan técnicas de extracción automática de datos que se conocen como “web scrapping” o “raspado de datos”.

Los programas de web scraping están diseñados para navegar a través de múltiples páginas web, extraer datos relevantes de las páginas y guardar los datos de forma estructurada para que puedan ser utilizados en el futuro.

La idea central es localizar datos puntuales de los sitios web y almacenarlos para su posterior utilización. Es decir, de un sitio web, puede ser interesante obtener sólo algunos datos, y el resto de lo que se analiza puede ser descartado.

Medición de la inflación (ODS 2)

Los precios online pueden ser utilizados para construir índices a través del scraping de los precios de distintos productos de los retailers más importantes emulando el Índice de Precios al Consumidor (IPC) del respectivo país.

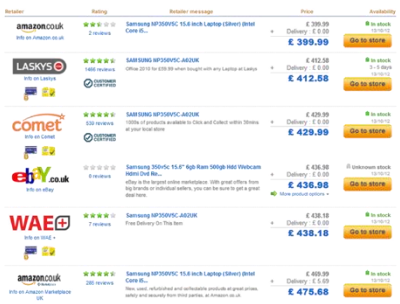
Cavallo et al. (2016) muestran cómo los precios online pueden ser utilizados para construir índices de precios con frecuencia diaria en múltiples países, evitando sesgos que distorsionan la evidencia sobre la rigidez de los precios y sus relatividades internacionales. Para dicho propósito, realizan un scraping de los precios de distintos productos de los retailers más importantes de cada país, emulando el IPC del respectivo país. Así encuentran que sus IPC online resultan ser muy parecidos a los desarrollados por las oficinas de estadísticas nacionales mejorando además su frecuencia y transparencia.

Vale remarcar que la metodología de precios online también tiene una serie de inconvenientes. Por ejemplo, dado que esta forma de obtención de datos se basa en el comercio online, su aplicabilidad se ve limitada en países menos desarrollados, que no presentan demasiadas plataformas de este estilo. Al mismo tiempo, la canasta obtenida puede no ser, necesariamente, muy representativa del consumo del país en cuestión. Al no presentar precios de servicios, dado que sus apariciones en páginas online son poco frecuentes, la canasta obtenida se reduce principalmente a productos de consumo. Otro posible inconveniente puede presentarse al considerar que los precios online quizás difieran de sus pares offline. Adicionalmente, los precios online tienden a presentar un solo precio para todas las localidades de un país mientras que sus pares offline presentan, en general, una dispersión significativa regional.

2.c Adoptar medidas para asegurar el buen funcionamiento de los mercados de productos básicos alimentarios …

Cada mes el INDEC observa una cantidad de 320.000 precios en los puntos de recolección. El relevamiento se realiza directamente mediante visita o contacto del encuestador a cada establecimiento u hogar seleccionado previamente.

- ***Cavallo et al (2016)*** - construyeron índices de precios y estimaron la tasa de inflación a partir de extraer miles de precios online.



Acceso a los datos: limitantes y potenciales soluciones

Para poner el big data al servicio del desarrollo sostenible, lo primero es lograr acceder a estos datos. Pero esto no resulta simple por diversas razones, incluyendo que las empresas privadas muchas veces consideran estos datos estratégicos para su negocio y temen perder competitividad si los hacen públicos. Por otro lado, está el problema, en este caso también compartido con los datos del sector público, de proveer datos asegurando que no se viola el derecho a la privacidad de las personas. Si bien, existen metodologías para anonimizar los datos, no son infalibles y muchas veces es posible re-identificar[37] los usuarios a los que corresponden los mismos[38].

A nivel internacional, el Grupo de Desarrollo de Naciones Unidas (UNDG) ha elaborado una guía[39] con principios éticos y aspectos vinculados a proteger la privacidad en la utilización del big data para el logro de la agenda 2030.

En el plano doméstico numerosos países han desarrollado normativas para la protección de los datos personales.

En la Unión Europea, en el 2016 se adoptó el Reglamento General de Protección de Datos (RGPD) que ha tenido un impacto significativo en todas las compañías que, de una forma u otra, se encuentran implicadas en el tratamiento de datos personales. El reglamento (GDPR) incorpora,

[37] https://www.nature.com/articles/s41467-019-10933-3/

[38] Las personas físicas pueden ser asociadas a identificadores en línea facilitados por sus dispositivos, aplicaciones, herramientas y protocolos, como direcciones de los protocolos de internet, identificadores de sesión en forma de “cookies” u otros identificadores, como etiquetas de radiofrecuencia. Esto puede dejar huellas que, en particular, al ser combinadas con identificadores únicos y otros datos recibidos por los servidores, pueden ser usados para elaborar perfiles de las personas físicas e identificarlas.

[39] https://unsdg.un.org/sites/default/files/UNDG_BigData_final_web.pdf

entre otras, el rol de un delegado en materia de protección de datos en las empresas y la realización de estudios de impacto en materia de procesamiento de datos

En Argentina, los datos personales se encuentran regulados en la Ley 25.326 "*Ley de Protección de los Datos Personales*" adoptada en el 2000. En aquella época el alcance de la Internet y el potencial del big data no se comparaban ni cercanamente con el que tienen en el presente, con los cual hay muchas aspectos que deben ser actualizados. Al respecto se han preparado proyectos ley pero al presente no se ha logrado sancionar ninguno[40].

Acceso a los datos a través de concursos (Dathatones)

Una manera de acceder a datos para proyectos de big data es a través de concursos (Datathon) en donde el sector público o privado ponen a disposición datos anonimizados para que se busquen aplicaciones. Algunos ejemplos incluyen:

- "Datathon[41] para el Bien Social" organizado por Telefónica en colaboración con MIT, la Campus Party de Londres y el Open Data Institute en 2013.

- Orange ha organizado en dos ocasiones (2013 y 2015) un reto mundial para el análisis de datos de móviles con fines humanitarios llamado "Datos para el Desarrollo" donde ha dado acceso a datos agregados y anonimizados de Costa de Marfil y Senegal a decenas de grupos de investigación a nivel mundial. El resultado han sido numerosos proyectos donde se ha utilizado el big data para entender el transporte y las ciudades, mejorar las estadísticas oficiales, contribuir a la salud pública, entre otras.

- Telecom Italia también ha organizado un reto mundial de análisis de datos con más de 1.000 participantes de países diferentes que presentaron más de ideas innovadoras que utilizaban el big data para varios fines, incluyendo fines sociales.

Nuevas alianzas público- privadas para el big data

El acceso a nuevas fuentes de datos requiere de nuevos tipos de alianzas entre el sector público y privado que pueden implicar diferentes modelos de colaboración[42]:

- La transferencia de conjuntos de datos: los conjuntos de datos son enviados directamente por su dueño al usuario final. Los datos en bruto se identifican, se muestrean y se agregan para evitar posible re-identificación. Un ejemplo de este modelo, son los concursos (mencionados previamente) que organizan empresas como la operadora de telefonía móvil Orange en las cuales dispone CDRs anonimizados a disposición equipos de investigadores del mundo.

- Acceso remoto: en el modelo de acceso remoto, los dueños (ej,. operadoras telefóncias) proporcionan a los usuarios finales acceso pleno a sus datos, manteniendo un estricto

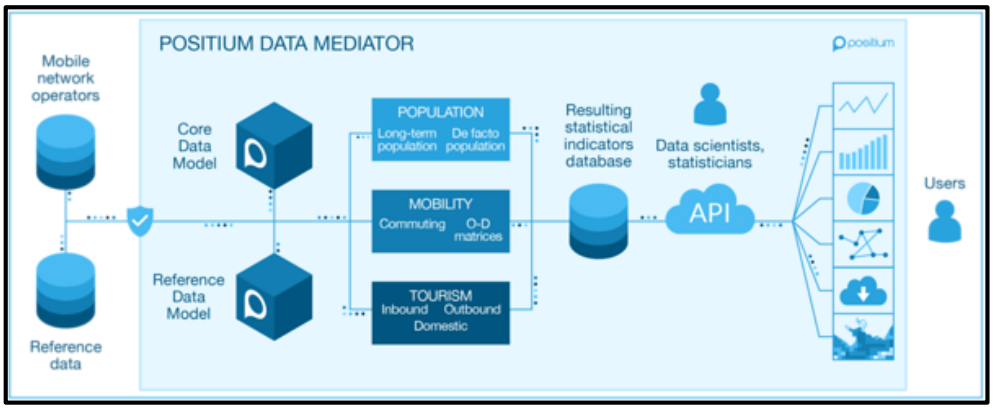
[40] El 2 de marzo de 2020 se presentó el Proyecto de Ley 0070-D-2020 para modificar la Ley de Protección de Datos Personales N° 25.326.
https://www.hcdn.gob.ar/proyectos/textoCompleto.jsp?exp=0070-D-2020&tipo=LEY

[41] http://news.o2.co.uk/?press-release=telefonica-the-open-data-institute-and-the-mit-set-da- ta-challenge-for-campus-party-2013

[42] https://nsdsguidelines.paris21.org/es/node/716

control sobre qué información puede ser extraída de sus bases y conjuntos. Los datos siempre se hospedan en la compañía de origen de los mismos.

- Transferencia de conjuntos de datos a un tercero de confianza: ni el dueño de los datos ni el usuario almacenan [hosting] los datos. Por ejemplo, en Estonia, a partir de una alianza privada-pública entre la empresa Positium y el Banco Central de Estonia, se establece una plataforma para procesado de datos que está parcialmente localizada en la compañía operadora de telefonía móvil, bajo su control, pero con funciones en las que también puede intervenir Positium, que es la compañia mediadora de los datos. El software de la plataforma permite prepara una base de datos con registros seleccionados al azar a partir de registros psudoanonimizados donde no es posible identificar al usuario del móvil. Todo este se prepara por fuera del sistema del operador, con lo cual sus datos están resguardados.



En Argentina, la plataforma en construcción, Palenque[43], sería un ejemplo de acceso a los datos bajo la modalidad de transferencia de datos a terceros. El proyecto surge en un contexto en el cual comienza a desarrollarse una nueva tendencia relacionada a la agricultura denominada "data drive agricultura", en la cual la tecnología de la información adquiere un rol central para la explotación agrícola. El concepto de data drive agriculture supone que las tecnologías vinculadas al sistema de geoposicionamiento satelital permiten obtener datos georreferenciales de distintos sitios de un lote para luego poder procesar esta información con sistemas especializados, lo que permite elaborar diferentes mapas o modelos con información precisa en sus diferentes áreas.

Palenque establece acuerdos con diferentes organismos del Estado, que generan datos estratégicos de valor agronómico, muchas veces a partir de disponer de equipamiento sofisticado y costoso, y que, de otra manera, o no estarían disponibles o su acceso resultaría demasiado complejo o costoso.

Las empresas AgTech, investigadores y demás interesados podrían acceder a los datos de una manera homogénea, centralizada y previsible.

En resumen, Palenque funcionaría como un concentrador de datos agrícola-ganaderos de valor estratégico, generados por diferentes organizaciones nacionales. A través de la plataforma se espera poder brindar soluciones tecnológicas basadas en grandes datos a los productores agropecuarios, así como al sector público y otros actores del sistema productivo y científico.

[43] http://palenque.org.ar/#/welcome

Generación de estadísticas globales desde un organismo internacional centralizado

Muchos de los datos digitales que se producen tienen un alcance supranacional o global. Esta característica que tienen algunas fuentes de big data, como las imágenes satelitales, genera oportunidades para reconsiderar los modelos de producción nacional de ciertas estadísticas, que normalmente están a cargo de las oficinas nacionales (como el INDEC).

Ir desde el modelo nacional actual hacia un modelo internacional colaborativo mejoraría la eficiencia en la generación de ciertas estadísticas como las de uso del suelo, marítimas y pesca (bajo las metas 14.2, 14.3, 15.1, 15.2, 15.3, 15.4). En lugar de replicar la producción de estadísticas en cada país a partir de imágenes satelitales que cubren todo el globo, se podría centralizar la producción en un único centro internacional, hecho que también generaría beneficios en términos de la consistencia y comparabilidad de estas estadísticas entre los diferentes países. Por otro lado, también es entendible que algunos países consideren que dejar la generación de estadísticas nacionales en manos de un tercero puede verse como una pérdida de soberanía que no están dispuestos a ceder.

Proveedores de datos en el sector público de la Argentina

Ya nos hemos referido previamente a las fuentes de datos en manos del sector privado como pueden ser las operadoras de telefonía móvil, las corporaciones que manejan plataformas online como Google y Twitter y las compañías que ofrecen productos satelitales.

El sector público también genera y almacena grandes cantidades de datos con un valor que aún no ha sido explotado. Aquí se mencionan tres fuentes de big data del sector público argentino: 1) los sistemas de gestión documental de la administración pública (GED y TAD); 2) Las historias clínicas electrónicas; y 3) Los datos de los usuarios de la tarjeta SUBE[44].

El sistema de Gestión Documental Electrónica (GDE y TAD)

Desde hace un par de años la Administración Pública Nacional cuenta con una plataforma, denominada Gestión Documental Electrónica (GDE), para facilitar la gestión de trámites y documentos. Su principal objetivo es habilitar la creación y procesamiento de documentos electrónicos, trazabilidad de todas las actuaciones, aportar robustez a las actuaciones por medio de la utilización de firma digital y la posibilidad de intercambio de información entre entidades, incluyendo la interacción directa con los ciudadanos. Esto último se hace a través de la Plataforma de Trámites a Distancia (TAD) la cual es el medio de interacción del ciudadano con el Sector Público, a través de la recepción y remisión por medios electrónicos de presentaciones, solicitudes, escritos, notificaciones y comunicaciones, entre otros

Los datos registrados a través del sistema TAD (Trámites a Distancia) y del GDE (Gestión Documental Electrónica) tienen un enorme potencial[45] para entender los procesos del estado, identificando los cuellos de botella y posibilitando de esta manera lograr una gestión más eficiente. Por supuesto que para ello hay que superar el inconveniente no menor que significa identificar y proteger la información sensible incluyendo los datos personales, previo a disponer los datos de estos sistemas para cualquier tipo de análisis externo.

[44] El sector público también cuenta con un monopolio en la recaudación de los impuestos, donde existen oportunidades para el análisis del big data.
[45] La basa de datos cuenta con 222 millones de documentos al 10/07/20

El big data sanitario[46] y el potencial de la historia clínica electrónica

El sector de la salud genera una enorme cantidad de datos que bien gestionados servirían como una herramienta clave en la prescripción de tratamientos, en la medición de la evolución de enfermedades, en la predicción de epidemias, en la detección precoz de patologías y en la decisión del profesional médico, entre otras.

La información que se produce en el sector incluye datos estructurados[47] (como las fichas que registran el nombre, edad, sexo y otros atributos de los pacientes) y datos no estructurados (como son las recetas de papel, las notas manuscritas de médicos y enfermeras, las grabaciones de voz, las radiografías, escáneres, resonancias magnéticas y otras imágenes médicas).

En algunos países, como España, se ha avanzado pon programas para poner a disposición de la investigación este tipo de datos. Al respecto, el Programa Público de Análisis de Datos para la Investigación y la Innovación en Salud (PADRIS[48], de sus siglas en catalán) se propone generar estructuras compartidas de datos sanitarios generados por el Sistema Sanitario Integral de utilización pública de Cataluña (SISCAT). Los datos provienen de radiografías, informes, recetas de farmacias, hospitales o centros de atención primaria. Se prevé que se incluyan datos genéticos. Los centros que pueden hacer uso de estos datos, que serán previamente anonimizados, serán centros de investigación que forman parte de la red de Centres de Recerca de Catalunya (CERCA), los agentes SISCAT y los centros de investigación universitarios públicos, así como la Administración sanitara. El manejo de los datos se hará de acuerdo con el marco legal y normativo, los principios éticos y de transparencia del programa hacia la ciudadanía, con el fin último de impulsar la investigación, la innovación y la evaluación en salud.

En Argentina muchos hospitales y clínicas registran las historias clínicas de sus pacientes en un soporte electrónico que queda almacenado en bases de datos. En CABA, por ejemplo, existe un Sistema Integrador de Historias Clínicas Electrónicas que se propone informatizar y unificar todos los datos médicos de pacientes que se atiendan en la Ciudad, tanto en sanatorios y clínicas privadas como en hospitales públicos.

El análisis de las bases de datos de historias clínicas electrónicas en Argentina para investigaciones con fines públicos tiene un enorme potencial para generar beneficios en el sector de la salud. No obstante, al igual que con los datos los datos del GDE o incluso aún en mayor medida por la sensibilidad de la información que incluye una historia clínica, resulta fundamental establecer un marco legal y normativo, así como principios éticos y de transparencia que regulen las investigaciones en base a estos datos. Para dar un ejemplo, una historia clínica de un paciente con una enfermedad crónica en manos de un potencial empleador podría limitar seriamente sus posibilidades de acceder al empleo.

Tarjeta SUBE y diseño de transporte

La tarjeta SUBE en Argentina registra datos de los movimientos de los ciudadanos en todos los modos de transporte público (colectivo, trenes y subtes) del Área Metropolitana de Buenos Aires

[46] https://www.red.es/redes/es/actualidad/sala-prensa/recursos-multimedia/imagenes/estudio-%E2%80%9Cbig-data-en-salud-digital%E2%80%9D-en

[47] Recordemos que un dato estructurado es un dato que puede ser almacenado, consultado, analizado y manipulado por máquinas, normalmente, en modo tabla de datos

[48] Más información de este programa en https://www.institutoroche.es/static/pdfs/MPP_EN_ESPANA_MAPA_DE_CCAA.pdf

(AMBA). Cabe destacar que SUBE procesa un total de 16 millones de transacciones por día y alcanza a 7 millones de usuarios.

El análisis de los datos de la tarjeta SUBE tiene un enorme potencial para entender en profundidad el comportamiento de los usuarios y así mejorar las políticas de movilidad.

Ventajas y desafíos del big data

A lo largo del documento se fueron mencionando ventajas y desafíos para el uso del big data en comparación con las fuentes de información tradicional. Aquí se resumen algunas de ellas.

Ventajas

- Costo-efectividad: permitan obtener información valioso a un costo mucho menor que por las metodologías de recolección de datos tradicionales.

- Mayor frecuencia temporal: la velocidad de estas nuevas herramientas y fuentes de datos permiten acceder a datos "frescos" que no solo ayudan a planificar mejor, y no con información de 2 o 3 años atrás, sino que permiten un monitoreo que lleva a corregir el rumbo de políticas, programas y proyectos cuando sea necesario y sin esperar meses o inclusive años. Ya que los metadatos móviles no procesados se encuentran disponibles casi instantáneamente, los Registros-en-Detalle de Llamadas (CDRs) de los operadores de telefonía móvil, por ejemplo, pueden producir estadísticas casi en tiempo real.

- Granularidad: Los datos del sector privado, los CDRs y datos geoespaciales en particular, ofrecen un gran nivel de detalle temporal, espacial, temático e individual. Además permiten obtener información sobre minorías y grupos vulnerables que suelen estar infrarrepresentados en los datos colectados por metodologías tradicionales. Finalmente, facilitan la producción de estadísticas desglosadas a nivel regional y sub-regional.

Desafíos

- Dificultades para acceder a los datos: muchas veces el acceso a estos datos se ve restringido porque las empresas que los poseen están obligadas a preservar la identidad de sus clientes o porque consideran esos datos como estratégicos para el desarrollo de su negocio y ponerlos a disposición de otros puede perjudicar su desempeño frente a sus competidores. Hay también preocupaciones de que los gobiernos podrían utilizar los datos para propósitos regulatorios o que la difusión de datos sobre los clientes de una organización podrían dañar su imagen pública.

- Dificultades para lograr series temporales consistentes: para su uso en estadísticas oficiales se requiere contar con un acceso metódico y permanente asegurando que los datos provistos respondan a los estándares de estructura y formato requeridos para su utilización. Aún cuando los datos estén disponibles, cambios en las tecnologías y algoritmos que las empresas usan para recolectarlos atentan contra la posibilidad de generar series estadísticas que sean consistentes a lo largo del tiempo. Adicionalmente, está la problemática de poder acceder a los algoritmos con los que se generan los datos,

que es fundamental para poder asegurar su consistencia estadística, pero que nuevamente muchas empresas consideran información estratégica que no están dispuestas a brindar.

- Representatividad de los datos: muchas veces se considera erróneamente a la superabundancia de los datos como sinónimo de representatividad, pero esto no es necesariamente cierto. Por ejemplo, los datos que se recopilan por canales digitales sólo son representativos de ciertos usuarios más activos y, en el mejor de los casos, sólo de aquellos que tienen acceso a tecnologías de información y comunicación. Todo esto genera dudas sobre la capacidad de realizar inferencias generalizables si los datos no representan adecuadamente la diversidad de la población bajo estudio. Otra cuestión a considerar es cuan real es la información de proviene de ciertos canales como las redes sociales. Al respecto, un alto porcentaje de los usuarios de estas redes no corresponde a personas reales y en muchos casos son administradas por robots.

Conclusiones y pasos a futuro

Los datos son esenciales para tomar decisiones y la materia prima para exigir responsabilidades. Sin datos, no podemos conocer el nivel de pobreza en una sociedad o cuántas mujeres han muerto víctimas de la violencia machista. A pesar de ello, en el presente muchas sectores y temáticas carecen de estadísticas fiables. Por otro lado, una gran parte de los indicadores de desarrollo sostenible existentes provienen de trabajosas encuestas domésticas, que insumen un tiempo considerable, con lo que a menudo las políticas públicas se basan en datos desactualizados.

En este contexto, las nuevas fuentes de información del big data representan un desafío y una oportunidad extremadamente interesante.

Con respecto a la hipótesis inicial de este trabajo, es decir, que el "*el big data puede contribuir al monitoreo de los ODS en Argentina*", se concluye que en muchos sectores efectivamente puede hacer contribuciones al monitoreo, sobre todo con el objeto de adaptar y ajustar políticas en marcha en base a información granular y actualizada.

Por otro lado, la aplicación del big data para la generación sistemática de estadísticas oficiales[49] aún requiere superar diversas barreras como las dificultades para asegurar la representatividad de las muestras y la provisión de los datos en los formatos requeridos de manera consistente a lo largo del tiempo.

Con respecto a las preguntas planteadas

1- ¿Cuáles son las metas en las que el big data puede hacer aportes más significativos?

Existen diversas publicaciones que han abordado los potenciales aportes del big data al logro de los ODS *(LIRNEasia, 2017; MacFeely, 2019).* En este trabajo se han descrito ejemplos del uso del big data vinculados con las temáticas de salud, pobreza, trabajo, economía, cambio climático,

[49] Salvo contados casos, como el de la utilización de imágenes satelitales para el monitoreo de los bosques,

recursos acuáticos, bosques, ciudades sostenibles, entre otros, pero vale aclarar que es un campo de investigación muy dinámico en el que permanentemente surgen nuevas aplicaciones.

¿Cuáles son las técnicas/algoritmos que se están utilizando para la preparación y utilización de los datos?

A lo largo de esta publicación se describieron metodologías de trabajo para el uso del big data en proyectos que se basan en datos de telefonía móvil, imágenes satelitales e internet (tuits, motores de búsqueda y páginas web). También se mencionaron someramente los principales algoritmos, en el campo del aprendizaje automático, utilizados para transformar los datos en conocimiento.

¿Cuáles son las ventajas y desafíos de estos nuevos datos en comparación con los obtenidos a través de las metodologías tradicionales?

Entre las ventajas, se destaca la posibilidad de contar con información actualizada y granular en un tiempo menor y aun costo inferior que por las metodologías de recolección de información tradicionales. Las principales desventajas, se vinculan con las dificultades para acceder a los datos a la par de asegurar el derecho a la privacidad de las personas y lograr la consistencia metodológica necesaria para generar estadísticos oficiales a partir de estas fuentes de datos.

Para finalizar, la Argentina tiene la ventaja de contar con un sector de software y servicios informáticos que ha logrado posicionarse entre los más dinámicos de la región y además cuenta con una sociedad altamente conectada, incluyendo millones de usuarios con acceso a la web y a la telefonía móvil.

La materia prima, los datos y los recursos humanos están, resta una política pública que los articule de manera inteligente para lograr las metas de desarrollo sostenible. Esta política debería considerar fundamentalmente el establecimiento de canales apropiados para el acceso a los datos garantizando al mismo tiempo la privacidad de las personas.

Bibliografía

La Agenda 2030 y los ODS

- *Castillo Marin (2019).* 17 Objetivos para un mundo mejor: una guía para entender los ODS (191 páginas). Nazareno Castillo Marin. Amazon, 2019.

- Dodds et al (2017). Negotiating the Sustainable Development Goals: A Transformational Agenda for an Insecure World. Dodds F., Donoghue, D., & Roesch, J. London and New York: Routledge, 2017.

- *United Nations (2020).* The-Sustainable-Development-Goals-Report-2020. Disponible en: https://unstats.un.org/sdgs/report/2020/The-Sustainable-Development-Goals-Report-2020.pdf

- *Naciones unidas. Resolución 68/261*. Marco de indicadores mundiales para los Objetivos de Desarrollo Sostenible y metas de la Agenda 2030 para el Desarrollo Sostenible. Comisión de Estadística en relación con la Agenda 2030 para el Desarrollo Sostenible. A/RES/71/313.

Los ODS en Argentina

- *CNCPS (2020)* Segundo Informe Voluntario Nacional de la Argentina 2020. Primera ed. – Ciudad Autónoma de Buenos Aires. Consejo Nacional de Coordinación de Políticas Sociales.

- *CNCPS (2019)* Metadata de los indicadores de seguimiento de los ODS. Agenda 2030 Argentina. Segunda Versión - consolidada en septiembre de 2019.

Big data: conceptos generales

- *Sosa (2019).* Big data. Breve manual para conocer la ciencia de datos que ya invadió nuestra vida. Walter Sosa Escudero. Editorial Siglo XXI. Cuarta edición, 2019.

- *Lerena (2019)* .Métodos y aplicaciones de la ciencia de datos para las políticas de cti: redes sociales, minería de textos y clustering .Octavio Lerena. - 1a ed . - Ciudad Autónoma de Buenos Aires : ciecti, 2019. Libro digital, pdf.

- *World Bank (2019).* Information and Communications for Development 2018: Data-Driven Development. Information and Communications for Development. Washington, DC: World Bank. doi:10.1596/978-1-4648-1325-2. License: Creative Commons Attribution CC BY 3.0 IGO.

- *BID (2017).* El uso de datos masivos y sus técnicas analíticas para el diseño e implementación de políticas públicas en Latinoamérica y el Caribe / Patricio Rodríguez, Norma Palomino, Javier Mondaca.

- *Big Data:* Avances Recientes a Nivel Internacional y Perspectivas para el Desarrollo Local

Autores: Facundo Malvicinoa y Gabriel Yoguelb. Centro Interdisciplinario de Estudios en Ciencia Tecnología e Innovación Ministerio de Ciencia, Tecnología e Innovación Productiva. Ciudad Autónoma de Buenos Aires, Agosto de 2015.

- *Gonzalez (2018).* El reto Big Data para la estadística pública. Jesús Alberto González Yanes. Trabajo final del Master en Ingeniería de Sistemas de Decisión. Universidad Rey Juan Carlos. Madrid, 2017-2018
- *Data-Pop Alliance (2016).* Oportunidades y requerimientos para aprovechar el uso de Big Data para las estadísticas oficiales y los Objetivos de Desarrollo Sostenible en América Latina". Data-Pop Alliance (Harvard Humanitarian Initiative, MIT Media Lab y Overseas Development Institute). Mayo de 2016.

- *Asociación por los Derechos Civiles (2015).* Exploring State Practices and Uses of Big Data Technology. Tatiana Anabel Fij. Asociación por los Derechos Civiles. Octubre de 2015.

- *LUCA (2017).* Data as a Force for Good. Dr. V. Richard Benjamins, Dr. Pedro de Alarcon, Javier Carro and Florence Broderick. LUCA- Telefónica Digital España (2017).

- *Malhotra et al. (2018).* Applying Big Data Analytics in Governance to Achieve Sustainable Development Goals (SDGs) in India. Charru Malhotra, Rashmi Anand, Shauryavir Singh (2018)

- *Instituto Universitario de Investigación Ortega y Gasset (2017).* Manual sobre utilidades del big data para bienes públicos. Escuela de Política y Alto Gobierno.

- *MacFeely (2019).* The Big (data) Bang: Opportunities and Challenges for Compiling SDG Indicators. Steve MacFeely. Global Policy Volume 10 . Supplement 1 . January 2019.

- *ElMassah et al. (2018).* Big Data and Localizing the Sustainable Development Goals (SDGs). Suzanna ElMassah y Mahmoud Mohieldin. Cairo University and World Bank Group. Preliminary Draft October 2018.

- *Lokanathan et al. (2016)* Mapping Big Data Solutions for the Sustainable Development Goals. Sriganesh Lokanathan, Thavisha Perera-Gomez, Shazna Zuhyle. LIRNEasia. March 2017. Etzion, D. & Aragon-Correa, J.A., 2016. Big data, management, and sustainability: Strategic opportunities ahead, Organization & Environment, 29 (2), 3-10.

- *Bauer (2018).* Smart Planet Governance. Analyzing the role of big data for monitoring the Sustainable Development Goals. Tim Bauer. Aster Thesis Series in Environmental Studies and Sustainability Science. Lund University Centre for Sustainability Studies. Submitted May 15, 2018.

Aplicación del big data a los ODS

- *MacFeely (2019).* The Big (data) Bang: Opportunities and Challenges for Compiling SDG Indicators. Steve MacFeely. Global Policy Volume 10 . Supplement 1 . January 2019.

- *LIRNEasia (2017).* Mapping Big Data Solutions for the Sustainable Development Goals. LIRNEasia. Sriganesh Lokanathan, Thavisha Perera-Gomez, Shazna Zuhyle. March, 2017.

Telefonía móvil

- *Wesolowski, et al. (2015).* Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile phone data. Proceedings of the National Academy of Sciences, 112(35), 11114-11119.

- *Tatem et al. (2009).* The use of mobile phone data for the estimation of the travel patterns and imported Plasmodium falciparum rates among Zanzibar residents. Malaria journal, 8(1), 1.

- *Ruktanonchai et al. (2016).* Identifying malaria transmission foci for elimination using human mobility data. PLoS Comput Biol, 12(4), e1004846.

- *Global pulse.* Mobile phone network data for development. Global Pulse. Disponible en www.unglobalpulse.org

- *Lu et al. (2012).* Predictability of population displacement after the 2010 Haiti earthquake. Proceedings of the National Academy of Sciences, 109(29), 11576-11581.

- *Calabrese et al. (2011).* Estimating Origin-Destination flows using opportunistically collected mobile phone location data from one million users in Boston Metropolitan Area. IEEE Pervasive Computing, 99.

- *Samarajiva et al. (2015).* Big Data to Improve Urban Planning. Economic & Political Weekly, 50(22), 43.

- *Toole et al. (2014).* The path most travelled: mining road usage patterns from massive call data. arXiv preprint arXiv:1403.0636.

- *Rojas et al. (2016).* Comprehensive Review of Travel Behavior and Mobility Pattern Studies. That Used Mobile Phone Data. Mario B. Rojas, Eazaz Sadeghvaziri, and Xia Jin. Transportation Research Record: Journal of the Transportation Research Board, No. 2563, Transportation Research Board, Washington, D.C., 2016, pp. 71–79.

- *Wanga et al. (2019).* Relationships between mobile phone usage and activity-travel behavior: A review of the literature and an example. Yihong Wanga, Gonc¸alo Homem de Almeida Correiaa, Bart van Arema. Advances in Transport Policy and Planning, Volume 3. Chapter 4. 2019 Elsevier Inc.

- Ashas et al. (2010). Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. Ahas, Rein , Silm, Siiri , Järv, Olle , Saluveer, Erki andTiru, Margus (2010). Journal of Urban Technology, 17: 1, 3 — 27.

- *Tatem et al (2014).* Integrating rapid risk mapping and mobile phone call record data for strategic malaria elimination planning. Malaria Journal 2014, 13:52.

- *Wesolowski et al. (2015).* Impact of human mobility on the emergence of dengue epidemics in Pakistan. Amy Wesolowski et al. PNAS | September 22, 2015 | vol. 112 | no. 38 | 11887–11892

- *Deville et al. (2014).* Dynamic population mapping using mobile phone data. Pierre Deville et al. 15888–15893 | PNAS | November 11, 2014 | vol. 111 | no. 45.

- *LIRNEasia (2017).* Leveraging Mobile Network Big Data for Developmental Policy. Final Technical Report.

- *Blumenstock et al. (2015).* Predicting poverty and wealth from mobile phone metadata. Joshua Blumenstock, Gabriel Cadamuro, Robert On. Science 27 November 2015. Vol 350 Issue 6264.

- *Steele et al. (2017).* Mapping poverty using mobile phone and satellite data. J. R. Soc. Interface 14: 20160690.

- *Shengjie et al. (2019).* Exploring the use of mobile phone data for national migration statistics. Shengjie Lai et al. Palgrave Communicatios (2019) 5:34

- *Jingtao Ma et al. (2013).* Deriving Operational Origin-Destination Matrices From Large Scale Mobile Phone Data. Jingtao Ma et al. International Journal of Transportation Science and Technology · vol. 2 · no. 3 · 2013 – pages 183 – 204.

- *Chen et al (2016).* The promises of big data and small data for travel behavior (aka human mobility) analysis. Cynthia Chen et al. Transportation Research Part C 68 (2016) 285–299.

- *Pastor-Escuredo et al (2014).* Flooding through the Lens of Mobile Phone Activity. David Pastor-Escuredo et al. IEEE 2014 Global Humanitarian Technology Conference.

- *BID (2019).* Cómo aplicar big data en la planificación del transporte urbano: el uso de datos de telefonía móvil en el análisis de la movilidad. Nota técnica Nº IDB-TN-1773. BID, 2019.

- BID (2020) Nueva generación de modelos de transporte a través del uso de big data: Caso San Salvador. Rendón, José Rodrigo; Hernández, Enrique; Del Río, Hernán.

- *Naciones Unidas (2019).* Handbook on the Use of Mobile Phone Data for Official Statistics. UN Global Working Group on Big Data for Official Statistics (Draft). September ,2019.

- Thomas et al. (2009) The association between socioeconomic status and exposure to mobile telecommunication networks in children and adolescents. Silke Thomas, Sabine Heinrich, Anja Kühnlein, Katja Radon. Bio Electro Magnetics - Volume31, Issue1. January 2010. Pages 20-27

- *Blumenstock et al (2010).* Mobile divides: gender, socioeconomic status, and mobile phone use in Rwanda. In Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development (p. 6). ACM.

- *Frias-Martinez et al. (2012).* On the relation between socio-economic status and physical mobility. Information Technology for Development, 18(2), 91-106.

- *Gutierrez et al. (2013).* Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets. arXiv preprint arXiv:1309.4496.

- *Decuyper, et al. (2014).* Estimating food consumption and poverty indices with mobile phone data. arXiv preprint arXiv:1412.2595.

- *Amy et al (2012).* Quantifying the impact of human mobility on malaria," Amy Wesolowski, Nathan Eagle, Andrew J. Tatem, David L. Smith, Abdisalan M. Noor, Robert W. Snow, Caroline O. Buckee, Science, October 12, 2012

Imágenes satelitales

- *Naciones Unidas (2017).* Earth Observations for Official Statistics. Satellite Imagery and Geospatial Data Task Team report. 5th December 2017.

- *Xi et al. (2011).* Using luminosity data as a proxy for economic statistics Xi Chena and William D. Nordhausa. PNAS | May 24, 2011 | vol. 108 | no. 21 | 8589–8594.

- *Hardwick et al. (2016).* Satellite observations to support monitoring of greenhouse gas emissions. Stephen Hardwick y Heather Graven. Grantham Institute. Briefing paper No 16. March 2016.

- *Matsunaga et al. (2018).* M. T. and Maksyutov S. (eds.) (2018) A Guidebook on the Use of Satellite Greenhouse Gases Observation Data to Evaluate and Improve Greenhouse Gas Emission Inventories, Satellite Observation Center, National Institute for Environmental Studies, Japan, 129 pp.

- *Buzai et al. (2020).* Megaciudad Buenos Aires: Cartografía de su última expansión y conurbación mediante el procesamiento digital de imágenes satelitales nocturna. Gustavo D. Buzai, Eloy Montes Galbán. Revista Cartográfica 100 - ISSN (impresa) 0080-2085 - ISSN (en línea) 2663-3981 - enero-junio 2020: 215-238

- *Jean et al (2016).* Combining satellite imagery and machine learning to predict poverty. Neal Jean, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, Stefano Ermon. Science august 2016 vol. 353 issue 6301.

- *Perez et al. (2017).* Poverty Prediction with Public Landsat 7 Satellite Imagery and Machine Learning. Anthony Perez, Christopher Yeh, George Azzari, Marshall Burke, David Lobell, Stefano Ermon. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

- *Letu et al. (2010).* Letu, Husi , Hara, Masanao , Yagi, Hiroshi , Naoki, Kazuhiro , Tana, Gegen , Nishio, Fumihiko and Shuhei, Okada(2010) 'Estimating energy consumption from night-time DMPS/OLS imagery after correcting for saturation effects', International Journal of Remote Sensing, 31: 16, 4443 — 4458

- *Xi et al. (2011).* Using luminosity data as a proxy for economic statistics. Xi Chena and William D. Nordhausa. PNAS | May 24, 2011 | vol. 108 | no. 21 | 8589–8594

- *Xie et al. (2009).* Transfer Learning from Deep Features for Remote Sensing and Poverty Mapping. Michael Xie, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

- *Elvidge et al. (2009).* A global poverty map derived from satellite data. Christopher D. Elvidge, Paul C. Sutton, Tilottama Ghosh, Benjamin T. Tuttle, Kimberly E. Baugh, Budhendra Bhaduri, Edward Bright. Computers & Geosciences 35(2009)1652–1660.

- *Scanlona et al. (2018).* Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data. Bridget R. Scanlona, Zizhan Zhangb, Himanshu Savec, Alexander Y. Suna, Hannes Müller Schmiedd, Ludovicus P. H. van Beekf, David N. Wieseg, Yoshihide Wadaf,h, Di Longi, Robert C. Reedya, Laurent Longuevergnej, Petra Dölld,e, and Marc F. P. Bierkens. E1080–E1089 | PNAS | Published online January 22, 2018

- *Richey et al. (2015).* Quantifying Renewable Groundwater Stress with GRACE. Alexandra S. Richey, Brian F. Thomas, Min-Hui Lo, John T. Reager, James S. Famiglietti, Katalyn Voss, Sean Swenson, and Matthew Rodell. Water Resources Research 51 · June 2015

- *Anderson et al (2017).* Katherine Anderson, Barbara Ryan, William Sonntag, Argyro Kavvada & Lawrence Friedl (2017) Earth observation in service of the 2030 Agenda for Sustainable Development, Geo-spatial Information Science, 20:2, 77-96.

- *Banskota et al. (2014).* Asim Banskota, Nilam Kayastha, Michael J. Falkowski, Michael A. Wulder, Robert E. Froese & Joanne C. White (2014) Forest Monitoring Using Landsat Time Series Data: A Review, Canadian Journal of Remote Sensing, 40:5, 362-384.

- *Kohli et al. (2012).* An ontology of slums for image-based classification. Divyani Kohli, Richard Sliuzas, Norman Kerle, Alfred Stein. Computers, Environment and Urban Systems 36 (2012) 154–163.

- *Dugoua et al. (2018).* Dugoua, Eugenie, Kennedy, Ryan and Urpelainen, Johannes (2018) Satellite data for the social sciences: measuring rural electrification with night-time lights. International Journal of Remote Sensing, 39 (9). pp. 2690-2701.

- *Bruederle et al. (2018).* Bruederle A, Hodler R (2018) Nighttime lights as a proxy for human development at the local level. PLoS ONE 13(9): e0202231. https://doi.org/10.1371/journal.pone.0202231

- *Ghosh et al. (2013).* Using Nighttime Satellite Imagery as a Proxy Measure of Human Well-Being. Tilottama Ghosh, Sharolyn J. Anderson, Christopher D. Elvidge and Paul C. Sutton. Sustainability 2013, 5, 4988-5019.

- *Engstrom et al. (2017).* Poverty from Space Using High-Resolution Satellite Imagery for Estimating Economic Well-Being. Ryan Engstrom, Jonathan Hersh and David Newhouse. Policy Research Working Paper 8284. World Bank Group - December 2017.

- *Bayle (2018).* Detección de villas y asentamientos informales en el partido de La Matanza mediante teledetección y sistemas de información geográfica. Tesis presentada para optar al título de Magister en Explotación de Datos y Descubrimiento del Conocimiento. Lic. Federico Bayle. UBA, 2018.

- *Marinone (2016)* Identificación de fuentes y sumideros de metano dentro del territorio nacional a partir de mediciones satelitales. Marinone Esteban. Universidad Nacional del Centro de la Provincia de Buenos Aires. Facultad de Ciencias Exactas. Trabajo final de Licenciatura en Tecnología Ambiental. Febrero de 2016.

- *IMF (2019).* Illuminating Economic Growth. WP/19/77. Yingyao Hu and Jiaxiong Yao (2019)

Internet

- *Ruiza et al. (2016).* Social Networks, Big Data and Transport Planning. Tomás Ruiza, Lidón Marsb, Rosa Arroyoa Ainhoa Sernac. XII Conference on Transport Engineering, CIT 2016, 7-9 June 2016, Valencia, Spain. Transportation Research Procedia 18 (2016) 446 – 452.

- *Enenkel et al. (2018).* Social Media Data Analysis and Feedback for Advanced Disaster Risk Management. Markus Enenkel, Sofía Martinez Sáenz, Denyse S. Dookie, Lisette Braman, Nick Obradovich, Yury Kryvasheyeu. Social Web in Emergency and Disaster Management February 9th 2018. Los Angeles, CA, USA.

- *Allaire (2016).* Disaster loss and social media: Can online information increase flood resilience?, Water Resour. Res., 52.

- *Efthymioua et al. (2012).* Use of social media for transport data collection. Dimitrios Efthymioua, Constantinos Antonioub. Procedia - Social and Behavioral Sciences 48 (2012) 775 – 785.

- *Kim et al. (2017).* Nowcasting commodity prices using social media. PeerJ Comput. Sci. 3:e126.

- *IMF (2018).* In search of information: use of Google trends' data to narrow information gaps for low-income developing countries. Futoshi Narita and Rujun Yin. IMF Working Paper. December 2018.

- *Global Pulse (2011).* Using social media and online conversations to add depth to unemployment statistics. Methodological White paper. UN Global Pulse. December 8, 2011.

- *Global Pulse (2014).* Mining Indonesian Tweets to Understand Food Price Crises. UN Global Pulse. Methods paper, February 2014

- *Cárdenas et al. (2015).* Cárdenas, R. J. A., Guataquí, R. J. C., & Montaña, D. J. M. (2015). Metodología para el análisis de demanda laboral mediante datos de internet: el caso colombiano. Revista de Economía del Rosario, 18(1), 93-126.

- Martínez et al. (2019). Análisis de técnicas de raspado de datos en al web – aplicado al portal del estado nacional argentino. Roxana Martínez, Rocío Rodríguez, Pablo Vera, Christian Parkinson. XXV Congreso Argentino de Ciencias de la Computación. Río Cuarto, 14 al 18 de Octubre de 2019.

- *Russell et al. (2019).* Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Instagram, GitHub, and More (Inglés) 3rd Edición. Matthew A. Russell, Mikhail Klassen. O'Reilly Media; 432 páginas. Edición: 3 (14 de enero de 2019).

- *Vázquez et al. (2020).* Grandes datos, google y desempleo. Raymundo M. Campos Vázquez, Sergio E. López-Araiza B. Estudios Económicos, vol. 35, núm. 1, enero-junio 2020, páginas 125-151

- *Pavlicek et al. (2015)* Nowcasting Unemployment Rates with Google Searches: Evidence from the Visegrad Group Countries. Jaroslav Pavlicek, Ladislav Kristoufek. PLoS One. 2015; 10(5): e0127084. Published online 2015 May 22.

- *Marinescu et al. (2016).* Opening the Black Box of the Matching Function: the Power of Words, WP núm. 22508, NBER, Cambridge.

- *Sudhakar et al. (2014).* The Use of Google Trends in Health Care Research: A Systematic Review. Sudhakar V. Nuti, Brian Wayda, Isuru Ranasinghe, Sisi Wang, Rachel P. Dreyer, Serene I. Chen, Karthik Murugiah. Plos One October 2014 | Volume 9 | Issue 10 | e109583

- *Orellano et al. (2015).* Uso de la herramienta Google Trends para estimar la incidencia de enfermedades tipo influenza en Argentina. Pablo Wenceslao Orellano, Julieta Itatí Reynoso, Julián Antman, Osvaldo Argibay. Questoes metodológicas Cad. Saúde Pública 31 (4) Abr 2015.

- *Camusso et al. (2019).* Google Correlate y Google Trends como Herramientas para Realizar un Nowcast de las Ventas Minoristas. Camusso, María Florencia y Jorge, Ramiro Emmanuel. Asociación Argentina de Economía Política. LIV Reunión anual. Noviembre de 2019.

- *Cárdenas et al. (2015).* Metodología para el análisis de demanda laboral mediante datos de internet: el caso colombiano. Cárdenas, R. J. A., Guataquí, R. J. C., & Montaña, D. J. M. Revista de Economía del Rosario, 18(1), 93-126.

- *Cavallo et al. (2016).* The Billion Prices Project: Using online prices for measurement and research. Cavallo, A., & Rigobon, R. (2016). The Journal of Economic Perspectives, 30(2), 151-178.

Otros

- *Bazzano et al. (2016).* Palenque. Plataforma de Grandes Datos para el Agro Arquitectura de Big Data. Agustina Bazzano, Lautaro Chiarle, Ernesto Mislej, Carlos Lizarralde, and Nicolas Higgs. AGRANDA 2016, 2º Simposio Argentino de Grandes Datos

- *Big Data en salud digital*. Fundación Vodafone España y Red.es. Recuperado a través de https://www.red.es/redes/es/actualidad/sala-prensa/recursos-multimedia/imagenes/estudio-%E2%80%9Cbig-data-en-salud-digital%E2%80%9D-en