SEDESOL 2018

Mónica Zamudio López

Título del proyecto: USO DE DATOS MASIVOS PARA LA EFICIENCIA DEL ESTADO Y LA INTEGRACIÓN REGIONAL

Clave: ATN/OC 15822-RG

Puesto: Científico de Datos Junior

Recolección y limpieza de información

Entregable número: 1

Acrónimo del proyecto:	Estimación de Ingreso
Nombre completo del proyecto:	USO DE DATOS MASIVOS PARA LA EFICIENCIA DEL ESTADO Y LA INTEGRACIÓN REGIONAL
Referencia:	ATN/OC 15822-RG
URL del Proyecto:	http://www.plataformapreventiva.gob.mx

Tipo de Entregable:	Reporte (R)
Fecha de Entrega Contractual:	Mayo - 2018
Fecha de Entrega	3 de Mayo de 2018
Número de Páginas:	14
Keywords:	estimación ingreso ciencia datos ingesta
Autor:	Mónica Zamudio López, Laboratorio de Datos, SEDESOL

Resumen

La Secretaría de Desarrollo Social (SEDESOL) es una entidad del gobierno mexicano destinada al apoyo de la población para el mejoramiento de sus condiciones de vida.

Un problema importante para SEDESOL es la correcta distribución de sus recursos por lo que es importante contar con una metodología que permita la generación de una focalización correcta para así poder ayudar a aquellos que realmente están en condiciones vulnerables.

En este reporte se detalla el proceso de obtención e integración de distintas fuentes de información que puedes ser de utilidad en dicho proceso, ya sea de manera directa o como auxiliar de las primeras. Las partes fundamentales para la realización de esto son, el estudio y evaluación de información generada por la misma Secretaría como lo es el Cuestionario Único de Información Socioeconómica (CUIS) y el Sistema de Focalización de Desarrollo (SIFODE), además de datos auxiliares que permitan la georreferenciación como archivos poligonales de carreteras y caminos. Para esto es importante llevar a cabo un sistema semiautomatizado para almacenar de manera adecuada todas las fuentes de datos para su futuro uso así como la información de los orígenes y naturaleza de las mismas.

Índice general

Índice de figuras

Índice de tablas

Lista de Acrónimos

CUIS Cuestionario Único de Información Socioeconómica

SIFODE Sistema de Focalización de Desarrollo

SEDESOL Secretaría de Desarrollo Social

ENIGH Encuesta Nacional de Ingresos y Gastos de los Hogares

PEA Población Económicamente Activa

LGDS Ley General de Desarrollo Social

INEGI Instituto Nacional de Estadística y Geografía

SISI Sistema de Información Social Integral

PUB Padrón Universal de Beneficiarios

AWS Amazon Web Services

1. Introducción

1.1. Sección Uno

1.2. Sección Dos

Fuentes de datos

Fuentes principales

Para una tarea como la de identificar la falsa representación del estatus socioeconómico de los usuarios -activos y potenciales- de los distintos programas, es primordial contar con toda la información socioeconómica reportada que sea posible, e integrarla con la información disponible sobre las estrategias de focalización de los programas. En SEDESOL, la mayor parte de esta información es integrada por la DGAE y DGAIP, y compartida con el Laboratorio de Datos a través de nuestros recursos de almacenamiento en AWS[1], salvo las excepciones que precisaremos conforme sea necesario.

CUAPS

El **Cuestionario Único de Aplicación a Programas** es una herramienta utilizada por SEDESOL para obtener información de diseño y focalización de los programas sociales. Para la integración de los datos, la DGAE envía anualmente un formato en Excel a los responsables de programas. A partir de 2017, se ha buscado migrar gradualmente a un llenado a través de la plataforma de SISI, por lo que la DGAE integra manualmente ambas fuentes de datos, y los envía al equipo del Laboratorio de Datos en formato *.xlsx* Los datos de CUAPS son entregados al laboratorio en tres tablas distintas:

- **Programas** - Esta tabla contiene los datos de diseño a nivel programa. Se especifica información administrativa del programa, sus objetivos generales y específicos de diseño, y los derechos sociales que el programa busca atender.
- **Componentes** - Para operar la política social, los programas definen apoyos que responden

[1] En particular, utilizamos el servicio S3 (Simple Storage Service) de Amazon Web Services. Esto es explicado con más detalle en el capítulo Capítulo 1.2

a los objetivos definidos en su diseño. Estos apoyos son la unidad más granular de suministro de beneficios utilizada por los programas, y pueden estar clasificados dentro de "bloques temáticos" llamados componentes.[2]

- **Focalización** - Esta tabla contiene información de la focalización de programas, a nivel criterio de focalización. Los programas pueden tener uno o más criterios de focalización para definir a su población objetivo, y estos criterios pueden estar lógicamente relacionados. A partir de la información de la tabla, es posible reconstruir esas matrices de diseño y delimitar la población objetivo de los apoyos otorgados por los programas.

Dadas las diferencias entre las unidades observacionales, los datos son ingresados a nuestro catálogo como tres tablas distintas.

PUB

El **Padrón Único de Beneficiarios** es el resultado de la consolidación de todos los padrones de los distintos programas sociales, conformado con el fin de monitorear los apoyos que reciben los beneficiarios. La actualización de los padrones es trimestral, y potencialmente corrige observaciones de trimestres anteriores al trimestre de actualización. En el PUB se tiene el nombre, sexo y grupo de edad del beneficiario, e información básica del programa del cual es beneficiario. Los datos del PUB son enviados por la DGAIP al equipo del laboratorio periódicamente, en formato .txt

CUIS-ENCASEH

El Cuestionario Único de Información Socioeconómica es el instrumento mínimo[3] utilizado por los programas para hacer levantamientos socioeconómicos, y es aplicado a nivel hogar. El CUIS puede aplicarse para procesos de identificación, recertificación, evaluación y verificación de los hogares, entre otros. Cada programa gestiona los datos para su operación, y la información que corresponde al CUIS se va integrando a lo largo del tiempo por la DGAE. La DGAE recibe la información del

[2]Como un ejemplo simple: un programa de Mejoramiento a la Vivienda puede tener componentes de pisos, muros y techos, y otorgar tres apoyos diferentes para la componente de muros: cemento, varillas y un subsidio al salario de trabajadores de la construcción.

[3]Los programas pueden hacer más preguntas de las que se incluyen en el CUIS, pero no menos.

CUIS de dos formas: en un flujo continuo, a través de aplicaciones móviles, y en *batch*, por parte de programas que cuentan con su propia infraestructura, como PROSPERA y LICONSA. Así, la dirección cuenta con datos históricos de los hogares que alguna vez han sido sujetos a levantamientos socioeconómicos, y a su vez integra periódicamente bloques de actualizaciones del CUIS. Los datos históricos son compartidos con el laboratorio a través de S3, en formato *.zip* , separando cada tabla en un archivo distinto, mientras que los bloques de actualizaciones forman parte de SIFODE. Las tablas que tenemos disponibles son:

- **domicilios**: contiene información no granular de la ubicación de la vivienda: estado, municipio y localidad para cada clave de hogar.
- **encuesta**: contiene información principalmente administrativa sobre la aplicación del cuestionario como el tipo de procedimiento, datos de la creación del registro en el sistema y el origen del levantamiento del cuestionario[4].
- **ids**: contiene columnas utilizadas para relacionar distintas tablas con información a nivel hogar y persona.
- **integrante**: contiene datos de cada persona como integrante del hogar, como su condición de residencia y su parentesco con el jefe del hogar.
- **persona**: contiene datos administrativos sobre cada integrante del hogar.
- **se_integrante**: contiene información socioeconómica a nivel persona, de cada integrante del hogar. Esto incluye temas de educación, discapacidad y ocupación, entre otros.
- **se_vivienda**: contiene información socioeconómica sobre el hogar en su conjunto, como las características de la vivienda, el acceso de sus integrantes a las instituciones de salud y los proyectos productivos que realizan los integrantes del hogar.
- **vivienda**: contiene información principalmente administrativa sobre la vivienda, como el tipo de vivienda y el total de personas en ella.

[4]El cuestionario puede ser levantado en aplicaciones móviles, cuestionarios en papel u otras modalidades.

Existen muchos programas que hacen preguntas adicionales al CUIS en sus levantamientos. Un caso de especial interés para este proyecto es el de PROSPERA, tanto por su tamaño[5] como por su dinámica: dentro de sus levantamientos, PROSPERA cuenta con un módulo adicional de verificaciones domiciliarias, que representa un insumo extremadamente importante para un proyecto como el nuestro. Más adelante se explica el contenido de este módulo con mayor detalle.

SIFODE

El Sistema de Focalización de Desarrollo es una herramienta diseñada para coordinar la operación de los programas sociales, administrada dentro de la DGAE. En esencia, su operación consiste en utilizar la información socioeconómica de los hogares con los que se cuenta para crear universos potenciales de personas beneficiarias de los programas sociales. Para esta tarea, se toman bloques temporales de datos del CUIS y del PUB, y se genera información sobre la situación de ingreso y carencias sociales enfrentada por los hogares encuestados. Estos elementos son los factores que componen la definición de pobreza multidimensional[6] del Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL):

- Ingreso corriente per cápita - el ingreso compone la detección de pobreza y vulnerabilidad a través de dos puntos de corte: la Línea de Bienestar Económico[7] y la Línea de Bienestar Mínimo.[8]
- Rezago educativo - El indicador de rezago educativo toma como elementos el acceso a la educación obligatoria correspondiente que tienen los distintos integrantes de los hogares. La persona se considera como no carente sólo si está en edad escolar y asiste a la escuela o si de acuerdo con su edad ha concluido la primaria o secundaria, considerando la legislación aplicable.
- Acceso a los servicios de salud - El indicador de acceso a los servicios de salud toma en consideración que las personas cuenten con adscripción o derecho a recibir servicios médicos en

[5] PROSPERA representa alrededor de la mitad de los beneficiarios registrados en el PUB.
[6] Ver **??** para más detalle
[7] Esta especifica el ingreso necesario para adquirir las canastas alimentaria y no alimentaria de bienes y servicios. Ver **??** para mayor detalle.
[8] Esta especifica el ingreso necesario para adquirir la canasta alimentaria.

alguna institución de salud, ya sea pública o privada. La persona se considera como no carente sólo si cuenta con adscripción o filiación directa o indirectamente a alguna institución de este tipo.

- Acceso a la seguridad social - El indicador de acceso a la seguridad social toma en cuenta el acceso de las personas a coberturas sociales mínimas que deben otorgarse a los trabajadores y a sus familias. La seguridad social puede definirse como *"el conjunto de mecanismos diseñados para garantizar los medios de subsistencia de los individuos y sus familias ante eventualidades como accidentes, enfermedades, la vejez o el embarazo."*[9]

- Calidad y espacios de la vivienda - El indicador de calidad y espacios en la vivienda toma en consideración que la vivienda cuente materiales que garanticen la estabilidad y firmeza de los pisos, techos y muros, así como que las personas que la habitan no vivan en condiciones de hacinamiento.

- Acceso a los servicios básicos en la vivienda - El indicador de acceso a los servicios básicos en la vivienda considera las condiciones sanitarias en las que viven las personas habitantes de la vivienda. En particular, se considera que dichas personas carecen de servicios básicos en la vivienda si la vivienda no cuenta con agua entubada dentro del terreno, drenaje y electricidad o si el combustible usado para cocinar es carbón, y la vivienda no cuenta con chimenea.

- Acceso a la alimentación - El indicador de acceso a la alimentación es construido a partir de la Escala Mexicana de Seguridad Alimentaria, y toma en consideración si las integrantes de la vivienda tuvieron alimentación basada en poca variedad, dejaron de tomar al menos una de las tres comidas del día, comieron menos de lo que piensan debían comer, se quedaron sin comida o sintieron hambre pero no comieron.

- Grado de cohesión social - El CONEVAL define tres espacios conceptuales para medir el desarrollo social de los mexicanos: el bienestar, los derechos sociales y el contexto territorial. El primero es medido a través de los cortes antes mencionados al ingreso, el segundo por los indicadores de acceso mencionados, y el último por el grado de cohesión social en el espacio del

[9] Ver **??**.

territorio. Las propuestas retomadas por CONEVAL con respecto a su conformación comprenden la desigualdad económica, la razón de ingreso entre la población en pobreza extrema y la no pobre y no vulnerable, la polarización social y las redes sociales. Sin embargo, no existe un consenso sobre los indicadores a utilizar, así que el grado de cohesión social constituye una dimensión teórica de la pobreza multidimensional pero no de su medición actualmente.

Una vez obtenidos esos indicadores de bienestar y carencias, esta información es integrada con los datos del PUB, para así generar universos de beneficiarios potenciales de programas y entonces optimizar la frecuencia de los levantamientos socieconómicos.

El SIFODE es integrado cada año por la DGAE, y compartido con el laboratorio a través de S3 en archivos *rar* separados por tablas. Si bien se cuenta con la información histórica de los cuestionarios aplicados a todos los hogares, el sifode toma únicamente la última ”fotografía” obtenida por el CUIS en cada hogar. Así, contamos tanto con tomas aisladas de CUIS e integradas a SIFODE, como con la información histórica de los hogares. Esto nos permite reconstruir la trayectoria de ingreso y carencias que han vivido los hogares a través del tiempo.

Módulo de verificaciones

Como se especificó anteriormente, PROSPERA es uno de los programas que tiene un cuestionario más completo que el CUIS. Uno de sus componentes más importantes es el módulo de verificaciones domiciliarias. Eso es especialmente relevante, porque no todos los cuestionarios son aplicados en el domicilio de las personas, y aún de serlo las personas pueden tener oportunidad de mentir en más de un aspecto de las características de su vivienda o su estilo de vida. La posibilidad de comparar entre datos reportados y datos verificados nos permite generar modelos considerando variables faltantes o latentes. Dado que PROSPERA es independiente de la DGGPB, los datos de sus cuestionarios y módulos de verificacion son compartidos con el laboratorio a través de un convenio de colaboración, en el que el laboratorio obtuvo los datos de verificaciones y cuestionarios anuales entre los años 2010 y 2017. Se tienen tablas separadas por tipo de visita (verificaciones u otro tipo de aplicación del cuestionario) y por año, en formato *DBF*.

Fuentes adicionales

Para complementar los modelos de ingreso, existen otras fuentes de datos que pueden explotar la relación espacial que guardan las distintas aplicaciones de cuestionarios, como lo son la red carretera de INEGI, los datos georeferenciados de INEGI y el PUB georeferenciado. Estas fuentes de datos generalmente provienen de servicios web que permiten descargar archivos comprimidos de tipo *ZIP, RAR, TAR*, etc., y contienen *shapefiles* con los polígonos que integran el mapa. Estos archivos son convertidos a formato *CSV* para su posterior ingesta.

El proceso de ingesta

Infraestructura

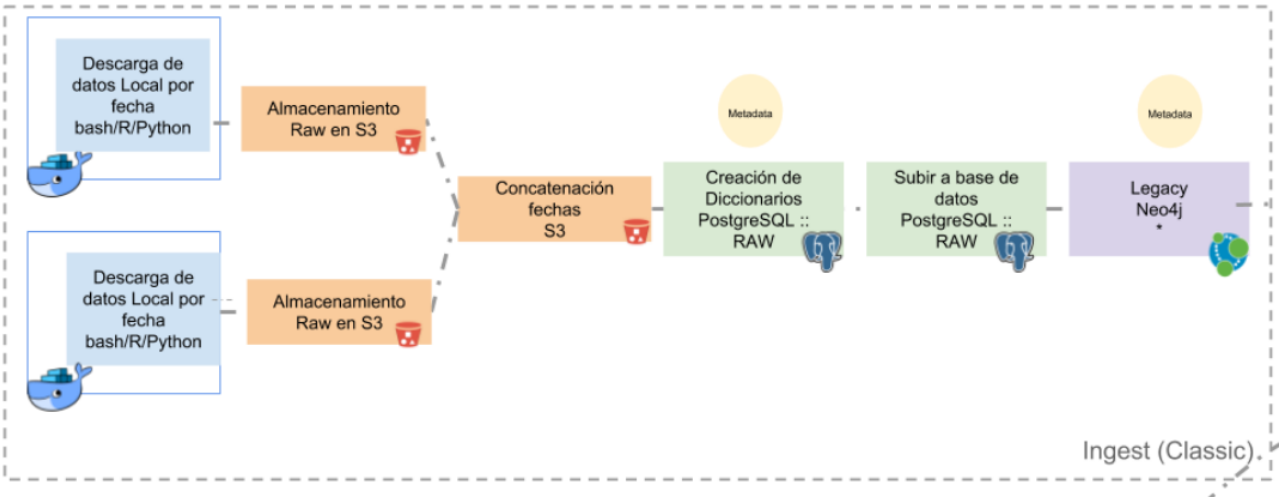
Como ya hemos explicado, la naturaleza del proyecto implica obtener, procesar, transformar e integrar múltiples fuentes de datos. Con eso en mente, hemos ensamblado una infraestructura orientada a fuentes de datos. Esa infraestructura responde a varias necesidades:

- **Escalabilidad** - las exigencias de memoria y procesamiento de nuestros flujos de extracción, transformación y carga (comúnmente llamados ETL, por sus siglas en inglés) pueden variar considerablemente con el tamaño de las bases de datos. Así, para datos como los padrones o cuestionarios largos sería óptimo correr 10 procesos en paralelo, mientras que para cuestionarios pequeños como CUAPS esa capacidad sería mal aprovechada.
- Portabilidad - es también importante que ningún proceso quede supeditado a ser ejecutado en una plataforma o sistema operativo en particular, para así disminuir las necesidades de mantenimiento del código.
- **Seguridad** - dado el tipo de información con el que contamos, debemos garantizar la seguridad y privacidad de los datos sin comprometer la flexibilidad de agregar nuevos colaboradores al proyecto de ser necesario.
- **Actualización periódica y eficiente** - al tener un flujo continuo de datos hacia nuestros modelos, necesitamos que esos procesos ETL puedan ser replicados sin mayor esfuerzo. Asimismo, queremos evitar que alguna tarea sea repetida innecesariamente.
- **Documentación** - para que pueda ser replicado con facilidad, es necesario que cualquier persona con acceso al código pueda entender el flujo de datos, la serie de procesos que deben ocurrir y las fuentes de datos necesarias para replicar el análisis.

Considerando estas necesidades, usamos varias partes móviles para el proceso de ETL. Todos nuestros procesos son ejecutados en un servidor de AWS EC2, que tiene montado un disco EBS para almacenamiento del código. En ese servidor se construye un contenedor de Docker a partir de una imagen privada con sistema operativo Linux Debian, que cuenta con Python, R y los paquetes necesarios para ejecutar el proyecto. La mayor parte de nuestras necesidades de almacenamiento de datos a través del flujo ETL se llevan a cabo en AWS S3, para al final escribirse a una base de datos AWS RDS en PostgreSQL. Por último, tenemos un registro del linaje de los datos en una base de datos orientada a grafos, en Neo4j.

Para facilitar la actualización periódica de los datos, todo el flujo está orquestado en Luigi con las tareas como sigue: Así, todas nuestras fuentes de datos son ingestadas concatenando el conjunto de

Figura 1.1: Flujo de datos



datos históricos antes de escribirse a la base de datos. Cada fuente de datos tiene su propio script de ingesta local, escrito ya sea en *bash, python* o *R* que toma como parámetros la fecha del dato, el nombre de la fuente y la ruta de descarga local al servidor.

Scripts de ingesta

Para ingestar los archivos de CUIS, es necesario utilizar los paquetes *unzip* y *unrar*??, que permiten la descompresión de archivos, y *csvkit* ??, un paquete de *python* con integración a *bash* que permite

aplicar distintas transformaciones a datos de tipo CSV, incluyendo la conversión a CSV desde archivos de Excel. Todos los códigos de ingesta están disponibles en Apéndice A.

2. Conclusiones

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Pellentesque rutrum nec dui nec dignissim. Pellentesque eget augue luctus nisl mollis malesuada. Vestibulum ultricies dapibus magna, vitae suscipit est condimentum sed. Vivamus sed sem ut mauris placerat aliquet nec quis ligula. Donec ac tortor erat. Duis ac augue sed urna laoreet finibus ac eu erat. Fusce placerat blandit arcu, ut mattis turpis luctus eu. Phasellus eleifend orci vulputate pharetra rhoncus. Nulla vitae interdum purus, vel posuere arcu. Donec at faucibus purus. Nam sodales tincidunt arcu ut vestibulum. Aliquam enim mi, tempor ac placerat in, gravida vitae augue. Donec lacinia accumsan justo non tincidunt. Curabitur porttitor quam in scelerisque tempus. Ut maximus elementum massa et egestas.

Ut interdum orci sed faucibus elementum. Maecenas vulputate metus non nulla finibus rutrum. Nullam eleifend ex id metus euismod, vel consectetur ex elementum. Sed efficitur libero non diam fermentum vehicula. Nunc dignissim suscipit libero, a feugiat nunc lacinia vitae. Sed at mollis nibh, mattis tincidunt dolor. Suspendisse maximus mi metus, vulputate hendrerit justo volutpat nec. Etiam vitae dui iaculis, pulvinar ante eu, congue leo. Fusce efficitur nunc massa, vel tincidunt velit accumsan in.

Vestibulum at nulla nunc. Sed sed lorem orci. Mauris ut porta diam. Quisque ac euismod quam. Phasellus tempor tempor tellus. Sed bibendum turpis at vehicula vulputate. Quisque condimentum efficitur lectus, vitae vehicula enim dapibus accumsan. Nullam bibendum, orci non scelerisque gravida, metus sapien pulvinar turpis, pulvinar egestas tellus sapien ut sem.

Nullam consequat elit quis ipsum imperdiet, ac placerat leo cursus. Proin cursus sit amet urna ut rutrum. Fusce congue elit non elit iaculis fermentum. Nullam quis ex nec dolor lobortis porta. Quisque enim lorem, pretium nec velit ut, eleifend laoreet magna. Vivamus eu mauris at purus sagittis condimentum. Nunc congue accumsan mauris, at convallis lacus pharetra in. Cras bibendum orci urna, nec elementum ex ornare sollicitudin. Praesent non ultricies velit. Curabitur volutpat malesuada enim, vel

maximus libero vestibulum vitae. Curabitur rhoncus erat nibh, quis viverra orci euismod nec. Donec magna lacus, tempor vitae justo sed, vulputate pretium massa. Aenean diam est, eleifend sit amet pellentesque vel, fermentum iaculis tellus. Etiam dignissim congue tellus, eget eleifend ipsum sagittis ut. Fusce convallis sapien dui.

Vivamus pretium mauris quis ligula vestibulum hendrerit. Ut maximus, erat vitae congue faucibus, eros nunc fermentum erat, id ultricies felis erat ac neque. Maecenas risus nunc, aliquam et leo quis, venenatis hendrerit nisl. Nunc viverra elementum ex, eu efficitur quam mollis nec. Mauris odio enim, fermentum non nibh eu, vestibulum venenatis felis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Cras ac velit varius, congue eros non, fringilla sapien. Fusce pretium porttitor urna a accumsan. Duis pharetra lorem vel mauris vehicula, non accumsan lectus mattis. Suspendisse posuere massa eget velit venenatis, eget gravida turpis vulputate. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Duis convallis accumsan urna, convallis facilisis ipsum feugiat id. Nunc gravida, magna eu euismod pharetra, nulla risus porta urna, a dapibus augue lorem blandit justo.

A. Apéndice

USO DE DATOS MASIVOS PARA LA EFICIENCIA DEL ESTADO Y LA INTEGRACIÓN REGIONAL

3 de mayo de 2018

FJR-1-ATN/OC 15822-RG

SEDESOL 2018