Informe técnico final

Anexo 3

Valoración de las fuentes de datos disponibles para la estimación de los saldos netos migratorios a nivel de entidad federativa

En este documento se describe las actividades realizadas en relación con las tareas de homogenizar las bases de datos de INEGI y las tecnologías que se utilizarán para desarrollo de un sistema de reportes y acceso a las mismas bases de datos.

Bases de datos de INEGI

Para este proyecto se utilizan bases de datos que son proporcionadas por el Instituto Nacional de Estadística y Geografía (INEGI) a través de su sitio web (inegi.org.mx), específicamente en la sección de estadística. Dependiendo de la base de datos o encuesta de interés, los datos se pueden obtener en su mayoría utilizando alguno de los siguientes 3 métodos de descarga: la consulta interactiva, tabulados y/o microdatos.

La consulta interactiva de datos consiste de formularios de captura Web que permiten seleccionar las variables que se desean extraer de una base de datos particular. Este método puede resultar conveniente cuando se conocen las variables de interés y no se desea descargar la base de datos completa. Esto puede reducir considerablemente el tamaño del archivo de descarga. Sin embargo, existen algunos casos donde el gran número de variables disponibles pudiera hacer más compleja la consulta. El formato en el que se obtienen los datos con este método puede variar por base de datos. En la mayoría de los casos se obtienen en archivos de Excel, y en algunos otros se obtienen archivos de Word o PDF.

Otra sección que comúnmente se puede encontrar en las bases de datos o encuestas de INEGI es la sección de tabulados. Estos presentan información con un formato e interés particular dentro de una hoja de Excel o en PDF. El formato especial que se les da a estos archivos complica su integración a otra base de datos automáticamente.

INEGI también pone a disposición de los usuarios los microdatos que son las bases de datos completas. Estas bases de datos contienen los datos que se obtienen de las encuestas sin un mayor procesamiento. En algunos casos, los archivos de microdatos pueden ser muy grandes. Por ejemplo, se pueden obtener 6.67 GB de los microdatos (sin comprimir) del Censo 2010. En la mayoría de los casos, los microdatos se proporcionan en formato DBF (se describen más adelante) con archivos adicionales que proporcionan los catálogos u otros archivos que nos ayudan a interpretar los datos. Para algunos casos de bases de datos recientes, adicionalmente, se proporcionan archivos de SPSS (.SAV) y STATA.

DBF (dBASE III)

INEGI utiliza dBase como formato estándar para proporcionar los microdatos de los censos, conteos y encuestas. Los archivos que siguen el formato de este estándar tienen la terminación (extensión) ".dbf". Estos archivos se componen de un encabezado con información sobre la estructura de los datos y los datos mismos. A cada campo de dato se le asigna un tipo de dato (carácter, número, fechas, memo). En el caso de los archivos disponibles en INEGI, los caracteres son guardados con el código Windows ANSI. Es importante mencionar que la descripción de los títulos de cada variable o columna y el significado de los valores en estas bases de datos debe ser consultada en archivos, llamados catálogos, que se proporcionan en un archivo separado. Este catálogo puede ser descargado en archivo Excel, Word o PDF, dependiendo de la base de datos.

Descarga de bases de datos disponibles

Se realizó la tarea de descargar las bases de datos disponibles en línea del área de estadística de INEGI, principalmente microdatos y tabulados. Se lograron identificar 249 bases de datos diferentes que son mencionadas en el sitio web de INEGI. De estas, en total se descargaron 148 bases de datos diferentes (algunas formadas de varias tablas y/o catálogos), quedando 101 como no disponibles, solo disponibles bajo solicitud o que el enlace no estaba funcionando en el momento de la consulta. El espacio de disco duro que ocupan estas bases de datos descargadas es de 7.2 Gigabytes aproximadamente. Entre ellas están 118 bases de datos en formato DBF, 53 en archivos de Excel y 11 en archivos de SPSS (.SAV).

Gestión de bases de datos

Como ya se ha mencionado, las bases de datos de INEGI pueden obtenerse en diferentes formatos: DBF, Excel, SAV, STATA, SAS, PDF, entre otros. Una de las tareas importantes para este proyecto es la integración de las bases de datos en un formato que permite el acceso a ellas de forma homogénea. Es por esto que se identificó la necesidad de utilizar un sistema de gestión de bases de datos (DBMS de sus siglas en inglés), en este caso seleccionó MySQL. MySQL es un sistema de gestión de bases de datos relacional de software libre. Cuenta con una extensa comunidad de participantes y desarrolladores que lo utilizan.

Integración de archivos DBF a MySQL

Como la mayoría de las bases de datos identificadas hasta el momento se proporcionan en archivos DBF, se desarrolló un conjunto de rutinas que facilitan su integración a MySQL. Estas rutinas fueron necesarias debido a que no se encontró un aplicación o software que fuera lo suficientemente flexible para reprocesar nombres de variables y valores antes de su integración a MySQL. Estas rutinas o scripts fueron programados en el lenguaje Python con el apoyo de la biblioteca yDBF que permite la lectura de archivos DBF. Estas rutinas se describen a continuación en orden alfabético:

- analyze.py: muestra un descripción general de un archivo DBF, enlistando las variables con su respectivo nombre y tipo de dato.
- compare.py: enlista los nombres de variables de dos archivos DBF en dos columnas, donde los nombres de las variables están ordenadas alfabéticamente.
- getcvs.py: extrae de un archivo DBF los nombres de las variables en una primera línea y los valores de cada registro en los siguientes renglones. Cada nombre de variable y valor están separados por comas.
- getloadcommand.py: genera código SQL que incluye el comando LOAD DATA INTO y los datos necesarios para integrar un archivo CVS en una tabla de MySQL dada.
- gettable.py: genera código SQL que sirve para construir una tabla nueva en MySQL, basado en los datos y tipos de datos que contiene un archivo DBF.

El camino que se siguió para utilizar estas rutinas inicia utilizando analyze.py para observar las variables y tipo de datos que contiene un archivo DBF. En el caso de ser una base de datos que comparte un formato o una estructura con otras bases de datos, se puede utilizar compare.py para confirmar que la

cantidad y nombre de variables son idénticos. Después, se utiliza gettable.py para obtener el código que nos ayuda a construir la tabla de MySQL que almacenará los datos. Es importante analizar el código resultante, ya que en muchos casos, las variables están declaradas como tipo texto, siendo que contienen únicamente valores numéricos. Es importante analizar los tipos de datos y estos valores para obtener el mejor desempeño de MySQL. El siguiente paso es crear el archivo CVS con la rutina getcvs.py, el cual contiene los valores de cada registro separados por comas. Para integrar el archivo CVS y MySQL, es necesario el comando LOAD DATA INTO de MySQL, el cual puede ser generado por la rutina getloadcommand.py.

Análisis de las bases de datos de censos y conteos

Después de descargar y analizar algunos archivos de microdatos en formato DBF, se notaron algunas características que son importantes conocer para integrar todas las fuentes en un solo repositorio.

En el caso de los censos y conteos hay algunas inconsistencias entre los años que se han levantado. Por ejemplo, en la muestra del Censo de Población y Vivienda 2010, la variable sexo puede tener uno de dos valores: 1 para identificar si una persona es hombre y 3 para identificar si es mujer; en el caso de las muestras del censo del 2000 y los conteos de los años 1995 y 2005, se identifica a hombres con el valor de 1 y a mujeres con el valor de 2. De aquí surge la necesidad de reprocesar algunos valores antes de integrarlos en un solo repositorio, permitiendo tener una representación uniforme de los datos. Otro ejemplo de inconsistencias encontradas es el factor de expansión omitido en el conteo 2005. INGENI recomienda utilizar un factor de expansión de 10 para cada registro de este conteo. De nuevo, es necesario hacer ajustes, es decir, agregar una variable con el valor de 10 antes de integrar la muestra del conteo 2005 a un repositorio único. Asimismo, la muestra del conteo 1995 esta formada por el 0.3% (332,061) del total registros (102,827,600), que difiere del 10% (aproximado) del total de registros de las muestras del conteo 2005 y censos de los años 2000 y 2010. Otros factores que pueden complicar las tareas de homogenización de las bases de datos son los nombres de las variables que también varían entre los años censales, principalmente el conteo del 1995 que contiene nombres de variables basados en el número de pregunta de la encuesta.

Cabe mencionar que los ejemplos presentados son solo del análisis realizado para los censos y conteos. Se debe tomar en cuenta que en este proyecto se

planean utilizar otras bases de datos que probablemente tengan otro tipo de diferencias entre los años que se levantaron. Es recomendado analizar bien cada base de datos antes de integrarse en un solo repositorio.

Tecnologías para el desarrollo de software

Para este proyecto se utilizará el lenguaje de programación Python, un lenguaje de alto nivel que frecuentemente es usado en aplicaciones científicas. Se planea utilizar las bibliotecas (llamadas módulos en Python), NumPy y SciPy. NumPy permite el manejo de grandes arreglos numéricos y matrices en Python de forma eficiente y conveniente. SciPy consiste de un conjunto de funciones de alto nivel implementadas sobre NumPy.

Durante este proyecto se planea desarrollar un sistema que permita generar reportes y gráficas de forma interactiva. Este sistema pretende servir como apoyo al grupo de investigadores que participan en este proyecto con el objetivo principal de calcular los saldos netos migratorios entre entidades federativas de México. El sistema se desarrollará con tecnologías web (HTML, Javascript, CSS). Para agilizar el desarrollo del sistema se utilizará Django, un marco de trabajo (framework) de software Web para el lenguaje Python. Django simplifica el desarrollo de aplicaciones web proporcionando bibliotecas de programación que realizan tareas que frecuentemente se necesitan en este tipo de software.

Matplotlib (Python) y D3.js (Javascript) serán las bibliotecas que apoyen la creación de gráficas (véase Figura 1 y Figura 2). Matplotlib es utilizado en la comunidad científica junto a NumPy/SciPy para generar imágenes de gráficas de distintos tipos. Por otro lado, D3.js permite la creación de gráficas y visualizaciones cartográficas sin necesidad de crear imágenes. D3.js habilita el control directo de representaciones gráficas en un navegador, esto permite una visualización interactiva. Por ejemplo, se pueden cambiar la fuente de los datos en una gráfica y que estos cambios se reflejen sin necesidad de descargar una imagen diferente.

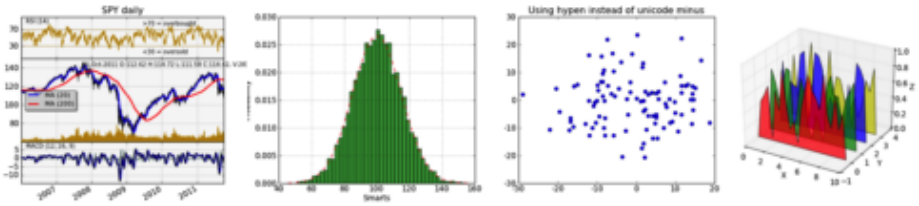

Figura 1. Ejemplos de gráficas generadas con Matplotlib.

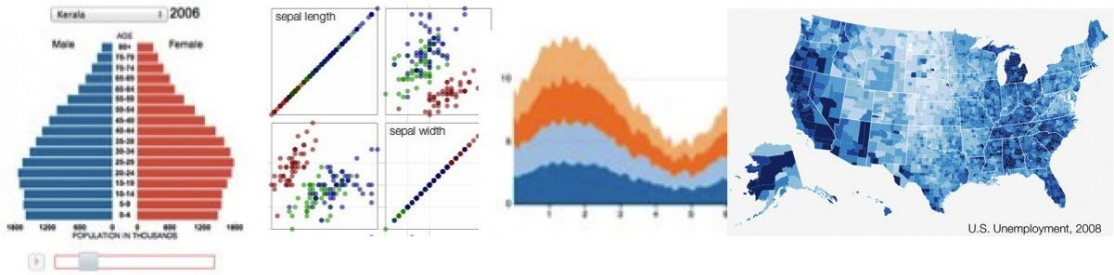

Figura 2. Ejemplos de gráficas y un mapa de coropletas generados con D3.js

Referencias

- Bostock, Michael (2012). D3.js, Data-Driven Documents. http://d3js.org/. Consultado en Mayo de 2012.
- Django Software Foundation (2012). Django documentation. http://docs.djangoproject.com. Consultado en Abril de 2012.
- INEGI (2011). Instituto Nacional de Estadística y Geografía (INEGI). http://www.inegi.org.mx/. Consultado en Mayo de 2012.
- Janert, Philipp (2011). Data analysis with open source tools. Sebastopol, CA: O'Reilly.
- Oracle (2011). MySQL 5.5 Reference Manual. Redwood City, CA: Oracle Inc.