La unión de bases de datos en ausencia de un identificador: El caso del Padrón de Beneficiarios de la SAGARPA (años) y el VII Censo Agrícola, Ganadero y Forestal 2007

Carlos Alberto Francisco Cruz
Juan Francisco Islas Aguirre
Jorge Lara Álvarez

Noviembre 13, 2014

SAGARPA
SECRETARÍA DE AGRICULTURA, GANADERÍA, DESARROLLO RURAL, PESCA Y ALIMENTACIÓN

Motivación

Realizar una evaluación de impacto de los Programas de la Sagarpa requiere de elementos mínimos:

- Padrón de beneficiarios: Se tienen registros administrativos en diferentes bases de datos que no era posible vincular debido a la ausencia de un identificador en común.

- Listado de posibles contrafactuales: Se determino que el VII Censo Agrícola, Ganadero y Forestal 2007 como una fuente viable para obtener información de personas no beneficiarias de los programas pero que podrían serlo.

Motivación

Tareas establecidas para la evaluación de impacto:

1. Conformar un padrón de beneficiarios de la SAGARPA

2. Vincular el padrón con el VII Censo Agrícola, Ganadero y Forestal 2007 y así obtener posibles controles para la evaluación de impacto

Limitante principal: No se tiene un identificador que permita unir ambas bases de datos

Base de datos 1

Nombre	A. Paterno	A. Materno	Entidad	Municipio	Localidad
Ma. Guadalupe	Hernández	Cruz	Chiapas	Arriaga	Arriaga
José	Bautista	Soto	Sonora	Agua Prieta	Agua Prieta

Base de datos 2

Nombre	A. Paterno	A. Materno	Entidad	Municipio	Localidad
ALBERTO	RAMÍREZ	ALTAMIRANO	GUANAJUATO	GUANAJUATO	GUANAJUATO
MARIA GUADALUPE	HERNANDEZ	CRUZ	CHIAPAS	ARRIAGA	ARRIAGA

Otras limitantes para lograr la vinculación

- Variaciones por deletreo. Ejemplo: Leydi vs Lady.
- Variaciones por pronunciación. Ejemplo: Carla vs Karla.
- Nombres compuestos. Ejemplo: Maria Guadalupe vs Ma. Guadalupe
- Nombres alternativos. Ejemplo: José vs Pepe
- Nombres sólo indicados con la primera letra. Ejemplo: José Juan vs J. J.

Alternativa para unir bases de datos:

Hacer uso de información contenida en las bases de datos

Stata tiene el comando “Reclink” que permite realizar el vinculo de información entre dos bases de datos a partir de cadenas de texto con base en un bigrama.

Un bigrama es utilizado como base para el análisis estadístico de texto y permite realizar la comparación entre textos a partir de su reconocimiento de voz.

El bigrama proporciona la probabilidad condicional de una palabra (W_n) dada una palabra precedente (W_{n-1}):

$$P(W_n|W_{n-1}) = \frac{P(W_n, W_{n-1})}{P(W_{n-1})}$$

Genera una probabilidad de similitud entre una cadena de texto de una base de datos en relación a otra cadena de texto de otra base de datos

reclink varlist **using** “filename” , option

idmaster(varname): identificador numérico de la primer base

idusing(varname) : identificador numérico de la segunda base

gen(newvar): Genera una nueva variable con la probabilidad de similitud

wmatch(match weight list) : Asigna pesos a cada una de las variables de texto (1 a 20)

wnomatch(non-match weight list): Cuando no logre vincular un dato se puede utilizar una asignación de pesos alternativa (1 a 20).

required(varlist): Indica las variables que deben ser idénticas en ambas bases

orblock(varlist): Indica que las observaciones deben ser idénticas en alguna de las variables señaladas

exclude(filename) : Excluye observaciones de otra bases de datos

minscore(#) : Especifica el valor mínimo para establecer el vinculo (0 a 1)

minbigram(#): Especifica el valor del bigrama mínimo para declarar como similares dos cadenas de texto (0 a 1)

Etapas para realizar la vinculación de las bases de datos a través del “Reclink”

I. Pre-Proceso: Eliminar duplicados y quitar caracteres especiales (¡ , ” , # , $, % , & , (, = , ? , ”, . , :)

II. Grupos o bloques: Si se trata de bases de datos con un número importante de observaciones es recomendable hacer grupos o bloques.

III. Comparación: Se hace uso del “Reclink” para la vinculación de las bases de datos

IV. Clasificación: Se realiza una revisión de los registros vinculados a partir de la probabilidad generada por el “Reclink”

V. Fusión: Una vez depurada los resultados obtenidos se procede a vincular las bases de datos.

Padrón de Beneficiarios de la SAGARPA

Se conformaron más de 30 componentes de los programas vigentes de la Sagarpa haciendo uso de las variables:

Nombre, CURP, RFC, entidad, municipio y localidad;

En suma los componentes tenían más de 3.5 millones de registros.

La vinculación obtuvo un padrón de beneficiarios de **2,683,713** beneficiarios.

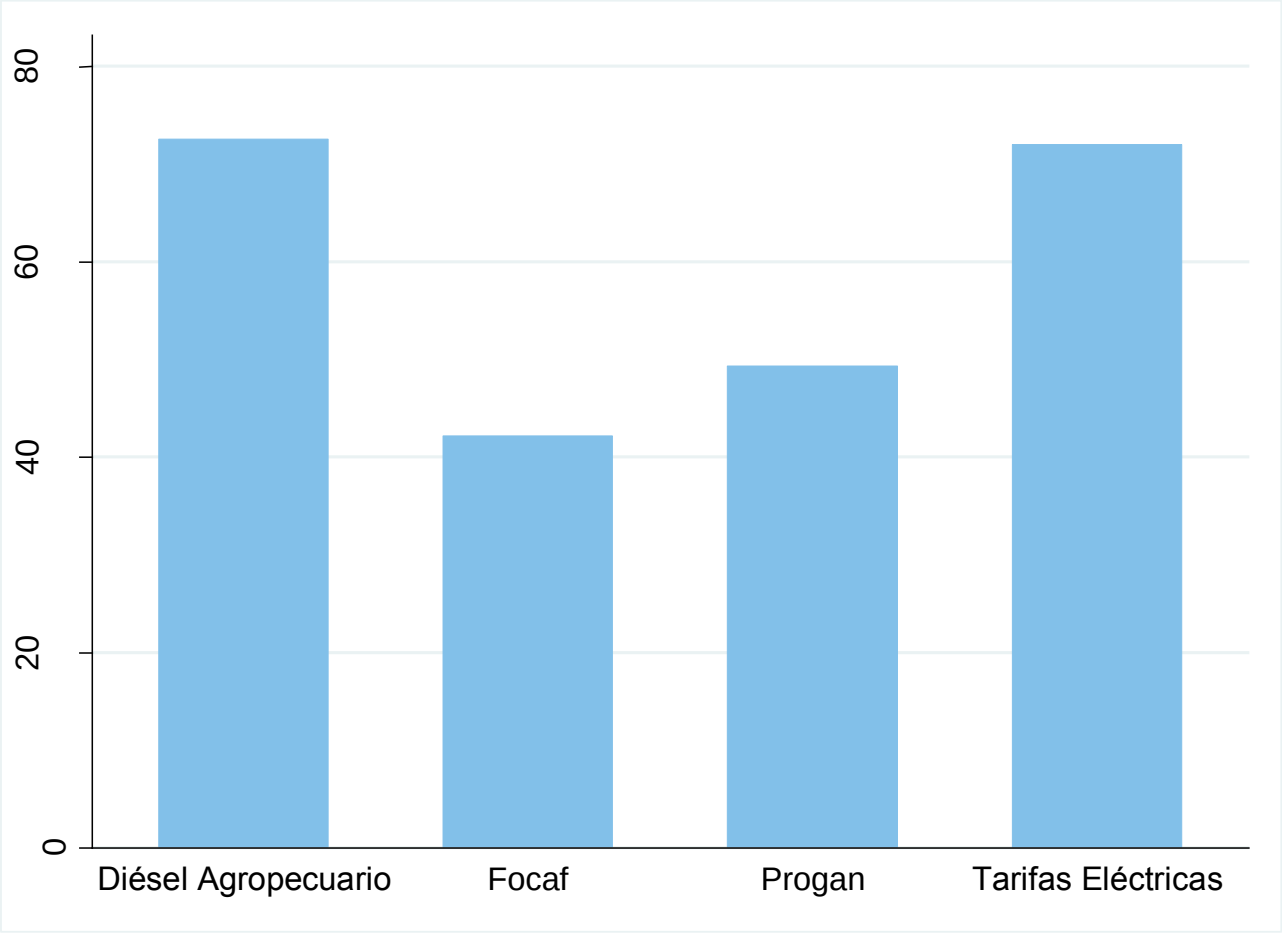
75.3% Hombres
24.6% Mujeres

La información permite identificar a los beneficiarios de más de 2 componentes.

Aproximadamente el 4% de los beneficiarios tiene más de 2 componentes

Entidad	Beneficiario	%	% Acum
Oaxaca	301,761	11.24	11.24
Chiapas	259,752	9.68	20.92
Veracruz	228,404	8.51	29.43
Puebla	175,775	6.55	35.98
Guerrero	160,602	5.98	41.96
México	144,796	5.40	47.36
Michoacán	132,234	4.93	52.29
Zacatecas	129,014	4.81	57.1
Hidalgo	117,839	4.39	61.49
Guanajuato	107,105	3.99	65.48
Jalisco	106,333	3.96	69.44
San Luis Potosí	99,336	3.70	73.14
Durango	91,171	3.40	76.54
Sinaloa	81,264	3.03	79.57
Chihuahua	74,068	2.76	82.33
Tamaulipas	68,621	2.56	84.89
Yucatán	50,952	1.90	86.79
Nayarit	45,180	1.68	88.47
Tabasco	40,717	1.52	89.99
Tlaxcala	39,205	1.46	91.45
Campeche	36,095	1.34	92.79
Coahuila	34,688	1.29	94.08
Querétaro	28,751	1.07	95.15
Morelos	25,787	0.96	96.11
Quintana Roo	24,462	0.91	97.02
Sonora	23,024	0.86	97.88
Nuevo León	21,853	0.81	98.69
Aguascalientes	14,538	0.54	99.23
Colima	8,008	0.30	99.53
Baja California	6,603	0.25	99.78
Baja California Sur	2,441	0.09	99.87
Distrito Federal	2,224	0.08	99.95
Total	2,683,713	100	

Gráfica 1: Beneficiarios de otros componentes que a su vez tienen Procampo (%)



VII Censo Agrícola, Ganadero y Forestal 2007

Tiene 5,548,845 unidades de producción dedicadas a actividades agrícolas, pecuarias o forestales.

Se realizó el vinculo del padrón de beneficiarios con el Censo, lo que implico un cruce de información de 2,683,713 contra 5,548,845 de registros.

Gráfica 2: Porcentaje de Beneficiarios identificados en el Censo