Mismos individuos

en diferentes bases de datos sin identificador común: la unión de las bases de datos de beneficiarios de la SAGARPA con una base universal

Carlos Alberto Francisco Cruz, Jorge Lara Álvarez, Juan Francisco Islas Aguirre, Ana Karen Díaz Méndez y Felipe Pérez Gachuz

Mexican laborers cut broccoli stalks for Smith Farms' crew A as.../Portland Press Herald/Getty Images

Este trabajo muestra una metodología para unir varias bases de datos con información de individuos idénticos entre ellas, pero sin un identificador universal. Con un mínimo de información de cada persona, las podemos unir. Lo que proponemos requiere una interfaz inteligente que emplea un algoritmo de pareo por registro (*record matching*), y su finalidad es detectar a los mismos individuos en diferentes bases de datos. Demostramos nuestra metodología pareando una base universal con diversas bases de datos de beneficiarios de la Secretaría de Agricultura, Ganadería, Desarrollo Rural, Pesca y Alimentación (SAGARPA).

Palabras clave: unión de bases de datos, SAGARPA, pareo por registro.

This paper shows a methodology to join several databases which hold information on identical individuals but without a universal identifier. With minimum information on each individual, it is possible to combine different databases. The proposed methodology requires an intelligent interface using a matching algorithm for each record (record matching) in order to detect the same individuals among different databases. Its usefulness is demonstrated by matching a universal database with several beneficiaries' databases from Agriculture, Cattle, Rural Development, Food, and Fishing Ministry of Mexico (*SAGARPA* in Spanish).

Key words: Joined Databases, SAGARPA, record matching.

Recibido: 29 de diciembre de 2014.
Aceptado: 26 de marzo de 2015.

Nota: se agradecen los valiosos comentarios de Leonardo Pérez Sosa y Natalia Eugenia Volkow Fernández, así como de los participantes del *2014 Mexican Stata Users Group meeting*.

1. Introducción

Una característica suficiente para unir dos bases de datos es que ambas compartan una clave que identifique a los individuos; no obstante, en los países en vías de desarrollo, como México, es común la ausencia de ésta debido, principalmente, a la falta de calidad en la recolección de los datos y a la diversidad de instituciones que los colectan. Esta carencia de un identificador es frecuente en las bases de datos de interés público: beneficiarios de programas sociales, padrones de productores, personas en situación vulnerable, entre otras. Existe un gran potencial de información que se está desaprovechando porque no es sencillo unir las diversas bases de datos; por lo tanto, su vinculación representa un reto y una gran oportunidad para mejorar las políticas públicas.

Actualmente, en México se cuenta con una cantidad importante de datos disponibles referente a la situación social y económica del país, los cuales son generados por diversas instituciones y oficinas de gobierno de manera independiente, y no se tiene el cuidado de vincular las observaciones a través de un identificador común. Leicester (2001) señala cuatro beneficios de unir bases de datos provenientes de distintas fuentes: uno de ellos es que con la base de datos obtenida es posible generar nueva inferencia estadística sobre un fenómeno en particular; la segunda es que permite incorporar a los estudios variables adicionales; la tercera amplía las opciones para estudios longitudinales; y la cuarta reduce considerablemente los costos de una investigación al utilizar información ya recabada. No obstante, como lo señalan Hernández y Stolfo (1998), unir bases de datos sin un identificador común es un trabajo que puede resultar complejo y laborioso.

La vinculación y unión está sujeta a la disponibilidad de un identificador que cumpla con dos funciones (Christen, 2006): 1) reconocer cada una de las observaciones al interior de cada base de datos y 2) distinguir a cada observación para su posible unión con diversas bases de datos que compartan el mismo identificador. Note que

cuando existe el mismo individuo en dos bases de datos y logramos unir la información contenida en ambas se dice que se realizó un *pareo*. Si se tiene el mismo identificador que permita la unión de los individuos en las dos, la tarea es fácil, pero cuando no se cuenta con éste se vuelve una labor compleja y, en ocasiones, imposible. La pregunta que motiva este documento es la siguiente: ¿cómo unir bases de datos que comparten algunos individuos cuando no se cuenta con un identificador como el que se menciona?

En los últimos años se han desarrollado diferentes metodologías y técnicas computacionales que buscan resolver este problema; por ejemplo, hay algunas que se sustentan en exploraciones a través de permutaciones, tal es caso de Reif (2010) y Barker (2012), quienes proponen dos algoritmos (implementados en Stata[1]) que permiten realizar una búsqueda de texto en otra base de datos. Con la permutación se puede encontrar una combinación similar a la palabra que se esté rastreando. No obstante, estos métodos tienen dos limitantes importantes: la primera es que se debe señalar cuáles serían las posibles diferencias entre las palabras a buscar y la segunda es que, una vez concluido el procedimiento, se debe realizar una revisión para detectar posibles errores. Actualmente se tienen estudios que se han beneficiado de la unión de diferentes bases de datos. Machado (2004) describe investigaciones llevadas a cabo en el campo de la salud infantil a partir de la vinculación de bases de datos provenientes de diferentes fuentes y que no cuentan con un identificador. Este mismo autor destaca la importancia de la recopilación y unión de la información existente.

En ese sentido, el objetivo de este trabajo es mostrar una propuesta que permita vincular y unir dos[2] bases de datos que no cuenten con un identificador. Nuestra metodología consiste en comparar variables comunes en ambas y diagnosticar si tenemos al mismo individuo en éstas. Asimismo, intenta utilizar toda la información disponible secuencialmente; además, tiene una interfaz inteligente para corroborar que el pareo se realice de manera correcta. Este método hace posible unir bases de datos de millones de observaciones. Hasta donde tenemos conocimiento, es el primero que se propone para unirlas sin un identificador común. Para demostrar su valor empírico, se presenta el caso de diversas bases de datos de beneficiarios de la SAGARPA y una base universal,[3] la cual contiene a los productores agropecuarios beneficiarios y a los que no lo son, por eso la llamamos así.

El primer paso fue unir entre ellas las bases de la Secretaría en lo que llamamos *padrón de beneficiarios*. La conformación de éste permite comparar a los beneficiarios de los diversos apoyos otorgados a través de programas sociales de la SAGARPA. Posteriormente, la vinculación del padrón con la base universal dio la pauta para un diseño riguroso de evaluación de impacto, a cargo de la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO, por sus siglas en inglés)-México, para tres programas de la Secretaría: PROCAMPO, PROGAN y Fomento Productivo al Café.

El resto del documento está desarrollado de la siguiente manera: la sección 2 explica la metodología y los problemas para su implementación, la 3 muestra un ejemplo real de nuestra propuesta y la 4 presenta conclusiones.

2. Vinculación y unión de bases de datos sin identificador

Como se mencionó antes, la unión de bases de datos está condicionada a la existencia de un identificador que, en primer lugar, permita reconocer de manera individual cada una de las observaciones en una base de datos y que, después, se pueda identificar información de las mismas en otra.

1 Es un *software* estadístico que permite realizar análisis cuantitativo y cualitativo. Debido a sus características y a la facilidad de su interfaz, es uno de los programas más utilizados en la actualidad.

2 No obstante, nuestra metodología se puede ampliar; por ejemplo, si se desea unir tres bases de datos (A, B y C), una opción sería juntar A con B y después la unión de éstas con la C.

3 Esta base universal incluye información de productores agropecuarios.

En caso de que la base de datos sea a nivel persona, lo ideal sería contar con un número de identificación nacional[4] para así poder ligar a cada persona con cualquier otra que lo contenga. En México, la Clave Única de Registro de Población (CURP) es una opción de identificador a nivel individual; lamentablemente, no toda la gente cuenta con este registro; además, está la situación de que un individuo tiene más de una CURP. En el 2014, de acuerdo con datos de la Secretaría de Gobernación, en la Base de Datos Nacional de la Clave Única del Registro de Población había 177 892 081 registros, cifra que supera al total de habitantes —a mediados de ese año, con base en la proyección del Consejo Nacional de Población (CONAPO), en México había 119 713 203—; esto se debe, principalmente, a que hay un registro acumulado y sin depurar, y con algunos duplicados, por lo cual es una alternativa que se encuentra limitada. Otra opción es hacer uso del Registro Federal de Contribuyentes (RFC), pero este dato es aún menos común que la CURP, debido a que sólo se genera para personas mayores de 18 años de edad y que se encuentren registrados en la Servicio de Administración Tributaria (SAT). Entre los regímenes fiscales más comunes destacan las personas que prestan servicios profesionales subordinados, aquellas que ofrecen servicios profesionales, o bien, las que tienen una empresa o negocio, de tal manera que la gente que cuenta con RFC debe cumplir con las obligaciones que establece la ley en el pago de impuestos.

Una opción cuando las bases de datos carecen de un identificador común es unirlas a partir de la información contenida en cada una de ellas; de nuevo, si ésta es a nivel individuo, posiblemente las bases de datos contengan variables que nos permitan distinguir al individuo. Éstas pueden ser numéricas, como la edad o número del domicilio o en texto, por ejemplo nombre de la persona o calle del domicilio. Entonces, el reto es usarlas en conjunto para suplantar a un identificador, permitiendo así unir diferentes fuentes de información. La combinación hace que el trabajo de unión sea posible pero, nótese, se debe hacer una selección para sustituir al identificador; además, las variables empleadas pueden contener errores de captura o diferentes maneras de registro, por lo que buscar la información en otra base datos debe permitir que ésta pueda no ser idéntica. Un error de captura puede ser una falta de ortografía o un número erróneo, un ejemplo es: 01/01/1981 y 01/01/81.

Entonces, la primera interrogante que surge es la siguiente: ¿es posible comparar la información de una variable capturada en dos bases de datos diferentes con posibles errores y heterogeneidad de captura? Si éstas son pequeñas, se puede hacer de forma manual; no obstante, si se desea comparar dos con más de cien observaciones cada una, el proceso se torna laborioso. En caso de tener millones de observaciones, una comparación manual se vuelve imposible.

Una alternativa acorde para vincular dos variables de distintas bases de datos que potencialmente están capturadas de una manera diferente o con error es un pareo por registro (Fellegi y Sunter, 1969), el cual intenta identificar entradas en diferentes bases de datos que corresponden a la misma unidad de observación. Existe una rutina en Stata para esto nombrada *reclink*, que emplea una cadena comparadora a partir de un bigrama (Blasnik, 2010), el cual es utilizado como plataforma para el análisis estadístico de texto y permite realizar la comparación entre textos (Collins, 1996). El bigrama proporciona la probabilidad condicional de una palabra (W_n), dada una palabra precedente (W_{n-1}):

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

Esta forma de vincular información entre dos bases de datos permite superar las variaciones de formato y los errores de captura. Sin embargo, su estructura y extensión pueden ocasionar que incluso la aplicación de este algoritmo lleve mucho

4 Por ejemplo, Documento Nacional de Identidad (DNI) en Argentina, Rol Único Nacional (RUN) en Chile y la Clave Única de Registro de la Población (CURP) en México.

tiempo y/o sea muy compleja. La complejidad radica en que puede resultar que la información para el mismo individuo no coincida exactamente en las dos bases de datos, sino que sólo sea muy semejante. Esto lo explicaremos más a detalle en la siguiente sección. Además, por lo general, debemos utilizar información contenida en más de una variable para asegurarnos de que se trata del mismo individuo. Esto dificulta la búsqueda.

Por estas razones, hemos diseñado una metodología que minimiza el tiempo computacional y maximiza la posibilidad de encontrar pareos. A continuación se describe y se presenta un ejemplo con el caso particular de los beneficiarios de la SAGARPA y nuestra base universal.

2.1 Etapas para unir bases de datos sin identificador a través de Stata

El objetivo es vincular aquellas que tienen diferentes fuentes de información a través de cinco etapas: 1) selección de variables a considerar en el pareo, 2) homogeneización, 3) división por grupos o bloques, 4) comparación, 5) interfaz inteligente y 6) pareo. Estos pasos son recomendables cuando se tienen bases de datos con un número importante de observaciones, en concreto más de 100 mil.[5] A continuación se describe cada una de las fases mencionadas.

Selección de variables

En esta parte del proceso se escogen las que se van a usar en cada base de datos para identificar a los individuos comunes entre ellas. Mientras más *única* sea la variable más valiosa va a ser; por ejemplo, la fecha de nacimiento (día, mes y año) es más útil que el estado de residencia. Si una base de datos sólo contiene variables muy *generales*, como sexo y edad en años, no va a ser posible unirla con otras bases usando este método.

Cabe señalar que las variables a comparar deben estar en ambas bases.[6] Por último, se recomienda depurar errores sistemáticos (caracteres adicionales, errores de ortografía comunes) en las que sean seleccionadas.

Homogeneización

Después de la selección, la siguiente etapa consiste en homogeneizar las variables de las dos bases de datos para el proceso de vinculación, es decir, el formato de captura en ambas bases. Esta fase depende, fundamentalmente, de la calidad de la información contenida en cada uno de los tabulados a vincular. Las variables de texto son las más complicadas de homogeneizar, esto debido a la diversidad de formatos en los que pueden ser capturadas y a las abreviaturas. Christen (2006) señala cinco posibles fuentes de heterogeneidad en los nombres propios en el momento de registrarse en una base de datos:

1. Variaciones por deletreo; ejemplo: Leydi *vs*. Laydi.
2. Variaciones por pronunciación; ejemplo: Paula *vs*. Paola.
3. Nombres compuestos; ejemplo: Marisol *vs*. María del Sol.
4. Nombres alternativos; ejemplo: Alfonso *vs*. Poncho.
5. Nombres sólo indicados con la primera letra; ejemplo: José Juan *vs*. J. J.

Además, podríamos agregar otro orden:

6. Nombre completo capturado en un solo campo, sin distinguir nombre de primer y segundo apellidos; ejemplo: Alfonso Rivera Gómez *vs.* Rivera Gómez Alfonso.[7]

5 Esto depende de la capacidad del procesador; 100 mil observaciones son suficientes en un procesador i9, a 8 Gb de RAM.

6 A manera de ejemplo: la base de datos A contiene las variables de sexo, nombre, apellido paterno y edad y la B, las de nombre, apellidos paterno y materno y edad. Entonces, no es posible emplear ni la variable apellido materno —dado que no está en la base A— ni tampoco la de sexo —pues no está en la B—; por lo tanto, sólo podríamos usar nombre, apellido paterno y edad.

7 Cabe señalar que todos los nombres aquí presentados son inventados; no obstante, ejemplifican situaciones reales.

Estas posibles variaciones son difíciles de detectar cuando se tienen muchas observaciones en las bases de datos. Una complicación adicional es que, debido a dichas variaciones y a la ausencia de un identificador, cada tabulado puede tener individuos duplicados capturados con una variación diferente. El cuadro 1 muestra un ejemplo de un registro duplicado en la misma base de datos; se observa que al interior de cada registro administrativo se pueden tener múltiples diferencias, es posible tener datos con letras mayúsculas y otros con minúsculas, además de las variaciones antes señaladas.

Cuadro 1

Ejemplo de un duplicado al interior de una base de datos

Nombre	Entidad	Municipio
María Guadalupe Luna Perea[a]	Chiapas	Arriaga
MA. G. LUNA PEREA	CHIAPAS	ARRIAGA

[a] Este nombre es sólo un ejemplo ilustrativo y no corresponde a una persona real.

Fuente: elaboración propia.

Ante esta situación, la etapa de homogeneización tiene dos objetivos específicos. El primero consiste en dejar todas las bases de datos en una condición similar; es recomendable substraer de las cadenas de texto los caracteres especiales —¡ , ” , # , $, % , & , (, = , ? , . , :—, además de los acentos y tener el texto en letras mayúsculas; otra recomendación, si se está trabajando con personas, es separar el nombre y los apellidos en celdas independientes (Herzog, Scheuren y Winkler, 2007). El segundo es revisar si una vez homogeneizadas las variables detectamos individuos duplicados dentro de cada base de datos. Alcanzar estos objetivos contribuye considerablemente a los resultados finales.

Una vez homogeneizadas las variables, podemos hacer un ejercicio de pareo perfecto, es decir, comparamos nuestras bases de datos para que coincidan con exactitud en todas las que seleccionamos en la primera fase. Dichos pareos perfectos los separaremos del grupo para reducir la carga computacional del pareo por registro que realizaremos en las siguientes etapas. Se relajará dicha exigencia hasta un nivel mínimo aceptable para considerar que dos observaciones corresponden a la misma persona. Esto se recomienda cuando: 1) las bases de datos contienen muchas observaciones y 2) se cuenta con varias variables útiles.

Grupos o bloques

Esta fase es importante cuando se van a unir tabulados con un número importante de observaciones. El método propuesto, basado en bigramas, puede ser muy tardado, sobre todo en bases de datos con miles o millones de observaciones, ya que cada una de las que contiene una base de datos es buscada en la otra. Una forma de reducir el esfuerzo computacional es generar grupos o bloques a partir de criterios comunes en ambas bases de datos (Baxter, Christen y Churches, 2003); por ejemplo, si en las dos se tiene información de las 32 entidades federativas del país, es recomendable generar una base de datos para cada una de ellas y realizar la vinculación con las bases de datos correspondientes a una entidad. Este planteamiento se puede realizar usando una sola variable o combinando varias; por ejemplo, podríamos agrupar por entidad federativa y sexo.

Comparación

En esta etapa se emplea la rutina de *reclink* (Barker, 2012). Comparamos las variables X_1, X_2, …, X_k de dos bases de datos al mismo tiempo. *Reclink* permite asignar un orden de relevancia a cada variable; mientras más *única* sea ésta más *peso* le adjudicaremos, por ejemplo, los apellidos paternos en México son más *únicos* que los nombres, por ello, les daríamos más relevancia a éstos. También, nos da la posibilidad de asignar un *peso* por no pareo, es decir, qué tanto nos importa si dos variables que estamos comparando no coinciden. Esta característica es clave en la rutina, pues se puede atribuir un gran peso de pareo a variables relevan-

tes, aunque tengan una alta probabilidad de ser capturadas con error, por ejemplo, la dirección del domicilio. A dicha variable le podríamos asignar un *alto peso* si coincide, porque es información única, pero un *bajo peso* en caso de no coincidir, porque puede tener mucha heterogeneidad en la captura.

Interfaz inteligente

A partir de los resultados de la comparación, es necesario establecer una medida o distancia que permita determinar qué tan similar es una observación en la primera base respecto a una observación en la otra (Bilenko, Mooney, Cohen, Ravikumar y Fienberg, 2003). En el caso de la rutina *reclink*, al basarse en un bigrama, es posible obtener la probabilidad de similitud entre una observación y otra. Lo que se obtenga depende, a su vez, de los pesos asignados a cada variable a considerar. Una probabilidad de 1 sugiere que los individuos de ambas bases de datos son idénticos; una de cero indica que no tienen relación.

Los registros con un valor inferior a 1 requieren de una interfaz inteligente, es decir, el discernimiento de una persona; es el usuario quien determinará si se trata del mismo individuo o no. El cuadro 2 muestra la sugerencia de los autores para clasificar las observaciones; se puede ver una posible clasificación de los registros comparados; cabe señalar que ésta se encuentra sujeta a una correcta elección de variables y sus respectivos pesos.

Pareo

Una vez finalizada la etapa de clasificación, es posible fusionar las bases de datos provenientes de fuentes de información distintas. El resultado es una base unificada confiable, a pesar de no contar con un identificador y de las dificultades de vinculación. Es importante destacar que mientras mejor sea la calidad de los datos y existan más variables en ambas bases de datos que permitan identificar a los individuos, más fácil será su vinculación.

3. Unión del padrón de beneficiarios de la SAGARPA y la base universal

En esta sección ejemplificaremos nuestra metodología con un caso práctico: la vinculación entre las bases de datos de beneficiarios de programas de la SAGARPA y de éstas con una base universal.[8] Una vez hecho el pareo, el análisis estadístico se hizo sobre variables que no nos permitían identificar a las personas.

En México, la SAGARPA es la principal entidad pública que se encarga de brindar apoyo a los sectores productivos agrícola, ganadero, de acuicultura y pesca, así como de postproducción. La población

8 Si bien, como explicamos arriba, es necesario usar variables como el nombre, sexo y edad, debemos dejar muy en claro que los investigadores que participamos en el presente estudio firmamos una carta compromiso para no extraer información personal de ninguna de las bases de datos utilizadas en esta investigación.

Cuadro 2

Clasificación de los registros con base en su probabilidad de similitud

Probabilidad	Clasificación
0.9500-0.9999	Se trata de la misma persona con pocas diferencias. Requiere de una revisión poco exhaustiva.
0.9000-0.9500	Puede tratarse de la misma persona, pero con varias diferencias. Requiere de una revisión exhaustiva.
0.8500-0.9000	Se tiene poca certeza de que sea la misma persona. Requiere de una revisión muy exhaustiva.
Menor a 0.8500	No se trata de la misma persona.

Fuente: elaboración propia.

objetivo a la que se orientan sus programas incluye personas físicas y morales, grupos, proyectos o cooperativas, dependiendo de las características del programa o componente en cuestión, de los montos de apoyo, así como de los criterios de elegibilidad. Hoy en día, en la Secretaría existen más de 50 programas sociales,[9] cada uno con su propia base de beneficiarios. Todos estos detalles e información adicional se modifican o conservan anualmente y se plasman en las reglas de operación, las cuales son publicadas en el *Diario Oficial de la Federación* en enero de cada año.

Por otro lado, en cuanto a la sistematización de la información de sus beneficiarios, no hay uniformidad, no se han establecido estándares de cómo llevar a cabo los registros. Cada subunidad de la SAGARPA opera su programa social; en algunos casos, ésta es la que enlista a los beneficiarios; en otros, la delegación estatal es la que lo hace. Cabe mencionar que no hay una vinculación entre las diferentes delegaciones estatales de la SAGARPA, así como con las subunidades responsables, las ventanillas de recepción de solicitudes, y otros agentes involucrados en la operación de los componentes. Entonces, la información recabada por cada componente no es homogénea y, por ende, resulta complicado para la Secretaría tener control respecto a quiénes se están destinando recursos públicos. En síntesis, esta instancia de gobierno no cuenta con un mecanismo preciso y eficiente para identificar a los beneficiarios entre sus componentes, pues no se vincula información entre los diferentes programas de apoyo.

Sus bases de datos cuentan con potenciales identificadores, como la CURP o el RFC. Sin embargo, existen muchos individuos sin observación para estas variables, además de que la calidad de captura no es buena. Un reto adicional es que cada una de ellas, en potencia, tiene individuos duplicados. Esto se debe a la mala calidad en la captura, puede ocurrir que una misma persona aparezca dos o más veces, pero no necesariamente la información capturada para cada uno de ellos es idéntica en sus duplicados, es decir, puede haber errores y heterogeneidad de captura para un mismo individuo.

Para identificar duplicados, depuramos los errores más comunes y, después, utilizamos las siguientes variables para comprobar si existieron errores sistemáticos de captura y contábamos con observaciones duplicadas:

1. Nombre, apellido paterno, apellido materno, entidad, municipio y localidad.[10]
2. Nombre y CURP.[11]
3. Nombre y RFC.

Una base de datos que unifique todos los listados de beneficiarios de la SAGARPA traería los siguientes beneficios:

- Un conteo preciso de esas personas.
- Un conteo del total de apoyos que recibe cada beneficiario.
- La relación entre los programas que recibe un beneficiario con varios apoyos.
- Un conteo del monto total de apoyo que recibe cada uno.

Como cada componente colecta datos sobre diferentes aspectos de un beneficiario, se podría lograr una caracterización más detallada de los apoyados por más de un componente. Entonces, el primer objetivo es unificar la información para 31 programas sociales de la SAGARPA, de los cuales pudimos tener acceso a sus bases de datos. Como regla general, usamos nuestro método de la siguiente manera:

a) Selección de variables a considerar en el pareo. Las que pudieran identificar al individuo, las cuales incluyeron, dependiendo de su disponibilidad, nombre, apellidos paterno y

9 Llamados *componentes*, divididos en nueve grandes ramas.

10 La información sobre la localidad permitiría detectar duplicados con facilidad. Sin embargo, en la mayoría de las bases siempre se reporta la localidad del beneficiario. Para ello, se aplicó el filtro a nivel municipal.

11 En algunos casos se usó la fecha de nacimiento obtenida de la CURP. Esto se hizo cuando de forma manual se detectaron errores de captura en la CURP. De manera similar, para el caso del RFC incorrectamente capturado, se utilizó la fecha de alta tributaria en el componente de la SAGARPA. Este procedimiento, además, permitió obtener la edad del beneficiario en el 2013 (o en cualquier otro año).

materno, fecha de nacimiento (obtenida de la CURP), año de nacimiento/edad (CURP), RFC, entidad, municipio y localidad.

b) Homogeneización de formato de ciertos campos. Pusimos todas las variables de texto en mayúsculas, la fecha de nacimiento en el mismo formato, el año de nacimiento lo pasamos a edad en el 2013 y la entidad, municipio y localidad los convertimos a clave usando el catálogo de claves de localidades del Instituto Nacional de Estadística y Geografía (INEGI). Intentamos un *pareo perfecto* con las siguientes opciones:
   1. Nombre, apellidos paterno y materno, entidad, municipio y localidad.
   2. Nombre y CURP.
   3. Nombre y RFC.

Esto siempre y cuando los individuos contaran con toda la información, es decir, si las personas no tenían CURP no usábamos la opción 2.

Parte recursiva. Ronda 1

Una vez seleccionadas las observaciones con las que íbamos a realizar un pareo por registro, la primera ronda era la más exigente:

3) División por grupos o bloques. Exigíamos que la edad, la entidad, el municipio y la localidad coincidieran. Le dábamos más relevancia a los apellidos que a los nombres.
4) Comparación. Comparábamos los registros de apellidos y nombre.
5) Interfaz inteligente. Cada registro menor a 0.93 de probabilidad de coincidencia se revisó.
6) Pareo. Se unía la información de los individuos que coincidían en ambas bases de datos.

Parte recursiva. Ronda 2

Una vez seleccionadas las observaciones con las que íbamos a realizar un pareo por registro, esta segunda fase relajaba que la edad y la localidad coincidieran. Esto porque notamos que la edad, sobre todo cuando no era calculada de una CURP disponible, estaba mal capturada. También, notamos un problema de calidad en la captura de los datos de localidad:

3) División por grupos o bloques. Exigíamos que la entidad y el municipio coincidieran. Le dábamos más relevancia a los apellidos que a los nombres.
4) Comparación. Comparábamos los registros de apellidos y nombre.
5) Interfaz inteligente. Cada registro menor a 0.95 de probabilidad de coincidencia se procedía a revisar.
6) Pareo. Se unía la información de los individuos que coincidían en ambas bases de datos.

Parte recursiva. Ronda 3

Por último, relajamos la restricción de que el municipio debería de coincidir. Sin embargo, fuimos más estrictos al revisar el nombre y la edad de las personas:

3) División por grupos o bloques. Exigíamos que la entidad y la edad coincidieran. Le dábamos más relevancia a los apellidos que a los nombres.
4) Comparación. Comparábamos los registros de apellidos y nombre.
5) Interfaz inteligente. Cada registro menor a 0.999 de probabilidad de coincidencia se procedía a revisar.
6) Pareo. Se unía la información de los individuos que coincidían en ambas bases de datos.

Con este método pudimos formar una base única de beneficiarios de la SAGARPA que contiene información de 31 programas sociales que se encuentran actualmente en funcionamiento. Nuestro resultado fue que existe un total de 2 683 713 beneficiarios, de los cuales 75.3% son hombres y 24.7%, mujeres.

La conformación de una base única permite obtener información que no sería completa si

se tuviera la de un solo componente. El cuadro 3 muestra la distribución de beneficiarios por entidad federativa; se puede observar con mayor precisión que Oaxaca es la entidad con mayor número, ya que concentra a más de 11.24%, le sigue Chiapas (9.68%), Veracruz de Ignacio de la Llave (8.51%), Puebla (6.55%) y Guerrero (5.98%). Por su parte, las entidades con la menor cantidad son el Distrito Federal (0.08%), Baja California (0.20%) y Baja California Sur (0.09%).

Cuadro 3

Beneficiarios de la SAGARPA

Entidad	Beneficiarios	%	% Acum.
Oaxaca	301 761	11.24	11.24
Chiapas	259 752	9.68	20.92
Veracruz de Ignacio de la Llave	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 de Ocampo	132 234	4.93	52.29
Zacatecas	129 014	4.81	57.10
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 de Zaragoza	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**	**100.00**	

Fuente: elaboración propia.

A su vez, la base de datos permite saber a cuántos programas sociales está inscrito cada uno de los beneficiarios. En el cuadro 4 se puede observar que más de 82.51% sólo recibe apoyos de un componente, mientras que 14.14% lo tiene de dos y 2.87%, de tres. También, es posible identificar el caso en el que se otorgan varios componentes por beneficiario. El mismo cuadro muestra que se tiene un beneficiario que recibe apoyo de ocho componentes; nueve, de siete y 139, de seis.

Cuadro 4

Número de programas sociales de la SAGARPA por beneficiario

Número de componentes	Beneficiarios	%
1	2 214 212	82.51
2	379 514	14.14
3	76 953	2.87
4	11 415	0.43
5	1 470	0.05
6	139	0.01
7	9	0.00
8	1	0.00
Total	**2 683 713**	**100.00**

Fuente: elaboración propia.

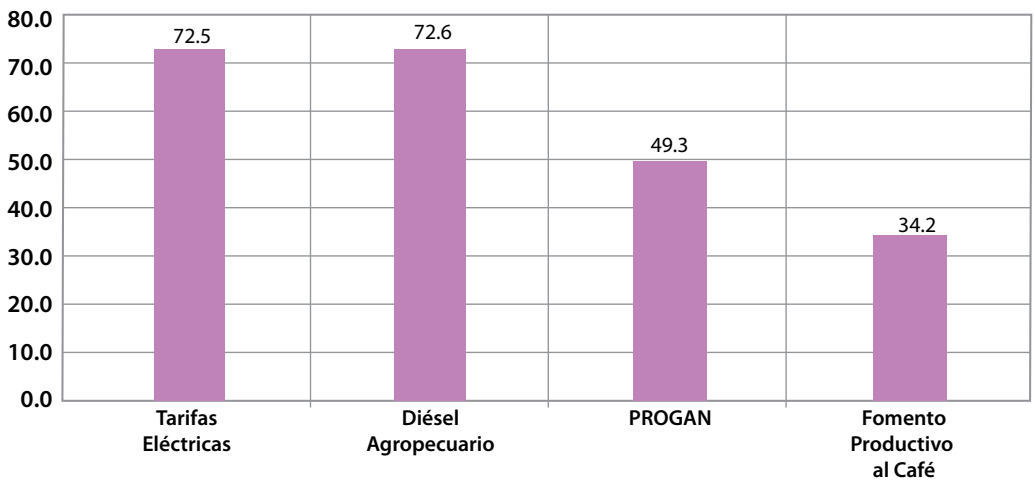
A partir de la información anterior es posible identificar el cruce que se puede tener entre dos componentes; por ejemplo, PROCAMPO es el que cuenta con mayor número de beneficiarios, por lo que se tiene un cruce importante de sus beneficiarios con otros componentes. La gráfica 1 muestra la cantidad de beneficiarios de Diésel Agropecuario, PROGAN, Fomento Productivo del Café y Tarifas Eléctricas que, a su vez, tienen PROCAMPO. En el caso de PROGAN, más de 150 mil (cerca de 50%) reciben, al mismo tiempo, PROCAMPO. A más de 34% de la población beneficiaria de Fomento Productivo del Café se le otorga, de manera simultánea, PROCAMPO. En el caso de Diésel Agropecuario y Tarifas Eléctricas son 72.6 y 72.5%, respectivamente.

3.1 Uniendo el padrón de beneficiarios de la SAGARPA y la base universal

Este padrón, que se mencionó al final de la introducción de este documento y que se formó mediante la unión de las bases de datos de la SAGARPA, dio la pauta para plantear un diseño de evaluación de impacto, por lo que, dadas las características de la información y la situación actual de los programas, se optó por un diseño cuasi-experimental. En este caso se requería de un marco muestral que permitiera ob-

Gráfica 1

Beneficiarios de PROCAMPO que simultáneamente son beneficiarios de otros programas (porcentajes)



Fuente: elaboración propia.

tener un grupo de control con el cual se pudiera establecer una comparación rigurosa con los beneficiarios del programa y, con ello, determinar el impacto de los programas de la SAGARPA en sus beneficiarios.

En ese sentido, una base universal que ofrece información detallada de los productores agropecuarios del país se determinó como la opción más viable para identificar un posible grupo de control. Para ello, fue necesario identificar en ésta a los beneficiarios de los programas de la SAGARPA. El reto era aún mayor, pues teníamos que vincular una base de datos de 2.6 millones de observaciones con una de cerca del doble.

La forma de vincular se llevó a cabo mediante el mismo método antes descrito, aunque es importante destacar que esta tarea se realizó sólo con los componentes de PROCAMPO, PROGAN y Fomento Productivo al Café.

Los resultados obtenidos son importantes (ver cuadro 5), ya que en el caso de PROCAMPO se identificaron 62% de los registros dentro de la base universal. En PROGAN se logró reconocer a 67% y en el de Fomento Productivo al Café, a 69 por ciento.

Cuadro 5

Beneficiarios de la SAGARPA identificados en una base universal

Componente	% en una base universal
PROCAMPO	62
PROGAN	67
Fomento Productivo al Café	69

Fuente: elaboración propia.

4. Conclusiones

La vinculación de la información proveniente de fuentes distintas permite un mayor aprovechamiento de los datos existentes, por lo que pueden disminuir considerablemente los costos de una investigación. La metodología descrita en este documento de trabajo brinda la posibilidad de unir dos bases de datos con individuos en común con un mínimo de información disponible e, incluso, con problemas en la calidad de captura de los datos. En ese sentido, el uso de este método, que emplea la rutina de Stata de *reclink*, hace posible la vinculación de dichas bases.

Nuestra metodología se ejemplificó con un caso real: la vinculación de información de la SAGARPA y una base universal, lo cual permitió realizar un diseño riguroso de una evaluación de impacto para los programas gestionados por la SAGARPA: PROCAMPO, PROGAN y Fomento Productivo al Café.

Fuentes

Barker, M. *Stata module to calculate the Levenshtein distance, or edit distance, between strings.* 2012. Consultado en: http://ideas.repec.org/c/boc/bocode/s457547.html

Baxter, R., P. Christen y T. Churches. "A Comparison of Fast Blocking Methods for Record Linkage", en: *CIMS Technical Report 03/139. CSIRO Mathematics, Information and Statistics*. 2003.

Bilenko, M., R. Mooney, W. Cohen, P. Ravikumar y S. Fienberg. "Adaptive Name Matching in Information Integration", en: *IEEE Intelligent Systems 18(5)*, 2003, pp. 16-23.

Blasnik, M. *RECLINK: Stata module to probabilistically match records*. 2010. Consultado en: http://EconPapers.repec.org/RePEc:boc:bocode:s456876.

Christen, P. "A Comparisonof PersonalName Matching: Techniques and Practical Issues", en: *Joint Computer Science Technical Report Series*. The Australian National University, 2006.

Collins, M. "A nex statistical parser bases on bigram lexical dependencies", en: *In Proceeding of the 34th Annual Meeting of the Association of Computational Linguistic*. Santa Cruz, CA, 1996, pp. 184-191.

Elmagarmid, A. K., P. G. Ipeirotis y V. S. Verykios. "Duplicate Record Detection: A Survey", en: IEEE Transactions on Knowledge and dataEngineering. Vol. 19, Núm. 1, enero de 2007, pp. 1-16.

Fellegi, I. P. y A. B. Sunter. "A Theory for Record Linkage", en: *Journal of the American Statistical Association*. 1969, pp. 1183-1210.

Hernández, M. A. y S. J. Stolfo. "Real-World Data is Dirty: Data cleasing and the merge/purge problem", en: *Data Mining and Knowledge Discovery*. 2, 1998, pp. 9-37.

Herzog, T. N., F. J. Scheuren y W. E. Winkler. *Data Quality and Record Linkage Techniques*. United States of America, Springer, 2007.

Leicester, G. *Methods for Automatic Record Matching and Linkage and Their Use National Statistics. National Statistics Methodological Series*. Oxford, 2001.

Machado, C. J. "A literature review of record linkage procedures focusing on infant health outcomes", en: *Cad. Saú de Pública, Rio de Janeiro*. 20(2), 2004, pp. 362-371.

Reif, J. Stata module to match strings base on their Levenshtein edit distance. 2010. Consultado en: http://ideas.repec.org/c/boc/bocode/s457151.html