Aprovechamiento estadístico de los registros administrativos

BID
Banco Interamericano de Desarrollo

Autor: Federico Segui Stagno

TABLA DE CONTENIDO

BASE CONCEPTUAL

Fuente: Segui, Federico (2019). Guía práctica sobre el uso estadístico de registros administrativos - Métodos y herramientas para la integración y explotación de registros administrativos con fines estadísticos.

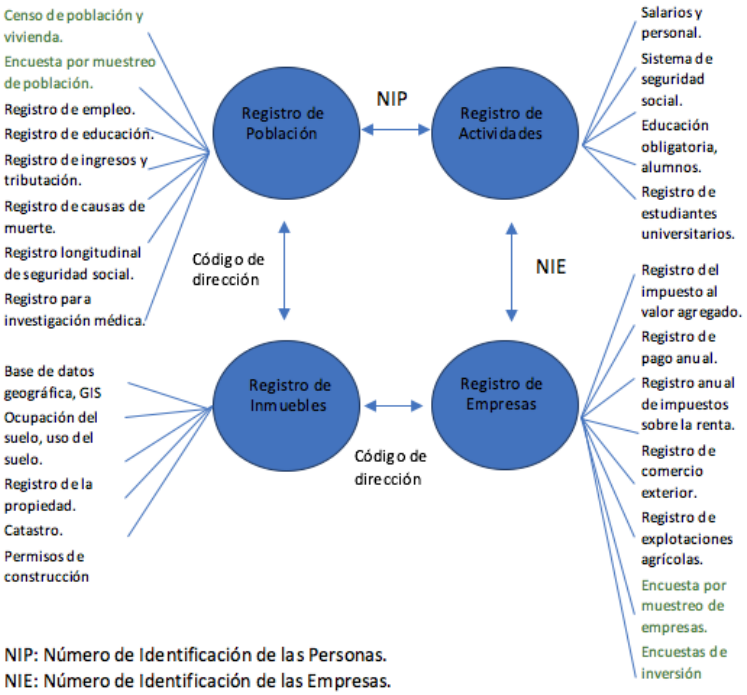
La vasta experiencia de los países nórdicos (Suecia, Noruega, Finlandia, Dinamarca, Islandia), pioneros en el uso de registros administrativos con fines estadísticos, está siendo replicada o adaptada en varios países alrededor del mundo.

Anders y Britt Wallgren en su trabajo “Estadísticas basadas en registros. Aprovechamiento estadístico de los registros administrativos” (traducido al español y publicado en 2012 por INEGI), describen la metodología sueca desarrollada para la producción de estadísticas basadas en registros administrativos.

La oficina de estadística de Suecia desarrolló un modelo de producción estadística basado en un sistema integrado de registros estadísticos conformado por cuatro registros base (población, empresas, inmuebles y actividades) integrados entre sí.

El modelo conceptual genérico de un sistema integrado de registros estadísticos planteado por Wallgren y Wallgren se ilustra en la siguiente figura.

Figura 1. Sistema de registros estadísticos por tipo de objeto y campo de estudio



Fuente: Anders Wallgren, Britt Wallgren. (2012). Estadísticas basadas en registros. Aprovechamiento estadístico de los registros administrativos. INEGI, México

Este modelo planteado por Wallgren y Wallgren ha sido ampliamente difundido en la región y se ha tomado como base conceptual para desarrollar la presente guía.

I. Elementos del sistema de registro

Los elementos, objetos, entidades o individuos del mundo real (personas, organizaciones, empresas, viviendas o inmuebles, vehículos, entre otros) se denominan objetos o unidades del sistema de registros y son los elementos que forman parte de la población.

La **población** la define el conjunto de objetos o unidades que la componen, incluyendo información sobre ubicación y tiempo.

La **población de interés** o **población objetivo** son todos los casos o unidades que forman parte del Registro Estadístico que cumplen con un conjunto de características particulares.

Las **variables** corresponden a una serie de atributos medibles de los objetos o unidades estadísticas.

Una **variable estadística** está definida por el tipo de objeto que presenta la característica (por ejemplo, superficie de la vivienda y superficie del predio son dos variables distintas), por el método de medición y la escala aplicada, y por el momento o período a los que se refiere la medición.

II. Categorías e interrelaciones de registros estadísticos

El sistema integrado de registros estadísticos está conformado por una serie de registros estadísticos:

Los registros base tienen como función definir los objetos y poblaciones del sistema de registros. Son la columna vertebral del sistema pues contienen los tipos de objetos y los vínculos más relevantes. La calidad del sistema está determinada por las definiciones de los objetos y la cobertura de los registros base.

- “Define tipos de objetos importantes.
- Define conjuntos de objetos o poblaciones estandarizadas importantes.
- Contiene vínculos con objetos de otros registros base.
- Contiene vínculos con otros registros relacionados con el mismo tipo de objeto.
- Es importante para el sistema en su conjunto, por lo que resulta esencial que sea de alta calidad y esté bien documentado.
- Es importante para el marco muestral.
- Se puede usar para estadísticas demográficas relacionadas con personas, actividades, inmuebles o empresas.
- Las fechas de nacimiento y defunción deben estar presentes en el registro base para producir estadísticas demográficas.”

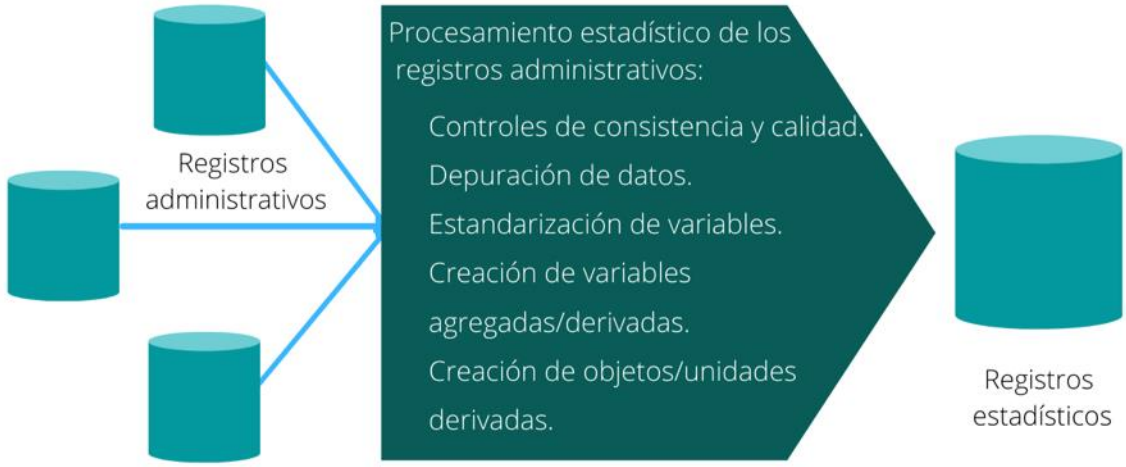
Los **registros primarios** son aquellos que se basan directamente en al menos una fuente administrativa.

Los **registros integrados** son creados combinando exclusivamente información ya existente en los registros estadísticos del sistema.

III. Transformación e integración de registros administrativos en registros estadísticos

El uso estadístico de datos administrativos requiere de una conversión del registro administrativo a un registro estadístico e implica una serie de procesos y subprocesos, como ilustra la siguiente figura.

Figura 2. Proceso de transformación de registros administrativos a registros estadísticos.



Fuente: Segui, Federico (2016-II). Marco conceptual y metodológico que sustenta el diseño, desarrollo e implementación de un sistema integrado de registros estadísticos de población e inmuebles. Proyecto Estadística de Población e Inmuebles a partir del uso de registros administrativos oficiales en la Comunidad Andina financiado por el BID

IV. Controles de calidad y depuración de datos

Los registros administrativos no han sido diseñados con fines estadísticos (como ya sabemos) y en muchos casos no aplican las buenas prácticas, recomendaciones o estándares metodológicos requeridos para producir estadísticas.

Los procesos de ingreso de datos, validación y control de consistencia de la información no siempre aplican los mismos criterios que utilizarían los INE, o las validaciones se aplican con mayor rigurosidad sobre variables administrativas (de interés para la fuente administrativa) que ciertas variables de interés estadístico (utilizadas por los INE) que gozan de menor severidad en sus controles, pues revisten menor importancia para la fuente administrativa.

Todos estos elementos provocan errores en los datos capturados en los registros administrativos, por lo que es necesario implementar controles de consistencia y calidad para detectarlos y proceder con el proceso de depuración de datos para minimizarlos. Los resultados de la validación y depuración también deben ser documentados.

V. Variables del sistema

A. Categorías de variables El acceso a registros administrativos por parte del INE debe estar garantizado por ley, al igual que la protección de la privacidad.

Los registros estadísticos están conformados por diferentes categorías de variables. Es importante comprender su utilidad y relevancia dentro del sistema de registros estadísticos.

- **Variables clave o llaves de identificación o variables identificadoras**, como su nombre lo indica, se usan para identificar objetos o unidades. Se usan para hacer la unión entre registros (método determinístico) con el mismo tipo de objetos o unidades.

- **Variables estadísticas**, son las variables provenientes del registro administrativo de interés estadístico, que se utilizan para generar estadísticas o hacer análisis estadístico.

- **Variables de contacto**, (identificadores explícitos) como nombre, dirección, teléfono, correo electrónico se usan cuando el INE necesita contactarse con el objeto o unidad al que corresponde el caso (en general cuando se utilizan cuestionarios).

- **Variables de ubicación geográfica**, utilizadas para asociar los objetos a ubicaciones físicas en el territorio. Entran en esta categoría las coordenadas geográficas determinadas por GPS y las variables codificadas correspondientes a nomenclátores de divisiones político-administrativas-geográficas del territorio de cada país.

- **Variables de referencia temporal**, indican el momento en que ocurre un evento relativo a los objetos o unidades del registro.

- **Variables derivadas o agregadas**, son variables que no están presentes en ningún registro administrativo, pero son necesarias para facilitar la producción de estadísticas.

- **Variables de uso administrativo interno**, se utilizan para indicar resultados de procesos internos como por ejemplo códigos de error de las variables estadísticas, valores editados, resultados de la depuración de variables, código de la fuente administrativa de las variables, código de la fuente o versión del código de clasificación de las variables codificadas (CIUO-08, CIIU-Rev4, etc.), fecha de actualización de las variables, estado de los objetos o unidades (personas: Activo, Inactivo por emigración, Inactivo por muerte, Eliminado por proceso administrativo, Inactivo por duplicado; viviendas: Activa, Inactiva por demolición, Eliminada por proceso administrativo, Inactiva por duplicado), entre otros.

- **Variables de unión o fusión**, (claves foráneas), son utilizadas para hacer la unión con otras tablas de una base de datos, describen relaciones entre diferentes tipos de objetos.

B. Estandarización de variables

Las definiciones de las variables de registros administrativos se ajustan a las necesidades administrativas y no siempre se corresponden con las definiciones estadísticas de acuerdo a un uso estadístico específico. Es así que resulta imprescindible estandarizar las variables de los registros administrativos para incorporarlas al sistema de registros estadísticos y poder utilizarlas en diferentes proyectos estadísticos.

El hecho de contar con variables estandarizadas minimiza errores, evita duplicidad de esfuerzos de documentación (quienes utilizan las variables estandarizadas no necesitan volver a documentarlas) y facilita su utilización.

La estandarización de variables implica, además, la transformación de formatos de datos y en algunos casos los datos en sí mismos. Por ejemplo, la variable fecha de nacimiento de un registro administrativo almacena los datos en el formato dd/mm/aaaa, pero la variable estandarizada fecha de nacimiento del registro estadístico de población se ha definido con el formato aaaa-mm-dd, lo cual implica que los formatos de datos originales de la variable del registro administrativo deben transformarse (siguiendo un algoritmo) para convertirlos al formato estandarizado de la variable del registro estadístico.

Otro ejemplo se puede ver en las variables categóricas donde los códigos de categoría y sus descripciones podrían cambiar al transformarlos a las correspondientes variables estandarizadas del registro estadístico.

C. Variables derivadas o agregadas

En la generación de estadísticas basadas en registros administrativos no se tiene la posibilidad de diseñar los instrumentos de captura de datos, no es posible definir las preguntas del cuestionario. En estos casos se debe apelar a la creación de variables agregadas utilizando las variables disponibles de los registros administrativos.

Las variables derivadas o agregadas se crean de seis modos diferentes:

1) Por medio de cálculos aritméticos o procedimientos lógicos.
2) Mediante agrupamiento de valores.
3) Mediante codificación.
4) Por asociación de otras variables.
5) Mediante agregación de otras variables.
6) Mediante modelos estadísticos.

Las variables derivadas se pueden generar combinando estos métodos, incluso aplicando cálculos más complejos con procedimientos basados en ciertas reglas.

D. Trazabilidad de variables

A efectos de disponer de una adecuada documentación de metadatos es fundamental mantener la trazabilidad de cada variable que conforma el registro estadístico, desde su origen en el registro administrativo, donde fue creada por la respectiva fuente administrativa, pasando por el proceso de validación, depuración y estandarización y justificación de su selección e inclusión en el registro estadístico. El sistema de documentación de metadatos debe garantizar la trazabilidad de todas las variables del sistema de registros estadísticos con sus respectivas fuentes administrativas (variables administrativas utilizadas para generar las variables estadísticas).

Debe existir una vinculación entre el inventario de registros administrativos y los metadatos del sistema integrado de registros estadísticos.

VI. Unión de registros

La unión de registros es la tarea de identificar de forma rápida y precisa los registros correspondientes a la misma entidad/objeto/individuo de una o más fuentes (archivos) de datos.

A lo largo de este documento se ha mencionado la necesidad de unir registros de diferentes fuentes administrativas para integrarlos y transformarlos en registros estadísticos (y entre registros

administrativos y registros estadísticos), para lo cual es necesario contar con claves únicas de identificación y en algunos casos claves foráneas también.

En un mundo ideal, o más parecido a la realidad de los países nórdicos, todos los registros administrativos deberían contar con una variable clave de identificación de los casos (en general es una única variable, pero hay casos donde se utiliza más de una variable para crear la clave de identificación). Además, esta clave debería estar estandarizada y utilizada por todos los registros administrativos referidos al mismo tipo de objetos o elementos. De esta forma se podrían unir filas de diferentes registros que refieren al mismo objeto o elemento, en los casos que coincida exactamente la clave de identificación de ambos registros (método determinístico).

Por ejemplo, en el caso del registro civil las personas tienen como clave la cédula de identidad o DNI y esta variable clave es utilizada en todos los otros registros de población como clave de identificación de cada persona. Se busca lograr una coincidencia exacta carácter a carácter de la variable clave, entre ambos registros.

La realidad en la región dista bastante de este escenario ideal. En todos los países existen varios casos de registros administrativos que no utilizan una clave única estandarizada común para identificar los casos, sobre todo en los registros de inmuebles.

En las situaciones donde sí se utiliza una clave de identificación común, ésta presenta ciertos problemas, como, por ejemplo: contiene duplicados, o la estandarización no es de alcance nacional entonces cada municipio tiene su propio código y procedimiento para crearlo.

Si se aplica el método determinístico de unión de registros sin evaluar la calidad de la variable clave utilizada, se corre el riesgo de unir filas o casos de ambos registros que no se tiene la certeza que correspondan realmente al mismo objeto o elemento. Asimismo, si se tienen duplicados no se sabrá cuál de las filas es la que corresponde efectivamente al elemento u objeto en cuestión. Además, están los casos cuya clave de identificación está en blanco o contienen datos inválidos y no será posible unirlos con otros registros.

Por estas razones se deben plantear métodos alternativos para la unión de registros. Los métodos probabilísticos de unión de registros utilizan algoritmos específicos para determinar con cierta certeza (probabilidad) que dos filas o casos de diferentes registros corresponden al mismo elemento u objeto. Utilizan otras variables del registro (aparte de las variables clave de identificación), combinándolas para lograr una pseudo-clave. Las variables de población más comúnmente utilizadas para estos propósitos son nombre, apellido, fecha de nacimiento o edad y otras dependiendo de la disponibilidad en el registro. En el caso de inmuebles se utiliza la variable de dirección del inmueble.

Las variables alfanuméricas que podrían ser utilizadas para hacer la unión de registros presentan una serie de problemas: errores, variaciones y datos en blanco; diferencias en definiciones, períodos, formatos de los datos capturados por los diferentes registros administrativos; cambios en los datos a lo largo del tiempo, por ejemplo, cambio de domicilio de las personas; errores de digitación, letras o palabras ingresadas en diferente orden; palabras fusionadas o divididas, palabras incompletas, letras faltantes o excedentes; puntuación o acentuación incorrecta; abreviaciones. Por lo tanto, es imprescindible realizar la depuración de datos y estandarización de variables antes de iniciar el proceso de unión de registros.

Los métodos de unión probabilística de registros implican el cálculo de pesos de unión estimados con base en todos los casos coincidentes y no coincidentes observados de los valores de la(s) variable(s) utilizada(s) para la unión.

Los métodos probabilísticos permiten obtener una mejor unión de registros que un simple método de unión determinístico. Además, pueden ser utilizados para detectar casos/filas duplicadas en un archivo de datos del registro administrativo (cuando dos o más casos/filas tienen diferentes valores en la clave de identificación, pero en realidad corresponden al mismo objeto/elemento/individuo).

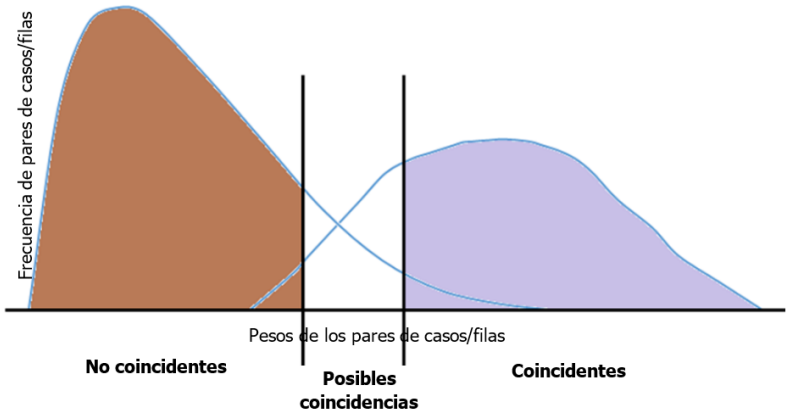
El proceso de unión probabilística de registros consta de tres fases:

1) Pre-unión. Esta etapa implica la depuración de datos y estandarización de variables, como se ha explicado en apartados anteriores.

Las variables de domicilios son un caso particular de variables alfanuméricas y se deben normalizar siguiendo los criterios de estandarización de cada país. En el caso que los domicilios se almacenen en más de una variable (calle, número, complemento, piso, apartamento, etc.) éstas se deberían concatenar en una sola variable siguiendo los mismos criterios de estandarización.

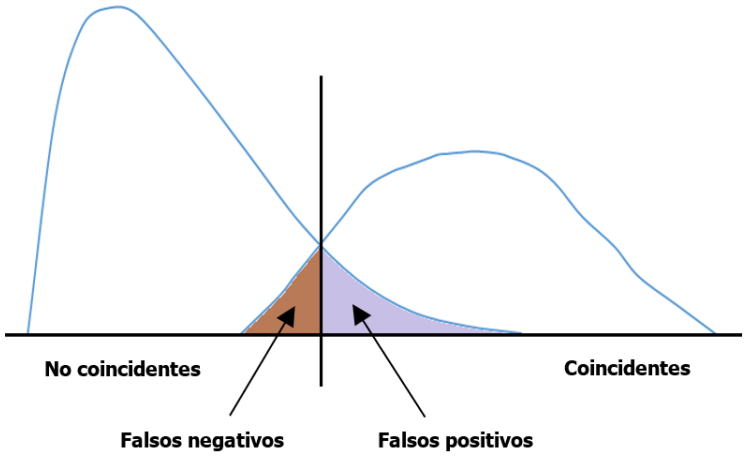
2) Unión. Se aplica la unión de registros para decidir cuándo dos casos o filas de diferentes registros coinciden (match), o sea pertenecen al mismo objeto o elemento, o no coinciden (no-match) es decir, se trata de diferentes objetos o elementos. También puede utilizarse para encontrar duplicados en el mismo archivo del registro administrativo. Se utilizan herramientas informáticas que proveen varios algoritmos para realizar la unión probabilística. Estas aplicaciones primero calculan los pesos para cada uno de los posibles casos de unión y luego se determinan los umbrales de los casos unidos y los no unidos.

Gráfico 1. Umbrales que delimitan las zonas correspondientes a los tres grupos de pares de casos/filas: no coincidentes, posibles coincidencias y coincidentes.



• Otra opción es seleccionar un único umbral de corte (L) y todos los puntajes que estén por encima serán considerados como coincidencias y los que estén por debajo de ese mismo umbral serán calificados como no coincidentes, sin dejar lugar a “posibles coincidencias” (ver siguiente figura), es decir, todo el proceso será automático.

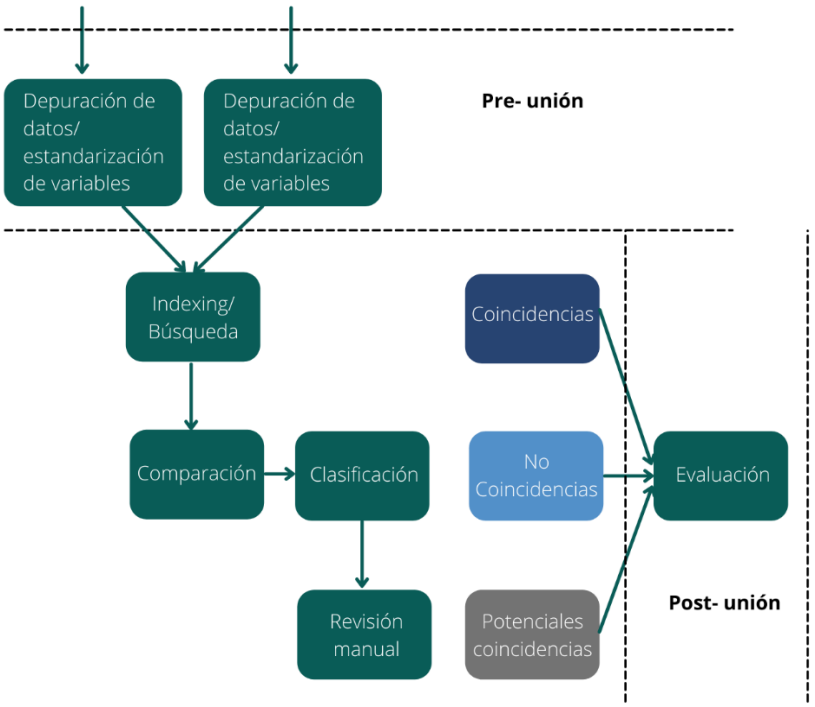
Gráfico 2. Fijación de un único umbral de corte para dividir en dos grupos de pares de casos/filas: no coincidentes y coincidentes.



3) Post-unión. Revisión manual de los casos/filas no vinculados (si se opta por esta estrategia). Evaluación de los resultados.

El diagrama de procesos de la siguiente figura representa esquemáticamente el proceso de unión de registros.

Figura 3. Diagrama resumen del proceso de unión de registros.



Los métodos de aprendizaje automático como árboles de decisión, redes neuronales, aprendizaje basado en ejemplos o instance-based learning, agrupamiento o clustering, entre otros, se utilizan ampliamente para la clasificación de patrones. Un algoritmo de aprendizaje automático construye un modelo a partir de información suministrada en forma de ejemplos, para generalizar comportamientos y reconocer patrones.

Los métodos de aprendizaje automático se clasifican en dos grupos: aprendizaje supervisado (se cuenta con información que especifica qué conjuntos de datos son satisfactorios para el objetivo del aprendizaje) y aprendizaje sin supervisión (encontrar patrones que permitan separar y clasificar los datos en diferentes grupos, en función de sus atributos).

APÉNDICE

Figura 1. Sistema de registros estadísticos por tipo de objeto y campo de estudio

Esta figura representa el sistema de registros estadísticos. Está formada por 4 círculos que representan a cada uno de los 4 registros base: población, inmuebles, empresas y actividades. Estos 4 círculos de cada registro base están vinculados entre sí a través de flechas dobles que representan la unión entre los 4 registros base por medio de claves o llaves de identificación, como son el NIP o Número de Identificación Personal, NIE o Número de Identificación de Empresas, y el código de dirección. El círculo del registro de población está vinculado con doble flecha con el círculo del registro de inmuebles por medio del código de dirección. El círculo del registro de inmuebles está vinculado con doble flecha con el círculo del registro de empresas por medio del código de dirección. El círculo del registro de empresas está vinculado con doble flecha con el círculo del registro de actividades por medio del NIE. El círculo del registro de actividades está vinculado con doble flecha con el círculo del registro de población por medio del NIP, El registro de población se crea y se actualiza a partir de varias fuentes administrativas y estadísticas: Censo de población y vivienda. Encuestas por muestreo de población. Registro de empleo. Registro de educación. Registro de ingresos y tributación. Registro de causas de muerte. Registro longitudinal de seguridad social. Registro para investigación médica. El registro de inmuebles se crea y se actualiza a partir de Bases de datos geográficas, sistemas GIS. Registros de ocupación del suelo, uso del suelo. Registro de la propiedad. Catastro. Permisos de construcción. El registro de empresas se crea y se actualiza a partir de fuentes administrativas y estadísticas, como son: registro del impuesto al valor agregado.

Registro de pago anual. Registro anual de impuestos sobre la renta. Registro de comercio exterior. Registro de explotaciones agrícolas. Encuesta por muestreo de empresas. Encuestas de inversión. El registro de actividades se crea y se actualiza a partir de registros de Salarios y personal. Sistema de seguridad social. Educación obligatoria, alumnos. Registro de estudiantes universitarios.

Figura 2. Proceso de transformación de registros administrativos a registros estadísticos.

Esta figura representa el Proceso de transformación de registros administrativos a registros estadísticos. A la izquierda, como entrada del proceso, se tienen los registros administrativos representados por cilindros, que es la figura que habitualmente representa un archivo o base de datos. Estos cilindros de los registros administrativos están unidos por una flecha con una figura en forma de gran flecha que representa el Proceso de transformación de registros administrativos a registros estadísticos. Y ésta a su vez está unida con un cilindro más grande a la derecha como salida del proceso, que representa a los registros estadísticos que se generan a partir del proceso

de transformación. Dentro de la figura con una gran flecha que representa al Proceso de transformación de registros administrativos a registros estadísticos, se describen las actividades que se ejecutan en este proceso de transformación: Controles de consistencia y calidad. Depuración de datos. Estandarización de variables. Creación de variables agregadas/derivadas. Creación de objetos/unidades derivadas. Unión de registros. Documentación de metadatos, y Aseguramiento de la calidad.

Gráfico 1. Umbrales que delimitan las zonas correspondientes a los tres grupos de pares de casos/filas: no coincidentes, posibles coincidencias y coincidentes.

Este gráfico representa los Umbrales que delimitan las zonas correspondientes a los tres grupos de pares de casos o filas: no coincidentes, posibles coincidencias y coincidentes. El eje vertical del gráfico representa la frecuencia de pares de casos o filas y el eje horizontal los pesos de los pares de casos o filas. Se tienen dos figuras similares a campanas que se intersectan. La campana de la izquierda representa a los casos no coincidentes, la campana de la derecha representa a los casos coincidentes. En medio de las dos campanas se trazaron dos líneas verticales que incluyen el área de la intersección entre ambas, para establecer dos umbrales y determinar los casos que posiblemente sean coincidentes.

Gráfico 2. Fijación de un único umbral de corte para dividir en dos grupos de pares de casos/filas: no coincidentes y coincidentes.

Este gráfico representa la forma en cómo se podría fijar un único umbral de corte para dividir en dos grupos de pares de casos/filas: no coincidentes y coincidentes. Se tiene el mismo gráfico anterior. El eje vertical del gráfico representa la frecuencia de pares de casos o filas y el eje horizontal los pesos de los pares de casos o filas. Se tienen dos figuras similares a campanas que se intersectan. La campana de la izquierda representa a los casos no coincidentes, la campana de la derecha representa a los casos coincidentes. Por el punto de intersección entre ambos gráficos se traza una línea que establece el umbral de corte. El área debajo de la gráfica de los casos no coincidentes que está a la derecha de la línea o umbral de corte representa a los falsos positivos y el área debajo de la gráfica de los casos coincidentes que está a la izquierda de la línea o umbral de corte representa a los falsos negativos.

Figura 3. Diagrama resumen del proceso de unión de registros.

Diagrama resumen del proceso de unión de registros. Es un diagrama de flujo que inicia con dos cilindros que representan dos archivos de entrada que se quieren unir. Cada uno de estos dos cilindros están unidos a un rectángulo que representa el subproceso de Depuración de datos y estandarización de variables, que sería la fase de pre-unión. A su vez, estos rectángulos están unidos al rectángulo que representa el subproceso de indexing y búsqueda, y éste se une con el rectángulo del subproceso de comparación que a su vez está unido al rectángulo de Clasificación. Del rectángulo del subproceso de clasificación salen 4 flechas, que están unidas con los rectángulos de los subprocesos de Coincidencias, No Coincidencias, Potenciales Coincidencias y Revisión Manual. El rectángulo de revisión manual se une con el rectángulo de potenciales coincidencias. Hasta aquí tenemos los subprocesos incluidos en la fase de unión. Los 3 rectángulos de coincidencias, no coincidencias y potenciales coincidencias se unen al rectángulo del subproceso de evaluación, que forma parte de la fase de post-unión. De este rectángulo del subproceso de evaluación sale una flecha que vuelve al subproceso de Clasificación, en el caso que se evalúe necesario volver a clasificar.