ACTUALIZACION DEL SISTEMA DE CODIFICACIÓN INFORMATIZADA (SICI)

Agosto 2019

En la Coordinación de Clasificadores y Nomenclaturas de la Dirección Nacional de Metodología Estadística del INDEC nos encontramos en un proceso de actualización del Sistema de Codificación Informatizada (SiCI) en su versión original, éste fue un desarrollo interdisciplinario[1] del Instituto. El mismo se utilizo por primera vez en oportunidad de codificar el Censo Nacional de Población y Vivienda 2001, y se utiliza en la actualidad de forma constante para la codificación de la Encuesta Permanente de Hogares (EPH) así como de diversos relevamientos que se presenten.

En este momento nos encontramos desarrollando la actualización del SiCI, enmarcado en los preparativos del próximo Censo Nacional de Población y Vivienda (CNPyV) 2020. Las actualizaciones que se están desarrollando involucran distintos aspectos del sistema:

1.- *incorporación de nuevos clasificadores al SiCI*, que están relacionadas concretamente al relevamiento censal y creados ad hoc:
- Clasificador de Carrera universitaria y terciaria no universitaria
- Clasificador de pueblos originarios

2.- Desde su origen y hasta la actualidad el SiCI fue desarrollado sobre la plataforma informática que utiliza el programa Clípper y se comenzó con su *migración a SAS*. Esta tarea se encuentra en pleno proceso de desarrollo y no se cuenta con resultados para compartir.

Y por ultimo y es el aspecto en que nos vamos a centrar en esta presentación:
3.- en la *aplicación de un clasificador distinto* para la codificación de rama de actividad económica, que no es el usado habitualmente en el INDEC para los relevamientos sociodemograficos.

El INDEC, dispone de dos clasificadores para la medición de la rama de actividad de los establecimientos, la Clasificación de la Actividad Económica para encuestas Sociodemograficas del MERCOSUR (CAES); y la Clasificación Nacional de Actividades Económicas (ClaNAE)[2], ambas

[1] Con la participación de la Dirección de Informática, Metodología estadística y la Coordinación de Clasificadores del INDEC durante los años 1998 y hasta el 2003.

[2] La CAES es un instrumento creado para ser utilizado por los países del MERCOSUR (Uruguay, Paraguay, Brasil y Argentina) que mantiene una correspondencia con la CIIU, revisión 3.1 en su primer versión y con la revisión 4 en su versión actual (CAES 1.0), es el instrumento que se utiliza para la codificación de la EPH, ENGHO y los censo de población desde el 2001. Este es un clasificador mucho más agregado que el utilizado en el INDEC para la codificación de los establecimientos en los relevamientos económicos que es la Clasificación Nacional de Actividades Económicas (ClaNAE). Existe una tabla de correspondencia entre los dos clasificadores nacionales y la clasificación internacional CIIU.

adaptadas de la Clasificación Industrial Internacional Uniforme de todas las actividades económicas (CIIU).
En el contexto de los preparativos del próximo CNPyV 2020 y a partir de una necesidad de la Dirección Nacional de Cuentas Nacionales del INDEC de contar con información de rama mas desagregada en algunos grupos, principalmente en los industriales, que la que le provee la información codificada con CAES se decide el uso de ClaNAE en lugar de CAES para este operativo.

La ClaNAE no se utilizaba en el SiCI desde el año 2004 (fecha del último censo económico), hubo que comenzar a actualizar el clasificador y en primer lugar reemplazar la ClaNAE 2004, que fue la versión utilizada en esa oportunidad y se corresponde con la CIIU Rev. 3.1 por la ClaNAE 2010 que remite a la CIIU Rev. 4. Por lo cual había que proceder a la actualización del SiCI, que recrea a través de diferentes métodos, todo el conjunto de procesos intelectuales que el codificador realiza cuando lee, interpreta, analiza y coloca el código a la frase que tiene delante de él.

Resumidamente[3] señalamos que este sistema se compone de tres elementos o procesos básicos:

1.-Proceso Diccionarios
2.-Procesos lingüísticos
3.-Procesos de codificación (diccionario de codificación)

1.- Los diccionarios: son listados inventariados de palabras o frases que conforman los instrumentos fundamentales del SiCI y que se originan en las respuestas empíricas relevadas en cada uno de los operativos que sirvieron de fuente. En el sistema conviven dos tipos de diccionarios: los que sirven para el tratamiento de las palabras y los diccionarios de codificación que conforman los procesos de codificación.
Estos elementos o proceso Diccio, crea una múltiple variedad de diccionarios que alimentan al SiCI. Es en la ejecución de este proceso donde se logro explicitar los pensamientos y decisiones que toma un codificador para la ejecución de su tarea y que sirvieron de modelo en el desarrollo del SiCI.

---

[3] El SiCI fue presentado en la V reunión del GTCI-CEPAL realizada en Panamá en el 2016 y se puede encontrar en el documento compilado por la Coordinación del GTCI en: "**DOCUMENTOS COMPARTIDOS POR ARGENTINA, EL SALVADOR, HONDURAS Y MÉXICO, SOBRE EL SISTEMA DE CODIFICACIÓN AUTOMATIZADO O ASISTIDO PARA CLASIFICAR OCUPACIONES O ACTIVIDADES**"

El objetivo de este proceso es el de armar los elementos que conforman los distintos diccionarios del Sistema. Tenemos:
a.- los **Diccionarios de frases**:
a.1.- *frases originales* que son las contenidas en los archivos fuente, a los que se normaliza, se eliminan espacios en blanco, signos, deja todo en mayúscula, y el
a.2.- de *frases corregidas codificadas*, una vez realizada la corrección ortográfica, quedan las frases correctas y son codificadas.

b.- Los **diccionarios de tratamiento de palabras** surgen de la descomposición de las frases y según su rol en la frase forman parte de los siguientes diccionarios:

- De palabras espurias
- De palabras anuladas
- De conectores
- De excepciones
- Corrector
- De palabras correctas
- De lectura

Estos toman las frases provenientes del campo y "acondicionan" la información de acuerdo a la función que cumple cada uno de ellos para que puedan ser utilizadas por el diccionario de codificación.

Al utilizar un diccionario para la codificación, se busca que la base a codificar y la del diccionario se asemejen lo más posible. Es necesario, entonces, armar los diccionarios con registros provenientes de campo, que se respondan en forma similar a la base que se pretenda codificar. Por ejemplo, si se quiere codificar una encuesta sociodemografica es conveniente utilizar principalmente registros que provengan de programas del mismo ámbito. Ello no implica que no se puedan utilizar registros que provengan de encuestas del área económica, y mucho menos que éstas no aporten nada al diccionario, sino que es más probable que las respuestas de dos encuestas de la misma área sean más parecidas. Además, se pueden realizar diccionarios paralelos, es decir, crear un diccionario que utiliza registros de una fuente para codificar cierto tipo de encuestas y otro que se compone de registros de otras fuentes para codificar otro tipo de encuestas.

2.- Los procesos lingüísticos: son aquellos que modifican los literales de las frases a codificar, permitiendo una simplificación del vocabulario y de la cantidad de palabras involucradas. Con literales o descriptores nos referimos a la frase que representa la respuesta original brindada por el informante.
Entre los procesos lingüísticos que operan sobre los literales nos encontramos con:

- *Proceso de normalizado*: consiste en sacar los caracteres no válidos que se encuentran en las frases de la base recibida con las tres variables (actividad, ocupación y tarea) y se convierten a mayúscula.

- *Campos semánticos o familiarizado*: consiste en asignar a una palabra tomada como referencia (denominada padre), una lista de palabras que serán tomadas como sinónimos (denominado hijos).

- *Proceso de estandarizado*: consiste en tratar todas las palabras del diccionario por número, género y truncamiento, según lo que sea más apropiado, a los efectos de lograr un diccionario de términos únicos (no repetitivos)

Hasta el momento, el SiCI se desarrolló fundamentalmente con la utilización de dos métodos de codificación que definen el **proceso de codificación** del Sistema:

a.-"microprocesos": en este caso se identifica una palabra clave que permite desarrollar un microprocesos. Las palaras clave son representativas de ocupaciones, a partir de los cuales se los puede reconocer y codificar independientemente de la posición en que se encuentre en la frase de donde se la eligió. Elegida la palabra, el microprocesos que genera determinara como se codificaran las descripciones que la contengan. Este es el método utilizado fundamentalmente para la codificación de ocupaciones; y

b.- el de "frases únicas", es un método de codificación automático o directo que permite la asignación de un código único sin intervención de los codificadores. Para esto, utiliza un diccionario de codificación formado exclusivamente por frases que ofrecen una única alternativa de código y que son independientes de las restantes variables del cuestionario. Es el método utilizado en la codificación de rama de actividad.

Estos métodos han demostrado ser fructíferos, con alrededor del 68 % de los casos codificados automáticamente. Los casos que no lo son, se someten a una revisión para tratar de obtener nuevas frases o palabras, según corresponda, para los respectivos diccionarios. Los dos métodos mencionados son deterministas: una vez detectada una frase o palabra clave, el SiCI aplica los diccionarios o microprocesos que llevan a tratar todos estos casos con una resolución preestablecida. Si se aplica el método de autofrase, la frase a codificar recibe el código que ya tiene asignado en el diccionario de codificación; si se aplica el método de microprocesos, se dispara el correspondiente al mismo (que tiene una o varias posibilidades de código, previamente asignadas).

La única manera de modificar o agregar nuevas autofrases o microprocesos es a través de la intervención directa y externa de los encargados de los clasificadores y de los sistemas de codificación; estos métodos no se corrigen a sí mismos.

Teniendo en cuenta que los diccionarios del sistema se nutre de la aplicación de un conjunto de frases anteriormente codificadas, de tal forma que aquellos casos que se repitan, se resuelvan de la misma manera y tal como se mencionara anteriormente, la utilización del Sistema con ClaNAE contaba con poco desarrollo y las frases que codificaban el SiCI estaban desactualizadas (solo se había utilizado en el 2004 y disponía de poca acumulación de frases previamente codificadas).

Por lo tanto la actualización, del Sistema consiste fundamentalmente en la incorporación de una nueva versión de un clasificador[4] con las tareas que se señalan a continuación:

1.- se incorpora la ClaNAE 2010 al SiCI

2.-dado que, la única base codificada a través del SiCI con ClaNAE como ya se mencionara estaba muy desactualizada y provenía de un relevamiento económico, se tomo el diccionario de codificación generado con la codificación con CAES en distintos relevamientos sociodemograficos, especialmente la EPH y se recodificaron las 30.000 "frases únicas" que conforman el diccionario de codificación con ClaNAE 2010.

3.- Con las frases únicas recodificadas se aplico el sistema a una base de la EPH de 270.000 casos conformada por varios trimestres de la Encuesta a modo de prueba. Observando que la codificabilidad automática entre un instrumento y otros (CAES y ClaNAE) variaba sustantivamente, codificando con CAES la parte automática era de un 65% y con ClaNAE de un 53%. Al analizar estos parámetros observamos que el proceso lingüístico de familiarización de las frases únicas tal cual estaba planteado, no servía para la ClaNAE.

La ClaNAE tiene una desagregación mucho más amplia que CAES, esta mayor desagregación se da, en varios grupos del clasificador. Por ejemplo: CAES no distingue entre comercio mayorista y minorista como si lo hace la ClaNAE; CAES no diferencia entre comercios de distintos tipos de alimentos cosa que si hace la ClaNAE, por lo cual con la familiarización desarrollada para un clasificador no servía, había que revisar muchos casos de la misma. Por ejm: la frase "venta de alimentos en una despensa", o "pescadería", con la familiarización se transformaban en "comercio de alimentos" con la utilización de CAES se podía codificar sin problema, ya que en ese clasificador el código no diferencia entre un comercio u otro. Sin embargo esta situación no se replica al utilizar la ClaNAE que si los diferencia, tal como se observa en la tabla que se presenta a continuación.

[4] Reemplazar ClaNAE 2004 por ClaNAE 2010.

ClaNAE		CAES
47	**COMERCIO AL POR MENOR, EXCEPTO EL COMERCIO DE VEHÍCULOS AUTOMOTORES Y MOTOCICLETAS**	**48**
47111	Venta al por menor en hipermercados	**4808**
47112	Venta al por menor en supermercados	**4808**
47113	Venta al por menor en minimercados	**4808**
47119	Venta al por menor en kioscos, polirrubros y comercios no especializados n.c.p.	**4808**
47190	Venta al por menor en comercios no especializados, sin predominio de productos alimenticios y bebidas	**4809**
47211	Venta al por menor principalmente de fiambres, quesos y productos lácteos	**4803**
47212	Venta al por menor de productos de almacén y dietética	**4803**
47213	Venta al por menor de carnes rojas, menudencias y chacinados frescos	**4803**
47214	Venta al por menor de huevos, carne de aves y productos de granja y de la caza	**4803**
47215	Venta al por menor de pescados y productos de la pesca	**4803**
47216	Venta al por menor de frutas, legumbres y hortalizas frescas	**4803**
47217	Venta al por menor de pan y productos de panadería y confitería	**4803**
47219	Venta al por menor de productos alimenticios n.c.p. en comercios especializados	**4803**

Por lo tanto hubo que anular esas familiarizaciones para tomar las frases diferenciando los tipos de comercio tal como venían de campo.

Por otro lado, comprobamos que el método de frases únicas tenía un techo en lo referente a aumentar los niveles de la codificación automática y se comenzó a desarrollar otro método para la codificación de rama que complementase la codificación de frase única.

En este momento se están desarrollando dos nuevos métodos para la codificación automática que serán incorporados al SiCI, éstos son: **1.- por contrastación de palabras y frases, y 2.- por la aplicación de arboles de clasificación.**

1.- En este caso se genera una **contrastación de palabras y frases** de una fuente de datos determinada a codificar, con diccionarios construidos en base a otras fuentes de datos con palabras y frases de textos ya codificados.

Este método requiere por un lado de esa **fuente de datos o archivo de input**, en los que encontramos textos libres, que es la información proveniente de campo y la que se va a codificar; y por otro lado **archivos con textos ya codificados**. Estos dos archivos son los que serán contrastados mediante algoritmos de text mining.

También se elaboran **diccionarios,** conformados por palabras o frases provenientes de distintos relevamientos previos ya codificados asociados a un código ClaNAE .
En nuestro caso funcionan como diccionarios:

a.- Censo Nacional Económico (CNE) 2004
b.- ClaNAE 2010
c.- CPC 2.1[5]
d.- Bases de la EPH[6]

Los archivos de inputs en primer lugar son sometidos a una serie de programas de corrección de palabras y de verbos. Se trata de manera diferenciada los textos con verbos que denotan servicios, o producciones que en verdad son servicios, por ejemplo: producción de audiovisuales, explotación de marcas, etc.

Una vez depurado, se realiza una partición de la base en función de los verbos que utiliza y a cada registro se lo ubica en cada una de ellas:

p: producción
v: venta
s: servicio
t: transporte

Estas particiones se realizan para los inputs y para los diccionarios, las mismas no son excluyentes, ya que algunos códigos de actividad quedan en más de una partición tanto en los diccionarios como en el input, esto es así para no restringir demasiado al punto de que no se encuentren registros de palabras compatibles con el input (lo que daría un match incorrecto al no encontrarse la palabra evaluada del input en el diccionario).

El texto no es leído ni interpretado semánticamente sino que es analizado informáticamente, entonces si aparecen en un texto verbos que pertenecen a más de una partición en su concepto, la depuración inicial hará que se pueda contrastar con todas las particiones correspondientes de los diccionarios, luego los otros procesos siguientes resolverán esa ambigüedad informática. Por

[5] En este caso se tomo el CNE 2004 y el CPC 2.1 porque lo primeros avances en este desarrollo se hicieron para aplicarlos a un próximo censo económico, posteriormente se comenzó a trabajar para el censo de población.
[6] Estas bases de EPH debieron ser recodificadas de CAES a ClaNAE 2010.

ejemplo podría decir “venta para la producción” y ser efectivamente una venta, aunque la producción se priorice sobre la venta. No se establecen particiones excluyentes sino que están hechos a los efectos de que no se puedan evaluar palabras en códigos imposibles de corresponder.

Los diccionarios tienen los sustantivos, verbos y ponderadores para cada uno de ellos, estos ponderadores están relacionados con la frecuencia con la que ocurrieron para cada código en el relevamiento que generó ese diccionario, es decir a mayor frecuencia mayor ponderación, lo que permite priorizar las palabras más frecuentes para la clasificación y a su vez permite eliminar las palabras ocurridas con muy poca frecuencia o mal escritas.

Una vez que la fuente de datos y los diccionarios fueron particionados se realiza la contrastación de cada partición del input con su partición equivalente en el diccionario y la ClaNAE que forma uno de los diccionarios, esta contrastación va a generar un score basado en las similitudes de las palabras (algoritmos de text mining) y en las frecuencias de aparición de las palabras y los verbos (excepto en la ClaNAE donde no hay frecuencias de ocurrencia por relevamiento pero en cambio se contrastarán también las frases completas de sus definiciones).

Si ningún código tuviera una puntuación aceptable para ningún diccionario ese registro quedara pendiente de codificar.

Cada registro de la tabla de input será contrastado con cada registro de un diccionario, lo que se denomina un producto cartesiano entre registros de tablas o un ‘todos contra todos’, tanto entre registros como entre palabras. Si una o más palabras perteneciente a un texto del input existe en un diccionario para un código específico o la diferencia de caracteres entre estas es mínimo, entonces ese texto será candidato a tener una asignación a ese código ClaNAE.

Conceptualmente y sintéticamente así funciona este método.

El otro método que estamos incorporando, por la aplicación **de árboles de clasificación.** En este caso consideramos dos archivos fuente, uno codificado con ClaNAE (por ejm. bases de EPH) cuyo rol es **de archivo de entrenamiento**, donde el algoritmo del árbol aprenderá en forma inductiva a través de las palabras ocurridas en los textos ya clasificados.
El otro **archivo fuente es el llamado de testing** donde vamos a aplicar los enunciados que el árbol haya desplegado aprendiendo del archivo de entrenamiento.

Este es un algoritmo que se despliega en forma descendente sin volver hacia atrás una vez elegido un camino, parte de un nodo inicial al que considera como el más discriminante, desde donde encuentra que puede partir el espacio de búsqueda inicialmente en dos, luego sigue en forma binaria dividiendo el espacio de las palabras más discriminantes, tanto en forma positiva como en forma negativa hasta completar un valor de parada o detención establecido previamente. Establece así una secuencia lógica de opciones, como proposiciones o enunciados, que llevan a definir cada código ClaNAE, de a dos, en forma binaria, cada vez por sí o por no, en función de

cómo ve que un código se va definiendo mediante palabras ocurridas en los registros que analiza, palabras que aportan y las que no aportan a cada código. Por ejemplo… el árbol puede llegar a la palabra cuaderno y a partir de ahí observar que hay otra palabra en el espacio de búsqueda, que es Enseñanza, que lo acerca a un código de Servicios de Educación. Sin embargo, si lo que va a clasificar por SI o por NO es una Venta en librería, el hecho de que encuentre la palabra Enseñanza actúa de discriminante negativa y lo va expresar de esa manera.

Todo se genera a partir de leer los registros del archivo de entrenamiento recorriendo las palabras y viendo cómo se forman los códigos a los que estas palabras están relacionadas. De esta forma se desplegara el árbol, como grandes enunciados, por un camino u otro.

Para conformar el archivo de entrenamiento, dado que en general hay una gran cantidad de palabras, entendemos que es importante, del mismo modo que en el otro método, genera un recorte por frecuencia de ellas, para contar únicamente con aquellas más significativas.

Siempre se debe considerar el sistema y los recursos físicos con los que se cuenta (memoria dedicada en el servidor especialmente). Teniendo conocimiento de las posibilidades se decidirá generar mayores o menores recortes del archivo de entrenamiento. Asimismo puede ser que se necesite recortar el archivo en cuanto a los registros, para esto debemos considerar contar con un mínimo aceptable de registros para cada código ClaNAE.

En el ambiente de testing se consideran las mismas palabras que ocurrieron en el archivo de entrenamiento, para poder aplicar el árbol. No todas las palabras estarán en ambos archivos. Lo importante es que estén las que el árbol encuentra como relevantes en la determinación de sus enunciados en el entrenamiento. De modo que se realiza una depuración para tener en el proceso de testing esas palabras, y las que no se presentan en testing quedarán como no ocurridas entonces luego no todos los enunciados del árbol podrán ser ejecutados.

Estando en proceso de testing, se va sucesivamente clasificando por cada código por sí o por no. A los registros clasificados por NO en el primer código a evaluar, se los toma para aplicarles el siguiente árbol de entrenamiento, su secuencia de enunciados, para la determinación del siguiente código a evaluar, también por SI o por NO, y así sucesivamente hasta terminar de evaluar todos los códigos.

La secuencia elegida de códigos para realizar estas operaciones es en orden de frecuencia de casos obtenidos en el relevamiento adoptado como entrenamiento. Por lo tanto el árbol comenzara a trabajar con códigos de poca ocurrencia en el entrenamiento y terminara con el de mayor ocurrencia. Cuantas más ocurrencias tenga, más material para evaluar la secuencia de enunciados tendrá.
Una vez terminado este proceso, tenemos a todos los registros clasificados con algún código en función de la aplicación de estos árboles binarios. Cada registro tendrá su código ClaNAE asignado.

En general, es importante que el árbol no sea demasiado particular para el archivo que está trabajando, sino que exista algún nivel de generalización, lo que se opera pidiendo que corte el proceso de despliegue en un punto determinado. Este punto se irá evaluando. Es lo que se conoce como “poda del árbol”.